Méthode d'interprétation d'une classification hiérarchique d'attributs-Modalités pour l"'explication" d'une variable ; application à la recherche de seuil critique de la tension artérielle systolique et des indicateurs de risque cardiovasculaire

Basavanepa Tallur

▶ **To cite this version:**

Basavanepa Tallur. Méthode d'interprétation d'une classification hiérarchique d'attributs-Modalités pour l"'explication" d'une variable ; application à la recherche de seuil critique de la tension artérielle systolique et des indicateurs de risque cardiovasculaire. [Rapport de recherche] RR-0123, INRIA. 1982. inria-00076437

HAL Id: inria-00076437
https://inria.hal.science/inria-00076437

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

CENTRE DE RENNES
IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP 105
78153 Le Chesnay Cedex
France
Tél.: 954 90 20

Rapports de Recherche

N° 123

MÉTHODE D'INTERPRÉTATION D'UNE CLASSIFICATION HIÉRARCHIQUE D'ATTRIBUTS-MODALITÉS POUR L' "EXPLICATION" D'UNE VARIABLE;

APPLICATION À LA RECHERCHE DE SEUIL CRITIQUE DE LA TENSION ARTÉRIELLE SYSTOLIQUE ET DES INDICATEURS DE RISQUE CARDIOVASCULAIRE

Basavanepa TALLUR

Mars 1982

METHODE D'INTERPRETATION D'UNE CLASSIFICATION HIERARCHIQUE D'ATTRIBUTS-MODALITES POUR L'"EXPLICATION" D'UNE VARIABLE ; APPLICATION A LA RECHERCHE DE SEUIL CRITIQUE DE LA TENSION ARTERIELLE SYSTOLIQUE ET DES INDICATEURS DE RISQUE CARDIOVASCULAIRE

Basavanepa TALLUR [+]

Publication interne n° 159 - Janvier 1982 - 34 pages

ésumé : A travers la classification hiérarchique par A.V.L. (Algorithme de la Vraisemlance des Liens) d'un ensemble des modalités-attributs, sur une population des consulants des Centres d'Examens de Santé, on étudie les liaisons entre la variable "à expliuer", la Tension Artérielle Systolique (TAS), et des variables "explicatives" biologiques t sociologiques. On propose une méthode permettant d'interpréter la classification hiérarchique pour expliquer une variable, de découvrir les facteurs de risques d'hypertenion artérielle et de définir le seuil critique de la TAS.

ummary :

A practical and easy-to-apply method for interpreting the hierarchical classification results obtained by applying I.C. Lerman's Likelihood Link Algorithm (L.L.A.) to a set of attribute-modalities of ordinal variables in view of "explaining" one of the variables is proposed. This method helps to find out which of the explanatory variables can be considered as "indicators" for the explained one. It makes use of the standard computer outputs and does not require any extra computations. This method is then applied to a survey in Preventive medecine whose objects were to determine the main factors of arterial hypertensive risk in a certain population, and to fix the safety limits for the systolic arterial tension.

+ IRISA Laboratoire de Statistiques, Campus de Beaulieu,
Avenue du Général Leclerc, 35042 RENNES CEDEX

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L.A. 227)
UNIVERSITÉ DE RENNES 1 I.N.S.A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

METHODE D'INTERPRETATION D'UNE CLASSIFICATION HIERARCHIQUE D'ATTRIBUTS-MODALITES POUR L'"EXPLICATION" D'UNE VARIABLE ; APPLICATION A LA RECHERCHE DE SEUIL CRITIQUE DE LA TENSION ARTERIELLE SYSTOLIQUE ET DES INDICATEURS DE RISQUE CARDIOVASCULAIRE

Principaux paragraphes

METHODE D'INTERPRETATION D'UNE CLASSIFICATION HIERARCHIQUE D'ATTRIBUTS-MODALITES POUR L'"EXPLICATION" D'UNE VARIABLE ; APPLICATION A LA RECHERCHE DE SEUIL CRITIQUE DE LA TENSION ARTERIELLE SYSTOLIQUE ET DES INDICATEURS DE RISQUE CARDIOVASCULAIRE (1)

Par B. TALLUR (2)

1. RESUME

A travers la classification hiérarchique par A.V.L. (Algorithme de la Vraisemblance des Liens) d'un ensemble des modalités-attributs, sur une population des consultants des Centres d'Examens de Santé, on étudie les liaisons entre la variable "à expliquer", la Tension Artérielle Systolique (TAS), et des variables "explicatives" biologiques et sociologiques. On propose une méthode permettant d'interpréter la classification hiérarchique pour expliquer une variable, de découvrir les facteurs de risques d'hypertension artérielle et de définir le seuil critique de la TAS.

2. INTRODUCTION

Il est fréquent que l'un des buts recherchés dans la pratique de l'analyse des données soit d'étudier comment une certaine variable, retenue d'avance, est liée à un ensemble de variables. En particulier, on cherche à "expliquer" une variable en fonction des autres variables dites "explicatives". Les

---

(1) Cette étude a été menée grâce à la Caisse Nationale d'Assurance Maladie qui l'a organisée dans le cadre d'un groupe d'études des Centres d'Examens de Santé ; et avec l'étroite collaboration de :
Dr Emile Abou, Dr Maurice Caillet, Dr Etienne Coste (CES de St-Brieuc, Rennes et Albi respectivement), Dr Bernard Dupont, M. Hubert Courcoux (Faculté de Médecine, Rennes) et Dr Louis Massé (Ecole Nationale de la Santé Publique de Rennes).

(2) IRISA, Laboratoire de Statistique, Campus de Beaulieu, Avenue du Général Leclerc, 35042 Rennes Cédex.

méthodes de régression permettent dans des situations particulières de résoudre ce problème. Nous proposons ici une méthode d'interprétation qui, en partant d'une classification hiérarchique de l'ensemble des modalités de l'ensemble de variables par l'algorithme AVL (I.C. Lerman) permet de dégager les variables expliquant celle expliquée.

Le problème nous a été posé par une étude sur l'hypertension artérielle (voir § 5) dont l'objectif a été de rechercher les facteurs responsables de risques cardiovasculaires parmi un ensemble de variables biologiques et sociologiques retenues par les médecins. Les aspects formels de la méthode proposée sont directement liés et ont été dégagés à partir de notre démarche dans l'interprétation de l'arbre de classification. Nous allons considérer deux types de données ; le premier est un tableau d'incidence (ou tableau disjonctif complet) et le second est une suite de tableaux de contingence, résultant du croisement des modalités de la variable à expliquer (Tension Artérielle Systolique, dans notre application) par l'ensemble des modalités de chacune des variables explicatives. Nous rappellerons brièvement le principe de l'A.V.L. pour ces deux types de données dans §3, et présenterons la méthode d'interprétation dans §4, et enfin, l'application à l'étude de l'H.T.A. sera exposée au §5.

3. RAPPELS SUR LA METHODE DE CLASSIFICATION PAR A.V.L.

Nous allons rappeler brièvement le principe de Classification Hiérarchique par AVL [I.C. Lerman, 1973] d'un ensemble fini D. Nous nous intéresserons, en particulier à deux types de tableau des données : (1) Le tableau d'incidence E x A où E est l'ensemble d'attributs de description ; l'ensemble D à classifier étant l'ensemble A ; (2) Le tableau de contingence K(I,J) de croisement de partitions de E définies par deux variables qualitatives dont les ensembles de modalités sont respectivement, I et J; plus généralement, une juxtaposition de tableaux de contingence $K(I,J_1 \cup J_2 \cup \ldots \cup J_L)$ de croisement de l'ensemble I des modalités d'une variable qualitative par la réunion des ensembles $J_1,\ldots,J_L$ des modalités des L variables qualitatives.

Au §3.1. nous présenterons l'expression de l'indice de proximité entre deux attributs descriptifs conforme à l'A.V.L., relativement à chacune des trois formes de l'hypothèse d'absence de lien (h.a.l.). Au §3.2. on donnera l'expression de l'indice de proximité entre les lignes (resp. colonnes) d'un tableau de contingence et son extension au cas de la juxtaposition des tableaux de contingence. Enfin, au §3.3., l'indice de proximité selon le prin-

cipe de l'A.V.L. entre deux classes d'éléments de l'ensemble D sera défini ; et quelques statistiques d'aide à l'interprétation seront présentées.

3.1. Tableau d'incidence E x A ; indice de proximité entre attributs descriptifs

Soient a et b $\in$ A, deux attributs descriptifs de l'ensemble fini E de sujets. L'attribut a(resp. b) sera représenté par la partie E_a(resp. E_b) de E formée par des sujets possédant l'attribut a(resp. b).

L'indice brut de proximité entre a et b est défini par

$$s_{ab} = \text{card}(E_a \cap E_b) \qquad (1)$$

= nombre de sujets possédant les attributs a et b simultanément

L'indice final de proximité entre a et b sera défini en situant la valeur de l'indice brut s_{ab} par rapport à la valeur attendue de la variable aléatoire $S_{ab} = \text{card}(X \cap Y)$ où X(resp. Y) est une partie aléatoire de E associée à E_a(resp. E_b) selon un modèle probabiliste à caractère uniforme et respectant la caractéristique cardinale de E_a(resp. E_b). L'indice définitif de proximité sera exprimé comme une vraisemblance, sous forme de la fonction de répartition de la v.a. S_{ab} associée à s_{ab} sous l'hypothèse N d'absence de lien (h.a.l.) (c'est-à-dire selon un modèle probabiliste définissant l'association $(E_a, E_b) \rightarrow (X,Y)$):

$$P_{ab} = \text{Prob}\left[S_{ab} \leq s_{ab}/N\right] \qquad (2)$$

En faisant l'approximation par la loi normale, on a :

$$P_{ab} = \pi\left[(s_{ab} - E(S_{ab}))/\sigma_{S_{ab}}\right] \qquad (3)$$

où π est la fonction de répartition de la loi normale N(0,1) ; $E(S_{ab})$ et $\sigma_{S_{ab}}$ sont respectivement l'espérance mathématique et l'écart-type de S_{ab} sous l'hypothèse N.

Nous allons considérer trois formes d'h.a.l. qui se distinguent dans leurs manières de respecter les caractéristiques cardinales :
$n_a = \text{card}(E_a)$, $n_b = \text{card}(E_b)$ et $n = \text{card}(E)$.

On trouve que pour chacun des trois modèles, $E(S_{ab})$ est toujours la

même :

$$E(S_{ab}) = n_a n_b/n \qquad (4)$$

3.1.1. h.a.l. N1:Modèle Hypergéométrique

En fixant E_a et en associant à E_b une partie aléatoire Y de E de cardinal n_b, ou bien en fixant E_b et en associant à E_a une partie aléatoire X de E de cardinal n_a on associe à s_{ab} deux v.a. S_a et S_b duales :

$$S_a = \text{card}(E_a \cap Y) \; ; \; S_b = \text{card}(X \cap E_b) \qquad (5)$$

suivant la même loi hypergéométrique.

