REVUE DE STATISTIQUE APPLIQUÉE

B. TALLUR

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

Revue de statistique appliquée, tome 31, n° 1 (1983), p. 25-43

<http://www.numdam.org/item?id=RSA_1983__31_1_25_0>

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

B. TALLUR[2]

1. RESUME

A travers la classification hiérarchique par A.V.L. (Algorithme de la Vraisemblance des Liens) d'un ensemble de modalité-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 (T.A.S.), 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 T.A.S.

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 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 A.V.L. (I.C. LERMAN) permet de dégager les variables expliquant celle à expliquer.

Le problème nous a été posé par une étude sur l'hypertension artérielle (voir § 4) dont l'objectif a été de rechercher les facteurs responsables de risques cardiovasculaires parmi un ensemble de variables biologiques et sociologiques

(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 (C.E.S. de St-Brieuc, Rennes et Albi respectivement), Dr Bernard DUPONT, M. Hubert COURCOUX (Faculté de Médecine, Rennes) et Dr Louis MASSE (Ecole Nationale de la Santé Publique de Rennes).

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

Mots Clés : **C.A.H., A.V.L., relation d'ordre, indice d'association.**

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.

Pour une présentation de l'Algorithme de la Vraisamblance du Lien (A.V.L.) de I.C. LERMAN et des indices de proximité compatibles avec l'A.V.L. pour différents types de tableaux de données, et en particulier aux deux types de tableaux ci-dessus, nous renvoyons les lecteurs aux références bibliographiques [3], [6], [7] et [9].

3. 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.. 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 à expliquer par les 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 3.1., nous présenterons les aspects formels de la méthode d'interprétation pour le premier type de données et au § 3.2. on exposera la démarche permettant l'interprétation pour le second type de données.

3.1. 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 avons retenu la "meilleure" partition selon le critère de la "statistique globale" (voir [6], [7]) de A en ℓ classes : $A = \{A_1, A_2, \ldots, A_\ell\}$. Chaque classe A_i est composée de zéro ou une 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 dèfini par chacune des variables $X_2, \ldots, X_m$, au moyen d'une mesure de similarité.

(Signalons que la relation d'ordre définie par chacune des variables $Y, X_2, X_3, \ldots, X_m$ sur l'ensemble des classes $\{A_k / 1 \leqslant k \leqslant \ell\}$ est généralement une relation d'ordre partiel).

3.1.1. 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 $X_{ij} (1 \leqslant j \leqslant n_i)$ à l'intérieur des classes $A_k (1 \leqslant k \leqslant \ell)$.

Cas 1

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

1-a) A_k ne contient aucune modalité $X_{ij} (1 \leqslant j \leqslant n_i)$ de X_i
1-b) A_k contient exactement une des modalités $X_{ij} (1 \leqslant j \leqslant n_i)$

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

Définition 1

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

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

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

$$A_{k(j)} < A_{k(j')} \Leftrightarrow j < j'$$

Cas 2

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

2-a) A_k ne contient aucune modalité $X_{ij} (1 \leqslant j, \leqslant n_i)$ de X_i
2-b) A_k contient exactement l'ensemble des modalités X_{ij} associées à r indices j consécutifs, soit à un ensemble d'indices de la forme :

$$[g(h) + 1, g(h) + 2, \ldots, g(h + 1)], \text{ avec } g(h + 1) = g(h) + r .$$

La fonction g ainsi définie est donc une fonction croissante de h, h variant de 0 à t, si les modalités de X_i appartiennent à t classes différentes (avec $g(0) = 0$, $g(t) = n_i$) ; les modalités associées aux indices $[g(h) + 1, \ldots, g(h + 1)]$ appartenant à une même classe, $g(h + 1) + 1$ appartenant à une classe différente.

Définition 2

Si l'on désigne par $A_{k(h)}$ la classe des attributs contenant exactement l'ensemble des modalités associées à un même intervalle des indices j de la forme $[g(h) + 1 \leqslant j \leqslant g(h + 1)]$ où g est la fonction entière définie ci-dessus, l'ordre associé à la variable X_i sur l'ensemble des classes $\{A_1, A_2, \ldots, A_\ell\}$ sera défini par :

$$A_{k(h)} < A_{k(h')} \Leftrightarrow g(h) < g(h')$$

Remarque : Le cas 1 est un cas particulier du cas 2 où chaque ensemble des modalitéx X_{ij} se réduit à une seule modalité $(g(h + 1) - g(h) = 1, \forall h)$.

