REVUE DE STATISTIQUE APPLIQUÉE

I. C. LERMAN

Introduction à une méthode de classification automatique illustrée par la recherche d'une typologie des personnages enfants à travers la littérature enfantine

Revue de statistique appliquée, tome 21, n° 3 (1973), p. 23-49

<http://www.numdam.org/item?id=RSA_1973__21_3_23_0>

INTRODUCTION A UNE MÉTHODE DE CLASSIFICATION (*) AUTOMATIQUE ILLUSTRÉE PAR LA RECHERCHE D'UNE TYPOLOGIE DES PERSONNAGES ENFANTS A TRAVERS LA LITTÉRATURE ENFANTINE

I. C. LERMAN (**)

1 – INTRODUCTION ET GENERALITES

On se propose dans ce texte d'exprimer le schéma de notre méthode de classification automatique qui a pour but d'extraire l'information pertinente des résultats d'une enquête par questionnaire en reconnaissant sur l'ensemble des questions une structure en classes et sous-classes ; cette classification est basée sur les ressemblances entre les comportements de la population étudiée vis à vis de chaque couple de variables descriptives défini par un couple de questions. Une telle classification a permis, dans les expériences réelles menées, de dégager clairement les tendances principales de l'attitude de la population étudiée et les diverses composantes de chacune de ces tendances. Ce succès est dû au développement d'une notion que nous avons mise en valeur et qui est celle de mesure de proximité entre structures de même type au sens algébrique du terme. Pour préciser cette notion on a utilisé une vieille idée bien établie de la statistique : la proximité entre deux structures α et β ne doit retenir que ce qui peut-être significatif par rapport à l'hypothèse N d'ignorance à priori de la position relative des deux structures mises en jeu. Nous allons, en nous appuyant sur un exemple réel, illustrer cette idée à tous les niveaux de la recherche d'une classification ou d'une hiérarchie de classifications.

L'exemple que nous allons traiter est du domaine de la Psycho-Sociologie et se situe dans le cadre d'une recherche menée par Mme Chombard de Lauwe où il s'agit de définir les modèles que les adultes créent et proposent aux enfants dans la littérature dite "pour enfants". La donnée est un ensemble de 1 500 personnages enfantins issus d'un échantillon d'ouvrages de cette littérature édités entre 1880 et 1960. Chaque personnage est décrit par un ensemble d'attributs qui établissent

a) **Son portrait**

Ses aptitudes (savoir, mémoire, très doué en général, dons particuliers dans certains domaines, . . .) ; les terrains où se manifestent les aptitudes (art, sport,

---

(*) Ce travail a été effectué au Centre de Mathématiques Appliquées et de Calcul de la Maison des Sciences de l'Homme et remis le 4/12/72.

(**) Actuellement maître de conférences à l'Université de Rennes 1.

commandement, vie pratique, secourisme, aventure, succès scolaires, farces, . . .) ; sa religion (chrétien, croit en Dieu, non mentionnée, . . .) sa psychologie (patience, colère, tenue parfaite, coquetterie, impulsif, paresse, gaité, franchise, . . .) ses relations avec autrui (soumission, égalité ou commandement envers l'adulte et envers un autre enfant, . . .) ; situation dans laquelle est impliqué le personnage (aventure, recherche, sauvetage, . . .) ; qualité du personnage par rapport à l'action (passif, actif) ; grands thèmes où il évolue (drames de l'enfance, aventure, leçon de morale, vocation de l'enfant, . . .) ; appréciation globale du sujet (personnage positif, d'abord négatif se transforme et devient positif, personnage foncièrement négatif)

b) **Son environnement physique**

Ville, château, véhicule-habitat (exemple : roulotte), véhicule-action (ex. automobile, avion), vit à l'étranger, nature sauvage, pistes-routes, . . .

c) **Son milieu social**

Ville, château, véhicule-habitat (exemple : roulotte), véhicule-action (ex. automobile, avion), vit à l'étranger, nature sauvage, pistes-routes, . . .

c) **Son milieu social**

Ouvrier, aristocratie, personnage "hors classe" (impossible à situer dans le contexte social décrit), . . .

d) **Son milieu familial**

Non décrit, absence de famille, famille incomplète ou famille de remplacement, milieu adulte (ex. armée, milieu professionnel), collectivité enfants, . . .

e) **Son entourage immédiat**

Père, mère, substitut père, substitut mère, policier (représentant l'"ordre"), traître-bandit (représentant le "méchant"), enfant concurent, ami enfant de même sexe ou de sexe opposé, animal, . . .

Quelques attributs tels que "positif" ou "passif dans l'acte" dont l'absence chez un sujet donné a été jugée, à priori, aussi significative que leur présence, ont été dédoublés donnant lieu à des attributs tels que "non positif" et "actif dans l'acte".

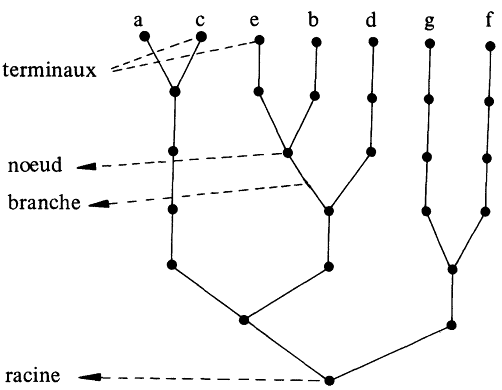
La description qui a nécessité 110 attributs a pour support un tableau T d'incidence des données de dimension 110 x 1 500 croisant l'ensemble A des attributs descriptifs et l'ensemble E des sujets enfants. A l'intersection de la ligne représentant l'attribut a et de la colonne représentant l'individu x on pose 1 ou 0 selon que a est présent chez x ou non. La donnée du tableau T est équivalente à celle d'une application de A dans l'ensemble $\mathcal{P}(E)$ des parties de E où à chaque attribut on associe le sous-ensemble des sujets qui le possèdent.

Par rapport à un même individu, deux attributs a et b sont dits avoir une association positive (resp. négative) si les deux attributs sont simultanément présents (resp. absents) chez l'individu.

Pour définir la typologie des personnages enfants, deux voies de travail se sont présentées au départ. La première consiste à définir une classification sur l'ensemble des sujets cherchant par la suite, pour une classe donnée, l'ensemble des attributs les plus fréquents dans cette classe, ce dernier définissant un type. La deuxième voie, plus directe, consiste à établir une classification sur l'ensemble des attributs, une classe formée exprimant un type. C'est cette seconde solution qui a été adoptée. D'ailleurs la notion de classification signficative sur l'ensemble des attributs est duale de celle sur l'ensemble des individus dans le sens suivant :

Toute classe A_i d'une classification "naturelle" sur A se réfère à une classe E_i d'une partition sur E telle que α) tout attribut de A_i est présent chez une proportion relattivement importante, mais non fixée, d'éléments de $E_i . \beta$) chaque individu de E_i posséde une proportion relativement importante, mais non fixée, d'attributs de A_i.

Plutôt que d'établir une seule partition, il est préférable de construire une hiérarchie de classifications qui nous rendra compte de la formation des classes, qui définira pour chacune des classes de la partition qu'on désire retenir, les sous-classes composantes (i. e. sous types du type que définit la classe) ; enfin qui permettra de faire apparaître les types les plus marqués dès les premiers niveaux de l'arbre des classifications et ceux correspondant à des classes de faible cohésion qui se forment plus tardivement.



NIVEAU	PARTITION
0	{{a},{c},{e},{b},{d},{g},{f}}
1	{{a,c},{e} {b},{d},{g},{f}}
2	{{a,c},{b,e},{d},{g} {f}}
3	{{a,c},{b,d,e},{g},{f}}
4	{{a,c},{b,d,e},{f,g}}
5	{{a,b,c,d,e},{f,g}}
6	{{a,b,c,d,e,f,g}}

Figure 1

Une hiérarchie de classifications se présente sous la forme d'une arborescence analogue à celle de la figure ci-dessus et qui sera interprétée comme la suite des partitions sur a, b, c, d, e, f, g écrite parallèlement ; dans cette chaîne la partition d'un niveau se déduit de celle du niveau précédent par réunion de deux classes ; il s'agit d'un arbre binaire.

Pour que les classes extraites de l'arbre des classifications répondent à la propriété exprimée intuitivement ci-dessus (conditions α) et β)) il est nécessaire d'établir la hiérarchie de partitions de façon à respecter au mieux les "ressemblances" entre attributs ; c'est-à-dire, de telle sorte que deux attributs se trouvent réunis à un niveau d'autant plus élevé que leur similarité est grande, cette dernière étant conçue à partir des vecteurs lignes du tableau T.

2 – INDICE DE PROXIMITE OU DE RESSEMBLANCE ENTRE ATTRIBUTS

L'information de départ d'une classification sur A nécessite la définition d'un indice de proximité entre attributs descriptifs, associant à chaque couple (a , b) de A x A, un nombre réel sensé mesurer la ressemblance entre a et b. Chaque attribut étant représenté par la partie des sujets qui le possèdent ; un couple (a , b) d'attributs définit la situation suivante

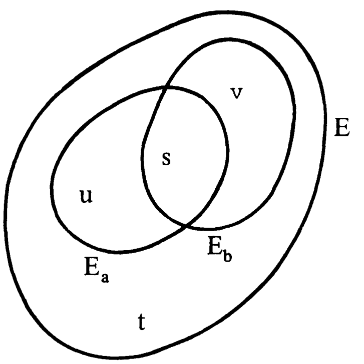


Figure 2

E_a (resp. E_b) est l'ensemble des sujets possédant a (resp. b), s = card ($E_a \cap E_b$) (resp. t = card ($E_a^c \cap E_b^c$)) est le nombre d'individus où les deux attributs sont simultanément présents (resp. absents) ; il s'agit du nombre d'associations positives (resp. négatives) entre a et b. u = card ($E_a \cap E_b^c$) (resp. v = card ($E_a^c \cap E_b$)) est le nombre de sujets possédant a et non b (resp. b et non a). On a $t = n - (s + u + v)$ où n est l'effectif de la population totale, soit n = card (E).

