RAIRO. RECHERCHE OPÉRATIONNELLE

M. L. TRICOT
M. DONEGANI

# Formules de réactualisation pour une famille d'indices de proximité inter-classe en classification hiérarchique

*RAIRO. Recherche opérationnelle*, tome 23, n° 2 (1989), p. 165-192

<http://www.numdam.org/item?id=RO_1989__23_2_165_0>

Recherche opérationnelle/Operations Research
(vol. 23, n° 2, 1989, p. 165 à 192)

# FORMULES DE RÉACTUALISATION POUR UNE FAMILLE D'INDICES DE PROXIMITÉ INTER-CLASSE EN CLASSIFICATION HIÉRARCHIQUE (*)

par M. L. TRICOT (1) et M. DONEGANI (1)

Résumé. — *Dans cet article, nous proposons un cadre mathématique général, englobant beaucoup d'indices de proximité inter-classe connus en classification hiérarchique. Nous en déduisons d'autres également intéressants. Nous reprenons les travaux de G. N. Lance et W. T. Williams, nous les généralisons en proposant une nouvelle formule de réactualisation s'appliquant parfaitement au cadre mathématique proposé. Nous étudions sur un exemple les conséquences pratiques de ces résultats.*

Mots clés : Classification hiérarchique, Formule de Lance et Williams, Indice de proximité inter-classe.

Abstract. — *We propose in this article a general framework which encompass many of the known inter-set proximities hierarchical clustering. It allow, by deduction, the developpement of other equally interesting indices. Developped from the works of G. N. Lance and W. T. Williams, a new generalized updating formula is also discussed. An example is studied.*

Keywords : Hierarchical Classification; Lance-William agglomerative formula; clustering criteria.

## 0. INTRODUCTION

Nous discutons dans cet article de certains problèmes relatifs à quelques algorithmes ascendants de classification hiérarchique.

On suppose donné un tableau $X$, $n \times p$, de réels (ce sont parfois des codes chiffrés) et un indice de dissimilarité $d$ sur les lignes $X_i$ de $X$. (Il existe également des cas où l'indice de dissimilarité est défini sur les colonnes : on considèrera alors le tableau $X^t$).

(*) Reçu décembre 1987.
(1) École Polytechnique Fédérale de Lausanne, Département de Mathématiques, MA (Ecubleus) CH-1015 Lausanne, Suisse.

Recherche opérationnelle/Operations Research, 0399-0559/89/02 165 28/$ 4.80

Un algorithme ascendant de classification hiérarchique permet d'identifier au plus $n$ partitions de l'ensemble des $X_i$, $P_1, \ldots, P_n$, telles que $|P_i| = i$ et que la suite des $P_i$ soit emboitée. Cet algorithme procède de la manière suivante :

Notons $\mathbf{P}(X)$ l'ensemble des parties de $X$.

On se donne un indice $D$ tel que :

$$D: \quad \mathbf{P}(X) \times \mathbf{P}(X) \to \mathrm{IR}$$

appelé *« indice de proximité inter-classe »*. On définit alors les $P_n$ de la manière suivante :

(i) Etape 1 : $P_n = \{X_i\}_{i=1}^n$, $k = 2$.

(ii) Etape $k$ : On cherche deux sous-ensembles $C_i$ et $C_j$ de $P_{k-1}$ tels que :
$D(C_i, C_j)$ soit minimal.
$P_k$ est obtenue en réunissant ces deux classes.

(iii) Si $k = n$, on arrête; sinon on passe à l'étape $k+1$.

Dans certaines applications, il peut arriver que la distance minimale soit réalisée par plusieurs sous-ensembles. Nous ne discuterons pas de ce cas, il a été traité en particulier par I. C. Lerman [11].

Un des principaux problèmes posé par cette classification hiérarchique est le choix de $D$. Nous proposons ici de réunir un grand nombre de tels indices dans une seule formule à paramètres (partie 1). En donnant une valeur précise à ces derniers, nous retrouvons des indices déjà connus.

L'intérêt d'une telle démarche est à notre avis double :

D'une part, elle donne une meilleure unité mathématique au sujet et met ainsi en évidence les principes fondamentaux qui ont guidé l'obtention de ces divers indices.

D'autre part, cette formule autorise la construction de nouveaux indices, d'interprétation plus simple ou encore mieux adaptés à des situations particulières.

Notre famille a également l'intérêt de s'adapter à n'importe quel indice de dissimilarité défini *a priori* sur le tableau de données initial. Elle permet également de traiter un tableau de distance.

Toutefois, une des premières propriétés requises pour un indice de proximité inter-classe est qu'il possède une formule de réactualisation. Rappelons qu'une telle formule permet de déduire très simplement le tableau de proximité inter-classe à l'étape $k$ de celui obtenu à l'étape $k-1$, ce qui diminue considérablement le temps de calcul.

Il existe, dans la littérature, et pour les indices auxquels nous nous intéressons ici, deux formules de ce type : la formule de Lance et Williams et la formule de Jambu. Nous donnerons les conditions nécessaires et suffisantes pour qu'un indice de notre famille soit Lance et Williams ou Jambu (partie 2). Il est remarquable que ces conditions soient tout à fait indépendantes de l'indice de dissimilarité utilisé.

Ceci nous amènera dans la partie 3 à définir une nouvelle formule de réactualisation dont nous étudierons les propriétés.

Finalement, dans la partie 4, nous appliquerons nos résultats à un jeu de données bien classique, les iris de Fisher.

Avant de passer à la suite, une petite précision de vocabulaire : dans ce texte, le terme de « classe » sera utilisé au sens de « sous-ensemble de $X$ », par référence au contexte de la classification. D'autre part, on notera $\| \quad \|_2$, la distance euclidienne classique dans $\mathrm{IR}^p$.

## 1. LES INDICES L

### 1.1. Définition

Le but fondamental de la classification automatique étant d'obtenir des classes bien homogènes et bien séparées, l'indice de proximité inter-classe utilisé en classification hiérarchique doit permettre de mesurer *simultanément :*

- le niveau de *séparation* de 2 classes
- le degré d'*homogénéité* de chacune de ces 2 classes.

Les méthodes de construction les plus répandues de tels indices sont basées sur les principes généraux suivants.

(i) On choisit un ensemble représentatif de chaque classe (« noyaux » de E. Diday).

(ii) On calcule un coefficient qui mesure la distance entre ces noyaux.

(iii) On calcule un coefficient qui mesure la dispersion de chacun de ces ensembles représentatifs.

(iv) Finalement on construit une fonction des trois coefficients ainsi obtenus, qui reflète bien la notion susdite de proximité inter-classe.

*Exemple.* — On choisit comme ensemble représentatif la moyenne $\bar{X}$ et $\bar{Y}$ de chaque classe. La dispersion d'un tel ensemble est nulle puisqu'il est réduit à un seul point. La proximité entre les deux ensembles représentatifs que l'on

fera coïncider avec la proximité inter-classe sera :

$$\| \bar{X} - \bar{Y} \|_2^2$$

On observe en pratique que les calculs d'ensembles représentatifs de classes sont généralement coûteux en temps de calcul, sans que cela rapporte grand'-chose au niveau des résultats. Ils se ramènent d'ailleurs souvent à un nouveau problème de classification. Il existe seulement deux exceptions :

1. L'ensemble représentatif est réduit à un point : médiane, moyenne, ...
2. L'ensemble représentatif est la classe tout entière.

Nous nous plaçons dans ce deuxième cas pour définir les indices L. Nous constaterons que dans le cadre de la norme $L2$, et lorsque l'ensemble représentatif est la moyenne, nous pouvons ramener le 1$^{er}$ cas à celui-là.

DÉFINITION 1.1.1 : Soit deux classes $C_i$ et $C_j$. On appellera « *indice L de proximité inter-classe* » tout indice de la forme :

$$\alpha_{ij} \Delta_{ii} + \beta_{ij} \Delta_{ij} + \gamma_{ij} \Delta_{jj}$$

où $\Delta_{ij}$ est défini par

$$\Delta_{ij} = \sum_{\substack{k \in C_i \\ l \in C_j}} d(X_k, X_l), \qquad \forall\, i, j.$$

$\alpha_{ij}$, $\beta_{ij}$, $\gamma_{ij}$ sont des réels et $d$ est un indice de dissimilarité quelconque définie sur les lignes de $X$.

On voit que $\Delta_{ii}$ mesure la dispersion de la classe $C_i$. Rappelons à ce propos que :

$$\Delta_{ii} = 2 n_i \sum_{i \in C_i} \| X_i - \bar{X} \|_2^2 \quad \text{lorsque } d \text{ est la distance euclidienne au carré.}$$

La quantité $\Delta_{ii}$ s'interprète naturellement comme une généralisation de la variance intra-classe classique à un indice d'inhomogénéité quelconque.

La quantité $\Delta_{ij}$ s'interprète aussi comme une mesure de la séparation entre les deux classes $C_i$ et $C_j$.

Finalement l'indice de proximité inter-classe est une combinaison linéaire de ces trois quantités, ce qui explique la dénomination « $L$ ».

### 1.2. Propriétés des coefficients

Nous précisons maintenant des conditions élémentaires sur les coefficients afin qu'un indice $L$ décrive correctement la notion de proximité inter-classe.

1. Les coefficients doivent être d'expression simple afin que l'indice soit aisément interprétable dans la pratique.

2. L'indice doit être d'autant plus grand que les classes $C_i$ et $C_j$ sont plus éloignées, c'est-à-dire que $\Delta_{ij}$ est plus grand. On prendra donc $\beta_{ij} > 0$.

L'indice doit être d'autant plus petit que les classes $C_i$ et $C_j$ sont plus homogènes, c'est-à-dire $\Delta_{ii}$ et $\Delta_{jj}$ sont plus petits. On prendra ainsi $\alpha_{ij}$ et $\gamma_{ij} < 0$.

3. Les trois coefficients doivent être de même « dimensionnalité » par rapport à $|C_i| = n_i$ et $|C_j| = n_j$.

4. L'effet homogénéité $\alpha_{ij}\Delta_{ii} + \gamma_{ij}\Delta_{jj}$ doit avoir le même poids que l'effet séparation $\beta_{ij}\Delta_{ij}$.