Cas 3

On est dans un cas différent des deux premiers. C'est-à-dire que chacune des classes A_i soit ne contient aucune modalité de la variable X_i, soit contient deux ou plusieurs modalités X_{ij} qui ne sont pas toutes contigües.

Dans l'ensemble de toutes les modalités de X_i appartenant à la classe A_k il peut exister des sous-ensembles tels que à l'intérieur de chaque sous-ensemble toutes les modalités soient contigües.

Définition 3 – Degré de neutralité d'une modalité

Le degré de neutralité d'une modalité $d \in D$ où D est l'ensemble de toutes les modalités à classifier, 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

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

où $s(d, c)$ est la valeur de l'indice de proximité centré et réduit entre les modalités d et c (cf.[6], chapitre 2, § IV 1.1) ; et $\bar{S}(d)$ est la valeur moyenne des proximités entre d et chacune des autres modalités :

$$\bar{S}(d) = \frac{1}{\text{card}(D) - 1} \sum_{c \in D - \{d\}} \bar{S}(c, d)$$

On dira que la modalité a est moins "neutre" que la modalité b si $V(a) > V(b)$. (cf. [6], chap. 3, III, pp. 184-186).

On affectera dans ce cas à la classe A_k la modalité la moins "neutre" ou la réunion des modalités appartenant au sous-ensemble des modalités consécutives contenant la modalité la moins "neutre".

Définition 4

Dans le cas 3 si l'on désigne par $A_{k(h)}$ la classe des atrributs à laquelle est associée soit la modalité de la variable X_i d'indice h, soit le sous-ensemble des modalités dont l'indice le plus petit est h, l'ordre sur $\{A_k / 1 \leqslant k \leqslant \ell\}$ associé à la variable X_i sera définie par

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

3.1.2. Comparaison des ordres sur $\{A_k/1 \leqslant k \leqslant \ell\}$ 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 $\{A_k/1 \leqslant k \leqslant \ell\}$ 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 le taux de Cholestérol, le taux de Triglycérides, le taux de Gamma GT, la Surcharge Pondérale... etc.

Définition 5

La variable $X_i (i \neq 1)$ sera appelée "indicateur" de la variable Y si l'ordre induit sur l'ensemble $\{A_k/1 \leqslant k \leqslant \ell\}$ des classes d'attributs par la variable X_i est, soit exactement identique, soit tout à fait opposé à celui défini par la varible 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. (cf. [6], chapitre 2, IV, § 4).

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

Les données de base servant à l'étude de liaison entre une variable "à expliquer" Y et un ensemble de variables "explicatives" $X_2, X_3, \ldots, X_m$ sont celles du tableau de contingence K_{IJ} :

$$K_{IJ} = \{k_{ij} \mid i \in I, j \in J\}$$

où

I = ensemble des modalités de la variable Y,

$J = J^{(2)} \cup J^{(3)} \cup \ldots \cup J^{(m)}$

$J^{(k)}$ $(2 \leqslant k \leqslant m)$ étant l'ensemble de modalités de la variable X_k, et k_{ij} = nombre d'individus ayant la modalité i de I et la modalité j de J.

K_{IJ} est en effet une juxtaposition de $(m - 1)$ tableaux de contingence, et l'indice de proximité défini par I.C. LERMAN et B. TALLUR [7] permet d'effectuer une classification hiérarchique par A.V.L. sur l'ensemble I et sur l'ensemble J.

Soient $I = \{I_1, I_2, \ldots, I_p\}$ et $J = \{J_1, J_2, \ldots, J_q\}$ les partitions retenues des ensembles I et J en p et q classes respectivement.