Les conditions minimales qu'on doit imposer à un indice de similarité est d'être conçu à partir des paramètres s, u et v de telle sorte, que sa valeur croît lorsque s augmente (u et v restant constants) qu'il soit symétrique en u et v et que sa valeur augmente lorsque celle de u diminue (s et v restant invariables). Cette façon de faire consiste à considérer comme *statistiquement équivalents* tous les éléments de la population étudiée. Il est clair que dans le choix de l'indice de proximité, la statistique s : nombre de personnages possédant simultanément les deux attributs, doit jouer un rôle important ; en effet, la présence commune, de deux attributs, chez un même individu, peut être significative de leur ressemblance. Mais la valeur brut de s est certainement un indicateur biaisé de la similarité ; il suffit en effet que les deux attributs a et b soient fréquents (resp. rares) pour trouver une valeur de s relativement grande (resp. petite) et ceci, indépendamment de la position relative des ensembles E_a et E_b, (cf. [1]). Conformément à ce qui a été annoncé au début de l'introduction nous allons adopter comme mesure de proximité entre les attibuts a et b un indice qui consiste à ne retenir dans la statistique s que ce qui peut être significatif par rapport à l'hypothèse N d'ignorance à priori de la position relative de E_a et E_b dans E, *compte tenu des cardinaux de* E_a *et de* E_b. A l'hypothèse N d'absence de liaison nous allons donner deux formes N_1 et N_2 ; pour N_1, on associe respectivement à E_a et à E_b deux parties aléatoires X et Y définies telles que chaque individu indépendamment ait la probabilité n_a/n (resp. n_b/n) d'appartenir à X (resp. à Y) ; pour N_2, on fixe E_b et on associe à E_a la partie X' : élément aléatoire de l'ensemble des parties de E de cardinal n_a, muni d'une mesure de probabilité uniformément répartie. Considérons alors les variables aléatoires (S_1 = card $(X \cap Y)$) et S_2 = card $(X' \cap E_b)$. Empressons nous de souligner que la distribution de S_2 est la même si au lieu de fixer E_b et d'associer à E_a la partie aléatoire X ; on fixait E_a et on associait de façon analogue une partie aléatoire Y. Les variables aléatoires S_1 et S_2 ont même moyenne

$$\mu = n_a \times n_b/n ,$$

la variance de S_1 est sensiblement égale à sa moyenne ; celle de S_2 se met sous la forme

$$\sigma_2^2 = n_a(n - n_a)\, n_b(n - n_b)/n^3 .$$

Pour répondre au souci exprimé, on peut prendre l'un des deux indices de similarité

$$\mathcal{S}_1(a, b) = (s - \mu)/\sqrt{\mu} \quad \text{ou } \mathcal{S}_2(a, b) = (s - \mu)/\sigma_2 ;$$

F. Nicolau (étudiant de 3[e] cycle) a vérifié que la première statistique donne plus de poids aux associations positives qu'aux absences communes, tandis que la deuxième est plus symétrique ; associations positives et associations négatives ont la même importance.

Au lieu de prendre $\mathcal{S}_1$ ou $\mathcal{S}_2$, il est plus précis de se reférer à une échelle de probabilité et de prendre

$$P_1(a, b) = \Pr^{N_1}\{S_1 < s\} \quad \text{ou bien} \quad P_2(a, b) = \Pr^{N_2}\{S_2 < s\} ;$$

le second membre désigne la probabilité dans l'hypothèse N_1, respectivement N_2, que la v. a. S_1, respectivement S_2, ait une valeur strictement inférieure à la valeur observée s = card $(E_a \cap E_b)$. Ici on introduit une notion de *vraisemblance* dans la notion de *ressemblance* : Les deux attributs a et b sont jugés d'autant plus voisins que la valeur du nombre de sujets les possédant également est invraisemblablement grande par rapport à l'hypothèse d'absence de liaison à laquelle nous avons donné deux formes N_1 et N_2. Le calcul de $P_1(a, b)$ ou de $P_2(a, b)$ est aisé si on tient compte de ce que la distribution théorique de $(\mathcal{S}_1 - \mu)/\sqrt{\mu}$

dans l'hypothèse N_1 ou de $(S_2 - \mu)/\sqrt{\sigma_2}$ dans N_2, est approximativement normale, centrée réduite ($\mathcal{N}(0, 1)$). Cependant, la variance des données observées peut être si grande, qu'il est nécessaire de remplacer les hypothèses N_1 et N_2 par N'_1 et N'_2 où $(S_1 - \mu)/\sqrt{\mu}$ et $(S_2 - \mu)/\sqrt{\sigma_2}$ suivent respectivement les lois normales $\mathcal{N}(0, \lambda_1)$ et $\mathcal{N}(0, \lambda_2)$. λ_1 et λ_2 ont été déterminés de telle sorte que la plus grande valeur observée de $\mathcal{S}_i(a, b)/\lambda_i$ ($i = 1$ ou 2) soit inférieure ou égale à 2,5 (cf. [10]), ceci permet un degré de discrimination suffisant à l'échelle définie par la loi normale en calculant les proximités par les formules

$$P'_1(a, b) = \int_{-\infty}^{(s-\mu)/\sqrt{\mu\lambda_1}} \frac{1}{\sqrt{2\pi}} \exp.\left(-\frac{x^2}{2}\right) dx ,$$

$$P'_2(a, b) = \int_{-\infty}^{(s-\mu)/\sigma_2\sqrt{\lambda_2}} \frac{1}{\sqrt{2\pi}} \exp.\left(-\frac{x^2}{2}\right) dx.$$

Introduisons ici une information plus faible concernant les ressemblances entre éléments de l'ensemble à classifier, ici A. Nous avons pu établir (cf. [5], Chap. 2) que l'Information la plus générale que peut nécessiter la recherche d'une classification, sans pour autant compromettre la fécondité des résultats, est de type suivant : étant donné un couple quelconque de paires d'attributs $(\{a, b\}, \{c, d\})$, savoir laquelle des deux paires $\{a, b\}$ ou $\{c, d\}$ est formée des attributs les plus proches ; cette proximité qui est compromise au niveau du comportement de la population se calcule par l'un des deux indices $\mathcal{S}_1(a, b)$ ou $\mathcal{S}_2(a, b)$ ou bien de façon tout à fait équivalente ici, par $P'_1(a, b)$ ou $P'_2(a, b)$. Cette information est un rangement des paires d'attributs par ressemblance décroissante ; ce préordre total sur l'ensemble des paires de A se nomme "préordonnance" sur A. Comme en général pour ce rangement, les paires, dont les composantes sont également proches, sont très rares ; on peut quasiment considérer que ce préordre total est un ordre total.

Si nous avons tenu à donner deux formes à l'hypothèse N et à associer à chacune d'entre elles un indice de proximité, c'est qu'on se propose de comparer les résultats obtenus pour chacun de ces deux indices qui sont finalement assez voisins.

3 – CARACTERE NEUTRE D'UN ATTRIBUT ET CLASSIFICABILITE

Nous avons déjà souligné dans l'introduction que notre classification permet de dégager les tendances principales du comportement et les diverses composantes de chacune d'elles. Mais cette décomposition en tendances peut être plus ou moins prononcée. Pour recouvrir cette préoccupation nous avons caractérisé et mesuré l'aptitude d'un ensemble A, à être organisé en une hiérarchie de classifications sensée respecter de façon satisfaisante le système des inégalités entre les ressemblances par paires défini par la préordonnance (cf. [5], Ch. 3). On montre que la donnée d'une chaîne de partitions, ordonnée par finesse décroissante, sur A, est équivalente à la donnée d'une préordonnance "ultramétrique" sur A ; c'est-à-dire, d'un préordre total sur l'ensemble B des paires d'éléments de A, qui est caractérisé par la propriété suivante : pour toute partie $\{a , b , c\}$ à trois éléments de A pour laquelle on a $\{a , b\} \leqslant \{b , c\} \leqslant \{a , c\}$ pour le préordre, la paire médiane $\{b , c\}$ et la paire supérieure $\{a , c\}$ sont dans une même classe du préordre (i. e. $\{a , c\} \leqslant \{b , c\}$). Le degré de classificabilité est défini par une distribution qui caractérise la distorsion de la structure de la préordonnance sur A, définie par un indice de proximité, par rapport à une structure ultra-métrique de cette dernière, comme suit : à chaque paire p de B, nous associons la proportion $\varphi(p)$ de triplets dont la paire médiane et la paire supérieure sont strictement séparées par p. La classificabilité est la suite décroissante des valeurs pondérées de la fonction $\varphi(p)$; la valeur α de φ étant pondérée par la proportion des paires dont $\varphi(p) = \alpha$.

Exemple : soit A = a , b , c , d , e et soit ω la préordonnance suivante sur A

$$\{a , d\} = \{a , c\} < \{a , e\} < \{c , e\} < \{b , d\} = \{c , d\} < \{b , c\} < \{d , e\} < \{a , b\} < \{b , e\},$$

la classificabilité peut être représentée par la courbe en escalier dont un palier

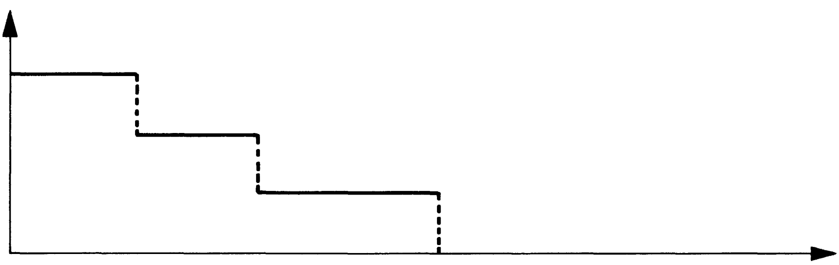


Figure 3

représente les paires qui interviennent pour séparer strictement la paire médiane de la paire supérieure *d'un même nombre de triplets* ; la largeur du palier représente la proportion de telles paires et sa hauteur la proportion de triplets dont une quelconque de ces paires sépare strictement la paire médiane de la paire supérieure.

La classificabilité de A est intimement liée à l'éloignement relatif de chacun des attributs par rapport à la suite des autres ; l'analyse des données préalable à une classification automatique (cf. [6]) conduit à mesurer le caractère neutre d'un attribut a, par rapport à une visée classificatoire, par la petitesse de la variance des proximités à a : $\{\mathcal{S}(a , x) / x \in A - \{a\}\}$, soit

$$\mathcal{V}(a) = \frac{1}{(m-1)} \sum_{\{x/x \neq a\}} [\mathcal{S}(a , x) - \mathcal{S}(a)]^2$$

où $m = \text{card}(A)$ et où $\mathcal{S}(a)$ est la moyenne des $\mathcal{S}(a , x)$.

Si dans un tableau de données il y a un petit sous ensemble de quelques attributs neutres qui perturbent la nature classifiable de l'ensemble, on peut les extraire à priori. La valeur de $\mathcal{V}(a)$ est d'autant plus grande que a intervient plus intimement dans la formation d'une classe. Nous étudions dans [7] une représentation géométrique obtenue par une analyse simultanée de la moyenne et de la variance des proximités à l'intérieur d'une même classe qui permet d'organiser géométriquement les éléments de cette classe à partir de ses éléments extrémaux les moins liés ; cette technique permet en outre de déterminer l'élément le plus central de la classe et une échelle d'attitude lorsqu'elle existe.

