Data Analysis and Informatics IV
E. Diday et al. (editors)
North Holland, 1986.

# ORGANISATION ET CONSULTATION D'UNE BANQUE DE "PETITES ANNONCES" A PARTIR D'UNE METHODE DE CLASSIFICATION HIERARCHIQUE EN PARALLELE

I.C. LERMAN et Ph. PETER
IRISA
Campus de Beaulieu
35042 Rennes Cédex
France

This research connects and develops in an original way, two fields : classification (hierarchical) and data banks. The problem is the organization and the consultation -in real time- of a large set (some thousands) of classified advertisements. For this purpose of structuring the data bank, we show the relevance of the classification tree reduced to its significant levels and obtained (on a large sample of ads) by the algorithm of the likelyhood of the links. More precisely it has been crucial to develop new and very general concepts and algorithms :
- Comparing objects described by qualitative preordonnance variables.
- Algorithm of hierarchical parallel classification which establishes a tree of which the leaves represent clusters of "reasonable" size.
- Assignement index of an object to a node of the classification tree.
Owing to the tree structure, the consultation, extraction and dynamic management of the data bank, are very fast.

## I. INTRODUCTION

Cette recherche est née de la rencontre d'un problème posé en matière de "base de données" et d'une approche méthodologique de la classification hiérarchique.

Le problème posé est celui de la construction, la consultation et la gestion dynamique -en temps réel- d'une banque de "petites annonces" (en l'occurrence immobilières), correspondant à un corpus de plusieurs milliers d'éléments.

Pour ce problème et en matière de base de données, les techniques informatiques classiques ne fournissent pas de solutions satisfaisantes. Trois raisons militaient a priori en faveur d'une organisation par la méthode considérée de classification hiérarchique : sa très grande généralité quant à la structure des données pouvant être prises en compte, son aptitude à ne pas se laisser éclater des classes de cohésion "moyenne" -mais cohérentes- devant des noyaux "forts" et son caractère très décisionnel (noeuds et niveaux "significatifs") dans les résultats qu'elle produit.

Toutefois, les exigences du problème posé ont conduit -à différents niveaux- à des développements spécifiques et originaux dont l'intérêt et la généralité dépassent très largement le cadre de l'application. Mais il sera plus que juste d'initialiser chacune des questions méthodologiques dans le cadre de la recherche appliquée qui nous concerne.

L'ensemble des unités de données est défini par un corpus de petites annonces. On retient pour la description un ensemble de variables qualitatives globales tels que :

- "nature de la transaction" dont les modalités sont : "vente", "achat", "offre de location", "demande de location".

- "objet de la transaction" dont les modalités sont : "maison", "pavillon", "appartement", "habitation", "studio", "chambre", "local", "garage", "terrain".

- "nombre de pièces" dont les modalités sont : "une", "deux", "trois", "quatre", "cinq", "six ou sept", "huit et plus", "non mentionné", "non lieu".

etc...

Une perception fine de la ressemblance entre deux observations (ici deux petites annonces) à travers l'ensemble des variables n'autorise plus -comme il est classique- en analyse de données- une représentation de nature géométrique, voire même ensembliste. Il y a en effet une relation de similitude sur l'ensemble M des modalités d'une même variable que peut traduire l'"expert" en termes d'une préordonnance totale sur M ou -de façon plus riche, mais plus arbitraire- en termes d'une matrice MxM de similarités, non nécessairement symétrique. A cet égard, nous développons au paragraphe II un coefficient -conforme à la classe de nos indices- de comparaison entre objets ainsi codés.

Devant l'importance de la taille du corpus à traiter (quelques milliers) pour une "bonne" estimation de la structure statistique de la banque et devant la nécessité de répondre à une annonce donnée par une classe de taille "raisonnable", nous avons mis au point une algorithmique originale de classification hiérarchique en parallèle pour le traitement de "gros" ensembles. Cette algorithmique -que nous étudions au paragraphe III- pose des problèmes méthodologiques et statistiques (découpage aléatoire de l'ensemble à classifier en "tranches", organisation hiérarchique partielle des tranches et règle d'arrêt basé sur un critère statistique objectif,...). D'autre part, notre algorithme est susceptible d'une implémentation sur calculateur parallèle, ce qui permet une très grande rapidité et un faible encombrement mémoire.

La consultation automatique de la banque permettant de répondre à une annonce donnée par une classe d'annonces "voisines" et complémentaires quant à la nature de la transaction, conduit à la nécessité d'une représentation adéquate d'une classe quelconque et d'une définition associée d'un indice d'affectation entre un élément et une classe. La représentation et l'affectation tiendront compte du codage adopté pour la description de l'ensemble des objets (paragraphe IV).

La consultation et la gestion dynamique (extraction ou ajout d'une annonce) se fait conformément à une recherche descendante dans un arbre dont les feuilles correspondent à des classes d'affectation. Cette descente -qui démarre du dernier niveau "significatif" procède par aiguillages successifs à partir de l'indice d'affectation entre l'objet ("petite annonce") et une même classe sous-tendue par un noeud. Pour une telle recherche, la réponse est quasi-instantanée (paragraphe IV et démonstration).

## II. INDICE DE PROXIMITE ENTRE OBJETS DECRITS PAR DES VARIABLES RELATIONNELLES

On désigne par E l'ensemble fini des objets de cardinal n. C est un ensemble fini de caractères descriptifs de E. On désigne par Q le cardinal de C. $J_q$ indiquera l'ensemble des codes des modalités de la variable qualitative $c_q$ ($c_q \in C$, $1 \leq q \leq Q$). De façon plus précise, on note $J_q = \{j_q / 1 \leq j_q \leq m_q\}$ où $m_q$ est le nombre de modalités de la variable $c_q$, $1 \leq q \leq Q$.

II.1. Structure de similarité sur $J_q$ (q fixé)

Tout en restant extrêmement générale, la structure descriptive la plus riche et la moins arbitraire d'une variable qualitative est fournie par une préordonnance totale sur l'ensemble de ses modalités (cf. aussi, dans un tout autre contexte, (CHAH(1984))).