L'étude de la liaison entre Y et les $X_i (2 \leqslant i \leqslant m)$ sera maintenant basée sur le tableau C croisant les deux partitions, le tableau C étant défini ci-dessous :

Définition 6

Etant donné les partitions $I = \{I_1, I_2, \ldots, I_p\}$ et $J = \{J_1, J_2, \ldots, J_q\}$ de l'ensemble des lignes et de l'ensemble des colonnes du tableau de régression K_{IJ}, le tableau de croisement C associé à ces classifications est défini par :

$$C = \{c_{rs} / 1 \leqslant r \leqslant p, 1 \leqslant s \leqslant q\}$$

avec $$c_{rs} = \Sigma \Sigma \{k_{ij} / i \in I_r, j \in J_s\}$$

Le tableau C permet de calculer des mesures d'association pour chaque couple de classes (I_r, J_s) $(1 \leqslant r \leqslant p ; 1 \leqslant s \leqslant q)$ ainsi que les "profils moyens" de chaque classe I_r (resp. J_s) à travers l'ensemble des classes J_s, $1 \leqslant s \leqslant q$ (resp. I_r, $1 \leqslant r \leqslant p$) ; le profil moyen de I_r étant le profil de la r-ième ligne du tableau C.

3.2.1. Mesures d'association entre une classe I_r et une classe J_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 trouver 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, notées χ^2_{rs}, sera d'autant plus forte que la classe I_r est plus particulièrement associée à la classe J_s :

$$\chi^2_{rs} = \frac{(C_{rs} - \frac{C_r . C_{.s}}{C..})^2}{(C_r . C_{.s}/C..)}$$

avec $$C_{r.} = \sum_s C_{rs} \; ; \; C_{.s} = \sum_r C_{rs} \quad C.. = \sum_r \sum_s C_{rs}$$

et $$\chi^2 = \Sigma \Sigma \{\chi^2_{rs} / 1 \leqslant r \leqslant p, 1 \leqslant s \leqslant q\}$$

On associe à chaque couple (I_r, J_s) le tableau de contingence à 2 lignes et 2 colonnes suivant.

	J_s	$\bar{J}_s$	
I_r	C_{rs}	$C_{r.} - C_{rs}$	$C_{r.}$
$\bar{I}_r$	$C_{.s} - C_{rs}$	$C.. - C_{r.} - C_{.s} + C_{rs}$	$C.. - C_{r.}$
	$C_{.s}$	$C.. - C_{.s}$	$C..$

La statistique du χ^2 calculée sur ce tableau, et que nous noterons $\chi^2(I_r, J_s)$ permet de déterminer si la liaison du couple (I_r, J_s) est significative, en la comparant à une variable χ^2 à 1 d.l.

On vérifie que

$$\chi^2_{(I_r, J_s)} = \chi^2_{rs} \Big/ \left\{ \frac{(C\,..\, - C_{r\,.})\,(C\,..\, - C_{.\,s})}{C^2\,..} \right\}$$

χ^2_{rs} étant défini comme précédemment.

Nous proposons une mesure d'association p_{rs} entre I_r et J_s :

$$p_{rs} = \text{Prob}\,[\chi^2_1 < \chi^2_{(I_r, J_s)}]$$

où χ^2_1 est une v.a. suivant la loi du χ^2 à 1 d.l.

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'assciations notées χ_{rs} pour chaque couple (r , s) :

$$\chi_{rs} = \left(C_{rs} - \frac{C_{r\,.}\,C_{.\,s}}{C\,..}\right) \Big/ \sqrt{\frac{C_{r\,.}\,C_{.\,s}}{C\,..}}$$

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.

3.2.2. Analyse factorielle des correspondances du tableau C

L'A.F.C. du tableau C permet la comparaison des profils moyens des classes I_r et J_s ainsi que l'étude de leurs liaisons. L'examen des indices d'association χ_{rs} et p_{rs} définis ci-dessus et celui des résultats de l'A.F.C. conduisent à des interprétations concordantes des liaisons entre les classes de modalités de Y est les classes de modalités des variables $X_2, X_3, \ldots, X_m$.

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

4.1. 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. pour la prévention de ces maladies. De 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..

La présente étude porte sur une population associée à quatre centres : Albi, Nice, Rennes et St-Brieuc. Parmi tous les consultants de l'année 1979 on a retenu 10 693 sujets non-médicalisés ; les sujets soumis à un traitement médical après un dépistage d'anomalie cardiovasculaire ayant été éliminés. Le but du travail 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.