4 – ARBRE DETAILLE PUIS CONDENSE DES CLASSIFICATIONS

La première étape de condensation de l'ensemble des attributs descriptifs se présente sous la forme d'une hiérarchie détaillée de classifications (cf. fig. 1, § 1), sur A, qui s'obtient pas à pas par réunions successives de classes, permettant de passer d'une partition à une suivante moins fine. La partition la plus fine est celle où toute classe contient un seul attribut, la partition la moins fine est celle où tous les attributs sont réunis dans une même classe. L'algorithme de classification va fonctionner à partir de la classification la plus fine en réunissant à chaque pas les deux classes les plus "proches". Pour établir l'algorithme il est donc nécessaire d'étendre la notion de proximité entre deux attributs descriptifs à celle *entre deux classes d'attributs.*

Rappelons (cf. § 2) que l'indice de proximité entre les deux attributs a et b se présente sous la forme d'une probabilité ; $P(a, b) = Pr^N\{S < s\}$ où S est la v. a. associée dans l'hypothèse N à s : nombre d'individus possédant simultanément les deux attributs. Cette mesure de proximité se généralise pour comparer deux classes d'attributs de A ; si C et D sont deux parties disjointes de A de cardinaux respectifs l et m, la proximité entre C et D sera définie par

$$P(C, D) = \{\max\{P(c, d)/(c, d) \in C \times D\}\}^{lm}$$

P(C , D) est également défini par rapport à l'hypothèse N d'absence de structure (cf. [8]) ; il s'agit de la probabilité dans cette hypothèse que la plus grande proximité P(c , d) entre un élément c de C et un élément d de D, soit inférieure à celle observée. L'algorithme associé à cette proximité a été appelé "Algorithme de la Vraisemblance du Lien" (A.V.L.). Ce qu'il y a de nouveau par rapport à d'autres méthodes de classification hiérarchique c'est qu'ici, la distribution de P(C , D) est appréhendée dans l'hypothèse N d'absence de liaison ; ce qui permet la référence à une échelle claire pour juger de la grandeur observée d'une mesure de proximité entre deux classes et pour comparer sans biais les mesures attachées à deux couples de classes.

Une seconde étape décisive de la méthode consiste à condenser l'arbre à ses niveaux les plus pertinents et ce, au moyen d'une statistique mesurant la proximité entre l'information de départ qu'est la préordonnance sur A, $\omega(A)$ et la classification *déjà formée à un niveau donné*. Pour cela, on regarde la partition sur A comme effectuant un classement des paires d'éléments de A en deux groupes : celui R des paires réunies (à composantes proches du point de vue de la partition) et celui S des paires séparées (à composantes éloignées du point de vue de la partition) ; en d'autres termes, la partition définit sur l'ensemble B des paires d'éléments de A un préordre total à deux classes R et S où $(p, q) \in R \times S \Rightarrow p < q$ et non $q < p$.

La base de la constitution de la statistique des niveaux est

$$\text{card}\{gr(\omega) \cap R \times S\}$$

où $gr(\omega) = \{(p, q)/(p, q) \in B \times B, p < q \text{ et non } q < p \text{ pour } \omega\}$ est le graphe de ω dans $B \times B$.

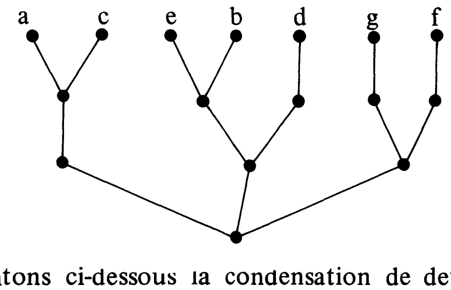
J.P. Benzecri a introduit ce cardinal sous la forme du "nombre d'inégalités entre les proximités spécifiés par la partition et compatibles avec ω".

ω étant un ordre total sur B, nous démontrons dans [8] que la distribution de card $\{gr(\omega) \cap R \times S\}$, lorsque la partition décrit uniformément l'ensemble des partitions de même type (i. e. dont les cardinaux des classes sont fixés), est asymptotiquement normale de *moyenne* $r.s/2$ et de *variance* $r.s(t+1)/12$ où $r = \text{card}(R)$, $s = \text{card}(S)$ et $t = \text{card}(B)$; $t = r + s$. C'est la statistique

$$\Sigma = \left\{\text{card}\{gr(\omega) \cap R \times S\} - \frac{r.s}{2}\right\} / \sqrt{r.s(t+1)/12} \qquad (1)$$

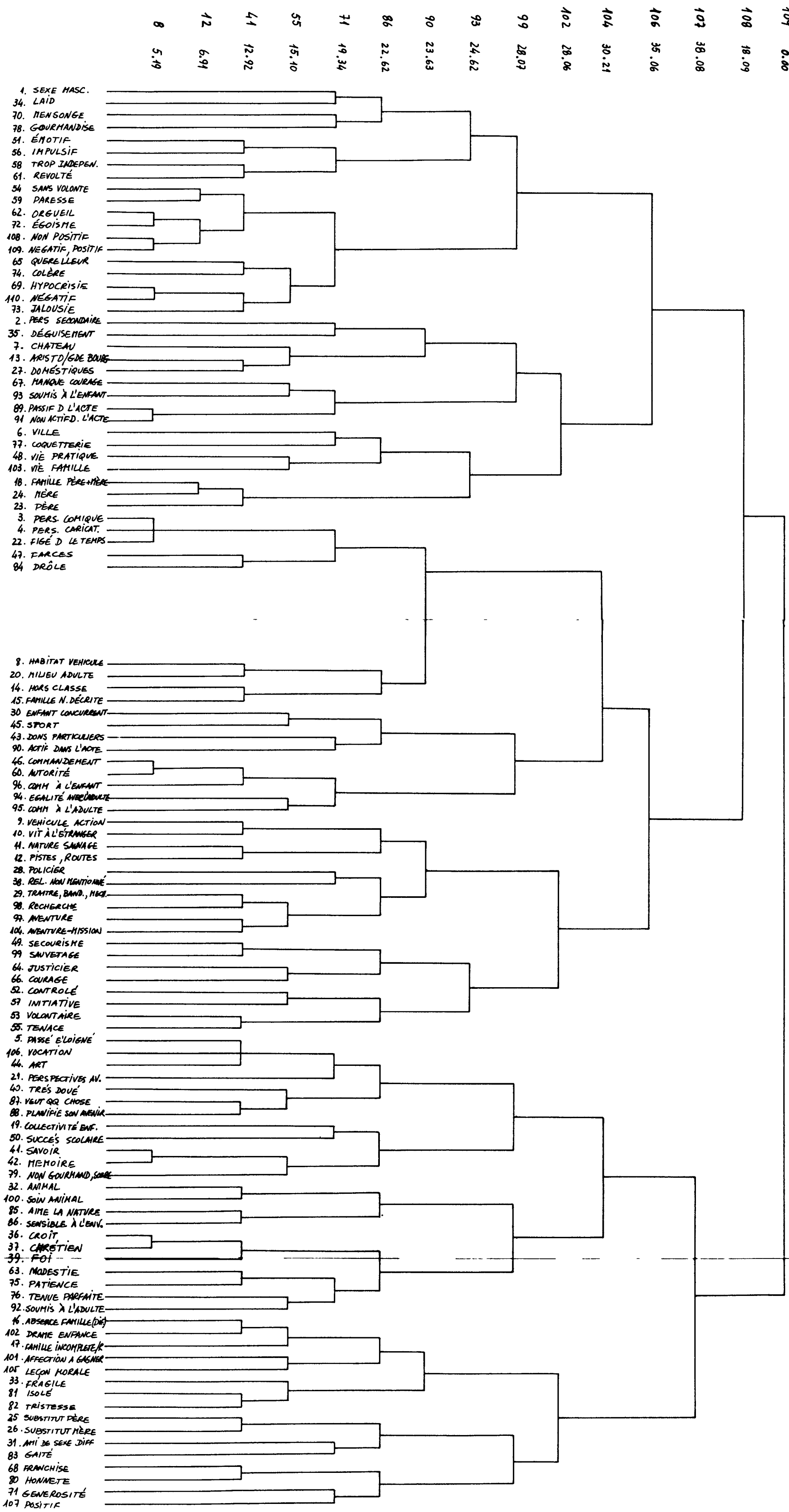
qui mesurera le degré de cohésion des classes formées à un niveau donné ; cette dernière sera appelée statistique "globale" des niveaux, elle est globale par rapport à celle "locale" θ, où on retient pour un niveau donné l'accroissement de (1) entre le niveau précédent et ce dernier. Nous retenons comme les plus significatifs les noeuds des niveaux correspondants aux maximum locaux de ce taux d'accroissement ; l'expérience a en effet montré que la distribution de la statistique locale le long de la suite des niveaux est telle que sa valeur augmente lorsqu'une classe en cours de formation se confirme et décroît sensiblement devant l'arrêt de constitution d'une classe ayant quelque consistance au profit de la naissance de l'embryon d'une autre classe. Cette dernière étape est très importante car elle permet sur l'arbre détaillé des classifications de définir les distinctions significatives entre classes et sous classes d'attributs.

La condensation de l'arbre de la figure 1 aux niveaux 2 et 4 donne



Nous présentons ci-dessous la condensation de deux arbres obtenus par l'algorithme de la vraisemblance du lien et associés respectivement aux hypothèses N'_1 et N'_2 présentées au paragraphe 2. Ces deux arbres ont été produits et interprétés(*) par Mme M.H. Nicolau dans le cadre d'une thèse de 3[e] cycle qui

(*) Avec l'aide précieuse de Cl. Bellan collaborateur de Mme Chombard de Lauwe.



- TYPOLOGIE ENFANTINE -

SIMILARITÉ 1 ... AGRÉGATION A.V.L., λ = 9 porte sur l'analyse de la statistique de proximité P(C , D) entre deux classes de variables que nous avions introduite (cf. ci-dessus). L'algorithme de Mme Nicolau permet de façon très générale de passer d'un indice de proximité entre parties disjointes de A à la représentation polonaise de l'arbre détaillé des classifications. Après l'interprétation de l'arbre détaillé puis condensé associé à N'_1 où on utilise l'indice de proximité P'_1 (a , b) entre attributs (cf. § 2) ; on comparera par rapport à ce dernier l'arbre condensé associé à N'_2 où on utilise P'_2 (a , b).

5 – ETUDE ET COMPARAISON DES DEUX ARBRES DEFINISSANT LA TYPOLOGIE DES PERSONNAGES ENFANTS DE LA LITTERATURE "ENFANTINE"

a) Arbre associé à l'A.V.L. établi par rapport à N'_1

L'arbre est binaire et comporte donc 109 niveaux (il y
Les niveaux retenus pour le condenser et qui correspondent à des maximum locaux du *taux d'accroissement* θ de la statistique globale Σ des niveaux sont ceux d'indice 8, 12, 41, 55, 71, 86, 90, 93, 99, 102, 104, 106, 107 qui donne le maximum absolu de la statistique globale et le niveau 108.