5. Si $n_i = n_j = 1$, l'indice doit valoir $d(X_i, X_j)$. Car dans ce cas $C_i = \{X_i\}$ et $C_j = \{X_j\}$.

6. Les coefficients sont fonction de la taille des classes $C_i$ et $C_j$.

7. $\beta_{ij} = \beta_{ji}$ et $\alpha_{ij} = \gamma_{ji}$, $\quad \forall\ i, j.$

Les trois premiers exemples qui s'imposent sont :

1. $$-\frac{1}{2}\Delta_{ii} + \Delta_{ij} - \frac{1}{2}\Delta_{jj}$$

2. $$-\frac{1}{2n_i}\Delta_{ii} + \frac{2}{n_i + n_j}\Delta_{ij} - \frac{1}{2n_j}\Delta_{jj}$$

3. $$-\frac{1}{2n_i^2}\Delta_{ii} + \frac{1}{n_i n_j}\Delta_{ij} - \frac{1}{2n_j^2}\Delta_{jj}$$

On les interprète de la manière suivante :

Pour le premier indice, chaque terme de la somme est le total des distances entre les points.

On rajoute le coefficient 1/2 pour tenir compte de la propriété 4.

Pour le deuxième indice, $1/n_i\Delta_{ii}$ peut s'écrire

$$\frac{1}{n_i}\sum_{l=1}^{n_i} d(X_l, C_i)$$

où $d(X_l, C_i) = \sum_{j=1}^{n_i} d(X_l, X_j)$. Ce dernier terme est la somme des distances de chaque point à la classe toute entière. On rajoute le coefficient 1/2 pour tenir compte de la propriété 4. D'autre part on pondère le terme du milieu par la moyenne arithémétique de $n_i$ et $n_j$ afin de conserver la symétrie.

Pour le troisième indice, chaque terme de la somme est la moyenne des distances entre les points. Par ailleurs, on constate aisément que si $d = \| \quad \|_2$, on retrouve l'indice des centres de gravité.

On voit que le premier indice et, dans une moindre mesure le deuxième, privilégient l'agrégation des petites classes alors que le troisième indice est insensible à « l'effet taille » des classes qu'il réunit.

*La plupart des indices classiques que l'on rencontre dans la littérature sont des indices L construits selon les règles énoncées plus haut ou certaines d'entre elles : indice de la moyenne des dissimilarités inter-classe, du total des dissimilarités inter-classe, de la moyenne sur l'ensembles des paires fusionnées, de l'inertie, des centres de gravité, de la variance, de l'augmentation d'inertie (Ward)...*

Toutefois il faut remarquer que certains indices bien connus n'appartiennent pas à la famille des indices $L$ : ce sont par exemple l'indice des centres de gravité pour la norme $L1$ ou encore les indices de Lerman dans l'algorithme de la vraisemblance du lien.

## 2. LES FORMULES DE RÉACTUALISATION ET LES INDICES L

### 2.1. La formule de Lance et Williams

C'est en 1967 que Lance et Williams ont montré que bien des indices de proximité inter-classe existant dans la littérature satisfaisaient à une formule de réactualisation simple :

$$\forall\, C_i, C_j, C_k \subset X, \exists \text{ des réels } E, F, G, H \text{ tels que :}$$

$$D(C_i \cup C_j, C_k) = ED(C_i, C_j) + FD(C_i, C_k) + GD(C_j, C_k) + H\,|D(C_i, C_k) - D(C_j, C_k)|.$$

On exprime ainsi la proximité entre les classes $C_i \cup C_j$ et $C_k$ comme une combinaison linéaire des proximités entre chaque couple de classes. On sait que cette formule permet de réactualiser rapidement le tableau des proximités inter-classe au cours du processus de hiérarchisation.

Pour la suite de cet article, nous supposerons que $H=0$, car ce terme ne sert qu'à intégrer les indices du lien maximal et minimal, que nous n'étudierons pas. Ce choix ne modifie en rien les développements qui vont suivre.

A ce stade, il est logique de se poser les questions suivantes :

*Un indice L peut-il toujours être exprimé à l'aide de la formule de réactualisation de Lance et Williams?*

*Étant donnée une formule de réactualisation de Lance et Williams, peut-on lui faire correspondre un indice L?*

La proposition 2.1.1 nous permettra de répondre à la première question; précisons auparavant cette convention de notation :

Si $D(C_i, C_j)=\alpha_{ij}\Delta_{ii}+\beta_{ij}\Delta_{ij}+\gamma_{ij}\Delta_{jj}$

Alors $D(C_i \cup C_j, C_k)=\alpha_{i\cup j,k}\Delta_{i\cup j,i\cup j}+\beta_{i\cup j,k}\Delta_{i\cup j,k}+\gamma_{i\cup j,k}\Delta_{kk}$

PROPOSITION 2.1.1 : *Un indice L satisfait à la formule de réactualisation de Lance et Williams si et seulement si* $\forall C_i, C_j, C_k \subset X$, *les trois conditions suivantes sont réalisées :*

(i) $$\alpha_{i\cup j,k}\left(1-2\frac{\alpha_{ij}}{\beta_{ij}}\right)-\frac{\alpha_{ik}}{\beta_{ik}}\beta_{i\cup j,k}=0$$

(ii) $$\alpha_{i\cup j,k}\left(1-2\frac{\gamma_{ij}}{\beta_{ij}}\right)-\frac{\alpha_{jk}}{\beta_{jk}}\beta_{i\cup j,k}=0$$

(iii) $$\left(\frac{\gamma_{ik}}{\beta_{ik}}+\frac{\gamma_{ij}}{\beta_{jk}}\right)\beta_{i\cup j,k}-\gamma_{i\cup j,k}=0.$$

De plus, l'expression des coefficients $E$, $F$, $G$ nous est donnée par :

$$E=2\frac{\alpha_{i\cup j,k}}{\beta_{ij}}, \qquad F=\frac{\beta_{i\cup j,k}}{\beta_{ik}}, \qquad G=\frac{\beta_{i\cup j,k}}{\beta_{jk}}.$$

*Démonstration.* — Nous avons d'une part :

$$\begin{aligned}
D(C_i\cup C_j, C_k)&=\alpha_{i\cup j,k}\Delta_{i\cup j,i\cup j}+\beta_{i\cup j,k}\Delta_{i\cup j,k}+\gamma_{i\cup j,k}\Delta_{kk}\\
&=\alpha_{i\cup j,k}(\Delta_{ii}+2\Delta_{ij}+\Delta_{jj})+\beta_{i\cup j,k}\Delta_{ik}+\beta_{i\cup j,k}\Delta_{jk}+\gamma_{i\cup j,k}\Delta_{kk}\\
&=\Delta_{ii}\alpha_{i\cup j,k} && (1)\\
&\quad+\Delta_{ij}2\alpha_{i\cup j,k} && (2)\\
&\quad\quad+\Delta_{jj}\alpha_{i\cup j,k} && (3)\\
&\quad\quad\quad+\Delta_{ik}\beta_{i\cup j,k} && (4)\\
&\quad\quad\quad\quad+\Delta_{jk}\beta_{i\cup j,k} && (5)\\
&\quad\quad\quad\quad\quad+\Delta_{kk}\gamma_{i\cup j,k} && (6)
\end{aligned}$$

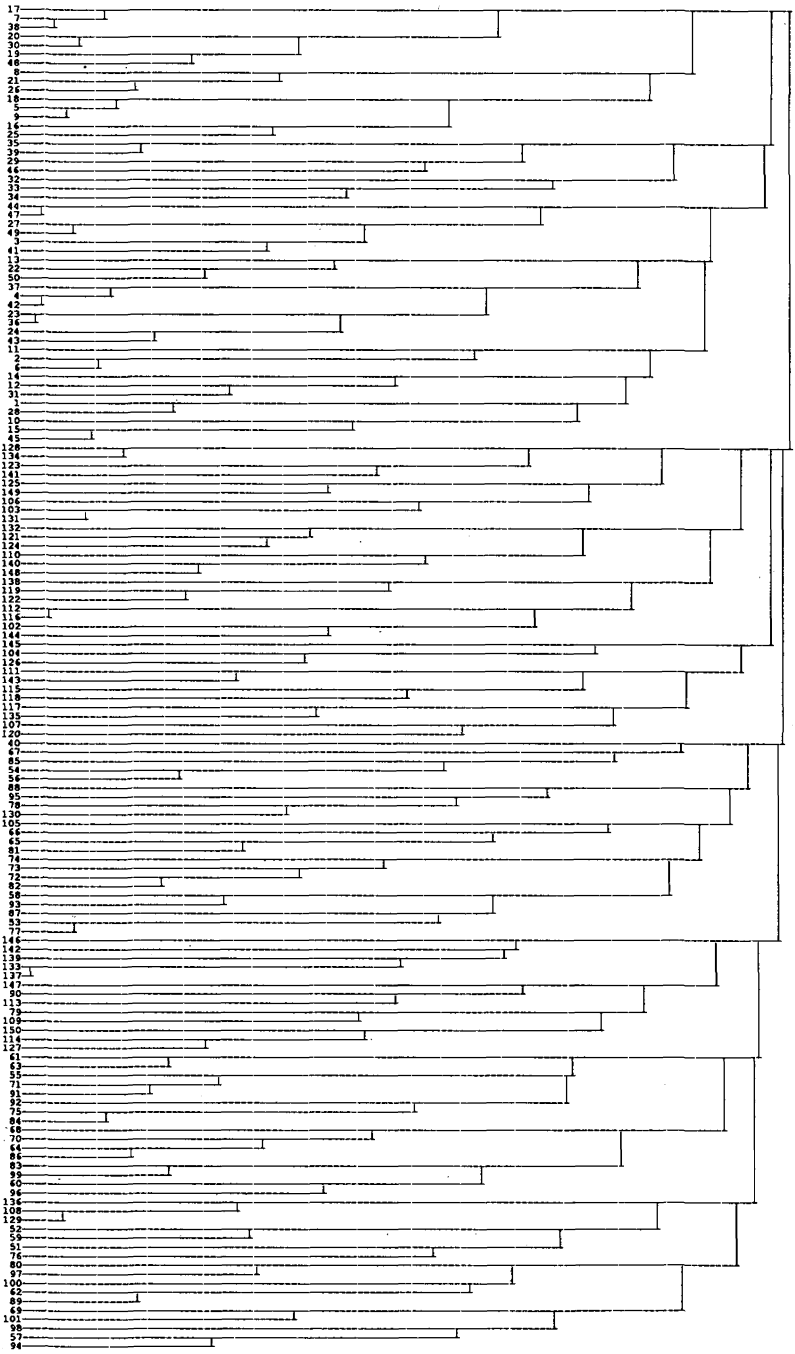

**Figure 1. — Indice de distance inter-classe : 1 distance : $L2$.**

D'autre part, on a :