4.2. 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.

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 0 – 1.

Les variables biologiques suivantes sont retenues :

1) Tension Artérielle Systoloique (8 modalités) ;
2) Taux de glycémie en g/1 (7 modalités) ;
3) Taux de cholestérol en g/1 (7 modalités) ;
4) Taux de l'Acide Urique en g/1 (7 modalités) ;
5) Taux de Gamma G.T. en nombre d'Unités Internationales/1 (7 modalités) ;
6) Taux de 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.

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.

4.3. Recherches 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 "T.A.S. < 10" et "T.A.S. 10 à 12,9" ; on lui associe l'intervalle "T.A.S. $\leqslant 12{,}9$" ;
- A_3 contient la modalité "T.A.S. 14 à 14,9" qui lui est associée ;
- A_4 contient les modalités "T.A.S. 13 à 13,9", "T.A.S. 15 à 15,9", "T.A.S. 16 à 16,9", "T.A.S. 17 à 17,9" et "T.A.S. $\geqslant 18$". Il s'agit d'associer soit la modalité "T.A.S. 13 à 13,9", soit l'intervalle "T.A.S. $\geqslant 15$". Selon le critère du degré de neutralité, c'est l'intervalle "T.A.S. $\geqslant 15$" qui est associé à A_4.

L'ordre défini sur l'ensemble des classes par la T.A.S. 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 T.A.S.

4.3.1. Remarques sur et interprétation des résultats (voir tableau 1)

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 :

$$A_2 < A_3 < A_4$$

TABLEAU 1

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

Variables	T.A.S.			Taux de cholestérol			Triglycérides			Glycemie			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

On remarque que Gamma G.T., bien que n'étant pas un indicateur de la T.A.S. 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.

4.3.2. 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 effectant à chaque modalité le rang de la classe – défini par la T.A.S. – 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)

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)

4.3.3. 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 T.A.S. 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 T.A.S., et qu'un certain nombre de paramètres biologiques sont faibles chez les fumeuses. L'interprétation dynamique de l'arbre aux niveaux supérieurs montre que (voir fig. 1) cette classe se réunit avec la classe A_2 qui est caractérisée par les valeurs moyennes de la T.A.S. ainsi que celles de tous les paramètres biologiques. Il s'agit donc d'une classe "normale". Remarquons que A_2 contient "non-fumeuses".

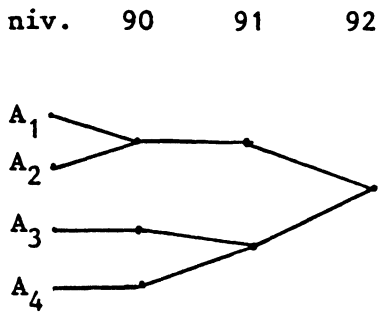


Figure 1

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 T.A.S. inférieures à 13 de celles qui sont supérieures ou égales à 13 ; ceci montre qu'il existe un seuil entre les valeurs faibles ou "normales" et les valeurs fortes de la T.A.S. et que ce seuil se situe autour de 13. Pour la médecine préventive, la connaissance de ce seuil est très importante, car elle permet de mieux surveiller les sujets au-delà de cette valeur accompagnés d'autres sysmptômes indicateurs de risques.

4.4. 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 proposé en [7] et l'algorithme A.V.L. ont 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).

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

On obtient l'arbre détaillé de classification hiérarchique suivant :

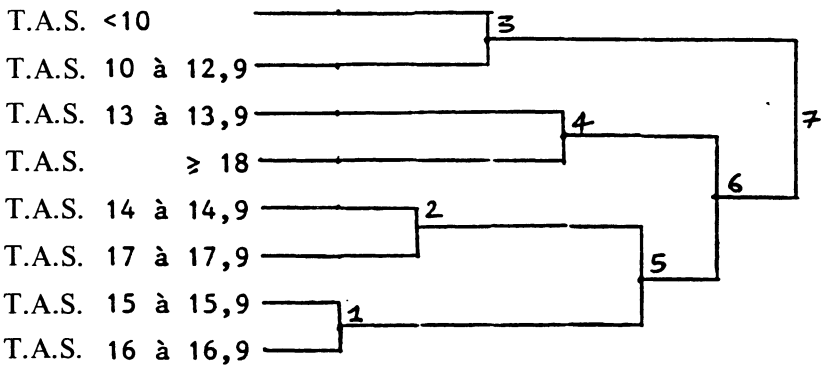


Figure 2. – Arbre détaillé de la classification des modalités de T.A.S.

