THESE

présentée à

L'UNIVERSITE DES SCIENCES ET TECHNOLOGIES DE LILLE

pour l' obtention du titre de

DOCTEUR

en Productique : Automatique et Informatique Industrielle

par

Baher Albert BARSOUM

CLASSIFICATION AUTOMATIQUE PAR AMINCISSEMENT DE L'HISTOGRAMME MULTIDIMENSIONNEL

Soutenue le 17 Juin 1994 devant la commission d'examen :

Messieurs

P.	*VIDAL*	*Président*	*Professeur à l'U.S.T.L.*
H.	*EMPTOZ*	*Rapporteur*	*Professeur à l'INSA de Lyon.*
J.P.	*ASSELIN DE BEAUVILLE*	*Rapporteur*	*Professeur à l'Université F. Rabelais de Tours.*
J.G.	*POSTAIRE*	*Directeur de thèse*	*Professeur à l'U.S.T.L.*
Y.	*EL GAMAL*	*Examinateur*	*Directeur école d'Ingénieur AMTA-Alexandrie Egypte.*
X.	*LOUCHEUR*	*Examinateur*	*Co directeur de la formation. G.H.C.*
D.	*BOURGOIS*	*Examinateur*	*Responsable de la cellule Recherche - Qualité Innovation - CCI Valenciennes.*

SOMMAIRE

CHAPITRE I
LA CLASSIFICATION

CHAPITRE II
L'HISTOGRAMME : UN MODE DE REPRESENTATION PRIVILEGIE DES DONNEES

CHAPITRE III
CONSTRUCTION D'UN HISTOGRAMME MULTIDIMENSIONNEL PAR BALAYAGE SPATIAL DE L'ESPACE DISCRETISE

CHAPITRE III
CONSTRUCTION D'UN HISTOGRAMME MULTIDIMENSIONNEL PAR BALAYAGE SPATIAL DE L'ESPACE DISCRETISE

CHAPITRE IV
L'AMINCISSEMENT DE L' HISTOGRAMME

CONCLUSION GENERALE

NOTATIONS UTILISEES

N ➞ Dimension de l'espace de représentation des données.

Q ➞ Taille de l'échantillon.

K ➞ Nombre de classes de l'échantillon.

C_k ➞ Classe k, $k = 1, 2, ..., K$.

D ➞ Pas de discrétisation de l'espace de représentation des données.

$C(n)$ ➞ Largeur du pas de discrétisation sur l'axe n de l'espace de représentation des données, $n = 1, 2, ..., N$.

X ➞ Vecteur des attributs, $X = [x_1 ; x_2 ; ... ; x_n ; ... ; x_N]^T$.

X_q ➞ Vecteur des attributs de l'observation q, $q = 1, 2, ..., Q$; ou l'observation multidimensionnelle, $X_q = [x_{1,q} ; x_{2,q} ; ... ; x_{n,q} ; ... ; x_{N,q}]^T$.

x_n ➞ Vecteur de l'attribut n, $x_n = [x_{n,1} ; x_{n,2} ; ... ; x_{n,q} ; ... ; x_{n,Q}]^T$.

$x_{n,q}$ ➞ Valeur de l'attribut n pour l'observation q.

$max_q(x_n,q)$ ➞ Valeur maximale de l'attribut x_n.

$min_q(x_n,q)$ ➞ Valeur minimale de l'attribut x_n.

$p(X)$ ➞ Fonction de densité sous-jacente à la distribution de l'échantillon.

$p(X \mid C_k)$ ➞ Fonction de densité de probabilité conditionnelle de la classe C_k.

$P(C_k)$ ➞ Probabilité a priori de la classe C_k.

$P(C_k \mid X_q)$ ➞ Probabilité a postériori de la classe C_k connaissant l'observation X_q.

S ➞ Matrice de covariance.

$\overline{X}$ ➞ Vecteur moyenne.

S_k ➞ Matrice de covariance de la classe C_k.

$\overline{X}_k$ ➞ Vecteur moyenne de la classe C_k.

$\overline{x}_n$ ➞ Moyenne de l'attribut x_n.

σ_{x_n} ➞ Ecart-Type de l'attribut x_n.

$\sigma^2_{x_n}$ ➞ Variance de l'attribut x_n.

Φ ➞ Fonction noyau de Parzen.

h_q ➞ Largeur du noyau de Parzen.

V_q ➞ Volume du domaine de Parzen et de centre l'observation X_q.

G ➞ Nombre des plus proches voisins.

$D_r(X_q)$ ➞ Domaine de volume unité et de centre l'observation X_q.

$D_r(X_q, w)$ ➞ Domaine d'observation homothétique de $D_r(X_q)$ dans une homothètie de centre X_q de rapport w.

D_k ➞ Domaine de concavité de la classe C_k.

$\overline{D_k}$ ➞ Domaine modal de la fonction de densité du mélange.

μ_{qk} ➞ Degré d'appartenance de l'observation X_q à la classe C_k.

A ➞ Adresse multidimensionnelle, $A = (a_1, a_2, ..., a_n, ..., a_N)$.

a_n → Elément n de l'adresse multidimensionnelle A, n = 1, 2,..., n, ..., N.

HP(A) → Hyperparallélépipède ou cellule d'adresse multi-dimensionnelle A.

S(A) → Sommet d'adresse multidimensionnelle A d'un hyperparallélépipède ou cellule.

C_e → Centre d'un hyperparallélépipède ou cellule.

S_o → Sommet d'un hyperparallélépipède ou cellule.

H(A) → Valeur de l'histogramme à l'adresse multidi-mensionnelle A d'un hyperparallélépipède ou d'un sommet.

$H_{Ce}(A)$ → Valeur de l'histogramme ou nombre d'observations assignées à l'hyperparallélépipède d'adresse multi-dimensionnelle A.

$H_{So}(A)$ → Valeur de l'histogramme ou nombre d'observations assignées au sommet d'adresse multidimensionnelle A.

$\propto$ → Adresse unidimensionnelle.

$HP(\propto)$ → Hyperparallélépipède ou cellule d'adresse uni-dimensionnelle $\propto$.

$S(\alpha)$ ➞ Sommet d'adresse unidimensionnelle α d'un hyperparallélépipède ou cellule.

$H(\alpha)$ ➞ Valeur de l'histogramme à l'adresse unidimensionnelle α d'un hyperparallélépipède ou d'un sommet ou cardinal du sous-ensemble des observations $O(\alpha)$.

$H_{Ce}(\alpha)$ ➞ Valeur de l'histogramme ou nombre d'observations assignées à l'hyperparallélépipède d'adresse unidimensionnelle α.

$H_{So}(\alpha)$ ➞ Valeur de l'histogramme ou nombre d'observations assignées au sommet d'adresse unidimensionnelle α.

$O(d)$ ➞ Sous-ensemble des observations assignées au rectangle d'adresse d d'un histogramme unidimensionnel.

$H(d)$ ➞ Valeur d'histogramme ou cardinal du sous-ensemble des observations $O(d)$.

$H(d_m)$ ➞ Valeur d'histogramme ou cardinal des observations du mode d'adresse d_m pour un histogramme unidimensionnel.

$O_{Ce}(d)$ ➞ Sous-ensemble des observations assignées au rectangle d'adresse d d'un histogramme unidimensionnel bâti sur la grille des centres.

$H_{Ce}(d)$ ➞ Valeur d'histogramme ou cardinal du sous-ensemble des observations $O_{Ce}(d)$.

$H_{Ce}(d_m)$ ➞ Valeur d'histogramme ou cardinal des observations du mode d'adresse d_m pour un histogramme unidimensionnel bâti sur la grille des centres.

$O_{So}(d)$ ➞ Sous-ensemble des observations assignées au rectangle d'adresse d d'un histogramme unidimensionnel bâti sur la grille des sommets.

$H_{So}(d)$ ➞ Valeur d'histogramme ou cardinal du sous-ensemble des observations $O_{So}(d)$.

$H_{So}(d_m)$ ➞ Valeur d'histogramme ou cardinal des observations du mode d'adresse d_m pour un histogramme unidimensionnel bâti sur la grille des sommets.

FH(d) ➞ Valeur d'histogramme "fictif".

RH(d) ➞ Valeur d'histogramme "réel".

$FH_{Ce}(d)$ ➞ Valeur d'histogramme fictif bâti sur la grille des centres.

$RH_{Ce}(d)$ ➞ Valeur d'histogramme réel bâti sur la grille des centres.

$FH_{So}(d)$ ➞ Valeur d'histogramme fictif bâti sur la grille des sommets.

$RH_{So}(d)$ ➞ Valeur d'histogramme réel bâti sur la grille des sommets.

$O(\alpha)$ ➞ Sous-ensemble des observations assignées à l'adresse unidimensionnel α d'un hyperparallélépipède ou d'un sommet.

CHAPITRE I

LA CLASSIFICATION

CHAPITRE I
LA CLASSIFICATION

I - 1 . INTRODUCTION

La classification existe vraisemblablement depuis l'origine de la communication humaine. Pour l'être humain, classer est une activité normale et nécessaire pour appréhender le réel, y mettre de l'ordre, organiser son savoir, donc pour survivre dans un environnement parfois hostile.

Classer permet tout à la fois, en groupant les objets naturels, les objets artificiels ou les concepts en ensembles plus ou moins inclusifs, d'établir des procédures de reconnaissance en fonction des données ou de caractères constants ou transitoires, communs ou uniques. L'attribution d'un nom à chacun des groupes reconnus complète le processus. Une classification, quelle que soit la technique utilisée est, en définitive, un moyen pratique de stocker, de structurer et de transmettre une information sélectionnée et codifiée, qui devient ainsi directement utilisable pour peu que la symbolique employée soit commune à celui qui transmet l'information et à celui qui la reçoit.

Depuis l'Antiquité, en suivant des méthodes différentes, la classification se fonde sur l'exploitation d'ensembles d'attributs très variés en utilisant différentes méthodologies.

Au XVIII^{ième} siècle, et jusqu'à une époque peu éloignée, on recourait à des "systèmes" pour classer. Ces systèmes s'apparentent à ce que l'on nomme aujourd'hui des "clés de détermination". Ils consistent à utiliser un petit nombre d'attributs judicieusement choisis pour permettre un arrangement des groupes. En systématique, science des classifications biologiques, l'exemple le plus célèbre est sans doute le "système sexuel" des plantes préconisé par Linné [Linné, 1753]. Il se fondait exclusivement sur l'analyse d'un nombre limité d'attributs liés à l'appareil reproducteur, les étamines et le pistil, pour construire la classification des plantes. Pour la première fois, Linné codifiait sous une forme condensée la description des plantes et des animaux et introduisait la nomenclature binominale en désignant les espèces par un nom double en latin, véritable nom propre à chaque espèce. La classification devenait une méthode pratique pour codifier "l'ordre naturel".

Même si Linné reste le grand nom que l'on retient dans l'histoire de l'utilisation de la classification biologique ou systématique, il était l'héritier d'une longue série d'érudits qui, depuis l'antiquité, notamment avec Aristote, cherchaient, par l'inventaire de la diversité et par l'analyse des traits exhibés par les plantes et les animaux, à établir les relations de toutes sortes qui pourraient exister entre ces derniers. Le but ultime, était bien entendu, de mettre en évidence, au sein de la diversité, un ordre sous-jacent qui traduisit un supposé dessein d'origine divine ou des relations d'une

autre nature, ce qui fut le cas lorsque l'évolution s'imposa comme cause principale de cet ordre. Dans le contexte de ce travail, on supposera que tout objet peut être décrit par un nombre fini de caractères ou attributs, dont les valeurs peuvent-être connues par le biais de capteurs ou d'instruments de mesure. Ces valeurs représentent l'information à partir desquelles la classification consiste à définir des classes.

Le but principal de la classification est donc de condenser les informations multiples observées sur un ensemble d'objets, la description complexe et détaillée de chacun d'eux étant remplacée par son appartenance à une classe bien définie. Cette démarche est relativement difficile à formaliser, surtout quand on se place dans un contexte non-supervisé, c'est-à-dire quand on ne dispose d'aucune information a priori sur la structure de l'ensemble des évènements à classer. Dans certaines situations, il est même difficile de savoir, a priori, si les données sont réllement structurées en classes.

Avant d'aborder la classification automatique proprement dite, il convient de présenter les techniques de classement.

I - 2 . CLASSEMENT ET CLASSIFICATION

L'objectif d'une procédure de classement consiste à classer des données dans des classes connues a priori. Dans cette démarche, on voit apparaître deux notions fondamentales : la notion de donnée et celle de classe.

Dans ce travail nous considérons une donnée comme un ensemble d'attributs mesurés sur un objet et regroupés sous la forme d'un vecteur. En général, un très grand nombre d'attributs peuvent être retenus pour caractériser le même objet. Cependant, en pratique, on n'utilise qu'un nombre restreint d'attributs, jugés pertinents pour le problème à résoudre. Ce choix est essentiel, car il influence fortement la qualité de la discrimination entre objets appartenant à différentes classes [Fukunaga, 1990] et conditionne donc le succès de la procédure de mise en évidence des différentes classes [Fehlauer et Eisenstein, 1978 ; Pudil et Blaha, 1981 ; Lakshminarasimhan et Dasarathy, 1975 ; Leonard et Kilpatirck, 1974 ; Fukunaga et Ando, 1973 ; Eigen, Fromm et Northouse, 1974 ; Chen , 1976 ; Jain et Waller, 1978].

On peut considérer que les données proviennent de plusieurs sources aléatoires, chacune d'elles correspondant à une classe. Une classe est alors un ensemble de données "fortement semblables" entre elles et "relativement dissemblables" des données des autres classes.

Les méthodes de classement reposent sur la connaissance a priori des différentes classes entre lesquelles se répartissent les données à analyser. Cela implique que la liste des classes est supposée complète et donc qu'une observation appartient forcément à une des classes prédéfinies.

Toutefois, dans de nombreuses situations pratiques, l'analyste ne dispose pas d'informations a priori sur les classes en présence dans un ensemble de données à analyser. Dans ce cas, il doit avoir recours à une procédure de classification automatique préalable.

Pour classer des données on procède alors en deux étapes fondamentales et bien distinctes :

- une phase d'apprentissage, ou de classification, dans laquelle, après avoir acquis les données disponibles, on les analyse pour découvrir leur structure et identifier les classes en présence.

- une phase d'exploitation, ou de classement proprement dite, dans laquelle des données collectées ultérieurement sont affectées aux classes mises en évidence à travers la phase d'apprentissage.

I - 3 . L'APPRENTISSAGE PAR CLASSIFICATION

L'apprentissage de la structure des données peut être réalisé de deux manières différentes suivant le degré de connaissance disponible : par classification ou apprentissage supervisé, dit avec professseur, ou par classification ou apprentissage non supervisé, c'est-à-dire sans professeur.

I - 3 - 1 . Classification supervisée ou avec professeur

Elle consiste à analyser un ensemble de données dont on connait la provenance a priori. L'attribution d'une identité à chaque observation est effectuée sous le contrôle d'un professeur qui donne à chacune une étiquette indiquant son appartenance à une classe : c'est l'étiquetage.

I - 3 - 2 . Classification non supervisée ou sans professeur.

Dans ce cas, on dispose de données dont on ignore si elles peuvent être groupées en classes et, si oui, dans quelles classes. On peut se demander quel est l'interêt d'une telle démarche. Il y a trois raisons :

- il est en général difficile et coûteux d'étiqueter un ensemble d'apprentissage sous le contrôle d'un professeur.
- dans plusieurs applications, les caractéristiques des données peuvent changer avec le temps. Ces changements ne peuvent être pris en considération que si la classification est non supervisée.
- enfin, au cours du procesus de classification, il peut apparaître nécessaire de prendre en compte d'autres attributs sur les données afin

de découvrir des sous-classes et de parfaire ainsi leur analyse.

Au cours de la phase d'apprentissage, toute l'information contenue dans les données doit être exploité afin de mettre en évidence la structure de ces données de manière aussi précise que possible.

Cette étape peut-être elle-même divisée en deux étapes
- le prétraitement des données,
- la conception de la règle de décision.

I - 3 - 2 - 1 . Prétraitement des données

Au cours de cette étape, le choix du vecteur-observation est primordial. Ce vecteur doit contenir les attributs susceptibles de mettre en évidence les classes recherchées. Les attributs peuvent être soit des attributs informatifs, ayant un grand pouvoir de discrimination entre les classes et présentant une faible corrélation avec les autres attributs sélectionnés, soit des attributs peu informatifs, c'est-à-dire des attributs dont les valeurs diffèrent peu d'une classe à l'autre.

En analyse de données, un problème difficile à résoudre est la sélection des attributs à retenir [Fehlauer et Eisenstein, 1978 ; Pudil et Blaha, 1981 ; Lakshminarasimhan et Dasarathy, 1975 ; Leonard et Kilpatirck, 1974 ; Fukunaga et Ando, 1973 ; Eigen, Fromm et Northouse, 1974 ; Chen , 1976 ; Jain et Waller, 1978].

I - 3 - 2- 2 . Conception de la règle de décision

Durant cette étape, on conçoit une procédure de classification dont le but essentiel est d'identifier les classes entre lesquelles se répartissent les observations en découvrant des groupements dans l'espace de représentation. Il s'agit donc de mettre en œuvre des démarches exploratoires qui permettent de structurer un ensemble d'observations non étiquetées en classes, de telle sorte que chaque classe ne contienne que des observations dont les attributs sont relativement semblables alors que des observations assignées à des classes distinctes ont des attributs nettement différents.

Le concept de classes est généralement difficile à définir en dehors d'un cadre statistique. C'est la raison pour laquelle nous proposons maintenant une définition statistique d'une classe d'observations.

I - 4 . DEFINITION STATISTIQUE D'UNE CLASSE D'OBSERVATIONS.

I - 4 - 1 . Description statistique d'une classe

Avant de considérer des problèmes multiclasses, nous commençons par analyser les propriétés statistiques de la

distribution des observations provenant d'une seule source aléatoire et donc ne constituant qu'une seule classe.

Lorsque l'on considère des objets décrits par N variables x_1 , x_2 , ..., x_n , ..., x_N représentant chacune un attribut spécifique, ils peuvent être représentés par des points dans un espace Euclidien à N dimensions. Le vecteur des attributs, ou observation multidimensionnelle, est noté X_q , q = 1, 2, ..., Q, Q étant le nombre total d'observations disponibles :

$$X_q = [\, x_{1,q} \; ; \; x_{2,q} \; ; \; ..... \; ; \; x_{n,q} \; ; \; ..... \; ; \; x_{N,q} \,]^T \qquad \text{I-1}$$

Soit p(X) la fonction de densité de probabilité sous-jacente à la distribution des observations qui caractérise la classe considérée.

On définit le vecteur moyenne $\overline{X}$ par :

$$\overline{X} = [\, \overline{x}_1 \, , \overline{x}_2 \, , \, ..., \overline{x}_n \, , \, ..., \overline{x}_N \,] \qquad \text{I-2}$$

Chaque composante de ce vecteur est la moyenne de chaque attribut x_n , n = 1, 2, ..., N, qui peut-être estimée par

$$\overline{x}_n = \frac{1}{Q} \sum_{q=1}^{Q} x_{n,q} \qquad \text{pour } n = 1, 2, ..., N, \qquad \text{I-3}$$

On définit aussi la matrice de covariance S par :

$$S = \| s_{j,k} \| \quad j = 1, 2, ..., N \quad \text{et} \quad k = 1, 2, ..., N \qquad \text{I-4}$$

dont chaque élément $s_{j,k}$ peut-être estimé par la covariance :

$$s_{j,k} = \frac{1}{Q} \sum_{q=1}^{Q} (x_{j,q} - \bar{x}_j)\,(x_{k,q} - \bar{x}_k) \qquad \text{I-5}$$

Cette matrice de covariance est donc symétrique et ses termes diagonaux sont les variances des attributs, carrés des écarts type σ_{x_n} estimés par :

$$\sigma_{x_n}^{\,2} = \frac{1}{Q} \sum_{q=1}^{Q} (x_{n,q} - \bar{x}_n)^2 \qquad \text{I-6}$$

Les fonctions de densité de probabilité p(X) suivent, en général, des lois bien définies, telles que la loi de Bernoulli, la loi binomiale, ou la loi de Poisson, et encore la loi de Laplace-Gauss. Nous nous interessons plus particulièrement à cette dernière qui est l'une des plus utilisées en analyse des données de par sa généralité et sa relative simplicité.

I - 4 - 2 . Loi de Laplace-Gauss ou Loi normale

Quand on dispose d'un ensemble important constitué de Q observations à N dimensions, $X_q = [x_{1,q}; x_{2,q}; ...; x_{n,q}; ...; x_{N,q}]^T$, la forme de la fonction de densité multivariable peut être approchée par :

$$p(X) = \frac{1}{(2\pi)^{N/2} \, |S|^{1/2}} \exp\left[-1/2 \; (X - \overline{X})^T \; S^{-1} \, (X - \overline{X})\right] \qquad \text{I-7}$$

$\overline{X}$ étant le vecteur moyenne des attributs donné par l'équation I-2 et S la matrice de covariance donnée par l'équation I-4 et dont les termes diagonaux sont les variances des attributs données par l'équation I-6.

La loi de Laplace-Gauss prévoit une répartition symétrique des écarts $(X - \overline{X})$ autour du vecteur moyenne des attributs $\overline{X}$. La forme du regroupement des observations autour de $\overline{X}$ est déterminée par la matrice de covariance S. En effet les lieux des points de l'espace d'égale densité sont des hyperellipsoïdes définis par :

$$(X - \overline{X})^T \; S^{-1} \, (X - \overline{X}) = C \qquad \text{I-8}$$

Parmi les domaines d'équidensité, celui défini par :

$$(X - \overline{X})^T \; S^{-1} \, (X - \overline{X}) = 1 \qquad \text{I-9}$$

joue un rôle priviligié. En effet, il a été démontré que la fonction de densité est concave dans ce domaine et que la probabilité qu'une réalisation X_q de la variable aléatoire X soit située dans ce domaine a une valeur indépendante des paramètres de la distribution [Postaire, 1987].

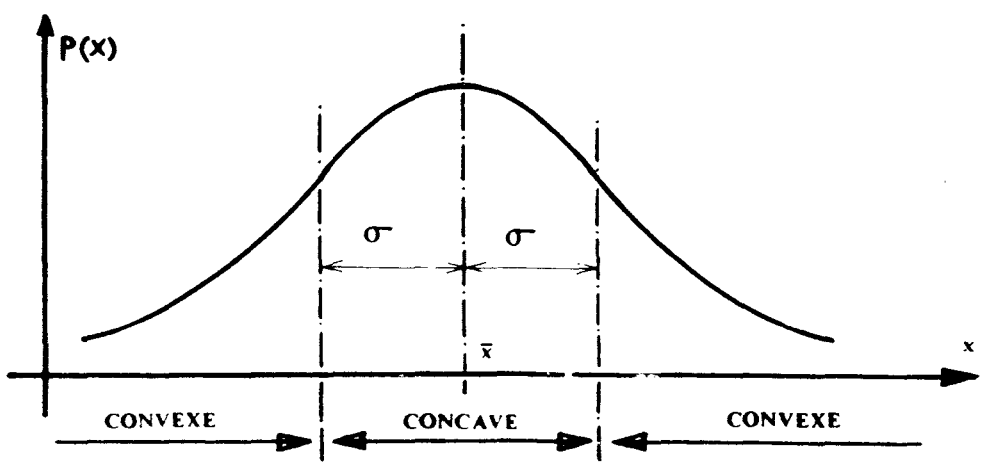
Pour le cas unidimensionnel où N = 1, ce domaine de concavité, noté D_k , est défini à l'intérieur de l'intervalle $[\ \bar{x} - \sigma\ ,\ \bar{x} + \sigma\ ]$ où 68,3 % des observations se groupent autour de $\bar{x}$. La probabilité pour que l'écart $\varepsilon = x_q - \bar{x}$ soit compris entre -E et +E est donnée par la fonction d'erreur

$$\text{erf}\left(\frac{E}{\sigma\sqrt{2}}\right),$$

avec

$$\text{erf}(u) = \frac{2}{\sigma\sqrt{n}} \int_{o}^{u} \exp(-t^2)\, dt \qquad \text{I-10}$$

La figure I.1 illustre cette fonction de densité normale monovariable.

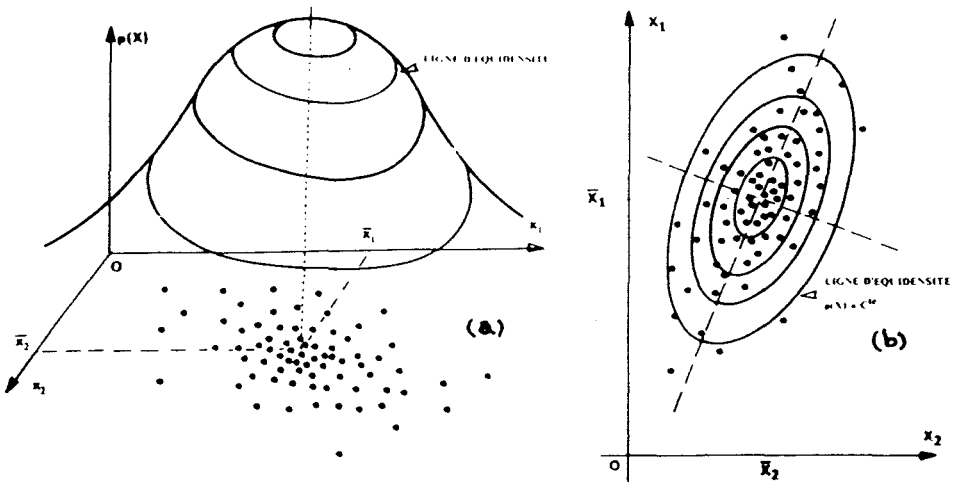


Fonction de densité normale monovariable

Figure I.1

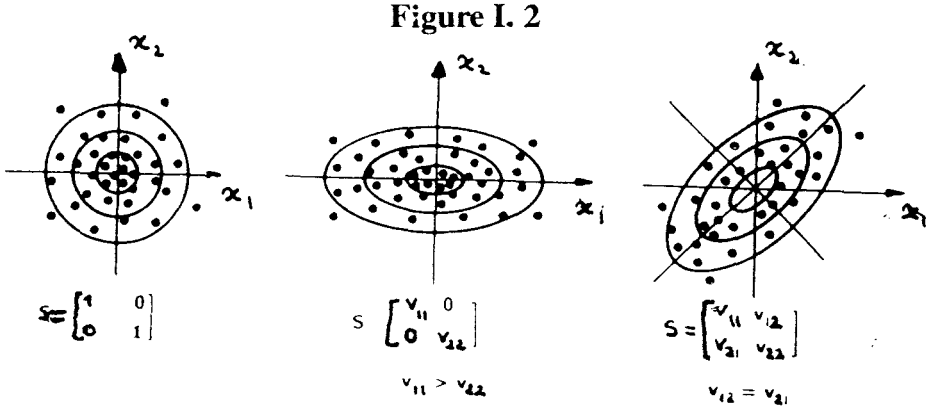
La figure I.2 donne deux représentations des lignes d'égale densité d'une fonction de densité normale bi-dimentionnelle.

La figure I.3 montre comment la forme du groupement des observations autour de la moyenne dépend de la matrice de covariance.



Représentation spatiale des lignes d'équidensité (a) et leur projection (b)

Figure I. 2



Différentes formes de groupements d'observations bidimensionnelles autour du vecteur moyenne en fonction de S.

Figure I.3

Certains auteurs définissent le domaine modal de la distribution comme l'intérieur de l'hyperellipsoïde d'équation $(X - \overline{X})^T \; S^{-1} \; (X - \overline{X}) = 1$, c'est-à-dire le domaine de l'espace R^N où la fonction de densité de probabilité est concave [Postaire, 1987]. La taille et la forme de ce domaine dépendent donc de la matrice de covariance. A partir de la définition de S, on constate que plus les observations sont proches de $\overline{X}$, plus les valeurs propres de S sont petites et plus le domaine modal est de taille réduite.

I - 5 . NOTION DE MELANGE

Supposons maintenant que l'on soit en présence d'un échantillon dont les observations proviennent de plusieurs sources aléatoires et constituent, de ce fait, un mélange de plusieurs classes. La fonction de densité de probabilité conditionnelle $p(X \mid C_k)$ définit la distribution du vecteur X pour la classe C_k, $k = 1, 2, ..., K$, K étant le nombre de classes.

Supposons que l'on connaisse les proportions $P (C_k)$, $k = 1, 2...., K$, de chacune des classes constituant le mélange telles que :

$$\sum_{k=1}^{K} P (C_k) = 1 \qquad \text{I-11}$$

La fonction de densité de probabilité de toutes les observations est alors un mélange pondéré des fonctions de densité associées aux différentes classes en présence. Cette fonction de densité de probabilité du mélange est alors de la forme :

$$p(X) = \sum_{k=1}^{K} p(X \mid C_k)\ P(C_k) \qquad \text{I-12}$$

La connaissance de la valeur du vecteur X_q permet de calculer la probabilité d'être en présence d'une observation de la classe C_k. En effet, d'après la règle de Bayes, cette probabilité, dite probabilité a postériori, s'exprime sous la forme :

$$P(C_k \mid X_q) = \frac{p(X_q \mid C_k)\ P(C_k)}{p(X)} \cdot \qquad \text{I-13}$$

On peut alors décider qu'une observation X_q fait partie de la classe C_i, pour laquelle

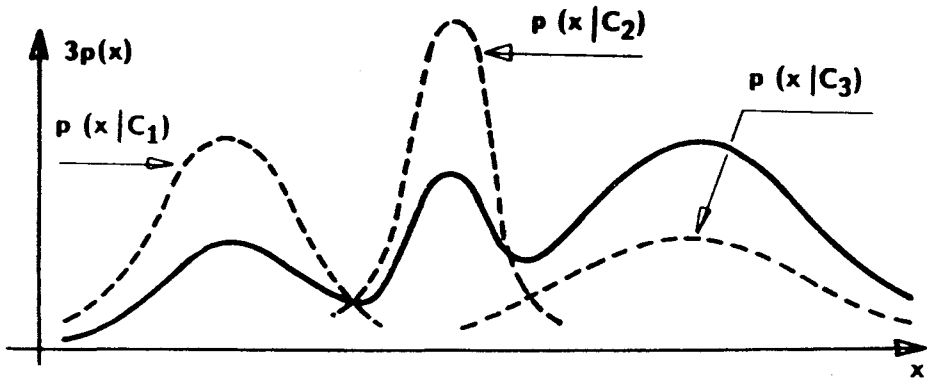
$$P(C_i \mid X_q) = \max \quad k = 1, 2, \ldots, K \qquad \text{I-14}$$

La probabilité d'erreur associée à cette règle de décision est :

$$P(\text{erreur} \mid X_q) = \sum_{\substack{k=1 \\ k \neq i}}^{K} P(C_k \mid X_q) \qquad \text{I-15}$$

Cette règle de décision est optimale en ce sens qu'aucune autre procédure ne peut conduire à un taux d'erreur plus petit. C'est la raison pour laquelle nous comparerons tous les résultats des procédures de classification avec le taux d'erreur théorique minimum correspondant à l'utilisation de cette règle.

Nous illustrons cette notion de mélange sous l'hypothèse normale. La figure I.4 représente les fonctions de densité conditionnelles normales monovariables pour trois classes ainsi que leur mélange donnant la fonction de densité pour toutes les données.



Fonctions de densité conditionnelles normales monovariables pour trois classes ainsi que leur mélange

Figure I.4

I - 6. LES PROCEDURES STATISTIQUES DE CLASSIFICATION PARAMETRIQUES

Dans ce type de méthodes, la forme analytique des modèles de distributions des observations est supposée connue a priori. Sous cette hypothèse paramétrique, le problème de l'analyse de données peut-être ramené à celui de la détermination des paramètres d'un mélange de fonctions de densité représentant les distributions des observations provenant de chacune des classes en présence dans l'échantillon analysé.

En l'absence de connaissance a priori, on ne peut faire que des hypothèses sur les lois de densité de probabilité des classes, puis estimer leurs paramètres. Dans la pratique cette approche ne peut s'appliquer de manière réaliste que sous l'hypothèse gaussienne.

Considérons un échantillon dont les Q observations proviennent de K classes C_k , $k = 1, 2, ..., K$, chacune d'elles étant caractérisée par une fonction de densité de probabilité conditionnelle normale :

$$p(X \mid C_k) = \frac{1}{(2\pi)^{N/2} \mid S_k \mid^{1/2}} \exp\left[-\frac{1}{2} (X - \overline{X}_k)^T S_k^{-1} (X - \overline{X}_k) \right] \quad \text{I-16}$$

où S_k est la matrice de covariance de la classe C_k , et X_k le vecteur moyenne de cette classe.

La fonction de densité sous-jacente à la distribution de l'échantillon disponible est alors un mélange pondéré des lois de probabilité de ces différentes classes de la forme :

$$p(X) = \sum_{k=1}^{K} p(X \mid C_k) \quad P(C_k) \qquad \text{I-17}$$

où $P(C_k)$ est la probabilité a priori de la classe C_k.

En analyse de données, la règle de décision optimale est celle de Bayes pour laquelle la probabilité d'erreur de classification est inférieure à celle obtenue avec toute autre règle de décision. Elle suppose une connaissance totale des lois de densité de probabilité régissant la distribution des observations. Elle nous donne la règle de décision suivante :

X_q est associé à la classe pour laquelle

$$P (C_k \mid X_q) = \max \quad k = 1, 2, \ldots K \qquad \text{I-18}$$

Sous l'hypothèse gausienne, on obtient

X_q est associée à la classe pour laquelle

$$\frac{1}{2} \log \mid S_k \mid + \frac{1}{2} \quad (X_q - \overline{X}_k)^T \quad S^{-1}_k \quad (X_q - \overline{X}_k) \; - \; \log \; [P(C_k)] = \min$$

$$k = 1, 2, \ldots, K \qquad \text{I-19}$$

Cette règle de décision est celle de la procédure de classification "quadratrique par morceaux" [Bezdek et al., 1985].

Lorsque les matrices de covariance des classes sont toutes égales, on obtient la règle suivante :

X_q est associée à la classe pour laquelle

$$\overline{X^T_k}\ S^{-1}_k\ X_q - \frac{1}{2}\ \overline{X^T_k}\ S^{-1}_k\ \overline{X_k} + \log\ [P(C_k)\ ] = \max$$
$$k = 1, 2, ..., K \quad \text{I-20}$$

Cette règle de classification est celle de la procédure de classification "linéaire par morceaux" [Bezdek et al., 1985].

Dans le cas où les domaines modaux $\overline{D_k}$ de la fonction de densité du mélange, c'est-à-dire les domaines où elle est concave, sont très semblables aux domaines de concavité D_k des composantes C_k, $k = 1, 2, ... K$, le dénombrement des classes constituant le mélange peut s'effectuer en dénombrant le nombre de domaines modaux de sa fonction de densité.

De plus, la position des modes du mélange donne une information relativement précise sur la position des composantes. On peut alors admettre que les domaines modaux du mélange apportent de bonnes informations quant aux positions et dispersions des différentes classes constituant l'échantillon.

Cette propriété a été utilisée de manière systématique par Postaire et Vasseur pour identifier des mélanges gaussiens [Vasseur et Postaire, 1980 ; Postaire, 1982].

D'autres auteurs ont également tenté de retrouver la structure d'un mélange par détection des modes de la fonction de densité de probabilité sous-jacente [Asselin de Beauville, 1978 ; 1979 ; 1983 ; 1989 ; Oulad Haj Thamir et Asselin de Beauville, 1982]. Certaines approches proposées dans ce sens s'appuient sur des techniques de recherche de maximum en suivant les lignes de plus grande pente de la fonction de densité. Une solution originale consiste à estimer le gradient de cette fonction directement à partir des observations [Fukunaga et Hostetler, 1975].

Une autre solution consiste à détecter les contours des régions modales de la fonction de densité [Postaire et Touzani, 1990 ; Touzani et Postaire, 1989]. Cette méthode demeure quand même très sensible à toutes les irrégularités que peuvent présenter des distributions réelles car la détection des contours s'appuie sur la mise en œuvre d'opérateurs différentiels très sensibles au bruit.

Pour éviter l'utilisation de tels opérateurs, on peut aborder l'analyse de la fonction de densité dans le cadre de la morphologie mathématique, algèbre des ensembles définis sur des espaces Euclidiens [Matheron, 1985]. Les transformations morphologiques : érosion et dilatation ont été généralisées aux traitements de données multidimensionnelles pour la classification automatique [Botte-Lecocq et Postaire, 1991 a et b ; Postaire, Zhang et Botte-Lecocq, 1993 ; Zhang, Botte-Lecocq et Postaire, 1993].

Ce type d'approche présente un certain nombre d'avantages. Tout d'abord, elle ne nécessite pas le stokage de l'ensemble d'apprentissage. En effet, dans le cas gaussien, le vecteur moyenne et la matrice de covariance d'une classe sont suffisants pour résumer l'information statistique relative à cette classe. De plus, elles sont simples et rapides.

Cependant, il existe trois inconvénients majeurs à leur utilisation. Le premier est la nature paramétrique des règles de décision. En effet, il est extrêmement difficile de vérifier la validité du modèle posé a priori, surtout pour des lois de densité de probabilité gaussiennes multivariables [Smith et Jain, 1988]. De nombreux auteurs ont noté que la procédure linéaire est robuste par rapport à la non normalité des classes, tandis que la quadratique souffre fortement d'un écart par rapport à cette normalité [Raudys et Jain, 1991]. De plus, si les classes sont multimodales, ces approches ne sont plus utilisables.

Le second inconvénient est lié à la dégradation des résultats d'estimation des paramètres lorsque le rapport entre le nombre Q d'observations de l'échantillon d'apprentissage et le nombre N d'attributs est faible. Dans le même ordre d'idée, pour deux classes gausiennes, Fukunaga [Fukunaga, 1990], Raudys et Jain [Raudys et Jain, 1991] montrent que l'augmentation de l'erreur de classification pour ces procédures est proportionnelle à 1/K, K étant le nombre de

classes, et dépend principalement de la dimension N de l'espace de représentation. D'autres méthodes paramétriques réduisent ces dépendances comme celle basée sur un test d'hypothèse de Smolarz [Smolarz, 1987].

Le troisième inconvénient concerne l'estimation des probabilités a priori. Il arrive souvent, en effet, que celles-ci ne soient pas accessibles. Dans ce cas, l'analyste a tendance à simplifier le modèle probaliste en les considérant identiques pour toutes les classes. Celeux et Govaert [Celeux et Govaert, 1991] ont démontré que les résultats sont fortement affectés par cette simplification si elle ne correspond pas à la réalité.

De plus, il est nécessaire de souligner que plus les classes du mélange sont proches, plus la dégradation des performances est importante. Il y a plusieurs approches possibles pour contourner cette difficulté. Une première consiste simplement à interpréter les estimations des probabilités a posteriori comme des degrés d'appartenance à une partition floue [Gethet et Geve, 1989]. Le critère de classification peut alors être modifié d'une manière adaptée [Bezdek, 1981 ; Windham, 1987 ; Trauvaert et al, 1991]. Une deuxième solution consiste à considérer un modèle de classification partielle [Celeux 1987]. On se ramène alors à un problème d'analyse modale dans un contexte paramétrique.

Il y a aussi des méthodes pour rechercher de manière itérative les estimateurs du maximum de vraisemblance des paramètres statistiques nécessaires à la classification [Everitt et Hend, 1981 ; Titterington et al, 1985 ; Celeux, 1988 ; Celeux et Diebolt, 1989]. Plusieurs méthodes classiques peuvent être utilisées du type gradient, Newton, Raphson, etc.... Il existe également des algorithmes qui résolvent ce problème : l'algorithme EM, (Estimation - Maximisation), l'algorithme SEM, (Stochastique-Estimation-Maximisation), et l'algorithme SAEM, (Stochastique-Annealing-Expectation-Maximisation). Toutefois ces algorithmes nécessitent une connaissance a priori du nombre de classes en présence et une initialisation proche de la solution recherchée [Mourot, 1993].

I-7. LES PROCEDURES METRIQUES DE CLASSIFICATION

En général, les résultats de procédures de classification statistique sont sensibles à la valeur du rapport entre le nombre Q d'observations disponibles et la dimension N de l'espace où elles sont représentées. Pour de petites valeurs de ce rapport, cette classification n'est plus possible. Les méthodes métriques permettent par contre d'analyser des données où ce rapport est faible.

Ces méthodes peuvent être divisées en deux groupes :

- les méthodes de classification hierarchique
- les méthodes de classification par optimisation d'un critère.

I - 7 - 1 . La classification hierarchique

Dans l'approche hiérarchique, les observations ne sont pas affectées aux classes correspondantes en une seule étape. Il existe deux grandes familles : les méthodes ascendantes et descendantes. Un algorithme ascendant commence avec une partition en un nombre de classes contenant chacune une des observations à classer. L'algorithme fusionne à chaque étape les deux classes les plus semblables en maximisant un critère de similarité. La partition précédente est ainsi emboîtée dans la partition suivante. Un algorithme descendant exécute le processus dans le sens inverse. Toutes les observations sont d'abord groupées en une seule classe et, à chaque étape, on maximise un critère de dissimilarité pour permettre de diviser chacune d'elles en deux. Le processus est itéré et la partition finale est obtenue en respectant soit un nombre de classes préfixé, soit un critère prédéfini [Bayne et al, 1980 ; Lane et al, 1967 ; Lukasova, 1979]. Une autre méthode consiste à réduire la hiérarchie des parties ou l'arbre des classifications sous forme d'un ensemble de nœuds significatifs, avant d'effectuer la partition de l'ensemble des objets à classer [Gordon, 1987 ; Lerman et al, 1991].

Ces méthodes hiérarchiques étant séquentielles, l'affectation d'une observation à un groupe n'est jamais remise en cause dans les étapes ultérieures. Pour le cas ascendant, par exemple, la séquence de partitions emboîtées démarre par le calcul de la matrice des distances ou de similarité entre les différentes observations, et se termine par un dendrogramme, ou arbre hiérarchique, qui montre les étapes de fusions successives en passant par une actualisation de la matrice des distances.

Les critères les plus utilisés dans les méthodes ascendantes sont, entre autres, le critère du plus proche voisin ou "Single link" [Florek et al, 1951 ; Mc Quitty, 1957 ; Sneath, 1957 ; Johnson, 1961], le critère de la distance moyenne ou "group average" [Sokel et Michener, 1958 ; Lance et Williams, 1966 ; Everitt, 1977], le critère du centre de gravité ou "centroid cluster analysis" [Sokel et Michener, 1958 ; King, 1966 et 1967] et le critère de Ward [Ward, 1963].

Pour les critères descendants, nous avons entre autres, l'analyse associative [Lambert et Williams 1962 et 1966 ; Mac Naughton-Smith, 1965], et la méthode de détection interactive automatique ou "Automatic Interaction Detector Method" (A. I. D.) [Sonquist et Morgan, 1963 et 1964].

La classification hiérarchique se heurte à deux problèmes. Elle est difficilement utilisable pour un grand

nombre de dimensions et d'observations car elle nécessite le calcul et le stockage de la matrice des distances, et cela même avec les algorithmes les plus performants [Bruynooghe, 1978]. Le second problème est lié au choix laissé à l'utilisateur d'une partition en un nombre de classes prédéfini. Si ce choix est difficile, un critère d'arrêt doit être spécifié.

I - 7 - 2 . La classification par optimisation d'un critère .

Ces méthodes de classification sont également appelées méthodes de coalescence. Elles diffèrent des méthodes hiérarchiques du fait qu'elles permettent le transfert d'une observation d'un groupe à un autre à condition que la nouvelle partition soit meilleure, au sens du critère à optimiser, que la partition obtenue précédement. Le processus s'arrête quand aucun tranfert n'améliore plus la partition. Le choix du critère de qualité à optimiser n'est pas simple car il suppose qu'il existe une bonne définition des classes, celles-ci pouvant être de formes, de dispersions et de tailles quelconques dans l'espace de représentation des données. Pour la plupart des méthodes, le nombre K de classes recherchées est fixé initialement par l'utilisateur. Le problème du choix du nombre correct de classes a été étudié par de nombreux auteurs [Thorndike, 1953 ; Ling 1971 ; Sneath et Sokal, 1973 ; Milligan et Cooper, 1985 ; Boch, 1985 ; Jain et Moreau, 1987 ; Dubes, 1987 et Jain et Dubes, 1987].

Dans ces méthodes, on commençe par définir K points dans l'espace de représentation des données, qui agiront comme estimateurs initiaux des centres des K classes. Ces centres, ou descripteurs des classes, sont appelés des noyaux. Chacune des observations est affectée à la classe dont le centre est le plus proche, et cela en optimisant le critère choisi. L'estimation du centre peut-être réactualisée soit après l'affectation de chaque observation [Mac Queen, 1967] ou après l'affectation de toutes les observations.

Dans la méthode des nuées dynamiques [Diday et al., 1982] qui est une généralisation d'un algorithme plus ancien, l'algorithme des centres mobiles [Ball et Hall, 1965], il s'agit d'optimiser un critère qui exprime l'adéquation entre une partition des observations en K classes et un mode de représentation des classes de cette partition.

Le problème d'optimisation se pose alors en termes de recherche simultanée de la partition et de sa meilleure représentation possible au sens du critère choisi. Les noyaux peuvent prendre des formes diverses :

- centres de gravité des classes,
- groupes de points,
- lois de densité de probabilité (critère de vraisemblance classifiante),
- sous-espaces vectoriels.

La difficulté de ces méthodes est la nécessité de connaître a priori le nombre de classes. De plus, elles peuvent être coûteuses en volume de calcul et gourmandes en place mémoire. Ces méthodes sont également sensibles à l'initialisation et elles nécessitent que les effectifs des classes ne soient pas trop déséquilibrés.

Il faut noter l'existence d'algorithmes à nombre de classes variables au cours de l'exécution. Il existe deux algorithmes principaux : l'algorithme ISODATA [Chandon et Pinson, 1981] et l'algorithme LEADER [Brown et al, 1991]. Ces algorithmes peuvent utiliser différents critères de classification. A partir d'une partition initiale, l'algorithme ISODATA procède par classification selon la méthode des centres mobiles par fusions ou divisions des classes en utilisant des seuils difficiles à définir a priori en absence de connaissance sur les classes. Par contre, l'algorithme LEADER est un algorithme qui classe séquentiellement les observations sans nécessiter le choix d'un seuil pour créer de nouvelles classes.

Après ce bref panorama des procédures de classification paramétriques et métriques, nous abordons les procédures non paramétriques.

I - 8 . LES PROCEDURES DE CLASSIFICATION NON PARAMETRIQUES

Pour éviter d'avoir recours à des modèles statistiques paramétriques qui peuvent conduire à imposer une structure aux données plutôt qu'à découvrir leur organisation véritable, les méthodes non paramétriques font appel à des notions de statistiques basées sur l'estimation point à point des fonctions de densité de probabilité.

En général, l'existence de plusieurs maxima locaux de la fonction de densité de probabilité sous-jacente à la distribution des observations est considérée comme l'indication d'une population formée de plusieurs classes, d'autant plus différenciées que ces modes sont nettement séparés par des régions où la fonction de densité de probabilité est de très faible valeur : les vallées. Evidemment, il s'agit ici d'un modèle de classification non paramétrique, sans définition précise du concept de classe [Bock, 1989]. Dans ce contexte, et à l'opposé des méthodes paramétriques, la loi de densité de probabilité est estimée par des méthodes non paramétriques comme la méthode du noyau de Parzen ou celle des G plus proches voisins.

I - 8 - 1 . La méthode du noyau de Parzen

Cette méthode est basée sur l'estimation explicite des lois de densité de probabilité des classes. L'estimation non paramétrique d'une loi de densité au point X est de la forme :

$$p(X) = \frac{G_q / Q}{V_q} \qquad \text{I-21}$$

où V_q et G_q sont respectivement le volume d'un domaine entourant X et le nombre d'observations tombant dans ce domaine.

L'estimateur du noyau est donné par

$$p(X \mid C_k) = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{V_q} \Phi\left(\frac{X - X_q}{h_q}\right) \qquad \text{I-22}$$

Φ étant la fonction noyau et h_q la largeur de ce noyau.

La fonction Φ et le paramètre h_q déterminent respectivement la forme du domaine et sa taille [Fukunaga, 1990]. On remarque que le choix de la largeur de fenêtre h_q est très critique pour l'estimation de la fonction de densité de probabilité car ce paramètre est

fortement lié au nombre d'attributs et au nombre d'observations constituant l'ensemble d'apprentissage. Fukunaga et Hummels [Fukunaga et Hummels, 1987] ont d'ailleurs montré que la forme du noyau n'est pas un facteur aussi influent que sa largeur.

En effet l'ajustement du paramètre h_q, qui conditionne la qualité de l'estimation, est l'un des problèmes les plus délicats à résoudre en utilisant la méthode du noyau. Si h_q est trop grand, il y a sur-lissage. L'estimateur manque alors de résolution et ignore certaines variations significatives de la fonction de densité de probabilité. Si h_q est trop petit, les variations de l'estimateur deviennent erratiques et ne sont plus significatives. Dans la pratique, on constate même que des domaines de taille satisfaisante dans certaines régions de l'espace à forte densité d'observations s'avèrent trop petits dans les régions à faible densité et réciproquement [Postaire, 1987].

I - 8 - 2 . L'approche des G plus proches voisins.

Pour pallier les difficultés liées au choix du paramètre h_q dans la méthode du noyau, on peut éviter de fixer arbitrairement la taille des domaines. Une autre approche consiste à ajuster cette taille en fonction de la

densité des observations dans le voisinage de l'observation X_q où on estime la fonction de densité. Il s'agit de la méthode des G plus proches voisins.

Soit $D_r(X_q)$ un domaine de volume unité admettant X_q pour centre de symétrie, tel que :

$$V[D_r(X_q)] = 1 \qquad \text{I-23}$$

où $V[D_r(X_q)]$ indique le volume du domaine $D_r(X_q)$

Un tel domaine est appelé domaine de référence.

Soit $D(X_q, w)$, le domaine homothétique du domaine de référence $D_r(X_q)$ dans une homothétie de centre X_q et de rapport w, $w > 0$. On montre aisément que :

$$V[D(X_q, w)] = w^N \qquad \text{I-24}$$

La taille du domaine $D(X_q, w)$, appelé domaine d'observation, est donc une fonction monotone croissante du rapport d'homothétie w.

La méthode d'estimation des G plus proches voisins consiste à faire croître la taille du domaine d'observation centré en X_q jusqu'à ce qu'il englobe un nombre donné G d'observations. Ce nombre G est, bien entendu, fonction du nombre Q d'observations disponibles. Ces G observations

englobées dans le domaine $D(X_q, w)$ sont les G plus proches voisins de X_q. Si la densité des observations est élévée au voisinage de X_q, la taille du domaine $D(X_q, w)$ utilisé est relativement petite. Si, au contraire, la densité est faible, la taille du domaine utilisé est plus importante. Dans les deux cas et conformément à l'équation I-22 (Cf. § I-8-1.), G doit aussi bien converger vers l'infini lorsque Q tends vers l'infini, qu'augmenter lorsque $D(X_q, w)$ tend vers zéro et cela pour garantir la convergence de l'estimateur [Duda et Hart, 1973]. Les fonctions

$$G = G_o\sqrt{Q} \quad \text{ou} \quad G = G_o \, Ln \, Q \qquad \text{I-25}$$

satisfont ces propriétés et garantissent la convergence de l'estimateur.

I-8-3. Détection des modes par recherche des maxima locaux

Les modes, extréma locaux de la fonction de densité de probabilité, indiquent la présence des classes au sein de la population estimée. Ils peuvent être détectés en "remontant" les pentes de cette fonction selon la direction de son gradient [Asselin de Beauville, 1978 ; 1979 ; 1983 ; 1989]. La fonction de densité de probabilité est estimée par une des méthodes non-paramétriques évoquées ci-dessus : la méthode du noyau ou celle des G plus proches voisins.

L'estimation obtenue est alors utilisée pour déterminer son gradient. Une variante consiste à estimer directement ce gradient à partir des observations elles-mêmes [Mizoguchi et Shimura, 1976 ; Fukunaga et Hostetler, 1975 ; Kittler, 1976]. Comme toutes les approches basées sur l'utilisation d'opérateurs différentiels, ces méthodes sont sensibles aux irrégularités de la distribution des données et tendent à générer un nombre excessif de groupements. De plus, les algorithmes de détection des modes qui recherchent les maxima locaux de la fonction de densité sont mal adaptés à l'analyse d'échantillons où les classes présentent des densités très différentes car des modes d'amplitude élevée risquent de masquer les "petits" modes [Touzani et Postaire, 1989b]. L'introduction des techniques de relaxation permet de limiter les effets des disparités entre les tailles des modes [Touzani et Postaire, 1988].

I - 8 - 4 . Analyse de la convexité

Au lieu de considérer les modes comme des extréma locaux de la fonction de densité, on peut les assimiler à des régions de l'espace où cette fonction est concave [Vasseur et Postaire, 1980 ; Postaire et Vasseur 1981 ; Postaire 1982 et 1983]. L'analyse de la convexité de la fonction de densité est effectuée en intégrant cette dernière sur des domaines d'observation de taille variable. Cette analyse améliore considérablement la robustesse de la méthode par rapport

aux techniques faisant appel aux notions de gradient, mais elle reste encore sensible aux irrégularités de la distribution des données. Là aussi, la relaxation permet d'améliorer les performances de la procédure.

I - 8 - 5 . Extraction des contours des modes

Une autre approche pour extraire les régions modales [Postaire et Touzani, 1989] est basée sur la notion de contour des lois de densité de probabilité. Chaque mode est en fait considéré comme une région de l'espace de représentation où se concentrent les observations associées à ce mode, entourée d'une région caractérisée par une faible densité d'observations. A la limite entre ces deux types de régions, il existe de fortes variations spatiales de la loi de densité de probabilité [Touzani et Postaire, 1989 ; 1991 ; Gzesnalowicz et Postaire, 1990 ; Zhen, Botte-Lecocq et Postaire 1991 ; Oulad Haj Thami et Asselin de Beauville, 1992 ; Zhang et Postaire, 1994]. Les frontières situées à la limite des modes sont mises en évidence grâce à des opérations différentiels multidimensionnels. Ensuite, une procédure d'extraction de contours permet de délimiter chaque mode. Les observations situées à l'intérieur des domaines de concavité ou des contours constituent des noyaux de classes. Les observations restantes sont alors affectées à leur plus proche noyau.

Ces trois dernières approches de recherche des modes posent le même problème : l'estimation de la fonction de densité ou de son gradient nécessite un grand nombre d'observations. Pour un petit nombre d'observations, ces approches de classification ne sont plus utilisables.

I - 8 - 6 . Conclusion sur l'approche non paramétrique

L'objectif des méthodes non paramétriques est de détecter des classes dans un ensemble de données sans connaissance a priori sur le nombre de classes et leurs caractéristiques. Par conséquent, les classes détectées peuvent avoir des caractéristiques de formes, de dispersions, d'orientations et d'effectifs quelconques. En apprentissage non supervisé, cette capacité est essentielle.

Ces méthodes souffrent cependant de plusieurs inconvénients principaux. Tout d'abord, la précision de l'estimation dépend de la taille de l'ensemble d'apprentissage surtout dans un espace de grande dimension [Fukunaga, 1990]. De plus, les méthodes d'estimation sont très coûteuses en temps de calcul et en place mémoire.

Enfin, le problème posé par l'approche non paramétrique est la validation et l'interprétation des résultats obtenus. La difficulté provient de l'absence de critère à optimiser ou de modèles de classes bien définis.

En pratique, les situations de multimodalité sont mal maîtrisées dans le cadre multidimensionnel. Par exemple, un mélange de M lois normales de moyennes distinctes ne donne pas toujours une fonction de densité de mélange à M modes [Everitt et Hand, 1981]. Par conséquent, un algorithme détectant les modes ne sera pas toujours en mesure de trouver le nombre de composants. Celeux [Celeux, 1987] pense que les algorithmes de recherche de modes ne permettent de déceler que des structures en classes très séparées et aisement détectables par des algorithmes plus simples. Ils ne permettent pas de détecter des classes ayant une structure complexe.

I - 9 . METHODES DIVERSES

I - 9 - 1. Les réseaux neuronaux

Les réseaux de neurones sont utilisés en reconnaissance des formes et en classification avec professeur car ils se prêtent très bien à l'apprentissage quand on leur présente des données étiquetées. Les neurones, éléments de base des réseaux, se modélisent comme des systèmes caractérisés par des fonctions de transfert non linéaires à une sortie et plusieurs entrées.

Les modèles de réseaux de neurones sont nombreux : modèle de Hopfield, de Boltzmann, réseau multi-couches, etc ... [Vassiliadis, 1990 ; Lau, 1992]. Il en est bien sûr de même des algorithmes d'apprentissage qui permettent leur mise en œuvre [Vassiliadis, 1990].

Les réseaux de neurones permettent :

- d'engendrer des frontières de décision entre classes non linéairement séparables,
- de séparer la phase d'apprentissage de la phase de classification du réseau. La convergence des poids synaptiques au cours de l'apprentissage nécessite un effort de calcul mais pour l'exploitation du réseau, c'est-à-dire la classification de nouveaux échantillons, seul un petit nombre d'opérations simples implantées en parallèle sont nécessaires, ce qui les rend très performants,
- ils sont robustes aux données bruitées,

Parmi les inconvénients, nous pouvons citer les problèmes liés à :

- la difficulté du choix du modèle de réseau,
- la définition du type de cellule de base, du nombre de couches, et des connexions du réseau,

- l'initialisation des poids du réseau [Masson, 1992],
- l'ajustement des paramètres de l'algorithme d'apprentissage [Sorsa et Kiovo, 1991],
- la lenteur de l'apprentissage, surtout si les données sont bruitées [Venketasubamanian et Vaidyanathan, 1991].

De plus, la nécessité d'une phase d'apprentissage limite l'utilisation de ces réseaux à des problèmes de classification surpervisée.

I - 9 - 2 . Les arbres de décision

L'idée de base ayant donnée naissance aux arbres de décision est de décomposer une décision complexe en un ensemble de décisions plus simples [Safian et Landgrebe, 1991]. Un arbre de décision est constitué d'un nœud-racine, de nœuds non terminaux et de nœuds terminaux. Un nœud non terminal correspond à une décision partielle ou intermédiaire à laquelle un sous-ensemble de classes est associé. Une seule classe est associée à chaque nœud terminal.

N'importe laquelle des méthodes de classification précédentes peut-être utilisée pour effectuer la décision à un nœud de l'arbre.