A partir du $41^{ème}$ niveau quelques types intéressants commencent à se dégager ; c'est ainsi qu'on voit apparaître un groupe constitué par les attributs : orgueil et égoïsme, non positif et négatif évoluant vers positif, paresse et sans volonté ; définissant un type de personnage dont le caractère est "plein de défauts et faible" mais non foncièrement mauvais puisqu'il évolue vers le bon. Au $55^{ème}$ niveau une autre forme de négativisme se précise, composée d'abord du noyau : négatif, hypocrisie et jalousie auquel s'adjoint des manifestations extérieures désagréables : colère, querelleur ; il s'agit de l'"aspect le plus abject du négativisme". Ces deux classes se réuniront au $71^{ème}$ niveau, donnant lieu à un type négatif plus général. Parallèlement plusieurs noyaux se forment. A ce même niveau 71 l'"agressif" apparaît avec révolté et trop indépendant associés à impulsif et émotif. Au $86^{ème}$ niveau on dégage un groupe de "défauts mineurs" : gourmandise et mensonge curieusement associés à un garçon laid. Le "bourgeois-aristocrate" composé des attributs : aristocratie grande bourgeoisie, domestiques, château ; le "falot" : passif dans l'acte, non actif dans l'acte, manque de courage, soumis à l'enfant ; le "comique" : personnage comique, caricatural, figé dans le temps, farces, drôle ; l'"autoritaire" : autorité, commandement, commande à l'enfant, égalité avec l'adulte, commande à l'adulte ; l'"aventurier" : aventure, aventure-mission, milieu de traître-bandit, recherche ; le "célèbre" : passé éloigné, vocation, art, perspectives d'avenir, veut quelque chose, planifie son avenir, très doué ; le "brillant" : savoir, mémoire, sobre, succès scolaires, collectivité enfants ; le "sensible" : sensible à l'environnement, aime la nature, présence d'un animal, soin à animal ; le "bon chrétien" : chrétient, foi, croit, modestie patience, tenue parfaite, soumis à l'adulte ; le "malheureux" : tristesse, isolé, fragile, affection à gagner, leçon morale, drame enfance, absence famille (difficultés), famille incomplète ou de remplacement.

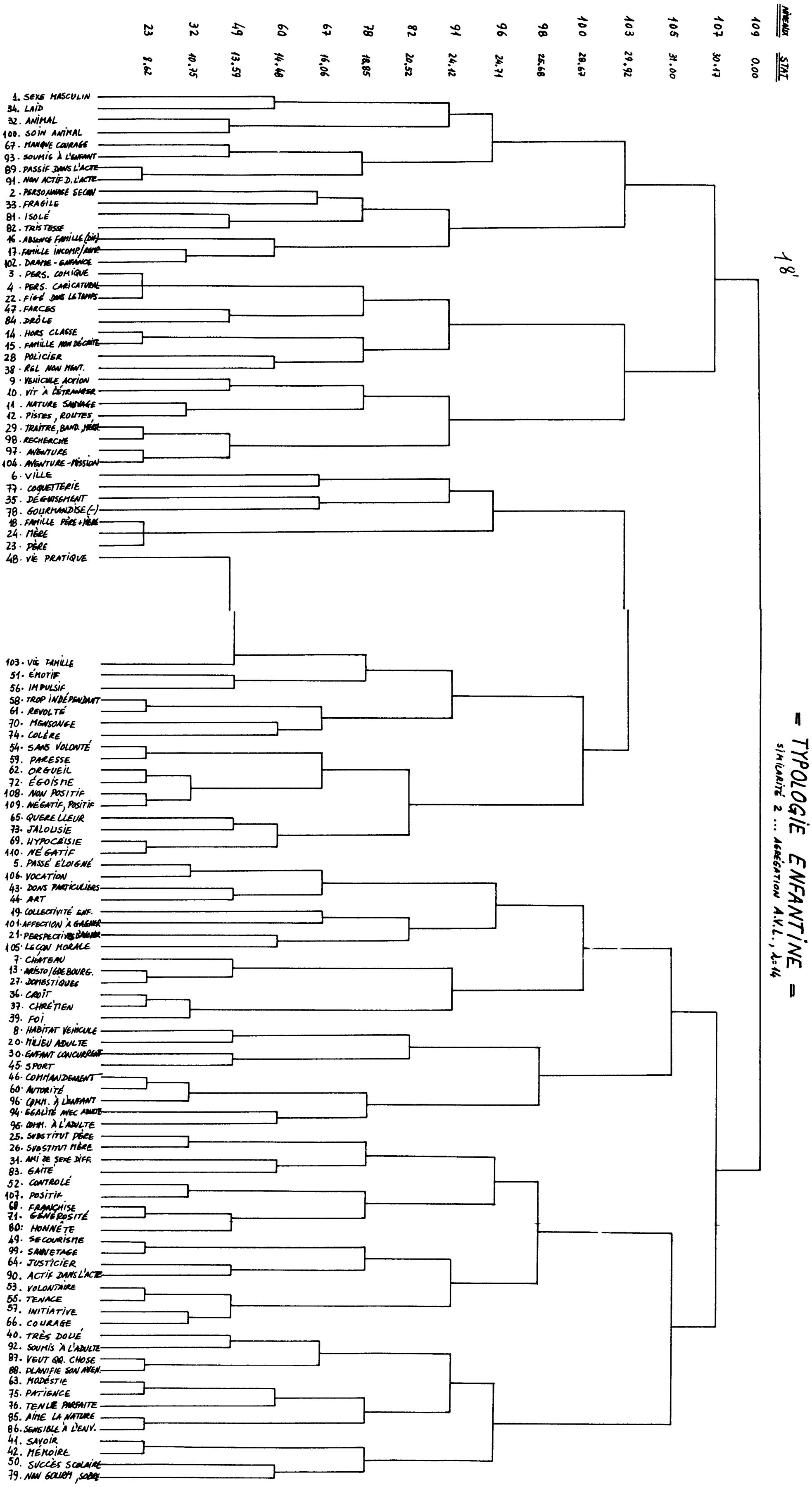
D'autre part sont apparues plusieurs agrégats moins prononcés formés pour la plupart d'attributs relativement neutres (cf. § 3) ; de plus, il faut signaler que quelques unes des réunions dont ont résulté les classes ci-dessus mentionnées, correspondent à des niveaux où la statistique globale Σ présente une diminution ; ce qui exprime que, considérée globalement, la partition de ce niveau est moins en accord avec l'ordonnance initiale que celle qui la précède. Dans ce cas, nous préférons ne pas insister sur la signification de tels groupements dont un exemple est l'association de mensonge et gourmandise à garçon laid. Entre les niveaux 86 et 99 les petits agrégats vont se réunir entre eux et donner naissance à quelques types plus généreux.

Le $99^{ème}$ niveau est celui le plus important avant le niveau le plus significatif (le $107^{ème}$) ; d'une part parce que l'accroissement θ de la statistique Σ à ce niveau est grande (par rapport aux valeurs de θ sur l'ensemble de tous les niveaux) ; d'autre part, parce que les réunions aux niveaux suivants (100 et 101) entraînent des décroissances très nettes de cette statistique. On prendra donc comme classifications les plus pertinentes celles fournies par le $107^{ème}$ et par le $99^{ème}$ niveau. A ce dernier s'impose d'abord une grosse classe résultant de l'agrégation du type "agressif" et des "défauts mineurs" au "négatif" pour former un type "négatif" assez général ; les types "falot", "comique" et "aventurier" se sont enrichis de quelques attributs ; le "sportif" et l'"autoritaire" apparaissent maintenant associés ; le commandement semble présent dans le sport, ou plus généralement dans l'action. "célèbre" et "brillant" se réunissent pour donner lieu au type de l'"enfant-modèle" ; par ailleurs, le "bon chrétien" se rattache au "sensible", le "malheureux" au "fragile". Deux derniers types se forment ; l'un qu'on appelera le "héros scout" défini par : secourisme, sauvetage, justicier, courage, contrôlé, initiative, volontaire, tenace ; l'autre, représentant l'"ami loyal, le bon camarade" composé de franchise, honnête, générosité, positif, gaité, ami de sexe différent, substitut mère, stubstitut père. Ces deux classes ont, en fait, une cohésion interne très faible ; les attributs qui les constituent étant assez neutres.

Enfin au $107^{ème}$ niveau apparait une classification en trois grandes classes. La première regroupe le "négatif" et le "falot" et représente donc un type "négatif" très général. La deuxième englobe le "comique", le "sportif-autoritaire", l'"aventurier" et le "héros scout", définissant le personnage type de la littérature d'évasion. La troisième classe rassemble l'"enfant modèle", le "bon chrétien", le "malheureux" et l'"ami loyal" définissant un type "normatif" général. Les derniers niveaux qui voient s'agréger ces classes deux à deux donnent lieu à des chutes très grandes de la statistique globale Σ.

b) **Arbre associé à l'A.V.L. établi par rapport à N'_2**

Nous allons examiner la condensation de cet arbre à ses niveaux les plus significatifs d'indices 23, 32, 49, 60, 78, 82, 91, 96, 98, 100, 103, 105 et 107. Le niveau le plus significatif est le $105^{ème}$; le $103^{ème}$ donne le dernier maximum local pertinent de la statistique globale Σ avant son maximum absolu. Au $103^{ème}$ niveau se dégage le type "négatif" composé des sous types suivants :