Dans ces conditions -relativement à la variable $c_q$- on introduit l'ensemble suivant des couples de modalités :

$$H_q=\{(j_q,h_q)/1\leq j_q\leq h_q\leq m_q\}, \quad (1)$$

sur lequel se trouve défini -par l'expert- un préordre total $\omega_q$ (i.e. préordonnance sur $J_q$) pour lequel un couple $(j_q,h_q)$ est d'autant plus grand -d'un point de vue ordinal- que la modalité $j_q$ ressemble à celle $h_q$ ; ainsi la dernière classe de ce préordre comporte $m_q$ termes de la forme $(j_q,j_q)$, $1\leq j_q\leq m_q$.

Exemple : Dans le cas de la variable "objet de la transaction", en codant par 1,2,...,9, la suite -dans l'ordre d'écriture- des modalités "maison", "pavillon", ..., "terrain" (cf. §I. ci-dessus), on peut proposer la préordonnance suivante : $15\sim16\sim17\sim18\sim19\sim25\sim27\sim28\sim29\sim36\sim37\sim38\sim39\sim46\sim47\sim48\sim49\sim57\sim58\sim59\sim67\sim68\sim69\sim26\sim 78\sim79\sim89<13\sim23\sim35\sim45<14\sim24\sim34\sim56<12<11\sim22\sim33\sim44\sim55\sim66\sim77\sim88\sim99$, où ij avec i<j indique le couple (i,j).

Dans l'introduction même de $H_q$, nous admettons le caractère symétrique de la notion de similarité. Sinon, comme cela pourrait se présenter dans un problème d'affectation, il suffit de définir le préordre total sur l'ensemble $K_q=J_q\times J_q$ et les mêmes considérations ci-dessous -conceptuelles et de calcul- restent valables.

A chaque élément de l'ensemble préordonné ($H_q$ dans notre cas) on associe un "rang". Pour définir précisément la fonction ordinale "rang", désignons par $(\ell_1,\ell_2,\ldots,\ell_k)$ la suite des cardinaux de la suite ordonnée des classes du préordre total. Le rang d'un élément appartenant à la j-ème classe, $1\leq j\leq k$, est posé égal à

$$\sum_{1\leq i\leq(j-1)} \ell_i+(\ell_j+1)/2$$

Ainsi, relativement à l'exemple ci-dessus, le rang de l'élément 24 est égal à 27+4+2=33. De la sorte, la somme de tous les rangs est -comme dans le cas totalement et strictement ordinal- égal à $L(L+1)/2$, où $L=\ell_1+\ell_2+\ldots+\ell_k$. La structure descriptive sera donc basée sur le tableau des rangs ainsi calculés :

$$\{r_{j_q h_q}/(j_q,h_q)\in H_q\}. \quad (2)$$

Une autre forme, plus riche mais moins générale de la relation de similarité sur $J_q$, pour une fine description de E, est fournie au moyen d'une table numérique indexée par $K_q=J_q\times J_q$, où le nombre qui se trouve à l'intersection de la ligne $j_q$ et de la colonne $h_q$ est sensé "mesurer" le degré de ressemblance entre les deux modalités $j_q$ et $h_q$. Cette table de nombres qu'on peut admettre -sans que cela soit nécessaire pour les calculs- symétrique, est supposée donnée par l'"expert". Nous l'écrivons sous la forme

$$\{p_{j_q h_q}/(j_q,h_q)\in J_q\times J_q\},$$

ou, plus simplement, en tenant compte de la symétrie,

$$\{p_{j_q h_q}/(j_q,h_q)\in H_q\}. \quad (3)$$

En réalité, la nature des calculs sera exactement la même qu'on travaille avec la table (2) des rangs, ou avec celle (3) des coefficients numériques, de sorte qu'on désignera par

$$\{s_{j_q h_q} / (j_q, h_q) \in H_q\}, \quad (4)$$

l'une ou l'autre des deux tables, étant entendu que l'étude concrète a été effectuée avec la table (2).

Nous commencerons par déterminer la contribution d'une même variable $c_q$ (dont l'ensemble des modalités est codé par $J_q$) à la similarité entre deux objets, puis -de façon égale et parallèle- nous intégrerons l'ensemble des variables.

## II.2. Contribution de $J_q$ à la ressemblance entre deux objets

L'indice q restant fixé dans ce paragraphe, nous l'omettrons pour des raisons de simplicité d'écriture.

L'élaboration de l'indice obéit dans notre approche à un principe statistique général de construction, où à partir d'un premier indice, localement défini, une normalisation est effectuée à partir de la distribution empirique de cet indice sur l'ensemble $P_2(E)$ des paires d'objets (ou parties à deux éléments) de E.

x et y désignant les deux objets à comparer, si $c(x)=j_o$ et $c(y)=h_o$, l'indice sera localement défini par le nombre $s_{j_o h_o}$ de la table (4). Il y a lieu par conséquent de préciser la distribution de $\{s_{jh}/(j,h) \in H\}$ sur $P_2(E)$. Cette distribution s'obtient très aisément à partir de la décomposition de $P_2(E)$ conformément à la partition de E déterminée par la variable qualitative c.

Plus directement, en désignant par n(j) le cardinal de la classe d'objets $E_j$, possédant la j-ème modalité du caractère c dont nous désignons par m le nombre de modalités,

$$\mathrm{card}(P_2(E)) = \sum_{1 \leq j \leq m} n(j)(n(j)-1)/2 + \sum_{1 \leq j < h \leq m} n(j)n(h). \quad (5)$$

$$= n(n-1)/2.$$

Désignons respectivement par

$$\rho_j = n(j)(n(j)-1)/n(n-1) \text{ et } \sigma_{jh} = 2n(j)n(h)/n(n-1), \quad (6)$$

la proportion de paires d'objets $\{x', y'\}$ dont les deux composantes sont dans la classe $E_j$ et celle pour lesquelles x' est dans la classe $E_j$ et y' dans celle $E_h$, $1 \leq j \leq m$ et $1 \leq j < h \leq m$.

Désignons encore par

$$\rho = \sum_{1 \leq j \leq m} \rho_j \text{ et } \sigma = \sum_{1 \leq j < h \leq m} \sigma_{jh}, \quad (7)$$

qui sont respectivement, la proportion de paires réunies et séparées par la partition $\{E_j / 1 \leq j \leq m\}$.