En prenant $S_{ab} = S_a$ ou S_b indifféremment, on trouve

$$E_1(S_{ab}) = n_a n_b/n$$

$$V_1(S_{ab}) = n_a(n-n_a)n_b(n-n_b)/n^2(n-1) \qquad (6)$$

et l'indice centré et réduit

$$q_{ab/N_1} = \frac{s_{ab} - n_a n_b/n}{\sqrt{n_a(n-n_a)n_b(n-n_b)/n^2(n-1)}} \qquad (7)$$

3.1.2. h.a.l. N2:Modèle Binomiale

Dans le cadre de ce modèle, on associe au couple (E_a,E_b) un couple (X,Y) de parties aléatoires indépendantes de E de cardinaux n_a et n_b respectivement.

On montre que la Loi de probabilité de la variable aléatoire $S_{ab} = \text{card}(X \cap Y)$ est binomiale de paramètre $(n, p=p(a)p(b))$ avec $p(a)= \frac{n_a}{n}$ et $p(b) = \frac{n_b}{n}$.

On a ainsi

$$E_2(S_{ab}) = n_a n_b/n$$

$$V_2(S_{ab}) = \frac{n_a n_b}{n}(1-p(a)p(b)) \qquad (8)$$

et l'indice centré et réduit devient

$$q_{ab/N_2} = \frac{s_{ab} - n_a n_b/n}{\sqrt{\frac{n_a n_b}{n}(1-p(a)p(b))}} \qquad (9)$$

3.1.3. h.a.l. N3 ; Modèle Poissonien

On considère ici que E est une réalisation d'un ensemble aléatoire ε dont le cardinal est une v.a. N de paramètre n. Au tri-uple (E,E_a,E_b) où E_a et E_b sont des parties de E, on associe un tri-uple (ε,X,Y) d'ensembles aléatoires où X et Y sont deux parties aléatoires indépendantes de ε.

On trouve que la v.a. $S_{ab} = \text{card}(X \cap Y)$, dans le cadre de ce modèle, suit la loi de Poisson de paramètre np où $p = p(a)p(b)$.

D'où

$$E_3(S_{ab}) = n_a n_b/n$$

$$V_3(S_{ab}) = n_a n_b/n \qquad (10)$$

et $$q_{ab/N_3} = \frac{s_{ab} - n_a n_b/n}{\sqrt{n_a n_b/n}} \qquad (11)$$

3.2. Tableaux de contingence

Le tableau de contingence K_{IJ} est le tableau de croisement de deux partitions définies, sur l'ensemble E des sujets, par les variables qualitatives V_1 et V_2, où I(resp. J) est l'ensemble des modalités de V_1(resp. V_2), indexant les lignes (resp. colonnes) du tableau.

Soit $K_{IJ} = \{k_{ij}/i \in I,\ j \in J\}$ (12)

avec card(I) = n, card(J) = m, et

k_{ij} = nombre de sujets possédant la $i^{\text{ème}}$ modalité de V_1 et la $j^{\text{ème}}$ modalité de V_2.

3.2.1. Indice de proximité entre colonnes (resp. lignes) d'un tableau de contingence K_{IJ}

La représentation mathématique adéquate de la structure des données d'un tableau de contingence est celle, géométrique, utilisée en Analyse des Correspondances. Chaque ligne i sera représentée par son profil f_J^i dans R^n, affecté du poids $P_{i.}$ avec

$$f_J^i = \{f_j^i / j \in J\} \quad ; \quad f_j^i = \frac{f_{ij}}{Pi.} \quad \text{où}$$

$$f_{ij} = \frac{k_{ij}}{k} \; ; \; Pi. = \sum_{j \in J} k_{ij}, \text{ et } k = \sum_{i \in I} \sum_{j \in J} k_{ij} \qquad (13)$$

Le nuage des profils des lignes, N(I) est :

$$N(I) = \{f_J^i, Pi. / i \in I\} \qquad (14)$$

L'indice de proximité entre les colonnes j et k sera équivalent à l'indice de proximité entre les variables numériques X_j et X_k qui leurs sont associées respectivement. La distribution de la variable X_j (resp. X_k) à travers les lignes est définie par la suite des valeurs de la $j^{ème}$ (resp. $k^{ème}$) coordonnée des points du nuage N(I) ; le point i étant affecté de la masse Pi.

$$X_j = \{X_{ij} / i \in I\} \quad \text{où } X_{ij} = f_{ij}/Pi. \text{ et}$$

$$X_k = \{X_{ik} / i \in I\} \quad \text{où } X_{ik} = f_{ik}/Pi. \qquad (15)$$

L'indice brut de similarité entre X_j et X_k sera

$$s_{jk} = \sum_{i \in I} X_{ij} X_{ik} \qquad (16)$$

Dans le cadre d'un modèle probabiliste combinatoire, l'h.a.l. associe à l'indice s_{jk} l'une des deux v.a. duales S_j et S_k suivantes

$$S_j = \sum_i X_{ij} X_{\sigma(i)k} \; ; \; S_k = \sum_i X_{\sigma(i)j} X_{ik} \qquad (17)$$

suivant la même loi, où $\{\sigma(1),\ldots,\sigma(n)\}$ est une permutation aléatoire tirée de l'ensemble $\mathfrak{S}$ de toutes les n! permutations de $\{1,2,\ldots,n\}$, muni d'une loi de probabilité uniformément répartie.

On montre que dans ce cadre, l'indice centré et réduit :

$$q_{jk} = \left[s_{jk} - E(S_{jk})\right] / \sigma S_{jk} \qquad (18)$$

où S_{jk} est la v.a. S_j ou S_k, est au coefficient $\sqrt{n-1}$ près, le coefficient de corrélation entre les variables X_j et X_k.

D'où l'expression suivante de l'indice centré et réduit de proximité entre les colonnes j et k :

$$q_{jk} = \rho(X_j, X_k) = \frac{\sum_{i \in I} (f_{ij}\, f_{ik}/P_{i.}) - P_{.j}\, P_{.k}}{\{[\sum_{i \in I} (f_{ij}^2/P_{i.}) - P_{.j}^2]\,[\sum_{i \in I} (f_{ik}^2/P_{i.}) - P_{.k}^2]\}^{1/2}} \qquad (19)$$

Enfin, l'indice de proximité sous forme de la vraisemblance sera :

$$P_{jk} = \text{Prob}[S_{jk} \leq s_{jk}] = \pi(q_{jk}) \qquad (20)$$

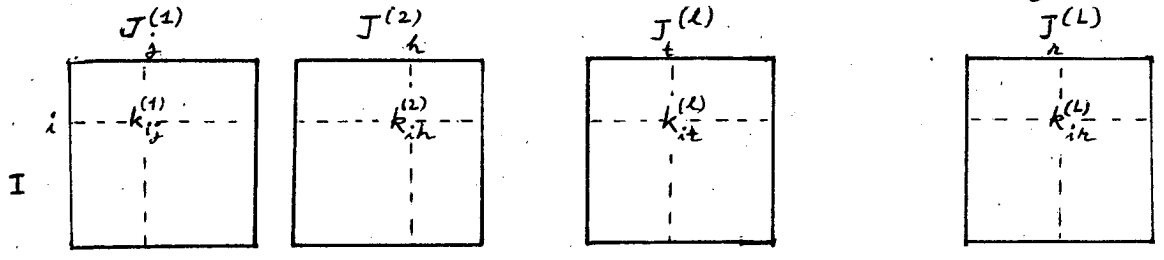
Grâce à la structure symétrique d'un tableau de contingence, on peut, en le transposant, définir l'indice de proximité entre les lignes.

3.2.2. Cas de juxtaposition des tableaux de contingence

L'indice défini au §3.2.1. a été généralisé au cas de juxtaposition des tableaux de contingence, [I.C. Lerman et B. Tallur, 1980]. Nous en rappellerons ici l'expression dans le cas particulier d'une juxtaposition horizontale de L tableaux de contingence $K_{IJ}(\ell)$, $\ell = 1,2,\ldots,L$; où

I est l'ensemble des modalités de la variable qualitative V indexant les lignes et

J(ℓ) est l'ensemble des modalités de la variable qualitative V_ℓ(ℓ=1,...,L),



Indice de Proximité entre les lignes i et i' :

En associant à la suite des tableaux de fréquence de la forme

$$\{f_{ij}^{(\ell)}/i \in I,\ j \in J^{(\ell)}\} \qquad \ell=1,\ldots,L \qquad (21)$$

avec $f_{ij}^{(\ell)} = k_{ij}^{(\ell)}/k^{(\ell)}$; $k^{(\ell)} = \sum_{i \in I} \sum_{j \in J^{(\ell)}} k_{ij}^{(\ell)}$, le tableau des pondérations de somme 1 :

$$\{f'_{ij}/i \in I,\ j \in J\} \quad \text{où } J = J^{(1)} \cup J^{(2)} \ldots \cup J^{(L)} \qquad (22)$$

avec $f'_{ij} = f_{ij}/L$ pour $j \in J$, $i \in I$ (23)

où on a noté $f_{ij} = f_{ij}^{(\ell)}$ si $j \in J^{(\ell)}$

En appliquant la formule du paragraphe précédent, l'expression de l'indice centré et réduit de proximité entre les lignes i et i' devient :

$$\rho(i,i') = \frac{\sum_{\ell=1}^{L} \sum_{j \in J^{(\ell)}} (f_{ij}\, f_{i'j}/P_{.j}) - L\, \hat{P}_{i.}\, \hat{P}_{i'.}}{\left[\{ \sum_{\ell=1}^{L} \sum_{j \in J^{(\ell)}} (f_{ij}^2/P_{.j}) - L\, \hat{P}_{i.}^2 \} \{ \sum_{\ell=1}^{L} \sum_{j \in J^{(\ell)}} (f_{i'j}^2/P_{.j}) - L\, \hat{P}_{i'.}^2 \} \right]^{1/2}} \quad (24)$$

avec $i \in I$, $\hat{P}_{i.} = \frac{1}{L} \sum_{\ell} P_{i.}^{(\ell)}$ (25)

où $P_{i.}^{(\ell)}$ est le i-ème terme de la marge colonne du tableau de fréquence (21)

Indice de Proximité entre les colonnes j et j'

Deux cas se présentent :

cas(a) j et j' appartiennent à un même $J^{(\ell)}$

On se ramène à la situation du paragraphe 3.2.1., où l'on considère l'indice (19) entre j et j' relativement au sous-tableau $K_{IJ}(\ell)$.

cas(b) $j \in J^{(\ell)}$, $j' \in J^{(\ell')}$ avec $\ell \neq \ell'$.