Cette approche est intéressante pour les raisons suivantes :

- des décisions complexes peuvent être décomposées en décisions plus simples qui correspondent, par exemple à des classes multimodales,
- la classification est rapide car contrairement aux autres méthodes où chaque observation est testée par rapport à chaque classe, dans celle-ci une observation est testée seulement par rapport à certains sous-ensembles de classes,
- pour les méthodes classiques, un seul ensemble d'attributs est utilisé pour discriminer toutes les classes. Pour les arbres de décision, on a différents attributs pour les différents nœuds internes de l'arbre. Cette possibilité peut réellement améliorer les performances en prenant en compte une information locale et en évitant les problèmes liés à un espace de grande dimension.

La conception d'une méthode hiérarchique nécessite les choix suivants :

- choix approprié de la structure de l'arbre,
- choix des attributs à utiliser à chaque nœud interne,

- choix de la règle de décision à utiliser à chaque nœud interne.

Ces problèmes sont difficiles à résoudre car il n'existe pas de solutions optimales [Safian et Landgrebe, 1991].

Quelques auteurs ont tenté des rapprochements entre les réseaux de neurones et les arbres de décision [Safian et Landgrebe, 1991]. Sethi et Jain [Sethi et Jain, 1991] utilisent un arbre de décision qu'ils transforment en un réseau de neurones multicouche.

I - 9 - 3 . La classification floue

La classification floue consiste à attribuer aux observations à classer des degrés d'appartenance par rapport aux classes floues mises en évidence. Si les degrés d'appartenance d'une observation aux classes floues sont tous inférieurs à des seuils d'appartenances, cette observation pourra être rejetée. La notion de degré d'appartenance permet de nuancer l'interprétation et de prendre en compte le phénomène d'évolution et le recouvrement des classes [Usai et Dubuisson, 1984 ; Béreau, 1986]. Le concept de fonction d'appartenance floue est utilisé pour quantifier le degré d'appartenance d'une observation X_q à une classes C_k qui est alors considérée comme un ensemble flou.

Pour l'ensemble d'apprentissage, une partition floue en K classes spécifie le degré d'appartenance de chaque observation à chaque classe.

La partition floue est caractérisée par une matrice μ :

$$\mu = [\mu_{qk}] \quad q = 1, 2, ..., Q \quad k = 1, 2, ..., K \qquad \text{I-26}$$

qui respecte les contraintes des conditions de non dégénérescence des classes et d'orthogonalité des données respectivement par :

$$0 < \sum_{q=1}^{Q} \mu_{qk} < n \qquad \text{I-27}$$

et

$$\sum_{k=1}^{K} \mu_{qk} = 1 \qquad \text{I-28}$$

La règle d'affectation utilisée est la règle du maximum d'appartenance définie par

X_q appartient à la classe C_k

$$\text{si } \mu_{qk} = \max \mu_{qj} \quad , j = 1, 2, ..., K \qquad \text{I-29}$$

Un certain nombre de fonctions d'appartenance ont été proposées, dont celles faisant intervenir la distance entre une observation et le prototype de la classe [Chaudhuri et Dutta Majumber, 1982]. Généralement, la distance utilisée est celle de Mahalanobis. Ces fonctions ont été étendues en considérant comme prototype d'une classe un noyau d'observations représentatif, ou les G plus proches voisins de cette observation, afin de ne pas être confronté aux problèmes d'estimation des paramètres pour le calcul de la distance de Mahalanobis [Dubuisson et al, 1986].

Jozwik [Jozwik, 1983] a proposé une extension floue de la méthode des G plus proches voisins et un algorithme d'apprentissage pour cette règle. Le degré d'appartenance d'une observation à une classe est la moyenne des degrés d'appartenance de ces G plus proches voisins à cette classe. Keller [Keller et al, 1985] a proposé une méthode d'affectation à chaque observation d'un degré d'appartenance unité à la classe à laquelle elle appartient et des degrés d'appartenance nuls pour toutes les autres classes. Bezdek [Bezdek et al, 1986] propose d'utiliser un algorithme flou de classification tel celui de C - moyennes (FCM) pour initialiser la méthode floue de G plus proches voisins. Cet algorithme est bien adapté à la détection de classes gaussiennes équiprobables sphériques.

Béreau [Béreau, 1986] a étendu la fonction caractéristique des G plus proches voisins afin que la

transition entre appartenance et non appartenance soit graduelle. La fonction d'appartenance floue choisie est inspirée de la fonction de distribution de Fermi-Dirac en thermodynamique statique. Cette règle de décision présente un avantage essentiel, la pondération de la distance de l'observation X_q à ses G plus proches voisins présente une décroissance moins rapide que les méthodes précédentes et, par conséquent, le voisinage d'une observation est mieux défini.

I - 9 - 4 . La fonction potentielle

Une classe est représentée dans l'ensemble d'apprentissage par l'ensemble des observations qui la composent. Toutes ces observations ne sont pas forcément représentatives de cette classe. La notion de potentiel introduite par Grenier et Gane [Grenier, 1984 ; Gane, 1987] permet d'évaluer la représentativité d'une observation. Le potentiel d'une observation X_q par rapport à la classe C_k est en fait la distance moyenne de ce point à ses plus proche voisins dans la classe [Dubuisson, 1990 a]. Ce potentiel décroit de manière strictement monotone à mesure que X_q s'éloigne de la classe.

Cette méthode permet une décision avec des classes non convexes [Dubuisson, 1990 a].

I - 10 . CONCLUSION

Nous avons présenté, dans ce chapitre, la plupart des approches paramétriques et non paramétriques pour la classification des données. Chaque méthode présente certains avantages et inconvénients qui favorisent son application plutôt à un type de données qu'à un autre. Il n'existe pas de procédure universelle adaptée à tous les types de données.

Nous proposons, dans ce mémoire, une nouvelle approche basée sur une technique d'amaincissement de l'histogramme selon une procédure de maximisation de la taille des regroupements des observations ou de leurs sous-ensembles. Notre objectif n'est pas de présenter "la méthode universelle" ou "la meilleure procédure" utilisable pour tous les types de données, mais de montrer l'intérêt d'une telle approche et son applicabilité à des distributions différentes tant en "forme" et "taille" des regroupements qu'au niveau de la "dimension des données".

Nous nous sommes placés dans un contexte non supervisé, sans aucune hypothèse paramétrique, afin de traiter des observations multidimensionnelles à valeurs réelles, représentables par des points repartis dans un espace Euclidien. Nous admettrons qu'il est possible de réaliser une partition de l'ensemble des données en classes disjointes.

La nouvelle approche pour la classification que nous présentons dans ce mémoire utilise l'histogramme des données. Dans le deuxième chapitre, nous présentons une étude rapide des histogrammes unidimensionnels et multidimensionnels, de leur construction et de l'effet du pas de discrétisation sur leur forme.

La plupart des procédures de classification utilisent des algorithmes de balayage de l'espace discrétisé. La dimension élevée N de l'espace de représentation de données entraine trop de lourdeurs au niveau de la programmation pour que ces algorithmes développés puissent être utilisables quelle que soit la dimension des données. D'autre part, les temps d'éxécution des procédures ainsi que la taille mémoire nécessaire sont proportionnels à cette dimension N. Pour améliorer ces temps d'éxécution, réduire la taille de l'espace mémoire et simplifier la structure des algorithmes, nous proposons dans le troisième chapitre une nouvelle procédure de balayage qui transforme mathématiquement le problème multidimensionnel en un problème à une seule dimension, quelle que soit la valeur de N. Cet algorithme, qui ne réduit pas physiquement la dimension de l'espace, ne prend en considération que les points de l'espace discrétisé situés dans les régions où des observations sont effectivement présentes.

Le quatrième chapitre de ce mémoire est consacré à la nouvelle approche du problème de classification automatique elle-même. Il s'agit d'une procédure d'amaincissement progressif, par étapes, des modes de l'histogramme multidimensionnel. La procédure, qui opére par regroupements successifs des observations, entraîne une réduction progressive de leur dispersion autour de chacun des modes de la distribution. Cette propriété résulte de la maximisation de la taille des regroupements itératifs des observations qui contribuent à donner naissance à chaque mode de la distribution.

Le chapitre cinq illustre cette nouvelle procédure de classification à l'aide d'ensembles de données générées artificiellement.

CHAPITRE II
L'HISTOGRAMME : UN MODE DE REPRESENTATION PRIVILIGIE DES DONNEES

CHAPITRE II
L'HISTOGRAMME : UN MODE DE REPRESENTATION PRIVILEGIE DES DONNEES

II - 1 . INTRODUCTION

Dans de nombreux domaines de l'activité humaine, la connaissance a été acquise par l'étude de collections d'objets. Pour découvrir l'organisation de ces objets, ceux-ci doivent être classés selon des critères bien définis. Toute classification comporte une phase initiale où il s'agit de collecter des informations sur les objets, suivie d'une phase de dépouillement qui consiste à passer des données brutes aux représentations qui se prêtent le mieux à l'analyse et à l'interprétation.

II - 2 . LES TABLEAUX STATISTIQUES

Les données brutes peuvent être mises sous la forme de tableaux permettant de regrouper sous la forme d'une seule représentation la description des objets par un ou plusieurs attributs. Ces tableaux peuvent être des tableaux d'effectifs ou des tableaux de fréquences.

Dans ce type de représentation, afin de rendre la description statistique de la distribution des objets commode à étudier, on regroupe les valeurs de chaque attribut en intervalles successifs et adjacents, de telle sorte qu'on ne différencie pas les valeurs de l'attribut qui sont comprises dans chaque intervalle.

Chaque intervalle peut être caractérisé par une seule valeur de l'attribut observé, généralement, celle du milieu de l'intervalle. Un exemple d'un tableau de fréquences à une seule variable représentant un attribut continu est donné par le tableau II-1. Cet exemple représente le temps nécessaire pour l'inspection d'un article. L'intervalle adopté est de une seconde et le nombre total d'articles est 100.

Intervalles (secondes)	Centre d'intervalle	Effectifs	Fréquences
[11 - 12 [	11,5	2	0,02
[12 - 13 [	12,5	16	0,16
[13 - 14 [	13,5	29	0,29
[14 - 15 [	14,5	27	0,27
[15 - 16 [	15,5	11	0,11
[16 - 17 [	16,5	6	0,06
[17 - 18 [	17,5	4	0,04
[18 - 19 [	18,5	2	0,02
[19 - 20 [	19,5	2	0,02
[20 - 21 [	20,5	1	0,01

Un tableau d'effectifs et de fréquences à une seule variable représentant un attribut continu

Tableau II - 1

II - 3 . LES GRAPHES

Les représentations graphiques permettent, par simple lecture, de voir les caractéristiques essentielles de la distribution des objets, telles que le nombre de classes qui la constituent et leurs caractéristiques. Un simple coup d'œil permet aussi de comparer des distributions différentes.

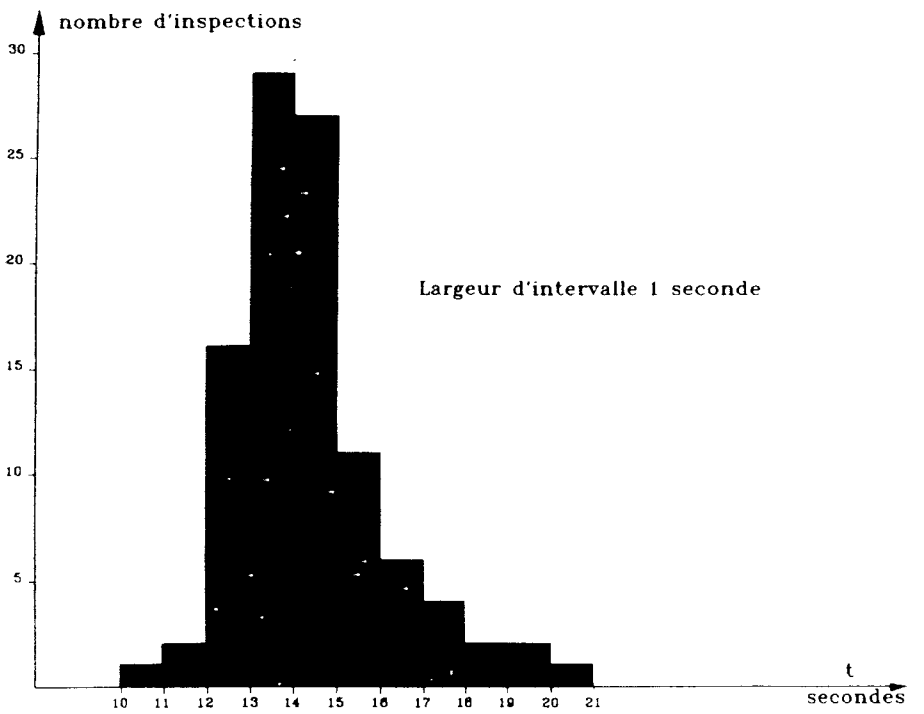
Lorsque l'attribut décrivant les objets est mesuré par une variable discontinue, on utilise un diagramme en "bâtons".

Dans le cas d'un attribut mesuré par une variable continue, on utilise une représentation graphique sous la forme d'un histogramme, qui constitue une généralisation du diagramme en bâtons.

La figure II.1 représente l'histogramme de l'exemple du tableau II-1. La plage de variation du temps d'inspection, qui s'étend de 11 à 21 secondes, est divisée en dix intervalles égaux d'une seconde chacun. L'histogramme est ainsi constitué de dix rectangles et on n'y distingue qu'un seul mode.

Le choix de la largeur des intervalles est généralement arbitraire. Sans connaissance a priori sur la variable à représenter, il est difficile d'ajuster ce paramètre fondamental qui conditionne grandement la forme de l'histogramme. Partant des données brutes du même exemple, nous pouvons bâtir plusieurs histogrammes des effectifs avec des intervalles de largeurs différentes.

Si la largeur de l'intervalle est agrandie à deux secondes, le nombre d'intervalles est de cinq et nous avons toujours un seul mode (Cf. figure II.2). Cet histogramme ne révèle pas toute la finesse de l'histogramme original de la figure II.1.

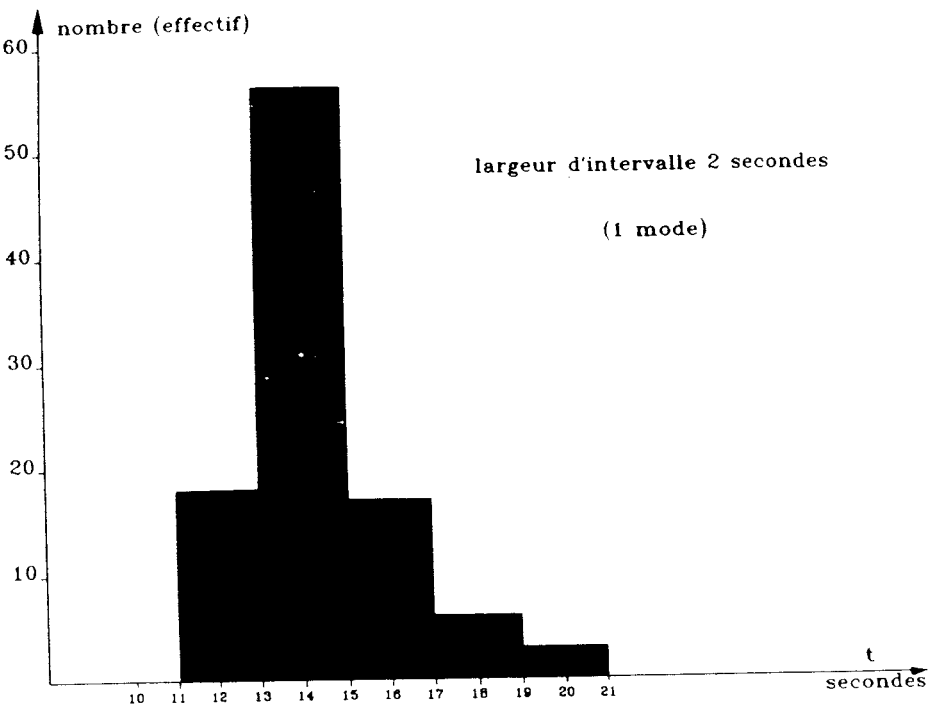


Histogramme de l'exemple du tableau II-1 pour une largeur d'intervalle de une seconde

Figure II.1

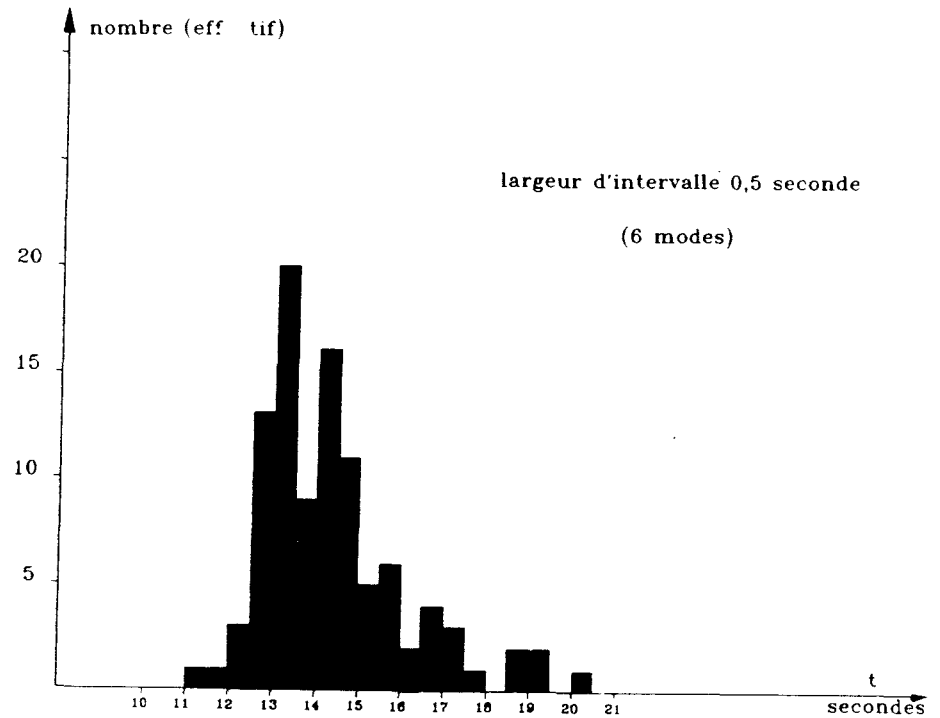
Si la largeur de l'intervalle est réduite à 0,5 seconde, l'histogramme correspondant devient hératique, donnant un effet de dents de scie qui n'apparait pas avec des intervalles plus larges. On dénombre six modes définis comme des maximum locaux de l'histogramme, (Cf. figure II.3).

Avec quarante intervalles d'une largeur de 0,25 seconde, l'histogramme présente une très forte variabilité. On y dénombre treize modes non significatifs qu'il ne faudrait pas interpréter comme une prolifération de sous-populations homogènes (Cf. figure II.4).



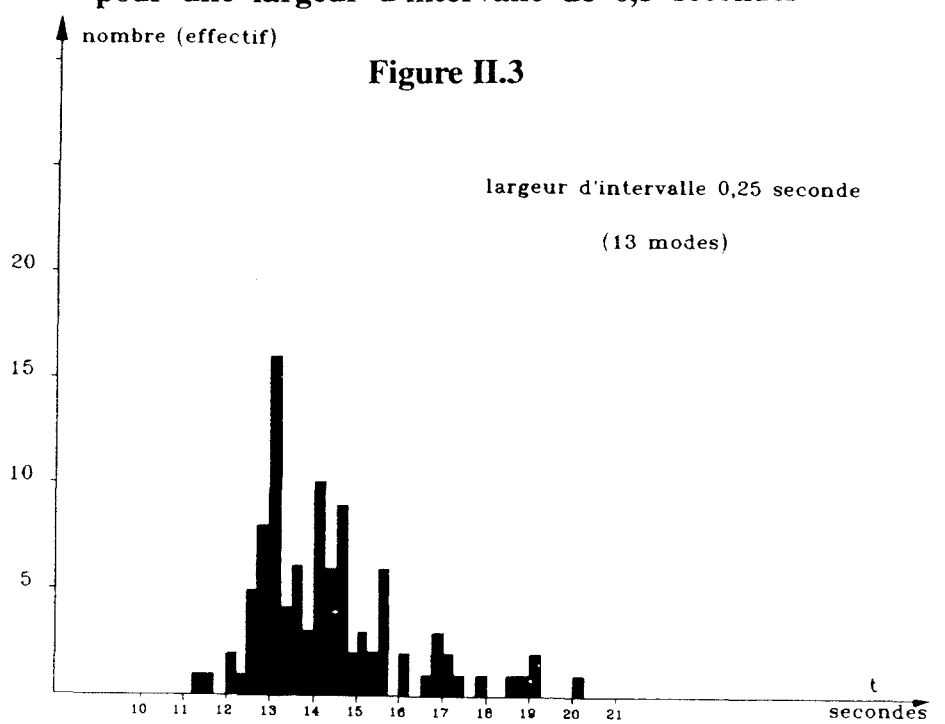
Histogramme de l'exemple du tableau II-1 pour une largeur d'intervalle de deux secondes

Figure II.2



Histogramme de l'exemple du tableau II-1 pour une largeur d'intervalle de 0,5 secondes

Figure II.3



Histogramme de l'exemple du tableau II-1 pour une largeur d'intervalle de 0,25 secondes

Figure II.4

II - 4 . CONSTRUCTION D'UN HISTOGRAMME MULTIDIMENSIONNEL

II - 4 - 1 . Discrétisation de l'espace de représentation des données

Comme nous l'avons indiqué au chapitre I, le vecteur attribut associé à chaque observation se présente sous la forme

$$X_q = [x_{1,q} ; x_{2,q} ; ..... ; x_{n,q} ; ..... ; x_{N,q}]^T \qquad \text{II-1}$$
$$q = 1, 2, ..., Q$$

La première étape pour générer l'histogramme consiste à discrétiser l'espace où sont représentées ces Q observations sur une grille hyperparallèlépipédique. Les plages de variation de chacun des attributs sont divisées en un même nombre D d'intervalles. Sur le $n^{ième}$ axe, on définit ainsi D intervalles dont la largeur C(n) est donnée par :

$$C(n) = \frac{\max_q(x_{n,q}) - \min_q(x_{n,q})}{D} \qquad \text{II-2}$$

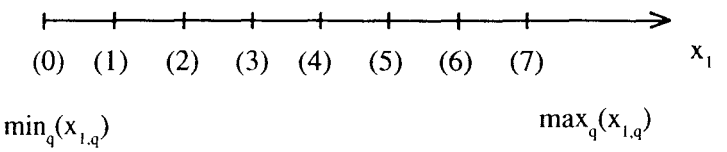
$$n = 1, 2 ,3,....., N \qquad \text{et} \qquad q = 1, 2, 3, ..., Q,$$

où $\max_q(x_{n,q})$ et $\min_q(x_{n,q})$ sont les valeurs maximum et minimum de l'attribut x_n. Les extrémités des intervalles sont repérées sur chaque axe par un index entier positif a_n variant de 0 à D.

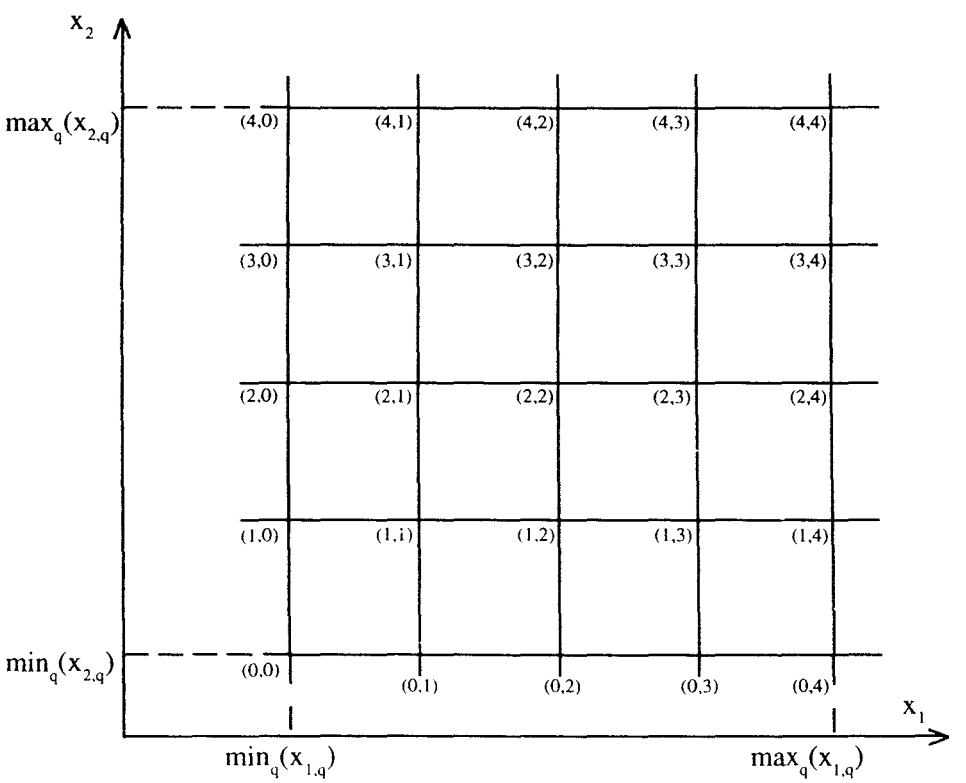
Cette division de chacun des axes en D intervalles permet de définir D^N hyperparallélépipèdes dont les sommets sont considérés comme les $(D + 1)^N$ points de discrétisation de l'espace. Ces cellules aux faces parallèles aux $[N (N - 1) / 2]$ plans définis par tous les couples d'axes définissant l'espace Euclidien E^N couvrent toute la partie utile de cet espace.

Etant donné que les plages de variations $[\max_q(x_{n,q}) \ldots \min_q(x_{n,q})]$, $n = 1, 2, 3, \ldots N$ des attributs x_n sont généralement inégales, les côtés des cellules sont de longueurs également inégales.

Conformément à cette discrétisation des plages de variation des attributs, à chacun des 2^N sommets de chaque cellule on affecte un index vectoriel à N dimensions : $A = (a_1, a_2, \ldots, a_n, \ldots, a_N)$ appelé "adresse multidimensionnelle" A du sommet, noté lui-même S(A). Chaque élément a_n, $n = 1, 2, 3, \ldots, N$, de cette adresse est l'index définissant la position du sommet par rapport à la discrétisation selon la $n^{ième}$ dimension. La figure II.5 représente les points de discrétisation aux sommets des cellules pour un espace à une dimension et pour un espace à deux dimensions. L'adresse A de chaque sommet est précisée sur la figure.



Discrétisation aux sommets des cellules pour N = 1 , D = 7



Discrétisation aux sommets des cellules pour N = 2 et D = 4

Figure II.5

II - 4 - 2 . Génération de l'histogramme : règle générale

Chaque observation est en général contenue dans une cellule. La cellule renfermant une observation donnée est appelée la cellule "associée" à cette observation. Réciproquement, cette observation est dite "assignée" à cette cellule. Une cellule renfermant plusieurs observations est donc associée à toutes ces observations. Réciproquement, celles-ci sont toutes des observations assignées à cette même cellule.

Conformément à cette définition, le sommet "associé" à une observation est le sommet de la cellule associée qui est le plus proche de l'observation considérée. Un sommet peut être associé à plusieurs observations situées dans les 2^N cellules ayant ce sommet en commun. On dit aussi que ces observations sont "assignées" à ce sommet. On note H(A) le nombre d'observations assignées au sommet S(A) d'adresse multidimensionnelle A.

La deuxième étape vers la génération de l'histogramme consiste à trouver le sommet associé à chaque observation dans l'espace Euclidien.

Chaque élément a_n de l'index vectoriel $A = a_1, a_2, ..., a_n, ..., a_N$ du sommet A associé à une observation $X_q = [x_{1,q} ; x_{2q} ; ....x_{n,q},....,x_{N,q}]^T$ est donné par :

$$a_n = \text{partie entière de } \{ \frac{x_{n,q} - \min_q(x_{n,q})}{C(n)} + 0,5 \} \qquad \text{II-3}$$

avec

$$n = 1, 2, 3, ..., N \quad \text{et} \quad q = 1, 2, 3, ..., Q$$

Notons que l'on a

$$a_n \in \{0, 1, 2, ..., D\}. \qquad \text{II-4}$$

Tout sommet associé à une ou plusieurs observations est aussi appelé un sommet "occupé". Dans le cas contraire, c'est un sommet "non occupé".

La table qui contient les adresses A des sommets S(A) occupés ainsi que le nombre d'observations H(A) assignées à chacun d'eux définit l'histogramme multidimensionnel des effectifs. Pour obtenir cette table, on balaye l'espace de représentations des données. Pour chaque observation, l'adresse du sommet associé est déterminée et est comparée à celles de la table. Si cette adresse A est déjà incluse dans la table, la valeur H(A) de l'histogramme qui lui est associée est incrémentée d'une unité. Autrement, une nouvelle adresse A' est ajoutée à la liste avec une valeur de l'histogramme H(A') égale à un. Lorsque le balayage de l'espace est terminé, toutes les adresses des sommets occupés ainsi que les valeurs de l'histogramme égales aux nombres d'observations assignées à chacun

d'eux sont disponibles sous la forme d'un tableau de dimension maximum $(N+1) * (D+1)^N$.

II - 4 - 3 . Génération de l'histogramme : les cas particuliers

En dehors de ce cas général, plusieurs autres situations peuvent apparaître lors de l'assignation des observations à leurs cellules et leurs sommets respectifs.

1° - Une observation située au centre d'une cellule est assignée au sommet de cette cellule ayant le plus grand nombre d'observations assignées dans le cadre de la règle générale,

2° - Une observation située sur une face d'une cellule est assignée au sommet le plus proche,

3° - Une observation située au centre d'une face d'une cellule est assignée au sommet de cette face ayant le plus grand nombre d'observations assignées dans le cadre de la règle générale,

4° - Une observation située sur une arrête d'une cellule est assignée au sommet le plus proche,

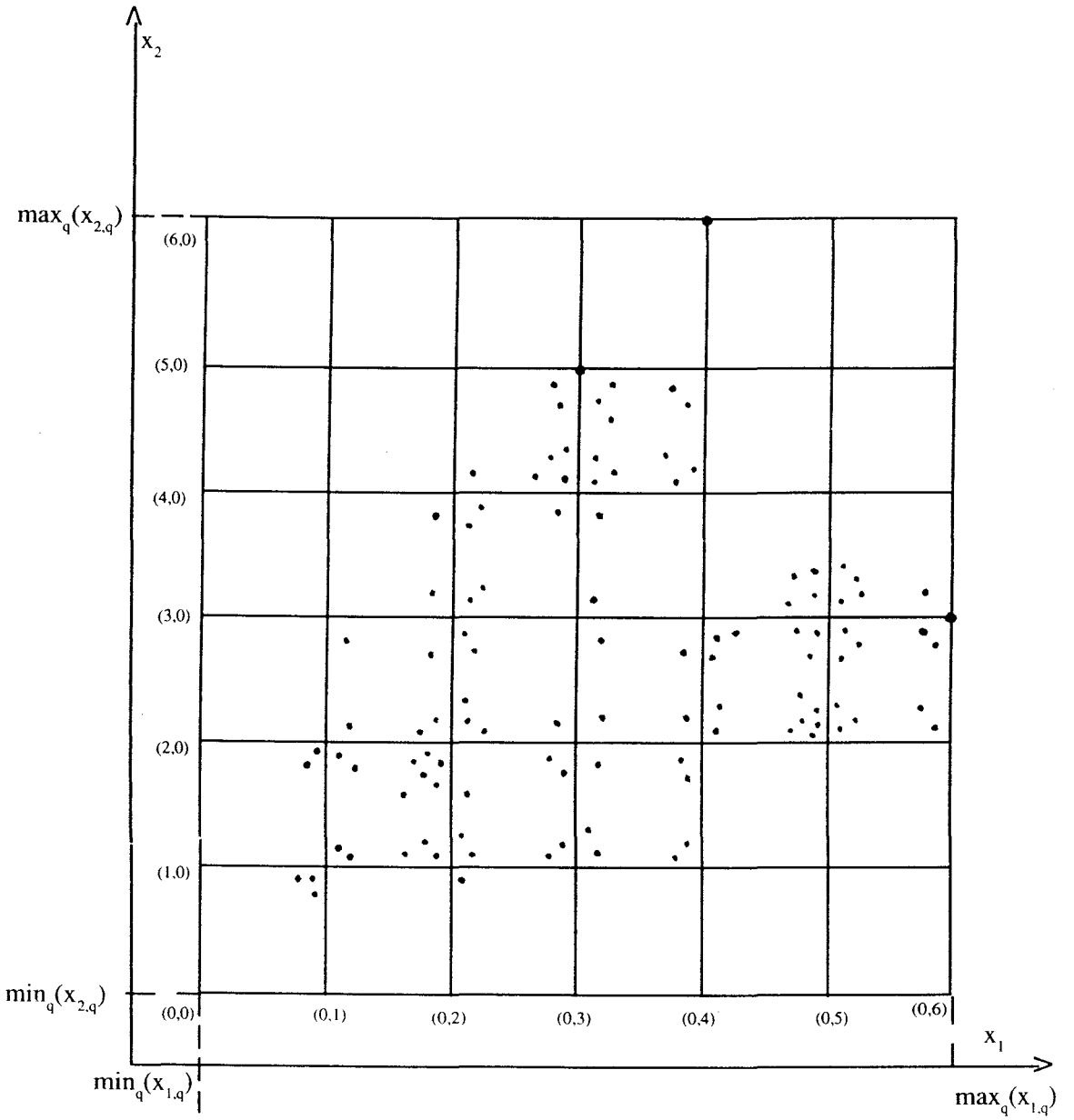
5° - Une observation située sur une arrête à mi-distance de deux sommets est assignée au sommet ayant le plus grand nombre d'observations assignées dans le cadre de la règle générale,

6° - Une observation se trouvant à égale distance de plusieurs sommets est assignée au sommet ayant le plus grand nombre d'observations assignées dans le cadre de la règle générale.

Dans les cas 1°, 3°, 5° et 6°, si les sommets ont le même nombre d'observations assignées, l'observation est assignée aléatoirement à n'importe lequel des sommets.

II - 5 . EXEMPLE DE CONSTRUCTION D'UN HISTOGRAMME

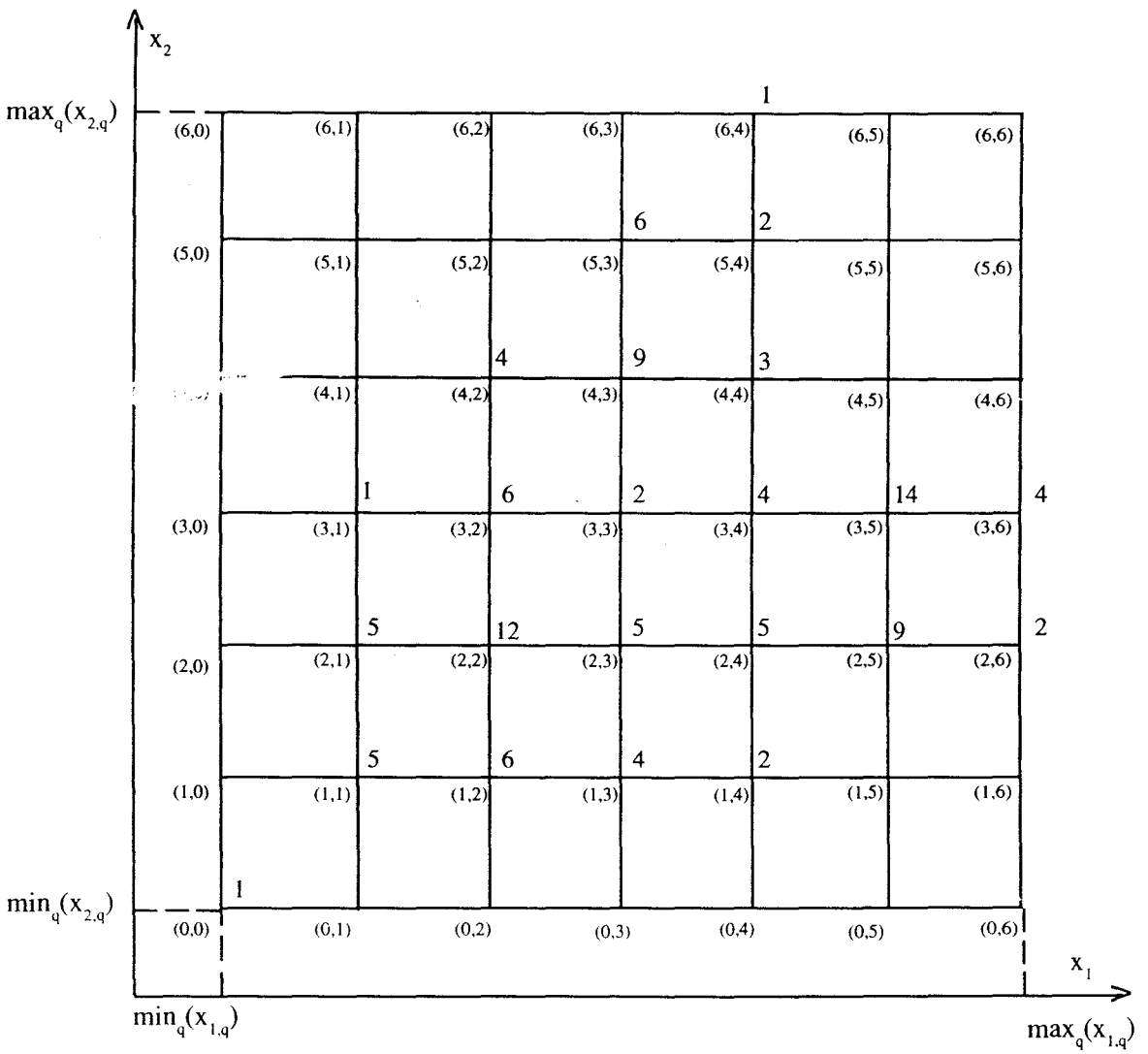
La figure II.6 représente un exemple à deux dimensions comportant 111 observations réparties en trois classes. La représentation numérique de l'histogramme pour cet exemple est donnée sur la figure II.7. Nous pouvons y distinguer deux types de sommets : les sommets occupés auquels des observations ont été assignées et les sommets non occupés auxquels aucune observation n'est assignée car ils se trouvent situés dans des zones de l'espace vides d'observations. Nous exploiterons ultérieurement cette distinction entre ces 2 types de sommets.



N = 2 D = 6

Discrétisation de l'espace à deux dimensions pour 111 observations réparties en trois classes

Figure II.6



Histogramme des 111 observations de la figure II.11

Figure II.7

II - 6 . CONCLUSION

Dans ce chapitre, nous avons montré comment un ensemble d'observations multidimensionnelles pouvait être représenté sous la forme d'un histogramme multidimensionnel. Cet histogramme reflète l'assignation des observations aux hyperparallélépipèdes ou cellules utilisés pour discrétiser l'espace de représentation des données.

La recherche des modes par utilisation de l'histogramme, au lieu de la manipulation des données brutes comme dans les méthodes hierarchiques par exemple, réduit considérablement le nombre d'opération à effectuer. Cela permet d'envisager la classification d'échantillon de grande taille.

Nous avons constaté que le choix du pas de discrétisation de l'espace de représentation des données conditionne grandement la forme de l'histogramme. Sans connaissance a priori sur les données à analyser, il est difficile d'ajuster la finesse de la discrétisation. Nous présentons, dans le chapitre V, une méthode pour définir ce choix.

Dans le chapitre IV, nous utilisons l'histogramme comme point de départ pour déterminer les modes de la distribution des observations dans un contexte non supervisé et ainsi affecter les observations aux différentes classes.

Le chapitre suivant est consacré à des aspects algorithmiques de la construction de cet histogramme.

CHAPITRE III
CONSTRUCTION D'UN HISTOGRAMME MULTIDIMENTIONNEL PAR BALAYAGE SPATIAL DE L'ESPACE DISCRETISE

CHAPITRE III
CONSTRUCTION D'UN HISTOGRAMME MULTIDIMENSIONNEL PAR BALAYAGE SPATIAL DE L'ESPACE DISCRETISE

III - 1 . INTRODUCTION

Comme dans bien d'autres techniques de classification de données multidimensionnelles, nous effectuerons, dans la technique que nous proposerons au chapitre IV, des balayages successifs de l'espace à N dimensions, discrétisé selon la procédure décrite au chapitre précédent. Rappelons que lorsque les plages de variation des attributs sont divisées en un même nombre D d'intervalles adjacents, l'espace de représentation des données est discrétisé sur une grille constituée de $(D + 1)^N$ points qui sont les sommets des D^N hyperparallélépipèdes définis au chapitre II. Plus la dimension N des observations et/ou le nombre d'intervalles D augmentent, plus le nombre de points de discrétisation de l'espace est important. Cette explosion exponentielle du nombre d'informations élémentaires à manipuler risque d'entraîner des temps de calcul prohibitifs et l'espace mémoire requis peut devenir très important.

Les techniques de balayage classiques d'un espace discrétisé résultent généralement de la répétition d'une séquence élémentaire associée à chacune des N dimensions de cet espace. Au niveau de la programmation, il s'agit donc de réaliser N boucles analogues.

Le fait que ce nombre N dépende des données à analyser entraîne trop des lourdeurs au niveau de la programmation pour que les algorithmes développés puissent être utilisables quelle que soit la dimension des données.

Pour améliorer les temps d'exécution des procédures, tout en réduisant l'espace mémoire nécessaire et en allégeant la programmation, nous proposons un algorithme de balayage de l'espace multidimensionnel discrétisé qui :

- réduit le problème multidimensionnel à un problème à une seule dimension,
- ne prend en considération que les points de l'espace discrétisé situés dans les régions où des observations sont effectivement présentes.

III - 2. REDUCTION DE L'ADRESSAGE MULTIDIMENSIONNEL DES POINTS DE DISCRETISATION À UN ADRESSAGE UNIDIMENSIONNEL.

Le but de la procédure proposée dans ce chapitre est de générer une adresse unidimensionnelle α, définie par un seul paramètre, à partir des éléments a_n de l'adresse multidimensionnelle A.

Nous avons vu, dans le chapitre II, que la première étape pour générer un histogramme multidimensionnel consiste à

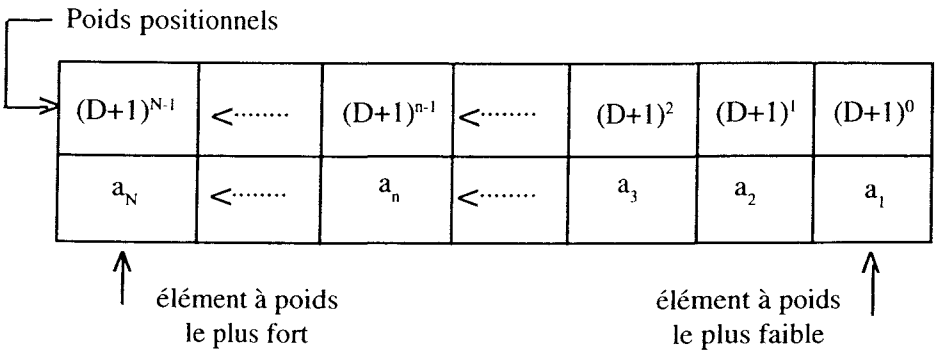
discrétiser l'espace à N dimensions sur une grille définissant des hyperparallélépipèdes adjacents.

Cette technique de discrétisation est très générale et a été utilisée dans de très nombreux algorithmes de classification multidimensionnelle. Avec cette procédure, chacun des 2^N sommets de chaque cellule élémentaire possède une adresse à N dimensions qui, selon les notations introduites auparavant, prend la forme $A = (a_1, a_2 ..., a_n, ... a_N)$ (Cf. § II - 4 - 1). Rappelons que chaque élément a_n, $n = 1, 2, 3, ..., N$, de cette adresse est un entier positif tel que :

$$0 \leq a_n \leq D \qquad \text{III-1}$$

qui, sous la forme d'un index, définit la position de ce sommet par rapport aux intervalles de discrétisation de la $n^{ième}$ dimension.

Cette adresse multidimensionnelle A peut être considérée comme exprimée dans un système de numération à poids positionnels à base (D+1). Chaque élément a_n est affecté d'un poids exprimé par $(D+1)^{n-1}$. L'élément affecté du poids le plus faible est a_1 et celui du poids le plus fort est a_N. Le tableau III-1 indique le poids de chacun des éléments a_n.



Les poids positionnels $(D+1)^{N-1}$ des éléménts a_n de l'adresse multidimensionnelle $A = (a_1, a_2, ...,a_n,...,a_N)$ exprimés dans la base (D+1)

Tableau III - 1

L'adressage unidimensionnel $\propto$ n'est donc rien d'autre que la conversion d'une adresse exprimée dans un système de numération à base (D + 1) en une adresse dans un système à base 10. Cette conversion est donnée par :

$$\propto = a_N(D+1)^{N-1} + ... + a_n(D+1)^{n-1} + ... a_3(D+1)^2 + a_2(D+1) + a_1 \quad \text{III-2}$$

Le tabeau III-2, donne la suite des adresses multidimensionnelles exprimées dans la base (D+1) pour N= 2 et D= 4 et leurs équivalences unidimensionnelles exprimées dans la base 10.

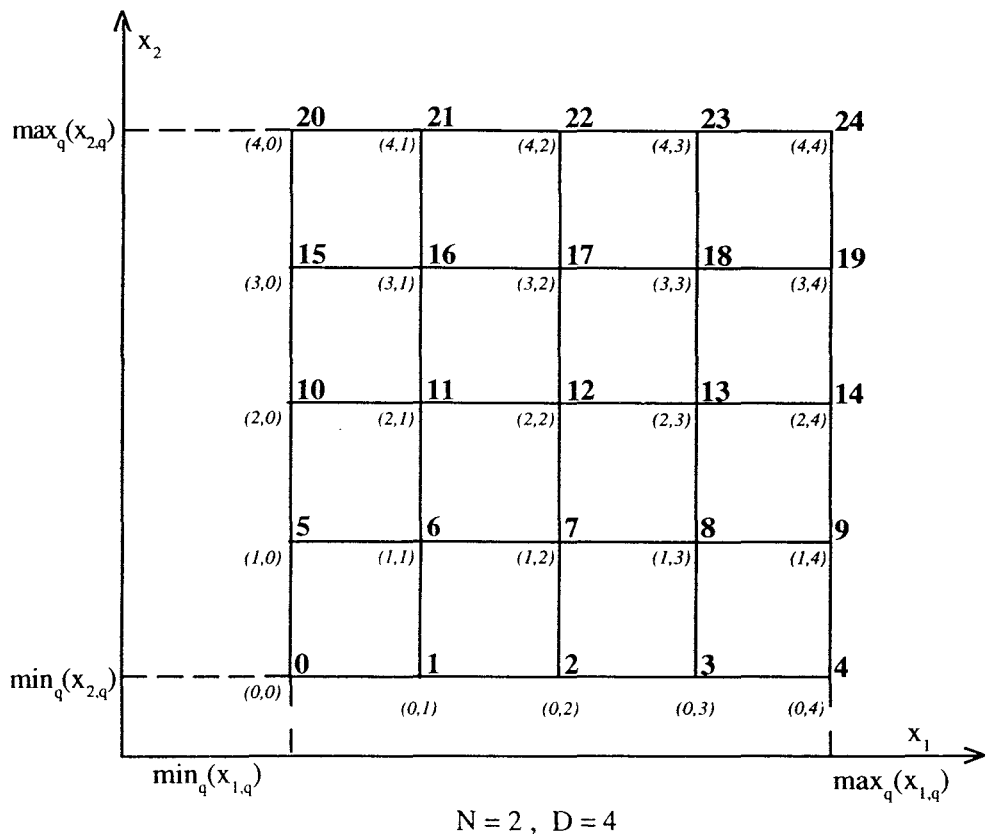
La figure III.1 indique les correspondances entre les adressages multidimensionnel et unidimensionnel aux sommets des cellules d'un espace discrétisé à deux dimensions.

La figure III.2 illustre ce codage pour un espace discrétisé à trois dimensions.

$(D+1)^1$ a_2	$(D+1)^0$ a_1	⟶	α
0	0	⟶	0
0	1	⟶	1
0	2	⟶	2
0	3	⟶	3
0	4	⟶	4
1	0	⟶	5
1	1	⟶	6
1	2	⟶	7
1	3	⟶	8
1	4	⟶	9
2	0	⟶	10
2	1	⟶	11
2	2	⟶	12
2	3	⟶	13
2	4	⟶	14
3	0	⟶	15
3	1	⟶	16
3	2	⟶	17
3	3	⟶	18
3	4	⟶	19
4	0	⟶	20
4	1	⟶	21
4	2	⟶	22
4	3	⟶	23
4	4	⟶	24

Suite des adresses multidimentionnelles et leurs équivalences unidimensionnelles pour D = 4 et N = 2

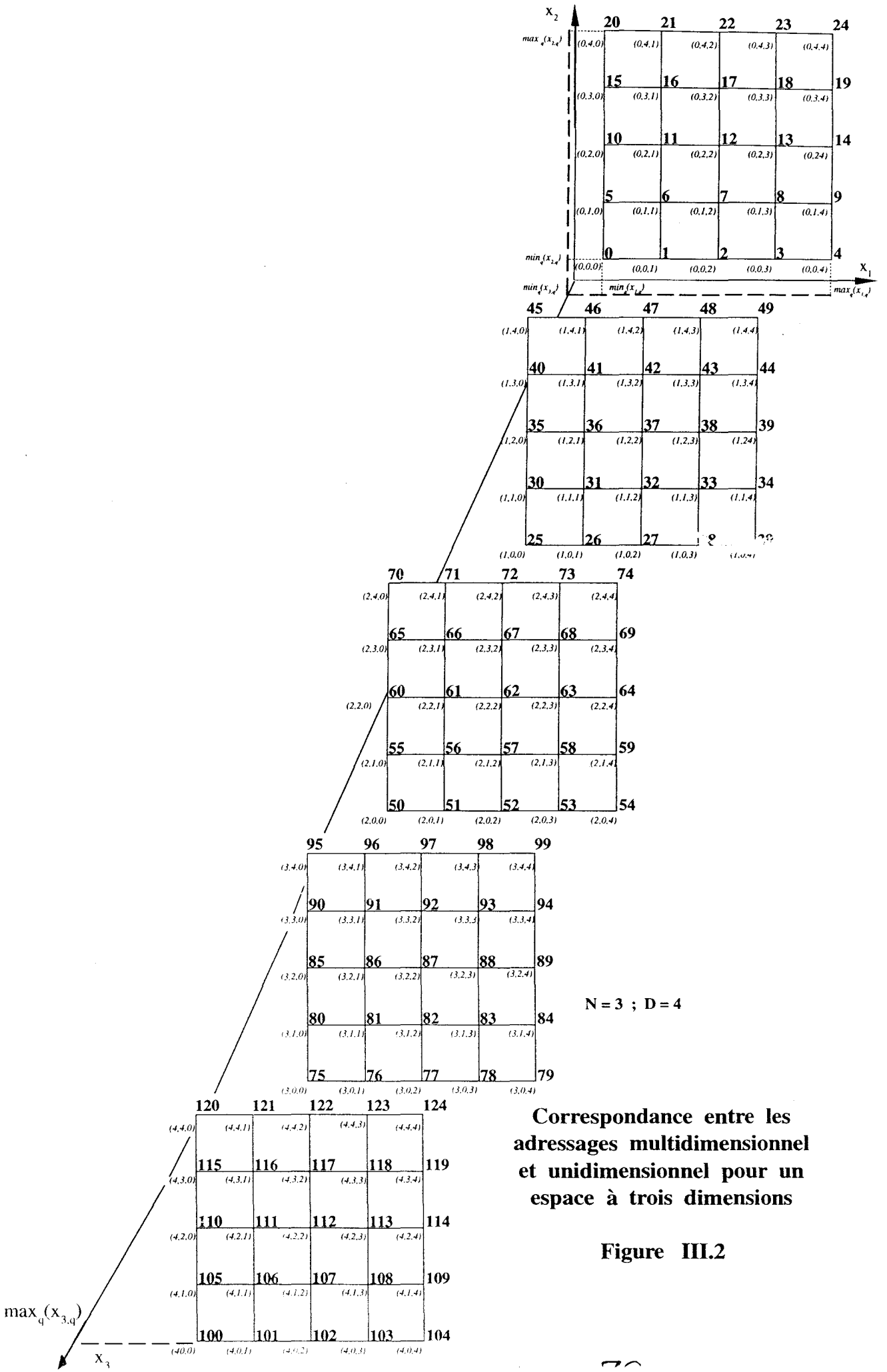
Tableau III-2



Lecture de la figure : 7 est l'adresse unidimensionnelle du sommet d'adresse bidimensionnelle (1, 2)

Correspondance entre les adressages multidimensionnel et unidimensionnel pour un espace à deux dimensions

Figure III.1



Correspondance entre les adressages multidimensionnel et unidimensionnel pour un espace à trois dimensions

Figure III.2

III - 3 . IDENTIFICATION DES ADRESSES UNIDIMENSIONNELLES DES 2^N HYPERPARALLELEPIPEDES SE PARTAGEANT UN MEME SOMMET.

Comme nous l'avons déjà précisé au chapitre II, dans un espace Euclidien à N dimensions, chaque sommet est commun à 2^N hyperparallélépipèdes. Afin d'identifier chacune de ces 2^N cellules, nous proposons de les repérer par des adresses unidimensionnelles de manière analogue au repérage défini pour les sommets. Chaque cellule n'ayant qu'un seul et unique centre, nous affectons l'adresse unidimensionnelle de l'hyperparallélépipède à ce centre. Tous les centres des cellules constituent une grille appelée "grille des centres".

Celle-ci est simplement obtenue par la translation de matrice $[+0{,}5\ C(n)]\ [I_N]$ de la grille des sommets, où $[I_N]$ est la matrice unité de taille N et C(n) est la largeur de l'intervalle sur le $n^{ième}$ axe, donnée par (Cf. § II-4-1) :

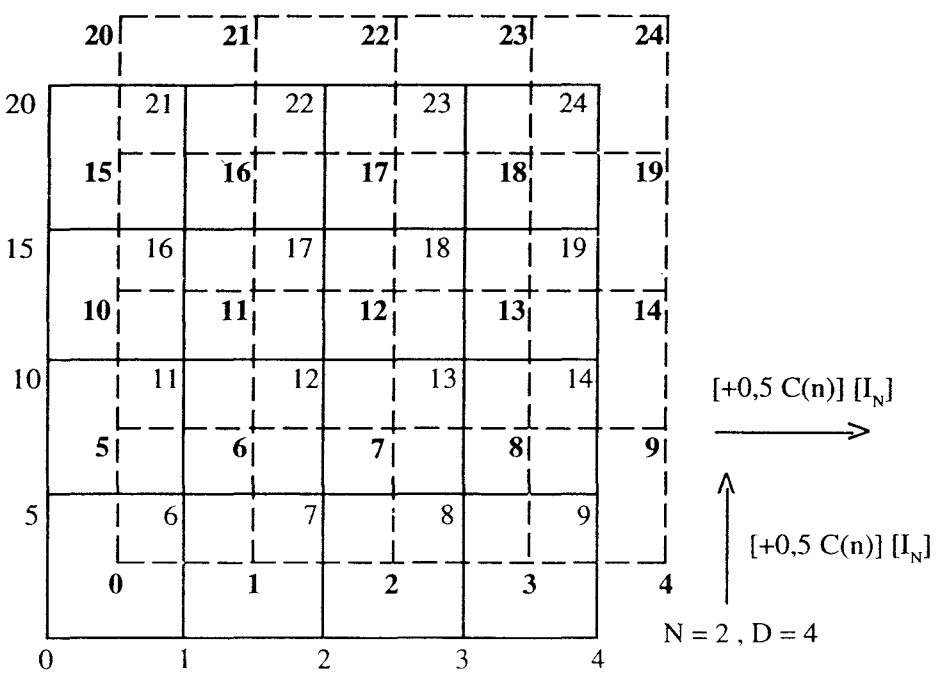
$$C(n) = \frac{\max_q(x_{n,q}) - \min_q(x_{n,q})}{D} \qquad \text{III-3}$$

Il est à noter qu'à chaque centre de cellule est associée pour chaque attribut, une valeur $x\ (a_n = d)$ donnée par

$$x\ (a_n = d) = \min_q\ (\ x_{n,q}\ ) + (d - \frac{1}{2})\ \frac{C(n)}{2} \qquad \text{III-4}$$
$$d = 1, 2, 3, ..., D$$

Nous remarquons que, sur la grille des centres, les adresses unidimensionnelles des cellules sont données par le même codage que celui utilisé pour repérer les sommets et défini par l'équation III.3 (Cf. figure III.3). Nous pouvons donc dire que les figures III.1 et III.2 peuvent représenter soit la grille des sommets, soit celle des centres.

Lors du déroulement d'une procédure de traitement de données nécessitant un balayage de tout l'espace, nous devons pouvoir définir les adresses unidimensionnelles des 2^N cellules se partageant un même sommet à partir de la seule connaissance de l'adresse unidimensionnelle de ce sommet.



Correspondance des adresses unidimensionnelles respectives sur les deux grilles des sommets et des centres

Figure III.3

Comme il apparaît de façon évidente sur la figure III.3, une des 2^N cellules qui se partagent un sommet possède toujours la même adresse unidimensionnelle que celle de ce sommet. Cette propriété résulte de la correspondance des 2 grilles par translation de matrice $[+0,5\ C(n)]\ [I_N]$. Donc, si l'adresse unidimensionnelle du sommet est α, une des 2^N cellules aura un centre de même adresse unidimensionnelle α. Pour trouver les adresses unidimensionnelles des autres cellules nous procédons par étapes.

La première étape de la procédure consiste à trouver l'adresse unidimensionnelle d'un deuxième centre de cellule ayant ce même sommet en commun. Cette adresse est obtenue en soustrayant la quantité $(D+1)^0$ de l'adresse unidimensionnelle α du premier centre trouvé. Ce nouveau centre aura donc pour adresse unidimensionnelle $[\alpha - (D+1)^0]$, soit $(\alpha - 1)$.

La deuxième étape de la procédure consiste à trouver les adresses unidimensionnelles de deux autres centres en soustrayant la quantité $(D+1)^1$ aux adresses unidimensionnelles des deux centres déjà trouvés. Ces deux nouveaux centres auront donc les adresses unidimensionnelles $[\alpha - (D+1)^1]$ et $[(\alpha - 1) - (D+1)^1]$.