= TYPOLOGIE ENFANTINE =
SIMILARITÉ 2 ... AGRÉGATION A.V.L., λ=14
18'
NIVEAUX STAT
109 0.00
107 30.17
106 31.00
103 29.92
100 28.67
98 25.68
96 24.71
91 24.12
82 20.52
78 18.85
67 16.06
60 14.48
49 13.59
32 10.75
23 8.62
1. SEXE MASCULIN
34. LAID
32. ANIMAL
100. SOIN ANIMAL
67. MANQUE COURAGE
93. SOUMIS À L'ENFANT
89. PASSIF DANS L'ACTE
91. NON ACTIF D. L'ACTE
2. PERSONNAGE SECON
33. FRAGILE
81. ISOLÉ
82. TRISTESSE
16. ABSENCE FAMILLE (bis)
17. FAMILLE INCOMP/ANOR
102. DRAME - ENFANCE
3. PERS. COMIQUE
4. PERS. CARICATURAL
22. FIGÉ DANS LE TEMPS
47. FARCES
84. DRÔLE
14. HORS CLASSE
15. FAMILLE NON DÉCRITE
28. POLICIER
38. REL. NON MENT.
9. VÉHICULE ACTION
10. VIT À L'ÉTRANGER
11. NATURE SAUVAGE
12. PISTES, ROUTES
29. TRAÎTRE, BAND. MÉCH
98. RECHERCHE
97. AVENTURE
104. AVENTURE - MISSION
6. VILLE
77. COQUETTERIE
35. DÉGUISEMENT
78. GOURMANDISE (-)
18. FAMILLE PÈRE + MÈRE
24. MÈRE
23. PÈRE
48. VIE PRATIQUE
103. VIE FAMILLE
51. ÉMOTIF
56. IMPULSIF
58. TROP INDÉPENDANT
61. RÉVOLTÉ
70. MENSONGE
74. COLÈRE
54. SANS VOLONTÉ
59. PARESSE
62. ORGUEIL
72. ÉGOÏSME
108. NON POSITIF
109. NÉGATIF, POSITIF
65. QUERELLEUR
73. JALOUSIE
69. HYPOCRISIE
110. NÉGATIF
5. PASSÉ ÉLOIGNÉ
106. VOCATION
43. DONS PARTICULIERS
44. ART
19. COLLECTIVITÉ ENF.
101. AFFECTION À GAGNER
21. PERSPECTIVES D'AVENIR
105. LEÇON MORALE
7. CHÂTEAU
13. ARISTO/GDE BOURG.
27. DOMESTIQUES
36. CROIT
37. CHRÉTIEN
39. FOI
8. HABITAT VÉHICULE
20. MILIEU ADULTE
30. ENFANT CONCURRENT
45. SPORT
46. COMMANDEMENT
60. AUTORITÉ
96. COMM. À L'ENFANT
94. ÉGALITÉ AVEC ADULTE
95. COMM. À L'ADULTE
25. SUBSTITUT PÈRE
26. SUBSTITUT MÈRE
31. AMI DE SEXE DIFF.
83. GAÎTÉ
52. CONTRÔLÉ
107. POSITIF
68. FRANCHISE
71. GÉNÉROSITÉ
80. HONNÊTE
49. SECOURISME
99. SAUVETAGE
64. JUSTICIER
90. ACTIF DANS L'ACTE
53. VOLONTAIRE
55. TENACE
57. INITIATIVE
66. COURAGE
40. TRÈS DOUÉ
92. SOUMIS À L'ADULTE
87. VEUT QQ. CHOSE
88. PLANIFIE SON AVEN.
63. MODESTIE
75. PATIENCE
76. TENUE PARFAITE
85. AIME LA NATURE
86. SENSIBLE À L'ENV.
41. SAVOIR
42. MÉMOIRE
50. SUCCÈS SCOLAIRE
79. NON GOURM., SOBRE

"caractère plein de défauts et faible" trouvé dans l'arbre précédent ; le "mauvais" ; l'"agressif" qui semble plus heureusement se compléter avec l'attribut colère qui était associé au "mauvais" dans l'arbre précédent ; et enfin, un sous type difficile à définir qui semble représenter un encadrement de l'enfant, familial et citadin marqué par l'insouciance et la légereté (vie de famille, vie pratique, ville, déguisement, coquetterie, gourmandise, . . .). A ce 103ème niveau le "comique" et l'"aventurier" sont déjà réunis. Le "comique" a donné l'habitat-véhicule et milieu adulte au "sportif-autoritaire" et pris policier, religion non mentionnée à l'"aventurier". Il s'agit là d'attributs qui, selon le psychologue apparaissent dans la littérature associés à n'importe lequel de ces trois types. L'"aventurier" se compose comme précédemment de deux groupes ; le premier caractérisant l'atmosphére dans laquelle il vit : aventure, aventure-mission, traître-bandit, recherche ; et le second, son milieu physique : véhicule-action, vit à l'étranger, nature sauvage, pistes-routes. Le "malheureux" est ici un type plus global qui contient le "falot" et un groupe de faible cohésion : sexe masculin, laid, animal, soin animal. L'association du "falot" au "malheureux" s'expliquerait par la soumission dans le malheur ; l'ensemble représente un personnage malheureux, soumis et laid qui cherche une compensation en prenant soin d'un animal. Sexe masculin et laid qui faisaient partie du "négatif" dans l'arbre précédent ont sauté vers le "malheureux" ; ce qui correspond au fait que dans la Littérature enfantine, la laideur d'un petit garçon vient généralement associée au mauvais caractère ou au mauvais sort, parfois aussi au "comique" (le "clown" ; résultat d'un arbre non signalé ici). De même l'attribut : personnage secondaire qui dans le premier arbre se réunissait au "falot" et maintenant au "fragile" représente, selon le psychologue, un personnage "faire valoir" souvent associé à un personnage principal faible. Le "bon exemple chrétien" a éclaté en deux parties : l'ensemble des qualités {modestie, patience, sensible à l'environnement} s'agrège au "brillant" et au "doué" pour former le type de l'"enfant modèle" et l'embryon {croit, chrétien, foi} se réunit au "célèbre" et au "bourgeois aristocrate" donnant lieu à un type qui semble représenter le personnage des "biographies d'enfants célébres". Enfin, on rencontre réunis le "héros scout" et l'"ami loyal, bon camarade" rendant compte d'une autre sorte d'enfant modèle ; si la première forme est axée sur les qualités de l'esprit, la seconde l'est plutôt sur les qualités de coeur. Au niveau le plus significatif, le 105ème, nous avons une partition en cinq classes : le "malheureux", le "comique-aventurier", le "négatif", le "modèle" et un groupe résultant de la réunion du "célébre-bourgeois-chrétien" avec le "sportif-autoritaire" dont la constitution préalable avait été accompagnée d'un minimum local de la statistique globale.

Si nous nous intéressons maintenant aux suites des valeurs de Σ associées à chacun des arbres ; on constate que les groupements responsables des dissemblances les plus remarquables entre les deux arbres correspondent en général à des minimums locaux de Σ. Cela se produit par exemple lorsque

– {sexe masculin, laid} se réunit dans le premier arbre à {mensonge, gourmandise} et dans le second à {animal, soin animal}

– dans le premier arbre, "falôt" s'associe au "bourgeois-aristocrate" et lorsque l'ensemble se réunit à {ville, coquetterie, vie pratique, . . .} ; dans le second arbre, lorsque "falot" s'associe au "petit garçon laid . . .", puis au "fragile".

Dans le premier arbre la réunion de l'"ami loyal – bon camarade" au "malheureux" est également accompagnée d'un minimum de Σ, ainsi que celle du "malheureux" au "comique et aventurier" dans le second arbre.

Il s'agit là de réunions qui vont un peu à l'encontre de l'information sur les ressemblances définie par la préordonnance et où les attributs neutres jouent un rôle important. D'autres regroupements où Σ n'a pas de minimums locaux mais qui varient aussi d'un arbre à l'autre, sont d'autre part dûs à la présence de tels attributs "neutres" (cf. § 3) ; les classes donnent alors lieu à des interprétations également valables mais faiblement apparentes. On se rend compte que l'identification des attributs neutres dans l'ensemble A est d'une aide précieuse pour l'interprétation des résultats. Une expérience où on n'a retenu que les 75 attributs les plus discriminants a fait apparaître par l'A.V.L. les cinq types "négatif", "comique", "aventurier", "malheureux" et "modèle" ; lesquels apparaissent de façon très nette. Il est intéressant de noter que ce sont ces cinq types que nous avions découvert au niveau le plus significatif d'un arbre de classifications issu de l'algorithme "lexicographique" (cf. [5] Chap. 3) sur un ensemble de 67 attributs parmi les 75 considérés ici. L'algorithme "lexicographique" qui a précédé l'A.V.L. consiste à saturer la suite des sections commençantes de la préordonnance, définissant une suite de relations d'équivalence de moins en moins fines ; l'arbre des classifications s'obtient en associant à chaque relation d'équivalence la partition sur A qu'elle définit. On montre que cet algorithme revient à réunir à chaque pas les deux classes de variables les plus proche au sens de la plus grande proximité observée entre deux éléments appartenant respectivement aux deux classes. La classification la plus significative qui avait été obtenue, avec comme indice $\mathcal{S}_1(a , b)$, est la suivante :

Le "négatif"

Non positif, égoïsme, orgueil, paresse, sans volonté, hypocrisie, jalousie, querelleur, colère, trop indépendant, révolté, impulsif, mensonge, coquetterie, manque de courage, passif dans l'acte.

Le "comique"

Personnage comique, caricatural, figé dans le temps, drôle, hors classe, famille non décrite, farces.

L'"aventurier"

Aventure, aventure-mission, aventurier, traître-bandit, recherche, policier, nature sauvage, pistes-routes, véhicule-action, vit à l'étranger.

Le "malheureux"

Isolé, tristesse, absence famille (difficultés), drame enfance, fragile, affection à gagner, famille incomplète, substitut père, substitut mère.

Le "modèle"

Passé éloigné, célèbre, vocation, art, savoir, mémoire, veut quelque chose, planifie son avenir, très doué, sobre, croit, chrétien, foi, tenue parfaite.

6 – RETOUR AUX INDIVIDUS

Relativement à la classification définie ci-dessus, nous avons cherché à connaître les personnages les plus responsables de cette classification en affectant à chacun des 1 500 individus la valeur de la statistique

$$\frac{1}{k} \sum_{1 \leqslant j \leqslant k} (f_j - \bar{f})^2 \Big/ \frac{\bar{f}(1 - \bar{f})}{n} \tag{1}$$

où k est le nombre de classes, f_j la fréquence relative de présence de l'individu en question dans la classe de numéro j et $f = \frac{1}{k} \sum_{1 \leqslant j \leqslant k} f_j$. (1) est le rapport de la variance inter-classe sur celle globale de la fréquence relative d'attributs possédés par l'individu lorsqu'on ne tient pas compte de la richesse de description d'un type (i. e. nombre d'attributs par classe).

Signalons les cinq personnages associés respectivement à chacun des types ci-dessus et ayant la valeur la plus élevée de la statistique (1). Pour le type "négatif", "Malvina" de "Rose et Violette" roman de Mme Chabrière Reides (1929) ; petite pauvresse mal élevée, elle a tous les défauts.

Pour le "comique", "Lola" tiré des périodiques "Zéphir", "Bravo" et "Cabriole" (Editions Artima) ; petite fille indépendante, très drôle, très espiègle, fait tout le temps des farces. Pour l'"aventurier", "Mic, Mac et le Robotor" de "Lisette" (journal) ; frère et soeur de même caractère : aventureux, courageux, débrouillard ; aventures du genre policier (très inspiré de Tintin).

Il nous a été par la suite difficile de dégager un personnage "malheureux" qui ne soit "enfant modèle" ; inversement, les premiers individus normatifs apparus avaient des caractéristiques du "malheureux".

Citons quand même pour l'"enfant modèle", "Ligier Richier" de "Enfants d'Alsace et de Lorraine" par Emile Carpentier (1875) ; vocation de sculpteur d'un petit paysan Lorrain, triste, isolé, travaille seul et visite les églises et leurs chefs d'oeuvres. Au grand déplaisir de ses parents suit Michel Ange en Italie. Il devient célèbre.