Toutes les paires d'objets appartenant à $E_j$ (resp. $E_j * E_h$) ont la même valeur $s_{jj}$ (resp. $s_{jh}$) de l'indice local.

La distribution de la table des similarités $\{s_{jh}/(j,h) \in H\}$ sur $P_2(E)$ est donc définie par

$$\{(s_{kk},\rho_k),(s_{jh},\sigma_{jh})/k\in J \text{ et } (j,h)\in J^{\{2\}}\}, \quad (8)$$

où nous avons noté $J^{\{2\}}=\{(j,h)/1\leq j<h\leq m\}$.

Calcul de la moyenne et de la variance de la distribution (8).

Si $\mu$ et $\mu_2$ désignent la moyenne et le moment absolu d'ordre 2, on a :

$$\mu=\sum_{1\leq k\leq m}\rho_k s_{kk}+\sum_{1\leq j<h\leq m}\sigma_{jh}s_{jh}, \quad (9)$$

$$\mu_2=\sum_{1\leq k\leq m}\rho_k s_{kk}^2+\sum_{1\leq j<h\leq m}\sigma_{jh}s_{jh}^2 \quad (10)$$

On suppose -ce qui est naturel- que $s_{kk}$ est le même pour tout k=1,2,...,m. Notons s cette valeur commune.

$$\mu=\rho s+\sum_{1\leq j<h\leq m}\sigma_{jh}s_{jh}, \quad (9')$$

$$\mu_2=\rho s^2+\sum_{1\leq j<h\leq m}\sigma_{jh}s_{jh}^2 \quad (10')$$

Le calcul informatique de la variance que nous notons V, utilise

$$V=\mu_2-\mu^2 \quad (11)$$

Plus directement,

$$V=\rho\Big(\sigma s-\sum_{j<h}\sigma_{jh}s_{jh}\Big)^2 + \sum_{1\leq j<h\leq m}\sigma_{jh}\Big(s_{jh}-\rho s-\sum_{\ell<k}\sigma_k s_{\ell k}\Big)^2, \quad (12)$$

qui se met sous la forme

$$V=\sum_{1\leq j\leq h\leq m}\sigma_{jh}\Big(\sum_{1\leq \ell\leq k\leq m}\sigma_{\ell k}(s_{jh}-s_{\ell k})\Big)^2, \quad (13)$$

en ayant au préalable noté $\sigma_{kk}$ pour $\rho_k$, $1\leq k\leq m$.

Le numérateur de l'indice d'association s'écrit :

$$s_{j_o h_o}-\rho s-\sum_{1\leq j<h\leq m}\sigma_{jh}s_{jh}, \quad (14)$$

qu'on peut mettre sous la forme

$$1\leq \sum_{k\leq m}\sigma_{\ell k}(s_{j_o h_o}-s_{\ell k}). \quad (15)$$

L'indice d'association entre les deux objets x et y, relativement à la variable en question est égal à

$$S(x,y)=\frac{\sum_{1\leq \ell\leq k\leq m}\sigma_{\ell k}(s_{j_o h_o}-s_{\ell k})}{\Big\{\sum_{j\leq k}\sigma_{jh}\Big(\sum_{\ell\leq k}\sigma_{\ell k}(s_{jh}-s_{\ell k})\Big)^2\Big\}^{1/2}}. \quad (16)$$

Nous allons à présent établir une propriété d'invariance de cet indice lorsque le nombre de modalités de la variable qualitative est deux.

Dans ce cas là, on a

$H=\{(1,1),(1,2),(2,2)\}$ et on posera

$s_{12}=p$, $s_{11}=s_{22}=q$, où $p<q$.

Dans ces conditions, la moyenne et la variance de la distribution des indices locaux sont respectivement égaux à

$$\mu=p\sigma+q\rho$$
$$V=p^2\sigma+q^2\rho-(p\sigma-q\rho)^2 \qquad (17)$$

Si les deux objets à comparer possèdent deux modalités distinctes, la valeur de l'indice S s'écrit :

$$\frac{p-(p\sigma+q\rho)}{\sqrt{p^2\sigma+q^2\rho-(p\sigma+q\rho)^2}} \qquad (18)$$

qui -après calcul- se réduit à

$$-\sqrt{\rho/\sigma} \qquad (19)$$

Si les deux objets à comparer possèdent la même modalité, on remplacera le numérateur de (18) par $[q-(p+q)]$ et l'indice S se réduit à

$$+\sqrt{\rho/\sigma}\ . \qquad (20)$$

D'où l'énoncé du résultat :

Propriété. Si le nombre de modalités de la variable qualitative se réduit à deux, l'indice globalement réduit ne dépend plus que de la répartition des deux modalités sur l'ensemble des objets.

De façon précise, si n(1)(resp.n(2)) est le nombre d'objets possédant la modalité 1(resp.2), deux objets x et y possédant respectivement les modalités 1 et 2, ont pour valeur de l'indice S :

$$S(x,y)=-\sqrt{\frac{1}{2}\left[\frac{n(1)-1}{n(2)}+\frac{n(2)-1}{n(1)}\right]}\ . \qquad (19')$$

Si par contre les deux objets ont la même modalité :

$$S(x,y)=+\sqrt{2\ \frac{n(1)n(2)}{n(1)[n(1)-1]+n(2)[n(2)-1]}}\ . \qquad (20')$$

Cette propriété qui peut surprendre est en fait heureuse et naturelle : la perception des ressemblances mutuelles entre objets pris dans un même ensemble, face à une simple dichotomie, n'a plus à dépendre d'une "quantification" de cette dichotomie.

Considérons deux objets face à plusieurs variables dichotomiques. Les formules (19') et (20') montrent que la contribution d'une même variable à la ressemblance des deux objets qui en possèdent la même modalité (resp. qui n'en possèdent pas la même modalité) est d'autant plus élevée que les deux modalités se trouvent plus également réparties.

La propriété d'invariance ci-dessus -par rapport à la table (4) des similarités locales- n'est plus valable si la variable a plus de deux modalités.

Pour la plupart des applications, une information très générale de type "préordonnance" sur H est suffisamment fine pour une excellente reconnaissance des classes de proximité sur l'ensemble des objets.