On considèrera la définition de l'indice par rapport au sous-tableau $K_{IJ'}$ indexé par $I \times (J^{(\ell)}, J^{(\ell')})$ qui résulte d'une juxtaposition horizontale d'exactement deux tableaux de contingence.

Aux tableaux de fréquences

$$\{f_{ij}^{(\ell)}/i \in I,\ j \in J^{(\ell)}\} \quad \text{et} \quad \{f_{ij}^{(\ell')}/i \in I,\ j \in J^{(\ell')}\} \quad (26)$$

on associe le tableau des pondérations de masse totale 1

$$\{f'_{ij}/i \in I,\ j \in J^{(\ell)} \cup J^{(\ell')}\}$$

avec $f'_{ij} = \frac{1}{2} f_{ij}$; $i \in I$, $j \in J^{(\ell)} \cup J^{(\ell')}$ (27)

Relativement au tableau $K_{IJ'}$, où $J' = J^{(\ell)} \cup J^{(\ell')}$, on a

$P'_{.j} = \frac{1}{2} P_{.j}$ pour tout $j \in J^{(\ell)} \cup J^{(\ell')}$

et $P'_{i.} = \frac{1}{2} (P^{(\ell)}_{i.} + P^{(\ell')}_{i.})$ pour tout $i \in I$ (28)

où $P^{(\ell)}_{i.}$ (resp. $P^{(\ell')}_{i.}$) est la masse de i définie au niveau du tableau $I \times J^{(\ell)}$ (resp. $I \times J^{(\ell')}$).

Dans ces conditions, l'expression de l'indice centré et réduit de proximité entre les colonnes j de $J^{(\ell)}$ et j' de $J^{(\ell')}$ se met sous la forme :

$$\rho(j,j') = \frac{\{\sum_{i \in I} (f_{ij} f_{ij'} / P'_{i.}) - P_{.j} P_{.j'}\}}{\left[\{\sum_{i \in I} (f^2_{ij} / P'_{i.}) - P^2_{.j}\}\{\sum_{i \in I} (f^2_{ij'} / P'_{i.}) - P^2_{.j'}\}\right]^{1/2}} \quad (29)$$

3.3. Critère d'agrégation par A.V.L.

La notion de proximité entre deux éléments de l'ensemble D à classifier est étendue à celle, entre deux classes A et B formées des éléments de D. Le rôle de l'indice brut de proximité entre A et B sera joué par :

$$q(A,B) = \text{Max}[P(a,b)/(a,b) \in AxB] \quad (30)$$

Sous l'h.a.l. N, définie d'une manière appropriée, on associe à q(A,B) une v.a. Q(A,B) dont la valeur observée de la fonction de répartition sera l'indice définitif de proximité entre les classes A et B :

$$P(A,B) = \text{Prob}[Q(A,B) \leq q(A,B)/N] = [q(A,B)]^{\ell m} \quad (31)$$

où ℓ= card(A) et m = card(B).

L'Algorithme A.V.L. consiste à réunir à chaque pas (niveau) la paire (ou les paires) d'éléments, ou de classes de D qui réalisent le maximum de la fonction P définie par la formule (31).

3.4. Quelques statistiques d'aide à l'interprétation

Pour contrôler l'interprétation de l'arbre de classification, on associe à chaque niveau la statistique globale $\sum$ et la statistique locale τ. Relativement à un même niveau i de l'arbre, $\sum_i$ mesure l'adéquation de la partition de l'ensemble D obtenue à ce niveau ; c'est une statistique de proximité, obéissant au principe de l'A.V.L. exposé plus haut, entre le préordre total défini par l'indice de proximité sur l'ensemble de paires d'éléments de D et la partition de D obtenue au niveau i de l'arbre.

La statistique "locale" τ_i mesure le degré de signification de l'association qui se produit au niveau i par rapport à l'ensemble des paires restant séparées à ce niveau.

Les "noeuds significatifs" correspondent aux associations qu'accompagnent des maxima locaux de τ_i.

Le degré de neutralité d'un élément d de D par rapport à une visée classificatoire est d'autant plus grand que sa variance des proximités aux autres éléments de l'ensemble D est plus petite ; la variance des proximités de $d \in D$ étant définie par la formule :

$$V(d) = \frac{1}{\text{card}(D)-1} \sum_{c \in D-\{d\}} [S(d,c)-S(d)]^2 \qquad (32)$$

où S(d) est la valeur moyenne des $[S(d,c)/c \in D-\{d\}]$ et où nous avons noté S(d,c) la valeur de l'indice centré et réduit entre les éléments d et c.

La chaîne de programme calcule et imprime systématiquement le tableau des valeurs des statistiques "globale" et "locale", signale les noeuds significatifs, condense l'arbre de classification à ces noeuds significatifs et imprime le tableau des valeurs de la variance des proximités (32) dans l'ordre croissant permettant de repérer les éléments les plus neutres rapidement.

4. METHODE D'INTERPRETATION POUR L'EXPLICATION D'UNE VARIABLE

La méthode que nous allons présenter dans ce paragraphe est directement liée aux résultats d'une classification hiérarchique par A.V.L. dont les principes viennent d'être rappelés dans les paragraphes précédents. Le but de cette méthode est "d'expliquer" une variable choisie par un ensemble de variables "explicatives". Nous allons considérer deux types de données : (1) classification de l'ensemble des attributs-modalités associés à toutes les variables

d'étude - aussi bien la variable à expliquer que des variables explicatives; (2) Classification des lignes et des colonnes du tableau de "régression" (c'est-à-dire le tableau de contingence qui croise l'ensemble des modalités de la variable à expliquer par la réunion des ensembles des modalités de toutes les variables explicatives ; il s'agit en fait d'une juxtaposition des tableaux de contingence où chaque tableau croise les modalités de la variable expliquée par des modalités de chacune des variables explicatives). Nous supposerons que les variables sont qualitatives ordinales où l'ensemble des modalités de chaque variable est totalement ordonné. C'est le cas, en particulier, des variables quantitatives découpées en classes.

Au paragraphe 4.1., nous présenterons les aspects formels de la méthode d'interprétation pour le premier type de données et au §4.2. on exposera la démarche permettant l'interprétation pour le second type de données.

4.1. Méthode d'interprétation dans le cas de classification des attributs-modalités

Nous supposerons avoir obtenu une classification hiérarchique selon l'A.V.L. sur l'ensemble A des attributs-modalités de toutes les variables ; et que nous ayons retenu la "meilleure" partition selon le critère de la "statistique globale" (voir §3.4.) de A en ℓ classes : $A = \{A_1, A_2, \ldots, A_\ell\}$. Chaque classe A_i est composée de 0 ou 1 ou plusieurs modalités des différentes variables $Y = X_1, X_2, \ldots X_m$ où on a noté Y la variable "à expliquer" et $X_2, \ldots, X_m$ les variables explicatives.

Notre démarche consiste à

(I) Pour chacune des variables $Y, X_2, \ldots, X_m$:

i) associer une modalité ou un intervalle de modalités à une même classe d'attributs A_i,

ii) définir une relation d'ordre sur l'ensemble des classes à partir de celle définie sur l'ensemble des modalités, et

(II) Comparer l'ordre défini sur $\{A_1, \ldots, A_\ell\}$ par la variable Y avec l'ordre défini par chacune des variables $X_2, \ldots, X_m$, au moyen d'une mesure de similarité.

4.1.1. Intervalle des modalités associé à une classe et ordre sur l'ensemble des classes

Pour la partie (I) de notre démarche, considérons une variable X_i ayant n_i modalités totalement ordonnées $X_{i1}, X_{i2}, \ldots, X_{in_i}$ telles que

$$X_{i1} < X_{i2} < \ldots < X_{in_i}$$

Trois cas se présentent suivant la répartition des modalités $X_{ij}\ (1 \leq j \leq n_i)$ à l'intérieur des classes $A_k\ (1 \leq k \leq \ell)$.

Cas 1. Pour chacune des classes A_k, on a l'une des propriétés suivantes :

1-a) A_k ne contient aucune modalité $X_{ij}\ (1 \leq j \leq n_i)$ de X_i

1-b) A_k contient exactement une des modalités $X_{ij}\ (1 \leq j \leq n_i)$

Il existe alors n_i classes contenant chacune exactement une modalité de X_i $(\ell \geq n_i)$. On associera, dans ce cas, la modalité X_{ij} de la variable X_i à la classe A_k qui la contient.

DEFINITION 1. On dira qu'il y a covariation entre la typologie définie par $\{A_k / 1 \leq k \leq \ell\}$ et la variable qualitative ordinale X_i ssi chacune des classes A_k contient au plus une modalité X_{ij}

$(\forall j = 1, \ldots, n_i), \exists\ k = k(j)\ ;\ X_{ij} \in A_k$ et $\{X_{ij'} / j' \neq j\} \cap A_k = \phi$

Dans ce cas, l'ordre associé à la variable X_i sur l'ensemble des classes $A_k\ (1 \leq k \leq \ell)$ sera défini par :

$$A_k(j) < A_{k(j')} \iff j < j'$$

Cas 2. Pour chacune des classes A_k, on a l'une des propriétés suivantes :

2-a) A_k ne contient aucune modalité $X_{ij}\ (1 \leq j, \leq n_i)$ de X_i

2-b) A_k contient exactement l'ensemble des modalités X_{ij} associé à un même intervalle des indices j de la forme $[g(h)+1, \ldots, g(h+1)]$.

On associe, dans ces conditions, l'intervalle des modalités $(X_{ig(h)+1}, \ldots, X_{ig(h+1)})$ de la variable qualitative X_i à la classe A_k des attributs qui le contient.

DEFINITION 2. Si on désigne par $A_{k(h)}$ la classe des attributs contenant exactement l'ensemble des modalités X_{ij} associé à un même intervalle des indices j de la forme $g(h)+1 \leq j \leq g(h+1)$ où g(h) sont des entiers positifs avec $g(1)=0$ et $\sum g(h)=n_i$, l'ordre associé à la variable X_i sur l'ensemble des classes $\{A_1^h,\ldots A_\ell\}$ sera défini par :

$$A_{k(h)} < A_{k(h')} \iff h<h'$$

Remarque : Cas 1 est un tas particulier du Cas 2 où chacun des intervalles des modalités X_{ij} se réduit à une seule modalité de sorte que $g(h)=1$ pour $h \neq 1$.

Cas 3. Pour chacune des classes A_k, on a l'une des propriétés suivantes :

3-a) A_k ne contient aucune modalité X_{ij} $(1 \leq j \leq n_i)$ de la variable X_i

3-b) A_k contient l'ensemble des modalités X_{ij} associé au moins à deux des intervalles non connexes des indices j de la forme $[g(h)+1,\ldots g(h+1)]$, $[g(h')+1,\ldots,g(h'+1)]$ $[g(h''+1,\ldots,g(h''+1)]$...où $h' \neq h+1$, $h'' \neq h'+1$,... et $h<h'<h''<\ldots$ etc ; un ou plusieurs d'entre ces intervalles pouvant être réduits à un seul indice.