Interprétation

La modalité T.A.S. ⩾ 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.

4.4.2. Classification des 110 modalités des variables biologiques et sociales explicatives et 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 , . . . , X_9 , on a le tableau de croisement C (tableau 2)

Le tableau 3 ci-dessous résume les mesures d'association p_{rs}, les indices orientés χ_{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 χ_r :

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

TABLEAU 2

Tableau C de croisement des classifications

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 3

X_r \ Y_s	Y_1				Y_2				Y_3			
	χ_{rs}	CT_{rs}	$\chi^2(I_r,J_s)$	P_{rs}	χ_{rs}	CT_{rs}	$\chi^2(I_r,J_s)$	P_{rs}	χ_{rs}	CT_{rs}	$\chi^2(I_r,J_s)$	P_{rs}
X_1	0,8	8,11	2,35	0,85	0,2	0,57	0,06	0,2	-2,8	91,32	9,04	>0,995
X_2	-0,4	7,10	0,67	0,60	-0,2	1,25	0,04	<0,2	1,6	91,65	2,91	0,90
X_3	-1,5	15,86	8,14	>0,995	3,4	80,20	14,56	>0,999	-0,8	3,94	0,63	0,6
X_4	1,5	14,92	8,16	>0,995	-0,6	2,23	0,48	0,5	-3,6	83,85	15,78	>0,999
X_5	3,8	25,40	47,31	>0,999	-3,9	26,95	30,34	>0,999	-5,2	47,65	49,07	>0,999
X_6	-2,6	21,99	24,92	>0,999	2,1	14,64	11,44	>0,995	4,4	63,37	46,54	>0,999
X_7	-0,3	9,02	0,31	0,4	0,9	72,98	1,05	0,7	-0,4	18,00	0,23	0,36
X_8	0,2	10,79	0,10	0,25	-0,4	75,81	0,31	0,4	0,2	13,40	0,05	0,2
X_9	-2,2	18,50	17,59	>0,999	1,4	6,99	2,63	0,86	4,5	74,51	24,88	>0,999

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 pour un risque d'erreur 0,05).

– X_1 et X_4 sont négativement associées à Y_3 (T.A.S. très élevée).

– X_2 est positivement mais faiblement associée avec Y_3 ($p_{23} = 0{,}90$)

– X_3 est positivement associée avec Y_2 (T.A.S. assez forte), et négativement avec Y_1.

– X_5 et X_6 sont opposées ; X_5 étant positivement associée à la T.A.S. faible et négativement associée avec la T.A.S. 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 T.A.S. faible ou modérée.

On trouvera un résumé de toutes ces liaisons, ainsi que la description des différentes classes concernées dans le tableau 4.

4.4.3. 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_i. Le premier axe est une échelle de la T.A.S. (voir Fig. 3).

Remarques :

1) X_7 et X_8 dont les associations ne sont pas significatives avec les classes Y de la T.A.S., 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.

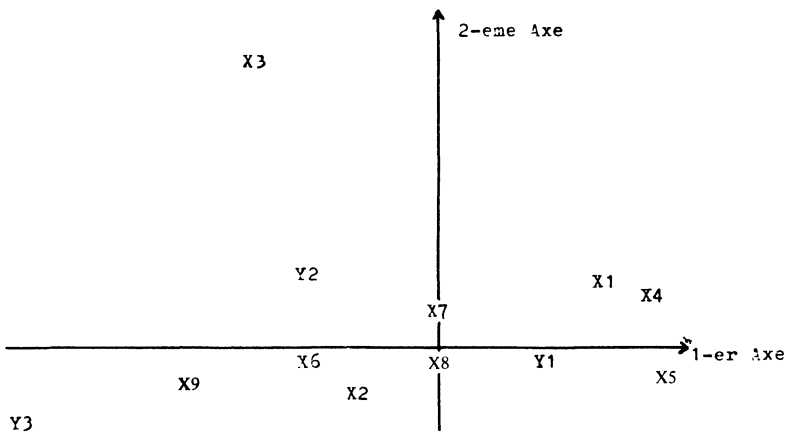


Figure 3. – Analyse des Correspondances du tableau C.