Pour le type "malheureux", "Wilhem" de "Le monde des enfants" ; contes moraux de Mme Thècle de Guimpert (1874) ; petit orphelin martyrisé par son tuteur, réalisera son rêve . être peintre grâce à son honnêteté, sa ténacité, sa foi en Dieu et son amour de l'art.

On peut s'étonner de constater que les personnages extraits sont assez peu connus, on peut même les accuser d'une certaine fadeur. A les examiner de près, on se rend compte qu'ils sont presque réduits à la pureté schématique d'un type et sont finalement assez pauvres. Ce sont d'ailleurs rarement des réussites littéraires. Les personnages les plus fameux sont certes dominés par un type et recueillant en conséquence une grande valeur de la statistique (1) ci-dessus, mais sont assez riches et empruntent divers attributs à divers types ; c'est ainsi que Tintin (le $50^{\text{ème}}$ pour la valeur de (1)) est surtout un aventurier qui possède quelques caractéristiques du type "normatif" telles que "sobre", "veut quelque chose" ; du type "comique" telles que "figé dans le temps", "famille non décrite" etc. . .

Il eût été également intéressant de considérer le comportement d'une statistique analogue à (1) mais où on tiendrait compte des cardinaux des classes ; elle a pour forme

$$\frac{1}{k} \sum_{1 \leqslant j \leqslant k} \alpha_j (f_j - f)^2 \Big/ \frac{f(1-f)}{n}$$

où α_j est la proportion d'attributs dans la $j^{\text{ème}}$ classe et où $f = \sum_j \alpha_j f_j$ est la proportion d'attributs de A possédés par l'individu en question.

On a pu observer de très grandes valeurs de la statistique (1) ; ceci étant dû à la nature du problème étudié où les personnages sont construits. Pour des expériences portant sur une population formée d'êtres vivant dans un certain contexte, l'intérêt de la découverte de ces individus les plus "typiques" peut être considérable car elle apporte au Sociologue la possibilité d'approfondir sa recherche par un retour sur le terrain pour l'examen plus précis de chacun des sujets extraits. On aura avec chacun d'eux des interviews en profondeur pour des histoires de vie ; on appliquera des tests projectifs, etc. . . Bref, on utilisera des techniques méthodologiques très riches en résultats mais seulement applicables sur un petit nombre de sujets, en raison de la dépense de temps qu'elles nécessitent.

Nous pouvons certes associer à la classification la plus significative sur l'ensemble des attributs, celle duale sur l'ensemble des individus, en affectant chacun d'eux à la classe j pour laquelle la fréquence relative f_j est la plus grande. Cette technique est par trop rigide et F. Nicolau propose de retenir pour chaque sujet, les trois fréquences f_j les plus élevés qu'on rapporte à $\sum_{1 \leqslant j \leqslant k} f_j$; puis, relativement à 3 classes d'attributs représentés respectivement par les trois sommets d'un triangle équilatéral, chaque individu est figuré de façon barycentrique par un point de coordonnées (p , q , r) où p, q et r sont de la forme

$$f_j \Big/ \sum_j f_j$$

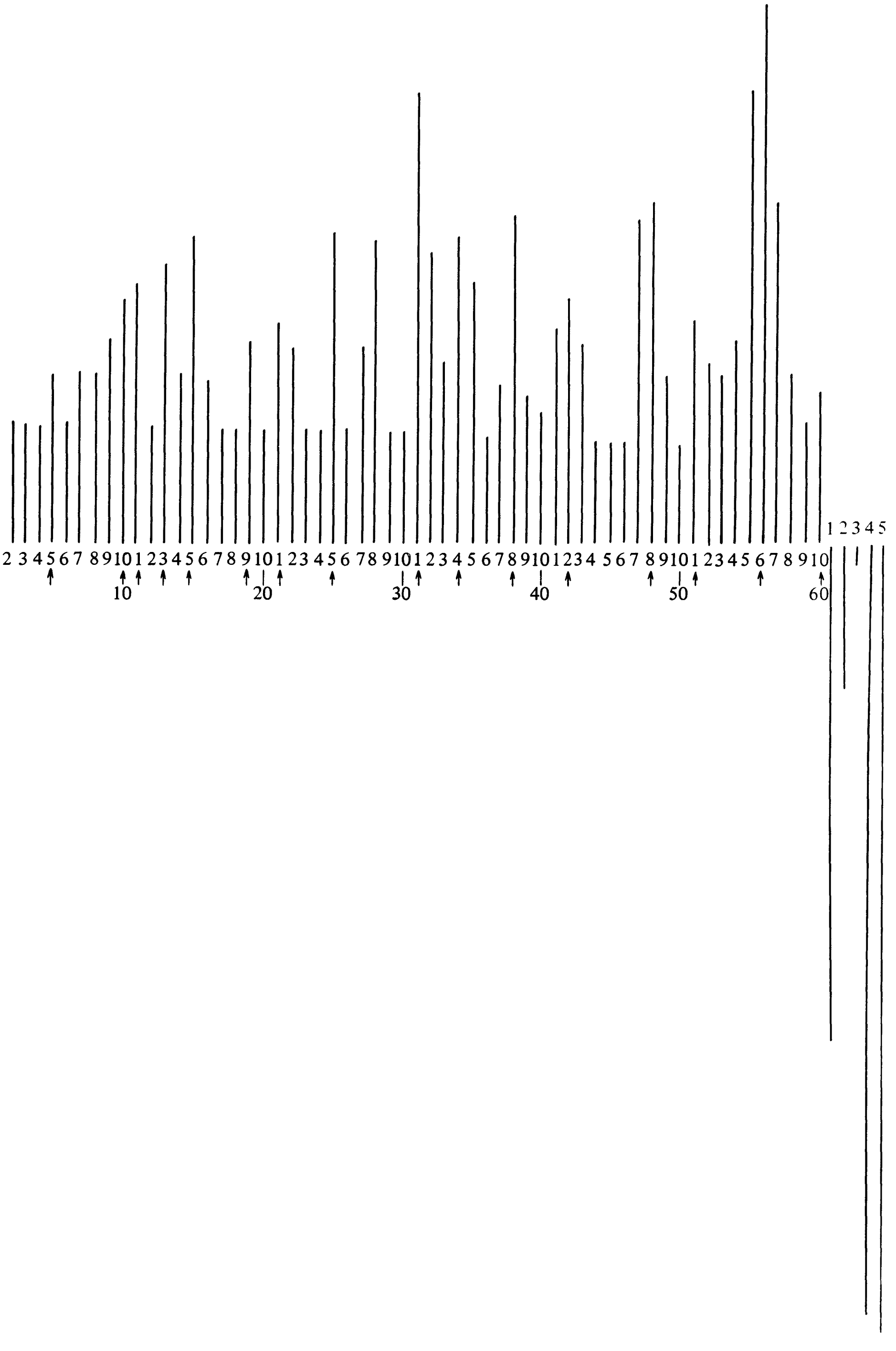
et définissent les degrés d'appartenance de l'individu à chacun des trois types.

Ce retour vers les individus doit permettre au chercheur de répondre aux questions que lui pose la formation des types de comportement.

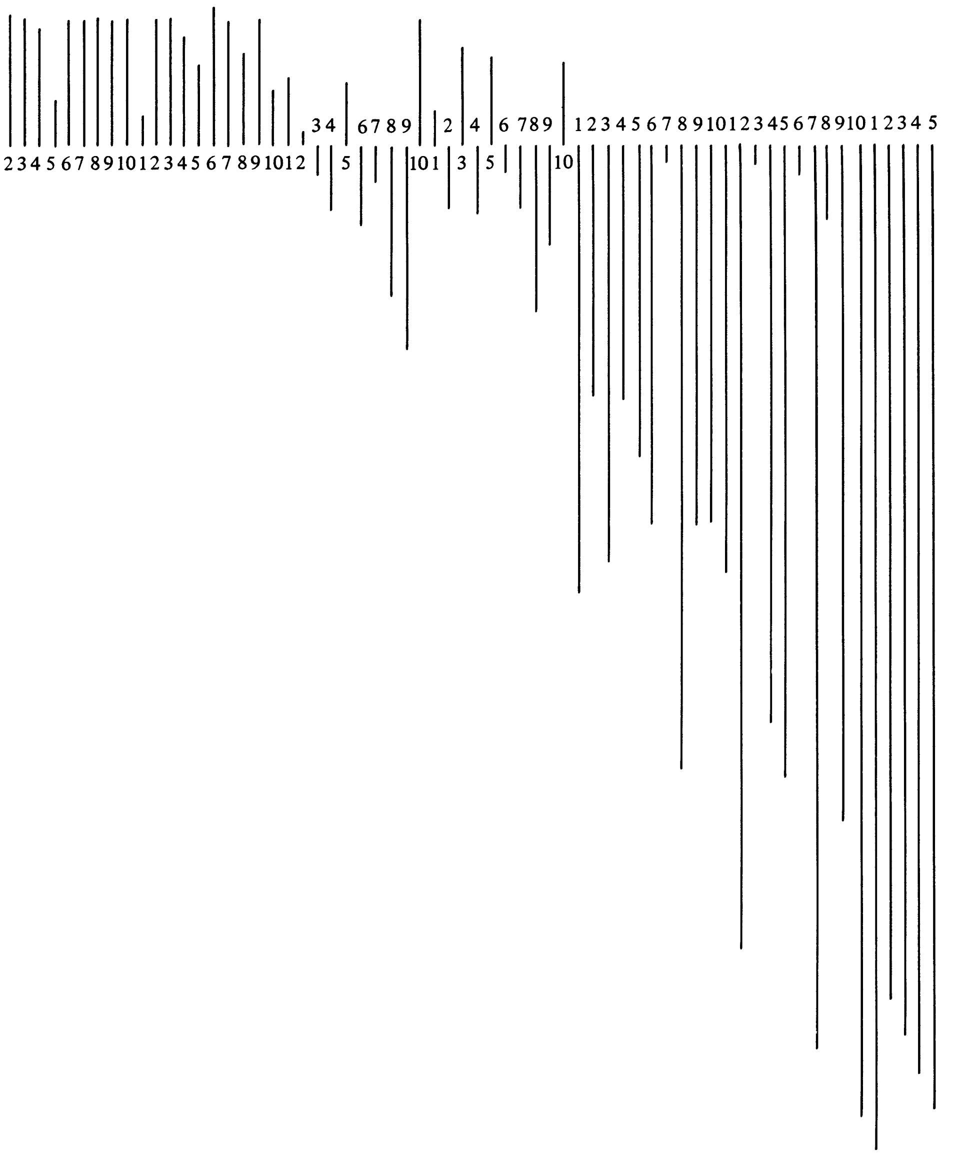
7 – CLASSIFICABILITE ET COHESION DES CLASSES FORMEES

Revenons à l'important problème de détermination des niveaux significatifs au moyen d'une statistique locale des niveaux. Relativement à l'ensemble des 67 attributs et de l'algorithme "lexicographique" considérés ci-dessus ; nous avons cherché à comparer la distribution d'une telle statistique le long de la suite des niveaux pour le cas réel étudié à celle, associée au cas artificiel de l'hypothèse N où à chaque vecteur de description observé de taille fixée (nombre