On associera, alors, la modalité ou l'un des intervalles des modalités

$$[X_{ig(h)+1},\ldots,X_{ig(h+1)}],\ [X_{ig(h')+1},\ldots,X_{ig(h'+1)}]\ \ldots.$$

contenant la modalité la moins "neutre" selon le critère de sa variance des proximités dont nous avons fait mention au §3.4.

La démarche où l'on associe non pas la modalité la moins neutre, mais l'intervalle des modalités auquel appartient cette dernière est justifiée par le fait que si l'association d'une modalité isolée à une classe peut être dûe "au hasard", celle d'une suite des modalités contigües à une même classe ne peut pas l'être.

DEFINITION 3. Si pour chacune des classes A_k, l'une des propriétés 3-a, 3-b est vérifiée, et en désignant par $A_{k(h)}$ la classe des attributs à laquelle est associé l'intervalle des modalités $[X_{ig(h)+1},\ldots,X_{ig(h+1)}]$, l'ordre sur l'ensemble des classes $\{A_1,\ldots,A_\ell\}$ associé à la variable X_i est défini par :

$$A_{k(h)} < A_{k(h')} \iff h<h'$$

4.1.2. Comparaison des ordres sur $\{A_k/1\leq k\leq \ell\}$ et recherche des indicateurs de la variable à expliquer

La deuxième partie (II) de notre démarche consiste à comparer l'ordre défini sur l'ensemble des classes $\{A_k/1\leq k\leq \ell\}$ par la variable "à expliquer" Y avec celui défini sur ce même ensemble par chacune des variables explicatives $X_2,\ldots,X_m$ au moyen d'un indice de proximité permettant de déterminer l'intensité de la liaison entre les variables associées. Dans notre application, la variable à expliquer est la T.A.S. et parmi les variables explicatives se trouvent notamment Taux de Cholestérol, Taux de Triglycérides, Gamma GT, Surcharge Pondéral...etc.

DEFINITION 4. La variable $X_i (i\neq 1)$ sera appelée "indicateur" de la variable Y si l'ordre induit sur l'ensemble $\{A_k/1\leq k\leq \ell\}$ des classes d'attributs par la variable X_i est, soit exactement identique, soit tout à fait opposé à celui défini par la variable Y. Dans le premier cas X_i est un "indicateur positif", et dans le second c'est un "indicateur négatif".

Remarque : La relation d'ordre définie par chacune des variables $Y,X_2,\ldots,X_m$ n'est pas nécessairement totale ; cela pourrait être formée d'une chaîne totale et des singletons. Dans le cas où l'ordre défini par la variable Y est partiel, on retiendra la restriction de chacun des ordres définis par les variables $X_2,\ldots,X_m$ à l'ensemble des paires de classes comparables pour Y.

Pour les variables explicatives qui ne sont ni "indicateurs positifs" ni "indicateurs négatifs", on pourra mesurer leur degré d'association avec Y par un indice de proximité entre préordres totaux, proposé par I.C. Lerman dans le cadre de l'A.V.L., et dont les principes et l'expression sont rappelés ci-dessus.

Indice de Proximité entre les préordres totaux

Chaque variable à ensemble totalement ordonné des modalités définit, sur l'ensemble E des individus, une partition et un ordre total sur l'ensemble des classes.

On désignera par ω le préordre total à k classes associé à une variable à k modalités et par $\nu_1,\nu_2,\ldots,\nu_k$ la suite des cardinaux des classes $E_1,E_2,\ldots,E_k$ rangée selon l'ordre quotient ω. Mathématiquement ω sera représenté par son graphe R_ω dans l'ensemble E_2 des couples d'éléments distincts de E : $R_\omega = \{(x,y)/x<y$ et non $y<x$ pour $\omega\}$

Indice brut de comparaison entre deux préordres ω et ω' à k et h classes respectivement sera le suivant

$$s_{\omega\omega'} = \mathrm{card}(R_\omega \cap R_{\omega'}) = \sum_{i=1}^{k-1} \sum_{j=1}^{h-1} \nu_{ij} \sum_{p>i} \sum_{q>j} \nu_{pq} \qquad (33)$$

où ν_{ij} est le cardinal de l'intersection de la i-ème classe de ω et de la j-ème classe de ω'. L'indice de proximité définitif $p_{\omega\omega'}$ entre ω et ω' selon le principe de l'A.V.L. sera le suivant :

$$p_{\omega\omega'} = \mathrm{Prob}\left[S_{\omega\omega'} < s_{\omega\omega'}/N\right] \qquad (34)$$

où $S_{\omega\omega'}$ est la v.a. associée à $s_{\omega\omega'}$ dans le cadre d'une hypothèse d'absence de lien (h.a.l.) N. Par une approximation à la loi normale, on a :

$$p_{\omega\omega'} = \pi\left[q_{\omega\omega'}\right]$$

où π est la f.r. de la loi normale N(0,1) et $q_{\omega\omega'} = \left[s_{\omega\omega'} - E(S_{\omega\omega'}/N)\right] / \sqrt{\mathrm{Var}(S_{\omega\omega'}/N)}$ (35)

I.C. Lerman [4] a clairement défini l'h.a.l. N et a étudié la distribution de $S_{\omega\omega'}$ sous N ; il montre que

$$E(S_{\omega\omega'}/N) = \lambda\mu \text{ et } \mathrm{Var}(S_{\omega\omega'}/N) = \lambda\mu + \rho_{ff}\sigma_{ff} + \rho_{cc}\sigma_{cc} + 2\rho_{cf}\sigma_{cf} + (\wedge M - \lambda^2 M^2)$$

avec $\lambda = \sum_{i<j} \nu_i\nu_j / \sqrt{n(n-1)}$, $\rho_{ff} = \sum_i \nu_i (\nu_i^f)^2 / \sqrt{n(n-1)(n-2)}$

$$\rho_{cc} = \sum_i \nu_i (\nu_i^c)^2 / \sqrt{n(n-1)(n-2)}, \quad \rho_{cf} = \sum_i \nu_i \nu_i^f \nu_i^c / \sqrt{n(n-1)(n-2)} \qquad (36)$$

$$\nu_i^f = \sum_{j>i} n_j - 1 \text{ et } \nu_i^c = \sum_{j<i} \nu_j - 1$$

$$\wedge = \sum_{i<j} \nu_i\nu_j \Big(\sum_{i'<j'} \nu_i\nu_j + \nu_i + \nu_j - 2n + 1 \Big) / \sqrt{n(n-1)(n-2)(n-3)}$$

$\mu, \sigma_{ff}, \sigma_{cc}, \sigma_{cf}$ et M ayant la même forme que $\lambda, \rho_{ff}, \rho_{cc}, \rho_{cf}$ et $\wedge$ respectivement où les ν_i $1 \leq i \leq k$ sont remplacés par les ν_i' $1 \leq i \leq h$, cardinaux des classes $E_1', \ldots, E_h'$ rangées selon l'ordre ω'.

4.2. Méthode d'interprétation pour l'explication d'une variable dans le cas de classification des lignes et des colonnes d'une juxtaposition des tableaux de contingence

Nous supposons avoir obtenu une classification hiérarchique sur l'ensemble des lignes I, indexé par les modalités de la variable "à expliquer" Y, ainsi qu'une deuxième classification hiérarchique sur l'ensemble des colonnes J indexé par la réunion des ensembles des modalités des variables explicatives $X_2,\ldots,X_m$; et que les "meilleures partitions" retenues selon le critère de la "statistique globale" sont $I = \{I_1,I_2,\ldots,I_p\}$ et $J = \{J_1,J_2,\ldots,J_q\}$ respectivement en p et q classes.

Etude de liaison entre Y et les X_i(i=2,..,m) à travers les deux classifications, ci-dessus, sera basée sur le tableau C défini ci-dessous de leur croisement.

DEFINITION 5. Etant donné les classifications suivantes $I = \{I_1,\ldots,I_p\}$ et $J = \{J_1,\ldots,J_q\}$ de l'ensemble des lignes et de l'ensemble des colonnes de tableau de contingence, le tableau de croisement C associé à ces deux classifications est le suivant :

$$C = \{Crs/1\leq r\leq p,\ 1\leq s\leq q\}$$

$$\text{où } Crs = \sum_{i\in I_r}\ \sum_{j\in J_s} k_{ij} \qquad (37)$$

Le tableau de croisement permet de calculer les mesures d'association pour chaque couple (I_r,I_s) $(1\leq r\leq p\ ;\ 1\leq s\leq q)$ de classes ainsi que les "profils moyens" de chaque classe I_r(resp. J_s) à travers les classes $J_s, 1\leq s\leq q$ (resp. I_r, $1\leq r\leq p$) ; les deux calculs conduisant à des interprétations, en général concordantes, des liaisons entre les lignes et les colonnes du tableau K_{IJ}.

4.2.1. Mesures d'association entre une classe I_r et une classe I_s

Une mesure globale d'association entre deux variables, où chacune des variables définit une partition sur l'ensemble des individus, est donnée par la statistique du χ^2 associée au tableau de contingence du croisement des deux partitions, c'est-à-dire, au tableau C défini ci-dessus. La valeur du χ^2 est d'autant plus élevée que le degré d'association entre les variables est plus fort. Dans l'hypothèse d'indépendance des variables, la probabilité de trou-

ver une valeur de la v.a. χ^2 à $(p-1)(q-1)$ d.d.l. inférieure ou égale à celle observée constitue une bonne mesure globale d'association dont la valeur est comprise entre 0 et 1. La contribution à la statistique du χ^2 du couple (I_r,J_s) de classes, notée χ^2_{rs}, sera d'autant plus forte que la classe I_r est plus particulièrement associée à la classe J_s :

$$\chi^2 = \frac{(C_{rs} - \frac{C_r.C._s}{C..})^2}{(C_r.C._s/C..)} \qquad \text{avec } C_{r.} = \sum_s C_{rs},\ C._s = \sum_r C_{rs} \qquad (38)$$

$$\text{et } C.. = \sum_r \sum_s C_{rs}$$

En associant à chaque case (r,s) du tableau C une v.a. de χ^2 à 1d.d.l. on peut mesurer l'association entre les classes I_r et J_s par la valeur observée de sa fonction de répartition :

$$p_{rs} = \text{Prob}[\chi^2_{rs} \leq \chi^2_1]$$

Mais l'inconvénient de ces mesures est qu'elles ne nous renseignent pas sur le sens des associations ; en effet ces mesures ne permettent pas de voir si l'association du couple (r,s) est négative ou positive. Pour remédier à cet inconvénient, nous calculerons en plus des p_{rs} les mesures orientées d'associations notées χ_{rs} pour chaque couple (r,s) :

$$\chi_{rs} = (C_{rs} - \frac{C_r.C._s}{C..}) \Big/ \sqrt{\frac{C_r.C._s}{C..}} \qquad (39)$$

Tandis que la valeur de p_{rs} mesure l'intensité de lien du couple (I_r,J_s), le signe de χ_{rs} en indique le sens.