TABLEAU 4

RESUME DU CROISEMENT DES CLASSIFICATIONS ET DES ASSOCIATIONS SIGNIFICATIVES

Composition des classes X_r	Composition des classes Y_s / X_r	TAS <10, 10 à 12,9 (Y_1)	TAS 13 à 13,9 (Y_2)	TAS 14 à 17,9 (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	−		+

5. 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.

Les études des trois autres populations, à voir, Femmes de 40 à 49 ans, Hommes de 30 à 39 ans et Hommes de 40 à 49 ans ont permis notamment de constater que le seuil de "sécurité" de la T.A.S. varie selon le sexe et la tranche d'âge. Il se situe pour les Femmes de 40 à 49 ans à T.A.S. = 14 et pour les Hommes il se déplace vers le haut d'une unité par rapport aux Femmes, dans les deux tranches d'âge. En ce qui concerne les facteurs de risque d'hypertension ils restent plus ou moins inchangés, surtout les facteurs biologiques.

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 5e 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.

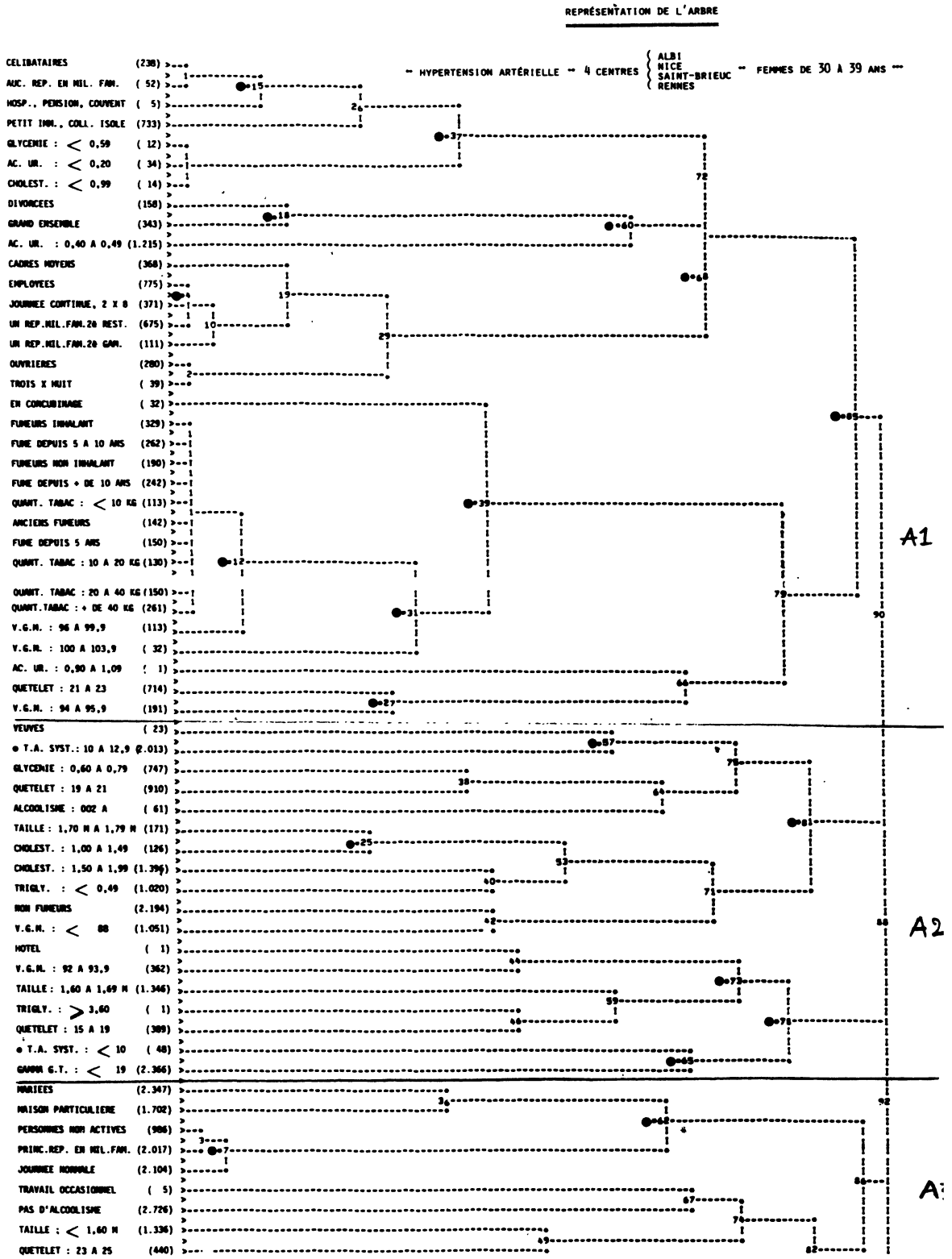
[6] I.C. LERMAN (1981). – *Classification et Analyse Ordinale des Données.* Dunod, Paris.