En itérant ce processus, la procédure de recherche des adresses unidimensionnelles des 2^N hyperparallélépipèdes

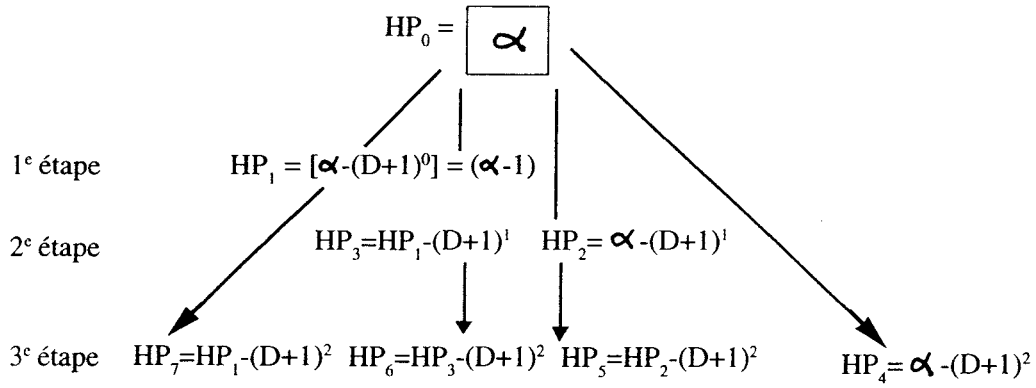
ayant en commun le sommet S(α) d'adresse unidimensionnelle α peut se résumer comme suit :

- à l'étape initiale, on considère le centre ayant la même adresse unidimensionnelle que le sommet considéré,

- à chaque étape, on identifie un certain nombre de nouvelles cellules se partageant le sommet considéré. Leurs adresses unidimensionnelles sont obtenues à partir des adresses unidimensionnelles des cellules déjà identifiées aux cours des étapes précédentes. Ainsi à la $m^{ième}$ étape, on calcule les adresses unidimensionnelles de 2^{m-1} nouveaux hyperparallélépipèdes à partir des 2^{m-1} adresses unidimensionnelles trouvées aux étapes précédentes, en soustrayant à chacune d'elles la quantité $(D+1)^{m-1}$,

- la procédure s'arrête lorsque le nombre d'étapes est égal au nombre de dimensions N, donc lorsque $m = N$.

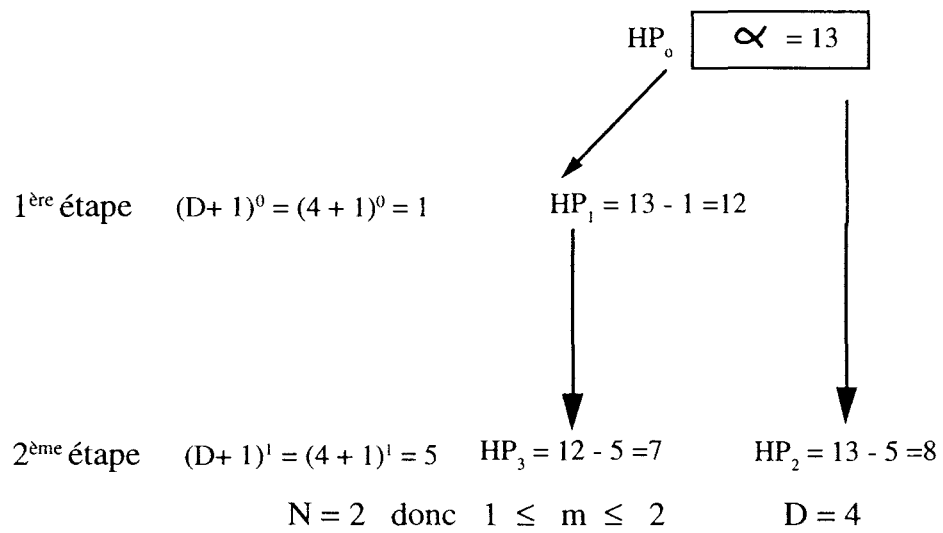
Cette procédure de chaînage peut-être représentée par la structure arborescente de la figure III.4 pour le cas général à trois dimensions. La figure III.5 correspond à la recherche des trois hyperparallélépipèdes se partageant le sommet d'adresse unidimensionnlle α = 13 de la figure III.3.



HP_m = $m^{ième}$ adresse trouvée d'un hyperparallélépipède

Procédure de chaînage déterminant les adresses unidimensionnelles des hyperparallélépipèdes ayant en commun le sommet S (∝)

Figure III.4



Détermination des adresses unidimensionnelles des hyperparallélépipèdes ayant en commun le sommet S(∝) d'adresse ∝ = 13 pour N = 2 et D = 4

Figure III.5

L'annexe III-1 indique le programme nécessaire à l'exécution de cette procédure de chaînage pour n'importe quel nombre de dimensions N. On remarque que ce programme ne comporte que peu de boucles, ce qui assure une simplicité et un temps d'exécution rapide. La complexité de l'algorithme et sa vitesse d'exécution seront analysées ultérieurement (Cf. § III-7-3).

III - 4. IDENTIFICATION DES ADRESSES UNIDIMENSIONNELLES DES 2^N SOMMETS D'UN HYPER-PARALLELEPIPEDE.

Un second problème, que l'on rencontre régulièrement lors du déroulement des procédures de traitement de données nécéssitant un balayage de tout l'espace, consiste à déterminer les adresses unidimensionnelles des 2^N sommets d'un hyperparallélépipède à partir de la seule connaissance de son adresse unidimensionnelle. La recherche des adresses unidimensionnelles de ces sommets se fera en N étapes comme pour le cas de la détermination des adresses unidimensionnelles des 2^N cellules se partageant un même sommet.

Comme nous l'avons précisé précédement (Cf. § III-3), cette procédure itérative d'identitification des adresses unidimensionnelles des 2^N cellules se partageant le même sommet débutait en obtenant la grille des centres par la translation de la grille des sommets de matrice $[+0,5\ C(n)]\ [I_N]$. Dans ces conditions, les adresses unidimensionnelles des 2^{m-1} cellules définies à la $m^{ième}$ étape de la procédure étaient

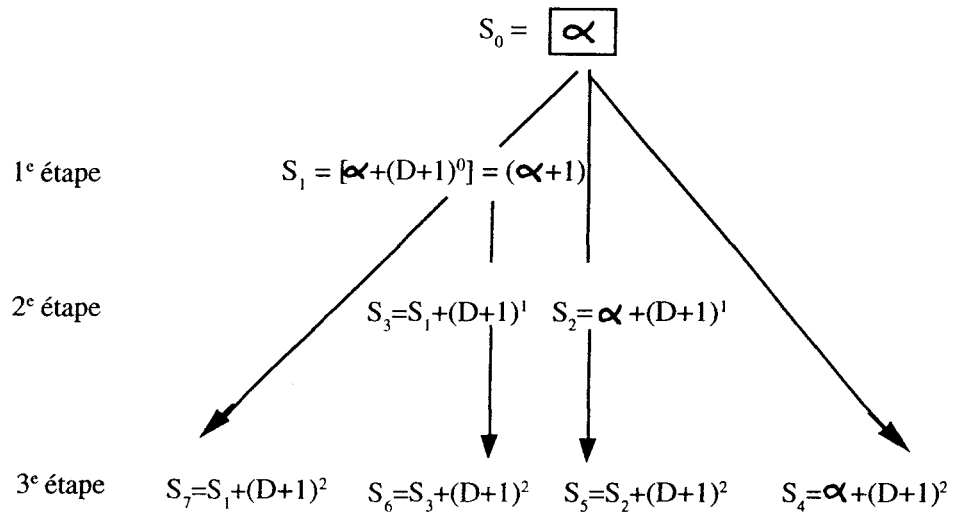
obtenues en soustrayant la quantité $(D+1)^{m-1}$ aux adresses unidimensionnelles des cellules identifiées aux étapes précédentes.

Comme l'illustre la figure III.3, le passage inverse de la grille des centres à celle des sommets est obtenu par une translation opposée de matrice [-0,5 C(n)] $[I_N]$. C'est pourquoi, à la $m^{ième}$ étape de la procédure, les adresses unidimensionnelles des 2^{m-1} sommets d'une même cellule sont obtenues en additionnant la quantité $(D+1)^{m-1}$ aux adresses undimensionnelles des sommets obtenus aux étapes précédentes. La procédure s'arrête lorsque le nombre d'étapes est égal au nombre de dimensions.

Cette procédure de chaînage peut être représentée par la structure arborescente de la figure III.6 pour le cas général à trois dimensions, et par celle de la figure III.7 pour le cas de l'espace à deux dimensions de la figure III.3, avec $\alpha = 13$ pour adresse unidimensionnelle de départ.

III - 5 . NOTION DE VOISINAGE

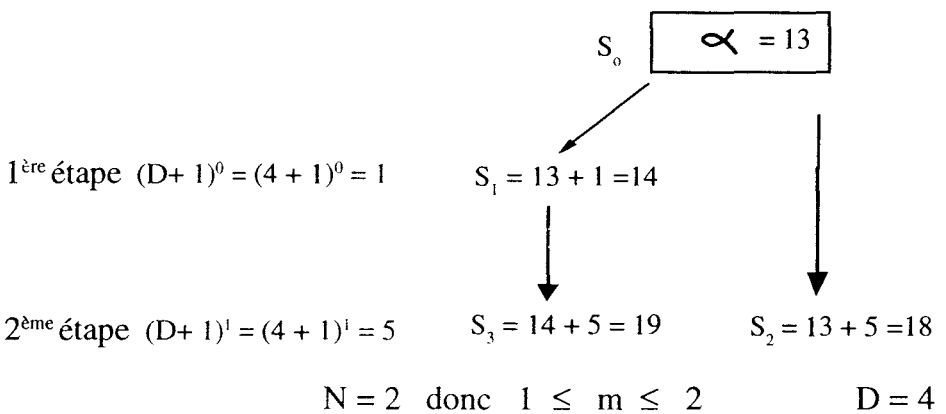
Sur la grille de discrétisation, on définit la notion classique de voisinage entre les cellules hyperparallélépipédiques élémentaires. Dans tout ce qui suit, nous appellerons "voisinage" de la cellule d'adresse unidimensionnelle α, l'ensemble des (3^N-1) cellules qui ont au moins un sommet en commun avec celle-ci. La figure III.8 représente l'ensemble des 26 cellules voisines de la cellule centrale pour un espace à trois dimensions.



S_m = m^ième adresse trouvée d'un sommet

Procédure de chaînage déterminant les adresses unidimensionnelles des sommets d'une cellule d'adresse unidimensionnelle α.

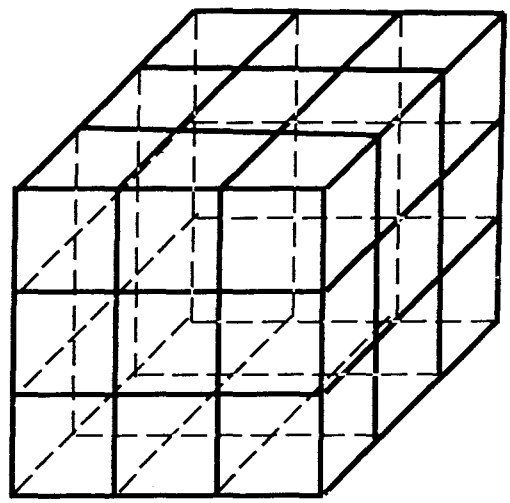
Figure III.6



N = 2 donc 1 ≤ m ≤ 2 D = 4

Détermination des adresses unidimensionnelles des sommets d'une cellule d'adresse α = 13 pour N = 2 et D = 4

Figure III.7



N = 3 la cellule centrale est entourée de 26 cellules voisines

Ensemble des 26 cellules voisines de la cellule centrale par N = 3

Figure III.8

III - 5 -1 . Identification des (3^N-1) hyperparalélépipèdes voisins de celui d'adresse unidimensionnelle α .

La procédure de recherche des adresses unidimensionnelles des cellules constituant le voisinage d'une cellule donnée se fait par étapes, selon un schéma comparable à celui de la recherche des cellules se partageant un même sommet.

La première étape consiste à trouver, à partir de l'adresse unidimensionnelle α de la cellule considérée, l'adresse unidimensionnelle des deux cellules adjacentes dans la direction de la première dimension x_1 de l'espace de représentation de données. L'une de ces deux adresses unidimensionnelles est obtenue en additionnant la quantité $(D + 1)^0$ à l'adresse unidimensionnelle α , l'autre en soustrayant cette même quantité de l'adresse unidimensionnelle α . Les deux nouvelles adresses unidimensionnelles ainsi obtenues sont $[\alpha + (D+1)^0]$ et $[\alpha - (D+1)^0]$, soient $(\alpha +1)$ et $(\alpha -1)$.

La deuxième étape de la procédure consiste à trouver les adresses unidimensionnelles de six autres hyperparallélépipèdes en explorant le voisinage dans la direction de la seconde dimension de l'espace, x_2. Trois de ces six adresses unidimensionnelles sont obtenues en additionnant la quantité $(D + 1)^1$ aux trois adresses unidimensionnelles précédement trouvées, les trois autres résultent de la soustraction de la même quantité $(D + 1)^1$ de ces mêmes trois adresses unidimensionnelles obtenues à l'étape précédente.

La procédure de définition des adresses unidimensionnelles des $(3^N - 1)$ cellules adjacentes à une cellule d'adresse unidimensionnelle α se résume comme suit :

- à l'étape initiale, on considère la cellule centrale du voisinage,

- à chaque étape, on identifie un certain nombre de cellules voisines d'une cellule considérée. Leurs adresses unidimensionnelles sont obtenues à partir des adresses unidimensionnelles des cellules déjà identifiées au cours des étapes précédentes. Ainsi, à la $m^{ième}$ étape, nous recherchons les adresses unidimensionnelles de $2(3^{m-1})$ nouveaux hyper-parallélépipèdes à partir des (3^{m-1}) adresses unidimensionnelles trouvées aux étapes précédentes. (3^{m-1}) de ces nouvelles adresses unidimensionnelles sont obtenues en additionnant la quantité $(D+1)^{m-1}$ à chacune des adresses unidimensionnelles trouvées aux étapes précédentes, les (3^{m-1}) autres résultent de la soustraction de cette même quantité $(D+1)^{m-1}$ de ces mêmes adresses unidimensionnelles obtenues précédement.

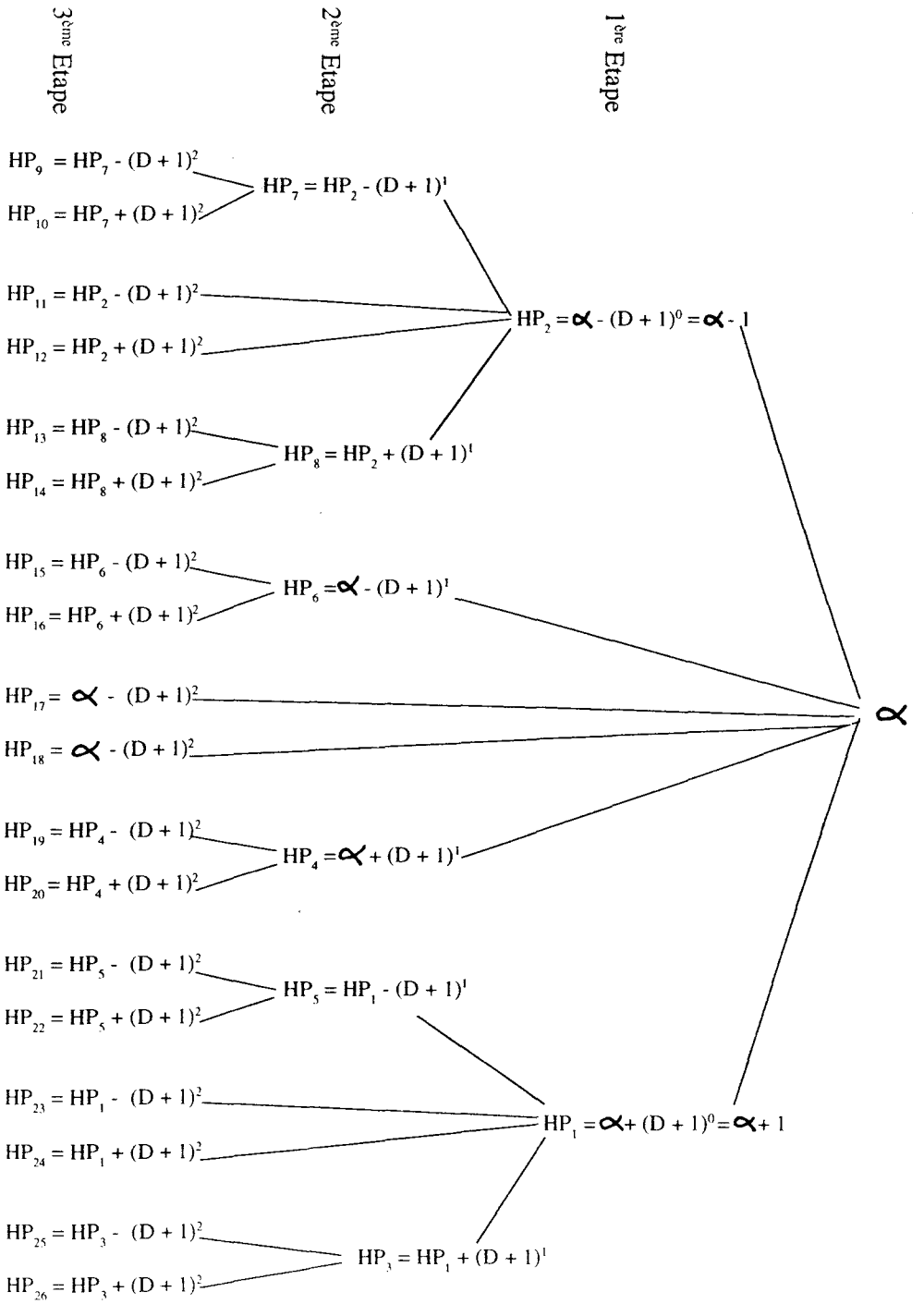
- la procédure s'arrête lorsque le nombre d'étapes est égal au nombre de dimensions N, donc lorsque $m = N$.

Cette procédure de chaînage est représentée par la figure III.9 pour le cas général à trois dimensions. La figure III.10 illustre cette procédure pour le cas de l'espace à deux dimensions de la figure III.1. La figure III.11 correspond au cas de l'espace à trois dimensions de la figure III.2.

L'annexe III.2 donne le programme nécessaire à l'exécution de cette procédure de chaînage pour n'importe quel nombre de dimensions N. On remarquera la simplicité de ce programme qui ne comporte que peu de boucles, ce qui assure un temps d'exécution rapide. La complexité de l'algorithme et sa vitesse d'exécution seront analysées ultérieurement (Cf. § III.7.3).

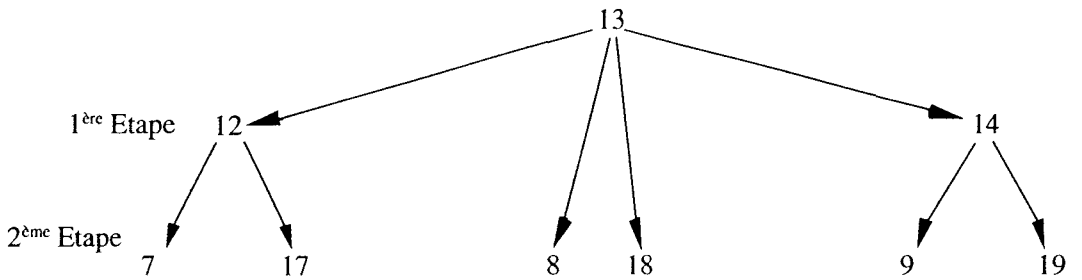
III - 5 - 2 . Effet de bord

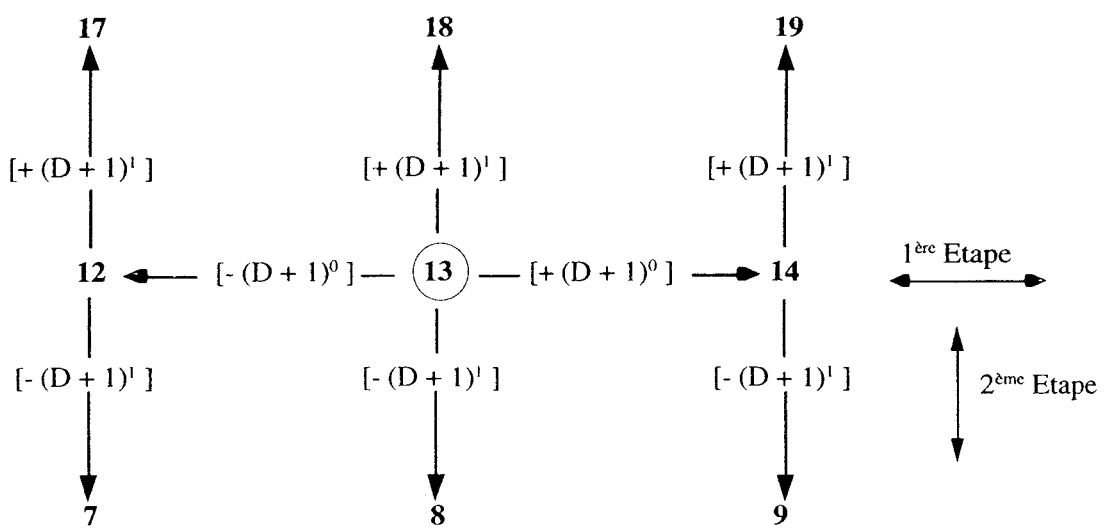
Comme nous l'avons mentionné au paragraphe III-2, l'adressage unidimensionnel des cellules établit une continuité de numérotation entre les cellules voisines selon l'ordre défini par l'équation III-2. Cependant, dans certains cas, ce système de codage donne deux adresses unidimensionnelles consécutives à deux cellules non voisines. Ainsi, sur la figure III.1, on constate que les cellules d'adresses unidimensionnelles 4 et 5, bien que n'étant pas voisines, ont deux adresses unidimensionnelles consécutives. Sur la figure III.2, les cellules d'adresses unidimensionnelles 24 et 25 sont dans le même cas. Cette continuité du codage entre cellules non voisines risque d'affecter la reconstruction du voisinage des hyper-parallélépipèdes en y intégrant des éléments non connexes.



Identification des (3^N-1) hyperparallélépipèdes voisins de celui d'adresse unidimensionnelle α .

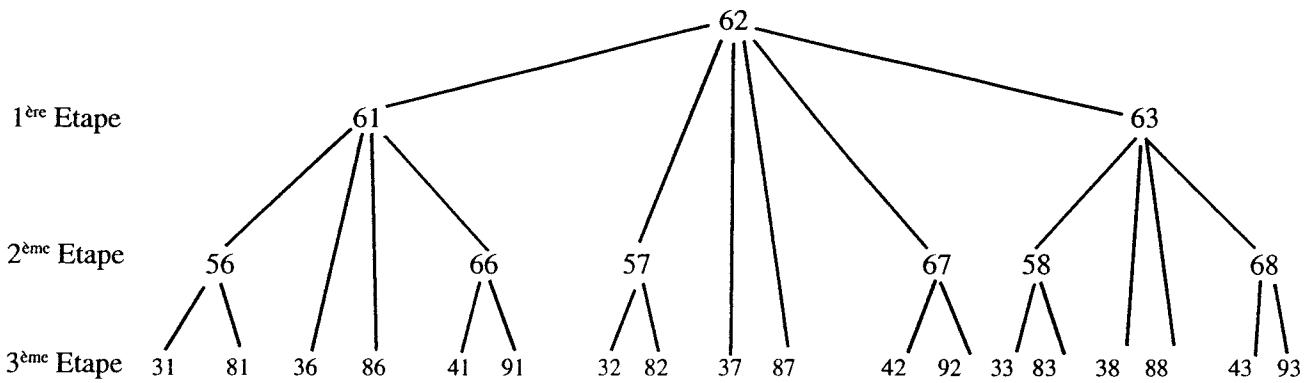
Figure III.9

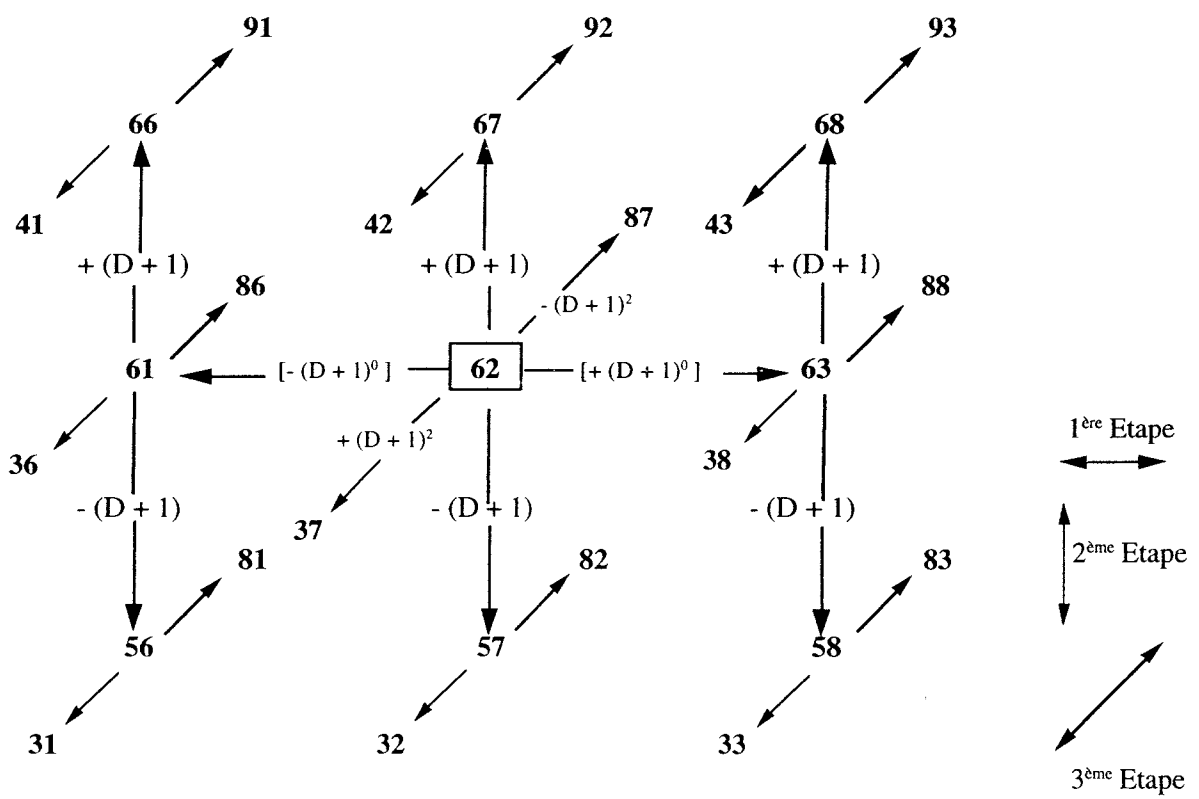




Identification des (3^N-1) hyperparalléiépipèdes voisins de celui d'adresse unidimensionnelle 13 pour N = 2 , D = 4

Figure III.10





Identification des $(3^N - 1)$ hyperparallélépipèdes voisins de celui d'adresse unidimensionnelle 62 pour N = 3 , D = 4

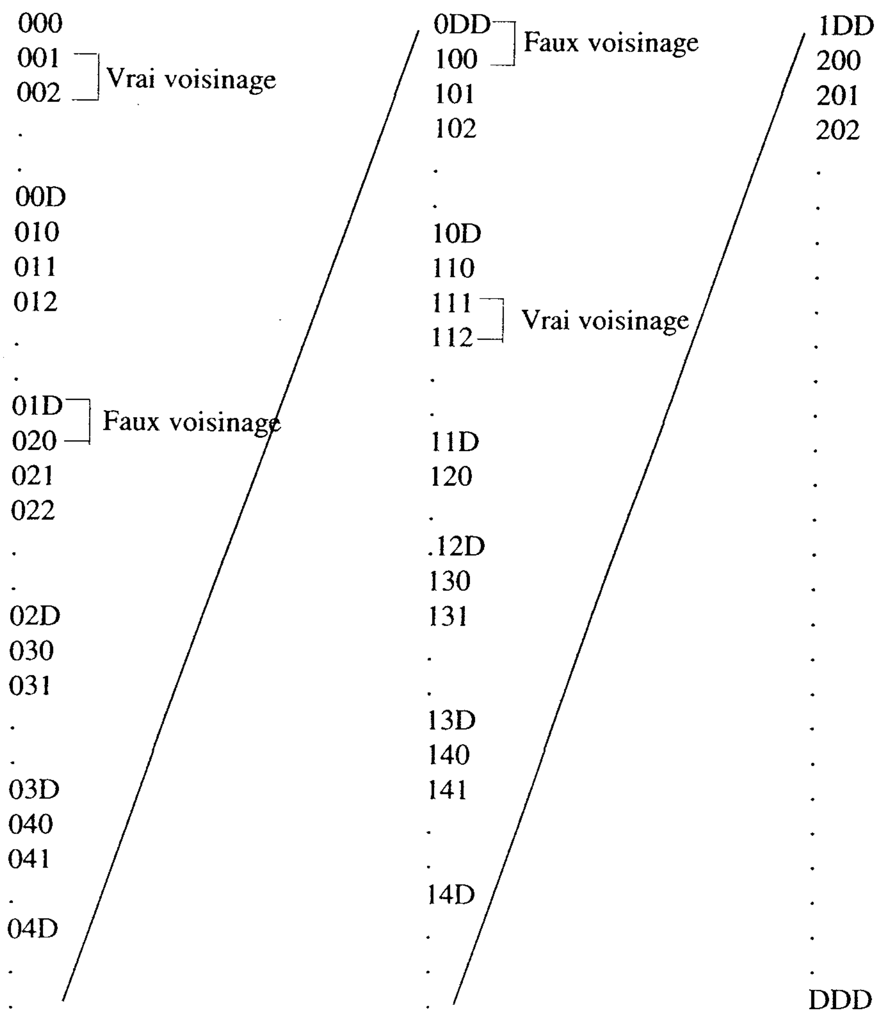
Figure III.11

En comparant la suite des adresses multidimensionnelles et leurs équivalences unidimensionnelles données par le tableau III-2 ainsi que leur affectation aux sommets des cellules de l'espace discrétisé à deux dimensions de la figure III.1, nous constatons que si deux cellules possèdent deux adresses unidimensionnelles consécutives, nous avons soit le cas de cellules connexes (Cf. § III-4-1), soit le cas de cellules non voisines. Nous appelons ce dernier cas "faux voisinage".

Le tableau III.3 donne la suite des adresses multidimensionnelles et leurs équivalences unidimensionnelles pour un espace à trois dimensions pour n'importe quelle nombre d'intervalles D choisi pour discrétiser cet espace. Sur ce tableau sont précisés quelques cas de vrai et de faux voisinages.

Pour les cas de voisinage réel entre deux cellules d'adresses unidimensionnelles consécutives, nous remarquons que tous les éléments de ces deux adresses multidimensionnelles sont identiques, sauf les éléments de poids le plus faible, a_1, qui différent d'une unité.

Si, par contre, ces deux adresses unidimensionnelles consécutives différent par le changement simultané des m premiers éléments de la première adresse de la valeur D à la valeur zéro et si en même temps l'élément de poids (m+1) augmente d'une unité, nous avons un cas de "faux voisinage" entre les deux cellules considérées.



Suite des adresses multidimensionnelles et leurs équivalences unidimensionnelles pour N = 3 et n'importe quelle valeur de D.

Tableau III - 3

Cette situation correspond à un changement de valeur des m premiers éléments de D à O, c'est-à-dire de la valeur associée à un bord extrême de l'espace discrétisé à la valeur associée au bord opposé. Pour cette raison, nous avons appelé ce phénomène "l'effet de bord".

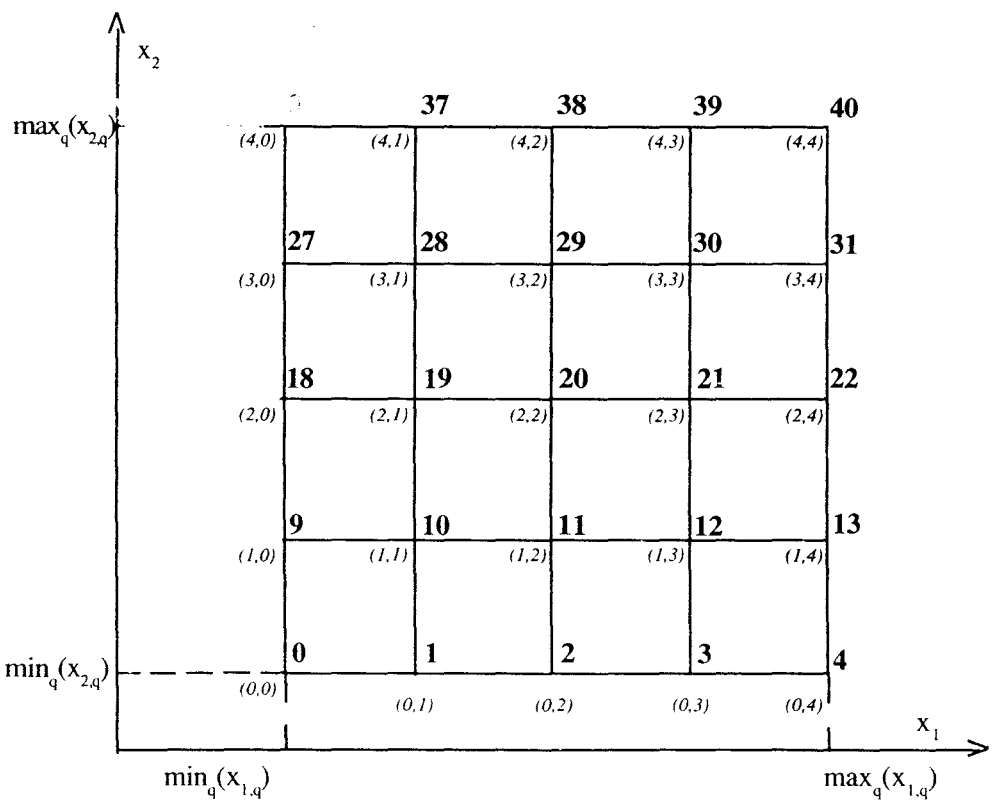
Pour remédier à cet effet, nous proposons d'introduire un écart artificiel de numérotation entre tout couple d'hyperparallélépipèdes se trouvant dans cette situation de faux voisinage. Pour cela, toutes les quantités (D+1) de l'équation III.2 définissant l'adresse unidimensionnelle à partir de l'adresse multidimensionnelle sont remplacées par la valeur (D+p), où p est un entier positif supérieur à l'unité. L'écart artificiel sera donc égal à (p - 1) et l'équation permettant le calcul de l'adresse unidimensionnelle sera donnée par :

$$\alpha = a_N(D+p)^{N-1} + \ldots + a_n(D+p)^{n-1} + \ldots + a_3(D+p)^2 + a_2(D+p)^1 + a_1 \qquad \text{III-5}$$

Nous remarquons qu'il n'y a aucune contrainte sur le choix du nombre p, si ce n'est qu'il doit être plus grand que l'unité.

Dans tout ce qui suit, p sera pris égal à 5. La figure III.1.2, présente la discrétisation de l'espace Euclidien à deux dimensions de la figure III.1 pour p = 5.

Nous remarquons que les cellules (4) et (5) de la figure III.1 ne peuvent plus être interprétées comme voisines car un écart artificiel égal à (p-1), c'est-à-dire égal à 4, apparaît entre leurs nouvelles adresses qui sont devenues respectivement égales à (4) et (9).



Discrétisation de l'espace
pour N = 2 , D = 4 et p = 5

Figure III.12

III - 6 . CONSTRUCTION DE L'HISTOGRAMME PAR SÉLECTION ET CALCUL DES ADRESSES DES SOMMETS ASSOCIÉS AUX OBSERVATIONS.

Dans le chapitre II, nous avons vu que la génération de l'histogramme consiste à assigner chaque observation à son sommet associé S (A) d'adresse multidimensionnelle A. Le nombre d'observations renfermées dans les 2^N cellules ayant ce même sommet S(A) en commun et donc lui étant assignées, indique la valeur H(A) de l'histogramme.

Le calcul des adresses multidimensionnelles A de ces sommets occupés S(A) se fait au moyen des équations II-2 (Cf. § II-4-1) et II-3 (Cf. § II-4-2) et ce à partir du fichier de dimension (Q x N) des vecteurs des attributs, ou observations multidimensionnelles X_q. A et H(A) sont mémorisés dans une table de dimension maximum $(N + 1) * (D + 1)^N$ (Cf. § II-4-2). A partir de cette table, les adresses unidimensionnelles et les valeurs correspondantes de l'histogramme $H(\alpha)$ sont calculées au moyen de l'équation III-4. (Cf. § III-5-2).

Quand la dimension des données est élevée, il faut réduire l'espace mémoire utilisé tout en minimisant le temps d'exécution.

Pour cela, nous avons :

- sélectionné uniquement les sommets associés à des observations : c'est-à-dire les sommets occupés. Les autres sommets sont ignorés, car aucune observation ne leur est associée : ils correspondent à des parties de l'espace vides d'observations.

- calculé l'adresse unidimensionnelle α des sommets occupés directement à partir du fichier des vecteurs des attributs sans avoir à définir et à mémoriser l'adresse multidimensionnelle A. Ce calcul direct, programmé en langage C, est décrit dans l'annexe III-3. Afin de ne pas avoir à manipuler d'adresses négatives, toutes les adresses intervenant dans cette procédure sont translatées d'une quantité positive qui dépend de la dimension N des données. Cette légère modification de l'algorithme est justifiée en détail dans cette même annexe. Il est à noter que la structure de ce programme est indépendante du nombre total Q d'observations à classer ainsi que de la dimension N des données.

La table qui contient les adresses α des sommets $S(\alpha)$ occupés ainsi que la valeur de l'histogramme $H(\alpha)$ définit l'histogramme multidimensionnel.

Pour sélectionner uniquement les sommets associés aux observations, on explore séquentiellement le fichier de données. On calcule pour chaque observation l'adresse unidimentionnelle du sommet qui lui est associé. Cette adresse est comparée aux adresses des sommets sélectionnés précédement au cours de la lecture sequentielle du fichier des données. Si cette adresse est déjà incluse dans la liste, la valeur de l'histogramme qui lui est assignée est incrémentée de une unité. Autrement, cette nouvelle adresse est ajoutée à la liste des adresses en lui assignant une valeur de l'histogramme égale à l'unité. Après avoir pris en considération toutes les observations disponibles, toutes celles affectées à un même sommet se trouvent regroupées et sont repérées par l'adresse commune de ce sommet, donnée par la liste des adresses.

Sachant que l'utilisation des histogrammes multidimensionnels pose des problèmes d'implantation quand il s'agit de traiter des données de dimension élevée et sachant que la procédure proposée a réduit le problème multidimensionnel à un problème unidimensionnel, quelle que soit la dimension N de l'espace, il est important d'évaluer la complexité de la procédure proposée, tant sur le plan de l'espace mémoire nécessaire, que sur celui des temps de calcul.

III - 7 . COMPARAISON ENTRE LES METHODES DE BALAYAGE CLASSIQUES ET LA METHODE DE BALAYAGE PROPOSEE

Pour démontrer l'intérêt de la méthode de balayage proposée, nous comparons une procédure de balayage d'un espace à N dimensions discrétisé suivant une technique classique et la procédure décrite ci-dessus. Cette comparaison porte sur 3 points :

- la taille de l'espace mémoire utilisé,
- la compléxité de la structure du programme d'exécution du balayage,
- la vitesse d'exécution ou temps de calcul.

III - 7 - 1 . L'espace mémoire utilisé

Lors de la construction de l'histogramme, si le nombre de sommets associés à une ou plusieurs observations est L, la table des adresses comporte un nombre L d'adresses auxquelles sont assignées des valeurs non nulles de l'histogramme. Dans les techniques de balayage classiques, si seulement les adresses des sommets associés sont considérées, la dimension de la table des adresses est (L * N) et celle des valeurs de l'histogramme multidimensionnel est (L * 1). L'espace mémoire total nécessaire est alors de dimension L * (N + 1) et donc proportionnel à la dimension N de l'espace des données.

	Eléments de l'adresse multidimensionnelle A						Valeur de l'histogramme
	a_N	.....	a_n	.....	a_2	a_1	H[A]
↑ L adresses ↓							

Table des adresses multidimensionnelles A et des valeurs de l'histogramme H(A) pour un espace à N dimensions avec L sommets occupés

Tableau III - 4

	Adresse unidimensionnelle	Valeur de l'histogramme
	α	H [α]
↑ L adresses ↓		

Table des adresses unidimensionnelles α et des valeurs de l'histogramme H (α) pour un espace à N dimensions avec L sommets occupés

Tableau III - 5

Le tableau III-4 montre la structure des tables des adresses et des valeurs de l'histogramme pour un espace à N dimensions avec L sommets occupés.

Par contre, avec la procédure de balayage proposée, la dimension de la table des adresses est L * 1, comme celle des valeurs de l'histogramme. La taille de l'espace mémoire total nécessaire est alors de taille 2 * L, et est donc indépendante de la dimension des données N. Le tableau III-5 indique la structure des tables des adresses et des valeurs de l'histogramme pour un espace à N dimensions avec L sommets occupés.

Le facteur de réduction de l'espace mémoire nécessaire avec la méthode proposée par rapport à l'espace mémoire utilisé avec la méthode classique est donné par :

$$\left(1 - \frac{2L}{L(N+1)}\right) \times 100\ \%$$

ou encore :

$$\frac{N - 1}{N+1} \times 100\ \% \qquad \text{III-6}$$

Ce facteur de réduction augmente avec la dimension N des données. Par exemple pour N = 4, ce facteur est de 60 %, alors qu'il atteint 77 % pour N = 8.

III - 7- 2 . Complexité de la structure du programme

Comme nous l'avons mentionné auparavant, les techniques de balayage classiques d'un espace discrétisé résultent généralement de la répétition d'une séquence élémentaire associée à chacune des N dimensions de cet espace.

Au niveau de la programmation, il s'agit donc de réaliser N boucles analogues. Si, par exemple, nous voulons connaître les valeurs de l'histogramme en tout point, nous avons un programme à **N** boucles de la forme :

```
for {a_N = 0 ; a_N ≤ D ; a_N ++}
    **********************
    for {a_n = 0 ; a_n ≤ D ; a_n++}
        ********************
        for {a_2 = 0 ; a_2 ≤ D ; a_2 ++}

            for {a_1 = 0 ; a_1 ≤ D ; a_1 ++}
                {
                    Instructions ;
                }                                   III-7
```

Cela implique, lors de l'écriture du programme, la définition a priori du nombre N de dimensions et la prévision d'une procédure spécifique pour modifier le nombre de boucles en fonction de N.

Les instructions sont donc exécutées $(D+1)^N$ fois.

Avec la méthode d'adressage unidimensionnel proposée pour l'exécution de la même procédure, et pour n'importe quelle valeur de N, nous n'employons qu'une seule boucle.

Le programme devient alors :

```
for { ∝ = 1 ; ∝ ≤ L ; ∝++}
    {
        Instructions ;
    }                                          III-8
```

L étant le nombre de sommets occupés. Les instructions sont donc exéxutées L fois.

III - 7 - 3 . Vitesse d'exécution et temps de calcul

Pour accéder à la valeur de l'histogramme en un point, nous devons chercher son adresse. Le temps d'accès au contenu d'une adresse à N index est plus grand que le temps d'accès à une adresse définie par un seul index. La vitesse d'accès à une information dans la procédure de balayage proposée est plus grande que celle des procédures classiques.

Aussi, puisque la méthode proposée ne prend en compte que les L sommets occupés, le balayage ne se fera que pour ces L points. La boucle définie par le programme III-8 donc exécuté seulement L fois.

Pour les méthodes classiques, le programme défini par l'équation III-7 nécessite $(D + 1)^N$ exécutions de la même procédure élémentaire. Le facteur de réduction du temps d'exécution est donc donné par :

$$[1- \frac{L}{(D + 1)^N}] \times 100\% \qquad \text{III-9}$$

Rappelons que $(D + 1)^N$ est le nombre total de sommets dans l'espace à N dimensions, quand chaque axe est divisé en D intervalles adjacents et égaux [illegible] est le nombre de sommets occupés. Donc $L \leq (D + 1)^N$ quels que soient L et D, ce qui implique que

$$0 < \frac{L}{(D+1)^N} < 1 \qquad \text{III-10}$$

Si, par exemple, les observations occupent la moitié des sommets, cette réduction est de 50 %, alors qu'elle atteint 90 % pour une occupation d'un dizième des sommets.

III - 8 . CONCLUSION.

La plupart des procédures de classification utilisent différents algorithmes de balayage de l'espace discrétisé. La complexité des algorithmes, le temps de calcul et la taille

mémoire nécessaire, sont tous trois dépendents de la dimension N de cet espace et à la finesse de la discrétisation.

Dans ce chapitre, nous avons introduit une méthode pour réduire le problème du balayage d'un espace multi-dimensionnel à un problème à une seule dimension. Il ne s'agit pas de réduire la dimension de l'espace en négligeant un certain nombre d'attributs, mais tout simplement d'effectuer une reformulation mathématique simplifiée, qui ne néglige aucun attribut, quelque soit leur nombre. Cette méthode est en fait une conversion des adresses multidimensionnelles des cellules et des sommets, exprimées dans un système de repérage naturel mais complexe, en des adresses scalaires exprimées en base 10.

D'autre part, dans le souci de réduire les temps de calcul, nous n'avons considéré que les points de l'espace discrétisé situés dans les régions où des observations sont effectivement présentes.

Nous avons montré que cette approche a considérablement réduit la complexité de l'algorithme de balayage et l'a rendu potentiellement intéressante pour des problèmes de classification dans des espaces de grande dimension, sans toutefois avoir à préciser a priori cette dimension. Enfin nous avons démontré que le temps d'éxécution ainsi que la taille mémoire nécessaire ont été considérablement réduits par rapport à l'utilisation de codages plus classiques et plus naturels, mais difficiles à manipuler.

ANNEXE III - 1

Identification des adresses unidimensionnelles des 2^N cellules se partageant un même sommet $S(\propto)$

Pour déterminer les adresses unidimensionnelles des 2^N cellules se partageant un sommet d'adresse $\propto$, nous devons procéder pas à pas à partir de cette adresse. A chaque étape de rang m, nous devons soustraire la quantité $(D+1)^{m-1}$ de toutes les adresses unidimensionnelles déjà trouvées. Ces adresses sont mémorisées dans une table à une dimension.

La procédure de soustraction des quantités $(D+1)^{m-1}$, m = 1, 2,....., N, N étant la dimension des données, se fait avec le formalisme du language C à partir de l'équation :

Ce[i] = Ce[j] - (pow ((D + 1) , (m - 1))) A III - 1 - 1

Ce[i] = adresses des cellules à trouver

Ce[j] = adresses des cellules déjà trouvées

avec :

Ce[1] = $\propto$.

Pour effectuer un nombre de soustractions conforme au rang de l'étape (Cf. § III-3), nous utilisons deux constantes : w et l.

Le programme prend fin lorsque le nombre d'étapes est égal à N. Le nombre de cellules trouvées est alors égal à 2^N.

```
C_e[1] = ∝ ;
w = 2 ;
l = 2 ;
m = 1 ;

do
{
    j = l ;
        for (i = w ; i ≤ l ; i ++)
            {
                Ce[i] = Ce[j] - pow ((D+1) , (m-1)) ;
                j ++ ;
            }
        m ++ ;
        w = l + 1 ;
        l = pow (2 , m) ;
}
While (m ≤ N) ;                                    A III - 1 - 2
```

ANNEXE III - 2

Identification des adresses des (3^N-1) hyperparallélépipèdes constituant le voisinage de l'hyperparallélépipède d 'adresse unidimensionnelle α .

Pour trouver les adresses des 3^N-1 hypercubes constituant le voisinage de la cellule d'adresse unidimensionnelle α , nous devons procéder pas à pas à partir de cette adresse α. A l' étape de rang m nous devons additionner la quantité $(D+1)^{m-1}$ une fois et la soustraire une autre fois à toutes les adresses déjà trouvées et à l'adresse α elle-même. Les adresses sont mémorisées dans une table à une dimension. La procédure d'addition et de soustraction des quantités $(D + 1)^{m-1}$, m = 1, 2, 3.... N, N étant la dimension des données, se fait à partir de l'équation :

$$Ce[i] = Ce[j] + a *(pow((D+1) , (m - 1))) - b *(pow((D + 1) , (m -1)))$$

A III - 2 - 1

Ce[i] = rang des adresses à trouver

Ce[j] = rang des adresses déjà trouvées

avec :

Ce[0] = α .

a et b sont des constantes à valeurs binaires. Elle permettent d'ajouter ou de soustraire la quantité $(D + 1)^{m-1}$.

Pour définir l'addition et la soustraction qui correspondent respectivement à {a = 1 , b = 0} et {a = 0 , b = 1} nous utilisons trois constantes c, w, l. En premier, on effectue les additions sur toutes les adresses déjà connues. Les soustractions sont exécutées dans un deuxième temps. Cette addition et cette soustraction sont exécutées chacune une fois à la première étape, trois fois dans la seconde, neuf fois dans la troisième, et ainsi de suite. Le programme prend fin lorsque le nombre d'étapes est égal à N. Le nombre de cellules trouvées, y compris la cellule centrale est alors égal à 3^N.

```
Ce [o] = ∝ ;
        w = 1 ;
         l = 3 ;
        m = 1 ;
do
        { j = 0  ;
          c = (l - w) / 2 ;
          for (i = w ; i < l ; i + +)
              { if (c > 0)
                    { a = 1 ;
                      b = 0 ;
                    }
                else
                    {
                        a = o ;
                        b = 1 ;
                    }
```

```
                    Ce [i] =  Ce [j]  +  a * (pow ((D + 1) , (m - 1)))
                    - b * (pow ((D + 1) , (m - 1))) ;
                    c - - ;
                    if (c = = o)  j = o ;
                    else  j++,
                }
            n + +  ;
            w = l ;
            l = 3l ;
        }
while (l ≤ (pow (3 ; N ))) ;                                    A  III - 2 - 2
```

ANNEXE III - 3

Modification de l'adressage des sommets par translation de l'adresse de l'origine des grilles

Dans l'annexe III-2, nous avons vu que pour le calcul des $(3^N - 1)$ voisins d'un hypercube, l'équation A III- 2-1 possède une partie négative qui est ajoutée à l'adresse Ce [j]. Si Ce [j] = 0, la nouvelle adresse d'un des voisins devient négative. Cette valeur négative d'une ou plusieurs adresses peut introduire des difficultés de calcul et une interprétation erronée lors du déroulement du programme. Ces valeurs négatives apparaissent lors de la soustraction des termes de la forme :

(pow ((D + 1) , (m - 1))).

Cette notion d'adresse négative apparaît dans l'exemple de la figure III.1. Les adresses des sommets adjacents sur une même ligne horizontale de la grille des sommets différent de ± 1. Par contre ceux des sommets adjacents sur une même ligne verticale de la grille différent de ± (D + p). Pour ne pas donner une adresse négative lors du calcul des sommets adjacents, on attribue l'adresse [(D + p) + 1] au point d'origine d'adresse zéro.

Pour respecter cette contrainte sur un espace à N dimensions, l 'équation III-5 (Cf. § III-5-2) est modifiée de la façon suivante :

$$\alpha = (a_N+1)\ (D+p)^{N-1} + \ldots + (a_n+1)\ (D+p)^{n-1} + \ldots + (a_2+1)\ (D+p)^1 + (a_1+1)$$

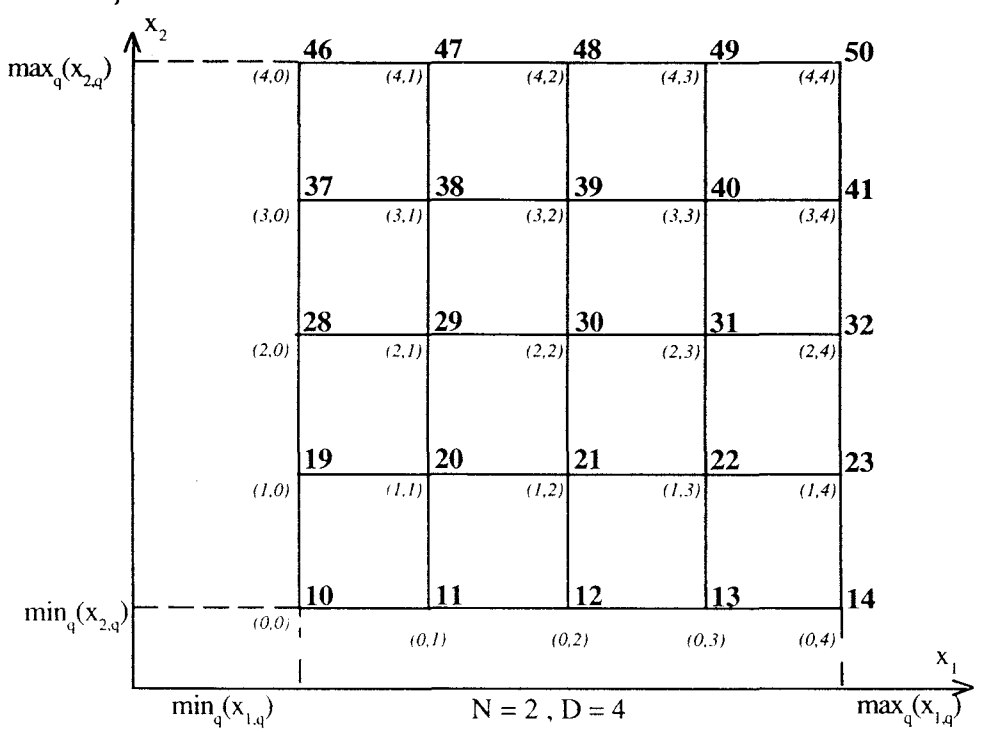
A III - 3 - 1

Les figures III.13 et III.14 donnent les adresses des sommets pour des espaces à deux et trois dimensions respectivement, en utilisant l'équation A III-3-1.

La partie du programme donnant les adresses des sommets occupés est donc donnée par :

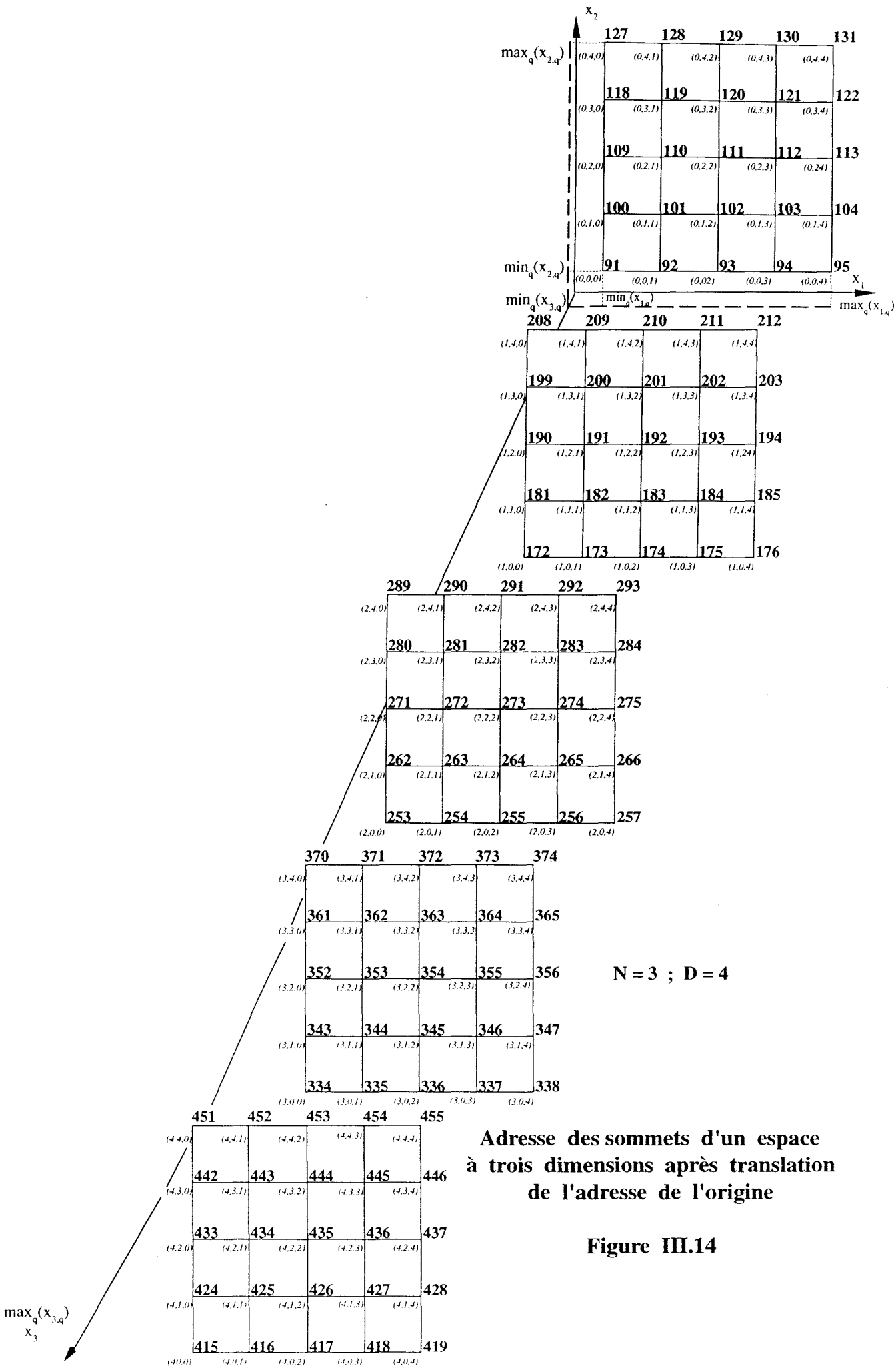
```
for (i = 1 ; i ≤ M ; i ++)
{
    ∝ [i] = 0 ;
    for (j = 1 ; j ≤ N ; j ++)
    {
        a [j] = partie entière de [ (X(j) - min x (j)) / C(j) + 0,5 ] ;
        ∝ [i] = ∝ [i] + ((a[j] + 1) * ((D + p) exp (j - 1)) ;
    }
}
```

A III -3 - 2



Adresse des sommets d'un espace à deux dimensions après translation de l'adresse de l'origine.