Nous signalons que les mesures orientées d'associations χ_{rs} sont proposées par I.C. Lerman [5] dans les différents cas de croisement de classifications.

4.2.2. Profil moyen pondéré de la classe I_r (resp. J_s) à travers les classes $J_s, 1 \leq s \leq q$ (resp. $I_r, 1 \leq r \leq p$)

Afin d'interpréter les liaisons entre les couples (I_r,J_s), nous calculerons les profils moyens pondérés des classes I_r à travers l'ensemble des classes J_s, $1 \leq s \leq q$.

DEFINITION 6. Le profil moyen pondéré de la classe I_r à travers l'ensemble J est le centre de gravité du sous-nuage $N_r(I)$ associé au tableau de contingence $K_{I_rJ} = \{k_{ij}/i \in I_r,\ j \in J\}$:

$$N_r(I) = \{(\frac{f_{ij}^{(r)}}{p_{i.}^{(r)}},\ p_{i.}^{(r)})/i \in I_r\} \qquad (40)$$

où $\{f_{ij}^{(r)}/i \in I_r,\ j \in J\}$ est une loi de probabilité définie sur $I_r \times J$ avec

$$f_{ij}^{(r)} = k_{ij}/k_r \quad i \in I_r,\ j \in J\ ;\ k_r = \sum_{i \in I_r} \sum_{j \in J} k_{ij} \text{ et } p_{i.}^{(r)} = \sum_{j \in J} f_{ij}^{(r)} \qquad (41)$$

Le centre de gravité du nuage $N_r(I)$ est :

$$\{p_{.j}^{(r)}/j \in J\} \qquad \text{avec } p_{.j}^{(r)} = \sum_{i \in I_r} f_{ij}^{(r)} \qquad (42)$$

En réunissant les modalités j appartenant à une même classe J_s, on obtient le profil moyen de la classe I_r à travers les classes $J_s, 1 \leq s \leq q$:

$$\{p_{.J_s}^{(r)}/1 \leq s \leq q\} \qquad \text{avec } p_{.J_s}^{(r)} = \sum_{j \in J_s} p_{.j}^{(r)} \qquad (43)$$

Analyse Factorielle des Correspondances du tableau C

A partir de la construction du tableau de croisement C des classifications, on peut noter que le profil moyen de la classe I_r à travers les classes J_s, $1 \leq s \leq q$ est le profil de la r-ème ligne associé au tableau C ; et que l'analyse des correspondances sur ce tableau permet de résumer toutes les liaisons entre les couples (I_r, J_s) des classes.

5. APPLICATION A L'ETUDE DE L'HYPERTENSION ARTERIELLE

5.1. Introduction

Plusieurs milliers de "bilans de santé" sont réalisés chaque année dans chacun des Centres d'Examens de Santé (C.E.S.) en France, dont les objectifs sont la prévention et le dépistage à temps des maladies. L'Hyper Tension Artérielle (H.T.A.) est considérée comme un des facteurs essentiels du risque cardiovasculaire ; et les maladies cardiovasculaires sont les causes du plus grand nombre de décès. D'où la préoccupation considérable des C.E.S. par la

prévention de ces maladies. Nombreuses études ayant démontré l'existence de la relation entre certaines variables biologiques et les valeurs tensionnelles, il est naturel de chercher les indicateurs de risques cardiovasculaires parmi les facteurs biologiques liés à l'H.T.A.

Le "bilan" consiste en un questionnaire détaillé rempli par chaque sujet sur ses conditions socio-professionnelles d'une part, et en une suite d'examens cliniques et de tests biologiques, d'autre part. Les résultats de toutes les analyses sont suivies de la conclusion du médecin qui reconvoquera le sujet en cas d'anomalie quelconque. L'exploitation de ces données précieuses se limite en général à l'étude de l'association des variables 2 à 2 par le test du χ^2, le coefficient de corrélation...etc.

Le présent travail est la suite logique d'une étude menée en 1979 sur l'ensemble des sujets examinés au C.E.S. de Rennes [7], dans le même but de la recherche des indicateurs du risque cardiovasculaire. Il a été confirmé le rôle de certains indicateurs connus tels que le taux du cholestérol, le taux des triglycérides et l'indice pondéral. D'autre part, le rôle du tabac ne nous avait pas paru évident ; les facteurs socio-professionnels ayant une liaison avec la tension varient avec l'âge et le sexe.

La présente étude porte sur une population élargie aux quatre centres: Albi, Nice, Rennes et St-Brieuc. Parmi tous les consultants de l'année 1979 ont été retenus 10.693 sujets non-médicalisés ; c'est-à-dire les sujets soumis à un traitement médical après un dépistage d'anomalie cardiovasculaire ont été éliminés. Le but en est de définir le seuil critique de la T.A.S. chez les sujets sains, au delà duquel on peut craindre les risques cardiovasculaires, ainsi que d'en dégager des facteurs ou indicateurs de risques.

5.2. Population, variables retenues et leur codage

La population de 10.693 sujets examinés est divisée en quatre sous-populations suivant le sexe et les tranches d'âge : 30 à 39 ans et 40 à 49 ans. Chaque sous-population a été étudiée séparément et les résultats sont comparés. Les effectifs de ces sous-populations sont résumés dans le tableau suivant.

HOMMES		FEMMES	
30 à 39 ans	40 à 49 ans	30 à 39 ans	40 à 49 ans
2.924	2.531	2.855	2.383

D'après les études antérieures, et en fonction des objectifs fixés, les spécialistes ont retenu un ensemble de variables relatives à la situation socio-professionnelle et au mode de vie familiale du sujet d'une part, et un ensemble de variables biologiques, d'autre part.

a. Variables sociologiques

C'est un ensemble de variables descriptives, toutes qualitatives à l'exception de "Quantité de Tabac" et "la durée de tabagisme" qui sont découpées en un certain nombre de modalités. Chaque modalité définissant un attribut de description, les variables sont codées en présence (=1) et absence (=0). Les variables suivantes sont retenues :

1. Situation de famille (7 modalités)
2. Catégorie socio-professionnelle (9 modalités)
3. Horaire de travail (7 modalités)
4. Type d'habitat (9 modalités)
5. Mode d'alimentation (4 modalités)
6. Catégorie de fumeur (4 modalités)
7. Durée de tabagisme (4 modalités)
8. Quantité cumulée de tabac (5 modalités)
9. Consommation d'alcool (5 modalités).

b. Variables biologiques

Elles sont toutes quantitatives et continues, mais afin d'obtenir des données du même type, elles ont été découpées en classes (après les études préliminaires), et ensuite codées en tant que les attributs de description. Les variables biologiques suivantes sont retenues :

1. Tension Artérielle Systolique (8 modalités)
2. Taux de glycémie en g/l (7 modalités)
3. Taux de cholestérol en g/l (7 modalités)
4. Taux de l'Acide Urique en g/l (7 modalités)
5. Gamma G.T. (7 modalités)
6. Taux des Triglycérides en g/l (8 modalités)
7. Volume Globulaire Moyen (V.G.M.) (7 modalités)
8. Taille (4 modalités)
10. Indice de Quetelet (c'est le rapport du poids en kg sur le carré de la taille en mètres (8 modalités).

Pour chaque sous-population, un tableau d'incidence (ou tableau disjonctif complet) croisant l'ensemble des sujets avec 118 modalités des variables est analysé. Nous avons également analysé les tableaux de contingence croisant les modalités de la variable T.A.S. avec celles de toutes les autres variables. La méthode utilisée est celle de la classification ascendante hiérarchique par l'A.V.L. (voir §3).

Nous exposerons en détail l'application de notre méthode d'interprétation aux résultats de la classification obtenue pour chacun des deux types de tableaux pour la population des consultantes âgées de 30 à 39 ans, dans les paragraphes suivants.

5.3. Recherche des facteurs de risque liés à la T.A.S. ; Analyse du tableau d'incidence

La classification par A.V.L. sur l'ensemble de 118 modalités pour la population des femmes de 30 à 39 ans a permis d'obtenir une partition en quatre grandes classes notées A_1, A_2, A_3 et A_4 au 89-ème niveau de l'arbre où le maximum de la "statistique globale" a été atteint. La variable à expliquer est, pour nous, la T.A.S. Pour la recherche des "indicateurs" de la T.A.S. qui constituent les facteurs de risque de l'hypertension, nous allons :

(i) associer à chacune des classes A_1, A_2, A_3, A_4 la modalité ou l'intervalle des modalités de chacune des variables explicatives,

(ii) définir l'ordre sur l'ensemble des classes $\{A_1,A_2,A_3,A_4\}$ selon la T.A.S. et selon chacune des variables explicatives dont les modalités sont totalement ordonnées, et

(iii) rechercher, parmi les variables explicatives, celles ayant défini l'ordre sur $\{A_1,A_2,A_3,A_4\}$ dont la restriction à l'ensemble des couples de classes comparables pour la T.A.S., est identique à celui défini par la T.A.S.

On voit que (voir, en annexe, l'arbre de classification) :

. A_1 ne contient aucune modalité de T.A.S.
. A_2 contient "TAS<10" et "TAS 10 à 12,9" ; on lui associe l'intervalle "TAS≤12,9".
. A_3 contient la modalité "TAS 14 à 14,9" qui lui est associée
. A_4 contient les modalités "TAS 13 à 13,9", "TAS 15 à 15,9", "TAS 16 à 16,9", "TAS 17 à 17,9" et "TAS≥18". Il s'agit d'associer soit la modalité

"TAS 13 à 13,9", soit l'intervalle "TAS≥15". Selon le critère du degré de neutralité, c'est l'intervalle "TAS≥15" qui est associé à A_4.

L'ordre défini sur l'ensemble des classes par la TAS est le suivant :

$A_2 < A_3 < A_4$

A_1 n'étant pas comparable aux autres classes. Le tableau 1 résume le résultat de ces démarches pour quelques unes des variables ordinales les plus liées à la TAS.

5.3.1. Remarques sur et interprétation des résultats

On constate que Taux de Cholestérol, Taux de Triglycérides, Taux de Glycémie et Indice de Quetelet sont des "indicateurs positifs" de la T.A.S. ; en effet chacune de ces variables ordonne les classes A_2, A_3 et A_4 de la même manière:

$A_2 < A_3 < A_4$

On remarque que Gamma G.T., bien que n'étant pas un indicateur de la TAS définit un ordre partiel sur l'ensemble des classes ($A_2 < A_4$ pour Gamma G.T.) qui a une association positive avec la T.A.S.