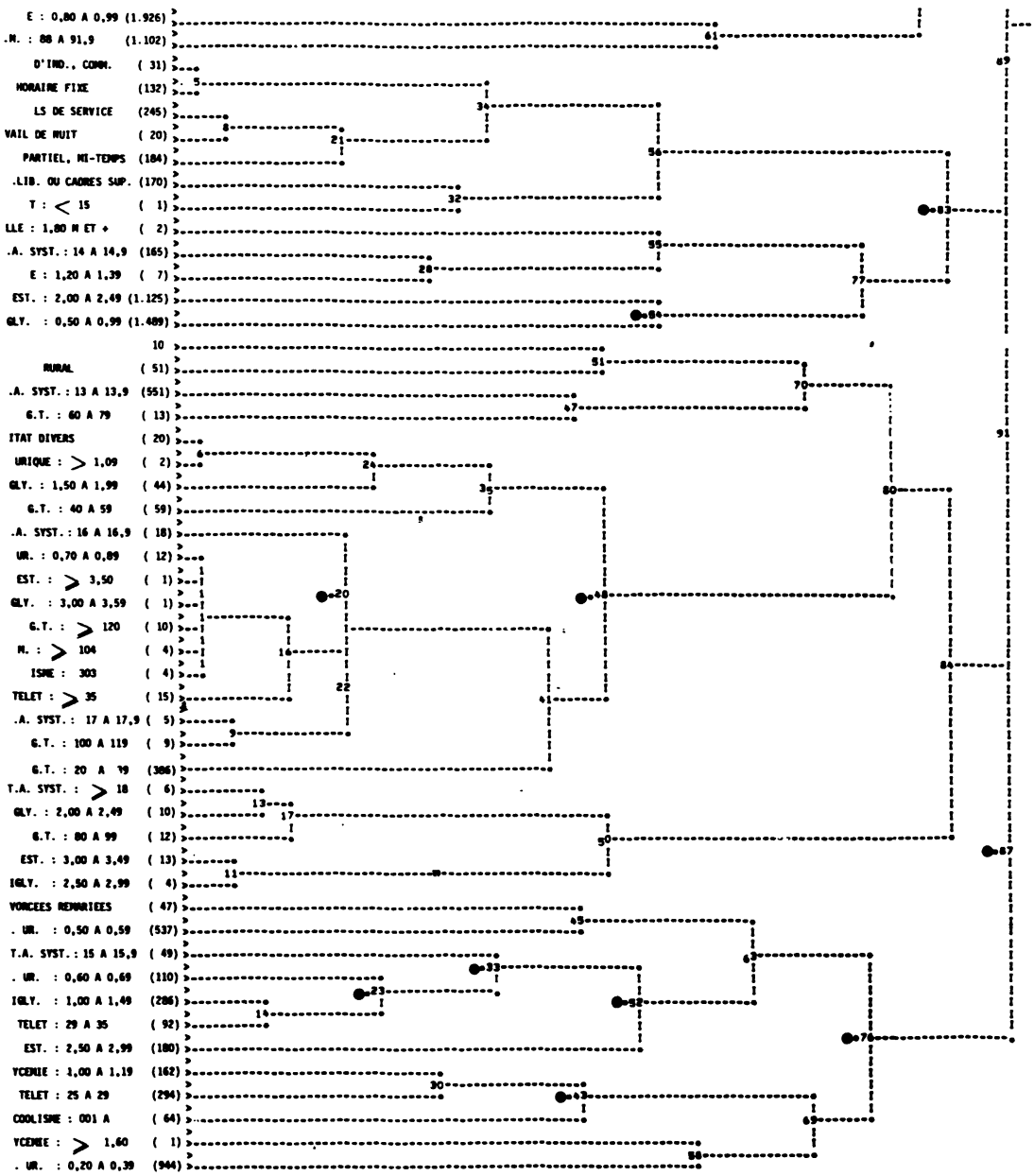
[7] I.C. LERMAN et B. TALLUR (1980). – *Classification des Eléments Constitutifs d'une Juxtaposition de Tableaux de Contingence.* R.S.A. Vol. XXVIII n° 3.