Figure III.13



N = 3 ; D = 4

Adresse des sommets d'un espace à trois dimensions après translation de l'adresse de l'origine

Figure III.14

CHAPITRE IV

L'AMINCISSEMENT DE L'HISTOGRAMME

CHAPITRE IV

L'AMINCISSEMENT DE L'HISTOGRAMME

IV - 1 . INTRODUCTION

La procédure de classification automatique de données proposée est basée sur une technique d'amincissements progressifs, par étapes, des modes de l'histogramme multi-dimensionnel dont la construction est décrite aux chapitres II et III. Cet amincissement est obtenu par une technique de réduction progressive de la dispersion des observations autour de chaque mode qui préserve leurs positions dans l'espace de représentation des données.

La réduction progressive de cette dispersion est effectuée par un regroupement itératif des observations qui contribuent à donner naissance à chaque mode de la distribution.

La première conséquence de cette réduction progressive de la dispersion est l'élargissement des vallées qui séparent les modes. Cet élargissement permettra de bien différencier les classes afin de les identifier.

La seconde conséquence de cette procédure est l'amplification des différences d'amplitudes entre les sommets des modes et les creux des vallées qui les séparent.

En itérant la procédure élémentaire d'amincissement, les modes se réduisent à de simples pics qui permettent d'identifier aisément les classes en présence dans les échantillons soumis à l'analyse. Les amplitudes de ces pics donnent le nombre d'observations assignées à chacune des classes et leurs emplacements indiquent la position des classes dans l'espace de représentation des données.

De plus, à la fin de la procédure d'amincissement, l'analyste dispose de la liste exhaustive de toutes les observations constituant chacune des classes, ainsi que tous les paramètres statistiques classiques de position et de dispersion de ces classes.

Cette nouvelle approche est basée sur l'hypothèse de correspondances entre les modes des histogrammes qui représentent les distributions des observations à classer et les classes en présence, déja évoquée au chapitre I, (Cf. § I-8-3).

IV - 2. SEPARATION ET RENFORCEMENT DES MODES PAR MAXIMISATION DE LA TAILLE DES REGROUPEMENTS.

La séparation des modes de l'histogramme multi-dimensionnel s'effectue par migration des observations vers chaque mode. Le but recherché est d'accroître l'amplitude de ces modes et de réduire leur dispersion

afin qu'ils puissent être identifiés sans ambiguïté. Au terme d'une procédure itérative, chaque mode sera réduit a un pic unique de très forte amplitude. Sa mise en évidence deviendra donc très aisée.

Afin de présenter simplement cette procédure fondamentale, nous commencerons par l'exposer sur une distribution unimodale et unidimensionnelle. Nous l'étendrons ensuite à des distributions multimodales et unidimentionnelles. Son application aux distributions multimodales et multidimensionnelles constituera la base de la procédure de classification automatique proposée.

La procédure de séparation des modes s'appuie sur les deux systèmes de discrétisation présentés au chapitre II, à savoir, la grille des centres des hyper-parallélépipèdes élémentaires et celle de leurs sommets. En effet, ces deux réseaux de points de discrétisation de l'espace de représentation des données vont nous permettre de modifier de manière itérative l'histogramme multidimensionnel représentatif de la distribution des données. A chaque itération, le résultat de la transformation précédente de cet histogramme sera mémorisé temporairement sur l'une des grilles, l'autre servant à accueillir l'histogramme modifié au pas suivant de la procédure itérative.

IV - 2 - 1 . Renforcement d'un mode d'une distribution unimodale et unidimensionnelle

On s'intéresse à la distribution d'une variable aléatoire x_l pour laquelle on dispose de Q observations $x_{1,1}$; $x_{1,2}$; ... ; $x_{1,q}$; ... ; $x_{1,Q}$

Le mode de la distribution est considéré comme la valeur de la variable aléatoire pour laquelle la fonction de probabilité sous jacente est maximum.

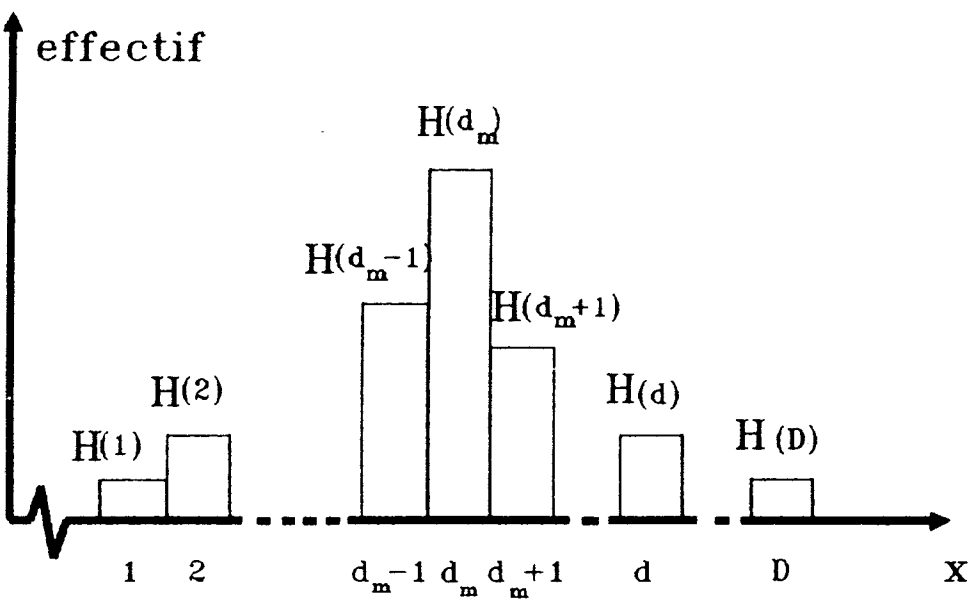
En termes d'histogramme, ce mode est donc le rectangle le plus haut. Il est défini par la position de ce rectangle et par sa hauteur (Cf. figure IV.1).

Supposons que l'on déplace les observations vers le mode en s'appuyant sur la structure de l'histogramme. Pour être plus précis, soit H (d) d = 1, 2, ..., D l'histogramme constitué de D rectangles, on rappelle que H(d) est le cardinal du sous- ensemble des observations O(d) telles que $x_{1,q} \in [x(d) - \frac{C(1)}{2}, x(d) + \frac{C(1)}{2}]$, C(1) étant la largeur de l'intervalle de discrétisation de l'espace sur l'axe du seul attribut $x_{1,q}$ considéré (Cf. § II-4-1) et x (d) est la valeur de l'attribut associé au centre de chaque cellule obtenue par la discrétisation de l'espace (Cf. § III-3). On peut donc écrire :

$$H(d) = CARD \{ x_{1,q} \mid x_{1,q} \in [x(d) - \frac{C(1)}{2} , x_d + \frac{C(1)}{2}] \} \quad IV\text{-}1$$

Soit H (d_m) le mode de cet histogramme d'adresse d_m tel que :

$$H(d_m) > H(d) \quad d = 1, 2, ..., D \; ; \; d \neq d_m \qquad \text{IV-2}$$



Histogramme constitué de D rectangles H(d), d = 1, 2, ..., D, avec H(d_m) comme mode.

Figure IV.1

A chaque rectangle H(d), $d \neq d_m$, on associe une procédure de déplacement en blocs des observations constituant le sous ensemble correspondant O(d).

Toutes les observations assignées au rectangle d'adresse d et de hauteur H(d) sont translatées en blocs de [+ C(1)]

si le mode est situé à droite du rectangle d'adresse d et de [- C(1)] si le mode est situé à gauche de ce même rectangle.

Lorsque cette procédure de déplacement élémentaire est appliquée à tous les rectangles de l'histogramme H(d), d = 1, 2,..., D, on obtient un nouvel histogramme $H^1(d)$, d = 1, 2, ... D, tel que :

$$\begin{cases} H^1(1) = 0 \\ H^1(d) = H(d-1) \quad \text{si} \quad 1 < d < d_m \\ H^1(d_m) = H(d_m) + H(d_m-1) + H(d_m+1) \\ H^1(d) = H(d+1) \quad \text{si} \quad d_m < d < D \\ H^1(D) = 0 \end{cases} \qquad \text{IV-3}$$

L'indice 1 de $H^1(d)$ indique le rang de l'itération de cette procédure. Le rectangle $H(d_m)$ constituant le mode joue un rôle particulier dans cette procédure puisqu'il reçoit les sous-ensembles d'observations $O(d_m - 1)$ et $O(d_m + 1)$ assignées aux rectangles voisins $H(d_m - 1)$ et $H(d_m + 1)$.

On montre, à l'annexe IV-1, que ce type de déplacement par blocs conduit, dans le cas d'une distribution symétrique, à réduire la variance de la variable aléatoire dont la distribution est représentée par l'histogramme H(d)

En itérant cette procédure de déplacement des observations, on aboutit à un nouvel histogramme ne comportant qu'un seul rectangle dont l'amplitude $H^F(d_m)$ est égale à :

$$H^F(d_m) = \sum_{d=1}^{D} H(d) \qquad \text{IV-4}$$

L'indice F indique le rang de l'itération finale.

Cette procédure se résume donc à un regroupement de toutes les observations constituant la distribution en un rectangle unique qui est situé à l'emplacement du mode initial.

Dans ce cas unimodal, on peut donc réduire facilement un histogramme à un seul rectangle aisement identifiable.

La procédure de classification proposée est basée sur une technique d'amincissements progressifs, par étapes, de l'histogramme en réduisant progressivement la dispersion des observations autour de chacun des modes, selon un schéma proche de celui que nous venons de présenter dans le cas simple, sinon trivial, d'une distribution unimodale et unidimensionnelle.

IV - 2 - 2 . Renforcement des modes d'une distribution multimodale et unidimentionnelle.

Lorsque la distribution considérée est multimodale, avec un nombre de modes non défini a priori, le problème principal est de savoir vers quel mode faire évoluer les observations afin de les regrouper. Le problème se pose surtout pour celles situées dans les vallées.

Si les paramètres statistiques des composantes constituant la distribution étaient connus, il est évident que la ligne de séparation indiquant vers quel mode déplacer chaque observation pour obtenir une classification optimale serait donnée par la théorie de décision. Celle-ci permettrait donc de regrouper autour de chaque mode les observations assignées à la classe correspondante en minimisant les risques d'erreurs.

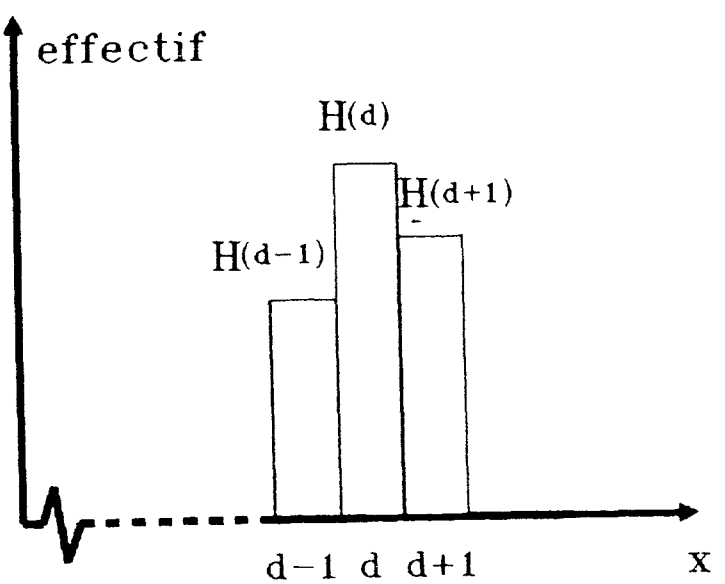
Dans un contexte non supervisé, on ne peut s'appuyer sur une telle procédure optimale. Nous proposons une procédure heuristique qui vise à déplacer les observations par blocs afin de renforcer les modes de leur distribution. La stratégie adoptée consiste à favoriser des déplacements qui créent des regroupements de taille maximale.

IV - 2 - 2 - 1 . Détermination des directions de déplacement des observations

Ne sachant où se trouvent les différents modes de la distribution, on commence selon une procédure de regroupents fictifs, par déplacer les sous-ensembles d'observations assignés à tous les rectangles de l'histogramme.

Chaque sous-ensemble peut être regroupé fictivement avec l'un de ses deux voisins. Pour ces deux regroupements possibles, l'un sera de taille supérieure ou égale à l'autre.

La direction dans laquelle se trouve le regroupement fictif le plus important indique la direction dans laquelle le sous-ensemble d'observations assigné au rectangle considéré sera déplacé.



Rectangles d'adresses (d - 1) , d , (d + 1)

Figure IV.2

De manière plus précise, on considère le rectangle H(d) d'adresse d, et entouré des rectangles H(d-1) et H(d+1) (Cf. figure IV.2). Pour déterminer la direction de déplacement de ce sous-ensemble O(d), on compare les deux regroupements fictifs possibles avec ses deux voisins O(d-1) et O(d+1) donnés par :

$$O(d) + O(d+1) = O^1\left(\frac{(d) + (d+1)}{2}\right) \qquad \text{IV-5}$$

et

$$O(d) + O(d-1) = O^1\left(\frac{(d) + (d-1)}{2}\right) \qquad \text{IV-6}$$

Les adresses $\left(\frac{(d) + (d+1)}{2}\right)$ et $\left(\frac{(d) + (d-1)}{2}\right)$ indiquent que ces regroupements fictifs sont positionnés entre les deux rectangles correspondants.

Si $\text{CARD}\left[O^1\left(\frac{(d) + (d+1)}{2}\right)\right] > \text{CARD}\left[O^1\left(\frac{(d) + (d-1)}{2}\right)\right]$ les observations constituant le sous-ensemble O(d) sont déplacées en bloc vers l'adresse $\left(\frac{(d) + (d+1)}{2}\right)$.

Par contre, si $\text{CARD}\left[O^1\left(\frac{(d) + (d+1)}{2}\right)\right] < \text{CARD}\left[O^1\left(\frac{(d) + (d-1)}{2}\right)\right]$, ces mêmes observations constituant le sous-ensemble O(d) sont déplacées en bloc vers l'adresse $\left(\frac{(d) + (d-1)}{2}\right)$.

IV - 2- 2 - 2 . Notations

On peut établir un lien directe entre les adresses de ces nouveaux rectangles $\left(\frac{(d) + (d+1)}{2}\right)$ et $\left(\frac{(d) + (d-1)}{2}\right)$ et le repérage des centres des cellules utilisés pour construire l'histogramme.

Il est possible de simplifier les notations en utilisant le concept de grille de centres et de grille de sommets introduit au chapitre III (Cf. § III-3). Si l'histogramme initial est déterminé sur la grille des centres, on notera :

$$\begin{cases} H(d-1) = H_{ce}(d-1) \\ H(d) = H_{ce}(d) \\ H(d+1) = H_{ce}(d+1) \end{cases} \qquad \text{IV-7}$$

L'indice "ce" indique que l'histogramme est déterminé sur la grille des centres.

Les sous-ensembles d'observations associés à ces rectangles seront donc notés : $O_{ce}(d-1)$, $O_{ce}(d)$ et $O_{ce}(d+1)$

Le regroupement fictif de $O_{ce}(d)$ avec $O_{ce}(d-1)$, localisé à $\frac{(d)+(d-1)}{2}$ est, selon les conventions du chapitre II, l'emplacement du sommet commun aux deux cellules d'adresses (d) et (d - 1). L'adresse de ce sommet est (d), (Cf. figure IV.2) on note :

$$O_{ce}(d-1) + O_{ce}(d) = O^{1}_{so}(d) \qquad \text{IV-8}$$

$$O_{ce}(d) + O_{ce}(d+1) = O^{1}_{so}(d+1) \qquad \text{IV-9}$$

L'indice "so" indique que ces regroupements fictifs sont localisés sur la grille des sommets.

A partir de tous ces sous-ensembles d'observations $O^1_{so}(d)$, d = 1,2,3,...,D; obtenus par tous les regroupements fictifs possibles, on peut définir un histogramme fictif $FH^1_{so}(d)$, d = 1,2,...,D, défini sur la grille des sommets, tel que :

$$FH^1_{so}(d) = \text{CARD } [O^1_{so}(d)] \qquad \text{IV-10}$$

Cet histogramme est appelé "fictif" car il ne sert qu'à indiquer dans quelle direction doivent migrer les observations.

IV - 2- 2 - 3 . Migration des observations par blocs

Chaque sous-ensemble $O_{ce}(d)$ d = 1, 2, 3,..., D ; peut donc être translaté en bloc, soit vers le sommet d'adresse d, soit vers celui d'adresse (d+1), selon laquelle des valeurs respectives de CARD $[O^1_{so}(d)]$ et CARD $[O^1_{so}(d+1)]$ qui forment l'histogramme fictif $FH^1_{so}(d)$, est la plus grande.

D'autre part le sous-ensemble O_{ce} (d-1) peut lui aussi être translaté en bloc soit vers le sommet d'adresse (d-1), soit vers celui d'adresse d, selon laquelle des valeurs respectives de CARD $[O^1_{so}(d-1)]$ et CARD $[O^1_{so}(d)]$ qui forment l'histogramme fictif $FH^1_{so}(d)$ est plus grande.

Effectuant les migrations dans les directions maximisant le regroupement des sous-ensembles, nous obtenons sur la grille des sommets, les valeurs de ces regroupements réels qui forment un nouvel histogramme réel cette fois-ci, $RH^{1}_{so}(d)$ $d = 1, 2, ..., D$.

A chaque sommet d'adresse d, quatre cas de figures sont possibles :

- $CARD\,[O^{1}_{so}(d)] > CARD\,[O^{1}_{so}(d-1)]$ et
  $CARD\,[O^{1}_{so}(d)] > CARD\,[O^{1}_{so}(d+1)]$

  Dans ce cas précis, les deux sous-ensembles $O_{ce}(d-1)$ et $O_{ce}(d)$ migrent vers le sommet d'adresse d, et se regroupent (Cf. figure IV-3-a) et nous aurons :

  $$RH^{1}_{so}(d) = CARD\,[O_{ce}(d-1) + O_{ce}(d)] \qquad \text{IV-11}$$

- $CARD\,[O^{1}_{so}(d)] > CARD\,[O^{1}_{so}(d-1)]$ et
  $CARD\,[O^{1}_{so}(d+1)] > CARD\,[O^{i}_{so}(d)]$

  Dans ce cas, le sous-ensemble $O_{ce}(d-1)$ migre vers le sommet d'adresse d et $O_{ce}(d)$ migre vers le sommet d'adresse (d+1) (Cf. figure IV-3-b), et nous aurons :

  $$RH^{1}_{so}(d) = CARD\,[O_{ce}(d-1)] \qquad \text{IV-12}$$

- $CARD\,[O^{1}_{so}(d)] < CARD\,[O_{ce}(d-1)]$ et
  $CARD\,[O^{1}_{so}(d)] > CARD\,[O^{1}_{so}(d+1)]$

  Dans ce cas, $O_{ce}(d-1)$ migre vers le sommet d'adresse (d-1) et $O_{ce}(d)$ vers le sommet d'adresse d (Cf. figure IV-3-c), et nous aurons :

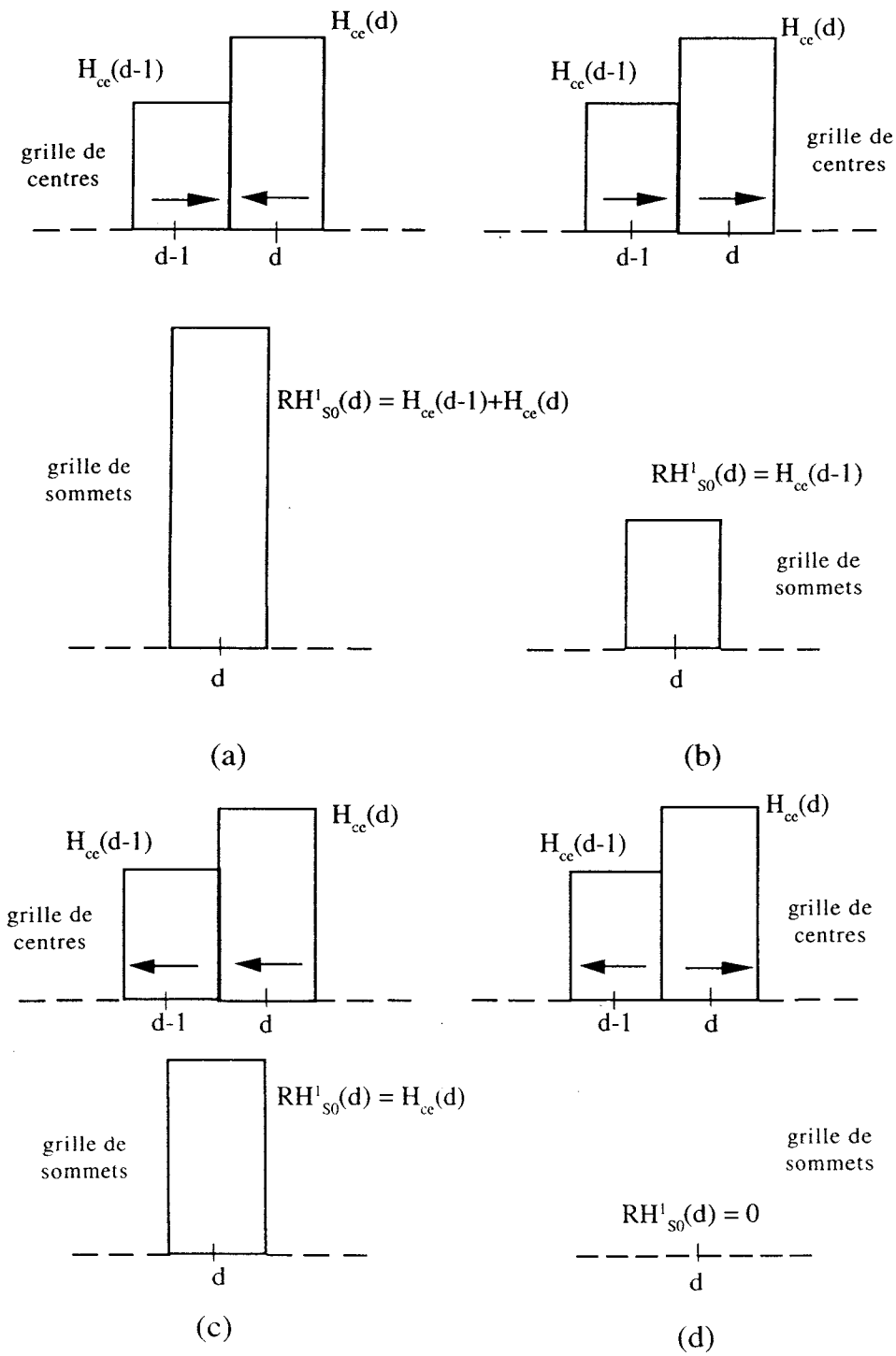
$$RH^{1}_{so}(d) = CARD\,[O_{ce}(d)] \qquad \text{IV-13}$$

- $CARD\,[O^{1}_{so}(d-1)] > CARD\,[O^{1}_{so}(d)]$ et
  $CARD\,[O^{1}_{so}(d+1)] > CARD\,[O^{1}_{so}(d)]$

  Dans ce cas, les deux sous-ensembles $O_{ce}(d-1)$ et $O_{ce}(d)$ migre vers les sommets d'adresses (d-1) et (d) respectivement (Cf. figure IV-3-d), et nous aurons :

$$RH^{1}_{so}(d) = 0 \qquad \text{IV-14}$$

Quand toutes ces migrations sont effectuées simultanément, on regroupe en chaque sommet soit aucun, soit un, soit deux, sous-ensembles d'observations. On obtient ainsi un nouvel histogramme réel cette fois-ci, $RH^{1}_{so}(d)$, $d = 1, 2, ..., D$ dont les valeurs résultent des déplacements par blocs de tous les sous-ensembles d'observations dans les directions favorisant les regroupements de taille maximum.



Les quatre cas de figure des migrations possibles des sous-ensembles

Figure IV.3

Si l'histogramme à partir duquel les regroupements fictifs sont effectués est bâti sur la grille des sommets (Cf. chapitre III-3), les valeurs de ces regroupements fictifs sont définies sur la grille des centres et vice-versa.

Nous définirons l'histogramme bâti sur la grille des sommets par ses valeurs $H^t_{so}(d)$, l'indice "so" indiquant que la grille est celle des sommets et l'indice t représente le rang de l'itération effectuée sur cette grille. Rappelons que la construction de l'histogramme initial correspond à $t = 1$. De même, l'histogramme bâti sur la grille des centres sera défini par les valeurs $H^t_{ce}(d)$. Ces valeurs seront soit des valeurs de regroupements fictifs et donc précédées de la lettre F, soit des regroupements réels favorisés par des déplacements par blocs et donc précédées par la lettre R. La figure IV.4 illustre les regroupements fictifs sur la grille des centres effectués à partir de l'histogramme bâti sur la grille des sommets. Les valeurs fictives de l'histogramme sont données par :

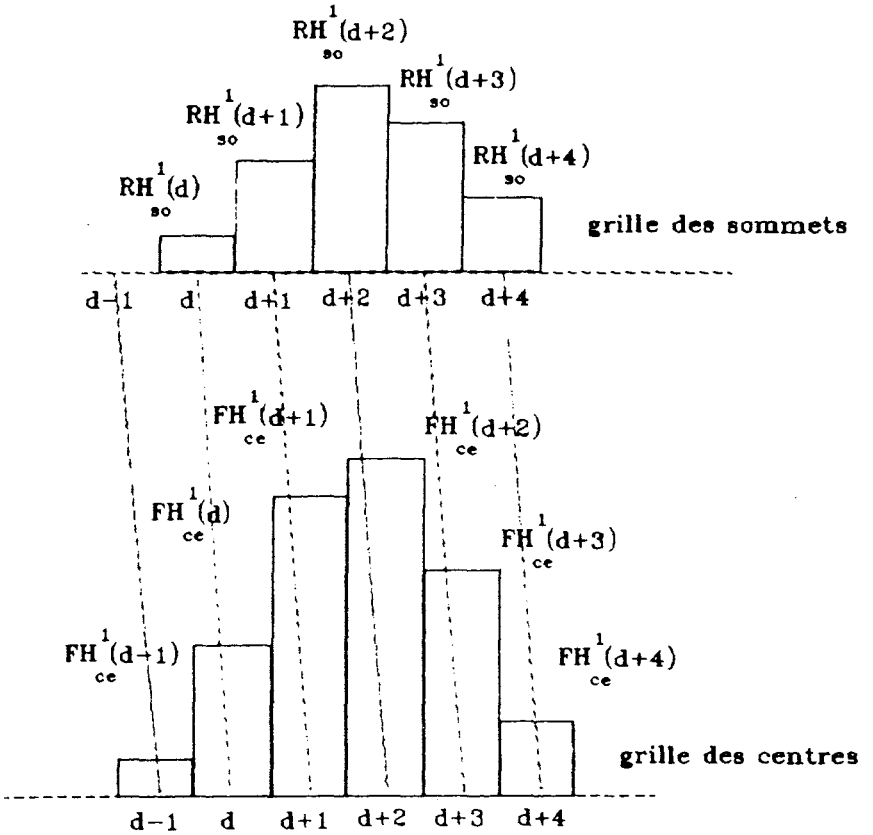
$$FH^1_{ce}(d) = RH^1_{so}(d) + RH^1_{so}(d+1) \qquad \text{IV-15}$$

Rappelons aussi que :

$$FH^1_{ce}(d+1) = RH^1_{so}(d+1) + RH^1_{so}(d+2) \qquad \text{IV-16}$$

A partir des équations IV-15 et IV-16 nous constatons que pour chaque rectangle, nous avons deux

voisins, donc deux regroupements fictifs possibles. Par exemple pour $RH^{1}_{so}(d+1)$ nous obtenons les deux regroupements fictifs possibles $FH^{1}_{ce}(d)$ et $FH^{1}_{ce}(d+1)$.

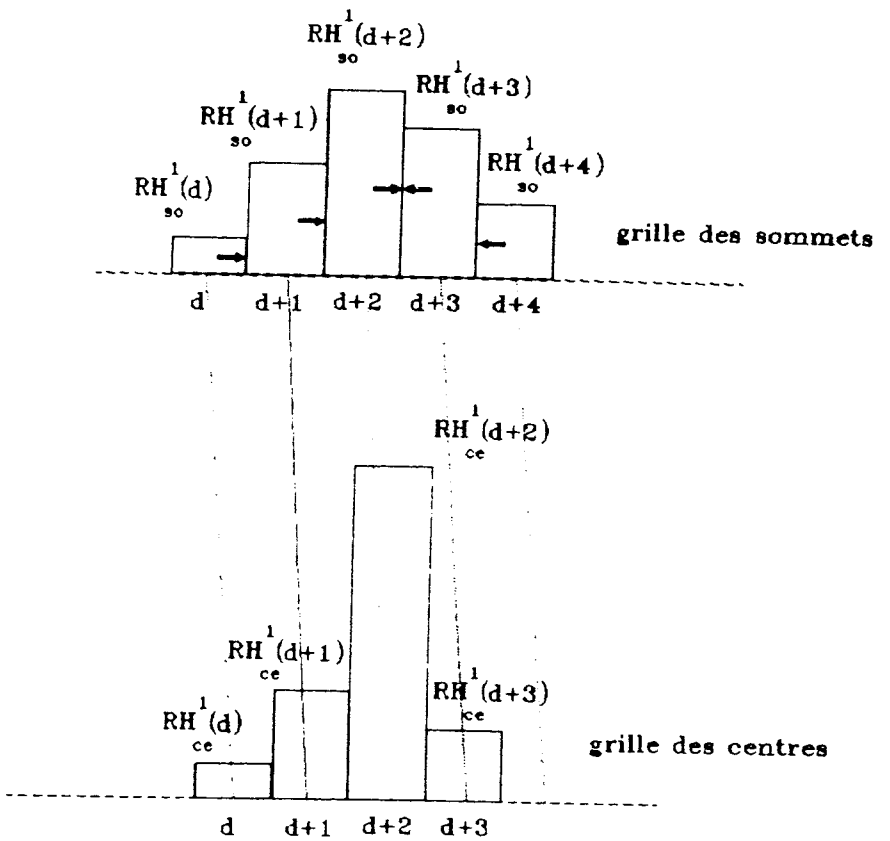


Groupement fictif sur la grille des centres à partir de l'histogramme bâti sur la grille des sommets

Figure IV.4

Celui de ces 2 regroupements fictifs qui contient le plus grand nombre d'observations indique dans quelle direction devront être effectivement déplacées les observations constituant le regroupement $O^{1}_{so}(d+1)$. La figure IV.5 illustre la direction de migration retenue pour chaque sous-ensemble.

Le résultat de cette migration d'observations est caractérisé par la valeur de l'histogramme $RH^{1}_{ce}(d)$ représentant le nombre d'observations ainsi regroupées réellement.

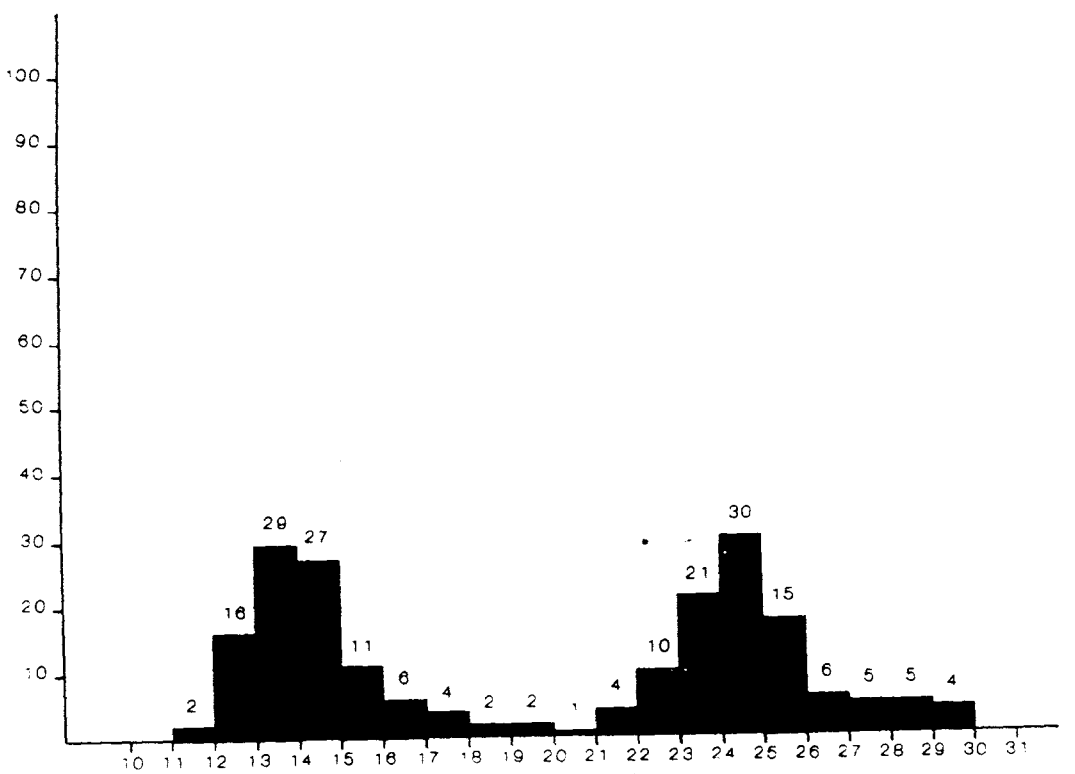


Direction de regroupement maximale

Figure IV.5

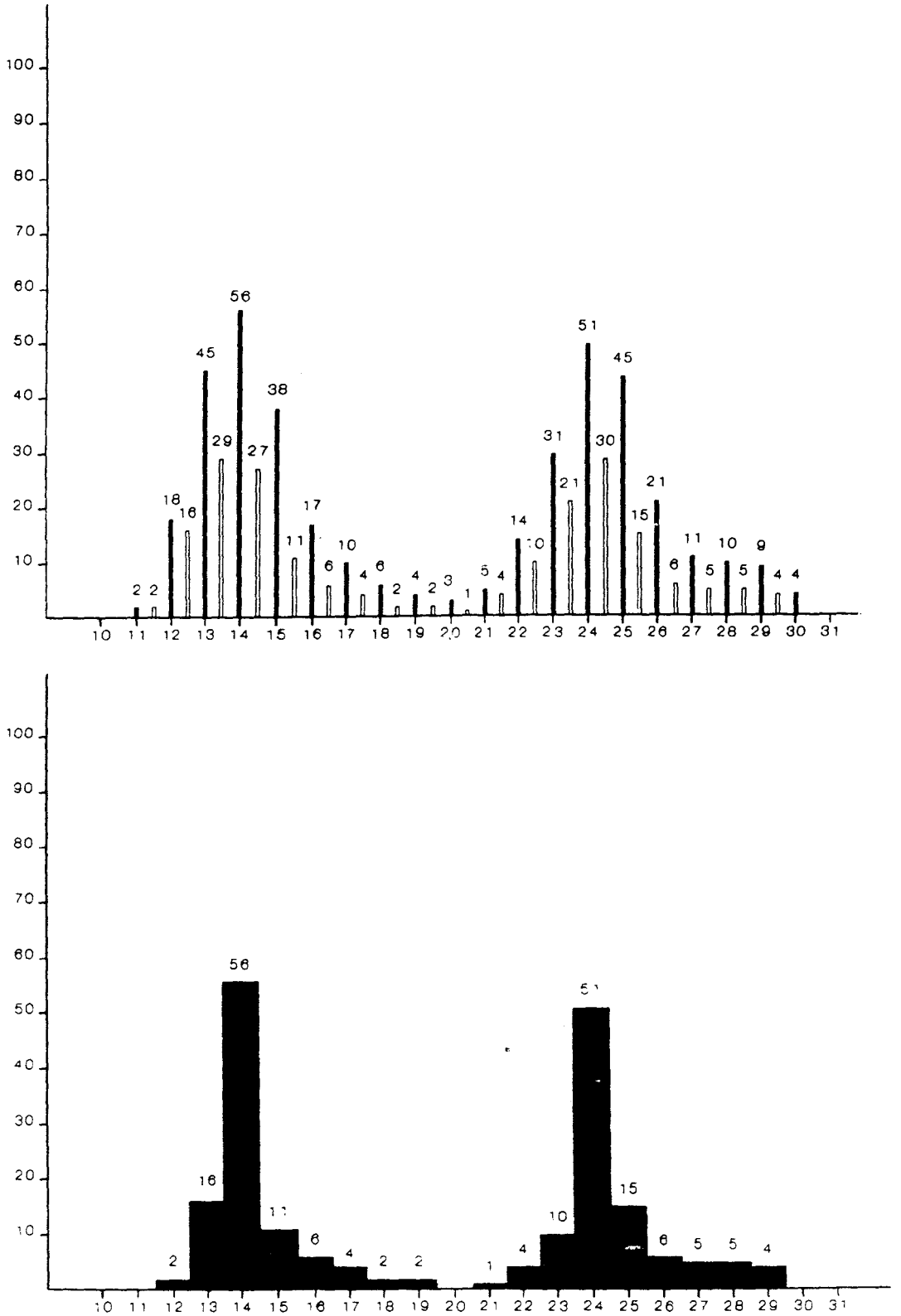
Cette technique est itérée jusqu'à ce qu'aucun autre regroupement ne soit possible en respectant les règles exposées ci-dessus. Le nombre de groupes obtenus après cet arrêt automatique de la procédure représente alors le nombre de classes en présence.

Pour démontrer l'efficacité de cette procédure, nous allons l'appliquer au cas d'un histogramme élémentaire bâti sur la grille des centres à partir d'une fonction de densité d'un mélange monovariable. Cette fonction présente deux modes reflétant la présence de 2 classes de 100 observations chacune (Cf. figure IV.6).



Le mélange monovariable

Figure IV.6



Le premier regroupement et l'apparition de la vallée

Figure IV.7

Pour la simplicité du dessin, lors de l'éxécution d'une itération, les rectangles de l'histogramme réel à partir desquels cette itération est exécutée et les rectangles de l'histogramme fictif sont représentés par des bâtons respectivement pleins ou creux. La figure IV.7(a) montre ces bâtons pour la première itération.

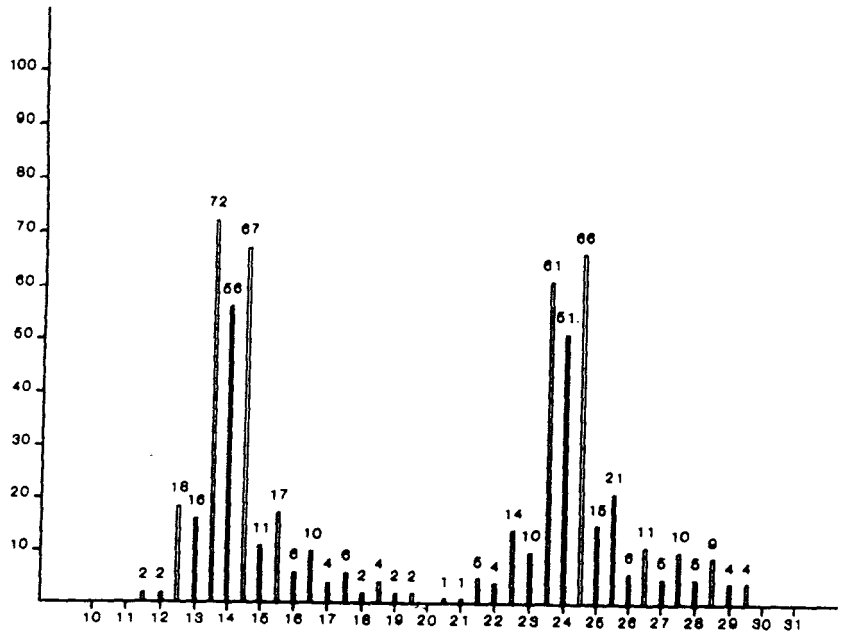
Après l'éxécution de cette première itération, nous remarquons les faits suivants (Cf. figure IV.7(b)) :

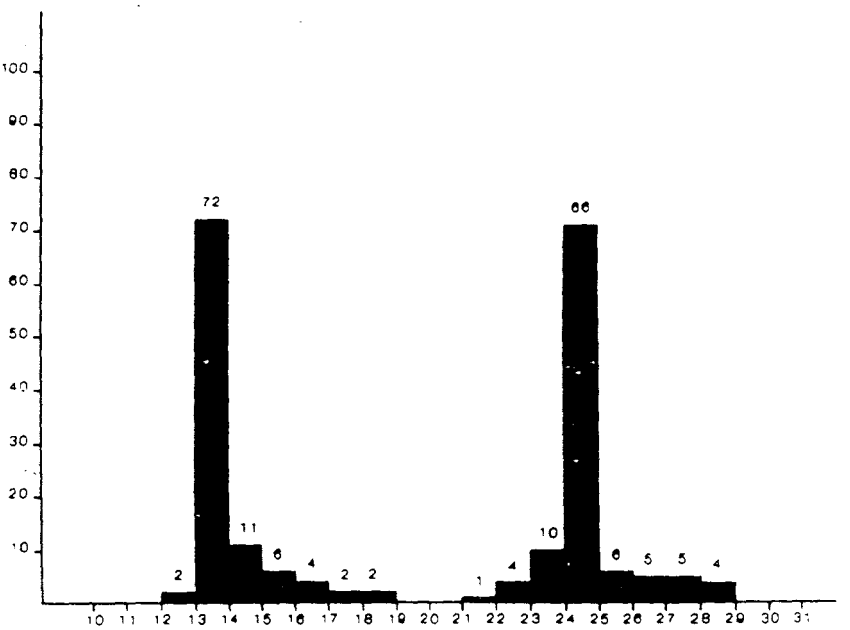
- Pour chacune des deux classes le premier regroupement s'est fait entre les deux sous-ensembles pour lesquels CARD $[O^{I}_{so}(d)]$ = CARD $[O^{I}_{ce}(d)]$ + CARD $[O^{I}_{ce}(d-1)]$, d = 1,2,...,D est maximum,

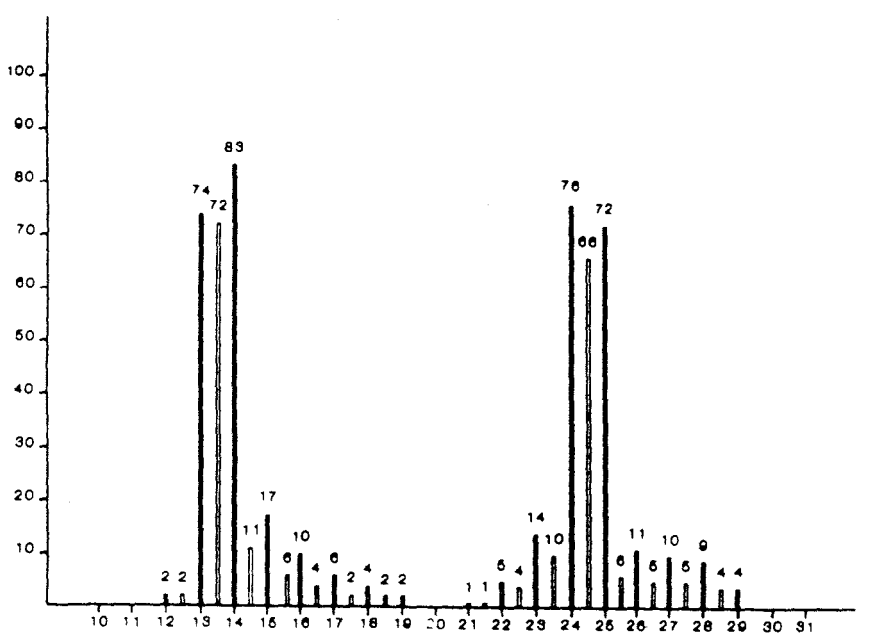
- La valeur maximale de chaque mode a été augmentée grâce à ce regroupement,

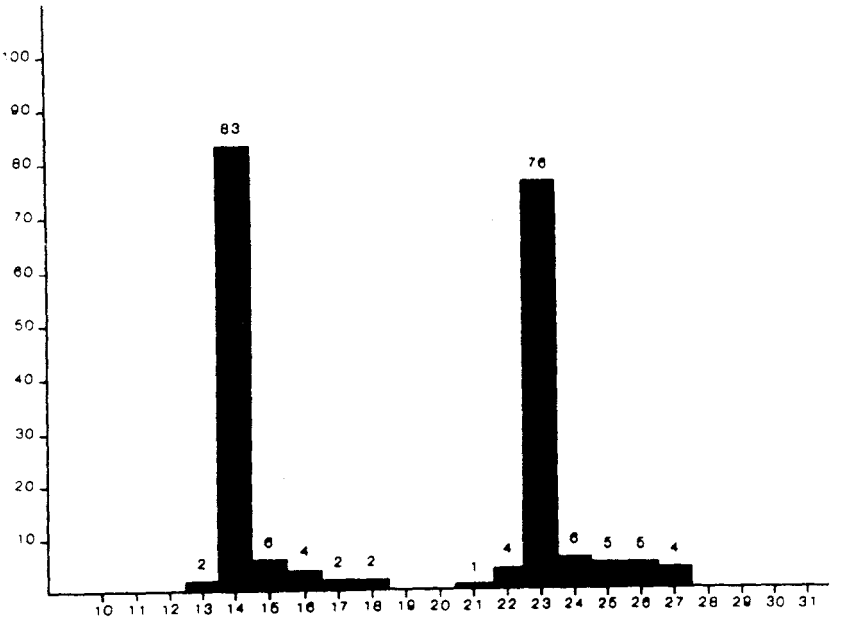
- Un creux, ou une vallée, s'est creusé entre les deux modes.

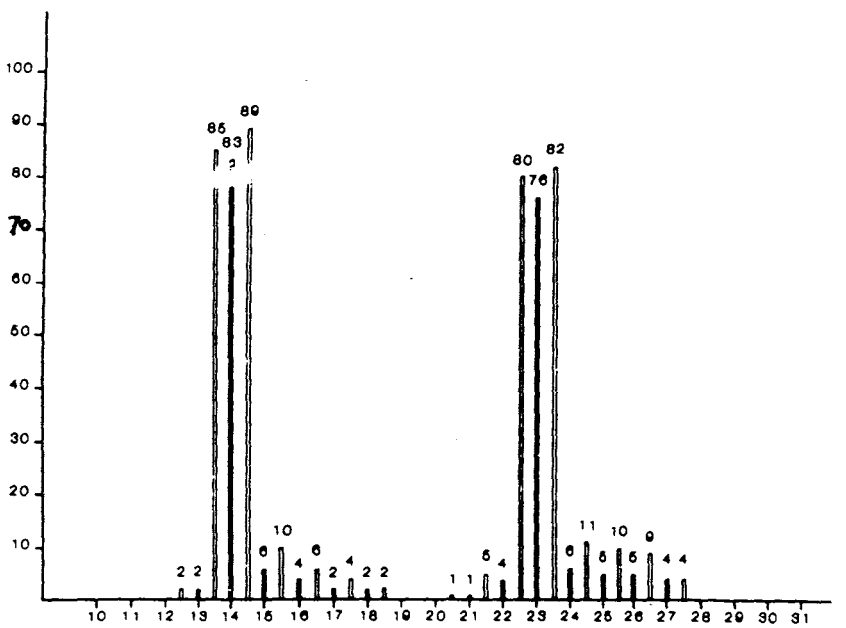
Les figures suivantes montrent la suite de la procédure itérative jusqu'à la classification finale par séparation des modes et indication du nombre d'observations assignées à chaque classe. La procédure de regroupement est arrêtée lorsque le nombre de pics reste constant entre deux itérations successives.

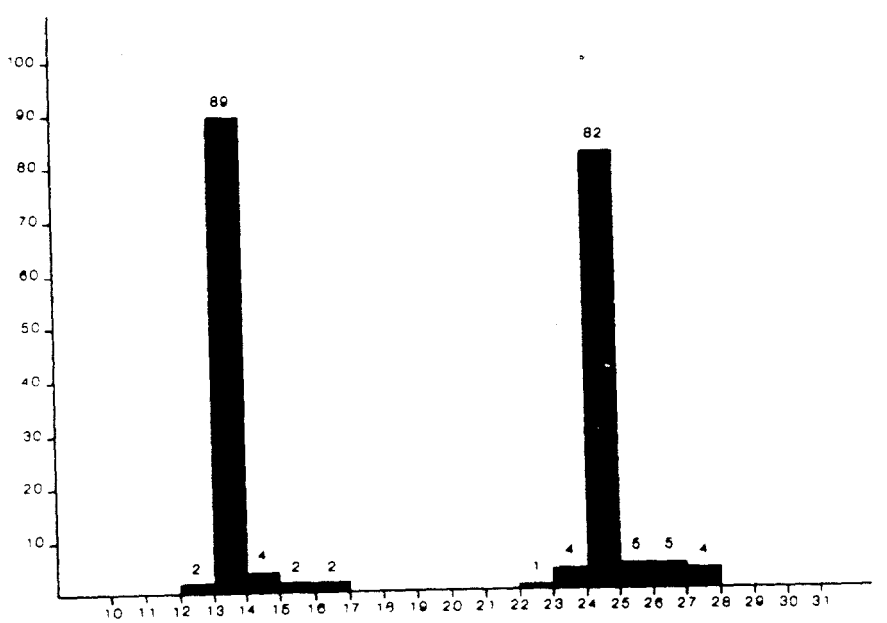


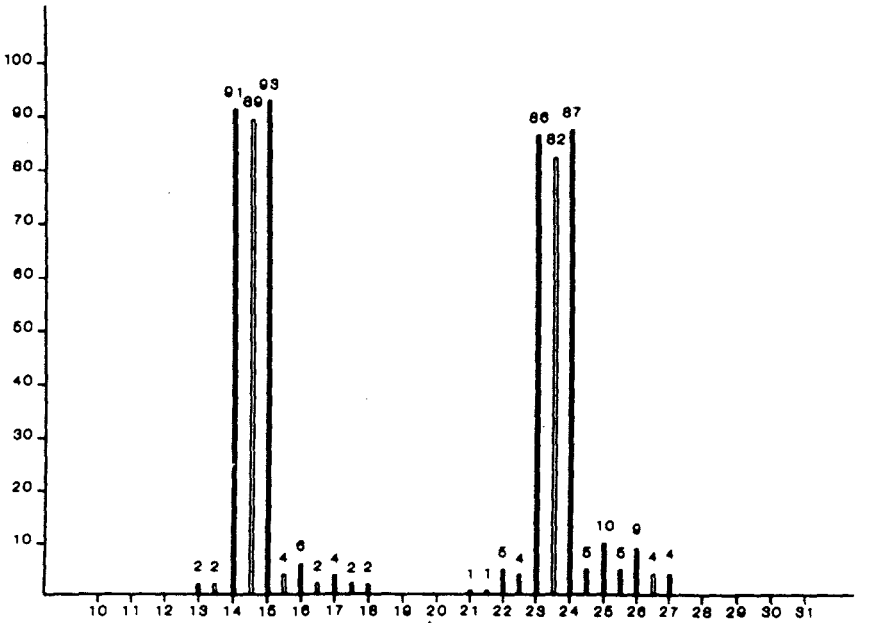


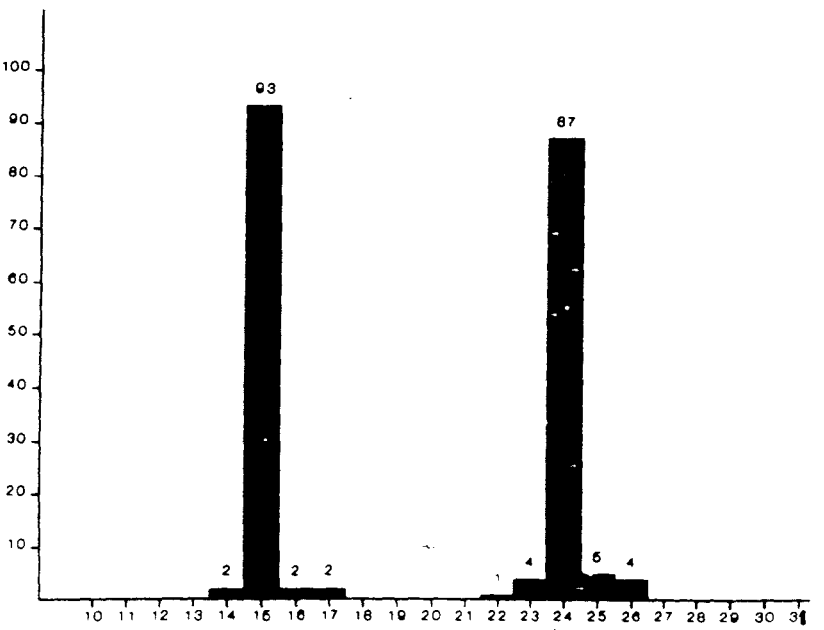


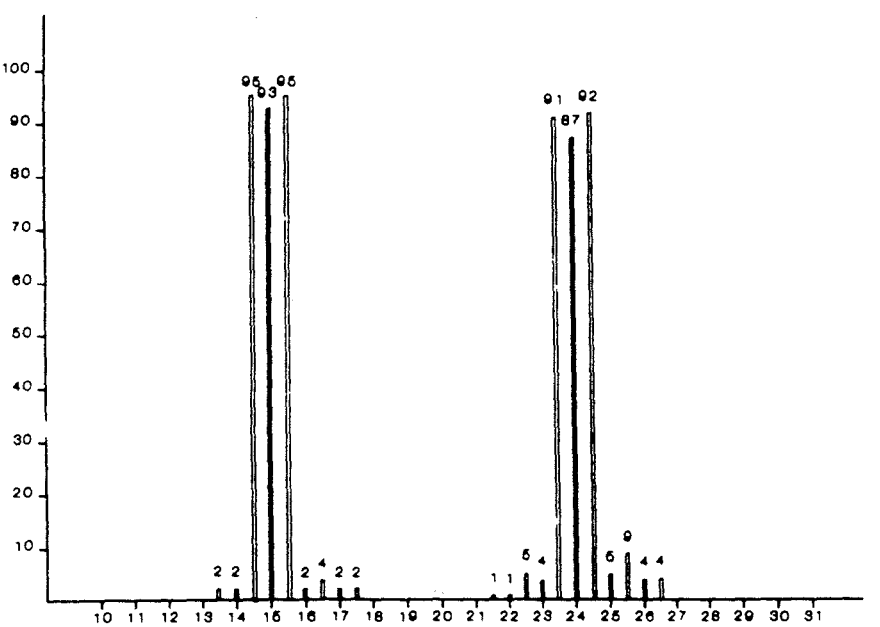

100
90
80
70
60
50
40
30
20
10
83
76
2
6
4
2
2
1
4
6
5
5
4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

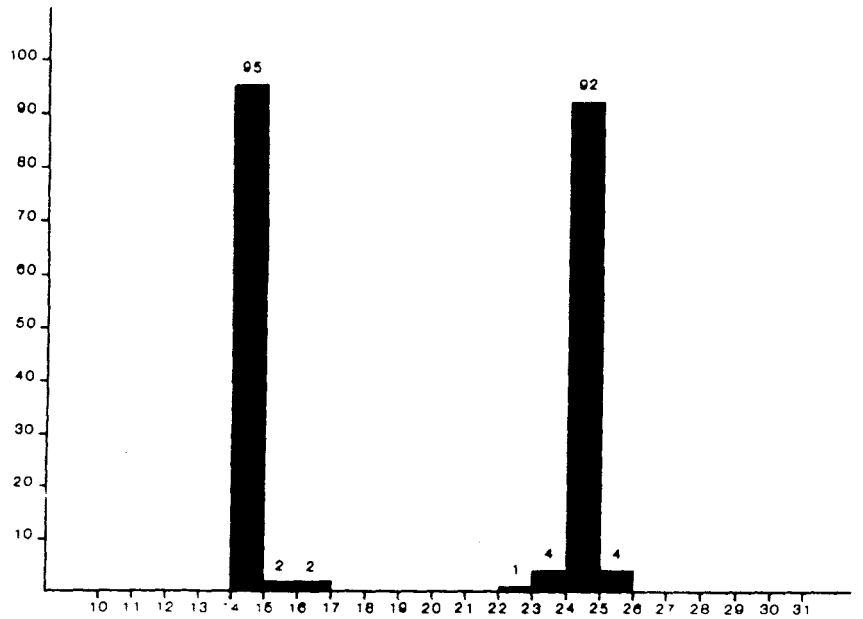

100
90
80
70
60
50
40
30
20
10
89
85
83
80
82
76
2 2
6
10
6
4
6
2
4
2 2
1 1
5
4
6
11
10
9
4 4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

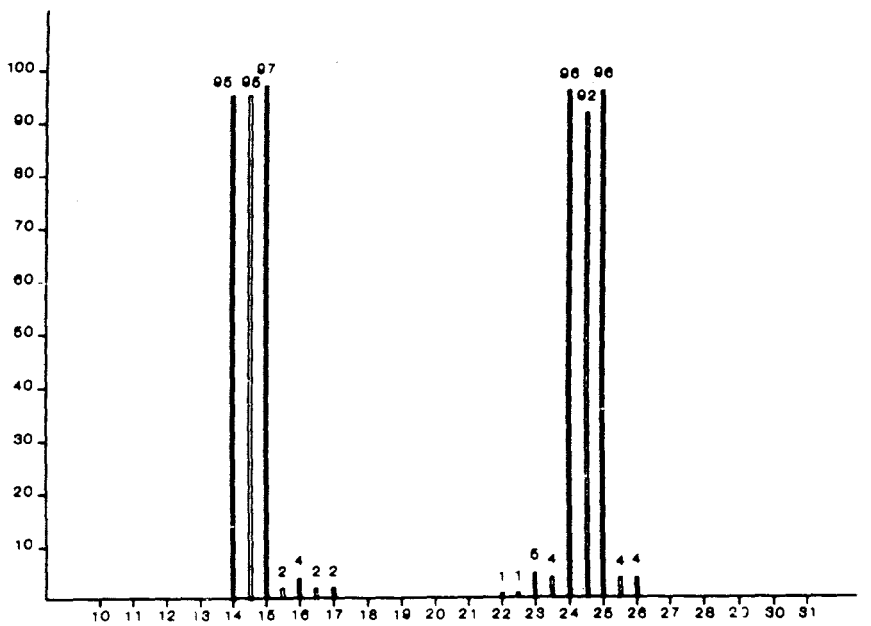

100
90
80
70
60
50
40
30
20
10
89
82
2
4
2
2
1
4
6
5
4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