### II.3. Indice d'association dans le cas de plusieurs variables

Désignons par $S_q(x,y)$ la contribution de la variable $c_q$ à la comparaison des deux objets x et y. $^q S_q(x,y)$ est donné par la formule $(16)^q$ ci-dessus relativement à la table (4).

Pour un même q, la moyenne et la variance de $S_q$ sur l'ensemble $P_2(E)$ sont respectivement égales à 0 et à 1.

Pour la définition de l'indice d'association, on tiendra également compte des différentes variables en proposant

$$T(x,y)= \frac{1}{\sqrt{Q}} \sum_{1\leq q\leq Q} S_q(x,y). \quad (21)$$

La réduction au moyen de $1/\sqrt{Q}$ se réfère à un modèle d'indépendance où les v.a. associés aux $c_q$, $1\leq q\leq Q$, ont une variance unité.

## III. UN ALGORITHME "PARALLELE" DE CLASSIFICATION HIERARCHIQUE DE "GROS ENSEMBLES"

### III.1. Idée générale et différentes phases de l'algorithme

Il y a plusieurs types de tableaux de données ExV: Objets x Variables (cf.(LERMAN (1981)), chap.2) et l'algorithme que nous allons présenter peut concerner le traitement de l'ensemble des lignes (resp.colonnes) de n'importe quel type de tables de données. Toutefois, nous l'avons effectivement implémenté dans 2 cas : d'abord celui d'un tableau d'incidence où V est formé de variables logiques de présence-absence (v(x)=1 (resp.0) selon que l'objet x possède (resp. ne possède pas) l'attribut défini par la variable v, v∈V), ensuite celui où les variables sont relationnelles (cf.§II).

Le cardinal de l'ensemble E des objets (card(E)) est le plus souvent plus important que celui de l'ensemble V des variables. Card(E) peut atteindre plusieurs milliers alors que card(V) dépasse rarement quelques centaines. De sorte que, bien que la notion de "gros" ensembles soit toute relative et dépende de la configuration informatique dont on dispose, l'algorithme concerne surtout le problème de la classification de l'ensemble E des objets.

On a vu ces dernières années se développer des algorithmes rapides de classification ascendante hiérarchique basés sur une limitation qu'on démontre possible dans la comparaison des indices de proximité -les plus utilisés- entre paires de classes propriété de "contractance" et graphes "réductibles" ((BRUYNOOGHE(1978)), (JAMBU(1978))), détermination des "voisins réciproques" ((RAHM(1980))),... . Ces algorithmes fournissent des résultats "exacts" ; c'est-à-dire, les mêmes que si l'algorithme classique -qui suppose la comparaison de toutes les paires de classes- est pratiqué.

Dans l'évaluation de ces algorithmes dits "rapides", on oublie trop souvent de tenir compte du temps nécessaire à l'établissement de la table des indices de proximité entre éléments de E ou de la préordonnance totale associée alors que pour l'algorithme classique appliqué dans son contexte de tailles "raisonnables" (quelques centaines sur de "grosses" machines), ce temps peut devenir très dominant si le nombre de variables (représentées ici par les colonnes) est "assez grand". Nous avons très clairement expérimenté ce fait en considérant le problème transposé de la classification hiérarchique de l'ensemble des variables où l'ensemble des objets

pouvait atteindre quelques milliers.

L'algorithme que nous développons ici n'est pas exact, il fournit une forme d'"approximation" de l'arbre complet. Chaque feuille de l'arbre construit est définie par une sous-classe homogène d'une grande classe correspondant à un même profil général qui "doit" nécessairement apparaître dans une classification hiérarchique exacte et totale. Ainsi, la structure principale en classes se retrouve -à partir de l'un des premiers niveaux- dans la structure de l'arbre que nous proposons. La qualité de l'approximation est de nature statistique et liée à la classifiabilité naturelle de l'ensemble décrit E.

La méthode que nous allons présenter de "Classification Automatique Hiérarchique Approchée en Parallèle (C.A.H.A.P.)" pourrait d'une certaine manière être considérée comme une illustration de la célèbre formule "diviser pour régner". Elle comporte quatre phases que nous reprenons en détail ci-dessous :

1- L'ensemble E des objets est découpé en sous-ensembles de tailles respectives comparables : E=E1+E2+..+Ej+..+Ek (somme ensembliste). Chaque sous-ensemble définira une tranche.

2- Chaque tranche fait l'objet d'une classification hiérarchique partielle ; c'est-à-dire, qu'on arrête à un niveau dépendant du nombre de classes formées et d'un critère d'adéquation. Une même classe formée à ce niveau de la tranche traitée définira ce que nous appellerons une "sous-classe".
Il est important déjà de remarquer que toutes les tranches peuvent simultanément être organisées en "sous-classes" sur un calculateur parallèle.

3- La troisième phase consiste à classifier -par la même méthode que celle utilisée en 2- l'ensemble de toutes les sous-classes obtenues à partir du traitement des différentes tranches (phase 2 ci-dessus).

4- La quatrième et dernière phase correspond à une édition des résultats acquis dans les phases 2 et 3.

Cet algorithme composé peut a priori être utilisé avec n'importe quel critère de formation ascendante hiérarchique des classes. Toutefois, il importe que le critère n'ait pas une tendance naturelle à former des classes par trop inégales en taille. Un critère tel que celui de la vraisemblance du lien répond parfaitement au problème compte tenu d'un certain équilibre dans les cardinaux des classes qu'il permet de construire ((F. NICOLAU(1980))).

L'idée de la réalisation de ce type d'algorithme remonte pour nous à l'année 1979 où, dans le cadre d'un stage de D.E.A. (C.N.E.T.-Lannion) ((RAPHALEN(1979)), (VALETTE & al.(1980))) le problème s'était posé de typer la charge d'un ordinateur (il s'agissait de l'IRIS 80) et ce, à partir de la classification de plusieurs milliers d'"unités de travail", caractérisées par différents paramètres d'utilisation de la configuration informatique.

De façon tout à fait indépendante et dans un tout autre contexte Madame Escofier (ESCOFIER(1979)) a étudié et traité le problème de l'approximation d'une analyse factorielle des correspondances portant sur un "grand" tableau à partir d'analyses partielles de même type portant chacune sur un sous-tableau, où l'ensemble des sous-tableaux définit une partition du tableau complet. Toutefois, les problèmes d'approximation posés dans le cadre factoriel sont de nature (technique ou statistique) très différente.