[8] 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.*

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

ANNEXE


REPRÉSENTATION DE L'ARBRE
** HYPERTENSION ARTÉRIELLE ** 4 CENTRES (ALBI NICE SAINT-BRIEUC RENNES) ** FEMMES DE 30 A 39 ANS **
CELIBATAIRES (238)
AUC. REP. EN MIL. FAM. (52)
HOSP., PENSION, COUVENT (5)
PETIT IND., COLL. ISOLE (733)
GLYCEMIE : < 0,59 (12)
AC. UR. : < 0,20 (34)
CHOLEST. : < 0,99 (14)
DIVORCEES (158)
GRAND ENSEMBLE (343)
AC. UR. : 0,40 A 0,49 (1.215)
CADRES MOYENS (368)
EMPLOYEES (775)
JOURNEE CONTINUE, 2 X 8 (371)
UN REP.MIL.FAM.2e REST. (675)
UN REP.MIL.FAM.2e GAR. (111)
OUVRIERES (280)
TROIS X HUIT (39)
EN CONCUBINAGE (32)
FUMEURS INHALANT (329)
FUME DEPUIS 5 A 10 ANS (262)
FUMEURS NON INHALANT (190)
FUME DEPUIS + DE 10 ANS (242)
QUANT. TABAC : < 10 KG (113)
ANCIENS FUMEURS (142)
FUME DEPUIS 5 ANS (150)
QUANT. TABAC : 10 A 20 KG (130)
QUANT. TABAC : 20 A 40 KG (150)
QUANT.TABAC : + DE 40 KG (261)
V.G.M. : 96 A 99,9 (113)
V.G.M. : 100 A 103,9 (32)
AC. UR. : 0,90 A 1,09 (1)
QUETELET : 21 A 23 (714)
V.G.M. : 94 A 95,9 (191)
VEUVES (23)
T.A. SYST.: 10 A 12,9 (2.013)
GLYCEMIE : 0,60 A 0,79 (747)
QUETELET : 19 A 21 (910)
ALCOOLISME : 002 A (61)
TAILLE : 1,70 M A 1,79 M (171)
CHOLEST. : 1,00 A 1,49 (126)
CHOLEST. : 1,50 A 1,99 (1.396)
TRIGLY. : < 0,49 (1.020)
NON FUMEURS (2.194)
V.G.M. : < 88 (1.051)
HOTEL (1)
V.G.M. : 92 A 93,9 (362)
TAILLE : 1,60 A 1,69 M (1.346)
TRIGLY. : > 3,60 (1)
QUETELET : 15 A 19 (389)
T.A. SYST. : < 10 (48)
GAMM G.T. : < 19 (2.366)
MARIEES (2.347)
MAISON PARTICULIERE (1.702)
PERSONNES NON ACTIVES (986)
PRINC.REP. EN MIL.FAM. (2.017)
JOURNEE NORMALE (2.104)
TRAVAIL OCCASIONNEL (5)
PAS D'ALCOOLISME (2.726)
TAILLE : < 1,60 M (1.336)
QUETELET : 23 A 25 (440)
A1
A2
A3



(Les chiffres entre parenthèses sont des effectifs).