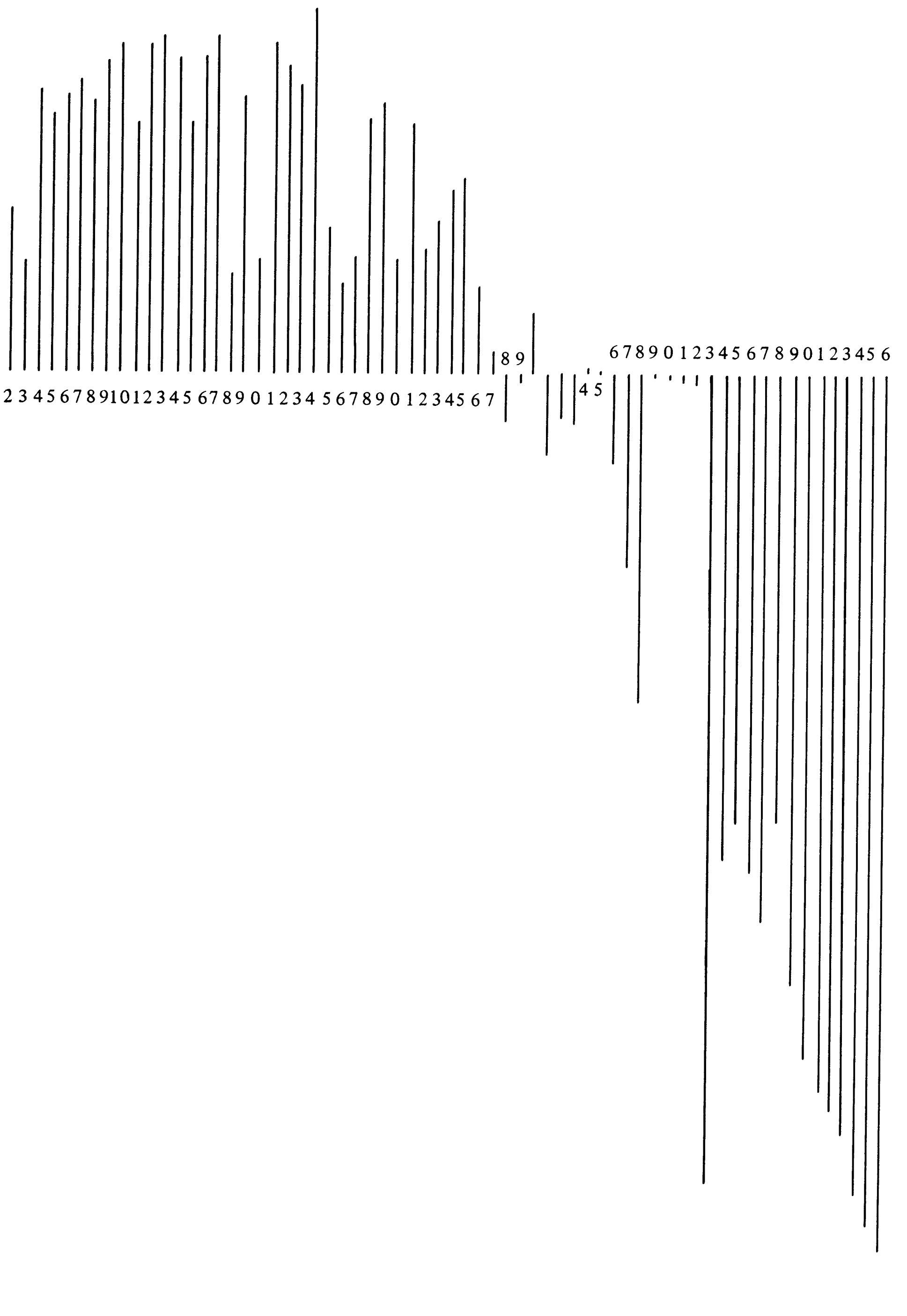
Cas réel



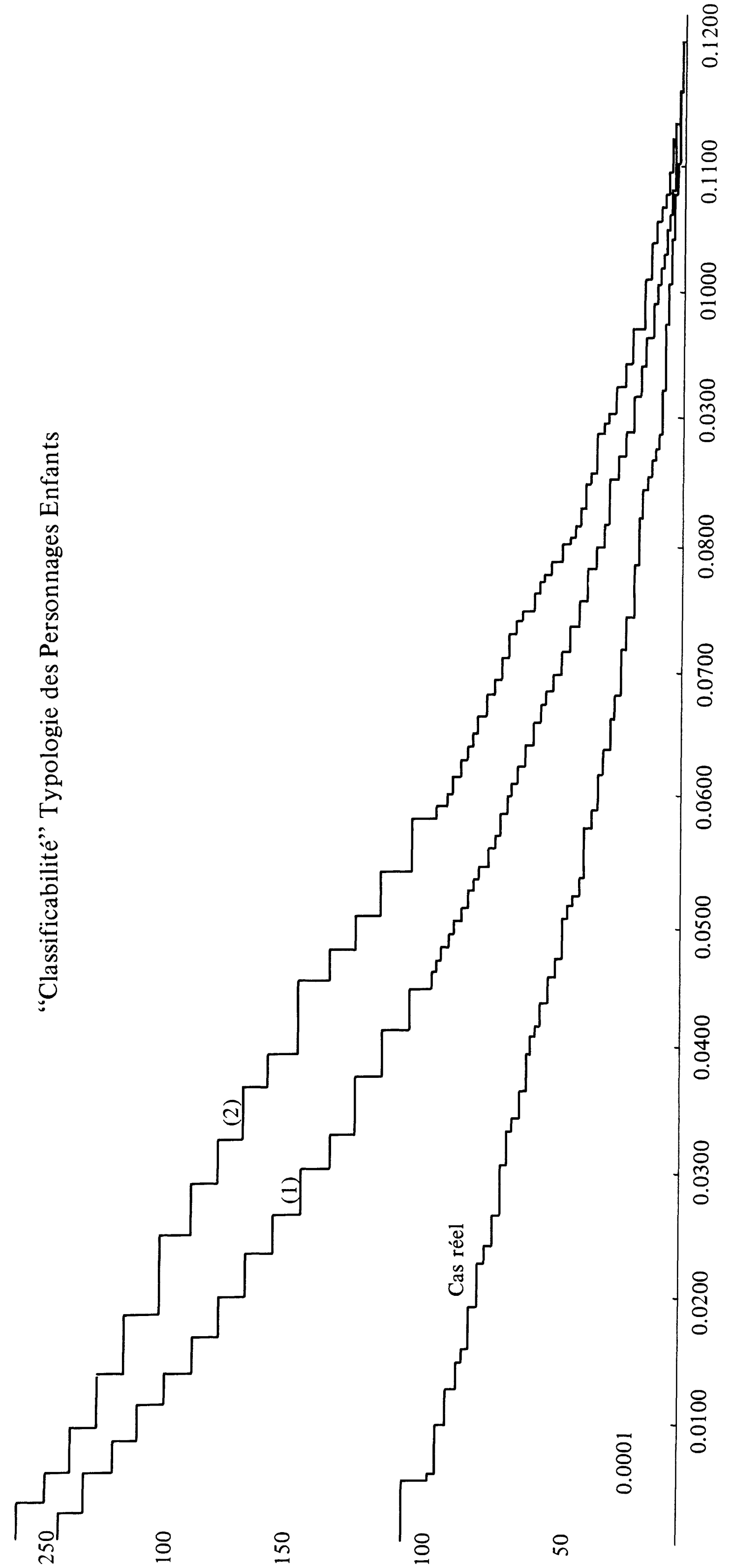
Cas simulé dans l'hypothèse N (1)



Cas simulé dans l'hypothèse N (2)



"Classificabilité" Typologie des Personnages Enfants



de 1 présents), correspond un vecteur logique de même dimension et même taille mais où les composantes égales à 1 sont également réparties. On utilise le même indice $\mathfrak{S}_1$ (a , b) dans les deux cas. La statistique locale des niveaux retenue alors n'est pas exactement θ mais est conçue à partir de la distribution dans l'hypothèse d'absence de liaison de

$$\text{card}\{gr(\omega) \cap (R' \times S)\} \qquad (\text{cf. § } 4)$$

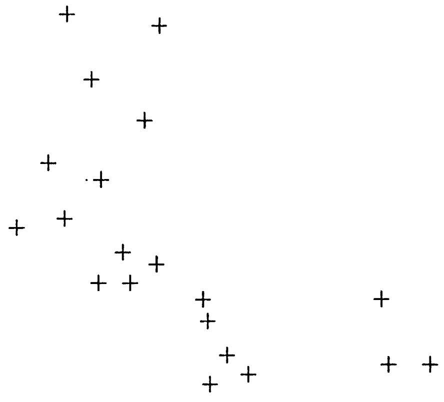
où R' est l'ensemble des paires qu'on vient d'agréger pour la première fois au niveau en question et S, l'ensemble des paires laissées séparées à ce niveau.

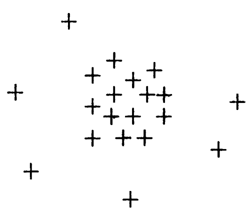
Nous avons effectué deux simulations du tableau d'incidence dans l'hypothèse N numérotées (1) et (2) ; les graphiques en bâtonnets suivants rendent compte de chacune des trois distributions. Ce qui est remarquable pour celle associée au cas réel est la persistance jusqu'au soixantième niveau des valeurs assez grandes de la statistique. Pour les cas simulés (1) et (2) les graphiques montrent une chute de la statistique vers les valeurs fortement négatives à partir du milieu de l'arbre environ ; on eût dit, par rapport au cas réel, que la positivité de la statistique "lâchait prise" à un niveau d'indice beaucoup plus petit.

On pouvait intuitivement pressentir qu'une telle population était bien classifiable ; nous avons néanmoins tenu à expérimenter la notion introduite (cf. [5] Chap. 4 et § 3 ci-dessus). Il ne faut pas être surpris de constater dans le graphique suivant que le cas (2), qui correspond à un comportement plus soutenu que dans le cas (1) de la statistique locale des niveaux, se trouve associé à la plus mauvaise classificabilité ; en effet, l'aptitude d'un ensemble de points pris au hasard à être organisé en classes n'induit pas nécessairement pour les classes formées une cohésion suffisante : condition nécessaire pour une grande valeur de la statistique globale des niveaux associée à une classification. Ainsi l'ensemble (1) des points ci-dessous est mieux apte à être organisé en classes que (2) et pourtant, la classification de (2) en 7 classes dont la première grouperait tous les points centraux et les 6 autres contiendraient chacune au point périphérique, présenterait une plus grande cohésion que toute classification de (1).