Comme nous l'avons signalé ci-dessus, nous allons reprendre et expliciter les quatre phases de notre algorithme.

III.2. Découpage en tranches

Compte tenu des possibilités informatiques de traitement, l'utilisateur du programme fournit le nombre m d'objets par tranche. Supposons que I={1,2,...,i,...,n} indexe la suite des éléments. Conformément à la division n=m*(k-1)+l, la j-ème tranche est formée des objets d'indices (m*(k-1)+1) à m*j, j=1,2,...,(k-2). La dernière tranche, formée des objets d'indices (m*(k-1)+1) à n, comporte $l$ termes et $l$ peut être inférieur à m. Toutefois, en jouant sur le diviseur m, on s'arrangera pour -des raisons de précision statistique- que la dernière tranche soit d'une taille $l$ "très comparable" à m.

Le choix des objets devant rentrer dans la composition d'une même tranche doit correspondre à un échantillon aléatoire exhaustif de la population P dont provient E. C'est automatiquement le cas si la suite (ei/1≤i≤n) des objets formant E est construite de telle sorte que ei (1≤i≤n) soit choisi indépendamment de (e1,e2,...,e(i-1)) uniformément au hasard dans (P-{e1,e2,...,e(i-1)}).

Relativement à cette décomposition aléatoire, des problèmes intéressants se présentent concernant la stabilité statistique de la classification hiérarchique de l'ensemble des sous-classes.

III.3. Classification hiérarchique partielle d'une tranche D=Ej

Le critère de base est celui de la vraisemblance du lien maximal (LERMAN(1970), (1981)). La programmation mise en oeuvre (LERMAN et PETER(1984)) profite du logiciel élaboré dans le cadre de MODULAD (LEREDDE et LERMAN(1983)). Toutefois, dans cette phase on arrête le déroulement de l'algorithme à un certain niveau dont nous préciserons bientôt la règle de détermination. Nous n'avons en effet besoin que d'une partition en sous-classes.

III.3.1. Règle d'arrêt

Des paramètres fournis par l'utilisateur se déduit le nombre maximum total NCL de sous-classes à retenir à partir de la classification ascendante hiérarchique d'une même tranche. On dispose d'autre part d'un critère très général C(π,p(D)) mesurant l'adéquation entre une partition π sur D et une structure de proximité p(D) -de caractère numérique ou ordinal- sur D (LERMAN(1983)), mais que nous ne pouvons rappeler ici par manque de place. Dans ces conditions, la règle adoptée est la suivante:

On désigne par NCP1 (resp.NCP2) le premier niveau de l'arbre pour lequel le nombre de classes de la partition obtenue est inférieur ou égal à NCL (resp.(NCL-t)) où t est un entier "assez petit" devant NCL. Dans le programme implémenté, on a travaillé avec différentes valeurs de NCL (120,350,...). On a bien entendu NCP1<NCP2. Le niveau de coupure NCP est alors choisi, compris entre NCP1 et NCP2 (NCP1≤NCP≤NCP2), de façon à optimiser le critère C(π,p(D)).

Cette technique permet de retenir la partition la plus "naturelle" dont le nombre de classes est inférieurement le plus voisin de NCL.

III.4. Associations entre "sous-classes"

III.4.1. Chaque sous-classe est représentée par un noyau formé d'un seul élément

C'est la solution la plus simple et qui a montré une très bonne efficacité dans le cadre de la méthode employée de la "Vraisemblance du Lien". Rappelons que l'algorithme de classification ascendante hiérarchique suppose la définition sur l'ensemble C à organiser d'un indice qui se réfère à une échelle (0,1) de probabilité (ou de fréquence mathématique) et dont la valeur reflète le complément à 1 d'un "degré

d'invraisemblance" de la relation observée.

Si {P(x,y)/{x,y} ∈ P2(C)} est la table des valeurs d'un tel indice d'association sur l'ensemble P2(C) des paires d'éléments distincts de C et si G et H représentent deux parties disjointes (i.e. deux classes) de C, l'indice de proximité entre G et H se met sous la forme :

$$(\max\{P(x,y)/(x,y) \in GxH\})^{Card(G)*Card(H)} \quad (1)$$

Pour des raisons de précision de calcul et compte tenu du fait que seule l'échelle ordinale induite par (1) importe pour la formation de l'arbre, on considère la fonction strictement croissante -Log -Log(.) ; ce qui donne pour le transformé de l'indice :

$$-Log[card(G)]-Log[card(H)]-Log[-Log(\max\{P(x,y)/(x,y) \in GxH\}) \quad (2)$$

Au paragraphe III.3. ci-dessus, l'ensemble C se trouvait défini par une tranche D=Ej du "gros" ensemble E à classifier, alors qu'ici C est l'ensemble des représentants des différentes "sous-classes" obtenues par les classifications hiérarchiques partielles des tranches.

III.4.1.1. - Choix du "meilleur" représentant d'une sous-classe

Compte tenu du fait que dans "A.V.L." nous travaillons avec un indice de proximité entre éléments (de l'ensemble à classifier) ayant une forme corrélative (avant la référence à une échelle de probabilité) et en se souvenant que la nature du critère maximisé en Analyse en Composantes Principales est une somme des carrés de corrélations, nous choisirons comme "meilleur" représentant d'une sous-classe, l'élément dont la somme des carrés des indices d'association avec les autres éléments de la sous-classe, est maximale.

Dans l'algorithme implémenté, chaque représentant est affecté du poids-cardinal unité. Mais rien n'empêche à ce qu'il puisse être affecté du poids cardinal de la sous-classe qu'il représente. Dans ce cas, la table des indices d'association entre représentants (de sous-classes) fournie à l'A.V.L. sera accompagnée de la table de cardinaux des sous-classes et sera -après transformation conformément à la formule (2) ci-dessus- considérée comme la table des indices d'association entre sous-classes.