$$
\begin{aligned}
D(C_i \cup C_j, C_k) &= ED(C_i, C_j) + FD(C_i, C_k) + GD(C_j, C_k) \\
&= E(\alpha_{ij}\Delta_{ii} + \beta_{ij}\Delta_{ij} + \gamma_{ij}\Delta_{jj}) + F(\alpha_{ik}\Delta_{ii} + \beta_{ik}\Delta_{ik}\Delta_{ik} + \gamma_{lk}\Delta_{kk}) \\
&\qquad + G(\alpha_{jk}\Delta_{jj} + \beta_{jk}\Delta_{jk} + \gamma_{jk}\Delta_{kk}) \\
&= \Delta_{ii}(E\alpha_{ij} + F\alpha_{ik}) \qquad (1)' \\
&\quad + \Delta_{ij} E\beta_{ij} \qquad (2)' \\
&\quad + \Delta_{jj}(G\alpha_{jk} + E\gamma_{ij}) \qquad (3)' \\
&\quad + \Delta_{ik} F\beta_{ik} \qquad (4)' \\
&\quad + \Delta_{jk} G\beta_{jk} \qquad (5)' \\
&\quad + \Delta_{kk}(F\gamma_{ik} + G\gamma_{jk}) \qquad (6)'
\end{aligned}
$$

Les coefficients des termes $\Delta_{kl} \forall k, l$, doivent être égaux, l'égalité étant vraie pour toute dissimilarité. On en déduit le système d'équations en $E$, $F$, $G$ suivant :

$$
\begin{aligned}
E\alpha_{ij} + F\alpha_{ik} &= \alpha_{i \cup j, k} \\
E\beta_{ij} &= 2\alpha_{i \cup j, k} \\
G\alpha_{jk} + E\gamma_{ij} &= \alpha_{i \cup j, k} \\
F\beta_{ik} &= \beta_{i \cup j, k} \\
G\beta_{jk} &= \beta_{i \cup j, k} \\
F\gamma_{ik} + G\gamma_{jk} &= \gamma_{i \cup j, k}
\end{aligned}
$$

Ce système de six équations à trois inconnues ne possède de solutions que si les trois conditions (i) (ii) et (iii) sont vraies et nous déduisons la proposition.

COROLLAIRE 2.1.2 : *Les indices L pour lesquels* $\alpha_{ij} = \beta_{ij}/2 = \gamma_{ij}$ *avec* $\beta_{ij} \neq 0$ *ne possèdent pas de formule de réactualisation du type Lance et Williams.*

*Démonstration.* — Si nous prenons par exemple la première condition de la proposition 2.1.1 et si nous supposons que $\alpha_{ij} = \beta_{ij}/2 = \gamma_{ij}$ alors nous concluons que $\beta_{i \cup j, k} = 0$ ; ce qui est une contradiction.

*Applications.* — 1. On constate que, selon ce corollaire, *l'indice de la variance :*

$$
\frac{1}{n_i + n_j} \sum_{x \in C_i \cup C_j} \| X - \bar{X} \|^2
$$

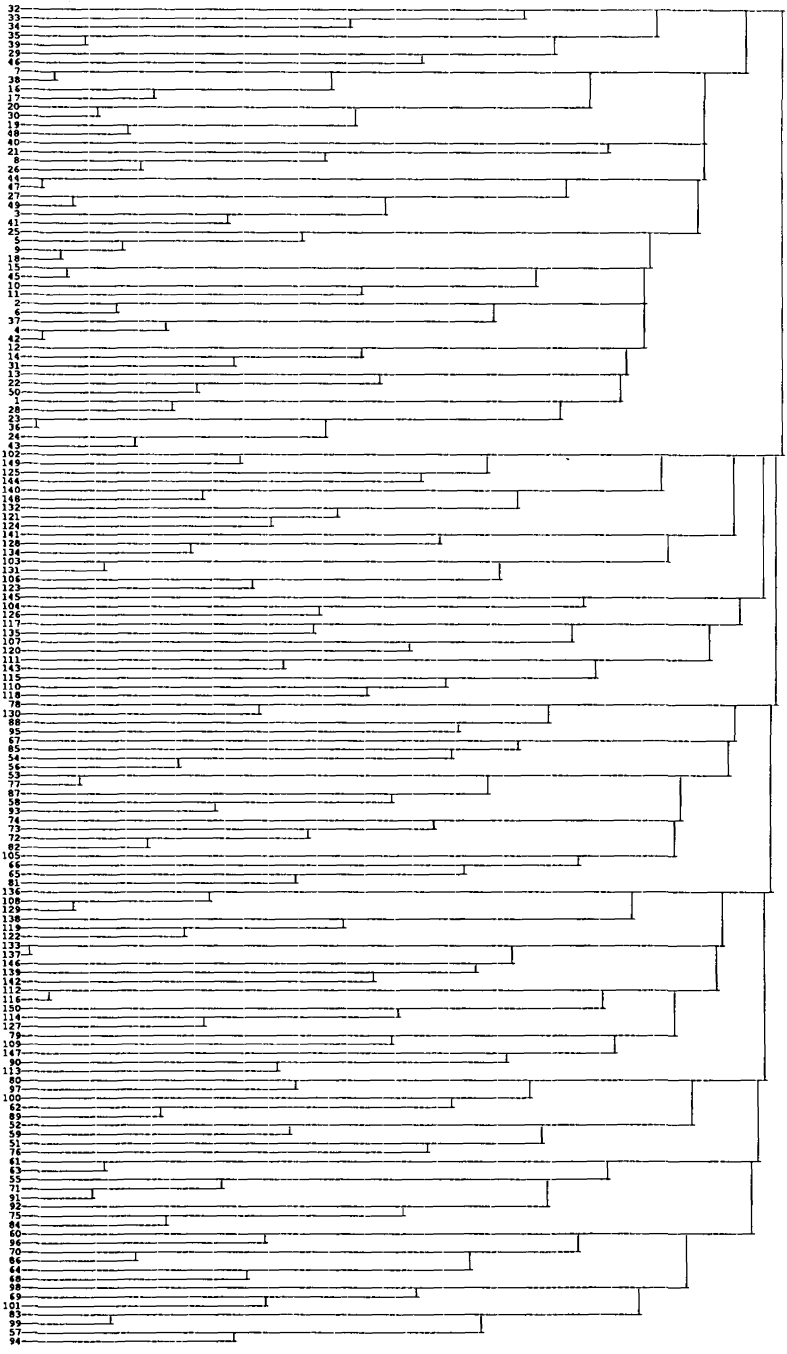

**Figure 2. — Indice de distance inter-classe : 2 distance : $L$ 2.**

où $\bar{X}$ est la moyenne de $C_i \cup C_j$, appelé, dans le manuel de SPSSX, indice « Waverage », qui a été défini pour la première fois, semble-t-il, par Anderberg et que l'on peut encore écrire :

$$\frac{1}{2(n_i+n_j)^2}(\Delta_{ii}+2\Delta_{ij}+\Delta_{jj})$$

ne peut donc être exprimé à l'aide d'une formule de réactualisation de Lance et Williams. Notons que l'implantation de cet indice dans ce logiciel est, de ce fait, *incorrecte.*

2. Nous prenons l'expression de *l'indice de la médiane* qui s'exprime dans le contexte de la formule de Lance et Williams par :

$$D(C_i \cup C_j, C_k) = -\frac{1}{4}D(C_i, C_j)+\frac{1}{2}D(C_i, C_k)+\frac{1}{2}D(C_j, C_k).$$

Supposons qu'il existe un indice $L$ satisfaisant à cette formule de réactualisation, alors, selon la proposition 2.1.1, selon l'expression de $F$ et $G$ :

quelles que soient les classes $C_i$, $C_j$, $C_k$ :

$$\frac{\beta_{i\cup j,k}}{\beta_{ik}}=\frac{1}{2} \quad \text{et} \quad \frac{\beta_{i\cup j,k}}{\beta_{ij}}=\frac{1}{2}.$$

Donc :

$$\forall i,j,k, \quad \frac{\beta_{ij}}{\beta_{ik}}=1 \Rightarrow \beta_{ij}=\beta_{ik}$$

$$\Rightarrow \text{ les } \beta_{ij} \text{ sont constants}$$

Ainsi, $\beta_{i\cup j,k}/\beta_{ij}=1$ ce qui est une contradiction.

Ceci montre qu'on ne peut pas associer à toute méthode hiérarchique basée sur la formule de Lance et Williams, un indice $L$.

3. Nous obtenons ainsi une démonstration originale permettant de déterminer facilement les coefficients de Lance et Williams pour les indices de Ward, des centres de gravité, ...