5.3.2. Mélange des variables explicatives, ordinales et non-ordinales

La présence des variables à l'ensemble des modalités non-ordonné n'affecte pas la méthode de la recherche des indicateurs ; on peut même envisager d'ordonner l'ensemble des modalités d'une telle variable, en affectant à chaque modalité le rang de la classe -défini par la TAS- qui la contient. Ainsi, pour la variable "situation de famille", les modalités sont réparties sur l'ensemble des classes de la façon suivante :

A_1 : célibataire, divorcée, en concubinage
A_2 : veuve
A_3 : mariée
A_4 : veuve remariée, divorcée remariée

Etant donné que $A_2 < A_3 < A_4$ pour la T.A.S., on peut ordonner les modalités de "situation de famille" par rapport à la T.A.S. comme suit :

Veuve<Mariée <(Veuve remariée = Divorcée remariée)

Variables	T.A.S.			Taux de cholestérol			Triglycérides			Glycémie			Indice de Quetelet			Gamma G.T.		
Classes	Modalité associée	Rang	Restriction du rang	Modalité associée	Rang	Restriction du rang	Modalité associée	Rang	Restriction du rang	Modalité associée	Rang	Restriction du rang	Modalité associée	Rang	Restriction du rang	Modalité associée	Rang	Restriction du rang
A_1	aucune	-		≤0,99	1		aucune	-		≤0,59	1		21-23	2		aucune	-	
A_2	≤12,9	1		1-1,99	2	1	≤0,49	1	1	0,60-0,79	2	1	15-21	1	1	≤19	1	1
A_3	14-14,9	2		2-2,49	3	2	0,50-0,99	2	2	0,80 à 0,99	3	2	23-25	3	2	aucune	-	-
A_4	≥15	3		≥2,50	4	3	1-3,59	3	3	1,00 à 1,19	4	3	≥25	4	3	≥20	2	2

TABLEAU 1 : Résultats des démarches conduisant à la découverte des indicateurs de la T.A.S.

La classe A_1 se joignant à A_2 au niveau 90 de l'arbre, on pourra éventuellement définir trois classes des modalités ordonnées :

(Célibataire=divorcée=concubinage=veuve)<Mariée

<(Veuve remariée = divorcée remariée)

5.3.3. Interprétation des classes

La classe A_1 regroupe toutes les modalités concernant les fumeuses, telles que "fume depuis 5 ans", "inhale la fumée", "quantité de tabac<10 kg",...etc. Cette classe ne contient pas de modalités de TAS. Il y a par contre quelques faibles valeurs des variables biologiques telles que cholestérol et glycémie. Il semble donc que le fait de fumer n'influence pas sur la TAS, et que certain nombre de paramètres biologiques sont faibles chez les fumeuses. L'interprétation dynamique de l'arbre aux niveaux supérieurs montre que (voir figure 1) cette classe se réunit avec la classe A_2 qui est caractérisée par les valeurs moyennes de la TAS ainsi que celles de tous les paramètres biologiques. Il s'agit donc d'une classe "normale". Remarquons que A_2 contient "non-fumeuses".

A_3 et A_4, qui se réunissent d'ailleurs au niveau 91, sont composées respectivement des valeurs élevées et très élevées de toutes les variables biologiques ainsi que celle de la T.A.S. La partition au niveau 91 en deux classes sépare toutes les valeurs de TAS inférieures à 13 de celles qui sont ≥13 ; ceci montre qu'il existe un seuil entre les valeurs faibles ou "normales" et les valeurs fortes de la TAS et que ce seuil se situe autour de 13. Pour la médecine préventive, la connaissance de ce seuil est très important, car elle permet de mieux surveiller les sujets au delà de cette valeur accompagnés d'autres symptômes indicateurs de risques.

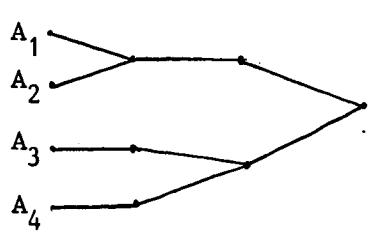


Figure 1

5.4. Recherche des facteurs de risques liés à la T.A.S. ; Analyse du Tableau de contingence

Le tableau de dépendance analysé croise les huit modalités de la variable T.A.S. par 110 modalités de l'ensemble des variables explicatives. L'application de l'indice de proximité défini au §5.1. et l'algorithme A.V.L. a permis d'obtenir une classification hiérarchique des lignes (c'est-à-dire des modalités de T.A.S.) et une autre sur les colonnes (c'est-à-dire des modalités des variables explicatives).

5.4.1. Classification des modalités de la T.A.S.

On obtient l'arbre détaillé de classification hiérarchique suivant :

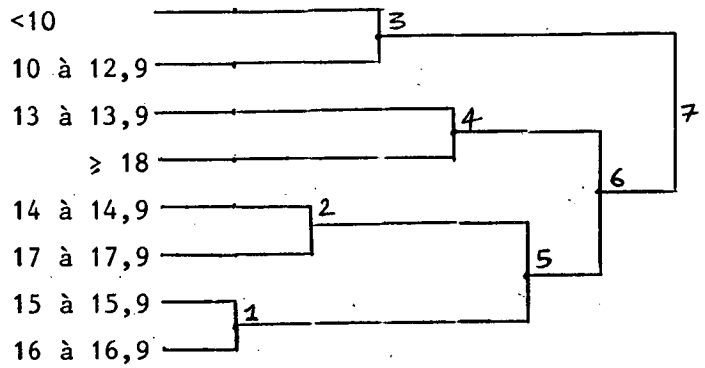


Figure 2. Arbre détaillé de la classification des modalités de T.A.S.

Interprétation

La modalité TAS≥18 étant très "neutre", on peut retenir une classification en trois classes produite au niveau 5 : 1) Faible ou moyenne (<10,10 à 12,9), 2) Assez forte (13 à 13,9), et 3) Très forte (14 à 14,9 jusqu'à 17 à 17,9). Au niveau 6, la partition en deux classes différencie nettement les valeurs de la T.A.S. inférieures à 13 de celles supérieures à 13.

5.4.2. Classification des 110 modalités des variables biologiques et sociales explicatives et de croisement des classifications

La partition optimale au sens du critère de la "statistique globale" est obtenue au niveau 89. Elle est constituée de 9 classes. Chaque classe étant composée des modalités des variables, aussi bien sociologiques que biologiques. Avant d'examiner leurs contenus, il est instructif et utile de construire le tableau de croisement des deux classifications et de calculer les indices d'association, afin de découvrir les associations entre la T.A.S. et d'autres variables. En désignant les trois classes de la T.A.S. par Y_1, Y_2, Y_3, et les

neuf classes des variables explicatives par $X_1, X_2, \ldots, X_g$, on a le tableau de croisement C suivant :

X \ Y	Y_1	Y_2	Y_3
X_1	1607	430	144
X_2	1499	406	197
X_3	415	158	46
X_4	1857	476	156
X_5	8514	2026	781
X_6	11541	3317	1524
X_7	2816	789	318
X_8	4479	1192	518
X_9	2243	670	344

Tableau 2. Tableau C de croisement des classifications

Le tableau 3 ci-dessous résume les indices d'association p_{rs}, les indices orientés x_{rs} ainsi que les contributions CT_{rs} en pourcentage de chaque classe Y_s à une statistique du χ^2 exprimée par une même classe X_r :

$$CT_{rs} = (\chi^2_{rs} / \sum_r \chi^2_{rs}) \times 100$$

	Y_1			Y_2			Y_3		
	X_{rs}	CT_{rs}	p_{rs}	X_{rs}	CT_{rs}	p_{rs}	X_{rs}	CT_{rs}	p_{rs}
X_1	0,8	8,11	0,5	0,2	0,57	0,2	-2,8	91,32	0,99
X_2	-0,4	7,10	0,3	-0,2	1,25	0,2	1,6	91,65	0,9
X_3	-1,5	15,86	0,9	3,4	80,20	>0,999	-0,8	3,94	0,5
X_4	1,5	14,92	0,9	-0,6	2,23	0,5	-3,6	82,85	>0,999
X_5	3,8	25,40	>0,999	-3,9	26,95	>0,999	-5,2	47,65	>0,999
X_6	-2,6	21,99	0,99	2,1	14,64	0,98	4,4	63,37	>0,999
X_7	-0,3	9,02	0,2	0,9	72,98	0,7	-0,4	18,00	0,2

X_8	0,2	10,79	0,2	-0,4	75,81	0,3	0,2	13,40	0,2
X_9	-2,2	18,50	0,98	1,4	6,99	0,8	4,5	74,51	>0,999

Tableau 3. Indices d'association attachés au tableau de croisement des classifications.

Interprétation du Tableau 3.

1. Le signe des indices α_{rs} permet d'identifier le sens positif ou négatif des associations entre les couples (X_r, Y_s) des classes.

2. La valeur de χ^2_{rs} pouvant être considérée comme une réalisation de la v.a. χ^2 à 1 d.d.l., on peut repérer les associations significatives. Par exemple, si χ^2_{rs} > 3,84 on pourra considérer que l'association entre X_r et Y_s est significative avec un risque d'erreur α=0.05.

Le tableau 4 montre les associations significatives avec leur sens.

	Y_1	Y_2	Y_3
X_1			-
X_2			
X_3		+	
X_4			-
X_5	+	-	-
X_6	-	+	+
X_7			
X_8			
X_9	-		+

Tableau 4. Signification des associations entre classes.

Observations

. Les classes X_2, X_7 et X_8 ne sont pas significatives (c'est-à-dire leurs associations avec Y_1,Y_2 et Y_3 ne sont pas significatives).

. X_1 et X_4 sont négativement associées à Y_3 (TAS très élevée).

. X_3 est positivement associée avec Y_2 (TAS assez forte)

. X_5 et X_6 sont opposées ; X_5 étant positivement associée à la TAS faible et négativement associée avec la TAS forte et très forte, alors que X_6 a des associations tout à fait contraires.

. X_9 a les mêmes liaisons que X_6 bien que son association avec Y_2 ne soit pas significative.

Donc, les classes X_3, X_6 et X_9 définissent les facteurs de l'H.T.A., tandis que X_1, X_4 et X_5 ceux de la TAS faible ou modérée.

On trouvera un résumé de toutes ces liaisons, ainsi que la description des différentes classes concernées dans le tableau 5.