Plus précisément, si les indices d'association entre représentants sont donnés -dans le cadre d'une table- sous la forme (21) (§II.3.) on commencera par une "réduction globale des similarités" où on substitue à la table

$$\{T(x,y)/\{x,y\} \in P2(C)\}, \quad (4)$$

où P2(C) est l'ensemble des parties à deux éléments de C, celle

$$\{Q(x,y)= \frac{T(x,y)-\bar{T}}{\sqrt{var(T)}}/\{x,y\} \in P2(C)\} \quad (5)$$

où $\bar{T}$ et var(T) sont la moyenne et la variance de la distribution observée (4).

A partir de (5), on obtient :

$$\{P(x,y)= \Phi[Q(x,y)]/\{x,y\} \in P2(C)\} \quad (6)$$

où $\Phi$ est la fonction de répartition de la loi normale centrée et réduite.

Soient D1 et D1' deux sous-classes quelconques représentées par x et x'. Si on veut tenir compte des cardinaux des sous-classes, dans leurs comparaisons mutuelles au moyen de leurs représentants respectifs, on remplacera la dernière partie de l'expression (2) par {-Log et la première partie par {-Log(card(D1))-Log(card(D1'))}.

Encore une fois, pour l'algorithme implémenté la suite se passe comme si card(D1) et card(D1') sont égaux à 1.

III.4.1.2. Usage du critère de l'inertie expliquée

Nous avons déjà signalé que -à la différence de l'algorithme du lien unique("single linckage")- l'A.V.L. ne subissait pas le phénomène bien connu du "chaînage" et avait une tendance naturelle à fournir des classes de tailles respectivement équilibrées. Toutefois, on peut parfaitement envisager le même processus avec d'autres critères. Le plus connu est sans doute celui de l'inertie expliquée (WARD(1963)) qui suppose une représentation euclidienne de l'ensemble des unités de données, ce qui ne peut être le cas dans le problème qui nous concerne ici.

III.4.2. Chaque sous-classe est prise globalement

C'est une solution qui peut être envisagée avec l'un ou l'autre des deux critères principaux ci-dessus présentés. La table des indices d'association entre "sous-classes" d'une même tranche est déjà établie dans les classifications hiérarchiques partielles. Il reste à organiser et à établir la table des indices d'association entre "sous-classes" de tranches distinctes. Pour cela, on aura besoin à un moment donné de garder en "mémoire centrale" la table de croisement entre deux tranches organisées chacune en "sous-classes". Le croisement d'une sous-classe G de l'une des tranches et d'une sous-classe H de l'autre tranche permet de déterminer la valeur de la quantité critère (2).

III.5. Présentation générale des résultats

III.5.1. Niveaux significatifs de l'arbre des classifications

Nous avons l'habitude dans une analyse "fine" de distinguer deux notions (cf. par exemple (LERMAN(1981) ou (1983))). La première concerne les "niveaux significatifs" et la seconde, les "noeuds significatifs". Les niveaux significatifs correspondent aux maxima locaux de la distribution observée sur la suite des niveaux d'une statistique "globale" qui évalue l'adéquation de toute une partition. Alors que les noeuds significatifs correspondent aux maxima locaux du taux d'accroissement de la statistique globale entre un niveau et le suivant.

Dans la présentation qu'on propose ici seuls les niveaux significatifs sont retenus.

III.5.2. Présentation de l'arbre "global" des classifications

Dans cet arbre général, chaque feuille se trouve définie par un sous-arbre obtenu à la phase deux (cf. paragraphe I et III). Ce sous-arbre se trouve représenté sur un seul niveau, mais où se trouvent marqués les numéros des niveaux des agrégations successives (cf. dans la figure ci-dessous le premier niveau du dernier arbre représenté).

Au delà de ce premier niveau qui reprend toutes les sous-classes "organisées", il est nécessaire d'augmenter artificiellement d'une constante, les niveaux de l'Arbre de Classification des Sous-Classes (ACSC) (i.e. des représentants), pour éviter

toute confusion entre les niveaux des sous-arbres et ceux de l'ACSC. Cette constante doit être un majorant de plus haut niveau atteint par un sous-arbre. On peut carrément prendre le cardinal d'une tranche complète. Finalement, pour le dessin de l'arbre global, les niveaux retenus sont :

- le plus haut niveau de l'arbre global,

- les niveaux significatifs de l'ACSC ; mais où se trouvent marqués latéralement les agrégations entre deux niveaux significatifs,

- un niveau fictif correspondant à un majorant du plus haut niveau atteint par un sous-arbre définissant une sous-classe, de manière à ce que les classes formées au premier niveau de l'arbre représenté sur le dessin correspondent aux sous-classes.

Les autres niveaux de l'arbre représenté sont réhaussés -pour le dessin de l'arbre- au niveau immédiatement supérieur retenu.

Précisons que l'édition de l'arbre s'effectue à partir de sa représentation polonaise où on distingue des entiers négatifs correspondant à des niveaux d'agrégation entre deux classes et des entiers positifs correspondant aux codes des feuilles.

## IV. BANQUE DE DONNEES ; ORGANISATION ET GESTION DYNAMIQUE

### IV.1. Introduction

Rappelons brièvement le problème initial : il s'agit de concevoir un système de gestion automatique des petites annonces, adapté à un type de marché tel que celui de l'immobilier et couvrant une région donnée. Ce système est supposé pouvoir se substituer à terme aux journaux spécialisés. Le système envisagé se propose d'aller au delà d'un simple service de consultation d'une banque de données en offrant un service plus "intelligent" qui suppose la compréhension des annonces telles qu'elles se trouvent formulées par les utilisateurs. Une fois l'annonce de l'utilisateur "comprise" et codée, il s'agira de puiser dans une banque d'annonces afin d'extraire les annonces présumées intéressantes, formant une classe d'"annonces complémentaires".