Nous étudions maintenant le rapport entre les indices $L$ et une autre formule de réactualisation : *la formule de Jambu.*

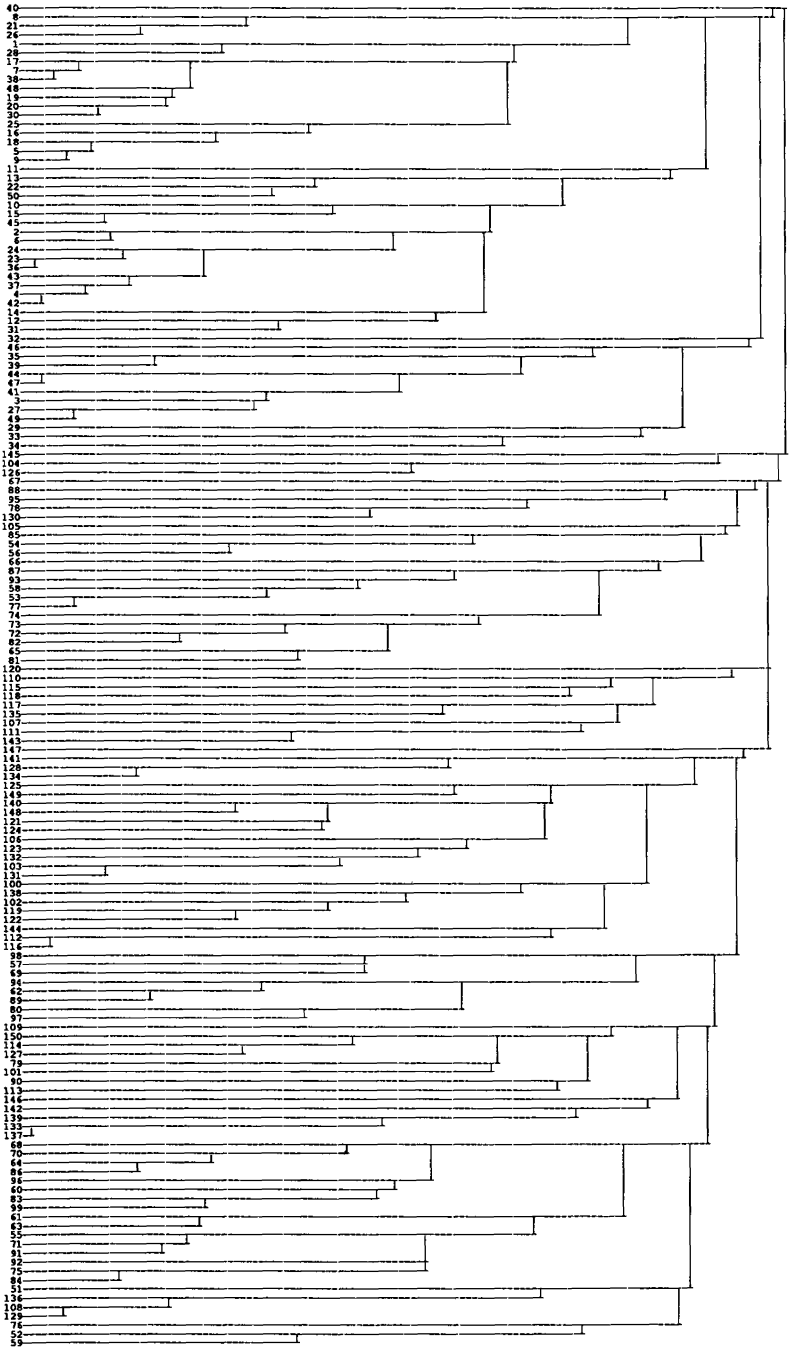

**Figure 3. – Indice de distance inter-classe : 3 distance : $L$ 2.**

## 2.2. La formule de Jambu

Nous rappelons que la formule de réactualisation de Jambu s'écrit :

$$D(C_i \cup C_j, C_k) = A f(C_i) + B f(C_j) + C f(C_k) + ED(C_i, C_j) + FD(C_i, C_k) + GD(C_j, C_k)$$

où $A$, $B$, $C$, $E$, $F$, $G$, sont des réels et où $f$ représente « l'indice de la hiérarchie ». En général :

$$f(C_i) = D(C_{i1}, C_{i2})$$

où $C_{i1}$ et $C_{i2}$ désignent les deux classes ayant servi à former la classe $C_i$. C'est ce choix de $f$ que nous considérons dès maintenant car c'est le seul qui nous semble présenter un intérêt pratique.

Nous sommes dès lors intéressés à savoir si un indice $L$ peut être exprimé à l'aide d'une formule de réactualisation de type Jambu. Les propositions suivantes répondent à cette question :

PROPOSITION 2.2.1 : *Si un indice L* possède une formule de réactualisation du type Lance et Williams, alors il possède une formule de réactualisation du type Jambu.

Dans ce cas, les coefficients $E$, $F$ et $G$ de la formule de Jambu sont identiques à ceux de la formule de Lance et Williams et $A = B = C = 0$.

*Démonstration.* — C'est une évidence compte tenu de l'écriture de ces deux formules.

PROPOSITION 2.2.2 : *Si un indice L* ne possède pas de formule de réactualisation de Lance et Williams, alors il possède une formule de réactualisation de Jambu si et seulement si :

$$\alpha_{ij} = \frac{\beta_{ij}}{2} = \gamma_{ij} \neq 0.$$

De plus, dans le cas où un indice $L$ possède une formule de récurrence de Jambu, nous avons les relations suivantes liant les coefficients de l'indice $L$ à ceux de la formule de réactualisation de Jambu :

$$A = -\frac{\beta_{i \cup j,\, k}}{\beta_{i\,1,\, i\,2}}, \qquad B = -\frac{\beta_{i \cup j,\, k}}{\beta_{j\,1,\, j\,2}}, \qquad C = -\frac{\beta_{i \cup j,\, k}}{\beta_{k\,1,\, k\,2}}$$

$$E = \frac{\beta_{i \cup j,\, k}}{\beta_{ij}}, \qquad F = \frac{\beta_{i \cup j,\, k}}{\beta_{ik}}, \qquad G = \frac{\beta_{i \cup j,\, k}}{\beta_{jk}}$$

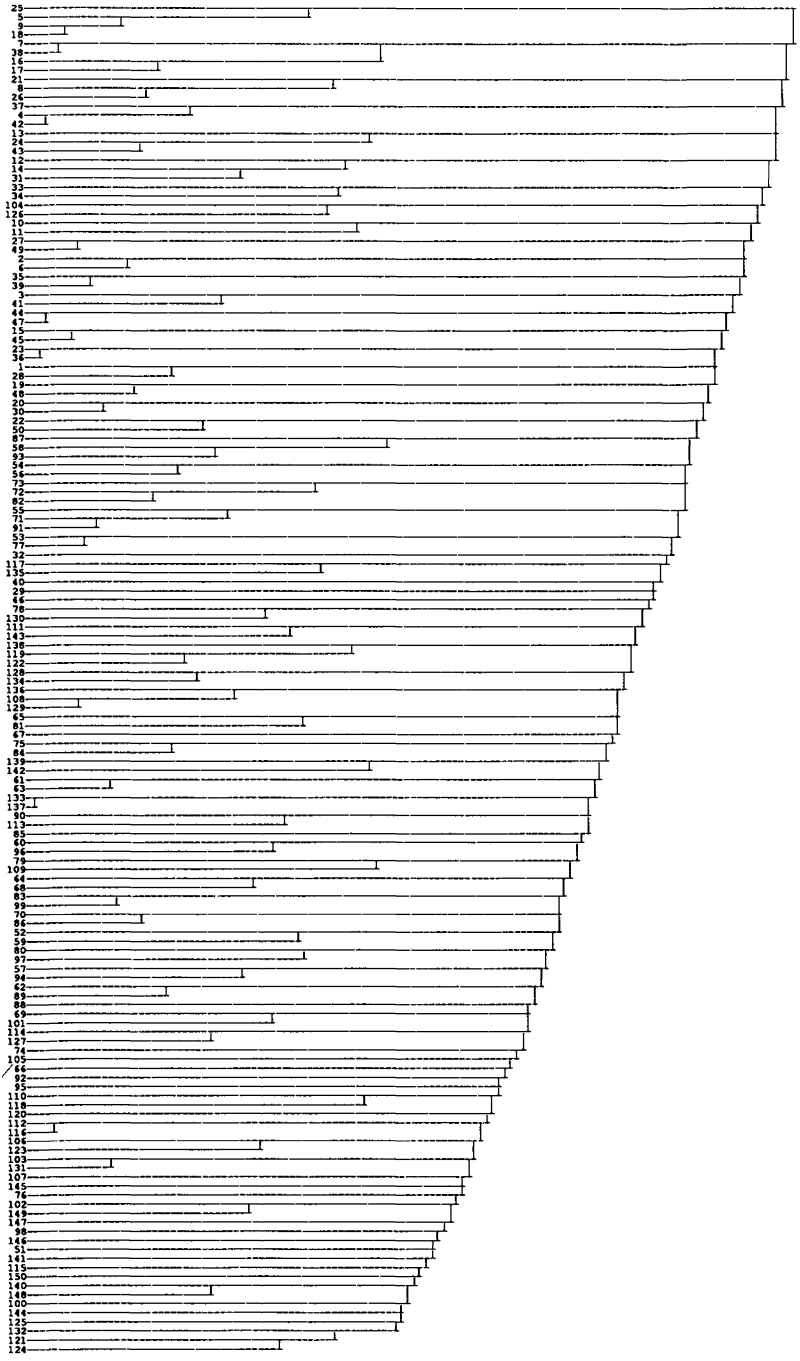

**Figure 4. — Indice de distance inter-classe : 1 distance : $L$ 1.**

*Démonstration.* — On écrit d'une part $D(C_i \cup C_j, C_k)$ en appliquant la définition d'un indice $L$ :

$$D(C_i \cup C_j, C_k) = \alpha_{i\cup j,k}(\Delta_{ii} + 2\Delta_{ij} + \Delta_{jj}) + \beta_{i\cup j,k}\Delta_{ik} + \beta_{i\cup j,k}\Delta_{jk} + \gamma_{i\cup j,k}\Delta_{kk}$$

où les termes $\Delta_{ii}\,\Delta_{jj}$ et $\Delta_{kk}$ peuvent encore être décomposé au niveau inférieur :

$$\Delta_{ii} = \Delta_{i1,i1} + 2\Delta_{i1,i2} + \Delta_{i2,i2}$$
$$\Delta_{jj} = \Delta_{j1,j1} + 2\Delta_{j1,j2} + \Delta_{j2,j2}$$
$$\Delta_{kk} = \Delta_{k1,k1} + 2\Delta_{k1,k2} + \Delta_{k2,k2}$$