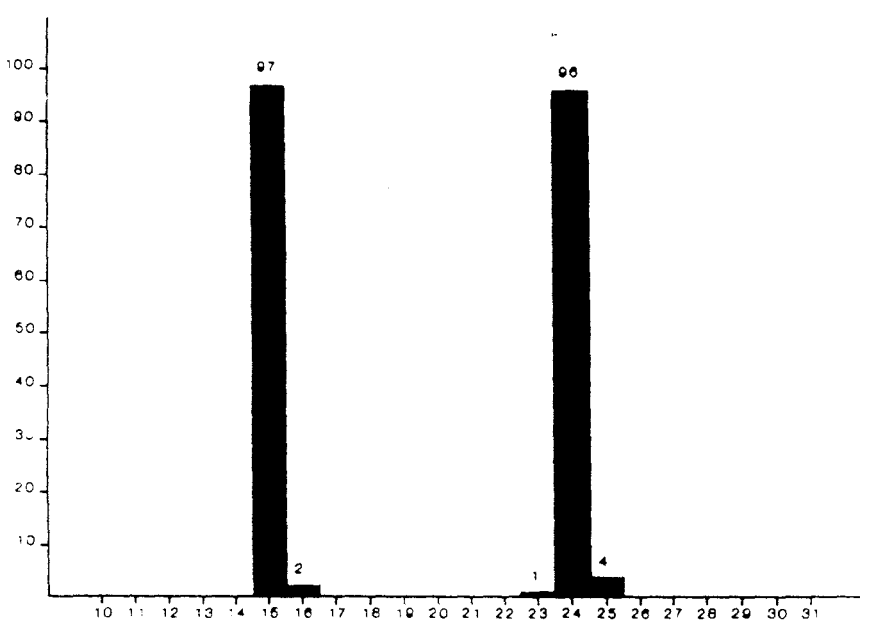

100
90
80
70
60
50
40
30
20
10
91
89
93
86
82
87
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

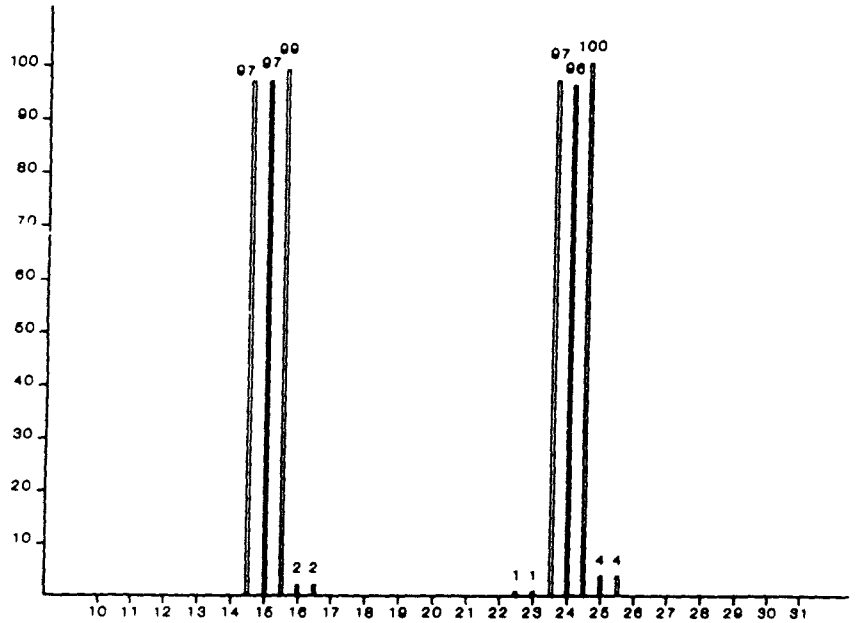

100
90
80
70
60
50
40
30
20
10
93
87
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

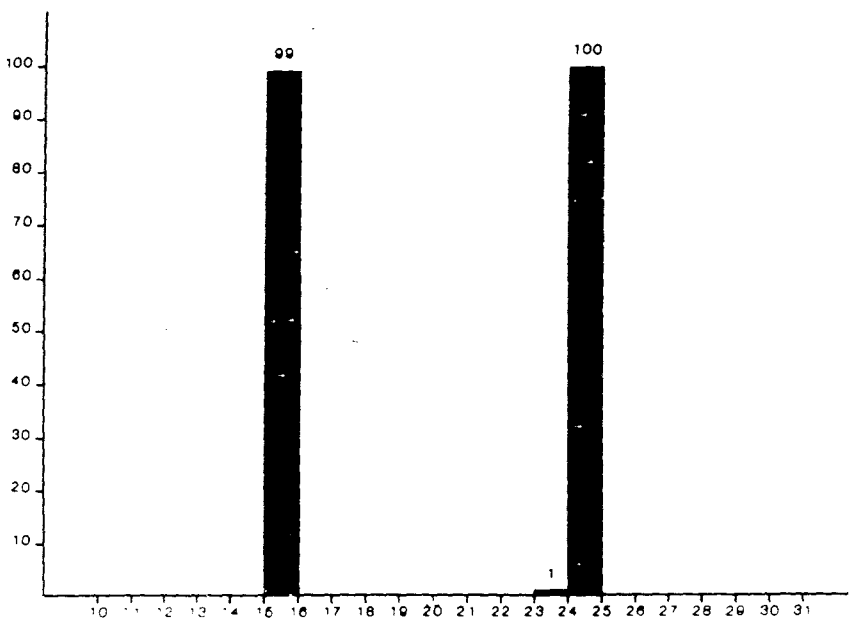

100
90
80
70
60
50
40
30
20
10
95
93
96
91
92
87
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

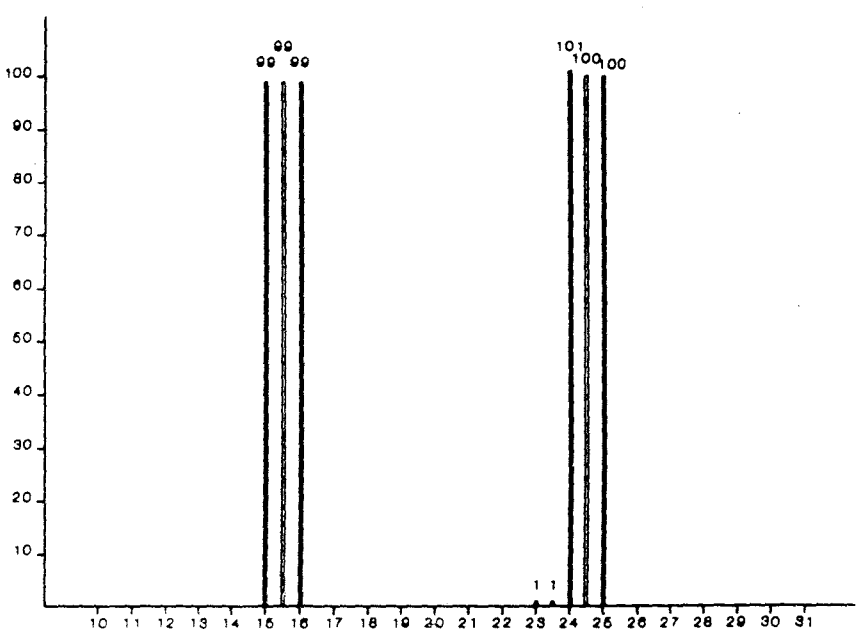

100
90
80
70
60
50
40
30
20
10
95
92
2 2
1
4
4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

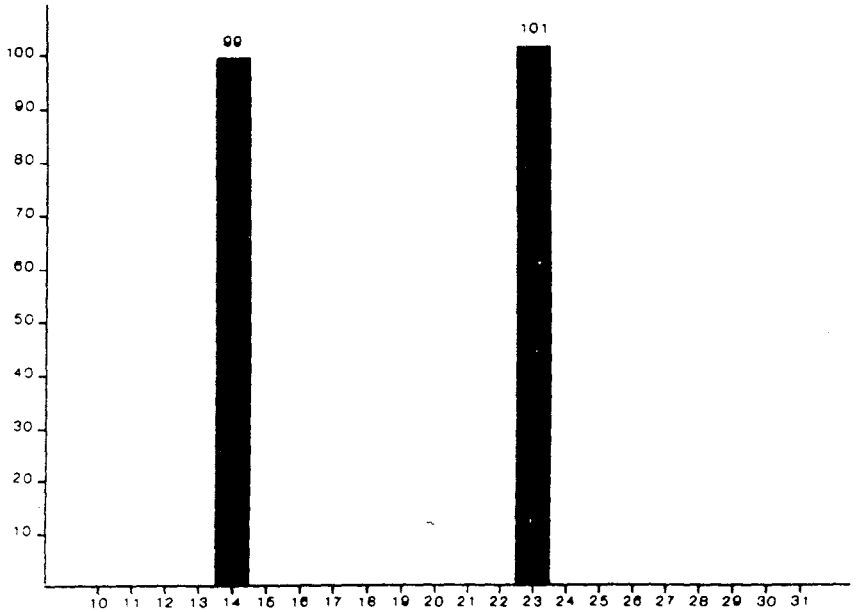

100
90
80
70
60
50
40
30
20
10
95 95 97
96 96
92
2 4
2 2
1 1
6 4
4 4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

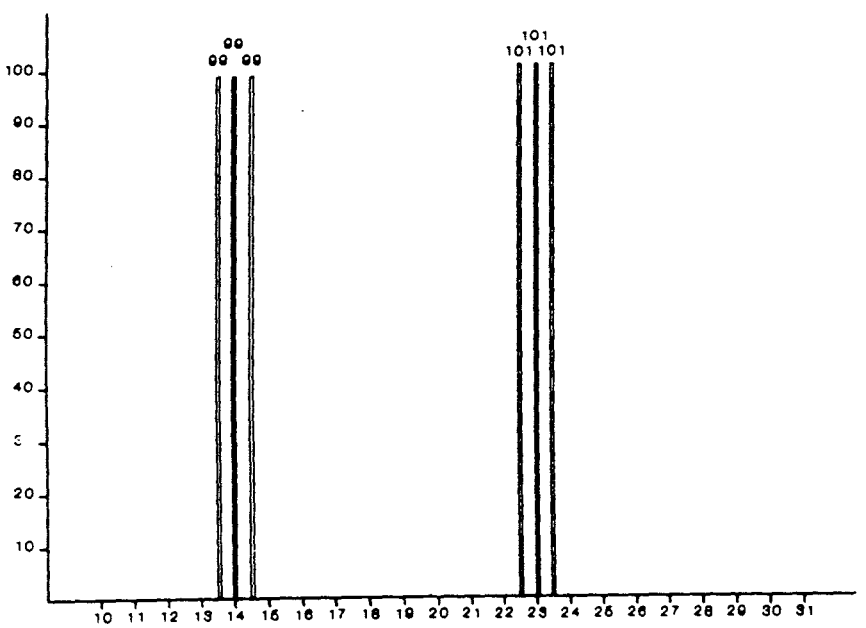

100
90
80
70
60
50
40
30
20
10
97
96
2
1
4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

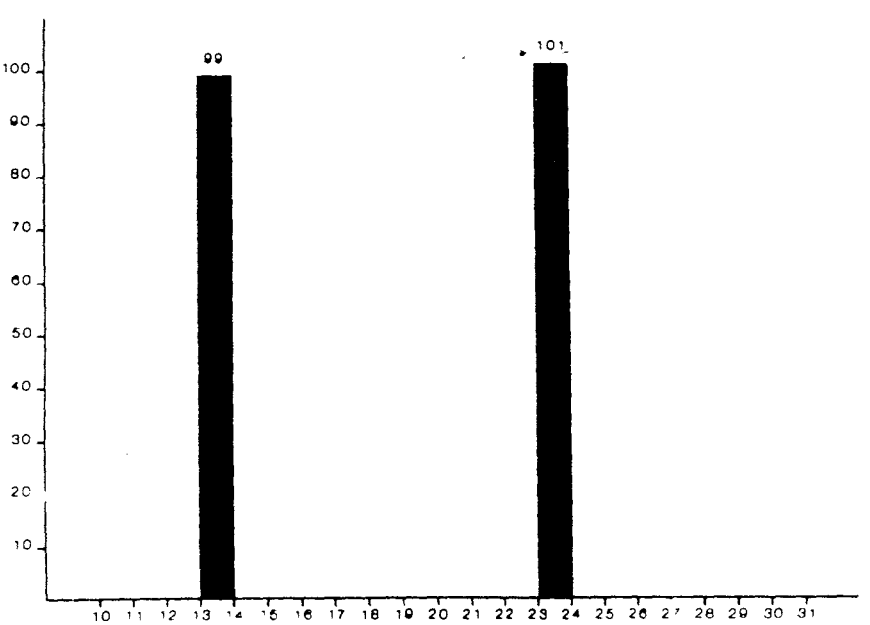

100
90
80
70
60
50
40
30
20
10
97 97 99
2 2
1 1
97 96 100
4 4
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

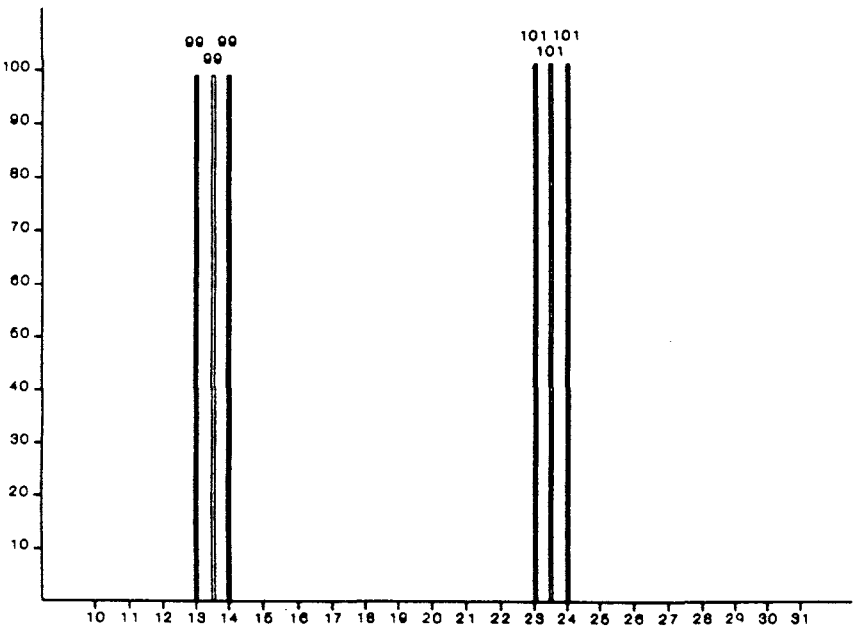

100
90
80
70
60
50
40
30
20
10
99
100
1
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

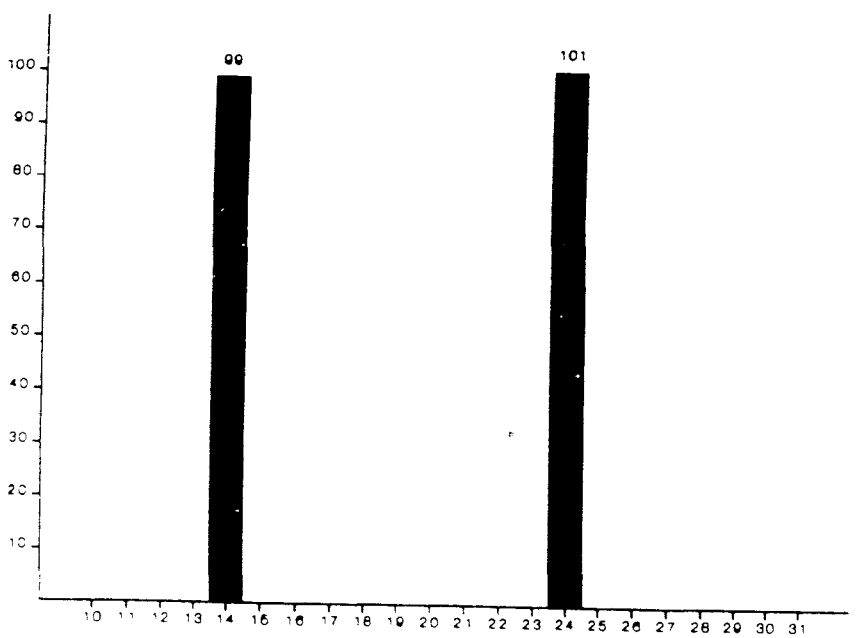

100
90
80
70
60
50
40
30
20
10
99 99 99
1 1
101 100 100
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31


99
101
100
90
80
70
60
50
40
30
20
10
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31


99
99 99
101
101 101
100
90
80
70
60
50
40
20
10
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31


99
101
100
90
80
70
60
50
40
30
20
10
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31





Itération de la procédure et classification finale

Figure IV.8

- soit au nombre d'hyperparallélépipèdes se partageant le même sommet si l'histogramme à amincir est déterminé sur la grille des sommets,

- soit au nombre des sommets de l'hyperparallélépipède si l'histogramme à amincir est déterminé sur la grille des centres.

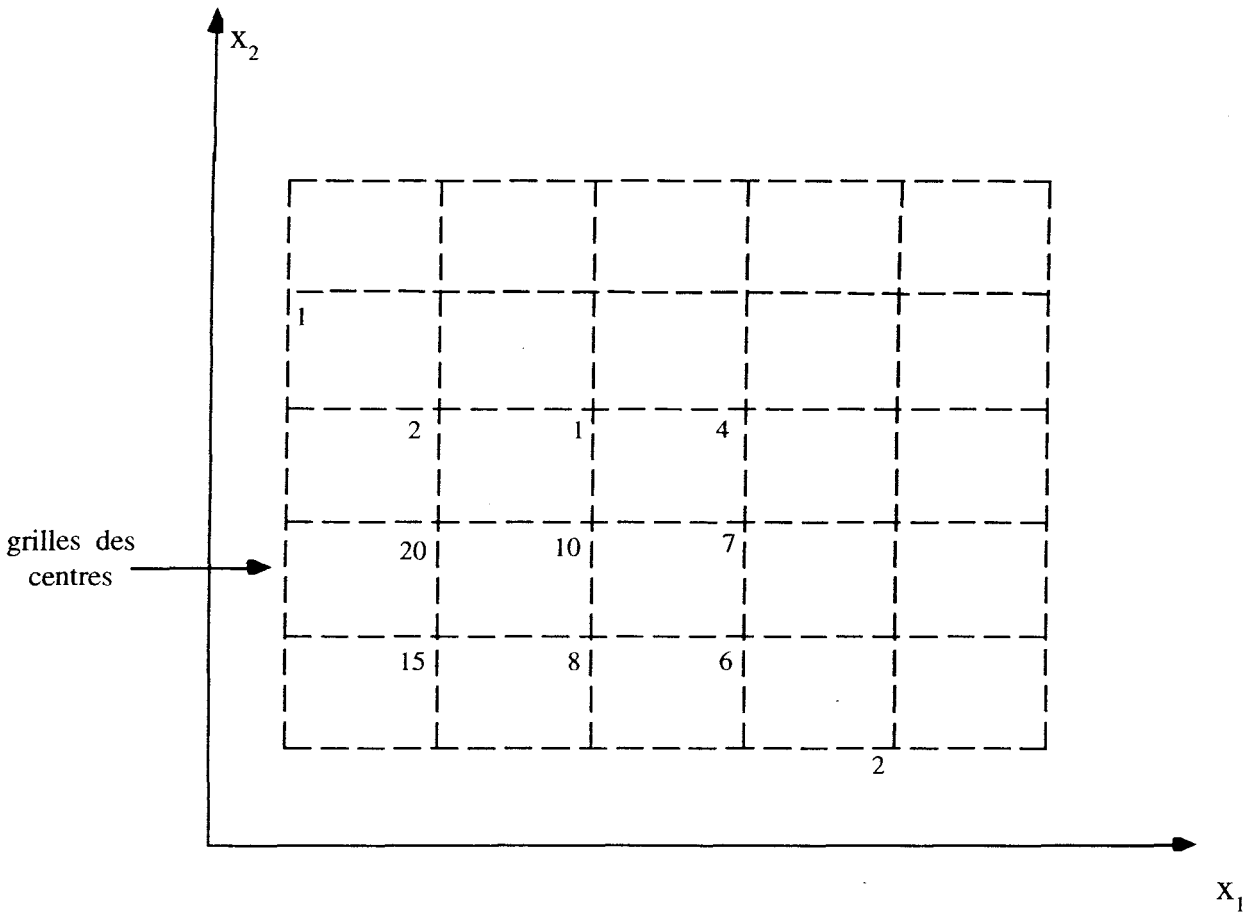
Parmi toutes les possibilités, on retiendra celle qui donne un regroupement fictif de taille maximale.

Cette procédure de regroupement par migration des observations basée sur l'amincissement de l'histogramme est une procédure itérative : le regroupement se fait une fois sur la grille des centres et une autre fois sur la grille des sommets. La procédure s'arrête d'elle-même lorsqu'aucun regroupement ne s'avère possible en respectant les règles exposées précédement.

Nous illustrons l'application de cette procédure sur deux histogrammes de distributions multidimensionnelles, l'un correspond à une distribution unimodale bidimensionnelle, l'autre à une distribution multimodale bidimensionnelle.

IV - 3 - 1 . Exemple 1 : distribution unimodale et bidimensionnelle.

La figure IV.9 représente l'histogramme d'une



L'histogramme unimodal bâti sur la grille des centres de l'exemple 1
N = 2 et D = 5

Figure IV.9

Nous remarquons :

- qu'à chaque itération la vallée se creuse de plus en plus et que les modes s'amplifient,

- que les emplacements des modes et le nombre d'observations par classe ont été préservés lors de la procédure de classification.

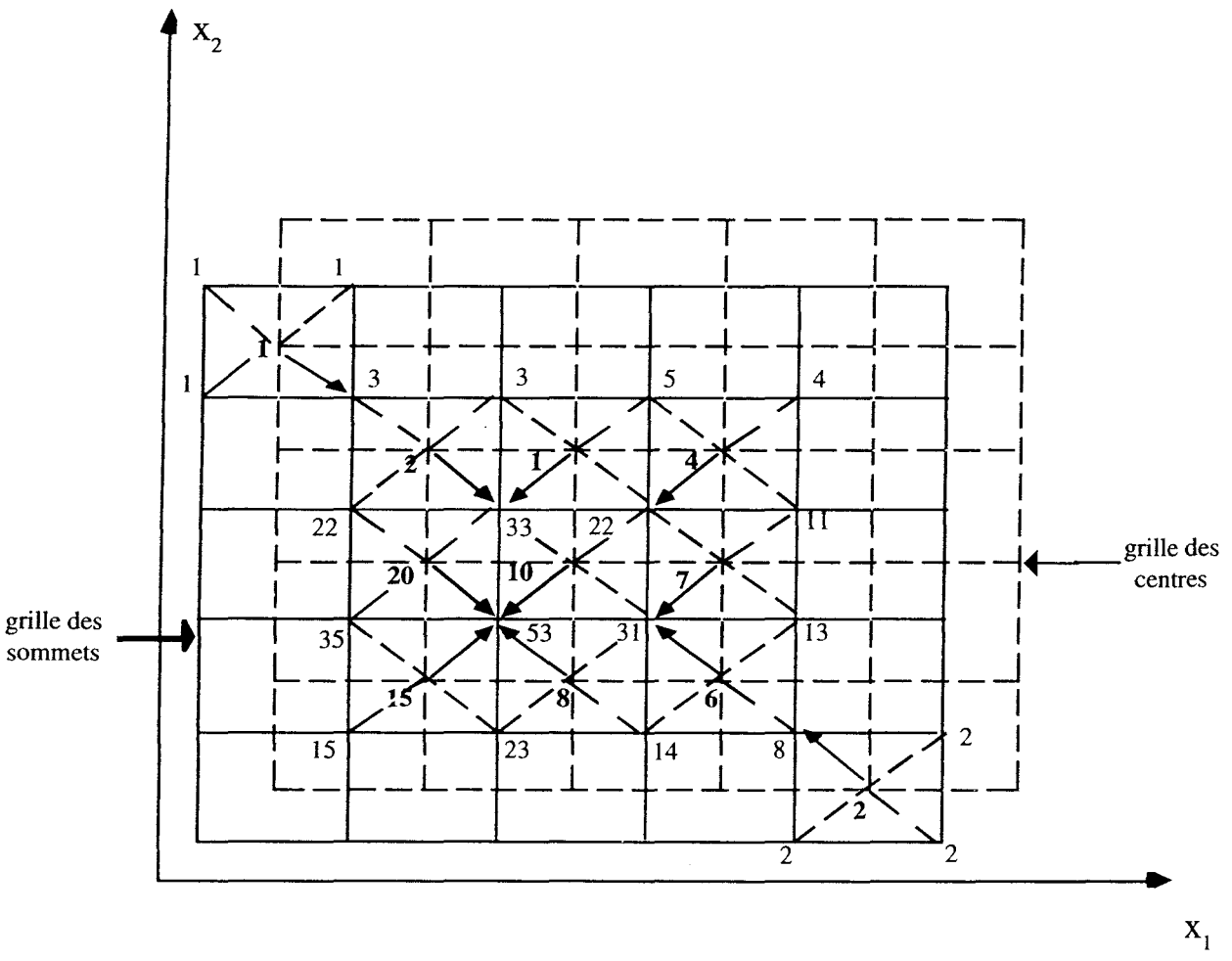
IV - 3 . RENFORCEMENT DES MODES ET CLASSIFICATION DE DONNEES MULTI-DIMENSIONNELLES

L'intérêt de la procédure de renforcement des modes présentée dans l'article précédent (Cf. § IV.2) est sa possibilité d'extension au cas d'un histogramme multi-modal représentant la distribution de données multidimen-sionnelles.

Le renforcement des modes d'une distribution multimodale multidimensionnelle s'effectue par déplace-ments par blocs des sous-ensembles d'observations $0(\alpha)$ d'adresse unidimensionnelle α définis par les valeurs de l'histogramme $H(\alpha)$. Chacun de ces sous-ensembles a le choix, dans un espace à N dimensions, entre 2^N possibilités de migration. Ces 2^N possibilités sont égales :

distribution bâti sur la grille des centres dans un espace bidimensionnel discrétisé avec cinq intervalles par axe. Cette distribution est inconnue a priori.

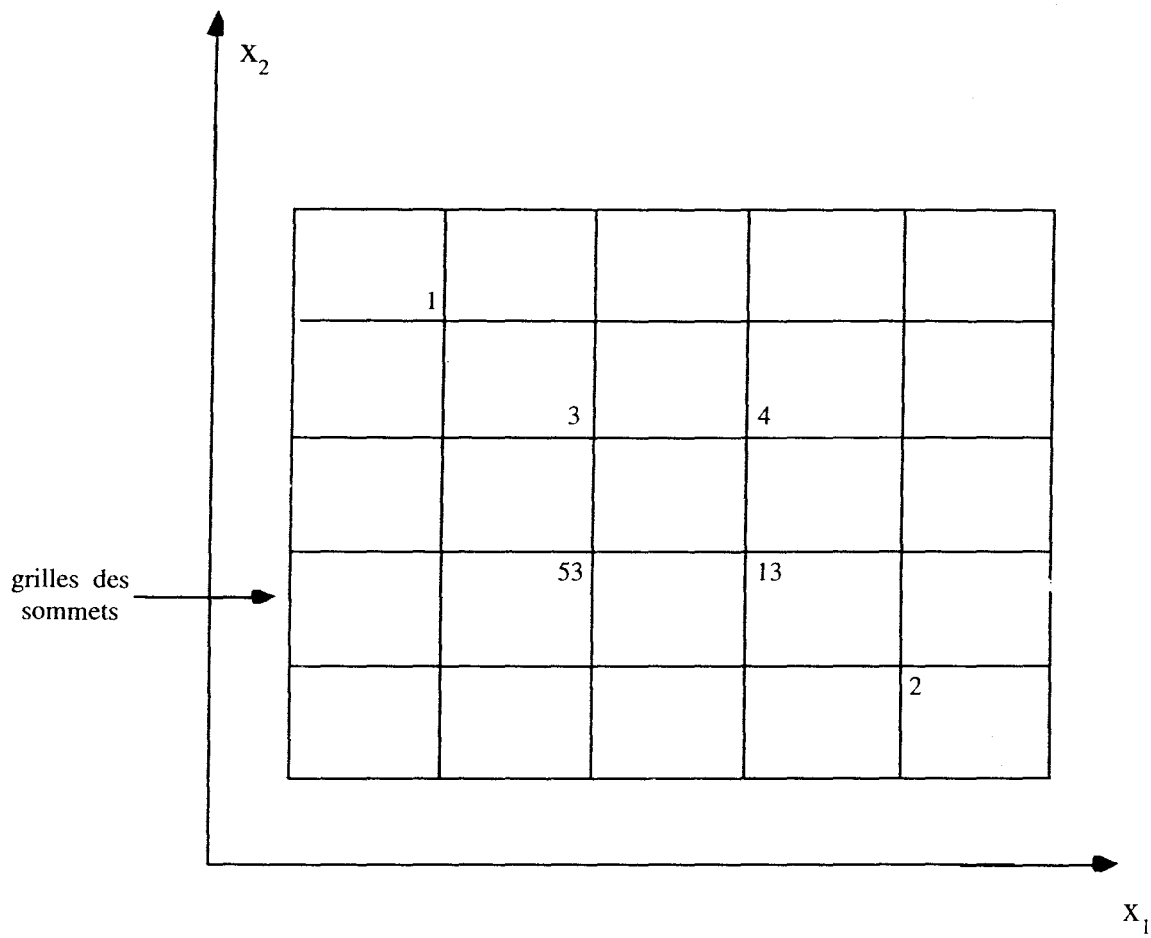
Chaque sous-ensembles $O^1_{ce}(\alpha)$ peut donc être translaté en bloc vers l'un des 2^2 sommets de l'hyperparallélépipède d'adresse unidimensionnelle α. Ces 4 sommets, conformément au système de numérotation introduit au chapitre III (Cf. § III-4), ont comme adresses unidimensionnelles α, $(\alpha + 1)$, $(\alpha + D + p)$, et $(\alpha + D + p + 1)$, où D est le nombre d'intervalles et (p-1) l'écart artificiel de numérotation entre tout couple de sommets se trouvant dans une situation de faux voisinage (Cf. § III-5-2). Les valeurs respectives des regroupements fictifs ainsi obtenus sur la grille des sommets forment l'histogramme fictif $FH^1_{so}(d)$ (Cf. figure IV.10). Les flêches pointillées indiquent les migrations possibles qui ne maximisent pas la taille des regroupements fictifs. Les migrations qui conduisent à des regroupements fictifs de taille maximale sont indiquées par les flêches pleines (Cf. figure IV.10). Ces dernières sont retenues pour déplacer les sous-ensembles d'observations $O^1_{ce}(\alpha)$. Nous remarquons que lorsque toutes ces migrations sont effectuées simultanément, on regroupe éventuellement et réellement en chaque sommet soit aucun, soit un, soit deux, soit trois, soit quatre sous-ensembles d'observations.



— — Migrations possibles sans maximisation de la taille des regroupements.

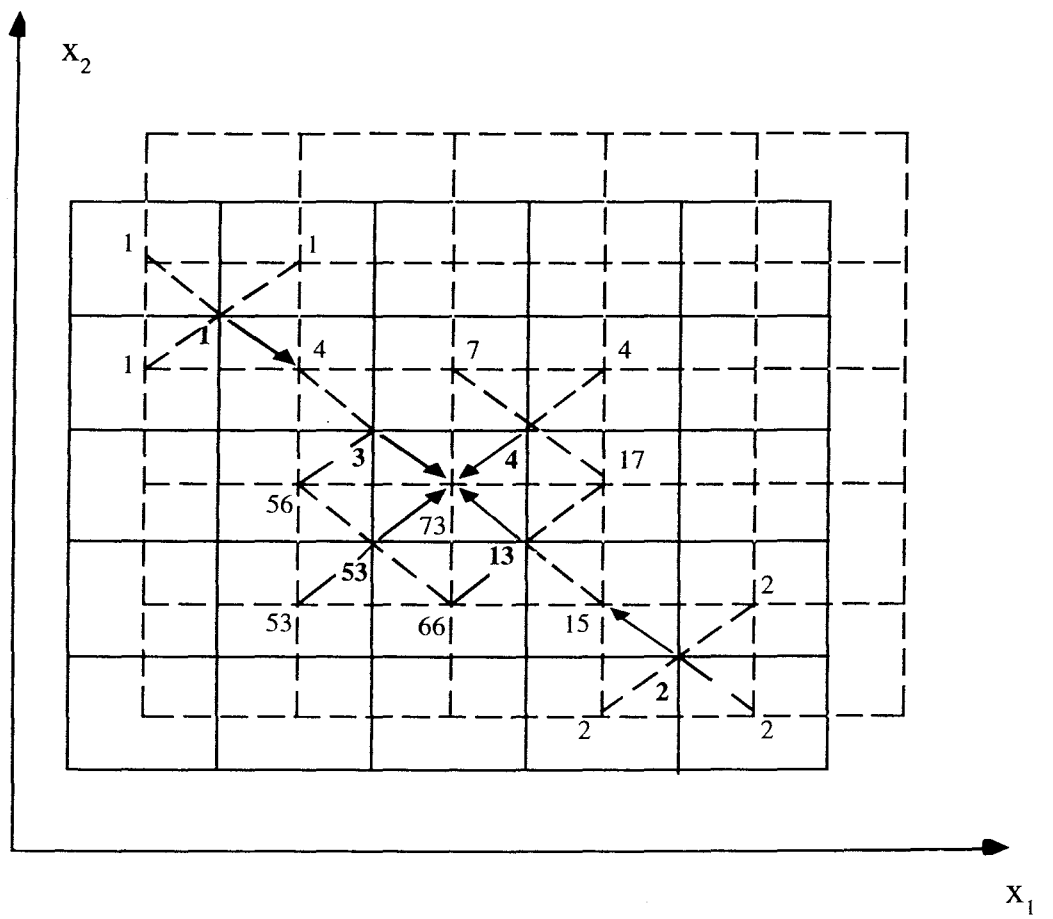
——► Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.10



Nouvel histogramme sur la grille des sommets
N = 2 et D = 5

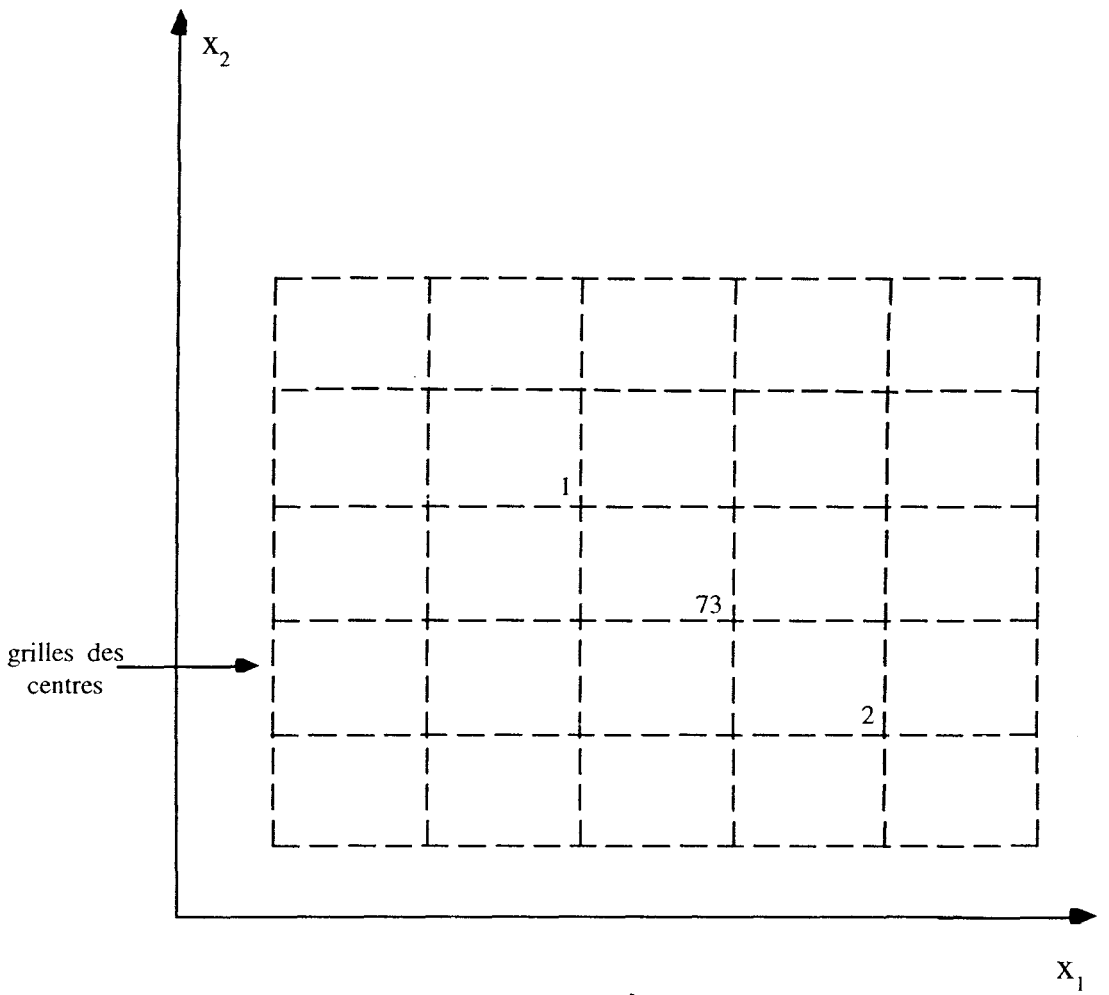
Figure IV.11



— — Migrations possibles sans maximisation de la taille des regroupements.

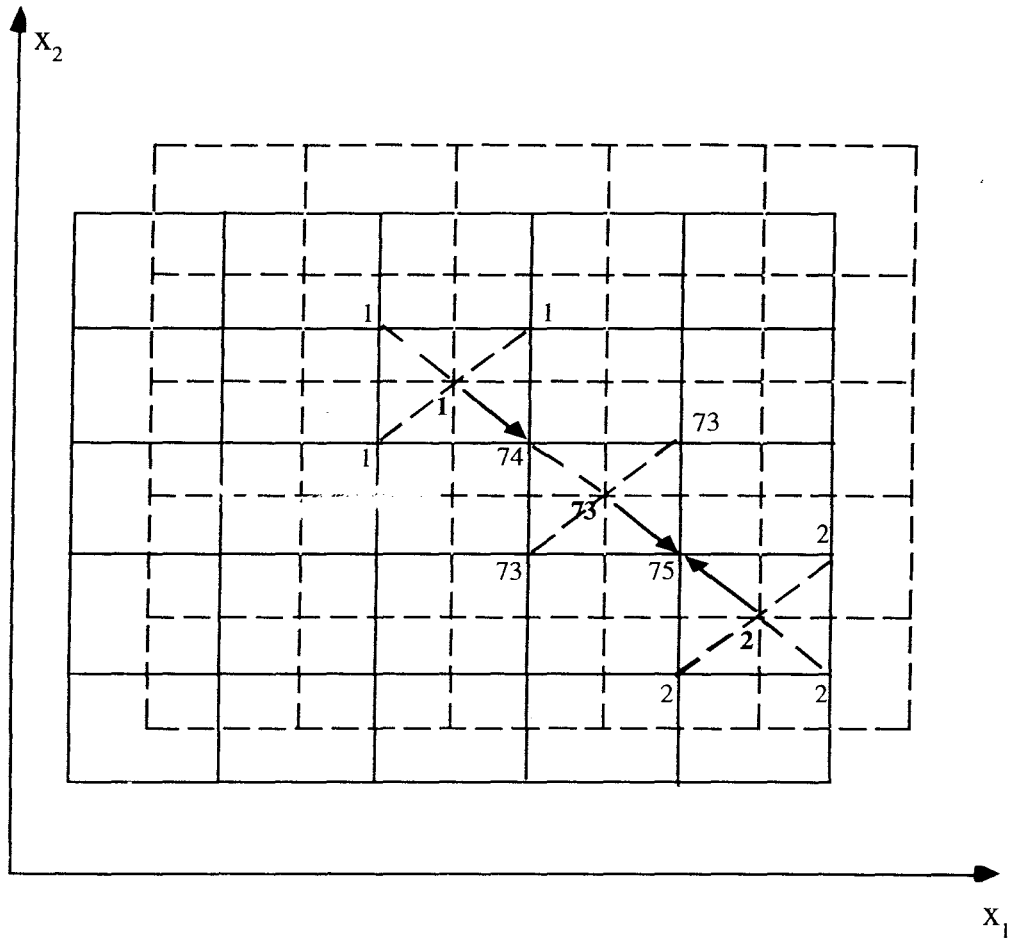
——▶ Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.12



Nouvel histogramme sur la grille des centres

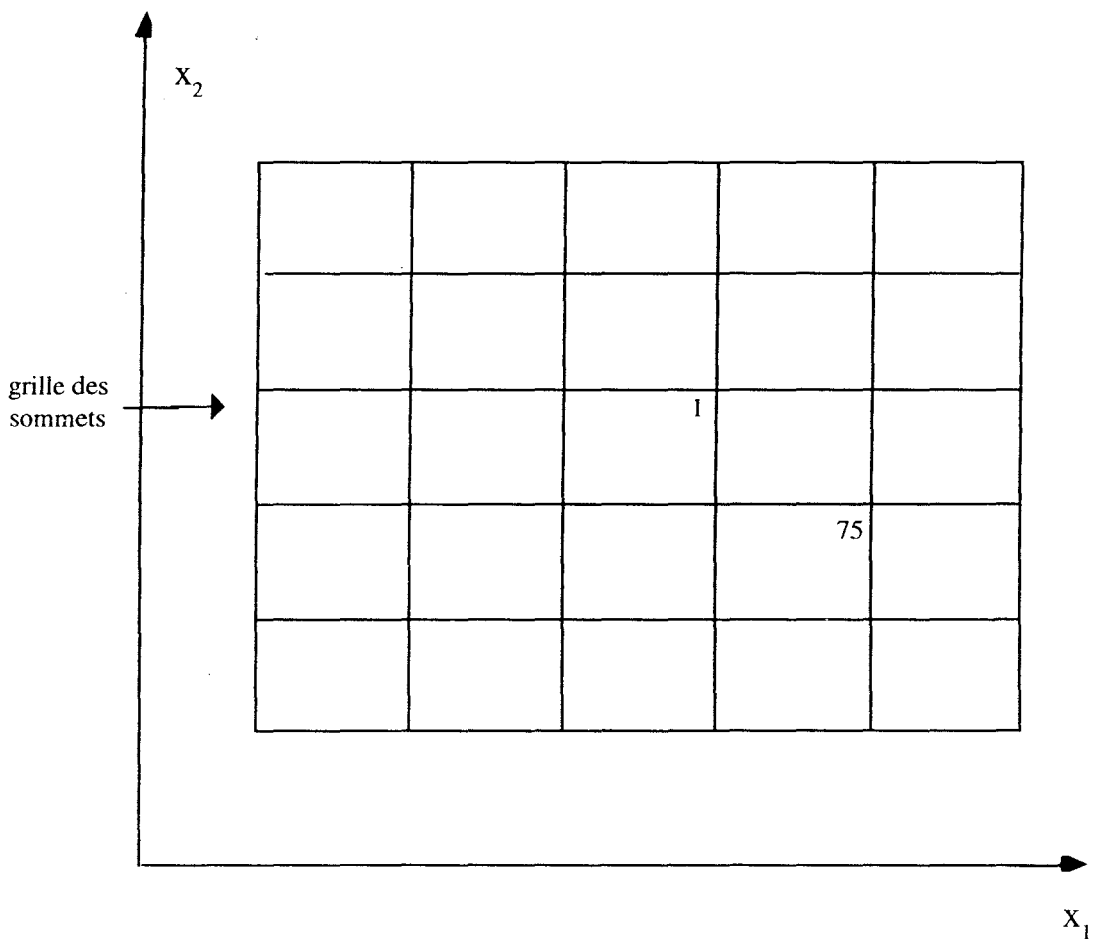
Figure IV.13



— — Migrations possibles sans maximisation de la taille des regroupements.

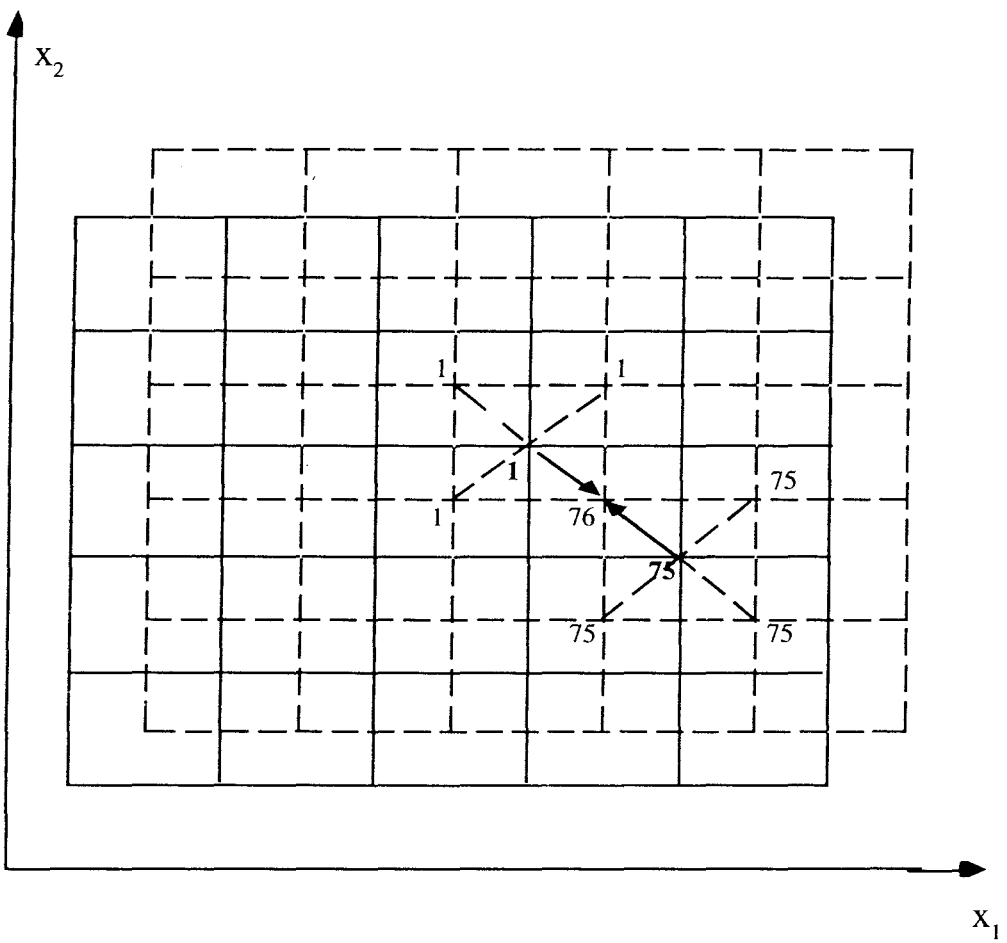
→ Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.14



Nouvel histogramme sur la grille des sommets

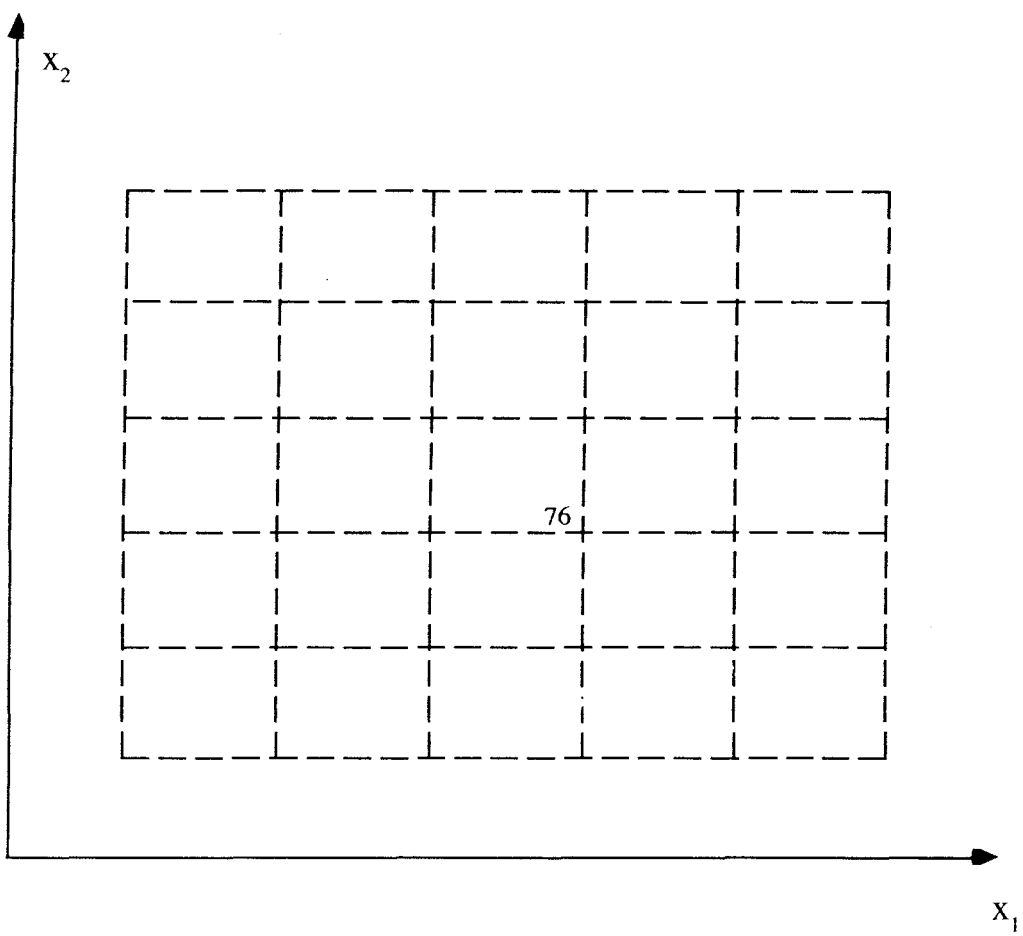
Figure IV.15



— — Migrations possibles sans maximisation de la taille des regroupements.

⟶ Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.16



Arrêt des itérations avec une seule classe regroupant 76 observations

Figure IV.17

On obtient ainsi le nouvel histogramme réel cette fois-ci $RH^{1}_{so}(\alpha)$ (Cf. figure IV.11) sur la grille des sommets dont les valeurs résultent des déplacements par blocs de tous les sous-ensembles d'observations qui ont favorisé les regroupements fictifs de taille maximum.

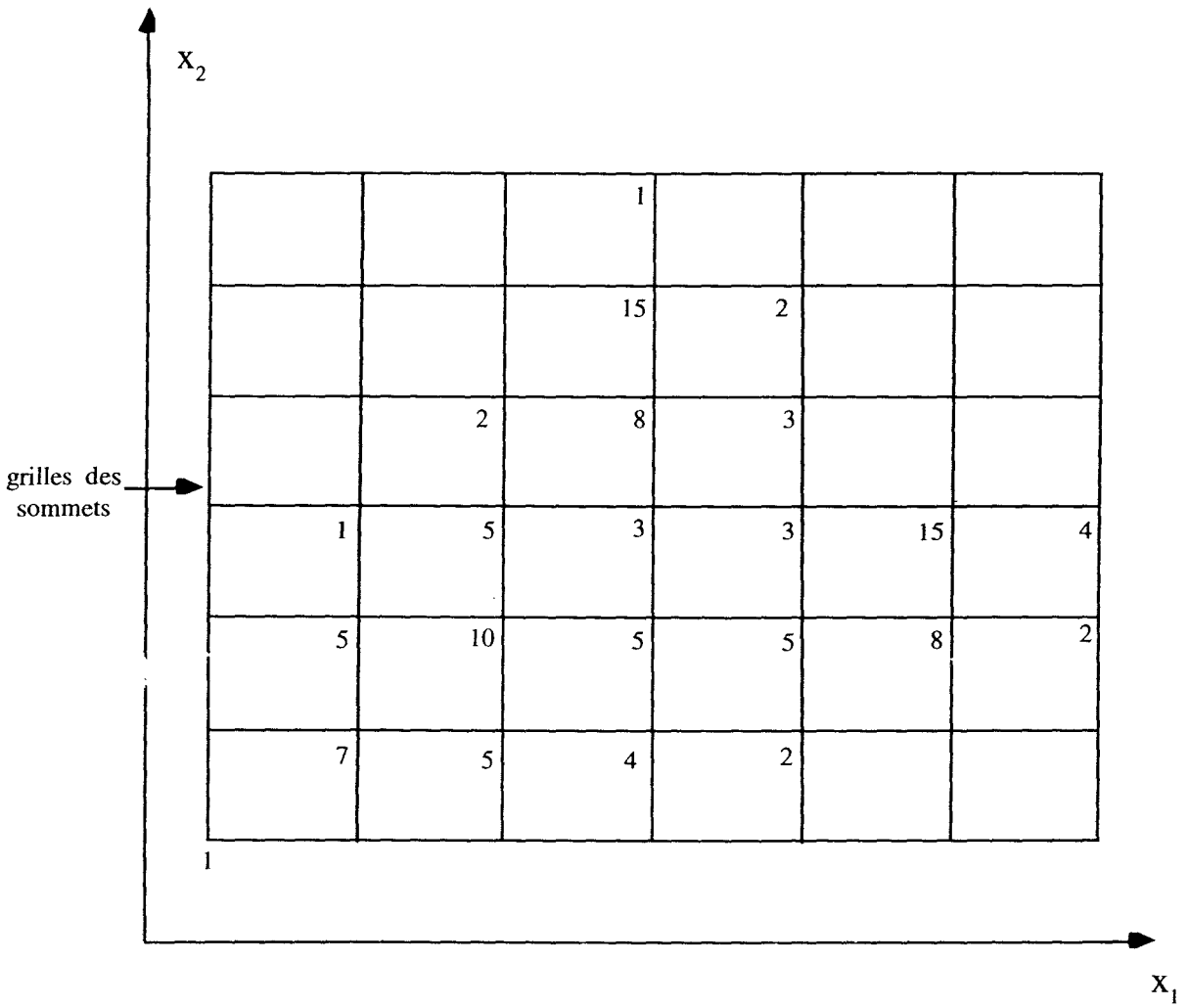
Cette technique est itérée jusqu'à ce qu'aucun autre regroupement ne soit possible. La séparation des modes et l'indication du nombre d'observations assignées à chaque classe sont alors déterminées.

Les figures IV.10 à IV.17 représentent toutes les étapes itératives de regroupements sur la grille des centres et celle des sommets qui conduisent à l'amincissement de l'histogramme.

IV - 3 - 2. Exemple 2 : distribution multimodale et bidimensionnelle.

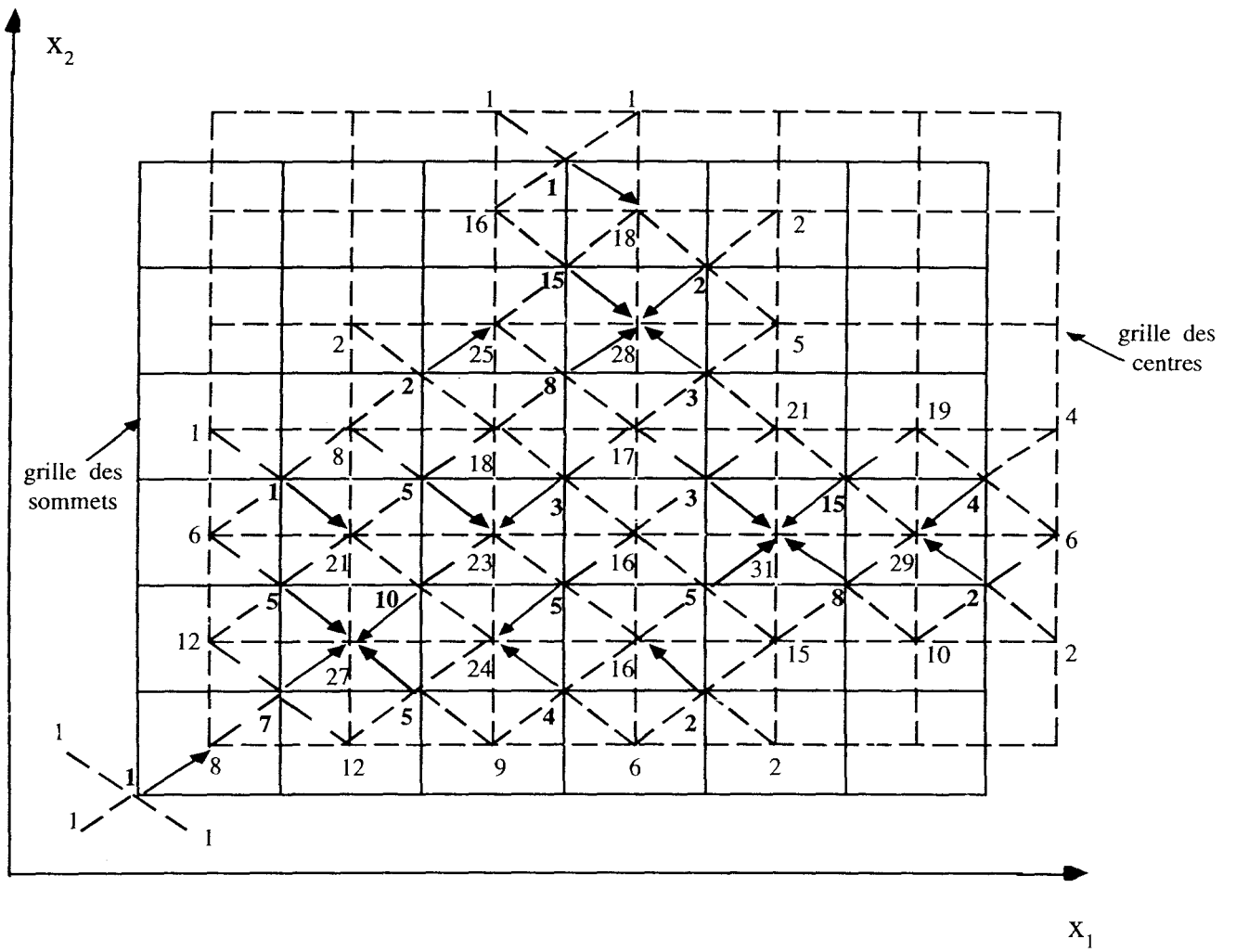
La figure IV.18 représente l'histogramme d'une distribution trimodale bâti sur la grille des sommets dans un espace bidimensionnel discrétisé avec six intervalles par axe. Cette distribution est inconnue a priori.

Les figures IV.18 à IV.28 représentent toutes les étapes itératives des regroupements sur les grilles des centres et des sommets à partir de la construction de l'histogramme des observations, jusqu'à leur classification en un nombre de classes inconnu a priori.



L'histogramme
trimodal bâti sur la grille des sommets de l'exemple 2.