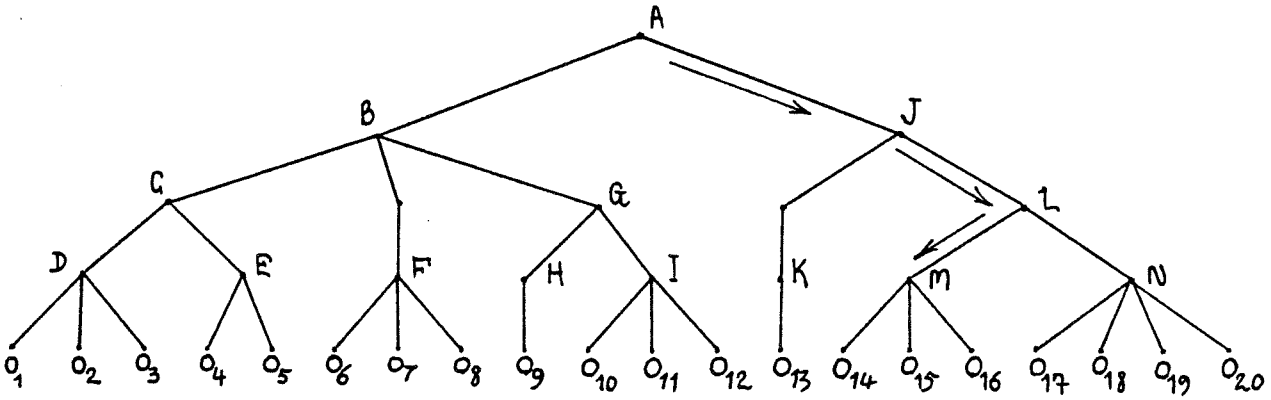
### IV.2. Construction de l'arbre de référence

La structure d'arbre est particulièrement adaptée au stockage d'un ensemble à organiser. Cette structure possède en outre l'avantage d'offrir un accès très rapide à n'importe quelle feuille (en O(Log(f)) si f est le nombre de feuilles).

Or nous avons vu -au paragraphe III précédent- comment construire un arbre de classification hiérarchique répondant aux nécessités du problème posé. On rappellera à cet égard le caractère judicieux de l'usage de l'Algorithme de la Vraisemblance du Lien dans le cadre d'une classification hiérarchique en parallèle ; ce qui permet d'obtenir des classes cohérentes et de taille "raisonnable".

Notre idée est alors de nous servir de cet arbre pour définir la base de données qui structure la banque de données. Toutefois, nous nous sommes vite rendus compte de la non pertinence de l'usage de tous les niveaux de l'arbre général dont les feuilles sont des sous-classes (cf.§III) et ce, d'autant que les agrégations survenant dans les derniers niveaux sont fragiles. Nous avons alors utilisé l'arbre condensé aux niveaux les plus "significatifs" qui correspondent aux maxima locaux de la statistique globale des niveaux. L'arbre condensé construit -à partir de l'algorithme de classification hiérarchique en parallèle (utilisant le critère de la vraisemblance du lien)- sur un échantillon de quelques milliers d'annonces, va déterminer une "estimation" au sens statistique du terme de la structure en arbre du marché.

Cette structure-arbre sera estimée dès lors qu'on aura représenté chacun des noeuds de l'arbre (e.g. A,B,...,M et N) de l'arbre ci-dessous figuré), à partir de la classe d'annonces codées qu'il soustend. On retiendra alors une "ossature" d'arbre dont chaque noeud possède une représentation. Cette structure servira alors directement pour la gestion dynamique de la banque de données. Il faut comprendre qu'une fois déterminée (la structure d'arbre munie d'une représentation des noeuds et feuilles), il devient indifférent de vider la banque de données du corpus ayant servi à la déterminer, avant son chargement effectif par la population totale étudiée, des petites annonces. Ce chargement et la consultation sont réalisés au moyen d'un algorithme d'affectation dont nous allons ci-dessous (§IV.3) préciser les modalités. Il reste -pour clore ce sous-paragraphe- à préciser la représentation qu'on considère d'une classe d'annonces codées.




Représentation d'une classe sous-tendue par une feuille ou un noeud

La représentation d'une même classe d'annonces (sous-tendue par une feuille ou un noeud) se fait très simplement en considérant la distribution empirique de chacune des variables qualitatives $c_q (1 \leq q \leq Q)$ (cf.§II) sur la classe d'annonces. Plus précisément, relativement à un noeud X (X=A,B,...,N dans l'exemple ci-dessus figuré), on établit la table suivante :

| $c_1$ | | $c_q$ | | $c_Q$ |
|---|---|---|---|---|
| $f^X_{11},\ldots,f^X_{1j_1},\ldots,f^X_{1m_1}$ | $\ldots$ | $f^X_{q1},\ldots,f^X_{qj_q},\ldots,f^X_{qm_q}$ | $\ldots$ | $f^X_{Q1},\ldots,f^X_{Qj_Q},\ldots,f^X_{Qm_Q}$ |

où $f^X_{qj_q}$ $(1 \leq j_q \leq m_q, 1 \leq q \leq Q)$ désigne la proportion d'annonces de la classe sous-tendue par X qui possèdent la modalité $j_q$ de la variable $c_q$.

IV.3. Affectation ; descente dans l'arbre

Pour enregistrer une annonce dans la base, il suffit de l'attribuer à la sous-classe représentée par un noeud terminal qui lui "ressemble" le plus. Ce noeud terminal est représenté dans la figure ci-dessus par D,E,F,H,I,K,M ou N. D'où la nécessité d'un indice d'association entre un élément et une classe qui correspondra pour nous à un noeud. Cet indice tiendra compte de la représentation ci-dessus d'une même classe.

Indice d'association entre une classe (ou noeud) et un seul élément

En désignant par O l'élément unique et par X le noeud, l'indice mis au point et qui s'inspire des considérations de calcul du paragraphe II ci-dessus, s'écrit

$$\mathcal{J}(O,X) = \sum_{1 \leq q \leq Q} \{ (( \sum_{1 \leq j_q \leq m_q} f^X_{qj_q} \times r_q(j_q, o_q)) - \mu_q)/\sqrt{V_q} \}. \quad (1)$$

Dans cette formule, il reste à préciser que $o_q$ est la modalité possédée par O pour la variable $c_q$, $\mu_q$ et $V_q$ sont la moyenne et la variance de (2) (du paragraphe II)

sur l'échantillon d'annonces qui servi à établir l'arbre (cf. formules (9') et (13) (§II)) et de rappeler que $r_q(j_q,o_q)$ est un rang calculé conformément à ce qui a été exprimé au paragraphe II.1. ci-dessus. Dans un tel indice d'association, la normalisation s'effectue en rapportant -pour chaque variable $c_q$- la moyenne observée des indices locaux $\{r_q(j_q,o_q)\}$ entre o et les éléments de X, à la distribution établie sur l'échantillon de la table (2) du paragraphe II.