8 – EXTENSION DE LA METHODE AUX AUTRES TYPES DE VARIABLES

Il n'a été jusqu'à présent question que d'un tableau de données où n'intervient qu'un seul type de variable, l'attribut descriptif. La méthode a été étendue avec la même précision théorique aux tableaux de données où l'ensemble des variables sont d'un même type algébrique, ce qui est le cas le plus fréquant ; à ce sujet nous distinguons principalement cinq types de variables que l'on va illustrer par des exemples

(1) 

(2) 

a) **Premier type de variable : l'attribut de description**

b) **Deuxième type de variable : le caractère présentant un ensemble sans structure de modalités**

E est défini par une population de malades mentaux le caractère est défini par un diagnostic médical qui peut être *névrose, psychose non schizophrénique* et *psychose schizophrénique.* Ce type de variable définit sur E une partition ici en trois classes N, P, S où N est l'ensemble des malades névrosés, P l'ensemble des malades ayant une psychose non schizophrénique et S celui des malades ayant une psychose schizophrénique. Il est à remarquer qu'on peut ramener, sans perte d'information ce type de variable au précédent en associant à chaque modalité un attribut de description. Ici par exemple on définira un ensemble de trois propriétés : être névrosé, avoir une psychose non schizophrénique et avoir une psychose schizophrénique. Le type de variable introduit ici est intéressant pour un traitement synthétique permettant de déterminer une classification du questionnaire en chapitres – chacun d'eux formé de questions dépendantes – au niveau du comportement de la population examinée.

c) **Troisième type de variable : l'item "total" à plusieurs modalités**

E est constitué par une population d'enfants en âge scolaire et la variable située dans le cadre d'une enquête psycho-pédagogique adressée aux parents est définie par la question suivante

A la maison, parle-t-il de l'école ou du lycée

– non

– oui, avec hostilité

– oui, incidemment

– oui, avec plaisir

Cette variable définit sur E un rangement en classes, la classe des parents ayant répondu "non", celle des parents ayant répondu "oui, avec hostilité", celle des parents ayant répondu "oui, incidemment" et enfin celle des parents ayant répondu "oui, avec plaisir". Ce type de variable se distingue du précédent par le fait que les classes sont ici ordonnées – et cette information serait perdue si on remplace l'item "total" à k modalités par k attributs associés respectivement aux k modalités.

d) **Quatrième type de variable : le "rang"**

E est formé de l'ensemble des élèves candidats à une épreuve donnée de Mathématique et on dispose d'une échelle de notes d'aptitude assez fine pour qu'il ne puisse exister deux sujets ayant la même quotation. La variable "note" permet de ranger sans exeaquo les différents sujets de E. elle définit un ordre total sur E.

e) **Cinquième type de variable : la "mesure positive"**

E est un ensemble de ménages consommateurs d'une région donnée et la variable "dépense d'un ménage pour un produit donné" associe à chaque élément de E un nombre positif qui est la valeur de la dépense pour le produit en question.

Pour pouvoir étendre la méthode aux différents types de variables, il y a lieu d'établir entre deux variables d'un même type algébrique un indice de proximité qui obéit au même principe qui a permis de comparer deux attributs descriptifs. La construction de tels indices est largement développée dans [8]. Signalons que nous nous inspirons des travaux de M.G. Kendall où il compare un couple (o , o) d'ordres totaux sur un ensemble fini E ; il considère une hypothèse N d'absence de liaison entre les deux variables de type d) qui tient compte du type de structure qu'induisent les variables sur E ; en effet il se place dans E x E et étudie la distribution de la statistique

$$\sum_{(x,y)\in E\times E} a(x,y)\,b(x,y) \quad \text{où} \quad a(x,y)\ (\text{resp. } b(x,y)) = \begin{cases} 1 \text{ si } x < y \text{ pour o (resp. o}') \\ -1 \text{ si } x > y \text{ pour o (resp. o}') \end{cases}$$

lorsque o′ parcourt l'ensemble, muni d'une mesure de probabilité uniforme, de tous les ordres totaux qu'on peut définir sur E, (cf. [4]). C'est cette forme que nous donnons à l'hypothèse N d'absence de liaison pour définir de façon "statistiquement pertinente" la proximité entre deux variables établissant le même type de structure sur E. Nous retrouvons de la sorte des statistiques connues dans les cas a), d) et e) ; il s'agit du coefficient d'association de K. Pearson pour a), du τ de M.G. Kendall qu'on peut obtenir en centrant et en réduisant card $R(o) \cap R(o')$ où $R(o)$ (resp. $R(o')$) est le graphe dans $E \times E$ de o (resp. o'), pour le cas d) et d'une statistique dont la distribution, dans l'hypothèse N, est l'objet d'un théorème important dû à A. Wald et J. Wolfowitz (cf. [14]). Mais, nous trouvons des statistiques nouvelles dans les cas b) et c) où le tableau de contingence est le support de l'information ; pour b), une marge du tableau indique une partition aux classes étiquetées et pour c), un préordre total. Dans le cas b) nous nous plaçons dans l'ensemble F des paires d'objets distincts de E et nous considérons la distribution de la statistique card$\{R(\pi) \cap R(\pi')\}$ où $R(\pi)$ (resp. $R(\pi')$) est l'ensemble des paires réunies dans la partition π (resp. π') ; dans le cas c) nous nous plaçons dans $E \times E$ et considérons la statistique card$\{R(\omega) \cap R(\omega')\}$ où $R(\omega)$ (resp. $R(\omega')$) est l'ensemble des couples (x , y) de $E \times E$ pour lesquels x précède strictement y pour le préordre total ω (resp. ω'). L'hypothèse N fixe dans le cas b) l'une des deux partitions et fait varier l'autre dans l'ensemble de toutes les partitions pour lesquelles la suite des cardinaux de la marge associée du tableau de contingence reste la même ; un théorème de dualité permet d'établir que cette distribution ne dépend pas de celle des deux partitions fixée. Le même théorème de dualité permet d'établir un résultat analogue pour la comparaison de deux préordres totaux où nous montrons notamment que la statistique que propose M.G. Kendall pour étendre son τ, est biaisée. Quel que soit le type algébrique du couple de variables envisagé, nous sommes amenés à nous référer à la loi normale centrée et réduite pour la distribution dans l'hypothèse N de la statistique de proximité entre les deux variables. La notion de proximité entre deux variables est ensuite étendue à celle entre deux classes de variables de même type (cf. § 4). Pour une classification portant sur un ensemble de variables de type c) où nous disposons d'ailleurs de résultats expérimentaux très poussés, la recherche d'une échelle d'attitude relativement à une même classe présentant une grande cohésion peut se justifier ; une telle échelle permet d'ordonner totalement les différentes modalités des différents items composant la classe. Nous avons développé dans [9] un algorithme de nature combinatoire qui permet de déterminer rapidement et de façon optimale l'échelle qui s'ajuste le mieux à l'ensemble des sommets observés (dont chacun représente un individu) dans le produit cartésien des différents préordres totaux définis par les différents items de la classe. Cet ajustement se fait au sens d'une distance statistiquement pertinente.

Bien qu'il s'agit de découvrir les "dimensions" sous-jacentes au comportement d'une population il n'a nullement été question ici des méthodes d'analyse factorielle inspirées de l'algébre linéaire.

Les méthodes signalées ici sont essentiellement combinatoires et statistiques ce qui permet d'abord de rester prés du langage du spécialiste, d'où une plus grande clarté dans la signification des résultats. Compte tenu de leur souplesse ces méthodes s'adaptent aisément à n'importe quel type de données ; elles permettent par conséquent de tenir compte de très prés de l'Information initiale posée par le Questionnaire. Enfin ces méthodes permettent de traiter facilement et rapidement de très grands tableaux de données.

BIBLIOGRAPHIE

[1] ACHARD P. – "Choix d'un indice de similarité en classification automatique", Séminaire sur "Les méthodes mathématiques de l'Archéologie" (C.A.D.A. – C.N.R.S.), Marseille, 4-8 octobre 1971.

[2] BENZECRI J.P. – "Construction ascendante d'une classification hiérarchique", Lab. Stat. Math. (I.S.U.P.) ; Univ. PARIS VI, 1968-69.

[3] BENZECRI J.P. – "Théorie de l'Information et Classification d'aprés un tableau de contingence", Lab. Stat. Math. (I.S.U.P.) ; Univ. PARIS VI.

[4] KENDALL M.G. – "Rank Correlation Methods", Charles Griffin, fourth-édition, 1970.

[5] LERMAN I.C. – "Les bases de la Classification automatique", Gauthier Villars, "collection programmation", Paris, 1970.

[6] LERMAN I.C. – "Analyse des données préalable à une classification automatique", *Rev. Math. & Sc. Hum.* n° 32, 1970.

[7] LERMAN I.C. – "Analyse du phénomène de la "sériation"", *Rev. Math. & Sc. Hum.* n° 38, 1972.

[8] LERMAN I.C. – "Etude distributionnelle de statistiques de proximité entre structures finies de même type ; application à la Classification Automatique. "rap. int. Maison des Sciences de l'Homme (C.M.A.C.), rap./49/Avril 72 ; Cahiers du B.U.R.O. n° 19.

[9] LERMAN I.C. – "Analyse hiérarchique" rap. int. Maison des Sciences de l'Homme (C.M.A.C.), rap./24/Déc. 69 ; Rev. Math. & Sc. Hum. n° 17, 1967.

[10] Mme NICOLAU M.H. – "Analyse d'un algorithme de classification" ; thèse de 3e cycle, Univ. PARIS VI (I.S.U.P.), Nov. 1972.

[11] NICOLAU F. – "Contributions au traitement automatique des données multidimensionnelles par l'analyse des correspondances et la classification automatique. Etude de données sociologiques et linguistiques" thèse de 3e cycle, Univ. PARIS VI (I.S.U.P.), Nov. 1972.

[12] REGNIER S. – "Sur quelques aspects mathématiques des problèmes de la classification automatique" *I.C.C. Bull.*, Vol. 4, (1965).

[13] De la VEGA W.F. – "Techniques de classification automatique utilisant un indice de ressemblance", *Rév. Française de Sociologie,* Déc. 1967.

[14] WALD A. & WOLFOWITZ J. – "Statistical tests based on permutations of the observations", *Ann. Math. Stat.,* Vol. 15, (1944). Ce travail développé peut être consulté dans l'ouvrage de D.A.S. Fraser., "Nonparametric Methods in Statistics, John Wiley, third edition (1963).

4