5.4.3. Analyse des correspondances du tableau de croisement

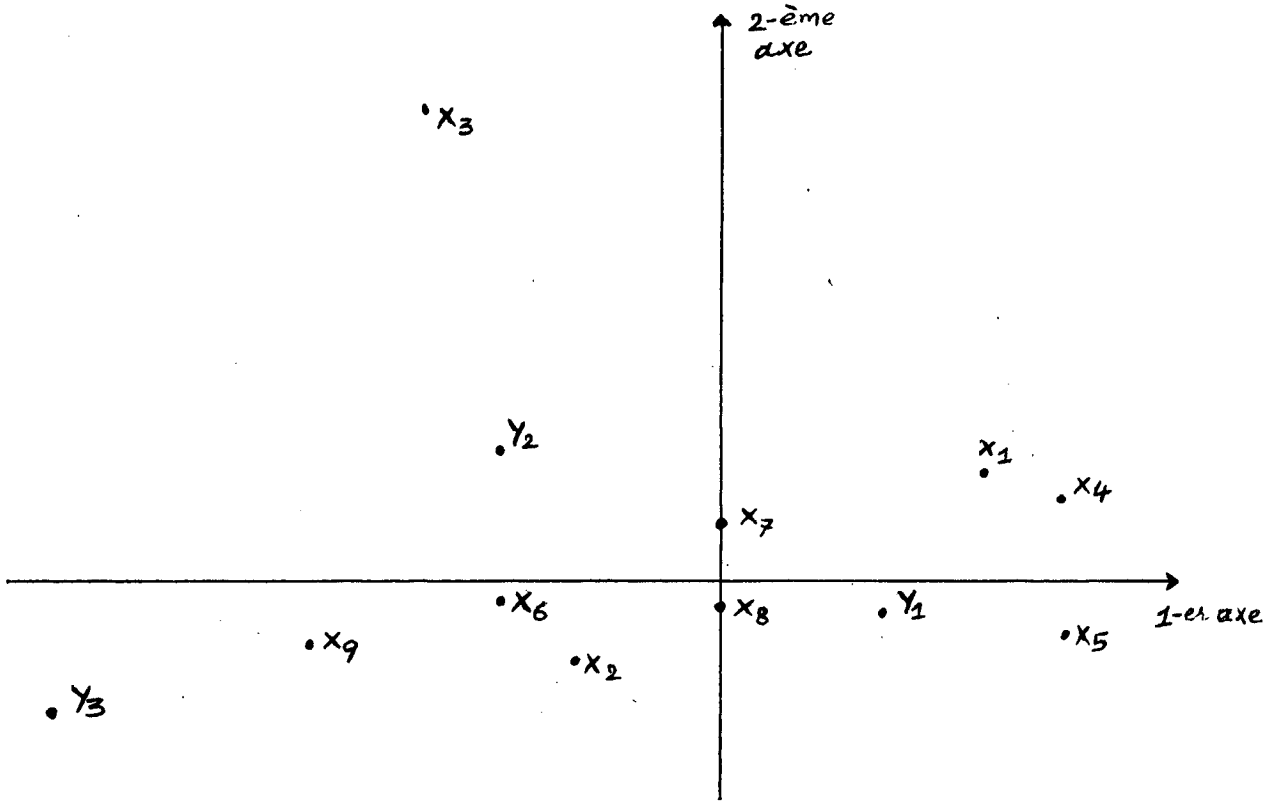
L'analyse des correspondances du tableau C permet une visualisation dans le plan des axes 1-2 des positions relatives des classes X_i et Y_j. Premier axe est une échelle de la T.A.S. (voir graphique 3).

Remarques : 1) X_7 et X_8 dont les associations ne sont pas significatives avec les classes Y de la TAS, comme on vient de le voir, avec les indices d'association, se trouvent près du centre de gravité.

2) Il y a opposition entre X_1,X_4 et X_5 d'une part et X_3,X_6 et X_9 d'autre part.

RESUME DU CROISEMENT DES CLASSIFICATIONS ET DES ASSOCIATIONS SIGNIFICATIVES

Composition des classes X_r	Composition des classes Y_s	TAS <10, 10 à 12,9	TAS 13 à 13,9	TAS 14 à 17,9
	X_r \ Y_s	Y_1	Y_2	Y_3
Célibataire, Profession Libérale ou Cadre Supérieur, Habite Petit Immeuble, Ancienne Fumeuse, Journée Continue ; Cholestérol 1,00 à 1,49 ; VGM 92 à 93,9	X_1			-
Veuve, Veuve Remariée, Patrons d'Industrie ou de Commerce, Travail Occasionnel, Logement Divers ; Cholestérol < 1 ; Glycémie < 0,60 ; Acide urique 0,20 à 0,39 ; Triglycérides < 0,50 ; Alcoolisme 303.	X_4			-
Divorcée, Divorcée Remariée, Cadre Moyen, Employée, Habite Grand Ensemble, Fumeuse inhalant, Fume depuis moins de 10 ans, Quantité de Tabac Cumulée>20 k, Taille 1m60 à 1m69 ; Cholestérol 1,50 à 1,99 ; Glycémie 0,60 à 0,79 ; Indice de Quetelet 15 à 20,9; VGM 96 à 99,9 ; Acude Urique 0,40 à 0,49 ; GGT<19.	X_5	+	-	-
Logement Rural, 2ème repas pris "à la gamelle" ; Indice de Quetelet 23 à 24,9 ; GGT 60 à 79.	X_3		+	
Mariée, Habite Maison Particulière, Non-Fumeuse, Taille<1m60 ; Cholestérol 2,50 à 2,99 ; Glycémie 0,80 à 0,99 ; Acide Urique 0,60 à 0,69 ; Triglycérides 1,00 à 1,99 ; Indice de Quetelet 25 à 34,9 ; GGT 20 à 39 ; VGM<88 ; VGM 100 à 103,9.	X_6	-	+	+
Personnel de Service, Non-actif, Temps Partiel, Principaux Repas en Milieu Familial ; Cholestérol 2 à 2,49 ; Glycémie 1 à 1,39 ; Acide Urique 0,50 à 0,59.	X_9	-		+



Graphique 3. Analyse des Correspondances du tableau C.

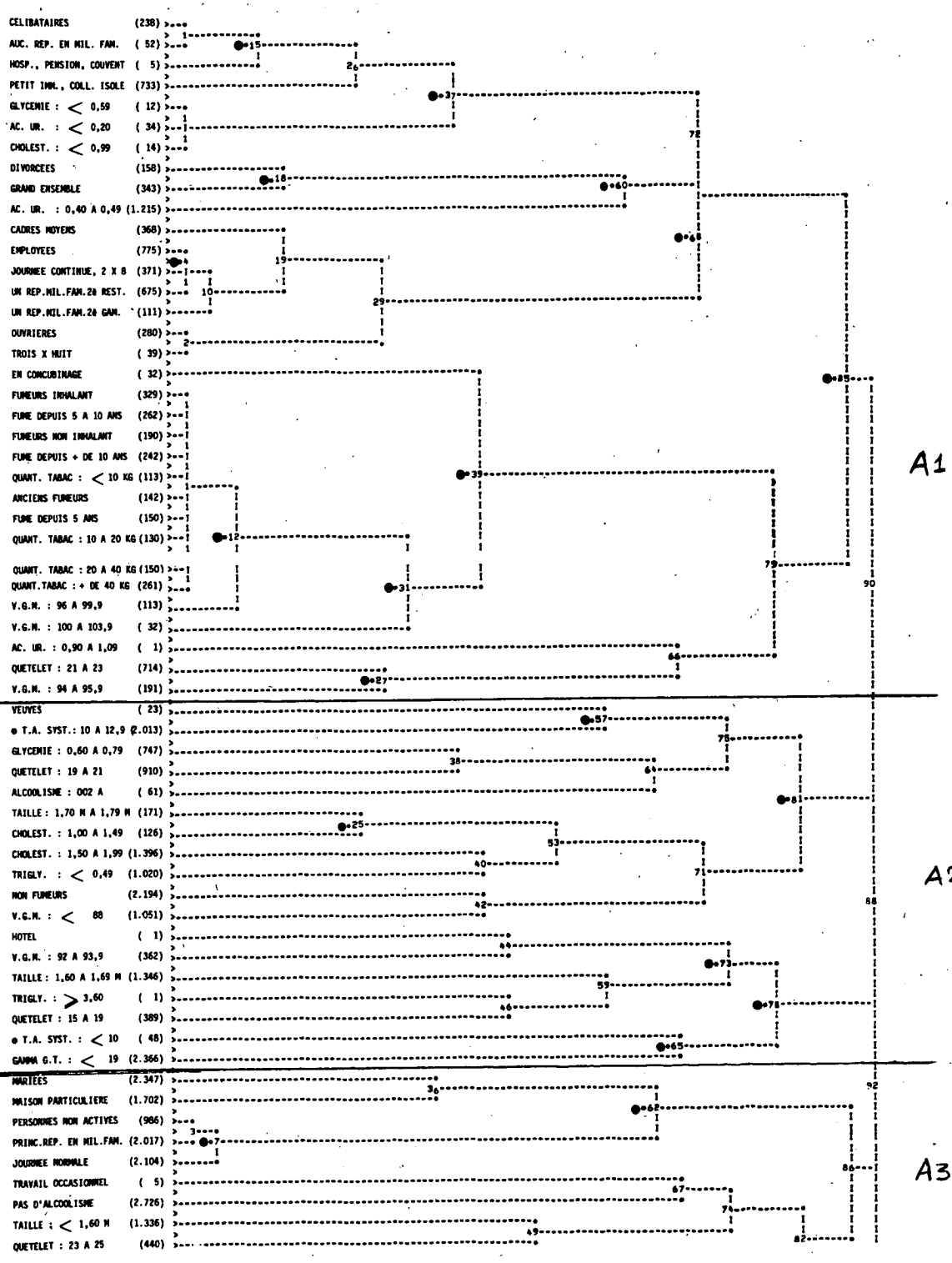
6. CONCLUSIONS

Les méthodes proposées au paragraphe 4 nous ont permis d'étudier les facteurs associés à l'H.T.A. par la classification hiérarchique par A.V.L. de l'ensemble des attributs du tableau d'incidence d'une part ; et par la classification des lignes et des colonnes d'une suite de tableaux de contingence (tableau de "régression") d'autre part. Les résultats trouvés ne sont pas seulement concordants avec notre précédente étude [7], mais en plus ils ont permis aux médecins de découvrir que le seuil de "sécurité" de la T.A.S. était de 13, alors que le seuil pathologique est de 16 selon l'O.M.S. Ce résultat non-négligeable permettra la mise sur pied d'un système de surveillance efficace aux Centres d'Examens de Santé dans un premier temps, et sa généralisation, après d'autres études de confirmation, à tous les médecins. Nous avons pu mettre en évidence certains facteurs sociologiques de l'H.T.A. tels que la profession, horaire de travail et habitat ; et en avons disculpé d'autres tel que le tabac.