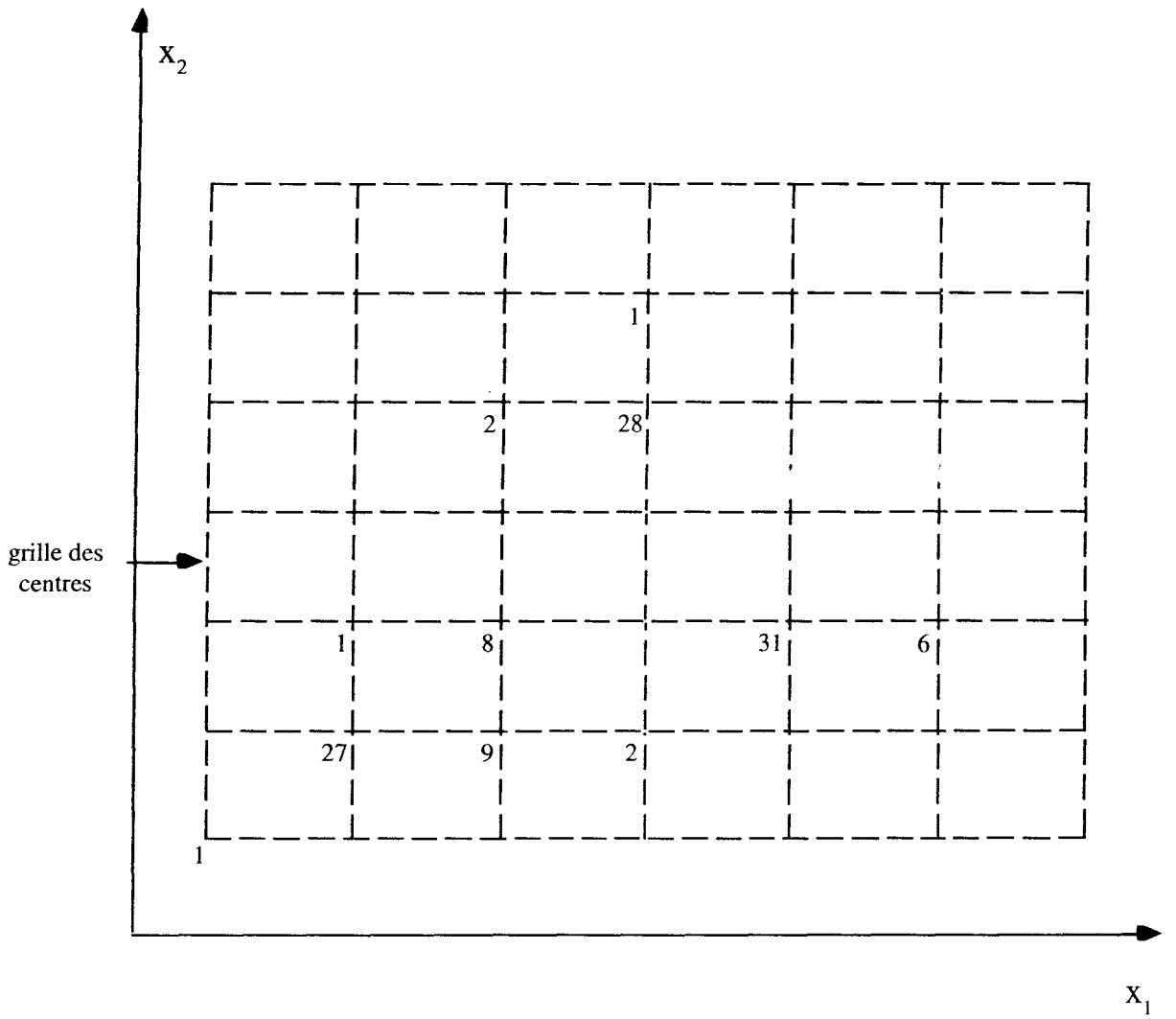
Figure IV.18



— — Migrations possibles sans maximisation de la taille des regroupements.

——► Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.19



Nouvel histogramme sur la grille des centres

Figure IV.20

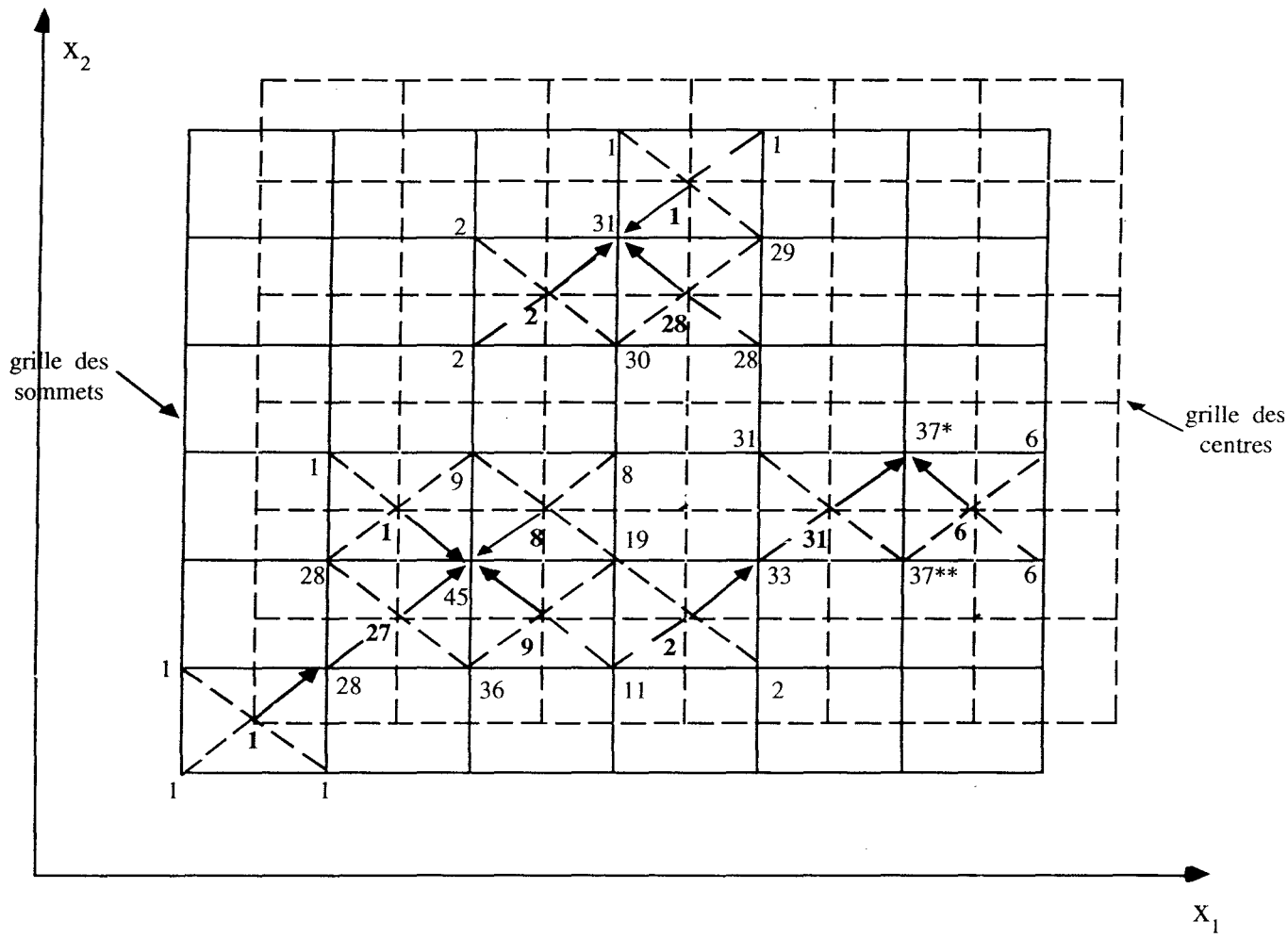
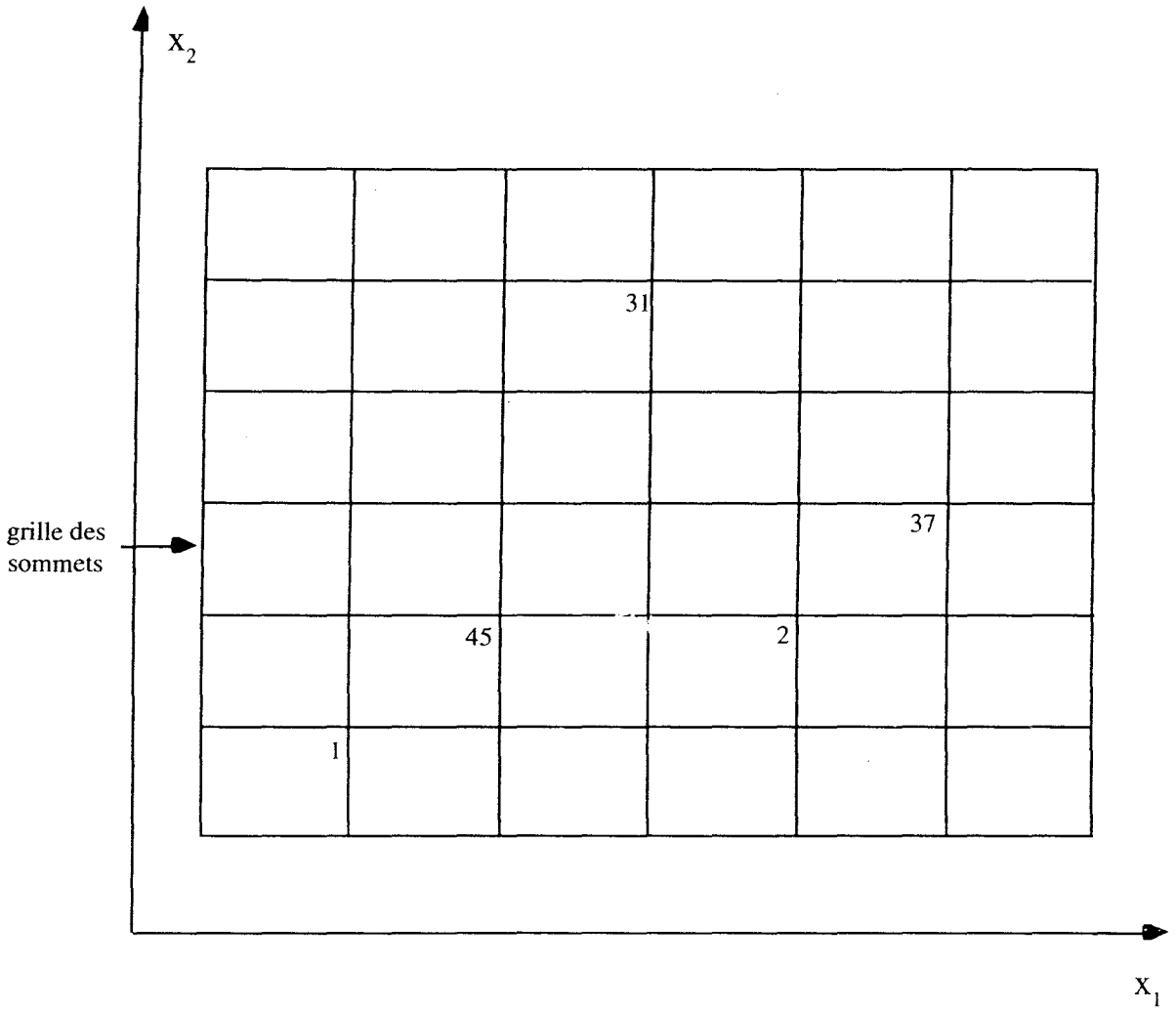

X2
X1
grille des sommets
grille des centres
1
1
2
31
1
29
2
28
2
30
28
31
37*
6
1
9
8
1
8
19
31
6
28
45
33
37**
6
27
9
2
1
28
36
11
2
1
1
1
Migrations possibles sans maximisation de la taille des regroupements.
Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.21



Nouvel histogramme sur la grille des sommets

Figure IV.22

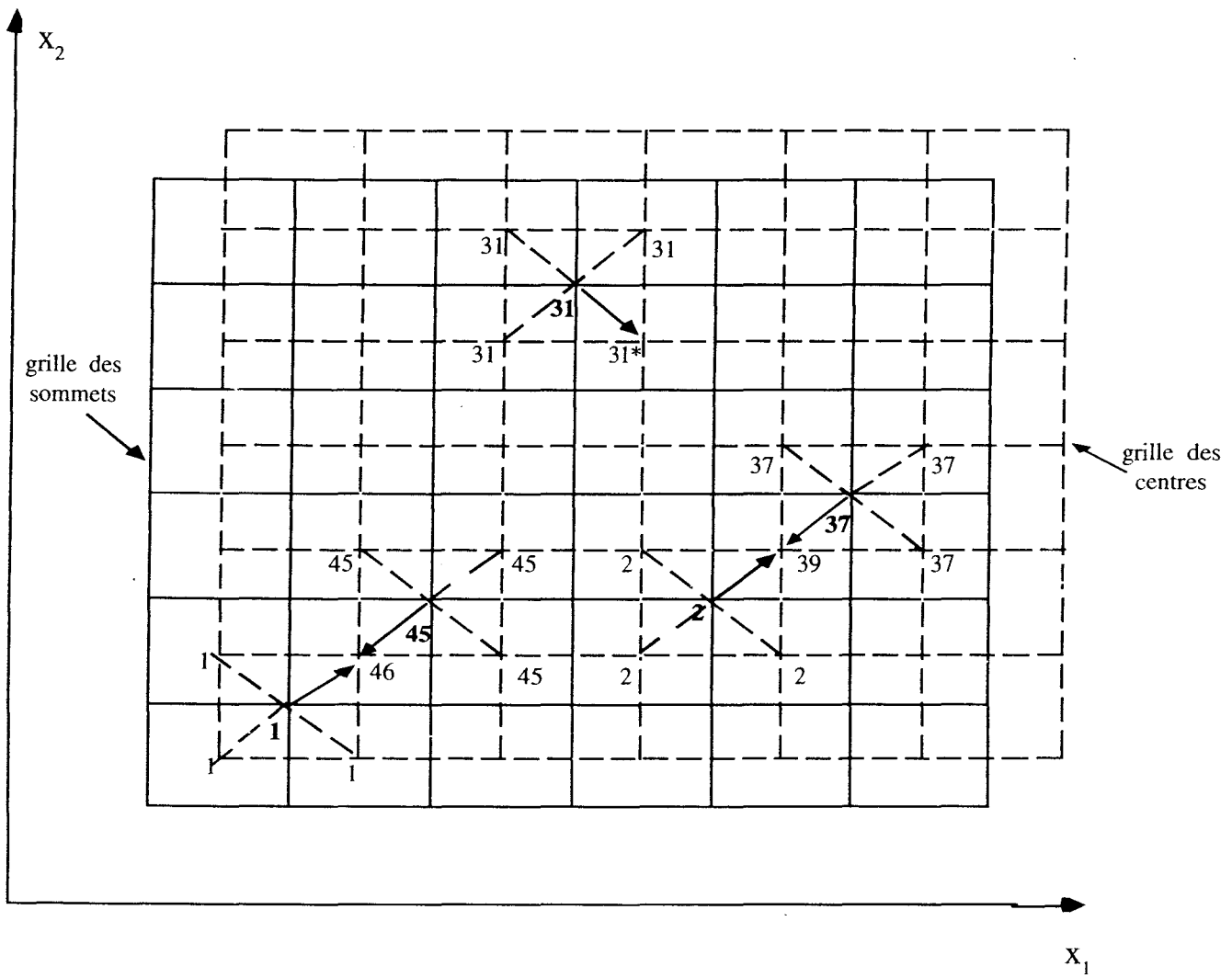
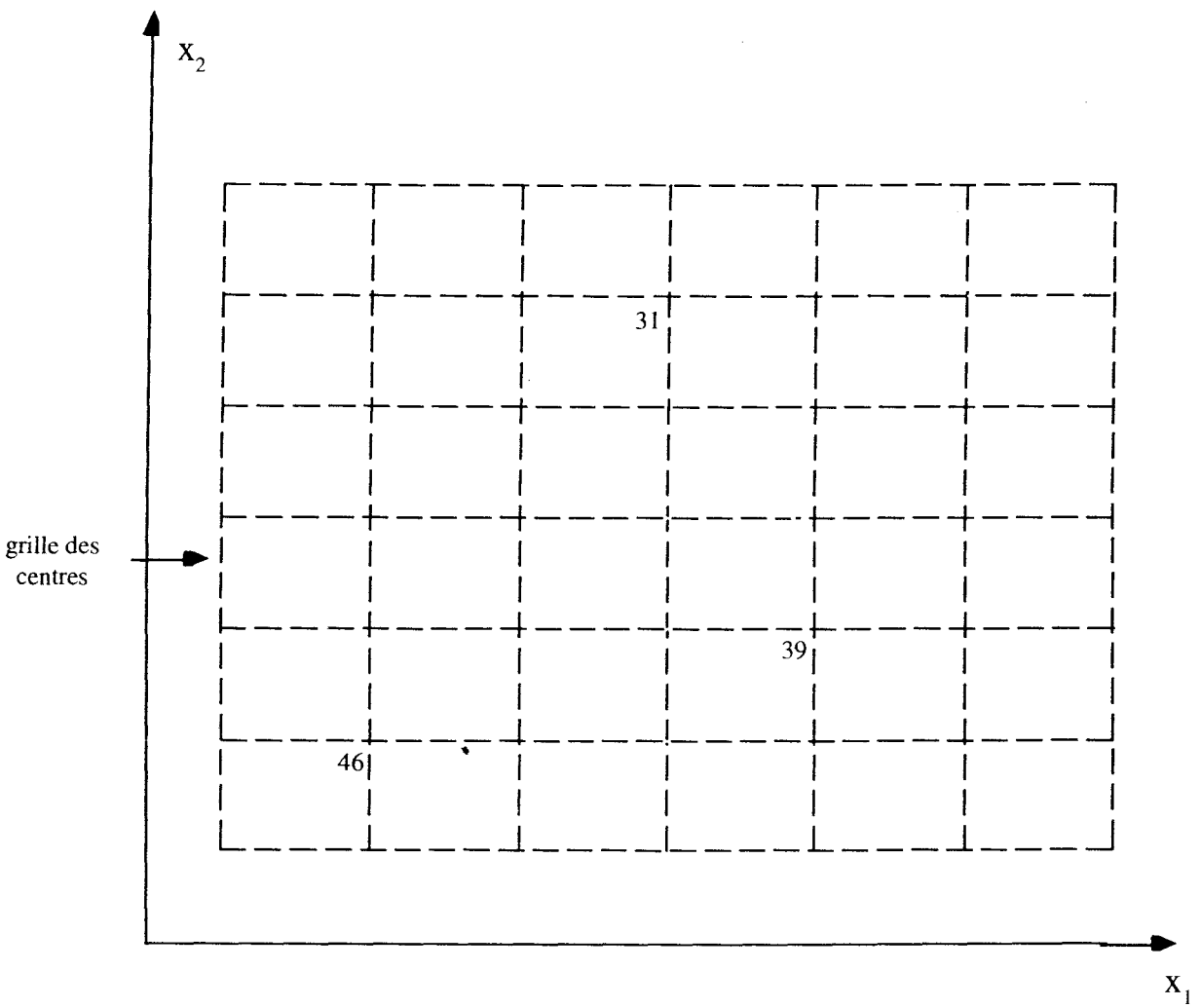

X2
grille des sommets
grille des centres
31
31
31
31
31*
37
37
37
37
45
45
2
39
37
45
2
45
1
46
45
2
2
1
1
1
X1
Migrations possibles sans maximisation de la taille des regroupements.
Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.23



Nouvel histogramme sur la grille des centres

Figure IV.24

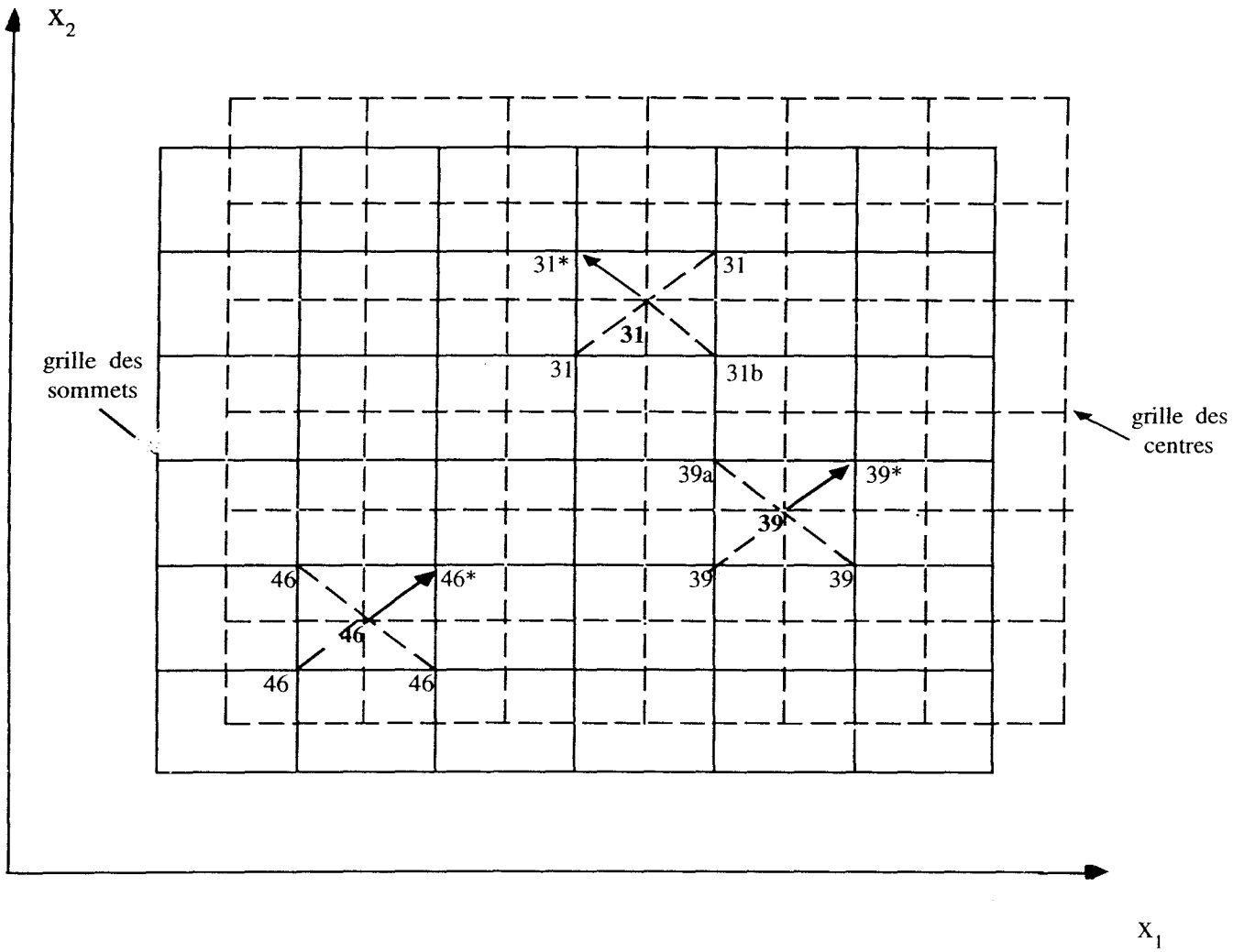
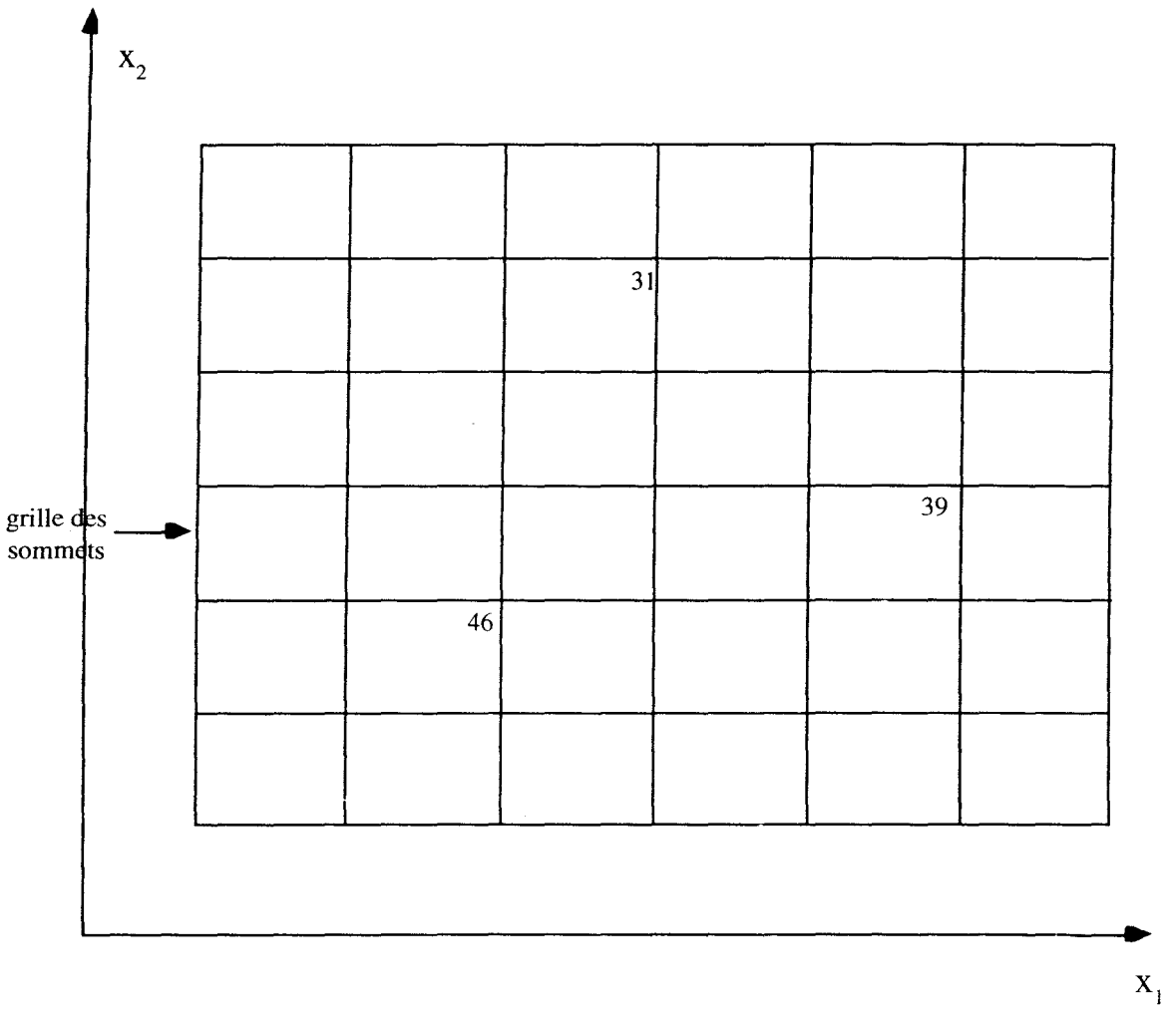

X_2
grille des sommets
31*
31
31
31
31b
grille des centres
39a
39*
39
46
46*
46
46
46
39
39
X_1
Migrations possibles sans maximisation de la taille des regroupements.
Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.25



Nouvel histogramme sur la grille des sommets

Figure IV.26

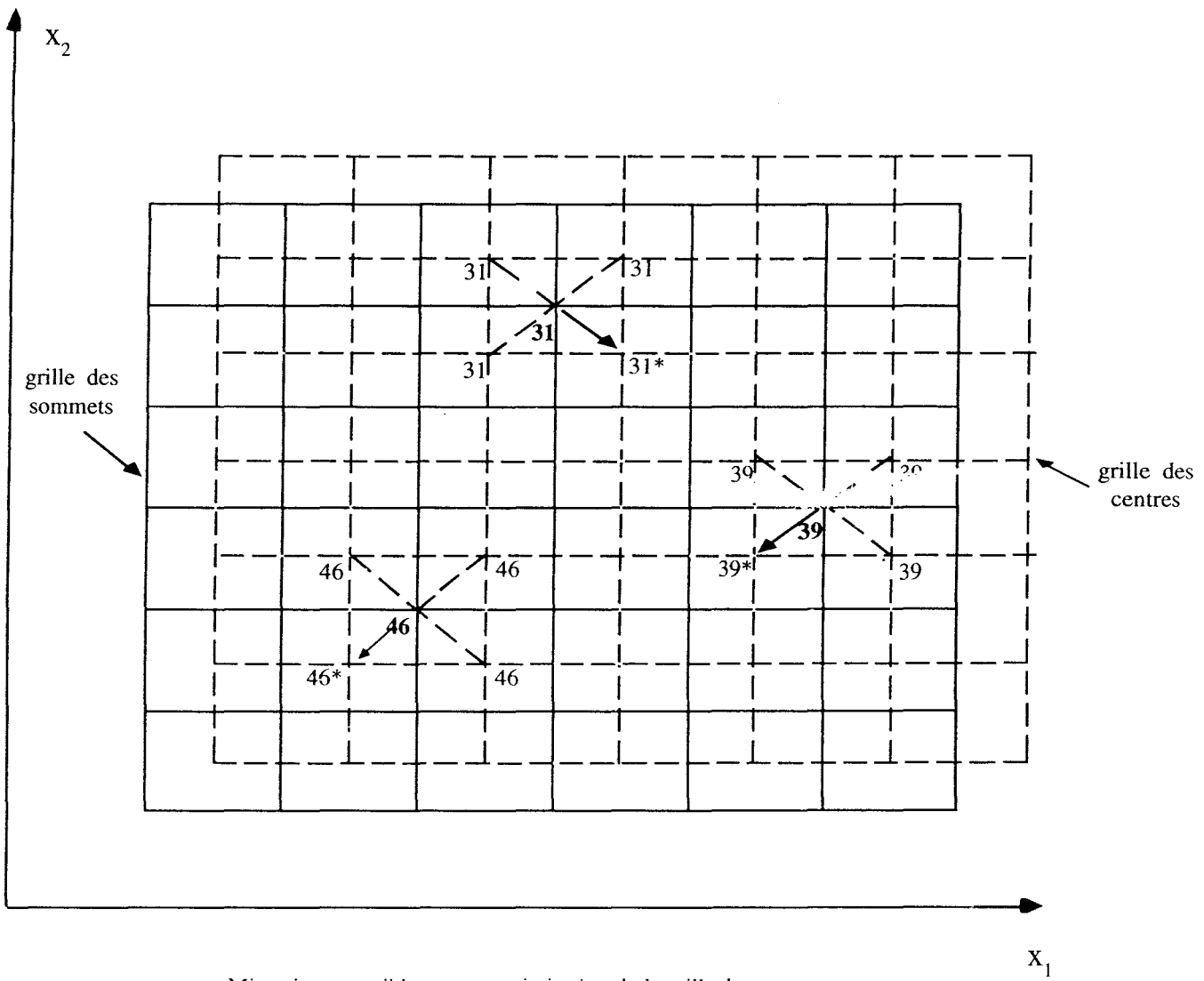
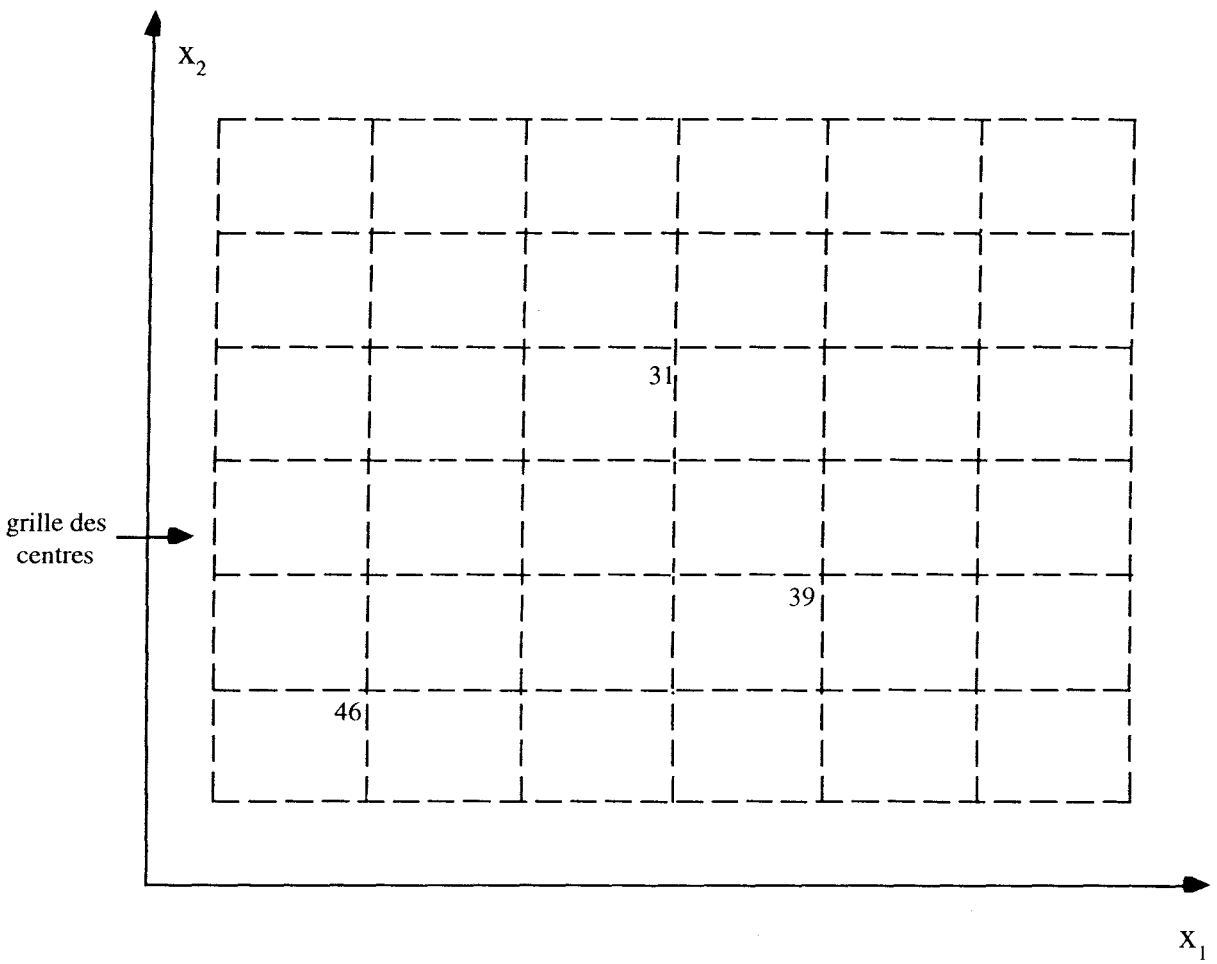

X2
grille des sommets
31
31
31
31
31*
grille des centres
39
39
39*
39
46
46
46
46*
46
X1
Migrations possibles sans maximisation de la taille des regroupements.
Migrations retenues avec maximisation de la taille des regroupements.

Figure IV.27



Résultat final de la procédure

3 modes mis en évidence

Figure IV.28

Nous profitons de cet exemple pour montrer comment certaines ambiguïtés au niveau des regroupements peuvent être levées. En effet, nous avons parfois à choisir entre plusieurs possibilités de migration qui maximisent toutes la taille des regroupements fictifs au même niveau. C'est le cas de la figure IV.21, où nous avons deux possibilités de regroupement de même taille égale à 37. Il en est de même de la figure IV.23 où le regroupement constitué de 31 observations possède quatre possibilités de migration d'égale taille maximale. Le nombre de ces possibilités ambiguïes de migration peut atteindre 2^N, nombre de déplacements possibles dans un espace à N dimensions. Dans tous ces cas, nous choisissons, la direction de migration qui lors de la première itération sur cette même grille avait favorisé le regroupement maximal. Cette règle prend en considération la distribution réelle des données dans l'espace Euclidien. Son application aux regroupements déjà définis dans une étape précédente stabilise ceux-ci et empêche leur migration aléatoire. Sur la figure IV.25, il est évident que si cette règle n'était pas appliquée et que le choix de migration des deux regroupements de 39 observations et de 31 observations s'orientait vers les sommets (a) et (b) respectivement, ces deux regroupements auraient fusionné lors de l'étape suivante. Cette stabilisation fait aussi partie du critère d'arrêt de la procédure automatique. Ce critère n'est pas un critère statistique ou de définition du nombre de classes, c'est un critère de comparaison.

Si deux itérations successives sur la même grille donnent le même nombre de classes, une même répartition des observations entre les classes en présence et un même emplacement des pics définissant le mode de chaque classe, la procédure de classification est arrêtée automatiquement.

IV - 4. COMPARAISON AVEC UNE METHODE CLASSIQUE DE MIGRATION DES OBSERVATIONS.

La méthode présentée dans ce chapitre consiste essentiellement à provoquer des déplacements élémentaires des observations de telle sorte qu'au cours de leurs migrations, elles se concentrent en des points privilégiés de l'espace pour constituer les classes recherchées.

Cette idée de faire migrer les observations dans l'espace de représentation des données n'est pas nouvelle. Fukunaga et Hostetler, [Fukunaga et Hostetler, 1975], ont proposé de déplacer les observations de manière itérative selon le schéma itératif

$$X^{i+1}_q = X^i_q + a\ Z\ (X^i_q) \qquad \text{IV-17}$$

qui indique que l'observation X_q est déplacée à chaque pas dans la direction du gradient $Z(X^i_q)$ de la fonction de densité sous-jacente estimé au point X^i_q.

Peu après, cette idée a été reprise par Koontz, Narendra et Fukunaga [Koontz, Narendra et Fukunaga, 1976], qui approchent les dérivées directionnelles de la fonction de densité de probabilité par de simples différences finies. Chaque observation est alors déplacée vers une de ses voisines dans la direction de la plus grande dérivée directionnelle calculée dans son voisinage immédiat.

IV - 4 - 1 . Méthode de la plus grande pente

Pour être plus précis, l'idée de base dans cette méthode est de regrouper les sous-ensembles formés lors de la construction élémentaire de l'histogramme multidimensionnel en les rattachant au sous-ensemble voisin le plus dense. Ce rattachement permet d'agréger les observations autour de modes locaux. Chaque groupement ainsi formé autour d'un mode est considéré comme une classe. Cet algorithme utilise le concept d'arbres ou de graphes orientés. L'idée de base de l'algorithme est de définir une règle de construction d'un ou plusieurs arbres orientés à partir des Q observations à classer. Chaque arbre définit une région modale. Pour cela, la valeur associée à chaque cellule de l'histogramme est examinée et comparée à celles de ses voisines. Trois cas peuvent se présenter :

- La valeur associée à la cellule centrale de l'histogramme est supérieure à toutes celles de ses voisines : cette cellule centrale est alors appellée racine, elle correspond à un mode local. Si d'autres cellules voisines ont la même valeur de l'histogramme que celle associée à la cellule centrale, la racine est choisie arbitrairement entre celles-ci.

- La valeur de l'histogramme associée à la cellule centrale est inférieure à celle d'une ou de plusieurs de ses voisines. Dans ce cas, cette cellule centrale est appelée nœud et doit être chaînée à sa voisine dans la direction de la plus grande pente de la distribution sous-jacente. Cette pente est approchée, dans chaque direction, par le quotient de la différence entre les deux valeurs $H(A_i)$ et $H(A_j)$ de l'histogramme asociées aux deux cellules considérées et de la distance d_{ij} les séparant :

$$G_{ij} = \frac{H(A_j) - H(A_i)}{d_{ij}} \qquad \text{IV-18}$$

où G_{ij} est la pente entre la cellule centrale d'adresse A_i et sa voisine d'adresse A_j ,

La cellule voisine à laquelle est chaînée la cellule centrale ou nœud est alors appelée le père ou le parent de cette dernière. La direction de la chaîne va du nœud vers le père :

- si les valeurs de l'histogramme de la cellule centrale et de celles de toutes ces voisines sont égales, nous sommes en présence d'un plateau.
  En pratique, pour chaque cellule A_i on calcule pour toutes ses voisines les pentes G_{ij}, et si

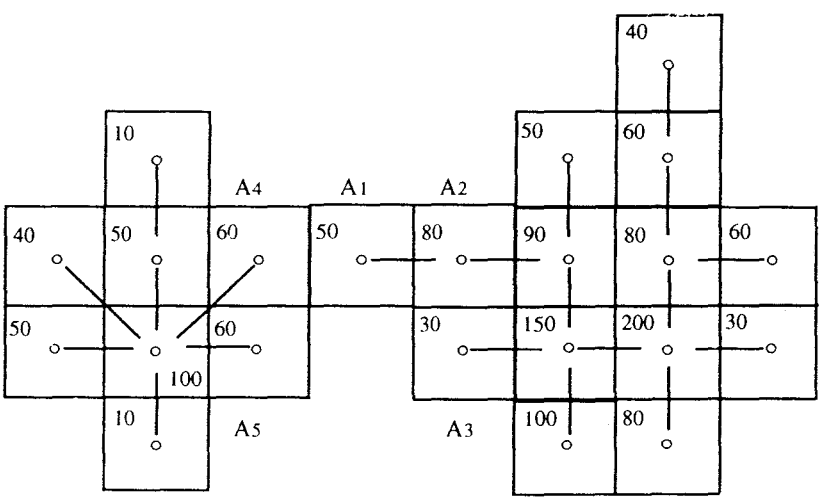
  $G_{ij} < 0$ pour toutes les voisines, la cellule considérée est une racine,

  $G_{ij} > 0$ cette cellule centrale est un nœud et a pour père sa cellule voisine pour laquelle G_{ij} est maximum et à laquelle elle est chaînée,

  $G_{ij} = 0$ nous sommes en présence d'un plateau.

Cette règle permet donc de construire un ensemble d'arbres dans lesquels chaque cellule est représentée une et une seule fois. Chaque racine constitue une estimation d'un mode de la fonction de densité sous-jacente à l'échantillon étudié, tandis que les arbres orientés ainsi obtenus sont des représentations des régions modales.

La figure IV.29 illustre une application de cet algorithme dans le cas d'un espace bidimensionnel. Chaque cellule contient la valeur associée de l'histogramme. Le voisinage considéré est défini par la 8-connexité. Pour chaque cellule, on compare les pentes de la distribution dans les directions de chacune de ses voisines. Elle est finalement chaînée à la cellule pour laquelle cette pente est maximum.



Méthode de la plus grande pente pour un espace à deux dimensions

Figure IV.29

IV - 4 - 2 . Différence entre la méthode de la plus grande pente et l'amaicissement de l'histogramme.

IV - 4 - 2 - 1 . Méthodologie

Comme nous l'avons déjà mentionné, les deux méthodes que nous comparons consistent à classifier les observations en les assignant au mode le plus proche. La méthode classique utilise pour ceci un calcul de pente, tandis que la méthode proposée se base sur une suite de regroupements qui favorisent les concentrations locales d'observations.

La classification par estimation de la pente préconise la définition de la valeur du paramètre "a" dans les limites définies par l'équation IV-17. Si "a" est choisi trop grand, la convergence de la procédure n'est pas garantie et la procédure peut générer un nombre important de classes non significatives [Fukunaga et Hostetler, 1975].

Quant à elle, la classification par maximisation de la taille des regroupements est affectée par le choix du pas de discrétisation. Nous présenterons au chapitre V une solution au problème de l'ajustement de ce pas.

IV - 4 - 2 - 2 . Effet collectif du voisinage

La méthode classique basée sur le calcul de la pente de la fonction de densité, est sensible à la notion de distance entre les cellules. Considérons la cellule d'adresse A_1 de la figure IV.30. Conformément aux règles de la méthode de la plus grande pente, cette cellule est un nœud dont le père est la cellule d'adresse A_2. Aussi les observations contenues dans cette cellule se trouvent à la jonction des deux classes de la distribution. Si l'on considère les observations se trouvant dans toutes les cellules entourant celle d'adresse A_1, nous remarquerons que la pente moyenne calculée pour les cellules d'adresses A_4 et A_5 est plus grande que celle des cellules d'adresses A_2 et A_3. C'est pourquoi ces 50 observations contenues dans A_1 doivent plutôt être chaînées aux cellules A_4 et A_5 et non à A_2. Elles intègrerons ainsi la classe de gauche et non celle de droite.

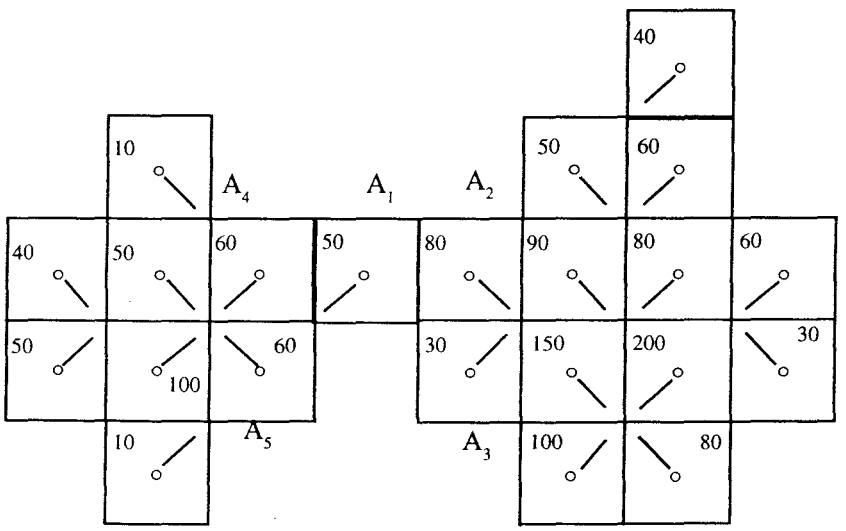
Pour un espace à N dimensions, le calcul des pentes et de leur moyenne devient laborieux.

La méthode proposée, n'étant point une méthode métrique, ne présente pas ces difficultés de calcul et, les 50 observations de la cellule d'adresse A_1

se dirigent vers cette classe de gauche suivant le principe de la maximisation de la taille des regroupements fictifs (Cf. figure IV.30).

Nous pouvons dire que la méthode de la pente est insensible à l'effet collectif du voisinage. La direction de déplacement d'un sous-ensemble d'observations est définie par le lien établit entre la cellule nœud de ce sous-ensemble et sa cellule père. Si la distribution autour de ces deux cellules change sans changer la cellule père, la direction du déplacement ne sera point changée. Toutefois, si cet effet collectif devait-être pris en considération par le calcul des pentes et de leur moyenne, le calcul deviendrait laborieux pour un espace à grande dimension.

Dans la méthode de l'amaicissement de l'histogramme par maximisation de la taille des regroupements, l'effet collectif du voisinage est pris en compte. Un changement de la distribution autour des modes, sans toutefois affecter ceux-ci, peut changer la direction de déplacement des observations et leur affectation à leurs classes respectives et surtout celles se trouvant dans les vallées entre ces classes.



Effet de voisinage

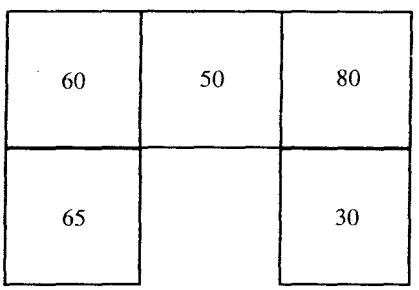
Figure IV.30

IV - 4 - 2 - 3 . Classe bimodale

Dans le cas d'une classe bimodale, la limite entre la classification de toutes les observations soit en une seule et unique classe, soit en deux classes distinctes est très ambigüe. La méthode de l'estimation du gradient donnera toujours deux classes même si la "vallée" entre les deux maximum de la classe bimodale n'est pas très prononcée. Dans ce même cas, la méthode de maximisation de la taille des regroupements nous donnera une seule et unique classe.

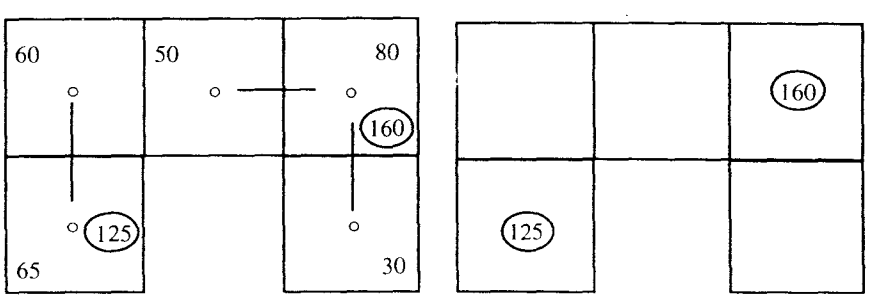
La figure IV.31 illustre un exemple où nous avons une classe bimodale car la "vallée"

entre les 60 et 80 observations n'est pas relativement prononcée (50 observations). La figure IV.32 nous montre que la méthode de l'estimation du gradient donne deux classes. La figure IV.33 montre que la méthode de maximisation des tailles de regroupement donne une seule et unique classe.



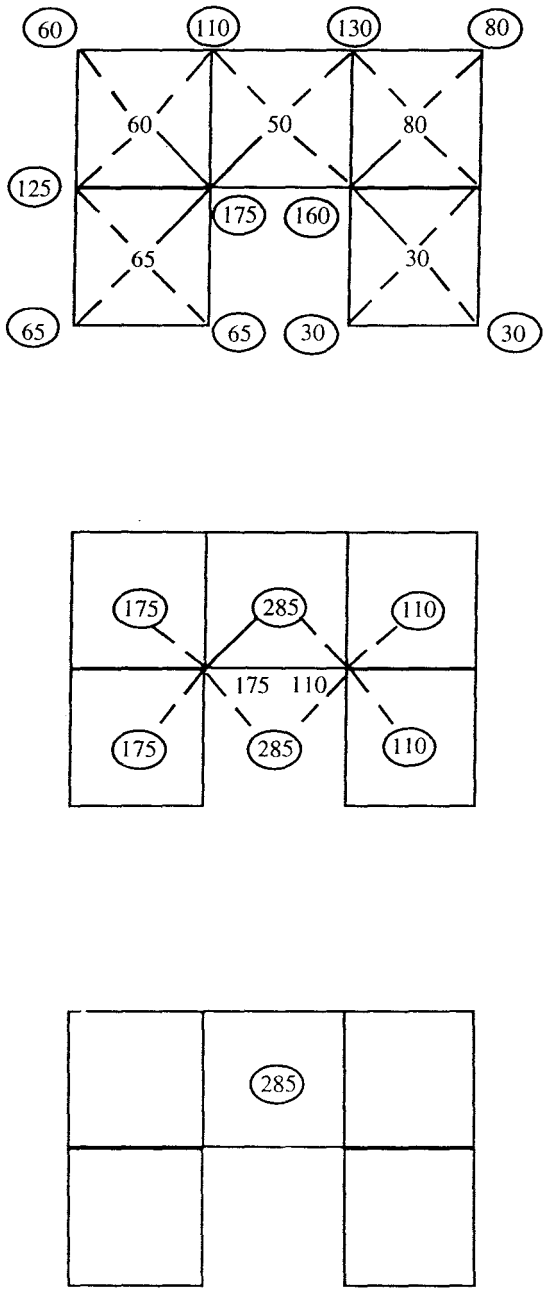
Une classe bimodale

Figure IV.31



Méthode de l'estimation du gradient (2 classes) pour une classe bimodale

Figure IV.32



Méthode de maximisation de la taille des regroupements

Figure IV.33

IV - 5 . CONCLUSION

Dans ce chapitre, nous avons présenté une méthode de classification qui est basée sur l'analyse de l'histogramme multidimensionnel. C'est une méthode totalement automatique qui ne nécessite aucune connaissance a priori sur les distributions analysées, que ce soit par une formulation mathématique de type paramétrique par exemple, ou encore une indication sur le nombre de classes en présence. Nous avons montré qu'en déplaçant les observations dans la direction des regroupements fictifs qui maximisent leur concentration spatiale, ces observations sont attirées vers les modes respectifs de la distribution. Cette migration des observations réduit la dispersion des classes, ce qui a pour effet d'élargir les vallées qui les séparent et d'accentuer les différences d'amplitudes entre les sommets et les creux de ces vallées. Cette procédure itérative est applicable à des problèmes de dimension élevée, car elle se base sur la technique de balayage spatial de l'espace discrétisé présentée au chapitre précédent. La rapidité d'éxécution de chaque itération de l'algorithme augmente au fur et à mesure que nous approchons du résultat final, car la procédure, qui est basée sur le regroupement des observations, ne prend en compte que les points de discrétisation de l'espace où se concentrent les observations.

La procédure de classification prend fin lorsque deux itérations consécutives donnent le même résultat. Chaque mode est alors défini par un seul pic dont l'amplitude indique le nombre d'observations assignées à la classe correspondante. En suivant la position de chaque observation au cours des différentes migrations et des différents regroupements éventuels et réels, on dispose finalement de la composition exhaustive de chaque classe.

Dans le chapitre V, nous utiliserons cette nouvelle approche pour analyser différents ensembles de données afin de montrer ses performances et de cerner ses avantages et ses limites.

Annexe IV-1

Réduction de la variance par déplacement des observations vers le mode

Soit un nombre Q d'observations distribuées symétriquement autour de leur mode et donc de leur moyenne. La variance de cette distribution est donnée par :

$$\text{Variance } [X] = E(X^2) - E(X)^2 \qquad \text{A-IV-1-1}$$

où :

$$E(x) = \frac{a_1 x_1 + a_2 x_2 + a_3 x_3 + ... + a_{\bar{x}}\bar{x} + ... + a_{L-2} x_{L-2} + a_{L-1} x_{L-1} + a_L x_L}{(a_1 + a_2 + a_3 + ... + a_{\bar{x}} + ... + a_{l-2} + a_{l-1} + a_l)}$$

A-IV-1-2

a_1 , a_2 , a_3 , ... , a_L étant les effectifs dans chaque intervalles de l'histogramme de largeur δ et de valeur moyenne d'intervalle x_1 , x_2 , x_3 , ... , x_L (Cf. figure A-IV-1).

Les observations étant distribuées symétriquement autour de leur mode et moyenne, $a_1 = a_L$, $a_2 = a_{L-1}$, $a_3 = a_{L-2}$ .... donc

$$E(x) = \frac{a_1 x_1 + a_2 x_2 + a_3 x_3 + ... + a_{\bar{x}}\bar{x} + ... + a_3 x_{L-2} + a_2 x_{L-1} + a_1 x_L}{(2a_1 + 2a_2 + 2a_3 + ... + a_{\bar{x}})}$$

A-IV-1-3

et donc

$$E(x^2) = \frac{a_1 x^2_1 + a_2 x^2_2 + a_3 x^2_3 + ...+ a_{\bar{x}}\bar{x}^2 + ...+ a_3 x^2_{L-2} + a_2 x^2_{L-1} + a_1 x^2_L}{(2a_1 + 2a_2 + 2a_3 + ... + a_{\bar{x}})}$$

A-IV-1-4

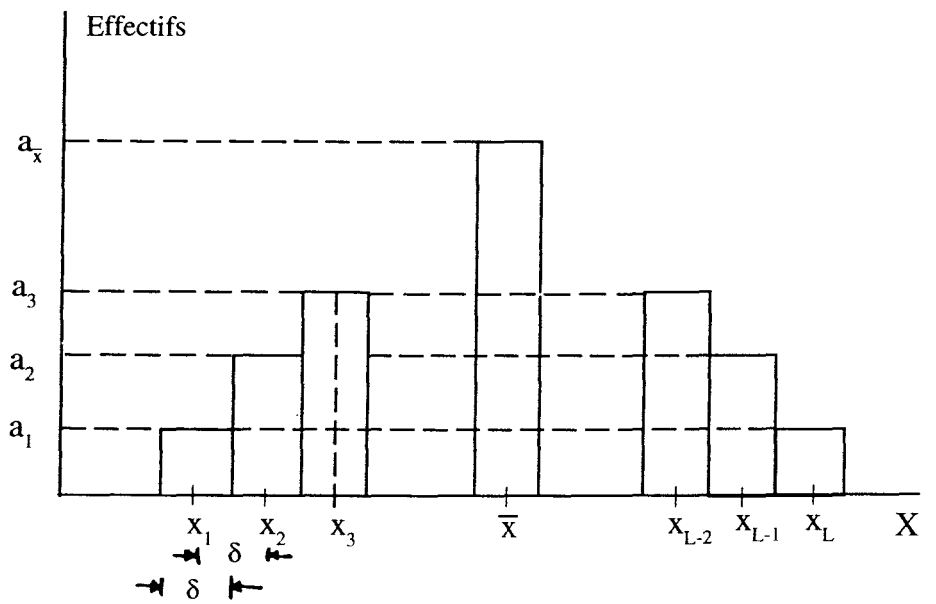


Figure A-IV-1

Déplaçant chaque observation vers le mode (ou la moyenne) de la distribution sans toutefois déplacer celui-ci, nous aurons une transformation de $Y = X \pm \delta$ selon que l'observation est située à droite (signe négatif) ou à gauche (signe positif) de cette moyenne.

La variance sera donnée par :

$$\text{Variance}[\,Y\,] = E(Y^2) - E(Y)^2 \qquad \text{A-IV-1-5}$$

Le calcul de E(Y) est donné par :

$$E(Y) = \frac{a_1(x_1+\delta) + a_2(x_2+\delta) + a_3(x_3+\delta) + \ldots + a_{\bar{x}}\bar{x} + \ldots + a_3(x_{L-2}-\delta) + a_2(x_{L-1}-\delta) + a_1(x_L-\delta)}{(2a_1 + 2a_2 + 2a_3 + \ldots + a_{\bar{x}})}$$

$$E(Y) = \frac{a_1x_1 + a_2x_2 + a_3x_3 + ... + a_{\bar{x}}\bar{x} + ... + a_3x_{L-2} + a_2x_{L-1} + a_1x_L}{(2a_1 + 2a_2 + 2a_3 + ... + a_{\bar{x}})} +$$

$$\frac{a_1\delta + a_2\delta + a_3\delta + ... - a_3\delta - a_2\delta - a_1\delta}{(2a_1 + 2a_2 + 2a_3 + ... + a_{\bar{x}})} \qquad \text{A-IV-1-6}$$

Des équations A-IV-1-6 et A-IV-1-3, nous déduisons

$$E(X) = E(Y) \qquad \text{A-IV-1-7}$$

Ce qui veut dire que ce déplacement de point vers le mode ou la moyenne ne change pas l'emplacement de celui-ci.