et d'autre part, on explicite le deuxième membre de la formule de Jambu sous forme d'un indice $L$ :

$$\begin{aligned} D(C_i \cup C_j, C_k) &= Af(C_i) + Bf(C_j) + Cf(C_k) + ED(C_i, C_j) \\ &\quad + FD(C_i, C_k) + GD(C_j, C_k) \\ &= AD(C_{i1}, C_{i2}) + BD(C_{j1}, C_{j2}) + CD(C_{k1}, C_{k2}) \\ &\quad + ED(C_i, C_j) + FD(C_i, C_k) + GD(C_j, C_k) \\ &= A(\alpha_{i1,i2}\Delta_{i1,i2} + \beta_{i1,i2}\Delta_{i1,i2} + \gamma_{i1,i2}\Delta_{i2,i2}) \\ &\quad + B(\alpha_{j1,j2}\Delta_{j1,j1} + \beta_{j1,j2}\Delta_{j1,j2} + \gamma_{j1,j2}\Delta_{j2,j2}) \\ &\quad + C(\alpha_{k1,k2}\Delta_{k1,k1} + \beta_{k1,k2} + \gamma_{k1,k2} + \gamma_{k1,k2}\Delta_{k2,k2}) \\ &\quad + E(\alpha_{ij}\Delta_{ii} + \beta_{ij}\Delta_{ij} + \gamma_{ij}\Delta_{jj}) \\ &\quad + F(\alpha_{ik}\Delta_{ii} + \beta_{ik}\Delta_{ik} + \gamma_{ik}\Delta_{kk}) + G(\alpha_{jk}\Delta_{jj} + \beta_{jk}\Delta_{jk} + \gamma_{jk}\Delta_{kk}). \end{aligned}$$

On remplace ensuite les termes $\Delta_{ii}\,\Delta_{jj}$ et $\Delta_{kk}$ par la même décomposition que ci-dessus.

En identifiant les coefficients des termes $\Delta_{ij}$, $\Delta_{i1,i1}$, $\Delta_{jk}\ldots$, nous obtenons alors les 12 équations suivantes :

$$\alpha_{i\cup j,k} = A\alpha_{i1,i2} + F\alpha_{ik} + E\alpha_{ij}$$
$$2\alpha_{i\cup j,k} = A\beta_{i1,i2} + 2F\alpha_{ik} + 2E\alpha_{ij}$$
$$\alpha_{i\cup j,k} = A\gamma_{i1,i2} + F\alpha_{ik} + E\alpha_{ij}$$
$$\alpha_{i\cup j,k} = B\alpha_{j1,j2} + E\gamma_{ij} + G\alpha_{jk}$$
$$2\alpha_{i\cup j,k} = B\beta_{j1,j2} + 2E\gamma_{ij} + 2G\alpha_{jk}$$

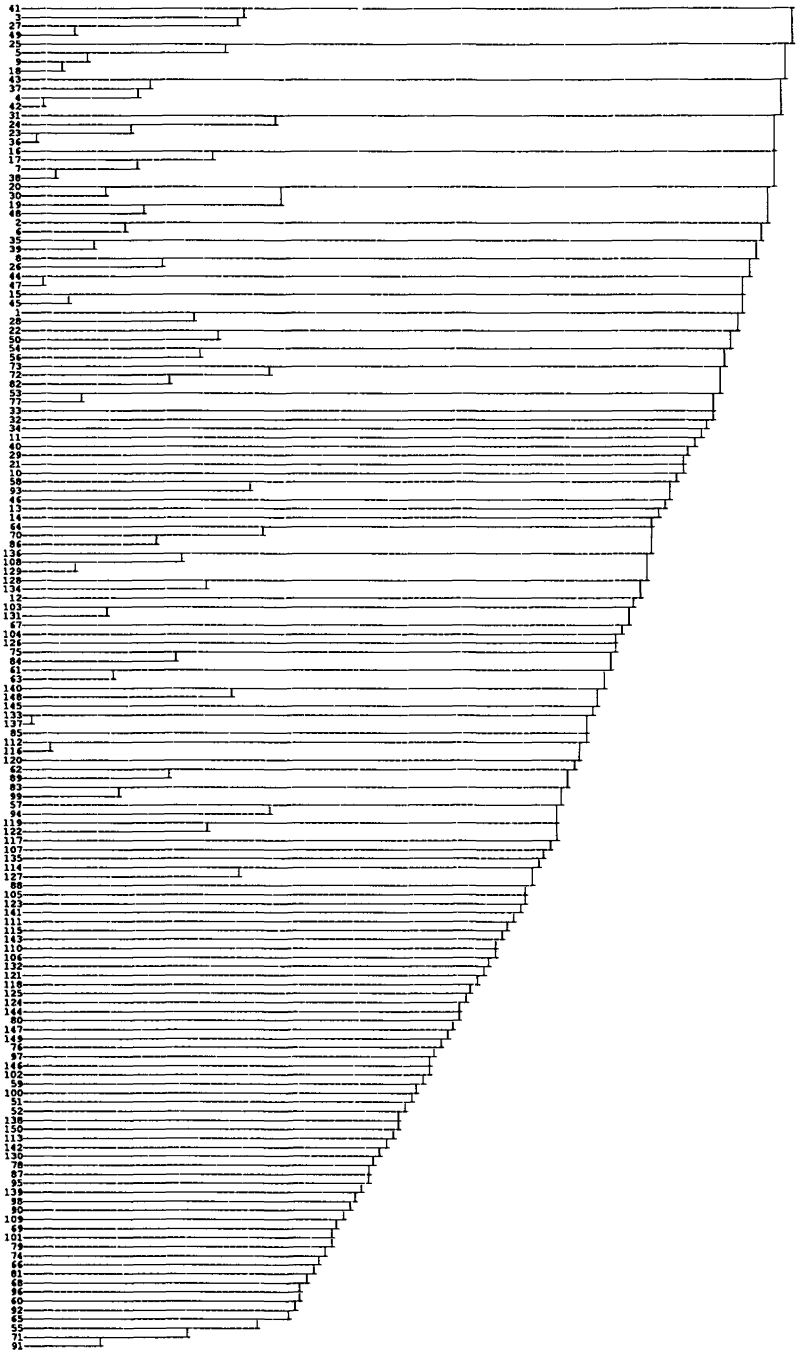

**Figure 5. — Indice de distance inter-classe : 2 distance : $L$ 1.**

$$\alpha_{i\cup j,\,k}=B\gamma_{j1,\,j2}+2E\gamma_{ij}+G\alpha_{jk}$$
$$\gamma_{i\cup j,\,k}=C\alpha_{k1,\,k2}+F\gamma_{ik}+G\gamma_{jk}$$
$$\gamma_{i\cup j,\,k}=C\beta_{k1,\,k2}+2F\gamma_{ik}+2G\gamma_{jk}$$
$$\gamma_{i\cup j,\,k}=C\gamma_{k1,\,k2}+F\gamma_{ik}+G\gamma_{jk}$$
$$\beta_{i\cup j,\,k}=F\beta_{ik}$$
$$\beta_{i\cup j,\,k}=G\gamma_{kk}$$
$$2\alpha_{i\cup j,\,k}=E\beta_{ij}.$$

En résolvant ce système de 12 équations à 6 inconnues et en utilisant la condition énoncée, nous obtenons les résultats annoncés dans le théorème.

On voit clairement la limite d'applicabilité de cette formule de réactualisation dans le contexte des indices $L$, la condition imposée : $\alpha_{ij}=\beta_{ij}/2=\gamma_{ij}$ est très restrictive. Cette formule permet de traiter toutefois l'indice de la variance cité auparavant.

Nous passons maintenant de manière naturelle à l'établissement d'une formule de réactualisation encore mieux adaptée aux indices $L$.

### 3. UNE NOUVELLE FORMULE DE RÉACTUALISATION

Nous donnons une formule de réactualisation intégrant tous les indices $L$ tels que $\beta_{ij}\neq 0$ :

PROPOSITION 3.1 : *Tout indice $L$ pour lequel $\beta_{ij}\neq 0$ peut s'exprimer à l'aide de la formule de réactualisation suivante :*

$$D(C_i\cup C_j,\,C_k)=A\Delta_{ii}+B\Delta_{jj}+C\Delta_{kk}+ED(C_i,\,C_j)+FD(C_i,\,C_k)+GD(C_j,\,C_k).$$

De plus, on a les relations suivantes liant les coefficients de l'indice $L$ à ceux de la nouvelle formule de réactualisation :

$$A=\alpha_{i\cup j,\,k}\left(1-2\frac{\alpha_{ij}}{\beta_{ij}}\right)-\frac{\alpha_{ik}}{\beta_{ik}}\beta_{i\cup j,\,k}$$
$$B=\alpha_{i\cup j,\,k}\left(1-2\frac{\gamma_{ij}}{\beta_{ij}}\right)-\frac{\alpha_{jk}}{\beta_{jk}}\beta_{i\cup j,\,k}$$
$$C=\gamma_{i\cup j,\,k}-\left(\frac{\gamma_{ik}}{\beta_{ik}}+\frac{\gamma_{ik}}{\beta_{jk}}\right)\beta_{i\cup j,\,k}$$