La méthode consiste alors à parcourir l'arbre de façon descendante, depuis sa racine jusqu'à une sous-classe terminale. Plus précisément, à chaque aiguillage (noeud non terminal) on choisit la branche qui mène au noeud dont l'indice (1) d'affectation avec l'élément à attribuer, est le plus grand. Ainsi, sur l'illustration ci-dessus, à partir de la racine A, on se dirige vers J parce qu'on suppose $\mathcal{J}(O,J) > \mathcal{J}(O,B)$ et ainsi de suite jusqu'à M où O sera affecté. Comme nous l'avons déjà ci-dessus mentionné, une telle procédure de recherche descendante est extrêmement rapide puisque est en log (nombre de feuilles de l'arbre), en rappelant que chacune des feuilles est en fait une "sous-classe".

IV.4. Consultation/Extraction/Réactualisation

IV.4.1. Consultation

C'est l'aspect le plus important. La technique est la même que pour l'enregistrement à cela près qu'on suppose que la base de données a été chargée -conformément à IV.3 ci-dessus- compte tenu de l'état du marché. On dispose donc maintenant d'une banque de données organisée par la structure de l'arbre ci-dessus dont les noeuds conservent la même représentation déjà établie par l'échantillon (cf.§IV.2.). Toutefois, dans cette recherche descendante, on aura soin :

- d'inverser le sens de la transaction de l'annonce ; en effet, si par exemple un utilisateur veut vendre, il y a lieu de lui proposer des acquéreurs et non des vendeurs !

- pour des raisons évidentes de stratégie commerciale, on s'interdira de descendre vers un noeud sous-tendant une classe qui se trouve vidée après la procédure d'affectation (cf.§IV.3.) et -éventuellement- un certain nombre de consultations.

En effet, au bout de la recherche descendante, on proposera à l'annonce sollicitatrice la classe d'annonces de la banque que supporte la feuille terminale qui a pu réaliser le maximum de (1). Encore une dernière fois, insistons sur l'extrême rapidité dans la réponse fournie à l'utilisateur (en log (nombre de feuilles-sous-classes)).

IV.4.2. Extraction d'une annonce

Cette opération ne présente aucune difficulté, il suffit de retirer l'annonce de la liste chaînée de la classe supportée par une feuille terminale à laquelle elle appartient.

IV.4.3. Réactualisation de la base

Ce problème réel est le plus délicat et mérite une étude expérimentale. Il s'agit en effet de déterminer la durée pendant laquelle, l'arbre structurant la banque de données -établi à partir d'un échantillon- reste représentatif de l'état du marché. Au bout de cette durée à déterminer compte tenu de la stabilité de la structure de l'offre et de la demande, on aura besoin de reprendre un échantillon représentatif pour bâtir une nouvelle structure d'arbre permettant la redistribution de la banque.

Qu'il nous soit permis de terminer en exprimant nos plus vifs remerciements à Monsieur L. TRILLING (professeur au département de Mathématiques et Informatique de l'Université de RENNES I et responsable de recherche à l'IRISA) pour les larges discussions que nous avons pu avoir avec lui et dont chacune enrichissait nos associa-

tions d'idées. C'est ainsi que nous avons été stimulés pour développer l'indice du paragraphe II et que nous avons été conduits au point de vue exprimé au paragraphe IV.

BIBLIOGRAPHIE

M. BRUYNOOGH (1978) "Classification ascendante hiérarchique de grands ensembles de données : un algorithme rapide fondé sur la construction des voisinages réductibles", Cahiers de l'Analyse des Données, vol. III, numéro 1, 7-33.

B. ESCOFIER (1979) "Stabilité et approximation en analyse factorielle", Thèse d'état, 8.10.1979, Université Paris 6.

M. JAMBU (1978) "Classification automatique pour l'analyse des données", tome 1, Dunod, Paris.

H. LEREDDE & I.C. LERMAN (1983) "CAHM1 : Méthode de classification hiérarchique", MODULAD, Bibliothèque Fortran pour l'Analyse des Données, Version 1.0.

I.C. LERMAN (1970) "Sur l'analyse des données préalable à une classification automatique. Proposition d'une nouvelle mesure de similarité", Revue Mathématique et Sciences Humaines, 8ème année, numéro 32.

I.C. LERMAN (1981) "Classification et analyse ordinale des données", Dunod, Paris.

I.C. LERMAN (1983) "Sur la signification des classes issues d'une classification automatique", NATO ASI Series, vol. G1, Numérical Taxonomy. Edited by J.Felsenstein, Springer-Verlag.

I.C. LERMAN et Ph. PETER (1984) "Analyse d'un algorithme de classification hiérarchique "en parallèle" pour le traitement de gros ensembles", PI IRISA n° 232, Rennes et Rap. Recherche INRIA Rennes n° 339.

I.C. LERMAN et Ph. PETER (1985) "Elaboration et Logiciel d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification". Rap. Int. IRISA (sous presse).

S. CHAH (1984) "Agrégation des préordonnances", Etude F-063, Centre Scientifique IBM de Paris, Mai 1984.

F. NICOLAU (1980) "Criterios de analise classificatoria hierarquica baseados na funçao de distribuiçao", Laboratoire de Statistique, Faculté des Sciences de Lisbonne.

C. de RAHM (1980) "La classification hiérarchique ascendante selon la méthode des voisins réciproques", Cahiers de l'Analyse des Données, vol. V, numéro 2, 135-144.

M. RAPHALEN (1979) "Caractérisation de la charge du calculateur IRIS 80 du C.N.E.T. Lannion, Classification des travaux soumis". Rapport de D.E.A., Université de Rennes I.

N. VALETTE, A. DUPUIS & A. LELIEVRE (1980) "Application des techniques d'analyse des données à la caractérisation de la charge de l'IRIS 80", Rapport C.N.E.T. RP/LAA/STI/15.

J.H., Jr. WARD (1963) "Hierarchical grouping to optimise an objective function", JASA, 58, 236-244.