Calculons $E(Y^2)$, nous avons :

$$E(Y^2) = \frac{a_1(x_1+\delta)^2 + a_2(x_2+\delta)^2 + a_3(x_3+\delta)^2 + ... + a_{\bar{x}}x^2 + ... + a_3(x_{L-2}-\delta x)^2 + a_2(x_{L-1}-\delta)^2 + a_1(x_L-\delta)^2}{(2a_1 + 2a_2 + 2a_3 + ... + a_{\bar{x}})}$$

donc

$$E(Y^2) = \frac{a_1x_1^2 + a_2x_2^2 + a_3x_3^2 + ... + a_{\bar{x}}\bar{x}^2 + ... + a_3x^2_{L-2} + a_2x^2_{L-1} + a_1x^2_L}{(2a_1 + 2a_2 + 2a_3 + ... + a_{\bar{x}})}$$

$$+ \frac{a_1\delta^2 + a_2\delta^2 + a_3\delta^2 + ... + ... + a_3\delta^2 + a_2\delta^2 + a_1\delta^2}{(2a_1 + 2a_2 + 2a_3 + ... + a_{\bar{x}})}$$

$$+ \frac{2a_1\delta x_1 + 2a_2\delta x_2 + 2a_3\delta x_3 + ... - 2a_3\delta x_{L-2} - 2a_2\delta x_{L-1} - 2a_1\delta x_L}{(2a_1 + 2a_2 + 2a_3 + ... + a_{\bar{x}})}$$

A-IV-1-8

De A-IV-1-4 en A-IV-1-8, nous déduisons :

$$E(Y^2) = E(X^2) + \frac{(2a_1 + 2a_2 + 2a_3 + ...)}{Q}\ \delta^2$$

$$- \frac{1}{Q}\ [\ 2a_1\delta(x_L - x_1) + 2a_2\ \delta(x_{L-1} - x_2) + 2a_3\delta(x_{L-2} - x_3) + \ ...\ ]$$

A-IV-1-9

La plage de variation de la variable x étant divisé en un nombre d'intervalles D de largeur δ, nous aurons :

$$\begin{aligned} X_L - X_1 &= (D-1)\,\delta \\ X_{L-2} - X_2 &= (D-3)\,\delta \\ X_{\bar{x}+\delta} - X_{\bar{x}-\delta} &= 2\,\delta \end{aligned} \qquad \text{A-IV-1-10}$$

donc :

$$E(Y^2) = E(X^2) + \frac{(2a_1 + 2a_2 + 2a_3 + \ldots.)}{Q}\ \delta^2$$

$$- \frac{1}{Q}\ [\,2a_1(D-1)\,\delta^2 + 2a_2\,(D-3)\,\delta^2 + 2a_3(D-5)\,\delta^2 + \ldots.\,]$$

donc :

$$E(Y^2) = E(X^2) - \frac{\delta^2}{Q}\Big\{\ [\,2a_1(D-1) + 2a_2\,(D-3) + 2a_3(D-5) + \ldots.\,]$$

$$-\,(\,2a_1 + 2a_2 + 2a_3 + \ldots.\,)\ \Big\} \qquad \text{A-IV-1-11}$$

On constate que puisque les termes (D-1), (D-3), (D-5), ... sont tous plus grand que l'unité le 2ème terme de l'équation A-IV-1-11 est toujours négatif donc :

$$E(Y^2) < E(X^2).$$

De là, en comparant les équations A-IV-1-1 et A-IV-1-5 puisque $E(Y)^2 = E(X)^2$ et $E(Y^2) < E(X^2)$, la variance [Y] < variance [X]. Donc déplaçant les points vers le mode ou la moyenne réduit la variance.

CHAPITRE V

RESULTATS EXPERIMENTAUX

CHAPITRE V

RESULTATS EXPERIMENTAUX

V - 1 . INTRODUCTION

Afin d'évaluer la procédure de classification par maximisation des regroupements des observations proposée dans le chapitre précédent, nous présentons, dans ce chapitre, plusieurs exemples bidimensionnels et multidimensionnels de classification de données générées artificiellement. Ces exemples sont variés : on trouvera des classes à faibles effectifs où les observations sont bien séparées, des classes à grands effectifs où les observations sont rapprochées, des classes non sphériques et non équiprobables, ainsi que des classes en forme de croissant.

Une étude de l'effet de la finesse de la discrétisation sur le nombre de modes détectés est aussi présentée.

V - 2 . EXEMPLE BIDIMENSIONNEL COMPOSE DE TROIS CLASSES NORMALES NON SPHERIQUES

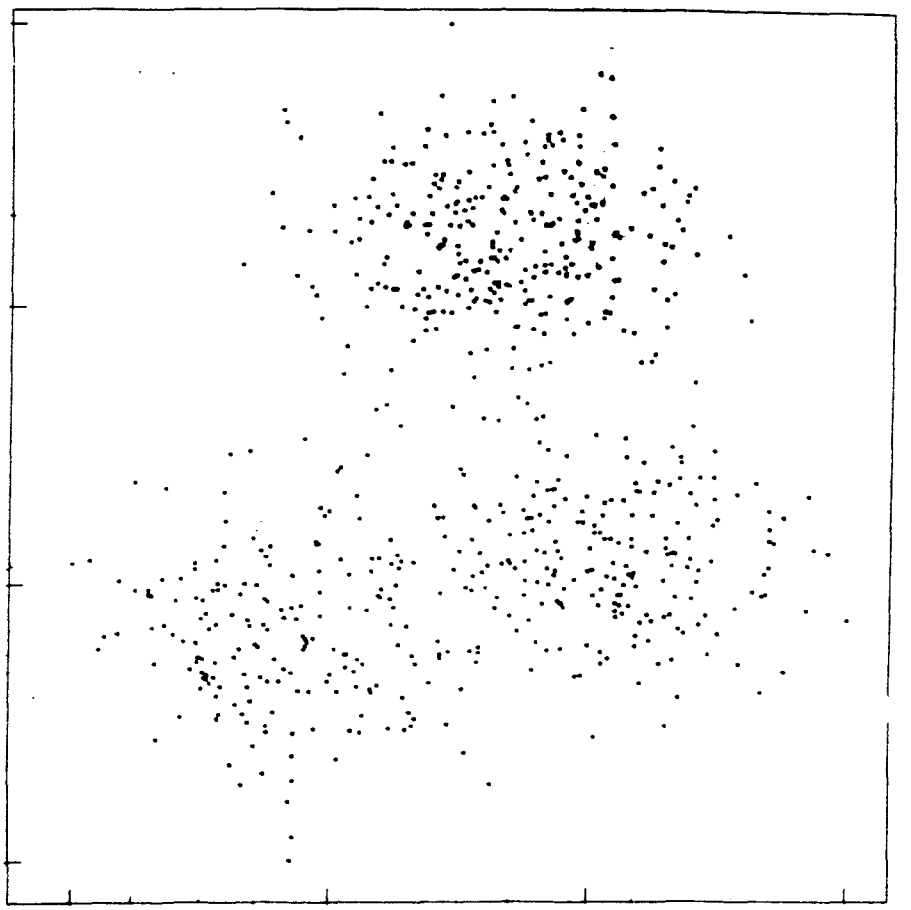
Pour le premier exemple, nous analyserons toutes les étapes de la classification jusqu'à la stabilisation du résultat.

Les données utilisées dans cet exemple sont des observations provenant de trois classes normales non sphériques et non équilibrables dont les paramètres statistiques, ainsi que le nombre d'observations par classe, sont précisés dans le tableau V.1. La figure V-1 présente l'échantillon tel qu'il a été simulé pour construire cet exemple.

Distribution	Nombre d'observations	Vecteur Moyenne	Matrice de covariance	Probabilité a priori
1	200	$\bar{X}_1 = \begin{bmatrix} 1,1215 \\ 0,954 \end{bmatrix}$	$S_1 = \begin{bmatrix} 1,896 & -0,042 \\ -0,042 & 2,173 \end{bmatrix}$	$Pb_1 = {}^2/_7$
2	200	$\bar{X}_2 = \begin{bmatrix} 5,8853 \\ 2,1096 \end{bmatrix}$	$S_2 = \begin{bmatrix} 1,589 & -0,133 \\ -0,133 & 1,984 \end{bmatrix}$	$Pb_2 = {}^2/_7$
3	300	$\bar{X}_3 = \begin{bmatrix} 3,8832 \\ 7,9213 \end{bmatrix}$	$S_3 = \begin{bmatrix} 2,137 & 0,104 \\ 0,104 & 1,718 \end{bmatrix}$	$Pb_3 = {}^3/_7$

Paramètres statistiques des classes de l'exemple 1

Tableau V - 1



Représentation graphique de l'échantillon de l'exemple 1

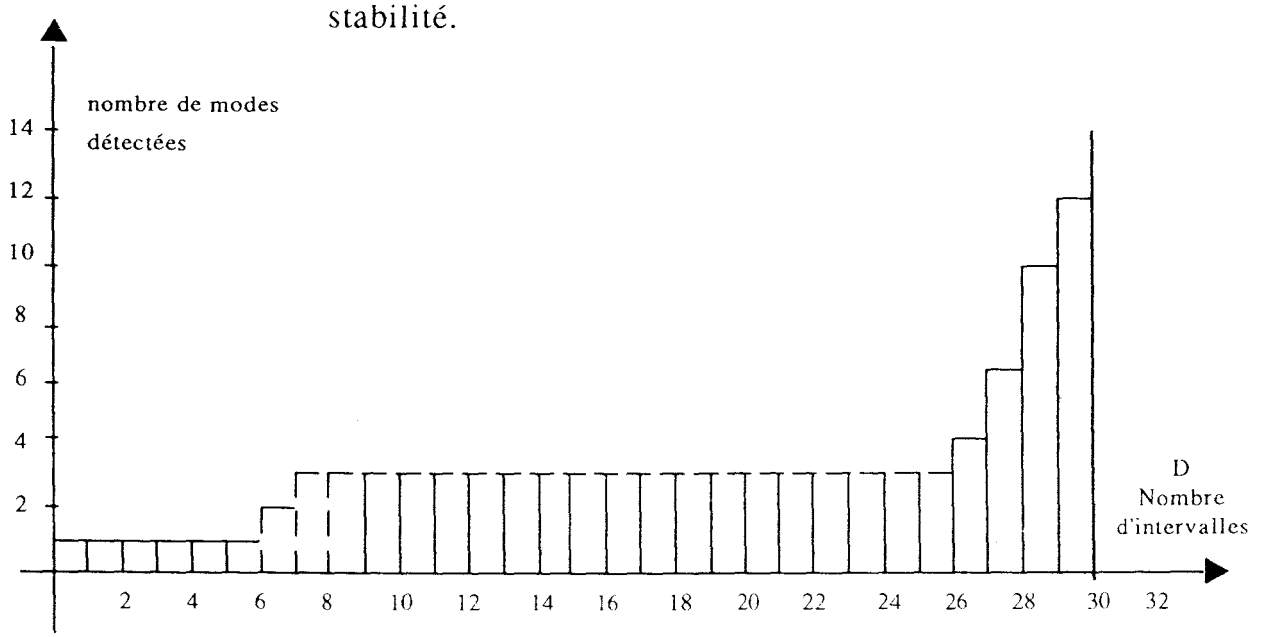
Figure V.1

Pour déterminer avec le maximum de fiabilité, le nombre de classes en présence, l'algorithme est lancé plusieurs fois pour un nombre différent d'intervalles de discrétisation D sur chaque axe. Cette procédure permet de retenir le nombre d'intervalles conformément au concept de stabilité du nombre de modes détectés [Fromm et Northouse, 1973 ; Eigen, Fromm et Northouse, 1974 ; Touzani et Postaire, 1988].

La figure V.2 indique les variations du nombre de modes mis en évidence à la fin de la procédure de classification itérative en fonction du nombre d'intervalles. On constate que :

(i) l'algorithme est peu sensible à l'ajustement de ce paramètre puisque le nombre de modes détecté reste stable et égal à trois pour $7 \leq D \leq 25$,

(ii) un éclatement des modes intervient brutalement dès que le nombre d'intervalles D dépasse la valeur 25, limite supérieure de la plage de stabilité.



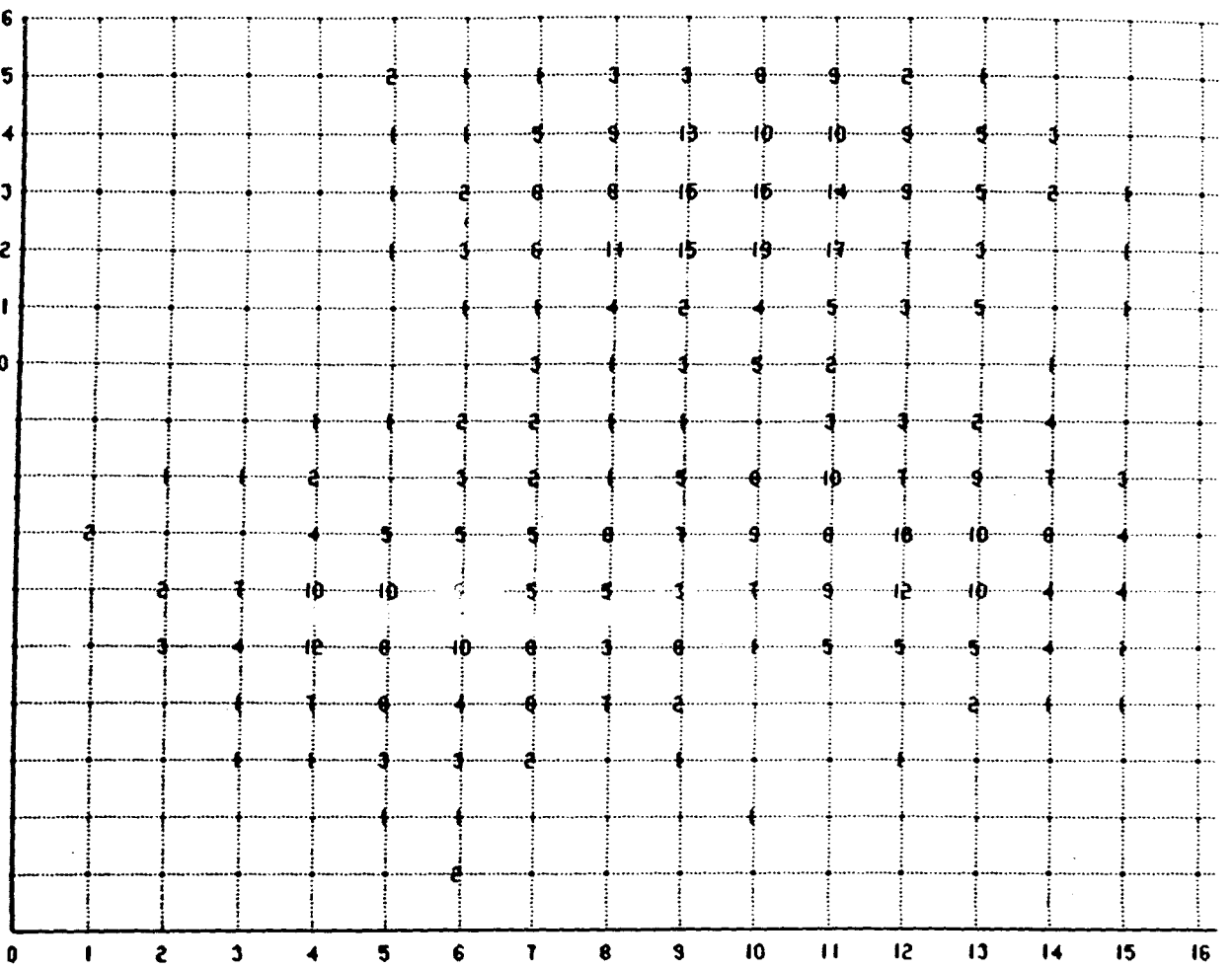
Nombre de modes détectés dans l'échantillon du tableau V-1 en fonction du nombre d'intervalles D

Figure V.2

Ayant déterminé cette plage de stabilité, l'algorithme calcule le nombre de classes en présence, assigne les observations à leurs classes respectives et détermine les paramètres statistiques pour un nombre d'intervalles égal à la valeur médiane de cette plage, c'est-à-dire D=16 pour cet exemple. Les étapes de cette procédure finale de classification automatique sont illustrées par les figures de V.3 à V.7.

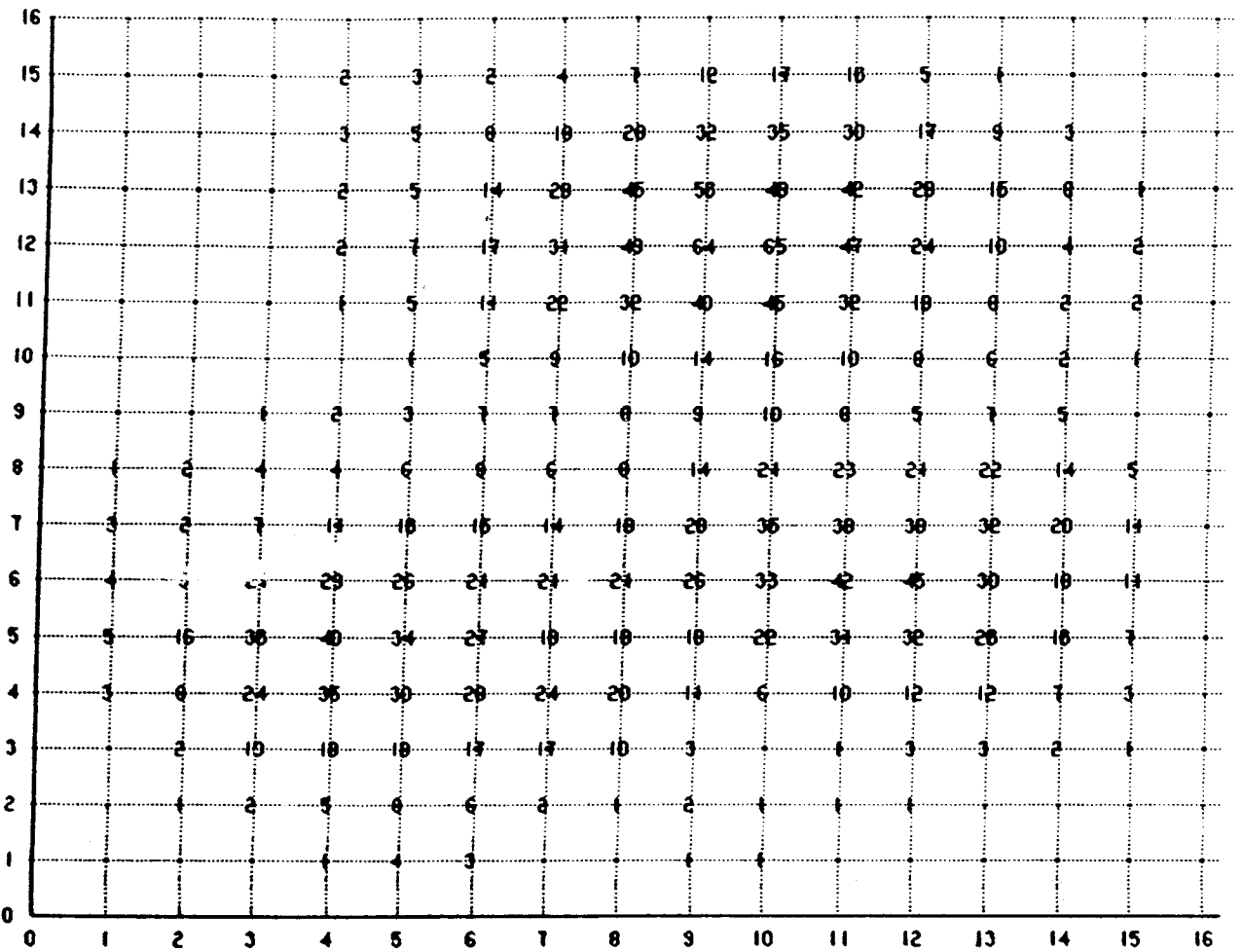
Conformément à la méthode de classifcation automatique par maximisation des regroupements présentée au chapitre IV, la figure V.3 présente l'histogramme multidimensionnel bâti sur la grille des centres et obtenu à partir de l'échantillon tel qu'il a été simulé. La figure V.4 représente le premier regroupement fictif bâti sur la grille des sommets. La figure V.5 représente le premier histogramme multidimensionnel réel bâti sur cette même grille. La figure V.6 illustre les étapes suivantes de cette classification automatique, tandis que la figure V.7 en représente l'étape finale.

Il est à noter que lors de la procédure de classification, un même indice est affecté aux observations formant le même sous-ensemble dans les regroupements successifs. Ainsi, à la fin de la procédure de classification, l'analyste dispose de la liste exhaustive de toutes les observations constituant chacune des classes, ainsi que des paramètres statistiques classiques de position et de dispersion des classes



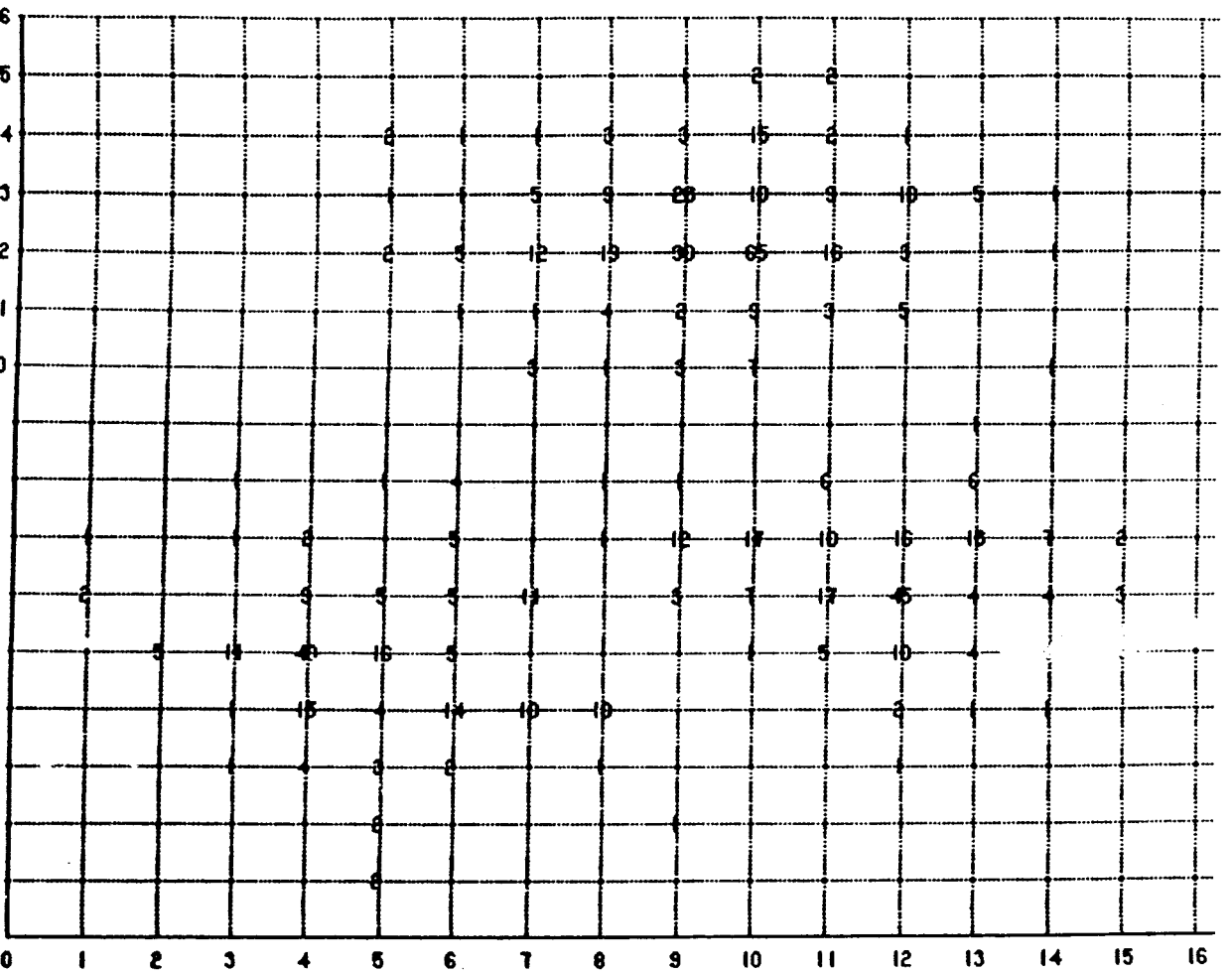
Histogramme multidimensionnel bâti sur la grille des centres

Figure V.3



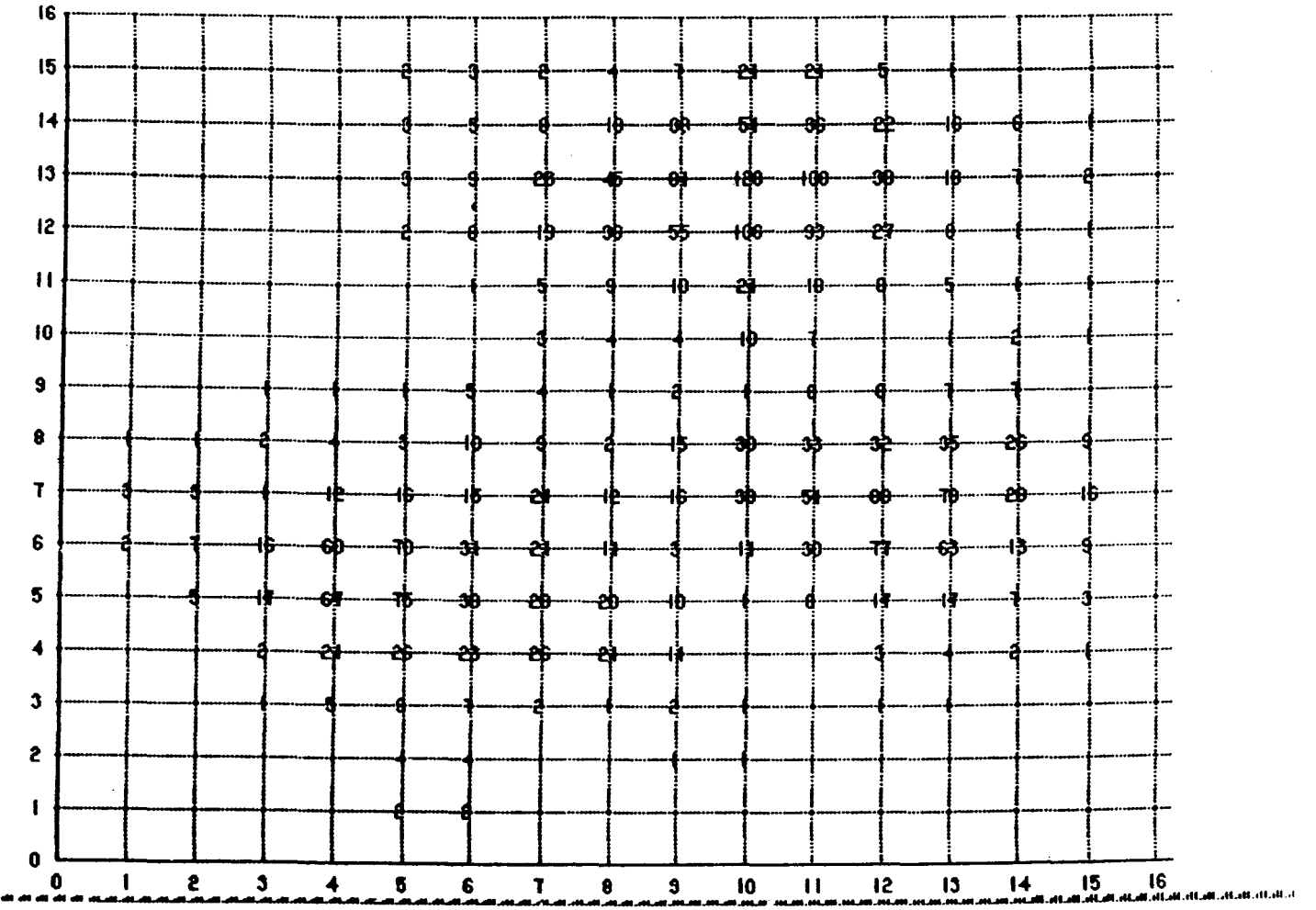
Premier histogramme multidimensionnel fictif bâti sur la grille des sommets

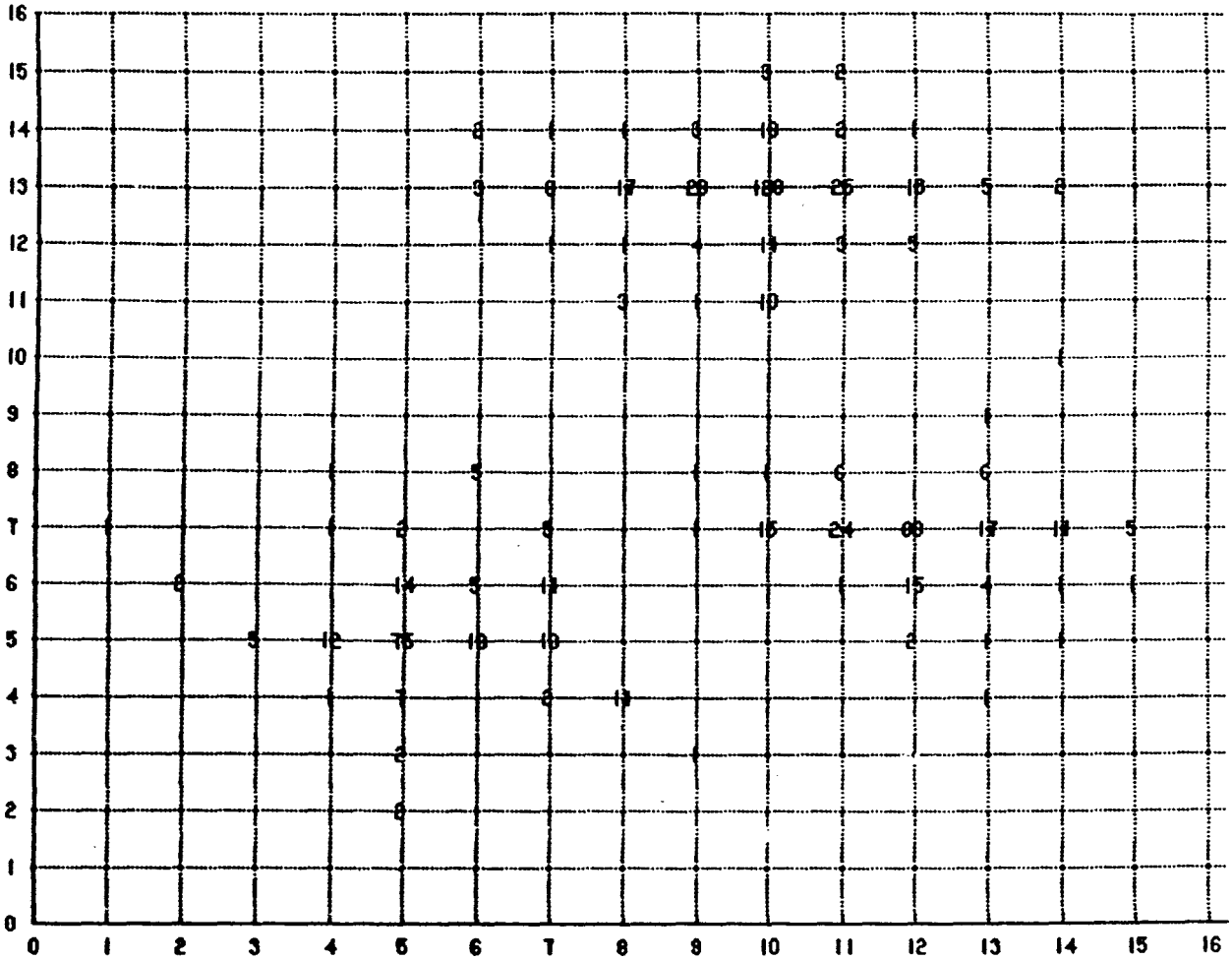
Figure V.4

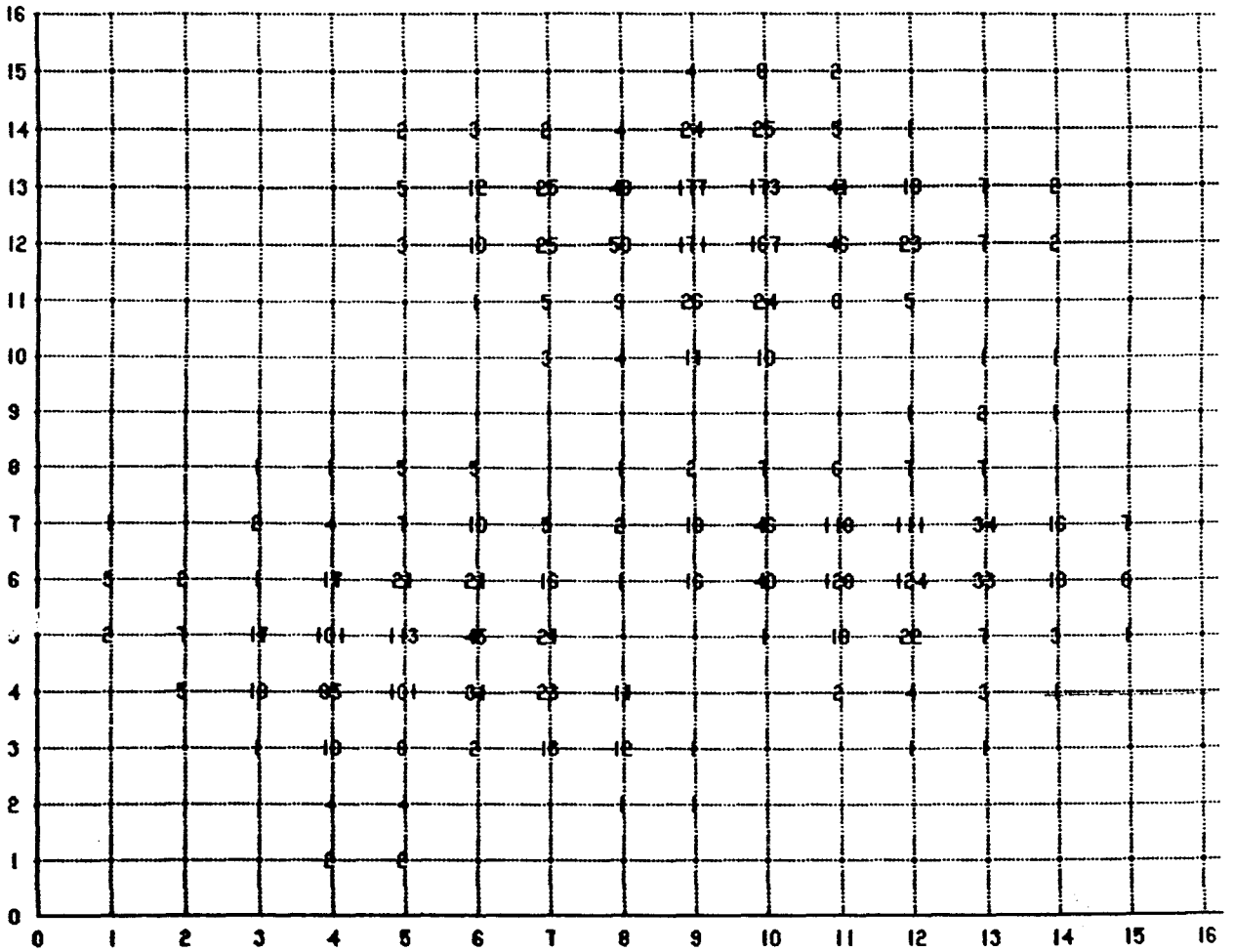


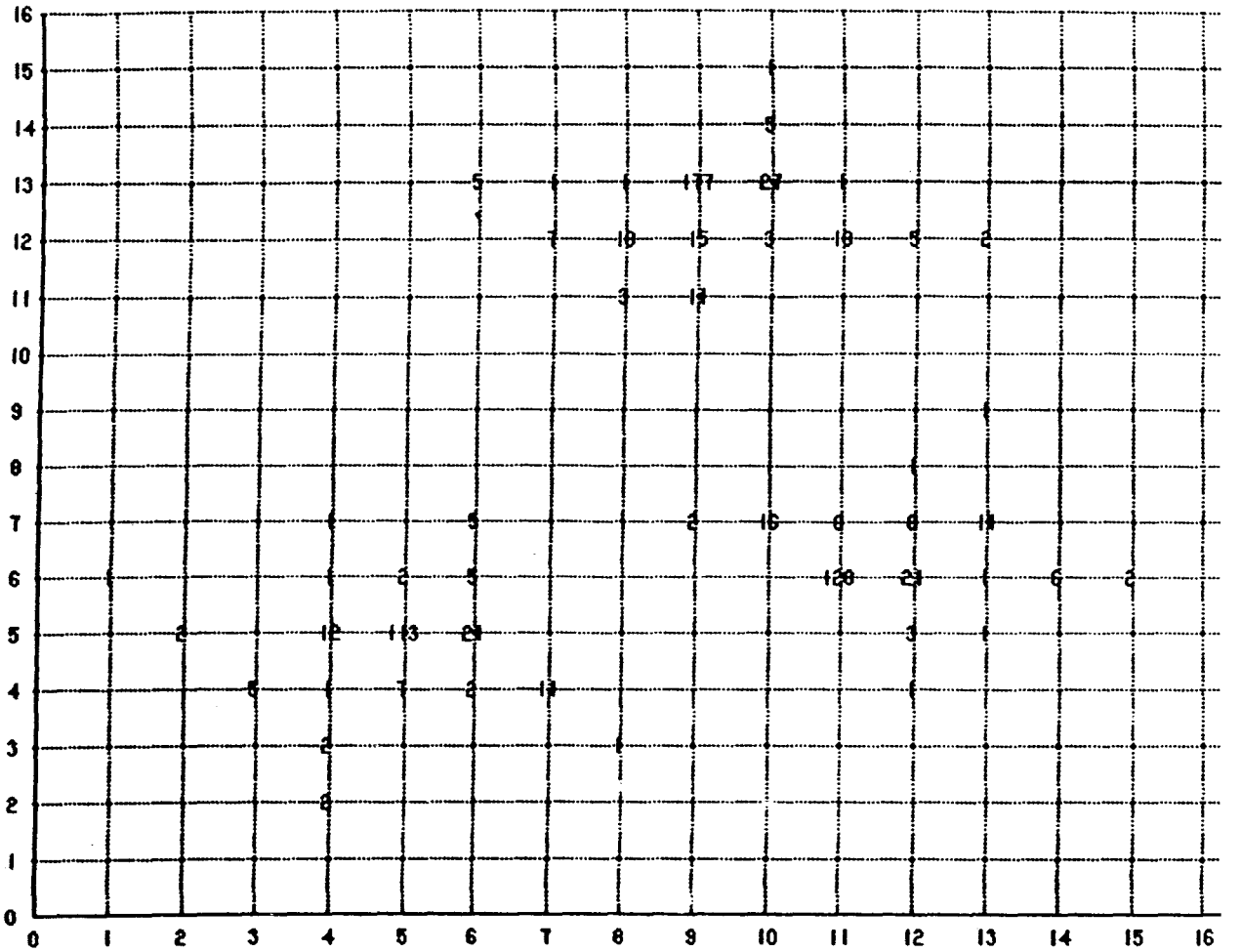
Premier histogramme multidimensionnel réel bâti sur la grille des sommets