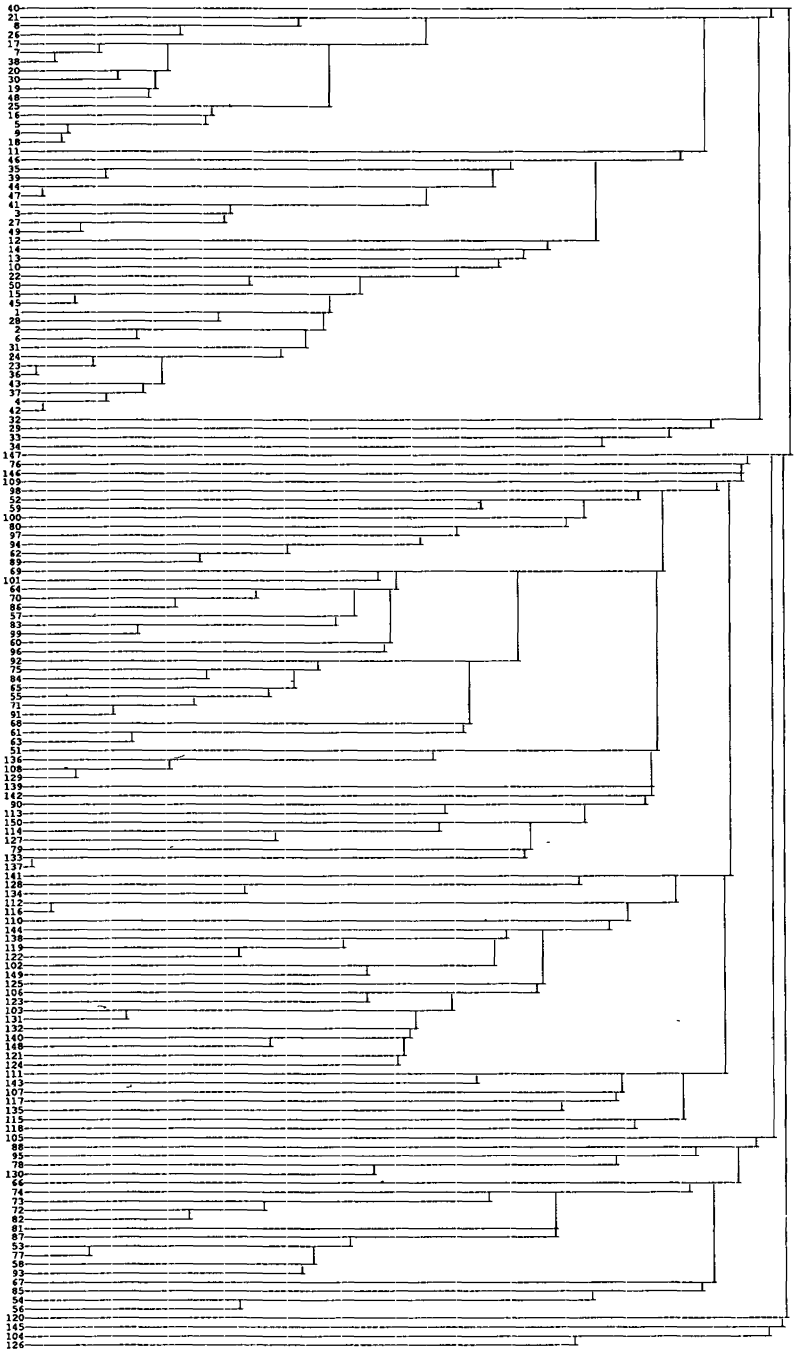

**Figure 6. – Indice de distance inter-classe : 3 distance : $L1$.**

$$E=2\frac{\alpha_{i\cup j,\,k}}{\beta_{ij}},\qquad F=\frac{\beta_{i\cup j,\,k}}{\beta_{ik}},\qquad G=\frac{\beta_{i\cup j,\,k}}{\beta_{jk}}$$

*Démonstration.* — On a par définition d'un indice $L$ :

$$D(C_i,\,C_j)=\alpha_{ij}\Delta_{ii}+\beta_{ij}\Delta_{ij}+\gamma_{ij}\Delta_{jj}$$

donc :

$$\begin{aligned}
D(C_i\cup C_j,\,C_k)&=\alpha_{i\cup j,\,k}\Delta_{i\cup j,\,i\cup j}+\beta_{i\cup j,\,k}\Delta_{i\cup j,\,k}+\gamma_{i\cup j,\,k}\Delta_{kk}\\
&=\alpha_{i\cup j,\,k}(\Delta_{ii}+2\Delta_{ij}+\Delta_{jj})+\beta_{i\cup j,\,k}\Delta_{ik}+\beta_{i\cup j,\,k}\Delta_{jk}+\gamma_{i\cup j,\,k}\Delta_{kk}\\
&=\Delta_{ii}\alpha_{i\cup j,\,k} &(1)\\
&\quad+\Delta_{ij}2\alpha_{i\cup j,\,k} &(2)\\
&\quad\quad+\Delta_{jj}\alpha_{i\cup j,\,k} &(3)\\
&\quad\quad\quad+\Delta_{ik}\beta_{i\cup j,\,k} &(4)\\
&\quad\quad\quad\quad+\Delta_{jk}\beta_{i\cup j,\,k} &(5)\\
&\quad\quad\quad\quad\quad+\Delta_{kk}\gamma_{i\cup j,\,k} &(6)
\end{aligned}$$

D'autre part, nous réécrivons la nouvelle formule de réactualisation en lui appliquant la définition d'un indice $L$.

$$\begin{aligned}
D(C_i\cup C_j,\,C_k)&=A\Delta_{ii}+B\Delta_{jj}+C\Delta_{kk}+ED(C_i,\,C_j)+FD(C_i,\,C_k)+GD(C_j,\,C_k)\\
D(C_i\cup C_j,\,C_k)&=A\Delta_{ii}+B\Delta_{jj}+C\Delta_{kk}+E(\alpha_{ij}\Delta_{ii}+\beta_{ij}\Delta_{ij}+\gamma_{ij}\Delta_{jj})\\
&\qquad+F(\alpha_{ik}\Delta_{ii}+\beta_{ik}\Delta_{ik}+\gamma_{ik}\Delta_{kk})+G(\alpha_{jk}\Delta_{jj}+\beta_{jk}\Delta_{jk}+\gamma_{jk}\Delta_{kk})\\
&=\Delta_{ii}(A+E\alpha_{ij}+F\alpha_{ik}) &(1)'\\
&\quad+\Delta_{ij}E\beta_{ij} &(2)'\\
&\quad\quad+\Delta_{jj}(B+E\gamma_{ij}) &(3)'\\
&\quad\quad\quad+\Delta_{ik}F\beta_{ik} &(4)'\\
&\quad\quad\quad\quad+\Delta_{jk}G\beta_{jk} &(5)'\\
&\quad\quad\quad\quad\quad+\Delta_{kk}(C+F\gamma_{ik}+G\gamma_{jk}) &(6)'
\end{aligned}$$

En identifiant les coefficients des termes $\Delta_{ij}$, $\Delta_{ik}$, $\Delta_{jk}\ldots$, comme nous l'avons fait pour les autres formules nous obtenons un système de six équations à six inconnues. La solution de ce système nous donne les résultats annoncés dans notre proposition.

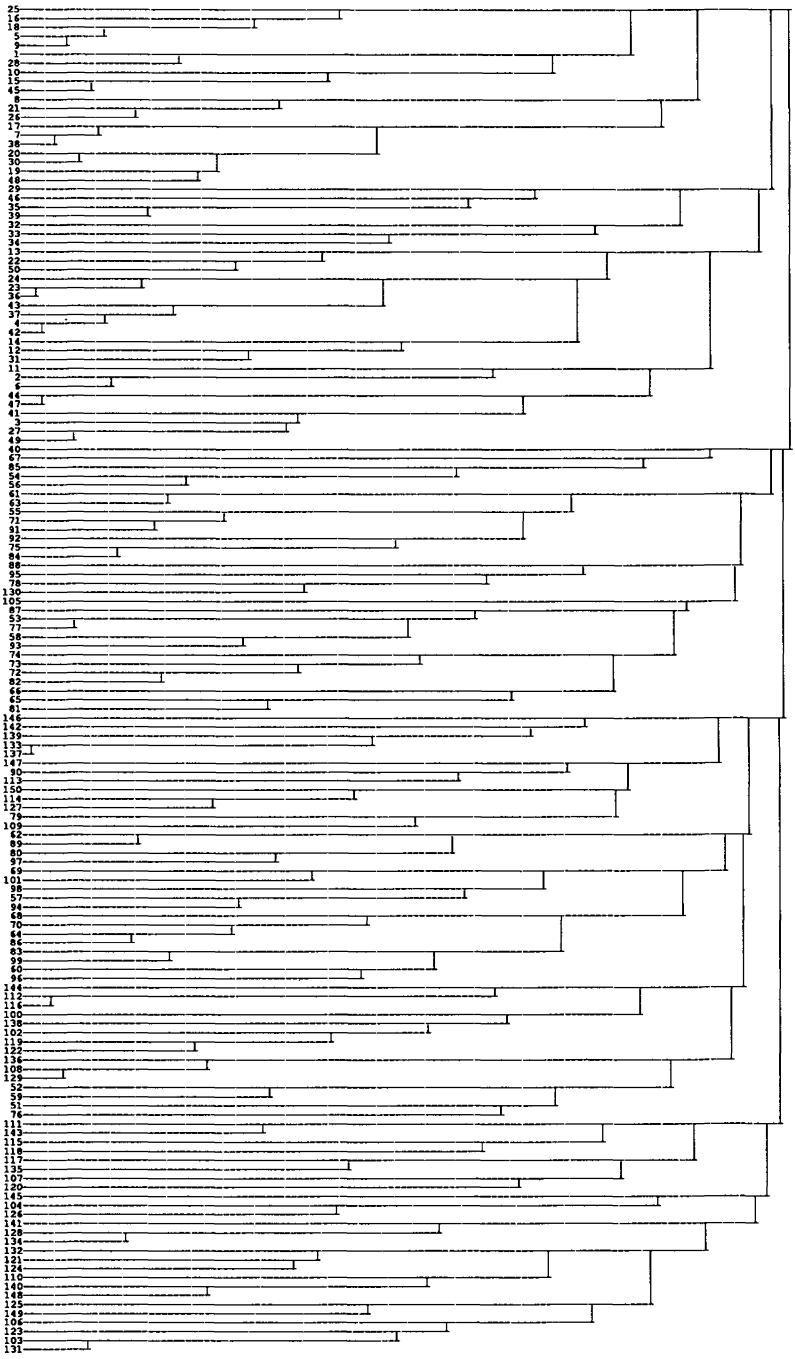

**Figure 7. – Indice de distance inter-classe de Ward distance : $L2$.**

Cette proposition suggère les remarques suivantes :

1. Si un indice $L$ peut s'écrire sous la « forme Lance et Williams », alors $A=B=C=0$ (dans la proposition 3.1), et on retrouve les coefficients de la proposition 2.1.1.