ANNEXE

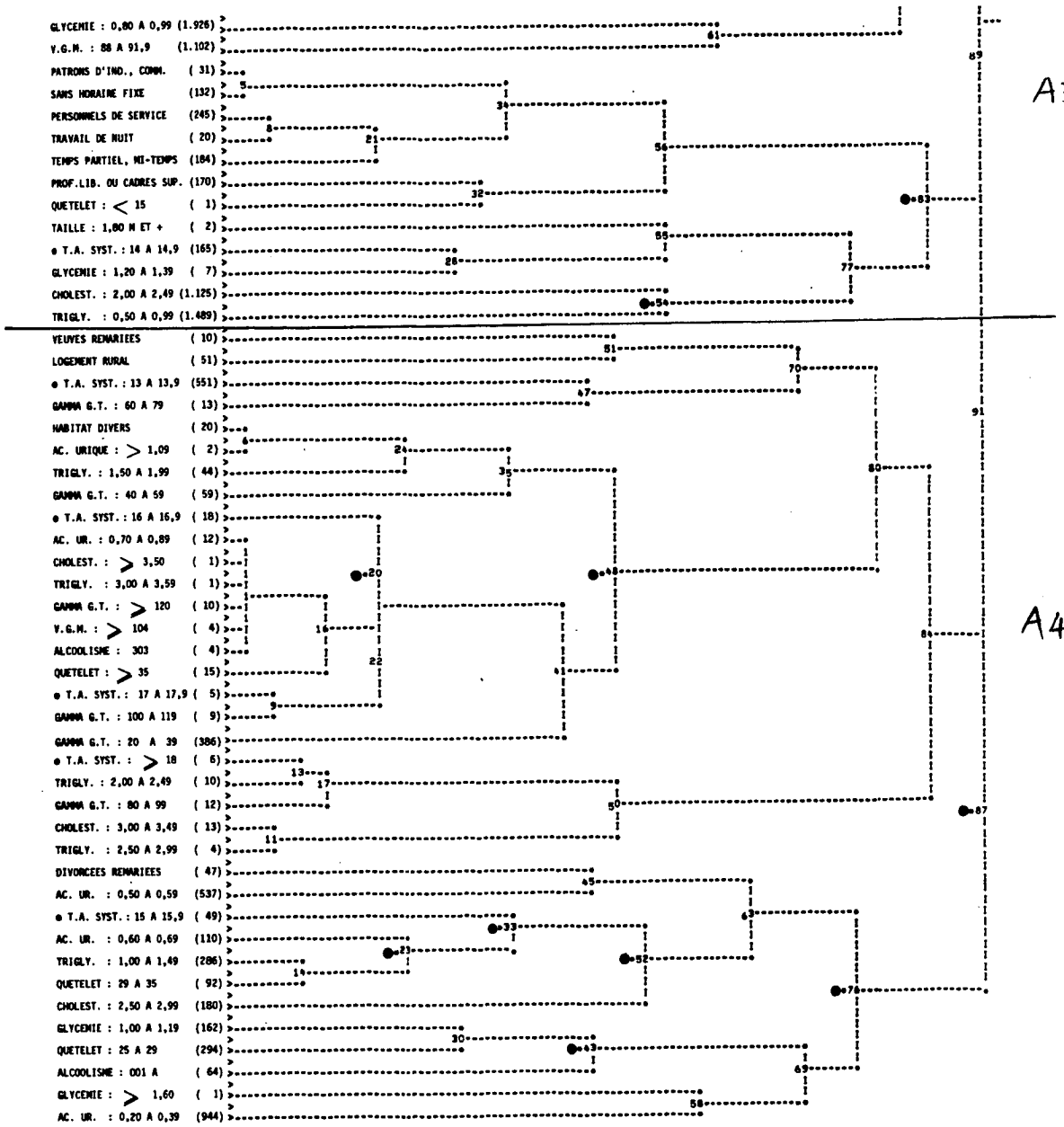
REPRÉSENTATION DE L'ARBRE

" HYPERTENSION ARTÉRIELLE " 4 CENTRES { ALBI, NICE, SAINT-BRIEUC, RENNES } " FEMMES DE 30 À 39 ANS "



(suite

(suite de l'arbre de classification)



(Les nombres entre parenthèses sont des effectifs)

BIBLIOGRAPHIE

[1] M. CAILLET, L. MASSE, H. COURCOUX, E. COSTE, E. ABOU, B. TALLUR et B. DUPONT (Juin 1981) :
"Importance du Niveau de la Tension Artérielle Systolique dans la sélection de populations-cibles en Médecine Préventive".
Communication du 5-ème Colloque National des Centres d'Examens de Santé, Bordeaux.

[2] H. COURCOUX (Déc.81)
Thèse de Docteur en Médecine.
Faculté de Rennes

[3] I.C. LERMAN (1973)
"Etude Distributionnelle de Statistiques de Proximité entre Structures finies de même type ; application à la classification automatique".
Cahiers du B.U.R.O. Paris n°19.

[4] I.C.LERMAN (1977)
"Reconnaissance et Classification des Structures Finies en Analyse des Données".
Vol.1. Théorie et Méthodes.
Rapport Interne IRISA, Rennes, n° 70.

[5] I.C. LERMAN (1979)
"Croisement de Classifications Floues".
Publication Interne IRISA, Rennes, n° 108.

[6] I.C. LERMAN et B. TALLUR (1980)
"Classification des Eléments Constitutifs d'une Juxtaposition de Tableaux de Contingence".
R.S.A. Vol. XXVIII n°3.

[7] L. MASSE, B. TALLUR, M. GALLOU
"Rapport sur l'Hyper Tension Artérielle au Centre d'Examens de Santé de la Caisse Primaire d'Assurance Maladie d'Ille-et-Vilaine".

[8] B. TALLUR (1978)
"Etude de l'Agriculture Régionale Française par une Méthode de Classification Automatique".
Publication Interne IRISA, Rennes, n° 103.

Liste des Publications Internes IRISA

PI 130 **Un système d'enchères simultanées réparti**
M. Banâtre, A.Couvert , 27 pages ; *Janvier 1980*

PI 131 **Méthodes Pseudo-Booléennes**
G. Boulaye , 14 pages ; *Janvier 1980*

PI 132 **Semaine d'étude internationale sur la détection et l'estimation temps-réel des contours et des mouvements dans les images**
Organisation : CCETT, INSA, IRISA (Rennes), , 260 pages ; *Septembre 1979*

PI 133 **Weak solutions for semi-martingales**
J. Pellaumail , 19 pages ; *Mars 1980*

PI 134 **Processus non sequentiel et leurs observations en univers non-centralisé**
Ph. Darondeau , 53 pages ; *Avril 1980*

PI 135 **Synchronisation and protection features for data abstraction**
D. Herman, M.Raynal , 21 pages ; *Juin 1980*

PI 136 **Elaboration et évaluation d'un graphe d'implication pour des données binaires**
I.C. Lerman, R. Gras, H. Rostam , 70 pages ; *Août 1980*

PI 137 **Rapport à la C.E.E. sur les techniques de programmation**
J. André, D. Coan, H. Geist, D. Law, Y. Letertre, H. Schoenen , 96 pages ; *Septembre 1980*

PI 138 **Définition d'un logiciel de transport adapté à la mise en oeuvre d'application burotiques sur des réseaux locaux**
S. Gaucher Cazalis, F. Krier, H. Le Goff , 70 pages ; *Septembre 1980*

PI 139 **Efficacité des algorithmes récursifs en présence de systèmes non-stationnaires.**
A. Benveniste, G. Ruget , 35 pages ; *Août 1980*

PI 140 **Structures de communication extensibles**
P. Le Guernic, M. Raynal , 60 pages ; *Octobre 1980*

PI 141 **Comparaison de tableaux de fréquence**
B. Escoffier , 16 pages ; *Octobre 1980*

PI 142 **Un lemme général de stabilité pour la commande adaptative en déterministe de systèmes non nécessairement à minimum de phase]:**
Cl. Samson , 40 pages ; *Novembre 1980*

PI 143 **Détection, Estimation de l'orientation et saisie d'une cible mobile par proximétrie optique**
B. Espiau , 142 pages ; *Janvier 1981*

PI 144 **Une contribution à l'étude de l'impact de l'informatique sur les organisations**
L. Breton, A. Prod'homme, J. Villard , 58 pages ; *Décembre 1980*

PI 145 **Rupture de modèles statistiques**
M. Basseville, A. Benveniste , 130 pages ; *Mars 1981*

PI 146 **Traitement des questionnaires avec non réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte**
B. Escoffier , 38 pages ; *Mars 1981*

PI 147 **Deux files d'attente à capacité limitée en tandem**
J. Pellaumail, J. Boyer , 19 pages ; *Juillet 1981*

PI 148 **Programme de classification hiérarchique : 1) Méthode de la vraisemblance des liens, 2) Méthode de la variance expliquée**
I.C. Lerman , 113 pages ; *Juin 1981*

PI 149 **Convergence des méthodes de commande adaptative en présence de perturbations aléatoires**
J.J. Fuchs , 46 pages ; *Juillet 1981*

PI 150 **Construction automatique et évaluation d'un graphe d'«implication» issu de données binaires, dans le cadre de la didactique des mathématiques**
H. Rostam , 112 pages ; *Juin 1981*

PI 151 **Réalisation d'un outil d'évaluation de mécanismes de détection de pannes]-]Projet Pilote SURF**
B. Decouty, G. Michel, C. Wagner, Y. Crouzet , 59 pages ; *Juillet 1981*

PI 152 **Règle maximale**
J. Pellaumail , 18 pages ; *Septembre 1981*

PI 153 **Corrélation partielle dans le cas « qualitatif »**
I.C. Lerman , 125 pages ; *Octobre 1981*

PI 154 **Stability analysis of adaptively controlled not-necessarily minimum phase systems with disturbances**
Cl. Samson , 40 pages ; *Octobre 1981*

PI 155 **Analyses d'opinions d'instituteurs à l'égard de l'appropriation des nombres naturels par les élèves de cycle préparatoire**
R. Gras , 37 pages ; *Octobre 1981*

PI 156 **Récursion induction principle revisited**
G. Boudol, L. Kott , 49 pages ; *Décembre 1981*

PI 157 **Loi d'une variable aléatoire à valeur $\mathbf{R}^*$** réalisant le minimum des moments d'ordre supérieur à deux lorsque les deux premiers sont fixés
M.Kowalowka, R. Marie , 8 pages ; *Décembre 1981*

PI 159 **Méthode d'interprétation d'une classification hiérarchique d'attributs-modalites pour l'«explication» d'une variable ; application à la recherche de seuil critique de la tension artérielle systolique et des indicateurs de risque cardiovasculaire**
B. Tallur , 34 pages ; *Janvier 1982*

PI 160 **Probabilite stationnaire d'un réseau de files d'attente multiclasse à serveur central et à routages dépendant de l'état**
L.M. Le Ny , 18 pages ; *Janvier 1982*

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique