M. JAMBU

Caractérisation des classifications arborescentes établies sur le critère du χ^2 ou de l'information mutuelle

Revue de statistique appliquée, tome 26, n° 2 (1978), p. 45-69
<http://www.numdam.org/item?id=RSA_1978__26_2_45_0>

CARACTÉRISATION DES CLASSIFICATIONS ARBORESCENTES ÉTABLIES SUR LE CRITÈRE DU χ^2 OU DE L'INFORMATION MUTUELLE

M. JAMBU

Attaché de recherches au CNRS
Université Pierre et Marie Curie (ParisVI)

INTRODUCTION

La difficulté d'interprétation des classifications arborescentes décourage parfois les taxinomistes. Pour remédier en partie à cette difficulté, on propose d'utiliser des algorithmes de classification fondés sur le critère du χ^2 ou de l'information mutuelle. Les classifications hiérarchiques établies à partir de ces critères permettent de formuler certains calculs propres à caractériser les classes obtenues par classification. On a illustré ces calculs sur un exemple de données portant sur une étude internationale de budgets familiaux.

1. INFORMATION MUTUELLE ET ENTROPIE – DEFINITIONS, NOTATIONS ET RESULTATS PRINCIPAUX (Cf. [1])

1.1. Lois de probabilités

Soit I et J deux ensembles finis et $k_{IJ} = \{k_{ij}, i \in I, j \in I\}$ le tableau des données matérialisant la correspondance entre les ensembles I et J. Dans l'exemple présenté, k_{IJ} est un tableau de contingence.

$$\text{Soit}\quad k_I = \left\{ k_i = \sum_{j\in J} k_{ij}\ ;\ i \in I \right\}$$

$$k_J = \left\{ k_j = \sum_{i\in I} k_{ij}\ ;\ j \in J \right\}, \qquad t = \sum_{i,j\in I\times J}$$

à partir de k_{IJ}, k_I, k_J, on calcule f_{IJ}, f_I, f_j

$$f_{IJ} = \{f_{ij} = k_{ij}/t\ ;\ i \in I\ ;\ j \in J\}$$

$$f_I = \{f_i = k_i/t\ ;\ i \in I\}$$

$$f_J = \{f_j = k_j/t\ ;\ j \in J\}$$

Mots-clés : **Analyse des données, Classification automatique, Théorie de l'information**

$$f_J^i = \left\{ \frac{f_{ij}}{f_j} \text{ avec } f_j \neq 0 \,;\, j \in J \right\} \quad \text{profil de i sur J}$$

$$f_I^i = \left\{ \frac{f_{ij}}{f_i} \text{ avec } f_i \neq 0 \,;\, i \in I \right\} \quad \text{profil de j sur I}$$

$f_{IJ}, f_I, f_J, f_I^j, f_J^i$ vérifient les axiomes des lois de probabilités

1.2. Lois de probabilités et partitions

Considérons une partition Q sur J et une partition K sur I. On définit les tableaux suivants :

$$f_{IQ} = \left\{ f_{iq} \,;\, f_{iq} = \sum_{j \in q} f_{ij} \mid i \in I, q \in Q \right\} \,;\, f_{KJ} = \left\{ f_{kj} \,;\, f_{kj} = \sum_{i \in k} f_{ij} \mid k \in K . j \in J \right\}$$

$$f_{KQ} = \left\{ f_{kq} \,;\, f_{kq} = \sum_{i \in k, j \in q} f_{ij} \mid k \in K \,,\, q \in Q \right\}$$

$$f_Q^i = \left\{ f_q^i \,;\, f_q^i = \frac{f_{iq}}{f_i} \,;\, f_i \neq 0 \,;\, q \in Q \right\} \qquad \text{profil de l'élément i dans Q}$$

$$f_K^j = \left\{ f_k^i \,;\, f_k^j = \frac{f_{kj}}{f_j} \,;\, f_j \neq 0 \,;\, k \in K \right\} \qquad \text{profil de l'élément j dans K}$$

$$f_I^q = \left\{ f_i^q \,;\, f_i^q = \frac{f_{iq}}{f_q} \,;\, f_q \neq 0 \,;\, i \in I \right\} \qquad \text{profil de la classe q de J dans I}$$

$$f_j^k = \left\{ f_j^k \,;\, f_j^k = \frac{f_{jk}}{f_k} f_k \neq 0 \,;\, j \in J \right\} \qquad \text{profil de la classe k dans J}$$

Ces tableaux définissent des lois de probabilités sur les ensembles notés en indices inférieurs.

1.3. Entropie d'une loi de probabilité f_I sur un ensemble fini I

Soit $i \in I$, i aléatoire de loi f_I on notera

$$H(I) \stackrel{\text{def}}{=} H(f_I) \stackrel{\text{def}}{=} \sum_{i \in I} - f_i \log_2 f_i$$

$H(f_I)$ est l'entropie associée à l'ensemble I muni de la loi de probabilité f_I.

On aura de même pour (i , j) a I x J, (i , j) aléatoire de loi f_{IJ}

$$H(IJ) \stackrel{\text{def}}{=} H(f_{IJ}) \stackrel{\text{def}}{=} \sum_{\substack{i \in I \\ j \in J}} - f_{ij} \log_2 f_{ij}$$

On a $H(f_{IJ}) \leqslant H(f_I) + H(f_J)$ (1)

1.4. Entropie conditionnelle et information mutuelle

On note H(J/I) l'information conditionnelle, transmise par $j \in J$ quand $i \in I$ est déterminé :

$$H(J/I) \stackrel{def}{=} \sum_{i \in I} f_i \, H(f_J^i)$$

$$H(J/I) = H(IJ) - H(I) \quad (2)$$

On note $H(f_{IJ} \, ; f_I \, f_J)$ l'information mutuelle entre les variables i de I et j de J

$$H(f_{IJ} \, ; f_I \, f_J) = H(I) - H(I/J) \quad (3)$$

$$= H(I) + H(J) - H(IJ) \quad (4)$$

Si $f_{IJ} = f_I \, . \, f_J \quad \Leftrightarrow \quad H(f_{IJ} \, ; f_I \, f_J) = 0$ (5)

1.5. Entropie relative et information mutuelle (Cf. [5])

Soient f_I et g_I deux lois de probabilités sur un même ensemble I fini.

On pose $H(f_I \, ; g_I) \stackrel{def}{=} \sum_{i \in I} f_i \log_2 (f_i/g_i)$ (6)

$$= \sum_{i \in I} g_i \frac{f_i}{g_i} . \log_2 \left(\frac{f_i}{g_i}\right) \quad (7)$$

$$= \sum_{i \in I} g_i \, \phi\left(\frac{f_i}{g_i}\right) \qquad \text{avec } \phi(x) = x \log_2 x$$

On définit l'information mutuelle à partir de l'entropie relative de la loi f_{IJ} et de la loi produit des lois f_I et f_J

$$H(f_{IJ} \, ; f_I \, f_J) \stackrel{def}{=} \sum_{i,j \in IxJ} f_{ij} \log_2 \frac{f_{ij}}{f_i f_j}$$

$$= \sum_{i,j \in IxJ} f_i f_j \, \phi\left(\frac{f_{ij}}{f_i f_j}\right) \quad (8)$$

1.6. Information mutuelle,entropie et partitions

Soit Q une partition de J et K une partition de I

On a les résultats suivants :

$$H(f_Q) \leqslant H(f_J) \quad (9)$$

$$H(f_K) \leqslant H(f_I) \quad (10)$$

$$H(f_{KQ}\ ;\ f_K \, . \, f_Q) \leqslant H(f_{KJ}\ ;\ f_K f_J) \leqslant H(f_{IJ}\ ;\ f_I f_J) \tag{11}$$

$$H(f_{KQ}\ ;\ f_K f_Q) \leqslant H(f_{IQ}\ ;\ f_I f_Q) \leqslant H(f_{IJ}\ ;\ f_I f_J) \tag{12}$$

2.INFORMATION MUTUELLE ET DISTANCE DU χ^2

On définit classiquement

$$||f_{IJ} - f_I \, . \, f_J||^2_{f_I f_J} = \sum_{\substack{i,\in I \\ j \in J}} (f_{ij} - f_i f_j)^2 / f_i f_j \tag{13}$$

$$= \sum_{\substack{i \in I \\ j \in J}} f_i f_j \left(\left(\frac{f_{ij}}{f_i f_j} \right)^2 - 1 \right)$$

$$= \sum_{\substack{i \in I \\ j \in J}} f_i f_j \, . \, \Psi \left(\frac{f_{ij}}{f_i f_j} \right) \quad \text{avec } \Psi(x) = x^2 - 1$$

D'après 8 on a

$$H(f_{IJ}\ ;\ f_I f_J) = \sum_{\substack{i \in I \\ j \in J}} f_i f_j \, \phi \left(\frac{f_{ij}}{f_i f_j} \right) \quad \text{avec } \phi(x) = x \log_2 x$$

Ces deux quantités

$$(||f_{iJ} - f_I f_J||^2 f_I f_J \text{ et } H(f_{IJ}\ ;\ f_I f_J))$$

l'une provenant de la statistique du χ^2 et l'autre provenant de l'entropie relative servent à comparer un eloi produit f_{IJ} à la loi définie par le produit des lois f_I et f_J. On étend ces définitions si I ou J sont suivi d'une partition (Q de J ou K de I). Ces deux quantités servent de base à l'élaboration de classifications ascendantes hiérarchiques dont nous allons rappeler les constructions.

3. INFORMATION MUTUELLE, DISTANCE DU χ^2 ET PARTITIONS

Soit Q une partition de J et K une partition de I

Soit N_I (Q) le nuage associé à la partition Q, et N_J (K) le nuage correspondant à la partition K. q une classe de Q, k une classe de k

$$N_I(Q) = \left\{ \begin{array}{l} q \in Q,\ i^{\text{ème}} \text{ coordonnée de } q \text{ dans } R_I = \dfrac{f_{iq}}{f_q}, \text{ masse affectée} \\ \text{au point } q = f_q \end{array} \right\}$$

$$N_J(K) = \left\{ \begin{array}{l} k \in K,\ j^{\text{ème}} \text{ coordonnée de } k \text{ dans } R_J = \dfrac{f_{ki}}{f_k}, \text{ masse affectée} \\ \text{au point } k = f_k \end{array} \right\}$$

$$N_I(Q) = \{f_I^q\ , f_q\} \subset R_I\ ; N_J(K) = \{f_J^K\ , f_k\} \subset R_J$$

$$N_K(Q) = \left\{ \begin{array}{l} q \in Q\ ;\ k^{\text{ième}} \text{ coordonnée de } q \text{ dans } R_K = \dfrac{f_{kq}}{f_q}, \text{ masse affectée} \\ \text{au point } q = f_q \end{array} \right\}$$

$$N_Q(K) = \left\{ \begin{array}{l} k \in K,\ q^{\text{ième}} \text{ coordonnée de } k \text{ dans } R_Q = \dfrac{f_{kq}}{f_k}, \text{ masse affectée} \\ \text{au point } k = f_k \end{array} \right\}$$

$$N_K(Q) = \{f_K^q\ , f_q\} \subset R_K\ .\ N_Q(K) = \{f_Q^k\ , f_k\} \subset R_Q$$

On définit : $\text{Lien}_2\ (Q\ , K) = ||f_{QK} - f_Q f_K||^2_{f_Q f_K}$

$$= \sum_{\substack{q \in Q \\ k \in K}} (f_{qk} - f_k f_q)^2 / f_q f_k \qquad f_k = \sum_{i \in k} f_i$$

$$f_q = \sum_{j \in q} f_j$$

$$= M^2\ (N_Q(K)) = M^2(N_K(Q)) \qquad f_{qk} = \sum_{\substack{i \in k \\ j \in q}} f_{ij}$$

$$\text{Lien}_1\ (Q\ , K) = H(f_{QK}\ ; f_Q f_K)$$

$$= \sum_{\substack{q \in Q \\ k \in K}} f_q f_k\ \phi\left(\frac{f_{qk}}{f_q f_k}\right)$$

$$= E(N_K(Q)) = E(N_Q(K))$$

4. CLASSIFICATIONS HIERARCHIQUES SELON LES CRITERES DU χ^2 OU DE L'INFORMATION MUTUELLE (Cf [1], [2])

4.1. Description mathématique des classifications – Rappel des notations principales

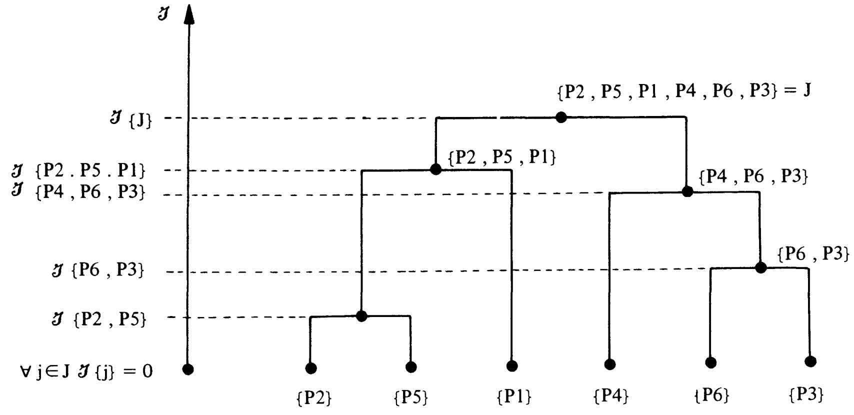

Figure 1 – Représentation d'une hiérarchie de parties totale et indicée. A(J) sur J

Soit J fini, $A \subset \mathcal{P}(J) = \phi$; A ordonné par la relation on note :

Som (A)	l'ensemble des éléments maximaux de A
Ter (A)	l'ensemble des éléments minimaux de A
Nod (A)	l'ensemble des éléments non minimaux de A
Pred (a , A)	l'ensemble des prédécesseurs de l'élément a dans A
Suc (a , A)	l'ensemble des successeurs de l'élément a dans A
Suci (a , A)	l'ensemble des successeurs immédiats de a dans A

$\mathcal{J}$(a) est l'indice de diamètre (ou d'agrégation, ou de stratification) de l'élément a de A.

$$\mathcal{J}\{j\} = 0 \qquad \forall j \in J$$

Les classifications hiérarchiques indicées totales se représentent comme la succession de partitions emboîtées et sont représentables graphiquement selon le dessin ci-dessus.

4.2. Construction des classifications hiérarchiques et critères d'agrégation

Le principe de la construction est d'élaborer pas à pas une succession de partitions emboîtées depuis la partition la plus fine $\{\{j\} ; j \in J\}$ jusqu'à la partition la plus grossière $\{J\}$. *La classification est effectuée sur l'un des deux ensembles en relation* (ici J)

On appelle Lien_1 (I , J) la fonction $H(f_{IJ} ; f_I f_J)$ et Lien_2 (I , J) la fonction $\| f_{IJ} - f_{IJ} \|^2 f_I f_J$ qui définissent la dépendance globale entre les ensembles I et J. Pour créer des partitions successives, il suffira d'agréger pas à pas des éléments de J qui minimise la variation de la fonction Lien entre deux partitions successives – selon le schéma suivant :

Etape de rang 1 – On cherche à construire sur J une partition, soit Som $(H_1) = Q$ définie de la façon suivante : Som (H_1) est la partition de J obtenue en considérant l'ensemble J lui-même auquel on a retiré deux éléments et auquel on a ajouté l'élément réunion des deux éléments retirés ;

La partition Q = Som (H_1) choisie sera celle qui minimise la fonction suivante :

$$\text{Lien (I . J)} - \text{Lien (I, Som } H_1)$$

l'indice de diamètre de la classe créée sera égale à la variation de cette fonction.

Passage de l'étape de rang (n – 1) à l'étape de rang (n). A l'étape de rang (n – 1) on a établi une partition Q_1 = Som (H_{n-1}). On cherche à établir une partition Q_1 = Som (H_n) de Som (H_{n-1}) selon la formulation précédente (agrégation de deux classes, extraction des deux classes qui servent à la réunion pour former un nouvel élément). On cherche la partition Som (H_n) telle que soit minimum la fonction suivante :

$$\text{Lien (I , som } H_{n-1}) - \text{Lien (I , Som } H_n)$$

l'indice de la classe nouvellement obtenue est égale à cette variation.

Dernière étape – On agrège les deux classes qui restent à réunir et l'indice de la dernière classe est égal à Lien (I , Som H_{n-1}) – Lien (I , $\{J\}$)

4.2.1. Cas du χ^2

On a les résultats suivants pour la hiérarchie construite sur J

$$\mathcal{J}\text{ (n)} = \mathcal{J}\text{ (sut)} = \text{Lien}_2 (I , Q_{n-1}) - \text{Lien} (I , Q_n)$$

$\mathcal{J}$ indice de la classe n
$s \cup t = n ; s \cap t = \phi$

$$= M^2(N_I(Q_{n-1})) - M^2(N_I(Q_n))$$

$$= \frac{f_t \cdot f_s}{f_t + f_s} \| f_I^s - f_I^t \|^2_{f_I}$$

$$= M^2(N_I(s \cup t)) - M^2(N_I(s)) - M^2(M_I(t)) \text{ (Cf. [2])}$$

4.2.2. Cas de l'information mutuelle

$$\mathcal{J}\text{ (n)} = \text{Lien}_1(I , Q_{n-1}) - \text{Lien}(I , Q_n)$$

$$= E(N_I(Q_{n-1}) - E(N_I(Q_n))$$

(E = entropie relative cf. 1.5)

avec

$$= H(f_{IQ_{n-1}} ; f_I f_{Q_{n-1}}) - H(f_{IQ_n} ; f_I f_Q)$$

5. TAUX D'INERTIE ET D'ENTROPIE ASSOCIES A UNE PARTITION Q D'UNE CLASSIFICATION HIERARCHIQUE

Soient $A_1(I)$, $A_2(I)$ les classifications hiérarchiques établies sur I à partir du critère d'entropie (pour $A_1(I)$) et du critère du χ^2 (pour $A_2(I)$).

Soient $B_1(J)$, $B_2(J)$ les classifications hiérarchiques établies sur J correspondantes à l'utilisation des mêmes critères.

Soient $\mathcal{I}_{I1}$ l'indice associé à $A_1(I)$, $\mathcal{I}_{I2}$ l'indice associé à $A_2(I)$, $\mathcal{I}_{J1}$ l'indice associé à $B_1(J)$, $\mathcal{I}_{J2}$ l'indice associé à $B_2(J)$.

En sommant les indices des noeuds d'une même classification hiérarchique, on obtient dans chacune des constructions les formules suivantes :

$$\mathrm{Lien}_1(I, J) = E(N_J(I)) = E(N_I(J))$$

$$= \sum_{n \in \mathrm{Nod}(A_1(I))} \mathcal{I}_{I1}(n) = \sum_{n \in \mathrm{Nod}(B_1(J))} \mathcal{I}_{J1}(n)$$

$$\mathrm{Lien}_2(I,J) = M^2(N_J(I)) = M^2(N_I(J))$$

$$= \sum_{n \in \mathrm{Nod}(A_2(I))} \mathcal{I}_{I2}(n) = \sum_{n \in \mathrm{Nod}(A_2(I))} \mathcal{I}_{J2}(n)$$

Ces formules permettent de calculer soit un taux d'inertie par noeud et par partition dans le cas de classification établie selon le critère du χ^2, soit un taux d'entropie par noeud et par partition dans le cas d'utilisation du critère d'information mutuelle.

Taux τ associé à une classe d'une classification hiérarchique

Taux τ d'entropie pour le noeud n de la classification $A_1(I)$

$$\tau_1(n) = \frac{\mathcal{I}_{I1}(n)}{\mathrm{Lien}_1(I, J)} \quad \left[\sum_{n \in \mathrm{Nod}(A_1(I))} \tau_1(n) = 1\right]$$

On calcule de même un taux d'inertie pour chaque noeud d'une hiérarchie établie selon le critère du χ^2.

Taux T associé à une partition Q d'une classification hiérarchique

Soit $A_1(Q_n)$ la classification hiérarchique déduite de $A_1(I)$, dont l'ensemble terminal est la partition Q_n ($\mathrm{Ter}\ A_1(Q_n) = Q_n$) correspondant au noeud de numéro $(n - 1)$.

Par définition on pose :

$$T_1(Q_n) = \frac{E(N_J(Q_n))}{E(N_J(I))} = \frac{\mathrm{Lien}_1(Q_n, J)}{\mathrm{Lien}_1(I, J)}$$

En posant $X = \sum_{n \in Nod(A_1(I))} \mathcal{J}_{I1}(n)$; $Y = \sum_{n \in Nod(A_1(I)) - Nod(A_1(Q_n))} \mathcal{J}_{I1}(n)$;

On a $\text{Lien}_1(Q_n, I) = X - Y$

d'où l'on tire $$T_1(Q_n) = \frac{X - Y}{X} = 1 - \frac{Y}{X} \text{ avec } \frac{Y}{X} \leqslant 1$$

On aura des calculs analogues avec Lien_2 (τ_2, T_2). D'un point de vuė pratique, ces calculs donnent une idée de l'importance d'une partition dans la classification hiérarchique.

6. ILLUSTRATION DES CALCULS SUR UN EXEMPLE

6.1. Présentation de l'exemple

L'objet principal d'une étude portant sur les budgets-temps est de déterminer une typologie tant des activités exercées par la population étudiée, que de la population observée, en ne tenant compte que du temps passé à exercer ces activités. Les tableaux de données illustrant un budget-temps se présentent alors ainsi :

Soit I la population étudiée et J l'ensemble des activités exerçables par la population I (exerçables pendant un temps fixé à l'avance, par exemple, une semaine, un mois, . . .). La case d'indice (i , j) du tableau des données I x J contient le nombre d'heures (ou de minutes) passées à exercer l'activité j de J, par le sujet i de I pendant le temps fixé de l'enquête. L'exemple qui nous sert d'illustration est un cas particulier de ce cas général. L'ensemble I est constitué de 28 classes de sujets dont les variables de base sont le sexe, le pays, l'activité professionnelle, le mariage. L'ensemble J est constitué de 10 classes d'activités regroupées pour les besoins de la comparaison des budgets-temps. La case d'indice (i , j) contient le nombre d'heures que la classe i de sujets de l'enquête a passé à exercer l'activité j de J pendant le temps fixé de l'enquête.

Composition de l'ensemble I

HAUS – hommes actifs des U.S.A.
FAUS – femmes actives des U.S.A.
FNAU – femmes non actives des U.S.A.
HMUS – hommes mariés des U.S.A.
FMUS – femmes mariées des U.S.A.
HCUS – hommes célibataires des U.S.A.
HAWE – hommes actifs des pays de l'Ouest
FAWE – femmes actives des pays de l'Ouest
FNAW – femmes non actives des pays de l'Ouest
HMWE – hommes mariés des pays de l'Ouest
FMWE – femmes mariées des pays de l'Ouest
HCWE – hommes célibataires des pays de l'Ouest

Composition de l'ensemble J

PROF – Travail professionnel
TRAN – Occupations dues ou liées au travail professionnel (transports . . .)
MENA – Travail ménager
ENFA – Occupations liées aux enfants
COUR – Les courses

HAES – hommes actifs des pays de l'Est
FAES – femmes actives des pays de l'Est
FNAE – femmes non actives des pays de l'Est
HMES – hommes mariés des pays de l'Est
FMES – femmes mariées des pays de l'Est
HCES – hommes célibataires des pays de l'Est
HAYO – hommes actifs de Yougoslavie
FAYO – femmes actives de Yougoslavie
FNAY – femmes non actives de Yougoslavie
HMYO – hommes mariés de Yougoslavie
FMYO – femmes mariées de Yougoslavie
HCYO – hommes célibatiares de Yougoslavie
FCUS – femmes célibataires des U.S.A.
FCWE – femmes célibataires des pays de l'Ouest
FCES – femmes célibataires des pays de l'Est
FCYO – femmes célibataires de Yougoslavie

TOIL – La toilette, les soins personnels
REPA – Les repas
SOMM – Le sommeil
TELE – La télévision
LOIS – Les autres loisirs

	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS
HAUS	610.	140.	60.	10.	120.	95.	115.	760.	175.	315.
FAUS	475.	90.	250.	30.	140.	120.	100.	775.	115.	305.
FNAU	10.	0.	495.	110.	170.	110.	130.	785.	160.	430.
HMUS	615.	141.	65.	10.	115.	90.	115.	765.	180.	305.
FMUS	179.	29.	421.	87.	161.	112.	119.	776.	143.	373.
HCUS	585.	115.	50.	0.	150.	105.	100.	760.	150.	385.
FCUS	482.	94.	196.	18.	141.	130.	96.	775.	132.	336.
HAWE	652.	100.	95.	7.	57.	85.	150.	807.	115.	330.
FAWE	510.	70.	307.	30.	80.	95.	142.	815.	87.	262.
FNAW	20.	7.	567.	87.	112.	90.	180.	842.	125.	367.
HMWE	655.	97.	97.	10.	52.	85.	152.	807.	122.	320.
FMWE	168.	22.	529.	69.	102.	83.	174.	825.	119.	392.
HCWE	642.	105.	72.	0.	62.	77.	140.	812.	100.	387.
FCWE	389.	34.	262.	14.	92.	97.	147.	848.	84.	392.
HAYO	650.	140.	120.	15.	85.	90.	105.	760.	70.	365.
FAYO	560.	105.	375.	45.	90.	90.	95.	745.	60.	235.
HMYO	650.	145.	112.	15.	85.	90.	105.	760.	80.	357.
FNAY	10.	10.	710.	55.	145.	85.	130.	815.	60.	380.
FMYO	260.	52.	576.	59.	116.	85.	117.	775.	65.	295.
HCYO	615.	125.	95.	0.	115.	80.	85.	760.	40.	475.
FCYO	433.	89.	318.	23.	112.	96.	102.	774.	45.	409.
HAES	650.	142.	122.	22.	76.	94.	100.	764.	96.	334.
FAES	578.	106.	338.	42.	106.	94.	92.	752.	64.	228.
FNAE	24.	8.	594.	72.	158.	92.	128.	840.	86.	398.
HMES	652.	133.	134.	22.	68.	94.	102.	762.	122.	310.
HCES	627.	148.	68.	0.	88.	92.	86.	770.	58.	463.
FMES	434.	77.	431.	60.	117.	88.	105.	770.	73.	229.
FCES	433.	86.	296.	21.	128.	102.	94.	798.	58.	379.

Figure 2 – Tableau des données

6.2 Classification hiérarchiques établies sur l'ensemble I des classes de sujets

On a effectué les classifications selon les critères du χ^2 et de l'information mutuelle sur l'ensemble noté I des sujets.

6.2.1. Cas de l'application du critère du χ^2

1	2	3	4	5	6	7	8
29	0	0	99	1	4	2	HAUS HMUS
30	0	0	99	8	11	2	HAWE HMWE
31	0	0	99	15	17	2	HATO HMTO
32	0	0	99	16	23	2	FATO FAES
33	0	0	99	28	21	2	FCES FCTO
34	0	0	99	22	25	2	HAES HMES
35	0	0	99	26	20	2	HCES HCTO
36	0	0	99	2	7	2	FAUS FCUS
37	0	0	99	30	13	3	HAWE HMWE HCWE
38	0	0	99	34	31	4	HAES HMES HATO HMTO
39	0	0	98	29	6	3	HAUS HMUS HCUS
40	0	0	98	18	24	2	FNAT FNAE
41	0	0	98	10	12	2	FNAW FMWE
42	0	0	97	9	32	3	FAWE FATO FAES
43	0	0	96	33	14	3	FCES FCTO FCWE
44	0	0	96	3	5	2	FNAU FMUS
45	0	0	95	42	27	4	FAWE FATO FAES FMES
46	1	0	94	38	37	7	HAES HMES HATO HMTO HAWE HMWE HCWE
47	1	1	93	44	41	4	FNAU FMUS FNAW FMWE
48	1	1	92	36	43	5	FAUS FCUS FCES FCTO FCWE
49	2	1	90	40	19	3	FNAT FNAE FMTO
50	2	1	89	46	35	9	HAES HMES HATO HMTO HAWE HMWE HCWE HCES HCTO
51	3	2	86	47	49	7	FNAU FMUS FNAW FMWE FNAT FNAE FMTO
52	3	2	84	39	50	12	HAUS HMUS HCUS HAES HMES HATO HMTO HAWE HMWE HCWE HCES HCTO
53	4	2	81	48	45	9	FAUS FCUS FCES FCTO FCWE FAWE FATO FAES FMES
54	21	15	66	52	53	21	HAUS HMUS HCUS HAES HMES HATO HMTO HAWE HMWE HCWE HCES HCTO FAUS FCUS FCES FCTO FCWE FAWE FATO FAES FMES
55	94	66	0	54	51	28	HAUS HMUS HCUS HAES HMES HATO HMTO HAWE HMWE HCWE HCES HCTO FAUS FCUS FCES FCTO FCWE FAWE FATO FAES FMES FNAU FMUS FNAW FMWE FNAT FNAE FMTO

Figure 3 – Description de la classification hiérarchique établie selon le critère du χ^2

- La colonne 1 consigne les numéros N des noeuds dans l'édification de la hiérarchie. Les noeuds sont numérotés de CARDI + 1 à 2∗CARDI – 1, (si CARDI désigne le nombre d'éléments de l'ensemble I).
- La colonne 2 consigne les valeurs $\{100 * \mathcal{J}_{I2}(N)\}$ où $\mathcal{J}_{I2}(N)$ est l'indice de diamètre de la classe N (Cf. § 4).
- La colonne 3 consigne les valeurs $\{100 * \mathcal{T}_2(N)\}$ où $\tau_2(N)$ est le taux d'inertie par classe N (Cf. § 5).
- La colonne 4 consigne les valeurs $\{T_2(N) * 100\}$ où $T_2(N)$ est le taux d'inertie associé à la partition Q_N (Cf. § 5).
- La colonne 5 consigne les aînés A(N) (un des successeurs immédiats) de chaque classe N.
- La colonne 6 consigne les benjamins B(N) (l'autre successeurs immédiat) de chaque classe N.
- La colonne 7 consigne les cardinaux P(N) de chaque classe N.
- La colonne 8 consigne les éléments de I qui constitue la classe N.

On donne une représentation graphique de la classification hiérarchique à la figure 4.

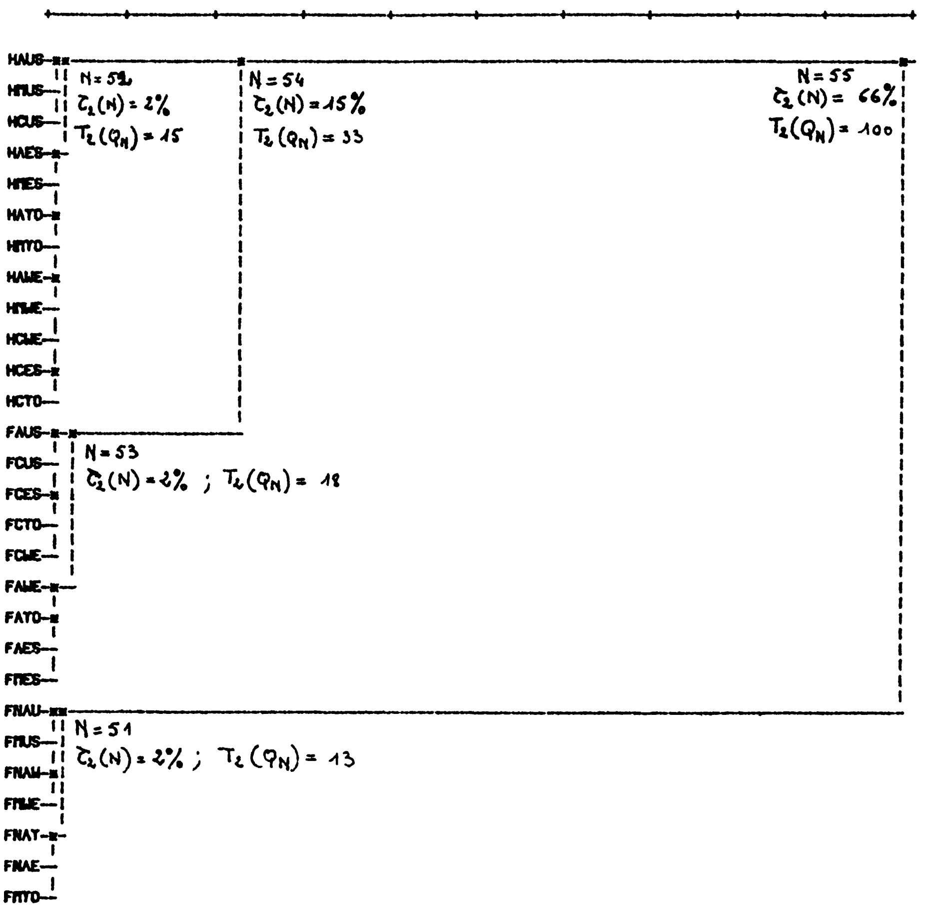

Figure 4 – Classification hiérarchique établie selon le critère du χ^2. Représentation des taux d'inertie afférant aux noeuds de la classification

6.2.2. Cas de l'application du critère de l'information mutuelle

1	2	3	4	5	6	7	8
29	0	0	99	1	4	2	HAUS HMUS
30	0	0	99	15	17	2	HATO HMTO
31	0	0	99	8	11	2	HAWE HMWE
32	0	0	99	16	23	2	FATO FAES
33	0	0	99	22	25	2	HAES HMES
34	0	0	99	28	21	2	FCES FCTO
35	0	0	99	2	7	2	FAUS FCUS
36	0	0	99	26	20	2	HCES HCTO
37	0	0	99	33	30	4	HAES HMES HATO HMTO
38	0	0	99	24	18	2	FNAE FNAT
39	0	0	98	31	13	3	HAWE HMWE HCWE
40	0	0	98	29	6	3	HAUS HMUS HCUS
41	0	0	98	5	12	2	FMUS FMWE
42	0	0	97	27	32	3	FMES FATO FAES
43	0	0	97	10	38	3	FNAW FNAE FNAT
44	0	0	96	42	9	4	FMES FATO FAES FAWE
45	0	0	96	34	14	3	FCES FCTO FCWE
46	1	0	95	3	43	4	FNAU FNAW FNAE FNAT
47	1	0	94	41	19	3	FMUS FMWE FMTO
48	1	1	93	37	39	7	HAES HMES HATO HMTO HAWE HMWE HCWE
49	1	1	92	35	45	5	FAUS FCUS FCES FCTO FCWE
50	2	1	90	36	48	9	HCES HCTO HAES HMES HATO HMTO HAWE HMWE HCWE
51	2	2	87	40	50	12	HAUS HMUS HCUS HCES HCTO HAES HMES HATO HMTO HAWE HMWE HCWE
52	3	2	85	49	44	9	FAUS FCUS FCES FCTO FCWE FMES FATO FAES FAWE
53	8	6	78	46	47	7	FNAU FNAW FNAE FNAT FMUS FMWE FMTO
54	20	16	61	51	52	21	HAUS HMUS HCUS HCES HCTO HAES HMES HATO HMTO HAWE HMWE HCWE FAUS FCUS FCES FCTO FCWE FMES FATO FAES FAWE
55	74	61	0	54	53	28	HAUS HMUS HCUS HCES HCTO HAES HMES HATO HMTO HAWE HMWE HCWE FAUS FCUS FCES FCTO FCWE FMES FATO FAES FAWE FNAU FNAW FNAE FNAT FMUS FMWE FMTO

Figure 5 – Tableau des paramètres de la classification hiérarchique établie selon le critère de l'information

Les colonnes 1, 2, 5, 6, 7, 8 ont la même signification que dans le cas précédent.

La colonne 3 consigne les valeurs $\{\tau_1(N) * 100\}$, où $\tau_1(N)$ désigne le taux d'entropie associé à la classe de numéro N.

La colonne 4 consigne les valeurs $\{T_1(N)) * 100\}$ où $T_1(N)$ désigne le taux d'entropie de la partition Q_N définie au niveau de noeud de numéro N.

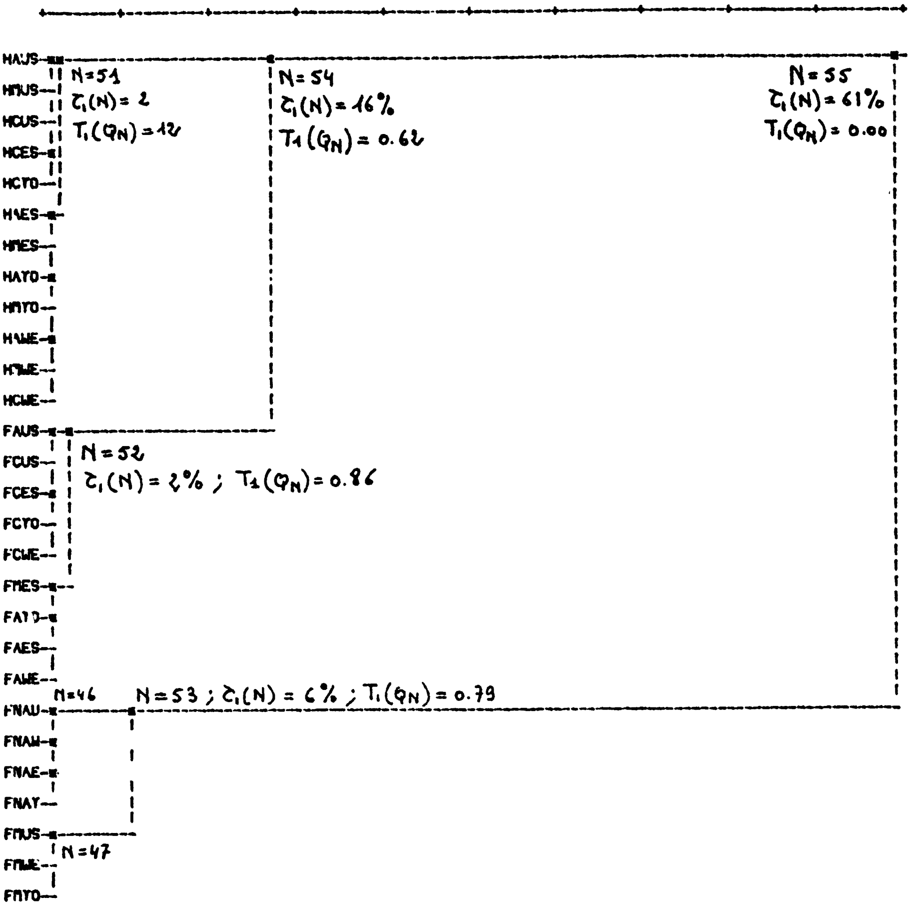

Figure 6 – Classification hiérarchique établie selon le critère de l'information mutuelle. Représentation des taux d'inertie afférant aux noeuds de la classification

7. CARACTERISATIONS DES CLASSIFICATIONS A L'AIDE DES CONTRIBUTIONS

On avait déjà proposé et programmé certains calculs de contributions aptes à caractériser les classifications arborescentes établies selon le critère de la maximisation du moment entré d'ordre 2 (cas du χ^2 (cf. [5]). On propose ici des calculs analogues pour les classifications hiérarchiques établies selon le critère de l'information mutuelle, calculs qui permettront de comparer totalement les deux types de classifications.

7.1. Contributions des éléments j de J aux éléments i de I

Les formules proposées proviennent de la décomposition de la variance du Nuage $N_J(I)$ (cas de la statistique du χ^2) et de la décomposition de l'entropie (cas de la théorie de l'information).

7.1.1. Cas de la statistique du χ^2

a) Formules

On a $$M^2(N_J(I)) = \sum_{\substack{i\in I \\ i\in J}} f_i \, \| f_J^i - f_J \|^2_{fJ} = \sum_{\substack{j\in J \\ i\in I}} f_i (f_j^i - f_j)^2/f_j$$

avec $$\rho^2(i) = \| f_j^i - f_J \|^2_{fJ} \ .$$

On s'intéresse ici à la contribution des éléments j de J à ρ^2 (i) c'est-à-dire à la distance d'un élément i de I au centre de gravité du Nuage N_J (I) (dont le centre est f_J). En utilisant la fonction ψ définie précédemment ($\psi(x) = x^2 - 1$), on aura :

$$\rho^2(i) = \sum_{j\in J} f_j \, \psi \, \left(\frac{f_{ij}}{f_j f_j}\right) = \sum_{j\in J} C\,(i\,,j)$$

$$\text{si} \quad f_{ij} \geqslant f_i f_j \qquad \text{alors} \quad \frac{f_{ij}}{f_i} \geqslant f_j$$

donc j intervient fortement dans ρ^2(i) mais avec $(f_{ij} - f_i f_j) > 0$

$$\text{si} \quad f_{ij} \leqslant f_i f_j \frac{f_{ij}}{f_i} \ll f_j$$

donc j intervient fortement dans ρ^2 (i) avec $(f_{ij} - f_i f_j) < 0$

Pour la lecture des contributions, on affectera C(i , j) du signe de $(f_{ij} - f_i f_j)$ [On les appelle les contributions absolues signées]. Pour plus de facilité on propose parfois de calculer les contributions signées relativement à ρ^2 (i).

b) Illustration sur l'exemple

	PRI(I)	HAUS	FAUS	FNAU	HMUS	FMUS	HCUS	FCUS	HAWE	FAWE	FNAW
PROF	12568.	0.9	0.0	-6.4	0.9	-2.4	0.6	0.0	1.4	0.1	-6.1
TRAN	2410.	0.5	0.0	-1.3	0.5	-0.6	0.1	0.0	0.0	-0.0	-1.1
MENA	7755.	-2.5	-0.0	2.6	-2.4	1.1	-2.8	-0.4	-1.8	0.0	4.5
ENFA	933.	-0.2	-0.0	2.6	-0.2	1.3	-0.5	-0.1	-0.3	-0.0	1.3
COUR	3043.	0.0	0.1	0.5	0.0	0.4	0.2	0.1	-0.4	-0.1	0.0
TOIL	2656.	0.0	0.1	0.0	-0.0	0.0	0.0	0.2	-0.0	0.0	-0.0
REPA	3306.	-0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.1	0.1	0.1	0.5
SOMM	21997.	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	0.1
TELE	2784.	0.9	0.0	0.5	1.0	0.3	0.4	0.2	0.0	-0.0	0.1
LOIS	9756.	-0.0	-0.1	0.3	-0.1	0.0	0.1	-0.0	-0.0	-0.3	0.0
	RHO(J)	5.1	0.5	14.2	5.2	6.1	4.8	1.1	4.1	0.8	13.7

	PRI(I)	HMWE	FMWE	HCWE	FCWE	HAYO	FAYO	HMYO	FNAY	FMYO	HCYO
PROF	12568.	1.4	-2.8	1.2	-0.1	1.3	0.4	1.3	-6.4	-1.2	0.9
TRAN	2410.	0.0	-0.8	0.1	-0.5	0.5	0.1	0.6	-1.0	-0.2	0.3
MENA	7755.	-1.7	3.1	-2.3	-0.0	-1.3	0.5	-1.5	10.1	4.8	-1.8
ENFA	933.	-0.2	0.5	-0.5	-0.2	-0.1	0.1	-0.1	0.2	0.3	-0.5
COUR	3043.	-0.4	-0.0	-0.3	-0.0	-0.1	-0.0	-0.1	0.2	0.0	0.0
TOIL	2656.	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
REPA	3306.	0.1	0.3	0.1	0.1	-0.0	-0.1	-0.0	0.0	-0.0	-0.1
SOMM	21997.	0.0	0.0	0.0	0.1	-0.0	-0.0	-0.0	0.0	-0.0	-0.0
TELE	2784.	0.1	0.0	0.0	-0.0	-0.1	-0.2	-0.1	-0.2	-0.2	-0.5
LOIS	9756.	-0.0	0.0	0.1	0.1	0.0	-0.5	0.0	0.0	-0.1	0.7
	RHO(J)	4.1	7.6	4.5	1.1	3.6	2.0	3.7	18.2	6.8	4.8

	PRI(I)	FCYO	HAES	FAES	FNAE	HMES	HCES	FMES	FCES
PROF	12568.	-0.0	1.3	0.6	-6.0	1.4	1.1	-0.0	-0.0
TRAN	2410.	0.0	0.5	0.1	-1.1	0.4	0.7	-0.0	0.0
MENA	7755.	0.1	-1.3	0.2	5.4	-1.1	-2.3	1.3	0.0
ENFA	933.	-0.0	-0.1	0.0	0.7	-0.1	-0.5	0.3	-0.1
COUR	3043.	0.0	-0.1	-0.0	0.3	-0.2	-0.1	0.0	0.1
TOIL	2656.	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0
REPA	3306.	-0.0	-0.0	-0.1	0.0	-0.0	-0.1	-0.0	-0.1
SOMM	21997.	-0.0	-0.0	-0.0	0.1	-0.0	-0.0	-0.0	0.0
TELE	2784.	-0.4	-0.0	-0.2	-0.0	0.1	-0.3	-0.1	-0.3
LOIS	9756.	0.2	-0.0	-0.6	0.1	-0.1	0.6	-0.6	0.0
	RHO(J)	0.8	3.4	1.8	13.6	3.3	5.6	2.4	0.5

INERTIE TOTALE DU NUAGE 0.143

Figure 7 – Tableau des contributions (X 1000) des paires (i , j) à l'inertie totale du nuage. La colonne PRI(I) contient les poids des éléments de l'ensemble I. La ligne RHO(J) contient les carrés des distances au sens du χ^2 des éléments de J au centre de gravité du nuage.

7.1.2. Cas de l'information mutuelle

a) Formules

on a

$$E(N_J(I)) \overset{\text{déf}}{=} H(f_{IJ} ; f_I f_J)$$

$$= \sum_{\substack{i \in I \\ j \in J}} f_i f_j \ \phi\left(\frac{f_{ij}}{f_i f_j}\right) \qquad \text{avec } \phi(x) = x \log_2 x \quad \text{(cf. § 1.5)}$$

$$= \sum_{i \in I} f_i H(f_J^i ; f_J)$$

On pose donc $\inf(i) = H(f_J^i ; f_J)$ = entropie relative de la loi f_J^i par rapport à f_J, centre de gravité du Nuage $N_J(I)$; $H(f_J^i ; f_J)$ est l'analogue de $\| f_J^i - f_J \|_{f_J}^2$ dans les formules provenant de la statistique du χ^2

On a

$$H(f_J^i ; f_J) \overset{\text{def}}{=} \sum_{j \in J} f_j \, \phi\left(\frac{f_j^i}{f_j}\right) \quad \text{cf}(1.5)$$

$$= \sum_{j \in J} f_j \, \phi\left(\frac{f_{ij}}{f_i f_j}\right)$$

Le calcul des contributions est analogue à celui provenant du critère du χ^2 avec la différence que la signature de la contribution apparaît directement

$$\text{si} \quad f_{ij} \gg f_i f_j \quad \phi\left(\frac{f_{ij}}{f_i f_j}\right) \text{ est très fortement positive}$$

$$\frac{f_{ij}}{f_i f_j} \gg 1 \quad (\Leftrightarrow (f_{ij} - f_i f_j) \gg 0)$$

$$\text{si } f_{ij} \ll f_i f_j \quad \phi\left(\frac{f_{ij}}{f_i f_j}\right) \text{ est très fortement négative}$$

l'interprétation est analogue à l'étude du cas précédent, On parlera de contribution $(c(i, j))$ à l'entropie relative.

$$c(i, j) = f_j \, \phi\left(\frac{f_{ij}}{f_i f_j}\right)$$

b) Illustration sur l'exemple

	PRI(I)	HAUS	FAUS	FNAU	HMUS	FMUS	HCUS	FCUS	HAWE	FAWE	FNAW
PROF	12568.	4.0	0.6	-0.8	4.2	-3.5	3.3	0.7	5.2	1.4	-1.3
TRAN	2410.	1.5	0.1	0.0	1.5	-0.7	0.7	0.2	0.3	-0.3	-0.4
MENA	7755.	-2.0	-0.5	6.2	-2.0	3.8	-1.8	-1.5	-2.2	0.7	8.7
ENFA	933.	-0.3	-0.1	2.8	-0.3	1.8	0.0	-0.2	-0.2	-0.1	1.8
COUR	3043.	0.3	0.8	1.6	0.1	1.4	1.0	0.8	-0.8	-0.5	0.1
TOIL	2656.	0.0	0.6	0.4	-0.1	0.4	0.2	0.9	-0.2	0.0	-0.1
REPA	3306.	-0.1	-0.4	0.3	-0.1	0.0	-0.4	-0.4	0.8	0.6	1.6
SOMM	21997.	-0.5	-0.2	-0.0	-0.4	-0.2	-0.5	-0.2	0.5	0.7	1.3
TELE	2784.	2.1	0.4	1.6	2.3	1.1	1.3	0.8	0.4	-0.2	0.6
LOIS	9756.	-0.7	-0.9	1.9	-0.9	0.5	0.8	-0.3	-0.4	-1.6	0.4
	INF(J)	4.4	0.3	14.0	4.3	4.6	4.7	0.8	3.4	0.6	12.7

	PRI(I)	HMWE	FMWE	HCWE	FCWE	HAYO	FAYO	HMYO	FNAY	FMYO	HCYO
PROF	12568.	5.3	-3.7	5.0	-1.1	5.2	2.7	5.2	-0.8	-3.0	4.2
TRAN	2410.	0.3	-0.7	0.5	-0.7	1.5	0.4	1.6	-0.5	-0.6	1.0
MENA	7755.	-2.2	7.0	-2.1	-0.2	-2.2	2.4	-2.2	14.3	9.1	-2.2
ENFA	933.	-0.3	1.0	0.0	-0.3	-0.3	0.3	-0.3	0.6	0.7	0.0
COUR	3043.	-0.8	-0.2	-0.7	-0.3	-0.4	-0.4	-0.4	0.9	0.2	0.1
TOIL	2656.	-0.2	-0.3	-0.3	0.1	-0.1	-0.1	-0.1	-0.2	-0.2	-0.1
REPA	3306.	0.8	1.3	0.5	0.7	-0.3	-0.4	-0.3	0.3	-0.0	-0.6
SOMM	21997.	0.5	0.3	0.6	1.7	-0.5	-0.8	-0.5	0.6	-0.2	-0.5
TELE	2784.	0.5	0.4	0.0	-0.3	-0.5	-0.7	-0.4	-0.7	-0.6	-0.8
LOIS	9756.	-0.6	0.7	0.9	1.1	0.4	-2.0	0.2	0.7	-1.1	3.2
	INF(J)	3.4	5.8	4.2	0.9	2.7	1.4	2.8	15.3	4.2	4.3

	PRI(I)	FCYO	HAES	FAES	FNAE	HMES	HCES	FMES	FCES
PROF	12568.	-0.3	5.2	3.1	-1.5	5.2	4.5	-0.3	-0.3
TRAN	2410.	0.1	1.5	0.5	-0.4	1.2	1.7	-0.2	0.0
MENA	7755.	0.9	-2.1	1.4	9.7	-2.1	-2.0	4.2	0.4
ENFA	933.	-0.2	-0.2	0.2	1.2	-0.2	0.0	0.8	-0.2
COUR	3043.	0.1	-0.6	-0.1	1.3	-0.7	-0.4	0.2	0.5
TOIL	2656.	0.0	-0.0	-0.0	-0.1	-0.0	-0.1	-0.1	0.2
REPA	3306.	-0.3	-0.4	-0.5	0.2	-0.3	-0.6	-0.2	-0.5
SOMM	21997.	-0.3	-0.5	-0.7	1.2	-0.5	-0.3	-0.2	0.3
TELE	2784.	-0.8	-0.1	-0.6	-0.3	0.5	-0.7	-0.5	-0.7
LOIS	9756.	1.4	-0.3	-2.1	1.1	-0.8	2.8	-2.0	0.7
	INF(J)	0.6	2.6	1.3	12.5	2.4	5.0	1.6	0.4

NEGUENTROPIE TOTALE DU NUAGE 0.122

Figure 8 – Tableau des contributions (X 1000) des paires (i , j) de I x J à la néguentropie totale du nuage $N_J(I)$; la colonne PRI(I) contient les poids des éléments de I ; la ligne INF(J) contient les entropies relatives des lois f_j^i par rapport aux lois f_I (I = {activités} ; J = {sujets})

7.2. Contributions des éléments j de J aux lasses q de I de la classification hiérarchique A(I) établie sur I

Les formules proposées au § 7.1. s'étendent facilement aux classes des classifications hiérarchiques. Elles proviennent de la décomposition de la variance du Nuage ($N_J(Q)$) où Q est une partition de I, et de la décomposition de l'entropie relative dans $N(_J(Q))$ où Q est muni de la loi f_Q, J de la loi f_J et Q_J de la loi f_{QJ}

7.2.1. Cas de la statistique du χ^2

a) Formules

On a $M_1(N_J(Q)) \overset{\text{def}}{=} \sum_{\substack{j \in J \\ q \in Q}} (f_{qj} - f_q f_j)^2 / f_q f_j = \text{Lien}_2(Q , J)$ Q partition de I

$$= \sum_{q \in Q} f_q \, \rho^2(q) = \sum_{q \in Q} f_q \, \| f_J^q - f_J \|_{fJ}^2$$

en écrivant ces formules avec la fonction ψ on aura :

$$\rho^2(q) = \sum_{j \in j} f_j \, \psi \left(\frac{f_{qj}}{f_q f_j} \right) = \sum_{j \in J} C(q , j)$$

On calculera naturellement les contributions (relatives ou absolues) signées (c'est-à-dire affectées du signe de $(f_{qj} - f_q f_j)$.

L'interprétation est analogue à celle indiquée au § 7.1

Si $f_{qj} \geqslant f_q \cdot f_j$ $\quad (f_{qj} - f_q f_j) > 0$

C(q , j) est élevé avec $f_{qj} - f_q f_i > 0$ j'intervient fortement dans la distance de la classe q au centre de gravité du nuage

si $f_{qj} \leqslant f_q f_i$ $\quad (f_{qj} - f_q f_j) < 0$

C(q , j) est observée $(f_{qj} - f_q f_j) < 0$ j intervient faiblement dans la distance de la classe q au centre de gravité du nuage

b) Illustration sur l'exemple – listage des calculs

N	RHO**2	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS
29	0.14	0.02	0.01	-0.07	-0.01	0.00	-0.00	-0.00	-0.00	0.03	-0.00
30	0.11	0.04	0.00	-0.05	-0.01	-0.01	-0.00	0.00	0.00	0.00	-0.00
31	0.10	0.04	0.02	-0.04	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	0.00
32	0.05	0.01	0.00	0.01	0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.02
33	0.02	-0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.01	0.00
34	0.09	0.04	0.01	-0.03	-0.00	-0.01	-0.00	-0.00	-0.00	0.00	-0.00
35	0.14	0.03	0.01	-0.06	-0.01	-0.00	-0.00	-0.00	-0.00	-0.01	0.02
36	0.02	0.00	0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00
37	0.12	0.04	0.00	-0.05	-0.01	-0.01	-0.00	0.00	0.00	0.00	-0.00
38	0.10	0.04	0.01	-0.04	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
39	0.14	0.02	0.01	-0.07	-0.01	0.00	0.00	-0.00	-0.00	0.02	-0.00
40	0.44	-0.17	-0.03	0.21	0.01	0.01	-0.00	0.00	0.00	-0.00	0.00
41	0.28	-0.12	-0.02	0.10	0.02	-0.00	-0.00	0.01	0.00	0.00	0.00
42	0.03	0.01	0.00	0.01	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01
43	0.01	-0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.01	0.00
44	0.27	-0.12	-0.02	0.05	0.05	0.01	0.00	0.00	-0.00	0.01	0.00
45	0.04	0.00	0.00	0.01	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01
46	0.10	0.04	0.01	-0.04	-0.01	-0.01	-0.00	0.00	-0.00	0.00	-0.00
47	0.27	-0.12	-0.02	0.07	0.04	0.00	0.00	0.00	0.00	0.01	0.00
48	0.01	-0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	0.00
49	0.34	-0.11	-0.02	0.18	0.01	0.00	-0.00	0.00	0.00	-0.00	0.00
50	0.10	0.04	0.01	-0.05	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00	0.00
51	0.28	-0.12	-0.02	0.11	0.02	0.00	-0.00	0.00	0.00	0.00	0.00
52	0.10	0.03	0.01	-0.05	-0.01	-0.00	-0.00	-0.00	-0.00	0.00	0.00
53	0.01	0.00	-0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00
54	0.03	0.01	0.00	-0.01	-0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.00

Figure 9 – Tableau des contributions absolues signées (c(N , j) * Signe $(f_{NJ} - f_N f_j)$) ; la colonne RHO**2 contient les $\rho^2(N)$ pour tous les noeuds N de la classification (J = {activités} ; N = {noeuds de la hiérarchie})

N	RHO**2	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS
29	0.14	0.17	0.10	-0.48	-0.05	0.00	-0.00	-0.00	-0.00	0.18	-0.01
30	0.11	0.34	0.01	-0.43	-0.07	-0.10	-0.00	0.03	0.00	0.01	-0.01
31	0.10	0.37	0.15	-0.38	-0.04	-0.02	-0.00	-0.01	-0.00	-0.02	0.00
32	0.05	0.26	0.04	0.18	0.03	-0.01	-0.00	-0.04	-0.01	-0.11	-0.32
33	0.02	-0.01	0.00	0.08	-0.09	0.03	0.00	-0.08	0.00	-0.55	0.15
34	0.09	0.41	0.14	-0.36	-0.02	-0.06	-0.00	-0.01	-0.00	0.00	-0.01
35	0.14	0.19	0.09	-0.40	-0.10	-0.00	-0.00	-0.03	-0.00	-0.07	0.12
36	0.02	0.04	0.01	-0.23	-0.06	0.20	0.21	-0.07	-0.00	0.13	-0.05
37	0.12	0.32	0.01	-0.46	-0.08	-0.09	-0.01	0.03	0.00	0.01	-0.00
38	0.10	0.40	0.15	-0.38	-0.03	-0.04	-0.00	-0.01	-0.00	-0.00	-0.00
39	0.14	0.16	0.08	-0.53	-0.07	0.01	0.00	-0.00	-0.00	0.15	-0.00
40	0.44	-0.39	-0.07	0.48	0.03	0.02	-0.00	0.00	0.00	-0.01	0.00
41	0.28	-0.41	-0.09	0.36	0.08	-0.00	-0.00	0.04	0.00	0.01	0.00
42	0.03	0.27	0.01	0.17	0.01	-0.03	-0.00	-0.01	-0.00	-0.10	-0.39
43	0.01	-0.05	-0.09	0.03	-0.18	0.00	0.01	-0.00	0.03	-0.41	0.20
44	0.27	-0.43	-0.09	0.18	0.20	0.05	0.00	0.00	-0.00	0.04	0.01
45	0.04	0.14	0.00	0.31	0.04	-0.01	-0.00	-0.01	-0.00	-0.09	-0.40
46	0.10	0.38	0.07	-0.44	-0.05	-0.06	-0.00	0.00	-0.00	0.00	-0.00
47	0.27	-0.44	-0.09	0.28	0.14	0.01	0.00	0.01	0.00	0.02	0.01
48	0.01	-0.00	-0.05	-0.04	-0.34	0.15	0.17	-0.06	0.01	-0.12	0.07
49	0.34	-0.34	-0.06	0.55	0.03	0.01	-0.00	0.00	0.00	-0.01	0.00
50	0.10	0.35	0.08	-0.46	-0.07	-0.04	-0.00	-0.00	-0.00	-0.00	0.01
51	0.28	-0.41	-0.08	0.41	0.08	0.01	-0.00	0.01	0.00	0.00	0.00
52	0.10	0.31	0.08	-0.51	-0.07	-0.01	-0.00	-0.00	-0.00	0.00	0.00
53	0.01	0.13	-0.00	0.24	-0.01	0.01	0.03	-0.05	0.00	-0.25	-0.29
54	0.03	0.41	0.08	-0.41	-0.08	-0.01	0.00	-0.01	-0.00	-0.00	-0.00

Figure 10 – Tableau des contributions relatives signées $(c(N, j)/\rho^2(N) *$ Signe $(f_{Nj} - f_N f_{j\cdot})$; la colonne RHO**2 contient les $\rho^2(N)$ pour tous les noeuds N de la classification hiérarchique

c) Interprétation des résultats

Les calculs issus de ces tableaux permettent de déterminer quelles variables j de J contribuent le plus à la formation des classes constituées en hiérarchie. Une première lecture du tableau des contributions relatives signées nous confirme l'importance des variables PROF (temps passé au travail professionnel) et MENA (travail ménager) comme activités discriminantes principales. (Il eut été surprenant compte tenu du choix des variables de base, qu'il en fût autrement !) Une seconde lecture nous fait nous intéresser aux contributions d'importance moyenne (de l'ordre du quart des contributions les plus fortes) et qui apporte une explication à la constitution des classes, surtout aux niveaux inférieurs. Ce sont précisément ces variables qui font la différence à ces niveaux. On peut citer les exemples suivants :

classe 29	HMUS – HAUS	C +	Télévision, transports
classe 30	HMWE – HAWE	C –	Temps passé à faire les courses
classe 31	HMYO – HAYO	C +	Transports
classe 32	FAES – FAYO	C –	Télévision, autres loisirs
classe 33	FCYO – FCES	C –	Télévision
		C +	Autres loisirs
classe 35	HCYO – HCES	C +	Autres loisirs
classe 36	FCUS – FAUS	C +	Les courses, la toilette, la télévision
classe 38	HMYO – HAYO		
	HAES – HMES	C +	Transports

classe 39	HCUS – HMUS – HAUS	C +	Télévision
classe 42	FAES – FAYO – FAWE	C –	Autres loisirs
classe 43	FCES – FCYO – FCWE	C –	Occupations liées aux enfants, télévision
		C +	Autres loisirs
classe 44	FMUS – FNAU	C +	Occupation liées aux enfants
classe 45	FAES – FAYO – FAWE – FMES	C –	Autres loisirs
classe 48	FCUS – FAUS – FCWE	C –	Occupations liées aux enfants, télévision
	FCYO – FCES	C +	Courses, les soins personnels
classe 53	Femmes célibataires Femmes actives FMES	C –	Autres loisirs, télévison

Sur ce résumé on n'a pas reporté les contributions des variables PROF et MENA. Dans le but de mieux illustrer les résultats, on propose de reporter ceux-ci sur la représentation de la hiérarchie en procédant de la façon suivante. A chaque noeud de la hiérarchie on associe leurs variables discriminantes positivement ou négativement en mettant + ou – ; quand les variables contribuent fortement on met V++ ou V– – (+ : taux fort, – : taux faible).

Cette représentation a l'avantage de faire porter sur un même graphique les éléments de I de l'ensemble J. Ceci permet de pallier le désavantage en classification automatique d'une absence de représentation simultanée simple. (Cf. figure 11).

7.2.2. Cas de l'information mutuelle

a) Formules – Soit Q une partition de I

on a $$E(N_J(Q)) = H(f_{QJ} \; ; f_Q f_J) = \text{Lien}_1\,(Q\,,J)$$

$$= \sum_{\substack{q \in Q \\ j \in J}} f_q f_j \; \phi\left(\frac{f_{qj}}{f_q f_j}\right) \text{ avec } \phi(x) = x \log_2 x$$

$$= \sum_{q \in Q} f_q \inf(q) = \sum_{q \in Q} f_q \; H(f_J^q \; ; \; f_J)$$

entropie relative de a loi f_J^q par rapport à f_J

Avec $$= H(f_J^q \; ; f_J) = \sum_{j \in J} f_j \; \phi\left(\frac{f_{qj}}{f_q f_j}\right)$$

avec $$C(q\,,i) = f_j \; \phi\left(\frac{f_{qj}}{f_q f_j}\right)$$

5

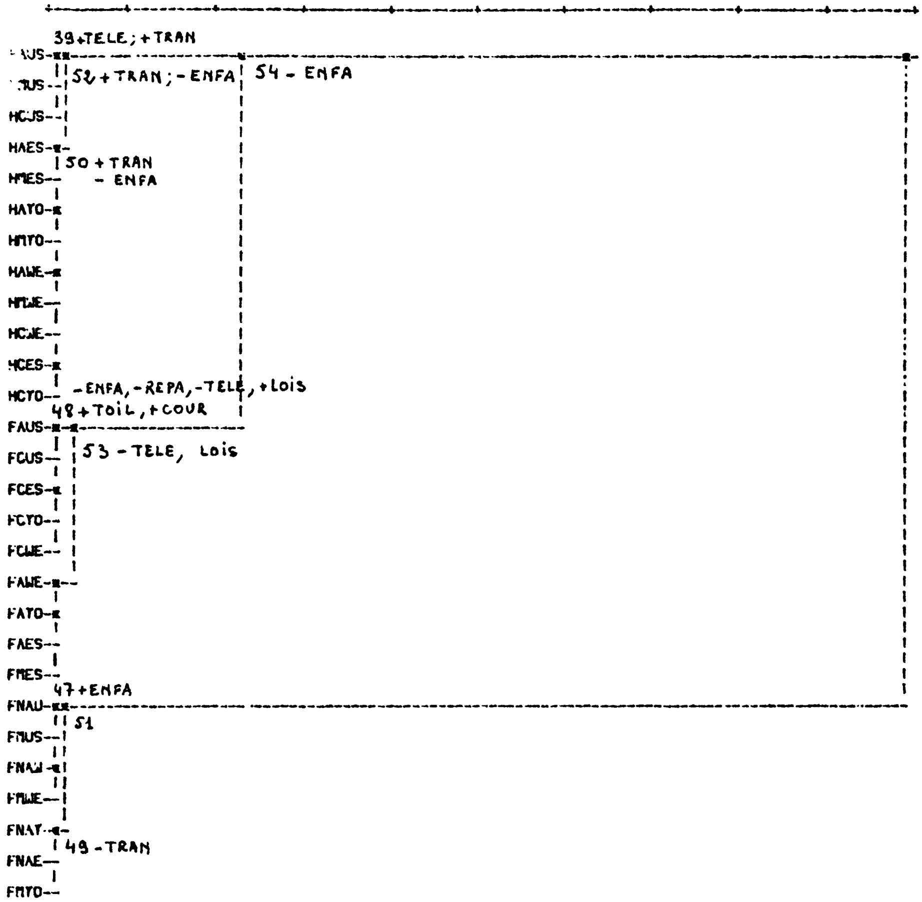

Figure 11 – Classification hiérarchique établie selon le critère du χ^2 – Représentation des éléments (ici les activités) de J les plus "significatives". N_j+ contribution positive de l'activité j de J à la classe N ; N_j – contribution négative de l'activité j de J à la classe N ; N numéro du noeud dans la hiérarchie

Le calcul des contributions, pour chaque classe q de la hiérarchie, permet donc de déterminer les éléments de j qui contribuent le plus à la formation d'une classe. On ne calcule pas de contributions signées, elles le sont par l'étude de la quantité $\left(\frac{f_{qj}}{f_q f_j}\right)$ qui intervient dans la fonction ϕ ; ϕ (x) = x $\log_2$ x

b) Illustration sur l'exemple – listage des calculs

N	INF(N)	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS
29	0.12	0.11	0.04	-0.06	-0.01	0.01	-0.00	-0.00	-0.01	0.06	-0.02
30	0.08	0.14	0.04	-0.06	-0.01	-0.01	-0.00	-0.01	-0.01	-0.01	0.01
31	0.10	0.15	0.01	-0.06	-0.01	-0.02	-0.01	0.02	0.01	0.01	-0.01
32	0.04	0.08	0.01	0.05	0.01	-0.01	-0.00	-0.01	-0.02	-0.02	-0.06
33	0.07	0.15	0.04	-0.06	-0.01	-0.02	-0.00	-0.01	-0.01	0.01	-0.02
34	0.01	-0.01	0.00	0.02	-0.01	0.01	0.00	-0.01	0.00	-0.02	0.03
35	0.01	0.02	0.00	-0.03	-0.00	0.02	0.02	-0.01	-0.01	0.02	-0.02
36	0.13	0.12	0.04	-0.06	0.0	-0.00	-0.00	-0.02	-0.01	-0.02	0.08
37	0.07	0.15	0.04	-0.06	-0.01	-0.02	-0.00	-0.01	-0.01	-0.00	-0.00
38	0.39	-0.03	-0.01	0.34	0.02	0.03	-0.00	0.01	0.03	-0.01	0.03
39	0.10	0.15	0.01	-0.06	-0.01	-0.02	-0.01	0.02	0.01	0.01	-0.00
40	0.12	0.11	0.03	-0.05	-0.01	0.01	0.00	-0.00	-0.01	0.05	-0.01
41	0.14	-0.10	-0.02	0.15	0.04	0.01	0.00	0.02	0.00	0.02	0.02
42	0.04	0.05	0.01	0.07	0.01	-0.00	-0.00	-0.01	-0.02	-0.02	-0.06
43	0.37	-0.03	-0.01	0.30	0.03	0.02	-0.00	0.02	0.03	-0.01	0.02
44	0.03	0.05	0.00	0.06	0.01	-0.01	-0.00	-0.01	-0.01	-0.01	-0.05
45	0.01	-0.02	-0.01	0.01	-0.01	0.00	0.00	-0.00	0.02	-0.02	0.03
46	0.37	-0.03	-0.01	0.27	0.04	0.03	-0.00	0.02	0.02	0.01	0.03
47	0.12	-0.10	-0.02	0.18	0.03	0.01	-0.00	0.01	-0.00	0.01	0.00
48	0.08	0.15	0.03	-0.06	-0.01	-0.02	-0.00	0.00	-0.00	0.00	-0.00
49	0.00	-0.00	-0.00	-0.01	-0.01	0.01	0.01	-0.01	0.01	-0.01	0.01
50	0.08	0.14	0.03	-0.06	-0.01	-0.02	-0.00	-0.00	-0.00	-0.01	0.01
51	0.09	0.13	0.03	-0.06	-0.01	-0.01	-0.00	-0.00	-0.01	0.01	0.01
52	0.00	0.02	-0.00	0.02	-0.00	0.00	0.00	-0.01	0.00	-0.01	-0.02
53	0.23	-0.09	-0.02	0.23	0.04	0.02	-0.00	0.01	0.01	0.01	0.02
54	0.02	0.08	0.02	-0.05	-0.01	-0.01	0.00	-0.00	-0.00	-0.00	-0.01

Figure 12 – Tableau des contributions des activités j de J à l'entropie relative de la loi f_J^N à la loi f_J INF(N) = $H(f_J^N \, ; f_J)$; $C(N, j) = f_j \, \phi \, (f_{Nj}/f_N \cdot f_j)$

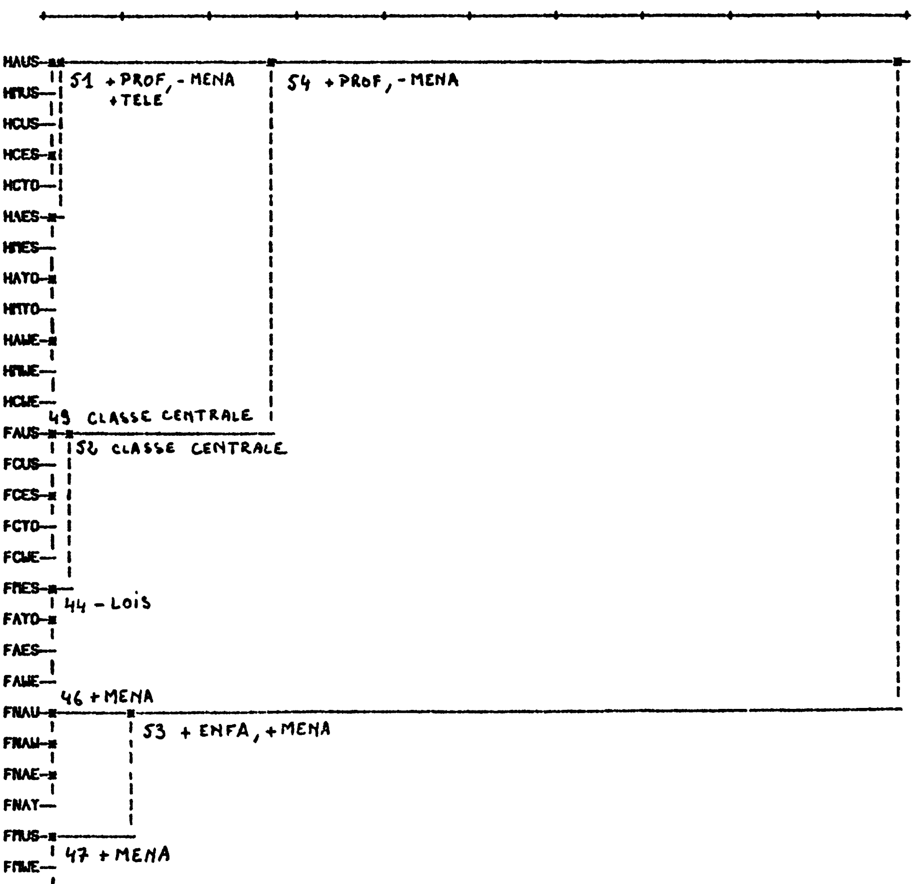

Figure 13 – Classification hiérarchique établie selon le critère de l'information mutuelle – Représentation des activités les plus "significatives" ; N numéro du noeud dans la classification hiérarchique ; N_j+ contribution positive de l'activité j de J ; N_j – contribution négative de l'activité j de J

N	INF(N)	PROF	TRAN	MENA	ENFA	COUR	TOIL	REPA	SOMM	TELE	LOIS
29	0.32	0.35	0.13	-0.17	-0.02	0.02	-0.00	-0.01	-0.04	0.19	-0.07
30	0.31	0.46	0.14	-0.19	-0.02	-0.04	-0.01	-0.02	-0.05	-0.04	0.02
31	0.31	0.47	0.03	-0.19	-0.02	-0.07	-0.02	0.07	0.04	0.04	-0.04
32	0.27	0.30	0.05	0.20	0.03	-0.02	-0.01	-0.05	-0.08	-0.06	-0.21
33	0.31	0.47	0.12	-0.19	-0.02	-0.06	-0.00	-0.03	-0.04	0.02	-0.05
34	0.11	-0.09	0.01	0.18	-0.05	0.07	0.02	-0.10	0.01	-0.19	0.28
35	0.15	0.12	0.02	-0.20	-0.03	0.15	0.14	-0.07	-0.04	0.11	-0.11
36	0.36	0.34	0.11	-0.17	0.0	-0.01	-0.01	-0.05	-0.03	-0.06	0.23
37	0.30	0.48	0.14	-0.20	-0.02	-0.05	-0.01	-0.03	-0.05	-0.01	-0.01
38	0.51	-0.07	-0.02	0.66	0.05	0.06	-0.01	0.01	0.05	-0.03	0.05
39	0.29	0.49	0.03	-0.21	-0.02	-0.08	-0.02	0.07	0.05	0.03	-0.00
40	0.30	0.36	0.11	-0.18	-0.02	0.04	0.00	-0.02	-0.05	0.18	-0.03
41	0.37	-0.27	-0.05	0.39	0.10	0.04	0.00	0.05	0.00	0.05	0.05
42	0.25	0.20	0.03	0.30	0.05	-0.01	-0.01	-0.05	-0.07	-0.07	-0.23
43	0.48	-0.07	-0.02	0.63	0.07	0.04	-0.01	0.04	0.06	-0.01	0.04
44	0.21	0.23	0.01	0.29	0.04	-0.03	-0.01	-0.03	-0.04	-0.07	-0.26
45	0.11	-0.15	-0.08	0.09	-0.06	0.01	0.02	-0.02	0.14	-0.16	0.27
46	0.45	-0.07	-0.02	0.60	0.10	0.06	-0.00	0.03	0.05	0.01	0.06
47	0.36	-0.27	-0.05	0.50	0.09	0.03	-0.00	0.03	-0.00	0.01	0.00
48	0.27	0.54	0.10	-0.23	-0.03	-0.07	-0.01	0.01	-0.01	0.00	-0.01
49	0.07	-0.04	-0.06	-0.10	-0.08	0.13	0.14	-0.08	0.10	-0.10	0.16
50	0.28	0.49	0.10	-0.22	-0.03	-0.06	-0.01	-0.01	-0.02	-0.02	0.05
51	0.27	0.49	0.11	-0.23	-0.03	-0.04	-0.01	-0.01	-0.03	0.02	0.03
52	0.09	0.21	-0.02	0.24	-0.01	0.02	0.05	-0.06	0.00	-0.12	-0.26
53	0.44	-0.20	-0.04	0.52	0.09	0.04	-0.00	0.03	0.03	0.01	0.04
54	0.17	0.48	0.09	-0.27	-0.04	-0.03	0.00	-0.02	-0.02	-0.01	-0.03

Figure 14 – Tableau des contributions signées relatives à la somme des valeurs absolues

$$C(N\,,j) = \text{Signe}\,(f_{NJ} - f_N\,.\,f_j) * \left| f_j\, \phi \left(\frac{f_{Nj}}{f_N\,.\,f_j}\right) \right|$$

$$\text{INF}(N) = \sum_{j \in J} f_j\, \phi \left(\frac{f_{Nj}}{f_N\,.\,f_j}\right)$$

c) Interprétation des résultats (figure 13)

Comme au § 7.2.1. c) on a projeté sur la classification hiérarchique établie selon le critère de l'information les cativités dont la contribution à l'entropie relative de la loi f_J^N/f_J est forte. On laisse le soin au lecteur de reprendre l'interprétation da la classification hiérarchique à partir des tableaux proposés dans les figures 8, 12 et 14.

CONCLUSION

Comme toujours en traitement de données sous l'angle de la classification hiérarchique, la seule consultation de la représentation arborescente de la hiérarchie ne suffit pas à interpréter les données. On doit assortir ces représentations de calculs complémentaires tels que ceux qui sont présentés dans cette note. Les taxinomistes doivent non seulement construire des classes, mais expliquer ou déterminer quelles variables contribuent à la formation de ces classes. Les calculs présentés dans cette note permettent de répondre en partie à cette préoccupation.

8. BIBLIOGRAPHIE

[1] J.P. BENZECRI – L'analyse des Données, Tome I, La Taxinomie, Dunod 1973. L'Analyse des correspondances, Tome II.

[2] M. JAMBU – Techniques de classification ascendante hiérarchique appliquées à des données "Sciences Humaines". Thèse 3è cycle, Université Pierre et Marie Curie, Paris VI. 1972.

[3] M. JAMBU – Programme de construction de classification ascendante hiérarchique (CAH). In l'Analyse des Données de J.P. Benzecri et Collaborateurs, Dunod, 2è édition, 1976.

[4] M. JAMBU – Quelques calculs utiles à l'interprétation simultanée d'une analyse des correspondances et d'une classification arborescente. Illustration sur des données de budgets temps. *Consommation,* n° 2, 1976, Dunod.

[5] J.C. BALDWIN – The dépendence capacity of finite Borel fields in *Information and control 9*, 380-392 (1966)

[6] J.P. BENZECRI – Note sur le mémoire de BALDWIN, LSM-ISUP, Octobre 1967.