2. Si un indice $L$ ne peut être écrit sous la « forme Lance et Williams » mais qu'il peut être écrit sous la « forme Jambu », on ne retrouve pas les coefficients de la proposition 2.2.1.

3. Afin de pouvoir passer à l'étape $k+1$, il faut en particulier calculer l'étape $k$, la quantité $\Delta_{i+j,\ i+j}$ qui est égale à $\Delta_{ii}+2\Delta_{ij}+\Delta_{jj}$. Les trois termes de cette dernière expression peuvent être déduits de l'étape $k-1$ puisqu'on y a calculé les quantités $\Delta_{ii}$, $\Delta_{jj}$ et $D(C_i, C_j)=\alpha_{ij}\Delta_{ii}+\beta_{ij}\Delta_{ij}+\gamma_{ij}\Delta_{jj}$. D'où :

$$\Delta_{ij}=\frac{1}{\beta_{ij}}\left(D(C_i, C_j)-\alpha_{ij}\Delta_{ii}-\gamma_{ij}\Delta_{jj}\right).$$

Nous avons ainsi les qualités informatiques recherchées : à l'étape $k$, il suffit de mémoriser les $D(C_i, C_j)$ et les $\Delta_{ii}$, soit $k(k+1)/2$ termes.

4. Cette nouvelle formule de réactualisation est, dans le contexte des indices $L$, plus générale que celle de Jambu dans la mesure où tout indice $L$ qui possède une formule de réactualisation Jambu satisfait également à la condition nécessaire que l'on vient de voir mais que la réciproque n'est pas vraie.

5. Nous donnons les coefficients des indices du paragraphe 1.2 dans leur forme de formule de réactualisation.

(*a*) Pour

$$D(C_i, C_j)=-\frac{1}{2}\Delta_{ii}+\Delta_{ij}-\frac{1}{2}\Delta_{jj}$$

$$A=-\frac{1}{2}, \qquad B=-\frac{1}{2}, \qquad C=\frac{1}{2} \quad E=-1, \qquad F=1 \quad G=1$$

(*b*) Pour

$$D(C_i, C_j)=-\frac{1}{2n_i}\Delta_{ii}+\frac{2}{n_i+n_j}\Delta_{ij}-\frac{1}{2n_j}\Delta_{jj}$$

$$A=\frac{-1}{2(n_i+n_j)}-\frac{1}{4n_i}+\frac{n_i+n_k}{2n_i(n_i+n_j+n_k)}$$

$$B=\frac{1}{2(n_i+n_j)}-\frac{1}{4n_j}+\frac{n_j+n_k}{2n_i(n_i+n_j+n_k)}$$

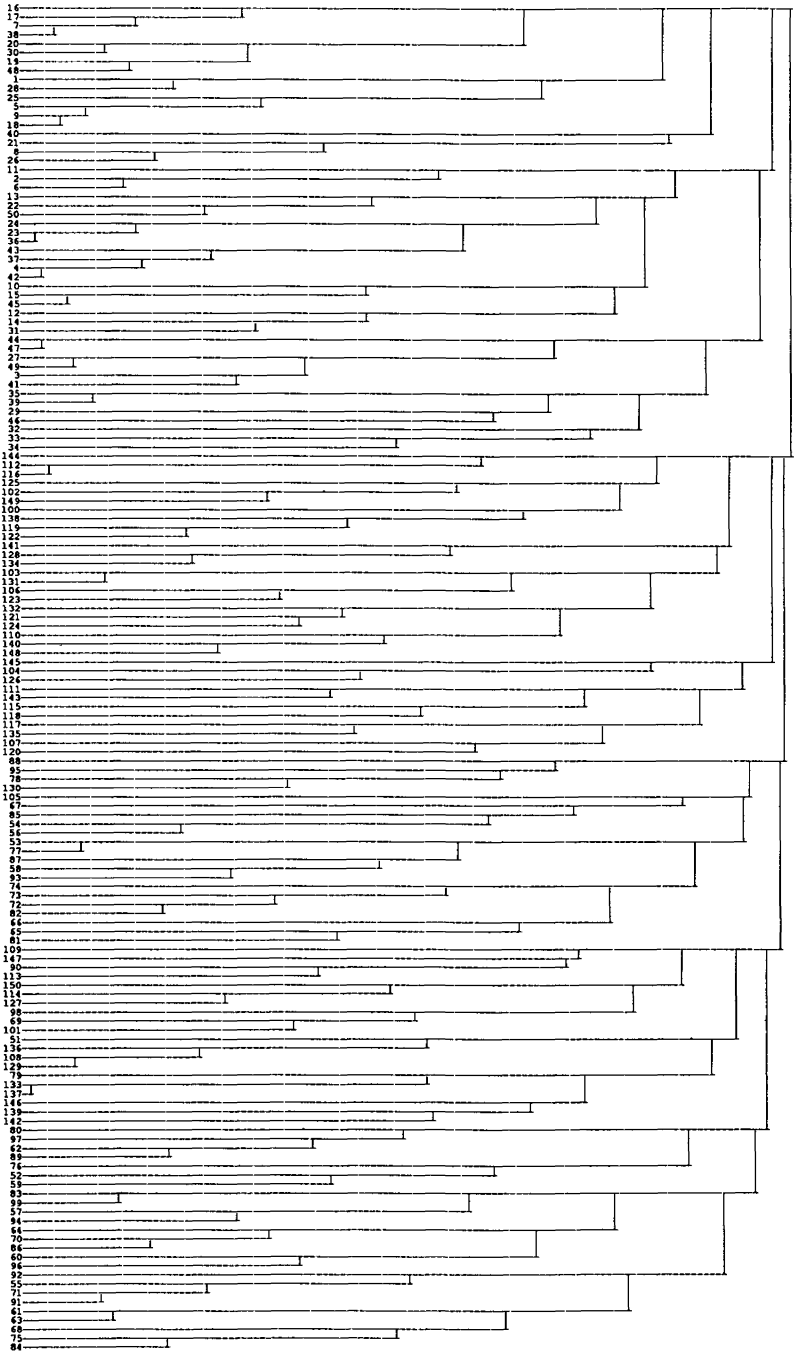

**Figure 8**

$$C = \frac{1}{2(n_i+n_j+n_k)}, \qquad E = -\frac{1}{2}$$

$$F = \frac{n_i+n_k}{n_i+n_j+n_k}, \qquad G = \frac{n_j+n_k}{n_i+n_j+n_k}$$

(*c*) Pour

$$D(C_i, C_j) = -\frac{1}{2n_i^2}\Delta_{ii} + \frac{1}{n_i n_j}\Delta_{ij} - \frac{1}{2n_j^2}\Delta_{jj}$$

$$A = B = C = 0$$

$$E = -\frac{n_i n_j}{(n_i+n_j)^2}, \qquad F = \frac{n_i}{n_i+n_j} \quad G = \frac{n_j}{n_i+n_j}$$

6. Afin que la formule de Lance et Williams soit complètement un cas particulier de notre formule dans le cadre des indices $L$, on l'écrira :

$$D(C_i \cup C_j, C_k) \doteq A\Delta_{ii} + B\Delta_{jj} + C\Delta_{kk} + ED(C_i, C_j) + FD(C_i, C_k) + GD(C_j, C_k) + H\left|D(C_i, C_k) - D(C_j, C_k)\right|$$

7. Nous aurions également pu étudier une formule de réactualisation du type :

$$D(C_i \cup C_j, C_k) = AD(C_i, C_i) + BD(C_j, C_j) + CD(C_k, C_k) + ED(C_i, C_j) + FD(C_i, C_k) + GD(C_j, C_k) + H\left|D(C_i, C_k) - D(C_j, C_k)\right|.$$

Mais il est aisé de montrer qu'un indice $L$ satisfait à une telle formule si et seulement si :

$$\alpha_{ii} + \beta_{ii} + \gamma_{ii} \neq 0 \qquad \text{et} \qquad \beta_{ij} \neq 0.$$

Elle est donc moins générale.

8. Remarquons que la démonstration de la proposition 3.1 est indépendante dans une très large mesure de la manière dont les $\Delta_{ij}$ ont été définis. Les conditions sur la définition de $\Delta_{ij}$ pour que la proposition soit valide sont que l'identification terme à terme des coefficients soit possible et que les termes de la forme $\Delta_{i\cup j,\, i\cup j}$ et $\Delta_{i\cup j,\, k}$ soient des fonctions linéaires respectivement de $\Delta_{ii}$, $\Delta_{ij}$, et de $\Delta_{ik}$, $\Delta_{jk}$.

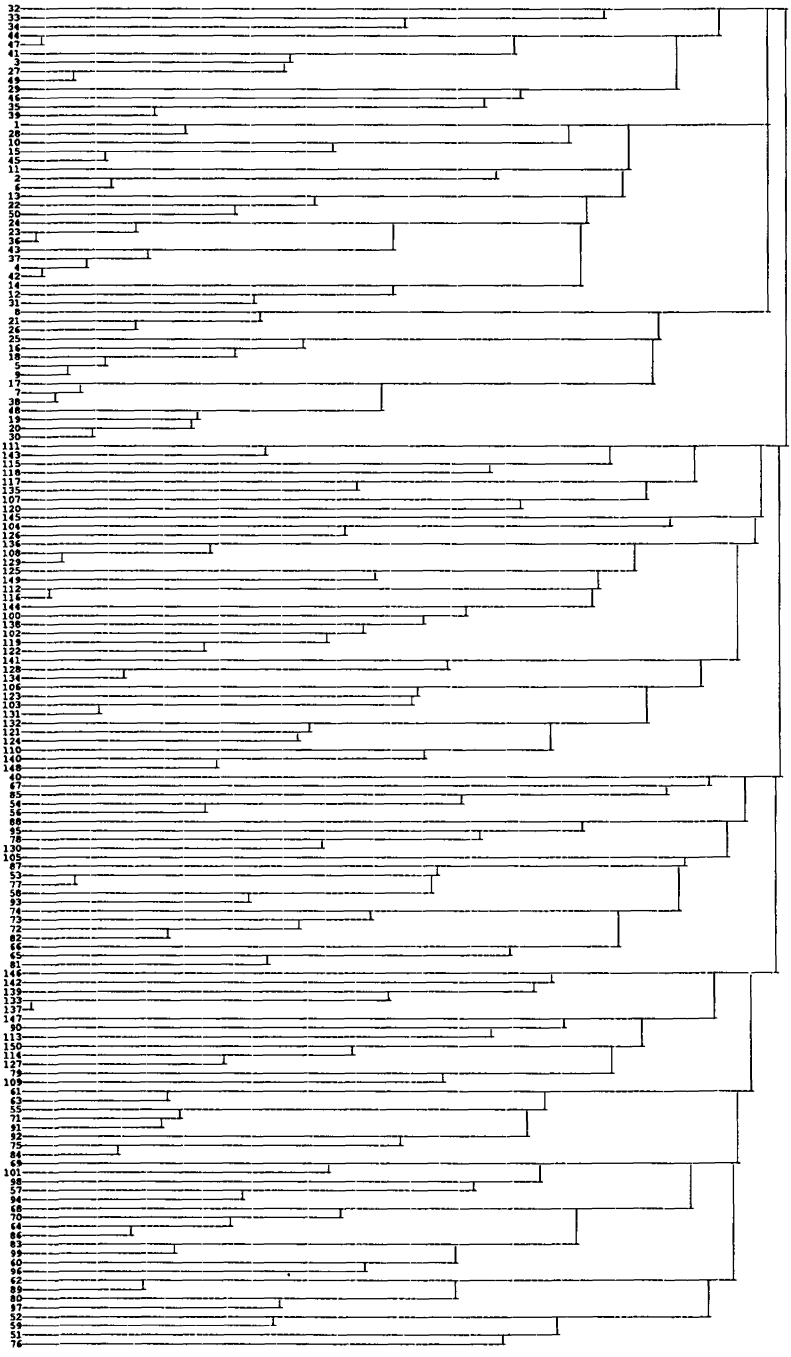

**Figure 9. — Indice de distance inter-classe : 2 modifié distance : $L1$.**

## 4. UN EXEMPLE

Nous allons maintenant étudier sur un exemple, l'intérêt pratique des indices $L$ et de notre nouvelle formule de réactualisation.

### 4.1. Exemple et objectif

Nous avons choisi comme exemple les bien connus iris de Fisher. Rappelons qu'il s'agit de 150 iris sur lesquels on a mesuré la longueur et la largeur des pétales, la longueur et la largeur des sépales. Ces iris sont divisés en trois espèces, 50 de chaque, soit iris setosa, iris versicolor et iris virginia.

L'objectif global du traitement est de voir si l'on retrouve cette classification préalable en trois groupes par la classification hiérarchique.

### 4.2. Traitement et analyse

Nous avons appliqué successivement les trois indices proposés en 1, avec la norme $L2$ (*fig.* 1, 2 et 3), puis $L1$ (*fig.* 4, 5 et 6) puis l'indice de Ward avec la norme $L2$ (*fig.* 7) et $L1$ (*fig.* 8). Finalement nous avons également appliqué un nouvel indice pour une raison que nous détaillerons plus bas. Il s'agit de l'indice 2 dans lequel on a pris la racine carrée des coefficients et la norme $L1$. Soit :

$$-\sqrt{\frac{1}{2n_i}}\Delta_{ii}+\sqrt{\frac{2}{n_i+n_j}}\Delta_{ij}-\sqrt{\frac{1}{2n_j}}\Delta_{jj}$$

Sur ces figures la hiérarchie est simplement indicée selon le numéro de l'étape de la classification et non selon l'indice de proximité inter-classe.

Les iris du premier groupe ont été numérotés de 1 à 50.

Ceux du deuxième, 51 à 100, ceux du troisième, 101 à 150.

Nous ne voulons pas analyser en détail les résultats mais faire part seulement de ces deux observations :

– On voit, sur les figures 1, 2, 3, 6, 7, 8, 9, une structure en quatre classes : la première contient les 50 iris du premier groupe, la seconde, environ 20 iris du second, la troisième, 30 du troisième et la quatrième, les individus restant des 2$^e$ et 3$^e$ groupes.

– Par contre, les figures 4 et 5 ne suggèrent aucune structure en classe.

### 4.3. Commentaires

Le choix de nos indices pour cet exemple a été fait de manière à illustrer l'évolution de la structure mise en évidence par la classification, en fonction de deux paramètres :

— *Le changement des coefficients dans l'indice de proximité inter-classe.*

— *Le changement de la distance entre les individus.*

Deux remarques importantes sur ces résultats s'imposent, à notre avis :

(A) La plupart des indices choisis retrouvent bien une unique structure des iris : un groupe 1 bien séparé, les groupes 2 et 3 se recouvrant selon environ la moitié de leurs éléments et étant bien séparés selon l'autre moitié. Il semble donc que cette structure est bien la vraie structure des données.

Cette première remarque montre que, sur cet exemple, le choix de tel ou tel de ces indices n'a pas une importance décisive pour identifier la structure des données, même si les coefficients sont très différents par rapport à la manière dont ils prennent en compte l'effet « taille » des classes.

(B) On voit que la structure des classes peut être sensible à l'effet de la dissimilarité. En effet l'indice de Ward et l'indice 3 donne la même structure aux données selon la norme $L1$ ou $L2$. Par contre, les indices 1 et 2 ne leur donne aucune structure intéressante selon la norme $L1$ (cf. *fig.* 4 et 5).

Dès lors, nous avons voulu voir si, du fait qu'utiliser la norme $L1$ revient à prendre la racine carrée des termes dans $\Delta_{ii}$, $\Delta_{ij}$, $\Delta_{jj}$ (par rapport à la norme $L2$), en prenant également la racine des carrés des coefficients de pondération, on modifié le résultat. Ceci a donné lieu au nouvel indice cité plus haut. On voit (*fig.* 9) que oui et que l'on retrouve la structure des autres situations.

Notre principale conclusion de cette étude est que, ni le choix de la dissimilarité ni celui des coefficients n'est indifférent en vue de déterminer une structure intéressante à interpréter pour l'utilisateur. Il conviendrait de le conseiller sur l'ensemble des choix possibles à faire parmi les dissimilarités et les indices de proximité inter-classe, pour obtenir une analyse complète et exhaustive de ses données : elle devrait lui permettre de conclure si, oui ou non, il peut y trouver cette structure intéressante.

## 5. CONCLUSION

Nous avons défini une famille très générale d'indices de proximités inter-classe ayant les bonnes propriétés pratiques requises (formules de réactualisation).

La définition de cette famille nous semble avoir un intérêt double :

Sur le plan théorique, elle permet d'unifier la collection hétéroclite d'indices que, jusqu'ici, l'utilisateur avait à disposition, en en faisant ressortir les caractéristiques mathématiques communes qui les sous-tendent : combinaison linéaire d'indices d'homogénéité et de séparation de classes (partie 1).

Sur le plan pratique, elle permet de s'intéresser à beaucoup plus d'indices qu'auparavant.

On pourrait imaginer, par exemple de conseiller à un utilisateur d'essayer des indices de la forme :

$$-\frac{1}{2n_i^{\alpha}}\sum_{k,\ l\in C_i} d\beta(X_1, X_k)+\frac{2}{(n_i n_j)^{\alpha/2}}\sum_{\substack{k\in C_i\\ l\in C_j}} d\beta(X_1, X_k)-\frac{1}{2n_j^{\alpha}}\sum_{k,\ l\in C_j} d\beta(X_1, X_k)$$

et de faire varier $\alpha$ et $\beta$ jusqu'à ce qu'il trouve une structure intéressante à interpréter.

Par ailleurs, nous n'avons pas traité des problèmes du genre :

Peut-on faire aisément une classification hiérarchique avec des indices de proximités inter-classe multivariés? ou la statistique de Fisher-Snedecor? ou l'indice :

$$\frac{\Delta_{ij}}{\sqrt{\Delta_{ii}\Delta_{jj}}}$$

REMERCIEMENTS

Nous tenons ici tout particulièrement à remercier I. C. Lerman ainsi que les rapporteurs de ce travail pour toutes leurs suggestions qui nous ont permis d'en améliorer sensiblement la qualité.

## BIBLIOGRAPHIE

1. M. R. Andergerg, *Cluster Analysis for Applications*, Academic Press, 1973.
2. F. Caillez et J. P. Pages, *Introduction à l'analyse des données*, SMASH, 1976.
3. J. L. Chandon et S. Pinson, *Analyse Typologique*, Masson, 1981.
4. E. Diday et coll. *Éléments d'analyse des données*, Dunod, 1982.
5. L. Hubert, *Some Applications of Graph Theory to Clustering*, Psychometrika, vol. 39, 1974, p. 283-309.
6. A. K. Jain et R. C. Duber, *Algorithms for Clustering Data*, Prentice Hall, 1988.

7. M. JAMBU, *Cluster Analysis and Data Analysis*, North-Holland, 1983.
8. G. N. LANCE et W. T. WILLIAMS, *A General Theory of Classification Sorting Strategies: Hierarchical Systems*, The Computer Journal, vol. 9, 1966, p. 373-380.
9. G. N. LANCE et W. T. WILLIAMS, *A General Theory of Classification Sorting Strategies : Clustering Systems*, The Computer Journal, vol. 10, 1967, p. 271-277.
10. I. C. LERMAN, *Classification et analyse ordinaire des données*, Dunod, 1981.
11. I. C. LERMAN, *Formules de réactualisation en cas d'aggrégations multiples*, Rapport N 409, mai 1988, IRIS, Rennes.
12. SAS *User's Guide*, Version 5, 1985.
13. G. A. F. SEBER, *Multivariate Observations*, Wiley, 1984.
14. SPSSX, *User's Guide*, Version 2.1, 1986.