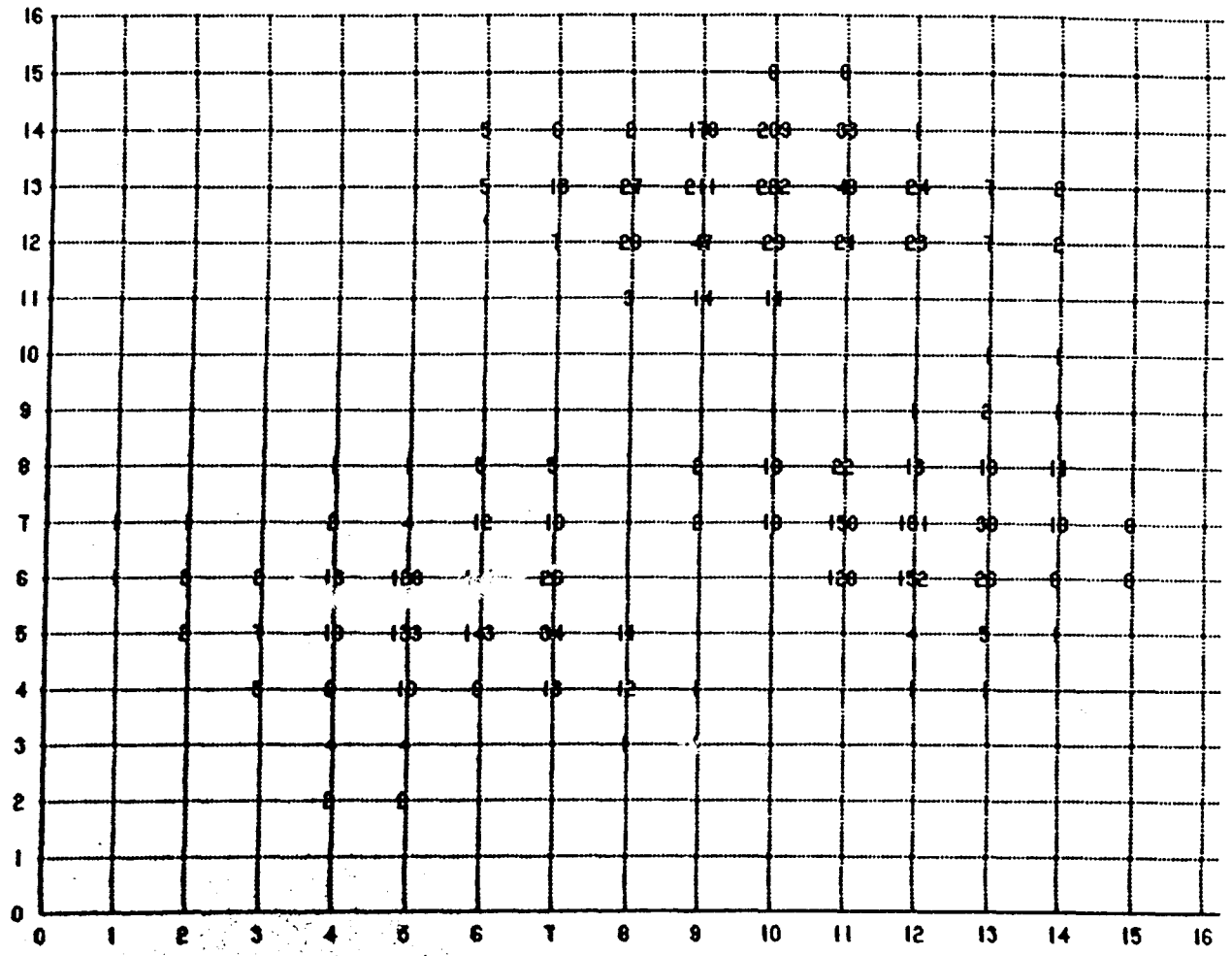
Figure V.5

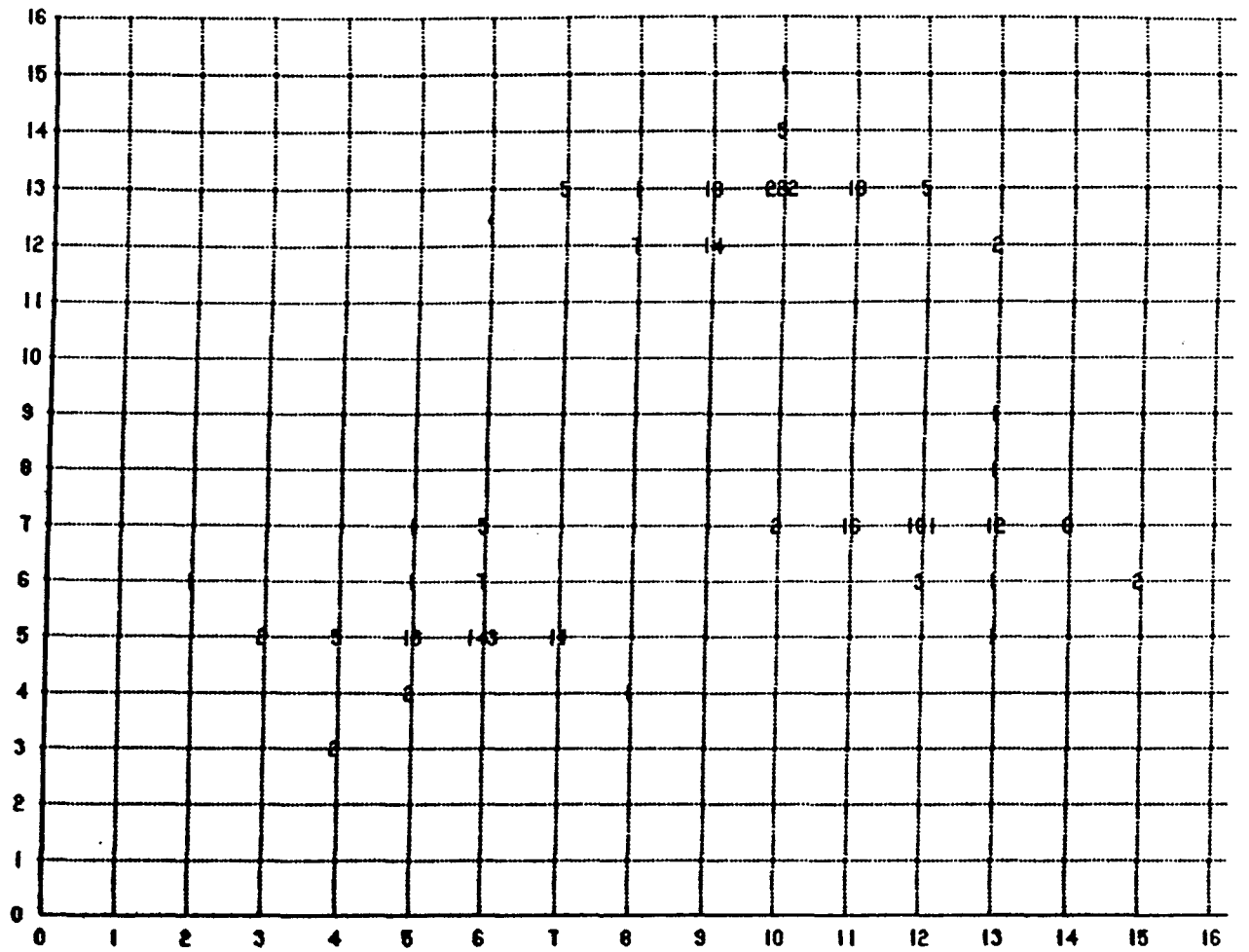


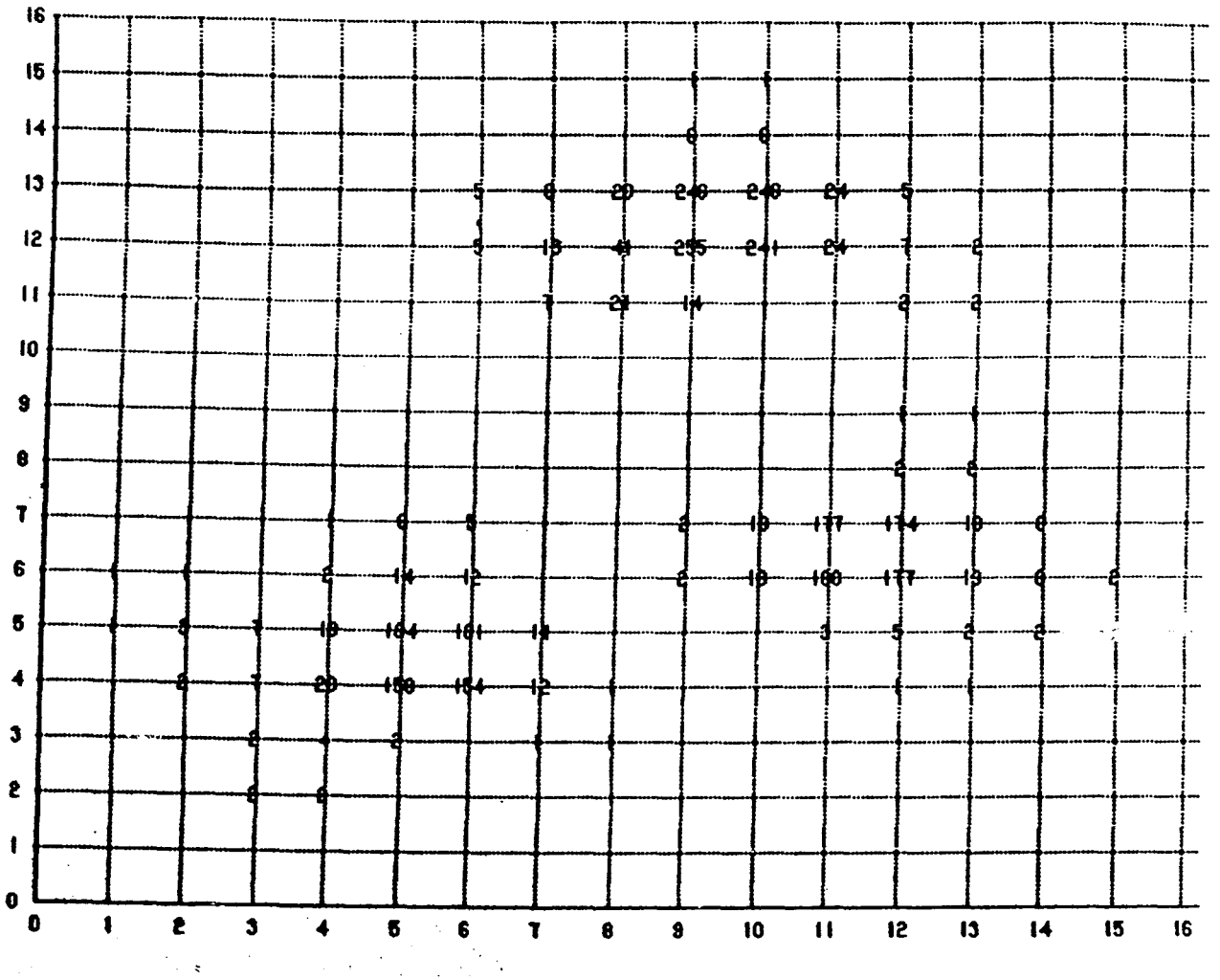


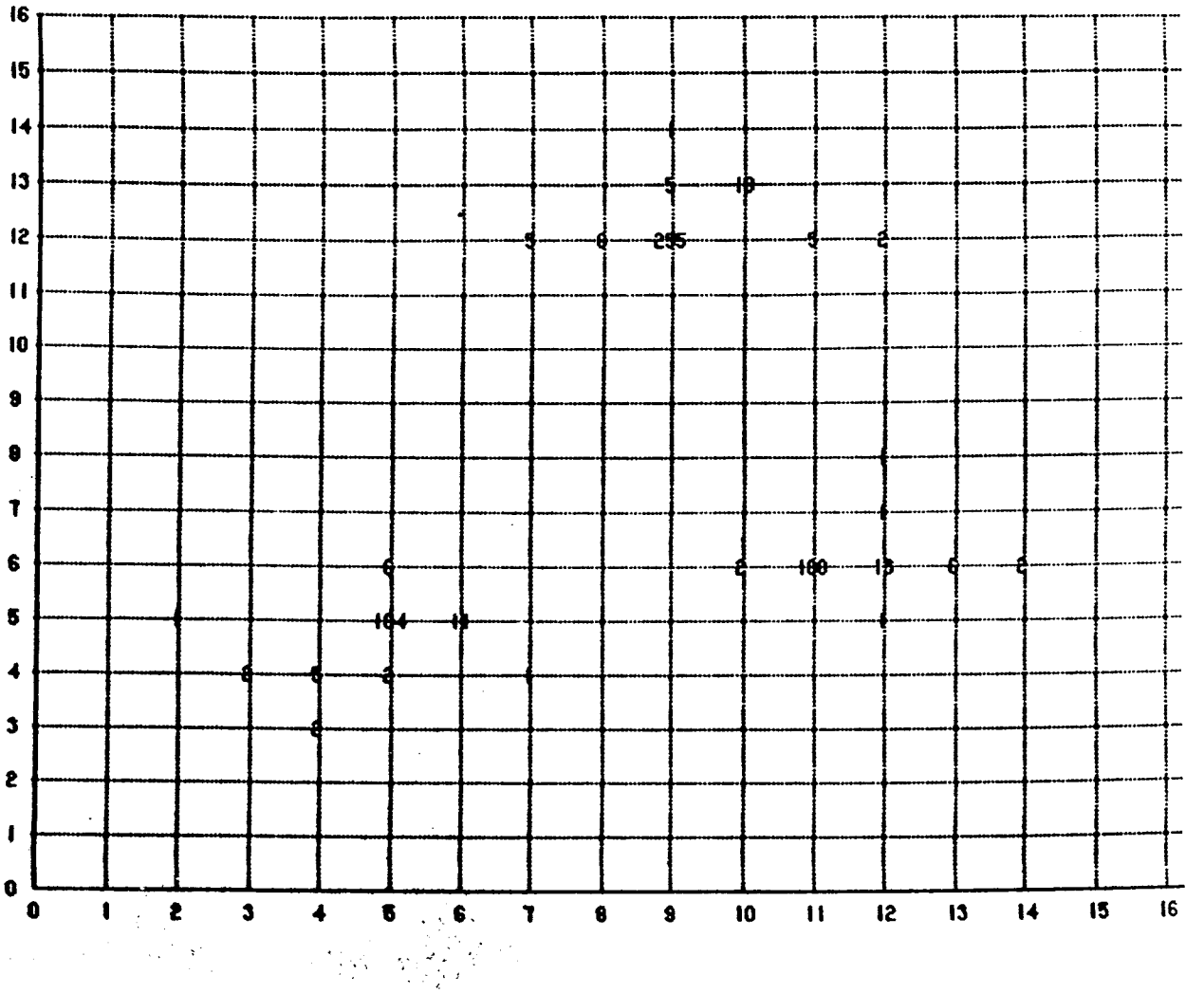


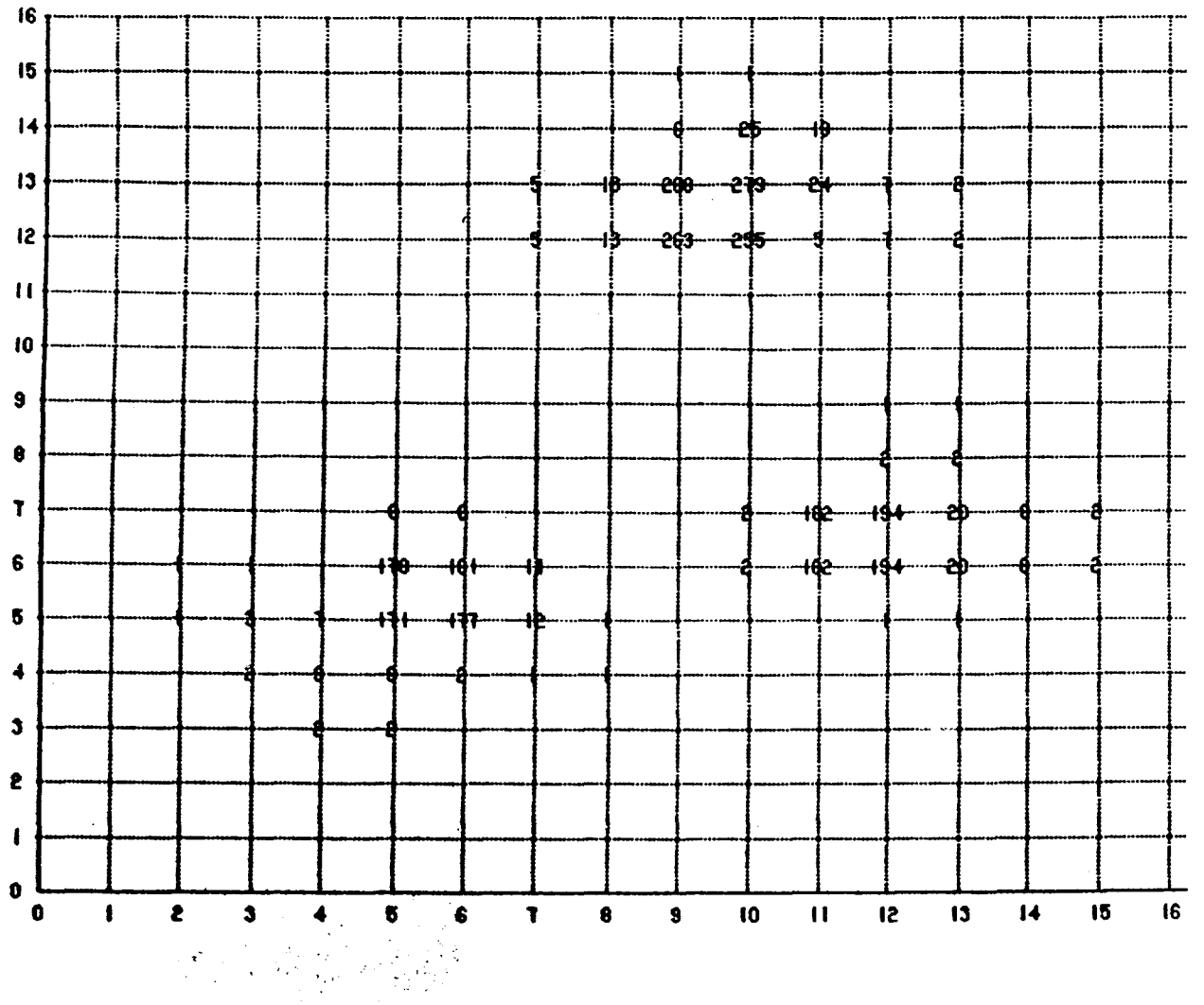


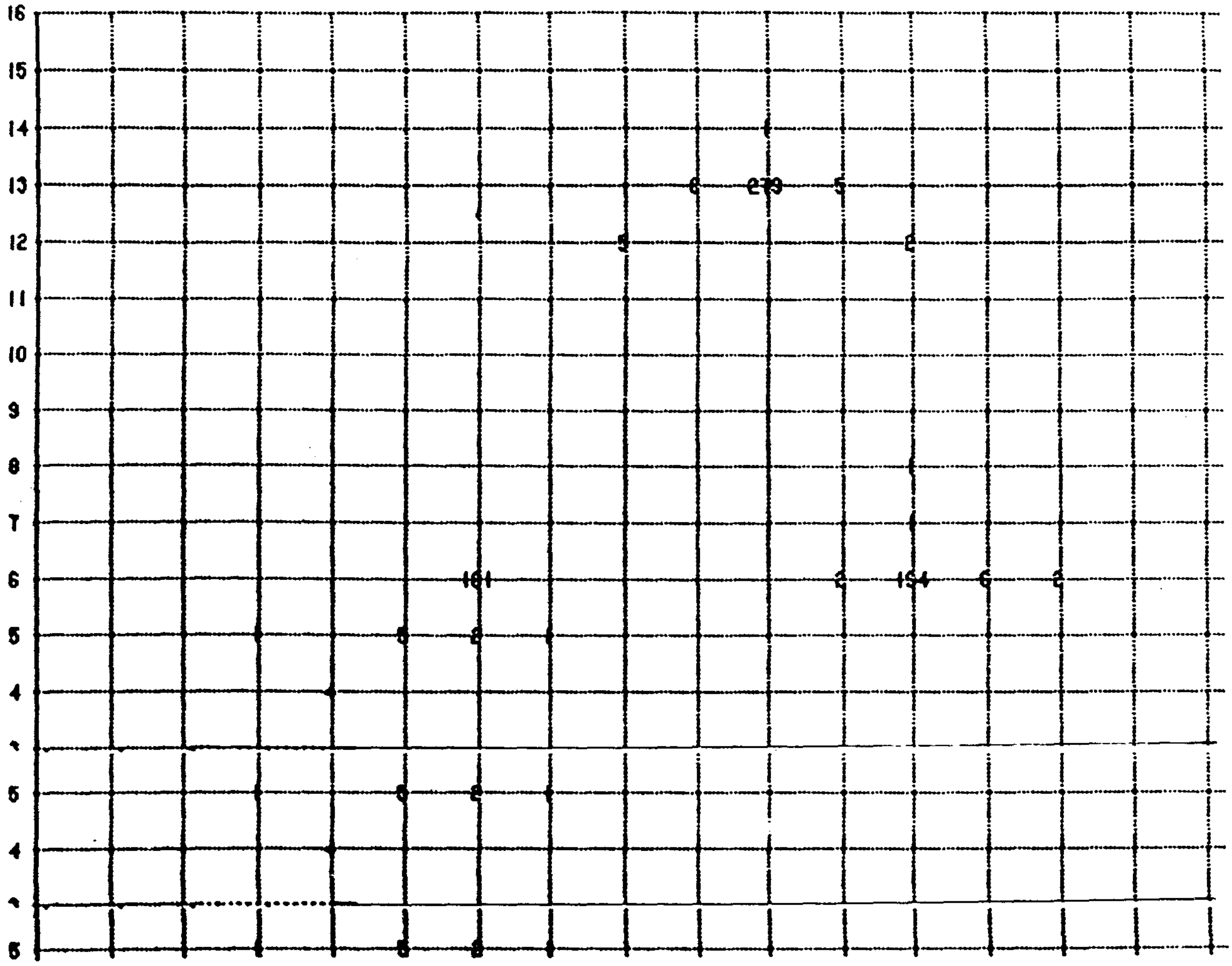


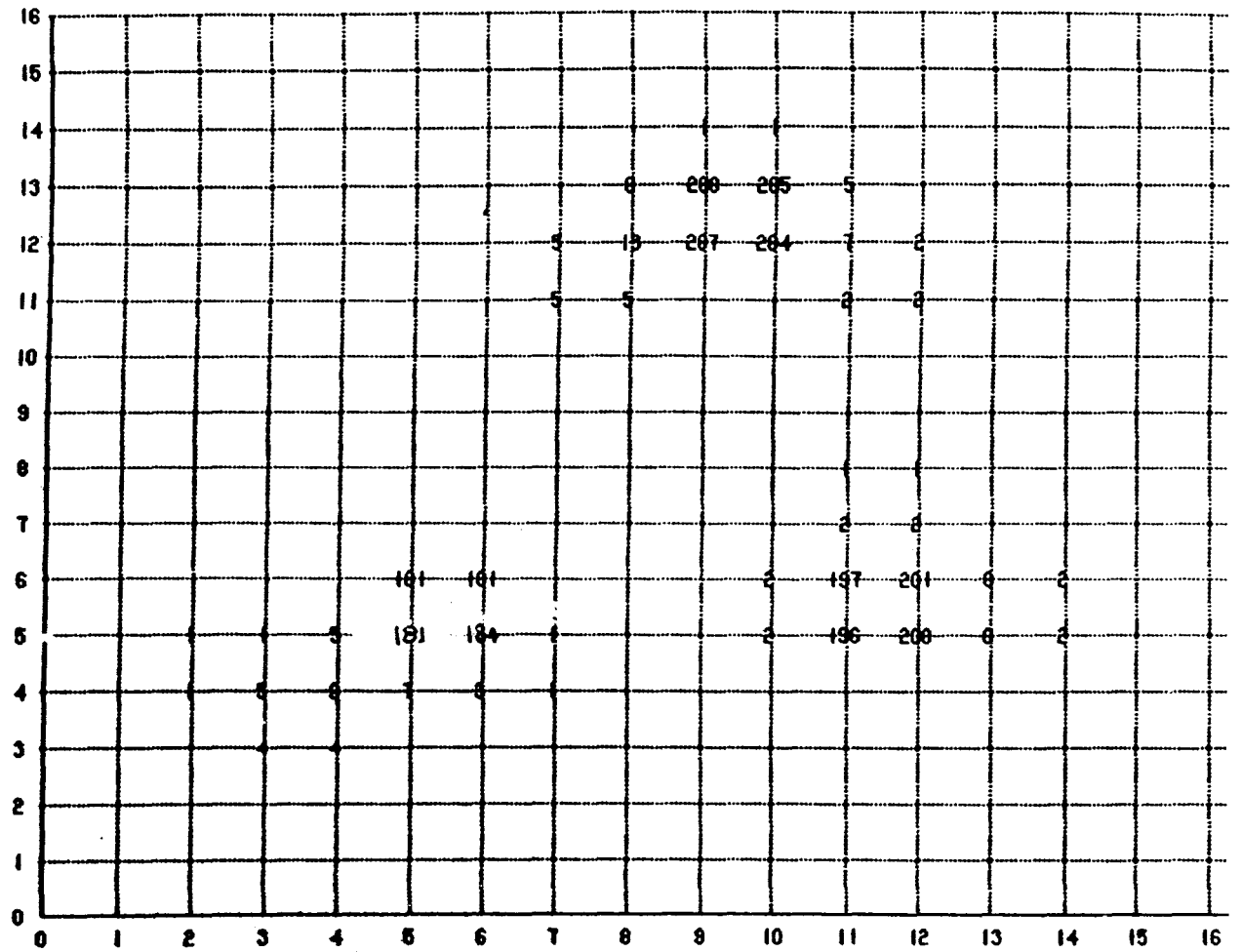


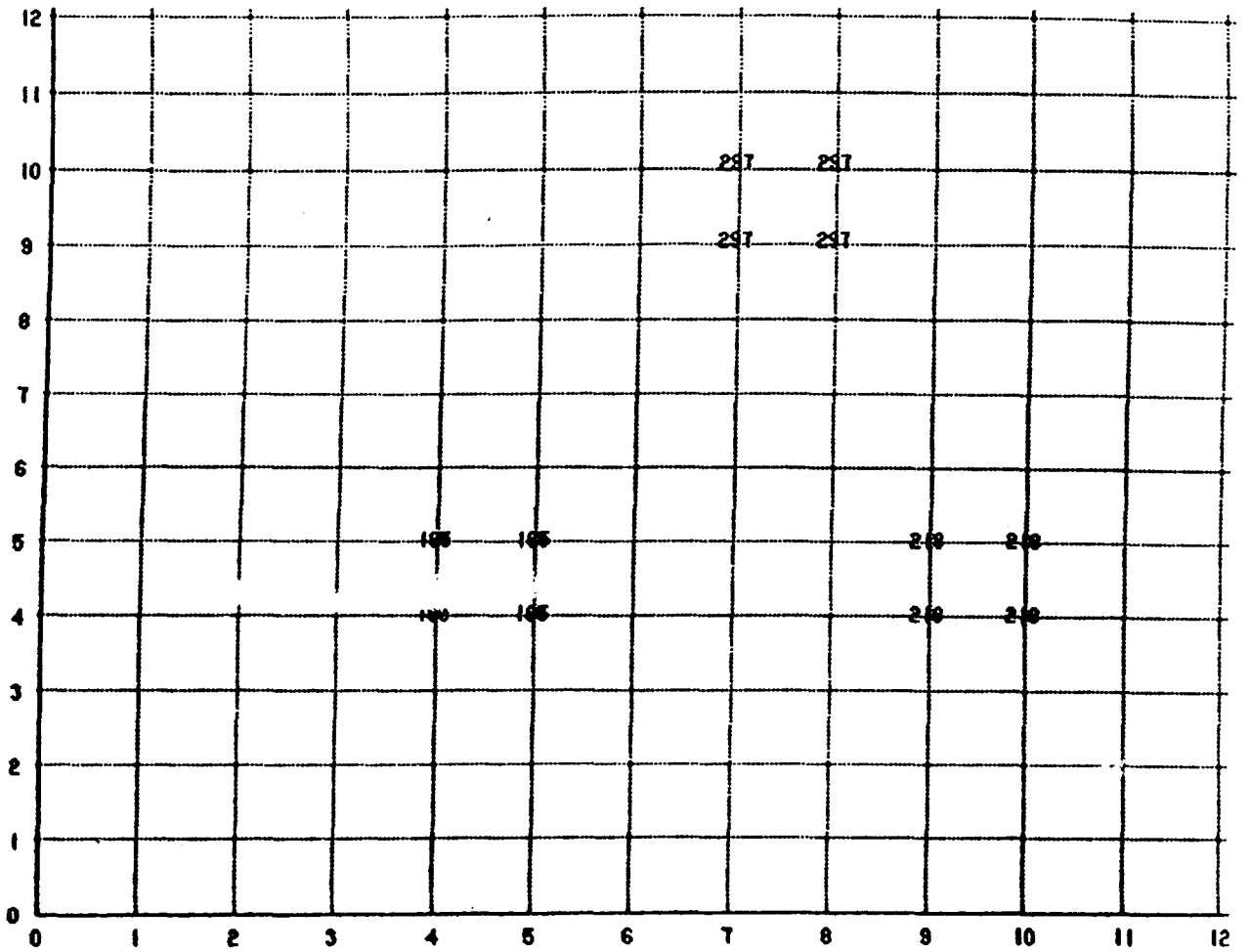



16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16



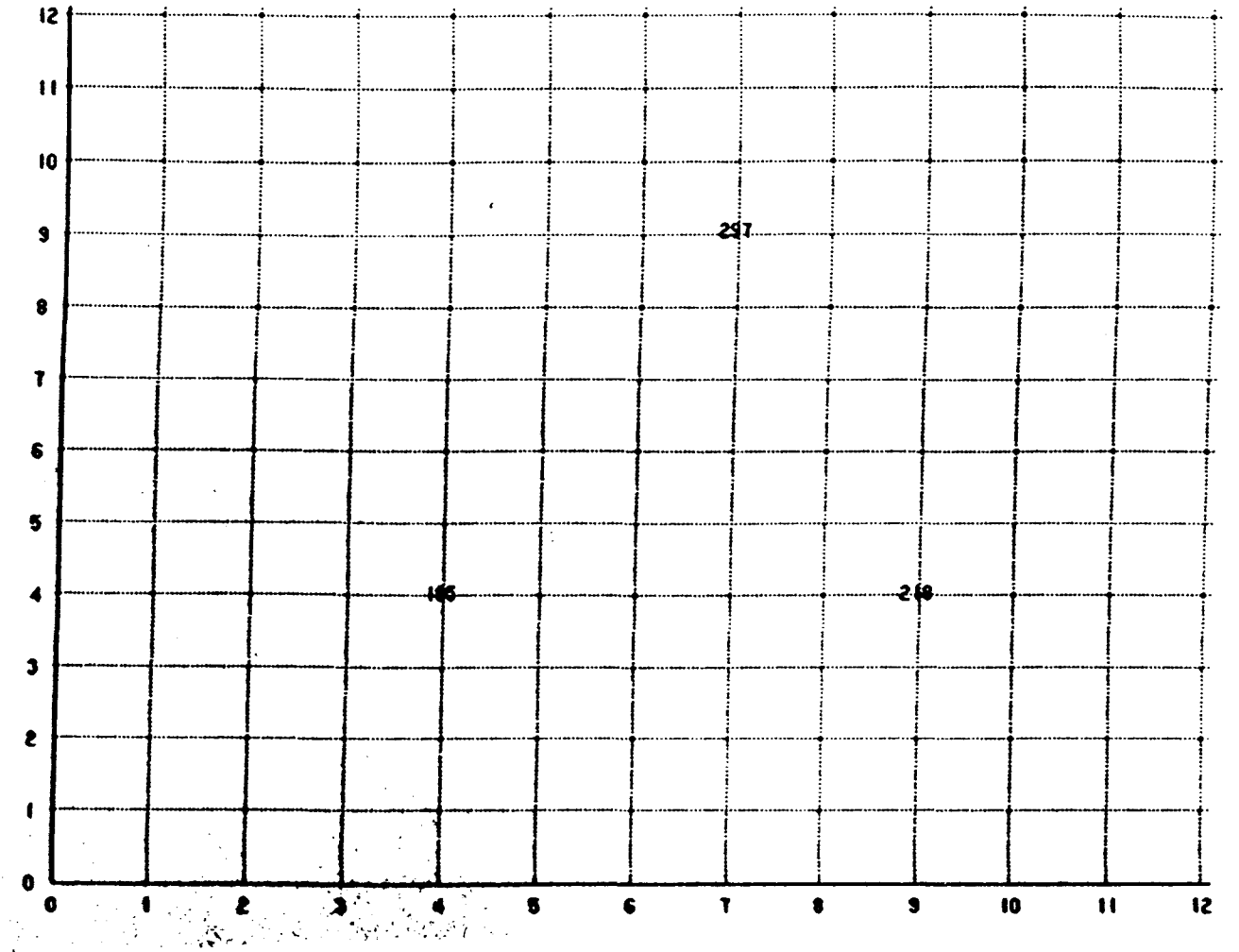



16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
8 200 205 5
5 13 207 204 7 2
5 5 2 8
2 8
161 161 2 197 201 8 2
5 181 184 2 196 200 8 2
6 6 7 6



Etapes itératives de la classification automatique

Figure V.6



Définition du nombre de modes et du nombre d'observations affectées à chacun

Figure V.7

Le temps d'exécution de cette procédure de classification varie avec la pas de discrétisation. Si l'on choisit d'exécuter cette procédure sur une station SUN spark 4 pour un nombre d'intervalles allant de zéro à trente, le temps d'exécution total, y compris la lecture du fichier, la définition du nombre de classes, l'affectation des observations à chaque classe et le calcul des paramètres statistiques, est de 390 secondes, ce qui correspond à une moyenne de 13 secondes par exécution de la procédure pour chaque pas de discrétisation.

Les résultats de la classification de ces 700 observations pour D=16 valeur médiane de la plage de stabilité, donnés par les tableaux V-2 et V-3, illustrent les performances de cette procédure. Le tableau V-2 indique les paramètres statistiques de chaque classe après la classification par la méthode proposée. Les effectifs des trois classes obtenues sont donnés dans la deuxième colonne du tableau V-2. Ces classes sont définies par leurs vecteurs moyennes et leurs matrices de covariance, donnés respectivement par la troisième et quatrième colonnes du même tableau.

Pour analyser les performances de la procédure proposée, nous avons calculé la matrice de confusion associée aux résultats de la classification, ainsi que le taux d'erreur (Cf. tableau V-3).

Chaque élément des 3 premières lignes et colonnes de la matrice de confusion représente le nombre d'observations appartenant à la classe indiquée par le numéro de la ligne de cet élément et affectées, lors de la procédure de classification, à la classe indiquée par la colonne de ce même élément. Par exemple, l'élément de la 1ère ligne, 2ème colonne de la matrice de confusion représente 15 observations appartenant originalement à la classe 1, mais affectées lors de la classification, à la classe 2. Le total de chaque ligne indique le nombre d'observations appartenant à la classe correspondante, tandis que le total de chaque colonne indique le nombre d'observations affectées à la classe correspondant à cette colonne lors de la classification.

Les éléments de la première colonne des taux d'erreur représentent le total des observations mal classées de la classe correspondant à la ligne.

Le taux d'erreur par classe est donné par le rapport entre le nombre d'observations mal classées de cette classe et le nombre total d'observations appartenant à cette classe.

Le taux d'erreur global est le rapport entre le nombre total d'observations mal classées par la procédure et le nombre total d'observations constituant l'échantillon

Distribution	Nombre d'observations	Vecteur Moyenne	Matrice de covariance
1	185	$\overline{X}_1 = \begin{bmatrix} 1,092 \\ 0,961 \end{bmatrix}$	$S_1 = \begin{bmatrix} 1,682 & -0,039 \\ -0,039 & 2,209 \end{bmatrix}$
2	218	$\overline{X}_2 = \begin{bmatrix} 5,538 \\ 2,211 \end{bmatrix}$	$S_2 = \begin{bmatrix} 1,873 & -0,129 \\ -0,129 & 2,014 \end{bmatrix}$
3	297	$\overline{X}_3 = \begin{bmatrix} 3,773 \\ 8,192 \end{bmatrix}$	$S_3 = \begin{bmatrix} 2,163 & 0,108 \\ 0,108 & 1,592 \end{bmatrix}$

Paramètres statistiques de l'exemple 1 après classification par la méthode proposée

Tableau V-2

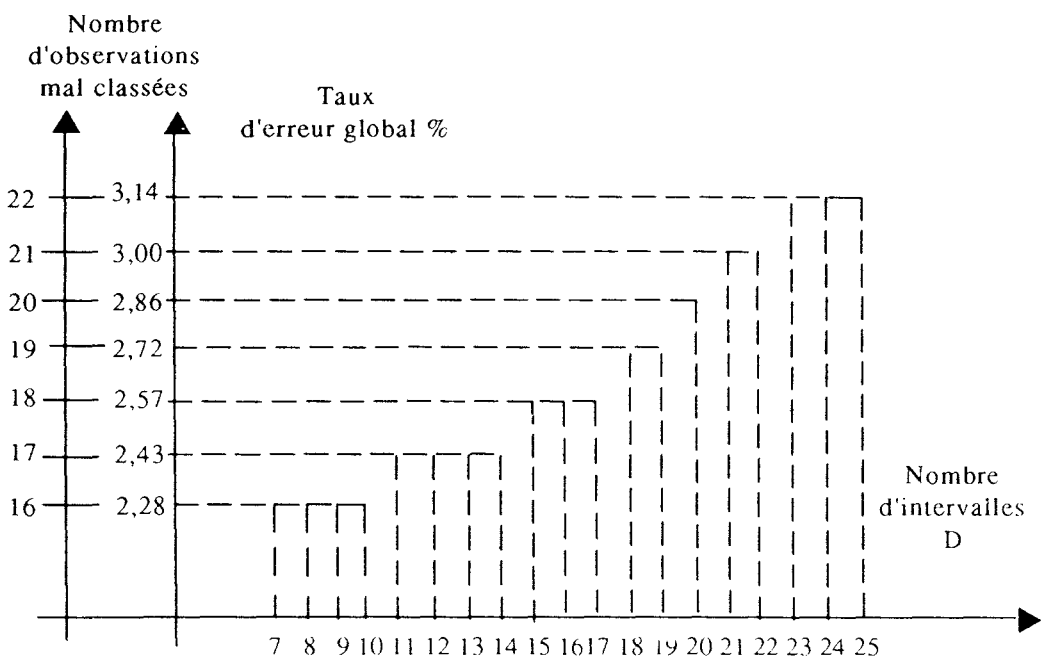
Matrice de confusion					Taux d'erreur		
Les classes	1ère classe	2ème classe	3ème classe	Total	Nombre total d'observations mal classées	Taux d'erreur par classe	Taux d'erreur global
1ère classe	185	15	0	200	15	15/200 = 0,075	0,02571 ou 2,571 %
2ème classe	0	200	0	200	0	0/200 = 0	
3ème classe	0	3	297	300	3	3/300 = 0,01	
Total	185	218	297	700	18		

Matrice de confusion et taux d'erreur de l'exemple 1

Tableau V - 3

soumis à l'analyse. Pour cet exemple pour D=16, ce taux d'erreur global est de 2,57 %, ce qui correspond à 18 observations mal classées sur le total de 700.

Le seul paramètre à ajuster par l'analyste est le pas de discrétisation des données. Si nous comparons les taux d'erreurs globaux en fonction de ce paramètre dans la plage de stabilité du nombre de classes, nous remarquons sur la figure V.18 que cette procédure de classification n'est pas très sensible à l'ajustement du nombre d'intervalles D dans cette plage. En effet, entre D=7 et D=25, l'accroissement du taux d'erreur global correspond à seulement 6 observations mal classés sur un total de 700.



Variation du taux d'erreur global avec le nombre d'intervalles D.

Figure V.8

D'autre part, nous remarquons qu'à la limite inférieure de la plage de stabilité, D=7, nous obtenons le plus petit nombre d'observations mal classées, 16 observations, et par conséquent le meilleur taux d'erreur global 2,28 % par rapport aux nombres d'intervalles.

Nous pouvons évaluer la performance de cette méthode en comparant ce taux d'erreur global au taux d'erreur théorique optimale défini par l'approche Baysienne. Pour cet exemple, ce taux théorique est de 3,06 %. Cette valeur comparée aux valeurs de la figure V.8, explicite la pertinence de la méthode de classification par maximisation de la taille de regroupements qui offre un taux d'erreur plus ou moins constant dans la plage de stabilité et conforme au taux théorique optimale.

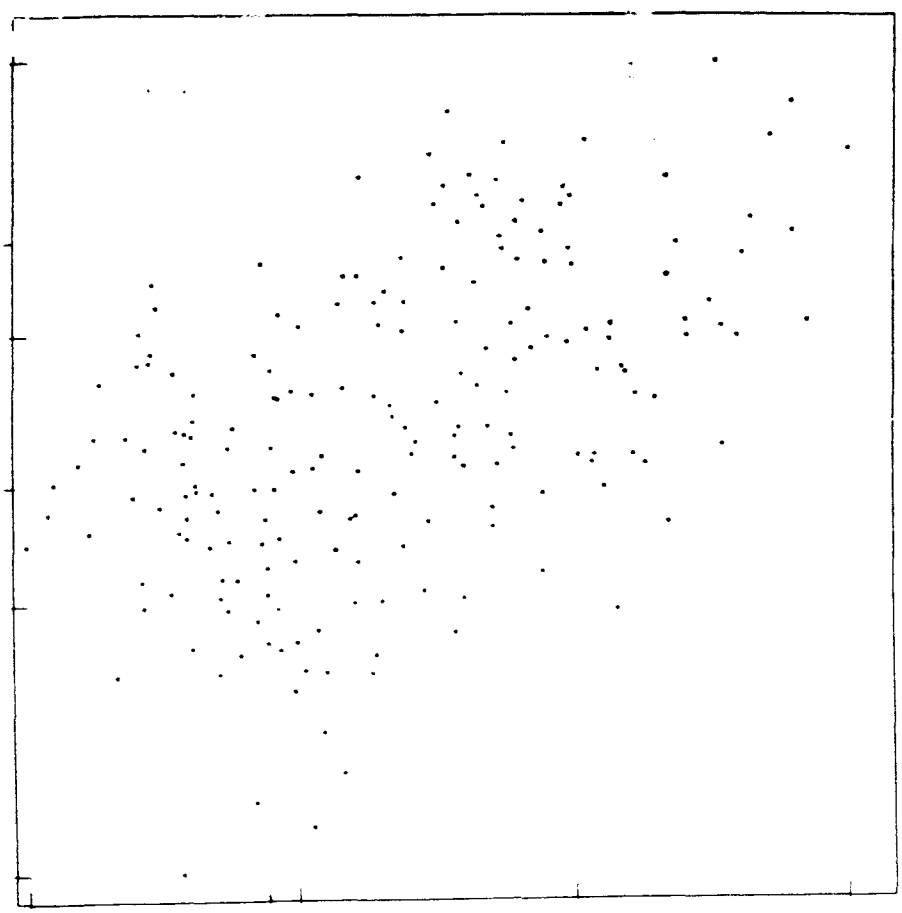
V - 3. EXEMPLE BIDIMENSIONNEL COMPOSE DE DEUX CLASSES NORMALES NON SPHERIQUES A FAIBLES EFFECTIFS DISPERSEES

Les données utilisées dans ce deuxième exemple sont des observations provenant de deux classes normales non sphériques et équiprobables dont les paramètres statistiques ainsi que le nombre d'observations par classe sont précisés dans le tableau V-4. La figure V.9 présente l'échantillon tel qu'il a été simulé.

Distri-bution	Nombre d'observations	Vecteur Moyenne	Matrice de covariance	Probabilité a priori
1	100	$\overline{X}_1 = \begin{bmatrix} 8,1062 \\ 7,6534 \end{bmatrix}$	$S_1 = \begin{bmatrix} 1,516 & -0,137 \\ -0,137 & 2,433 \end{bmatrix}$	$Pb_1 = {}^1/_2$
2	100	$\overline{X}_2 = \begin{bmatrix} 4,6169 \\ 4,9233 \end{bmatrix}$	$S_2 = \begin{bmatrix} 1,22 & -0,262 \\ -0,262 & 2,045 \end{bmatrix}$	$Pb_2 = {}^1/_2$

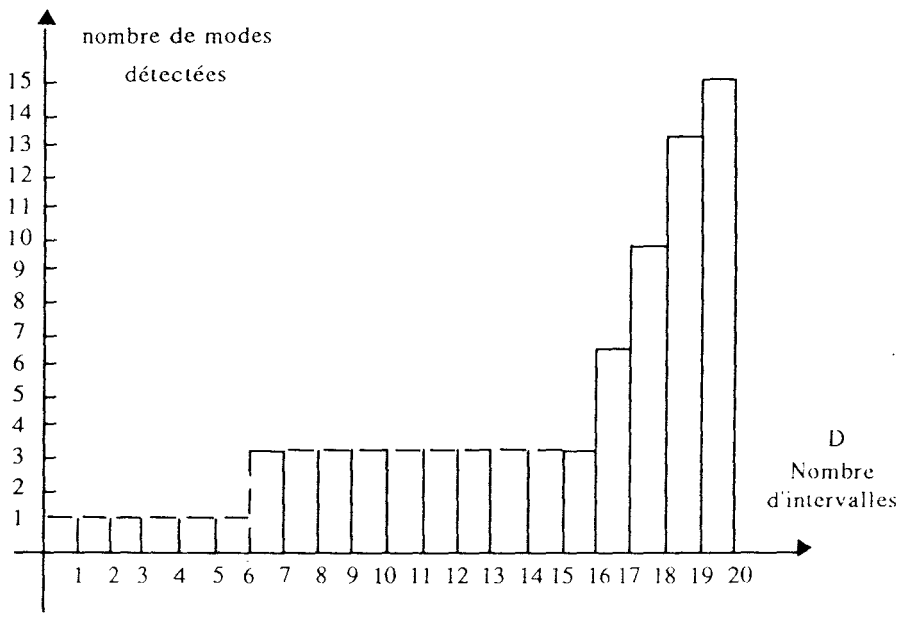
Paramètres statistiques de l'exemple 2

Tableau V - 4



Représentation graphique de l'échantillon de l'exemple 2

Figure V.9



Nombre de modes détectés dans l'échantillon de l'exemple 2 en fonction du nombre d'intervalles D

Figure V.10

La figure V.10 indique les variations du nombre de modes mis en évidence par la procédure de classification en fonction du nombre D d'intervalles de discrétisation. La plage de stabilité s'étend de D = 6 à D = 15.

Nous remarquons ici aussi qu'un éclatement des modes intervient brutalement dès que le nombre d'intervalles D dépasse la valeur 15, limite supérieure de cette plage de stabilité.

Le tableau V-5 indique les paramètres statistiques après la classification par la méthode proposée pour un nombre d'intervalles égal à la valeur médiane de la plage de stabilisaté, c'est-à-dire D = 11. Le nombre de classes, le nombre d'observations affectées à chacune d'elles, leurs vecteurs moyenne et leurs matrices de covariance sont donnés dans ce tableau.

Distribution	Nombre d'observations	Vecteur Moyenne	Matrice de covariance
1	106	$\overline{X}_1 = \begin{bmatrix} 8,4932 \\ 7,8359 \end{bmatrix}$	$S_1 = \begin{bmatrix} 2,01 & -0,141 \\ -0,141 & 2,98 \end{bmatrix}$
2	94	$\overline{X}_2 = \begin{bmatrix} 4,2091 \\ 4,7992 \end{bmatrix}$	$S_2 = \begin{bmatrix} 1,099 & -0,233 \\ -0,233 & 1,901 \end{bmatrix}$

Paramètres statistiques après classification pour l'exemple 2

Tableau V - 5

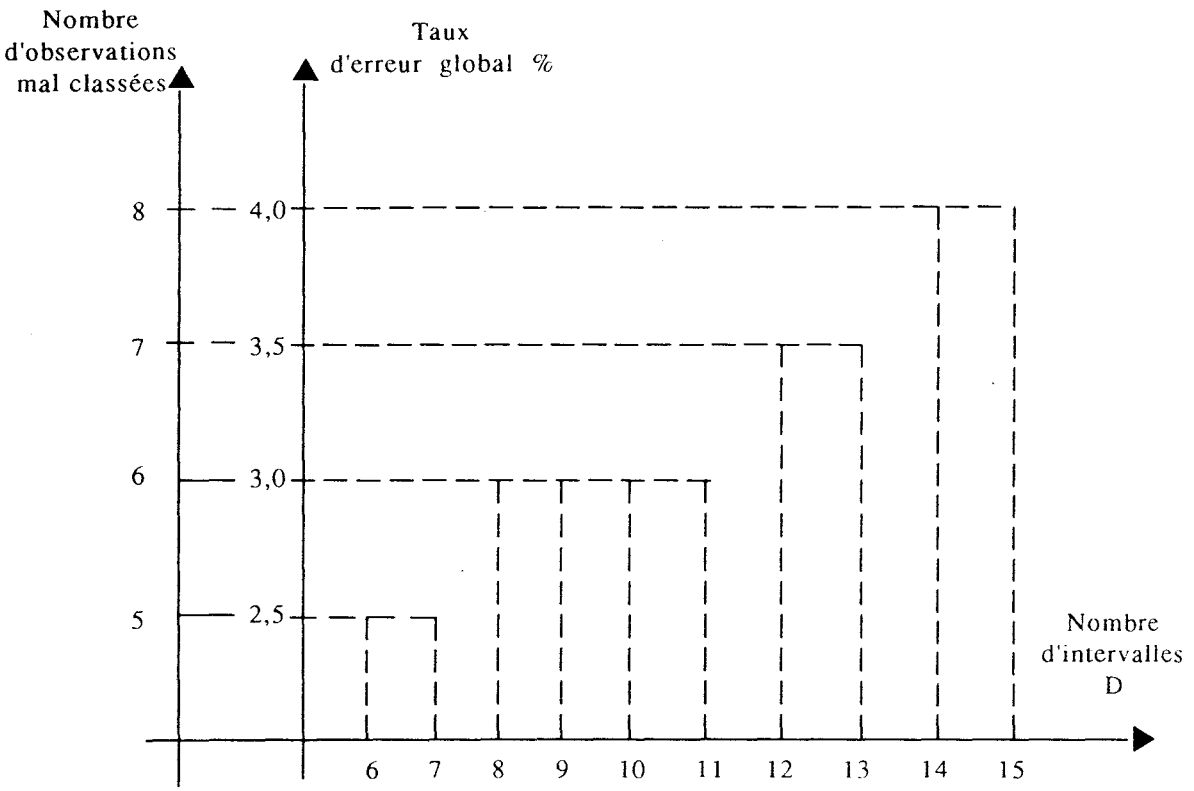
Le tableau V-6 donne la matrice de confusion associée à la classification par cette méthode, le taux d'erreur par classe et le taux d'erreur global pour D=11. Pour cet exemple, ce taux d'erreur est de 3 %.

Matrice de confusion				Taux d'erreur		
Les classes	1ère classe	2ème classe	Total	Nombre total d'observations mal classées	Taux d'er-	Taux d'erreur global
1ère class	100	0	100	0	0/100 = 0	0,03 ou 3 %
2ème classe	6	94	100	6	6/100 = 0,06	
Total	106	94	200	6		

Matrice de confusion et taux d'erreur de l'exemple 2

Tableau V - 6

La figure V.11 indique les variations du taux d'erreur en fonction du nombre d'intervalles dans la plage de stabilité. Nous remarquons que, comme dans l'exemple précédant, et malgré le faible effectif de chaque classe et leur grande dispersion, la procédure de classification est peu sensible au nombre d'intervalles. En effet, entre D = 6 et D = 15, l'accroissement du taux d'erreur global correspond à seulement 3 observations mal classées sur un total de 200.



Variation du taux d'erreur global avec le nombre d'intervalles D.

Figure V.11

D'autre part, nous remarquons, que comme dans l'exemple 1, qu'à la limite inférieure de la plage de stabilité, D=6, nous obtenons le plus petit nombre d'observations mal classées et le meilleur taux d'erreur global.

Le taux théorique optimale calculé pour cet exemple, est de 3,98 %. Cette valeur comparée aux valeurs de la figure V.11, affirme de plus en plus la pertinence

de la méthode de classification par maximisation de la taille des regroupements qui offre un taux d'erreur plus ou moins constant dans la plage de stabilité et conforme au taux théorique optimale.

V - 4 . EXEMPLE BIDIMENSIONNEL COMPOSE DE TROIS CLASSES EN FORME DE CROISSANT

Les données utilisées dans ce troisième exemple sont des observations provenant de trois classes équiprobables dont :

- Deux classes en forme de croissant. Les observations de ces classes sont définies par deux attributs donnés par $(A \cos \theta + B)$ et $(A \sin \theta + B_2)$, où A est l'amplitude de la classe, θ est la variable aléatoire normale appelée l'angle de la classe, et où B_2 et B_2 sont aussi des variables aléatoires normales. Les paramètres statistiques de ces deux classes ainsi que le nombre d'observations par classe sont précisés dans le tableau V-7(a).
- Et une classe normale non sphérique. Les paramètres statistiques et le nombre d'observations par cette classe sont précisés dans le tableau V-7 (b).

La figure V.12 présente l'échantillon tel qu'il a été généré.

Classe / Paramètres	En forme de croissant	
	Classe 1	Classe 2
Nombre d'observations	300	300
Amplitude de la classe	15	15
Moyenne de l'angle de la classe θ	0	3,14159
Variance de l'angle θ	0,707	0,707
Moyenne de B_1	5	5
Variance de B_1	10	10
Moyenne de B_2	5	5
Variance de B_2	10	10

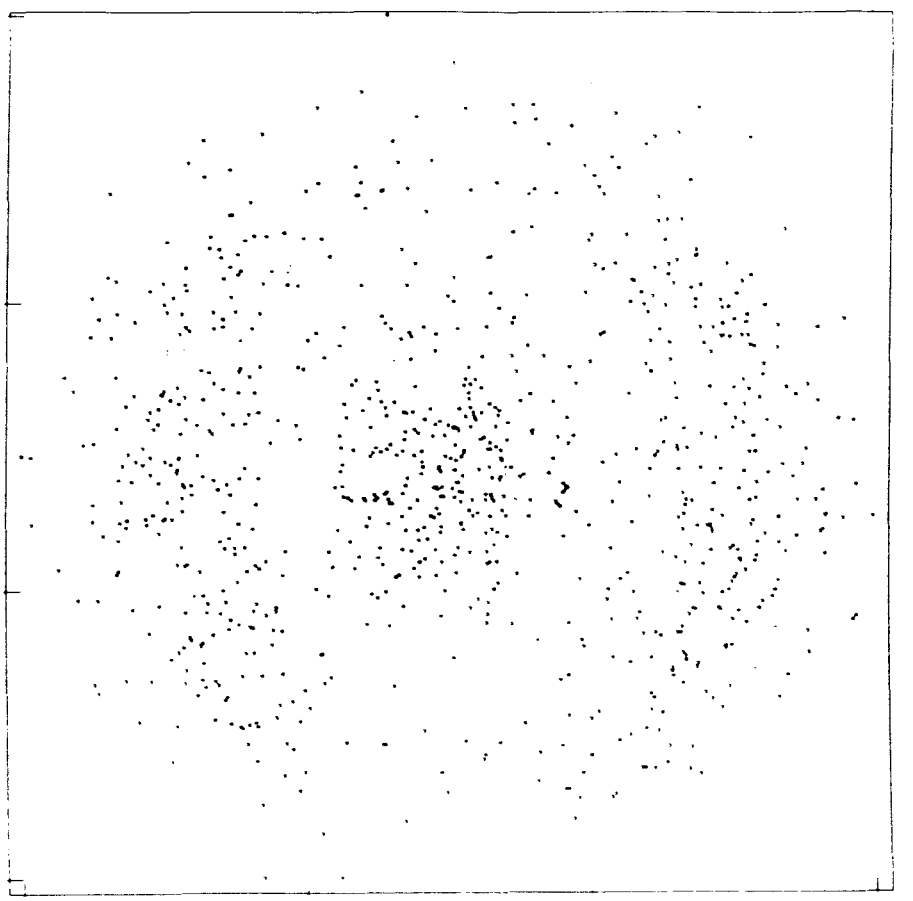
(a)

Classe / Paramètres	Normale non sphérique
Nombre d'observations	300
Moyenne de la première variable x_1	5
Variance de la première variable x_1	15
Moyenne de la seconde variable x_2	5
Variance de la seconde variable x_2	15

(b)

Paramètres statistiques de l'exemple 3

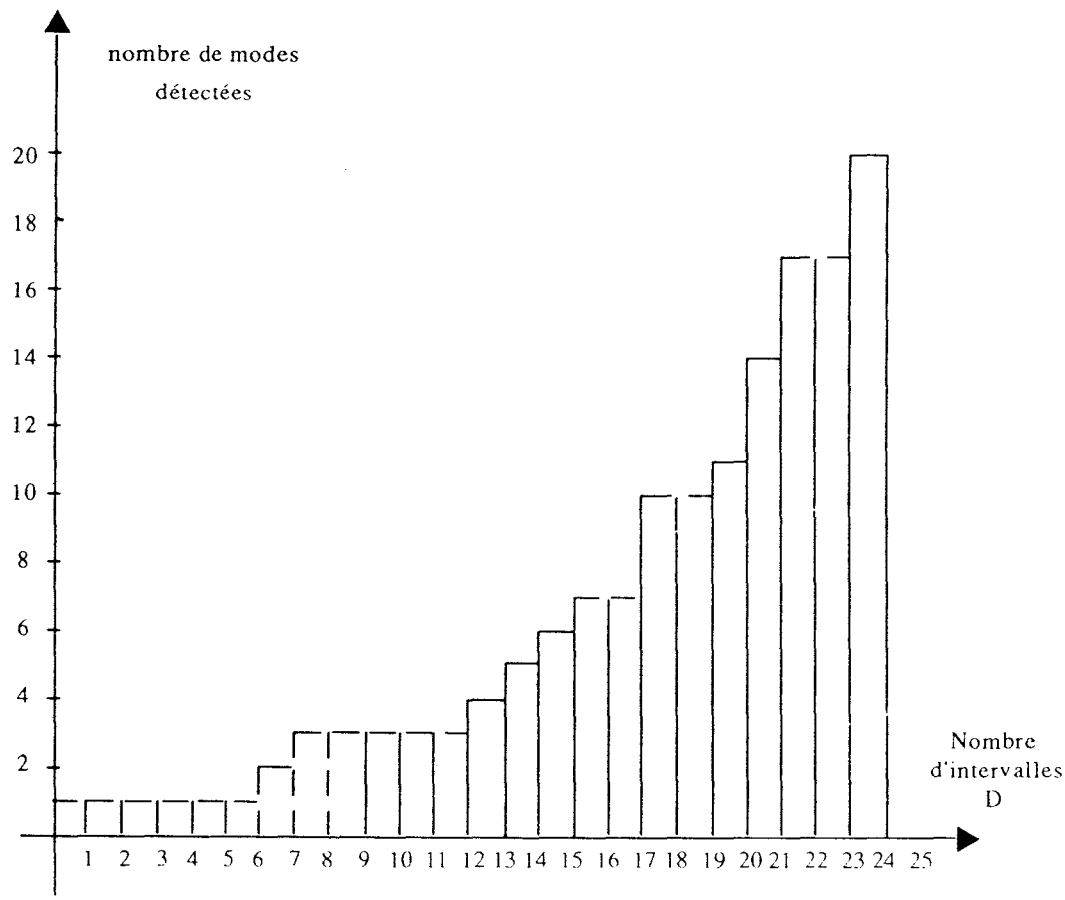
Tableau V - 7



Représentation graphique de l'échantillon
de l'exemple 3

Figure V.12

La figure V.13 indique les variations du nombre de modes mis en évidence par la procédure de classification en fonction du nombre d'intervalles. La plage de stabilité s'étend de D = 7 à D= 11.



Nombre de modes détectés dans l'échantillon de l'exemple 3 en fonction du nombre d'intervalles D

Figure V.13

V - 5 . EXEMPLE A QUATRE DIMENSIONS COMPOSE DE TROIS CLASSES NORMALES NON SPHERIQUES.

Les données utilisées dans ce quatrième exemple représentent des observations d'une distribution à trois classes dans un espace à quatre dimensions.

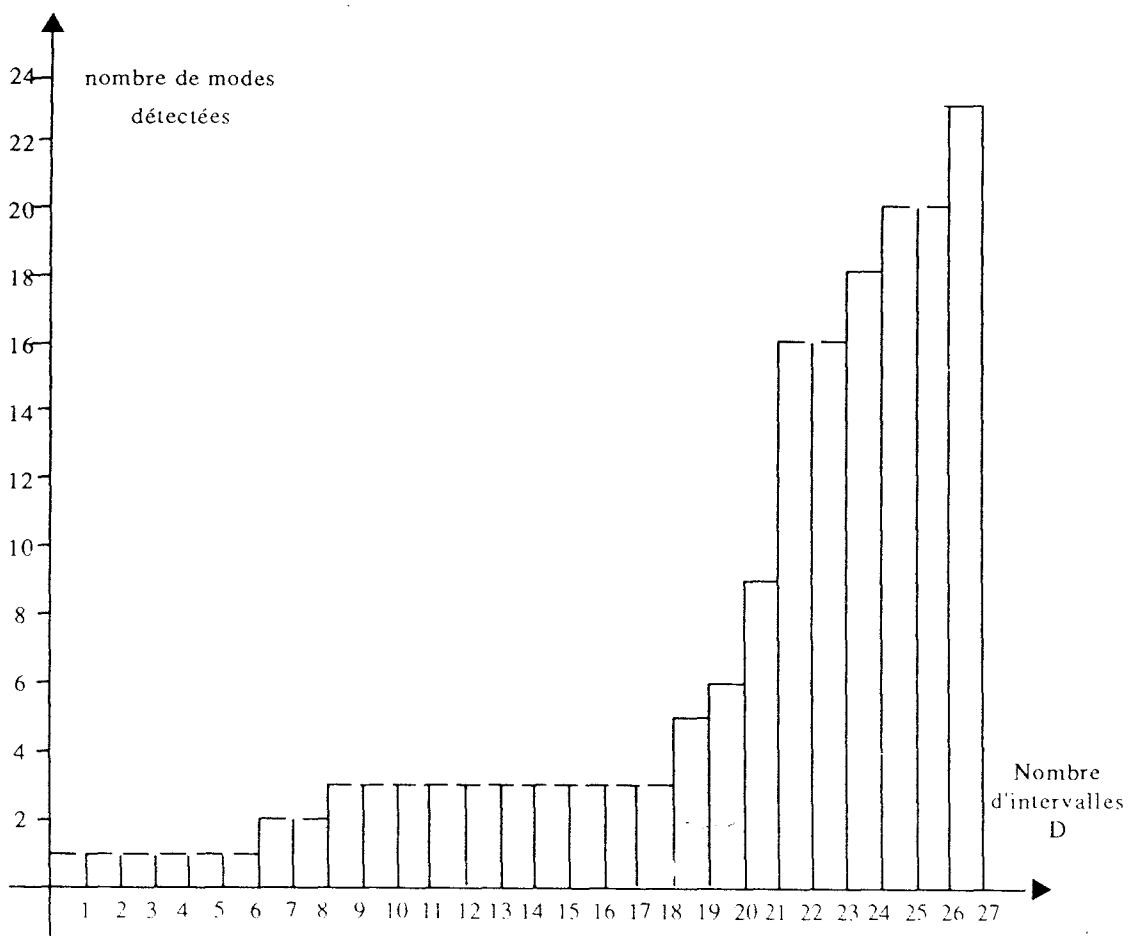
Les paramètres statistiques ainsi que le nombre d'observations par classe, tel qu'elles ont été générées, sont données par le tableau V-9.

Classe / Paramètres	Classe 1	Classe 2	Classe 3
Nombre d'observations par classe	200	100	200
Moyenne de la 1ère variable	8,000	1,000	5,000
Variance de la 1ère variable	2,000	1,000	1,000
Moyenne de la 2ème variable	7,000	3,000	4,000
Variance de la 2ème variable	1,000	0,500	1,000
Moyenne de la 3ème variable	5,000	2,000	3,000
Variance de la 3ème variable	3,000	0,500	1,000
Moyenne de la 4ème variable	4,000	7,000	3,000
Variance de la 4ème variable	2,000	2,000	1,000

Paramètres de l'exemple 4

Tableau V - 9

La méthode de classificattion par maximisation des regroupements a été appliquée à cet exemple pour différents nombres d'intervalles de discrétisation de l'espace à 4 dimensions allant de 1 à 30. La figure V.15 indique qu'une plage de stabilité s'étend de D=8 à D=17 et que, comme dans les autres exemples, un éclatement des modes intervient brutalement dès que le nombre d'intervalles dépasse la limite supérieure de la plage de stabilité.



Nombre de modes détectés dans l'échantillon de l'exemple du tableau V-10 en fonction du nombre d'intervalles D

Figure V.15

CONCLUSION GENERALE

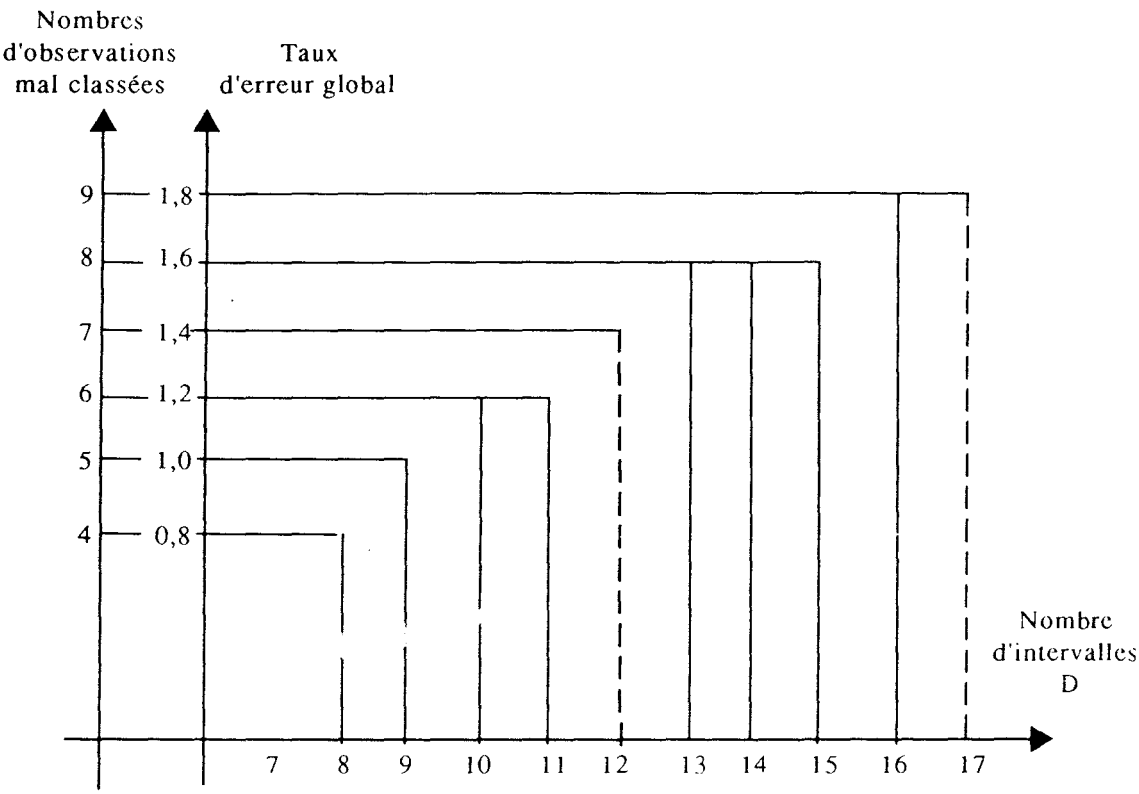
Le tableau V-10 nous donne la matrice de confusion associée à la classification par cette méthode, le taux d'erreur par classe et le taux d'erreur global qui est de 1,4 % pour D=12.

Matrice de confusion					Taux d'erreur		
Les classes	1ère classe	2ème classe	3ème classe	Total	Nombre total d'observations mal classées	Taux d'erreur par classe	Taux d'erreur global
1ère classe	196	0	4	200	4	4/200 = 0,02	0,014
2ème classe	0	98	2	100	2	2/100 = 0,02	ou
3ème classe	1	0	199	200	1	1/200 = 0,01	1,4 %
Total	197	98	205	500	7		

Matrice de confusion et taux d'erreur pour l'exemple 4

Tableau V-10

D'autre part, la figure V.16, indique que, comme dans les exemples précédents, les variations du taux d'erreur en fonction de D dans la plage de stabilité sont minimes. L'accroissement du taux d'erreur global correspond à seulement 5 observations mal classées sur un total de 500. Cela confirme que la procédure de classification est peu sensible au nombre d'intervalles dans la plage de stabilité.



Variation du taux d'erreur global avec le nombre d'intervalles D.

Figure V.16

Comme dans les autres exemples, nous remarquons que le nombre minimum d'observations mal classées qui donne le meilleur taux d'erreur globale est toujours à la limite inférieure de la plage de stabilité.

D'autre part, le taux d'erreur théorique optimale pour cet exemple d'une valeur de 1,46 % , comparée aux valeurs de la figure V.16, nous confirme la pertinence de la méthode.

V - 6 . CONCLUSION

Dans ce chapitre, nous avons présenté quelques résultats obtenus par la méthodologie de classification automatique par amincissement de l'histogramme multidimensionnel que nous proposons dans ce mémoire. Pour justifier les performances de la méthode, nous avons utilisé des exemples bidimensionnels et multidimensionnels constitués de classes de formes variées.

La comparaison systématique des résultats obtenus aux données réelles a permis de montrer l'intérêt d'une telle approche qui ne nécessite aucune connaissance a priori sur les données à classer et qui est très robuste par rapport à l'ajustement de l'unique paramètre laissé à l'initiative de l'analyste. L'étude de l'effet de la finesse de la discrétisation sur le nombre de modes détectés a permis de définir de grandes plages de stabilité qui correspondent à une mise en évidence fiable de la véritable structure des données.

Nous pouvons en toute confiance, préciser que la méthode de classification par maximisation des regroupements est caractérisée par :

- une vitesse d'éxécution rapide, si toutefois un matériel adéquat est utilisé,

- une plage de stabilité assez large, suivi d'un éclatement du nombre de modes,

- un meilleur taux d'erreur globlal à la limite inférieure de la plage de stabilité. C'est pourquoi nous proposons que la classification finale, après la détermination de la plage de stabilité, s'éxécute à cette limite inférieure bien que ce taux d'erreur est quasiment insensible au nombre d'intervalles de la plage,

- un taux d'erreur global comparable au taux d'erreur théorique optimale.

CONCLUSION GENERALE

Dans ce mémoire, nous avons proposé une nouvelle approche pour l'analyse de données multidimensionnelles. La méthodologie de maximisation de la taille des regroupements que nous avons développée a pour but, d'une part, d'extraire les modes associés aux différentes classes en présence dans l'échantillon soumis à l'analyse, d'autre part, de regrouper en classes les observations de l'échantillon analysé à partir de ces modes.

Nous avons montré que, dans un contexte non supervisé dans lequel on ne dispose d'aucune information a priori sur les données à classer, la méthode de maximisation de la taille des regroupements permet la détection des modes dans le cas d'échantillons constitués de classes de formes et de tailles variées.

D'autre part, comme la plupart des méthodes de classification basées sur l'analyse de l'histogramme multidimensionnel, cette approche est précédée d'une phase de discrétisation de l'espace de représentation des données. Le seul paramètre à ajuster durant cette étape est le pas de discrétisation. Nous avons proposé, dans ce mémoire, un moyen d'ajuster ce paramètre et cela en définissant une plage de

stabilité du nombre de classes détectée en fonction du nombre d'intervalles de discrétisation de l'espace.

Un point important considéré dans ce mémoire est celui de l'effet du nombre d'attributs décrivant chacune des observations de l'échantillon à analyser sur la complexité des algorithmes. Nous avons introduit une méthode pour réduire le balayage d'un espace multi-dimensionnel à un problème à une seule dimension par une reformulation mathématique simplifiée, qui ne néglige aucun des attributs utiles, et ce, quelque soit leur nombre.

Comme nous nous sommes situés dans ce contexte non supérvisé, le seul moyen d'évaluer les résultats obtenus par la méthode proposée était de les comparer aux données générées. Cette comparaison nous montre l'intérêt de cette approche par maximisation de la taille des regroupements pour des échantillons constitués de classes de formes et de tailles variées.

D'autre part, cette méthode est caractérisée par une grande vitesse d'éxécution, et un taux d'erreur global comparé au taux d'erreur théorique optimale défini par l'approche Bayesienne

Ce travail montre que la détection des modes par

cette méthode constitue une approche intéressante aux problèmes de classification automatique non supervisée.

Nous pouvons enfin, confirmer l'adaptabilité de la méthode de classification par maximisation de la taille des regroupements au traitement d'images [Vannoorenbergue, Wachowski, Barsoum et Postaire, 1994]. Cette application, présentée à la troisième conférence internationale du traitement d'images en Pologne ouvre de larges possibilités quant à l'utilisation de cette méthode dans différents domaines. Ce à quoi nous nous employons.

BIBLIOGRAPHIE

Bezdek J.C., Chuah S.K. Leep D., Generalized k-nearest neighbor rules. Fuzzy Sets and Systems, Vol. 18, pp. 237-256, 1986.

Bock H.H., On some significance tests in cluster analysis. Journal of Classification, Vol. 2, pp. 77-108, 1985.

Bock H.H., Probabilistic Aspects in Cluster Analysis. In Conceptual and Numerical Analysis of Data, Opitz O. (ed.) Springer Verlag, pp. 12-44, 1989.

Botte-Lecocq C., Postaire J.G., Itérations of morphological transformations for cluster separation in pattern recognition, Symbolic-Numeric data analysis and learning, Nova science Pub, New York. pp. 173-185, 1991a.

Botte-Lecocq C., Postaire J.G., Classification automatique par extraction des composantes connexes d'un ensemble discret binaire représentatif des données, 8ème congrès reconnaissance des formes et intelligence artificielle, RFIA 1991, Lyon, France, pp. 283-290, 1991b.

Bruynooghe M., Classification ascendante hiérarchique des grands ensembles de données. L'algorithme rapide fondé sur la construction de voisinages réductibles. Les Cahiers de l'analyse des données, 1978.

Celeux G., Reconnaissance de mélanges de densités de probabilité et applications à la validation des résultats en classification. Thèse de Doctorat d'Etat, Université de Paris IX Dauphine, 1987.

Celeux G., Le traitement des données manquantes dans le logiciel SICLA. Rapport Technique INRIA, n° 102, 1988.

Celeux G., Diebolt J., Une version de type recuit simulé de l'algorithme E.M. Rapport de Recherche INRIA, n° 1123, 1989.

Asselin de Beauville J.P., Estimation non paramétrique de la densité et du mode : exemple de la distribution Gamma. Laboratoire d'informatique appliquée, Université de Tours, Parc de Grandmont, Revue de statistique appliquée, Vol. 26, pp. 47-70, 1978.

Asselin de Beauville J.P., Un algorithme d'estimation du mode : étude de la convergence. Pub. Inst. Univ. Paris, Vol. 24, fax : 3-4, 1-29, 1979.

Asselin de Beauville J.P., L'estimation des modes d'une densité de propabilité multidimensionnelle : statistiques et analyse de données, Vol. 8 n° 7, pp. 16-40, 1983.

Asselin de Beauville J.P., Panorama sur l'utilisation du mode en classification automatique - Rairo - APII 23, pp. 113-137, 1989.

Asselin de Beauville J.P., Vitesse de convergence presque sûre d'une estimation du mode. Extrait des comptes-rendus de l'Académie des Sciences, T. 303, séries 1, n° 11, 1986.

Ball G.H., Hall D.J., Isodata, A novel method of data analysis and Pattern Classification, Technical Report, SRI Project 5533, Stanford Research Institute Manls Park, CA, USA, 1965.

Béreau M., Contribution de la théorie des sous-ensembles flous à la règle de discrimination des k plus proches voisins en mode partiellement supervisé. Thèse de Doctorat de l'Université de Technologie de Compiègne, 1986.

Bezdek J.C., Pattern Recognition with fuzzy objective function algorithms. Plenum Press, 1981.

Bezdek J.C., Hathaway R.J., Huggins V.J., Parametric estimation for normal mixtures. Pattern Recognition Letters, Vol. 3, pp. 79-84, 1985.

Bezdek J.C., Hathaway R.J., Howard R.E., Wilson C.A., Coordinate descent and clustering. Control and Cybernetics, Vol. 15, n° 2, pp. 196-204, 1986.

Celeux G., Govaert G., A classification EM algorithm for clustering and two stochastic versions. Rapport de Recherche INRIA, n° 1364, 1991.

Celeux G., Govaert G., Comparison of the mixture and the classification maximum likelihood in cluster analysis. Rapport de Recherche INRIA, n° 1517, 1991.

Chaudhuri B.B, Dutta Majumder D., On membership evaluation in fuzzy sets. In Approximate Reasoning in Decision Analysis, Gupta M.M. and Sanchez E. (Eds.), North-Holland, pp. 3-11, 1982.

Chen C.H., On information and distance measures, error bounds and feature selection, Information sciences, Vol. 10, pp. 159-173, 1976.

Czesnalowiez E., Postaire J.G., Détection des contours des modes sur l'estimation des k plus proches voisins. Application en classification automatique, 12ième journée statistique, Tours, pp. 82-83, 1990.

Daoudi M., Hamad D., Postaire J.G., Interactive Classification Through neural networks. International Conference on Neural Networks and Genetic Algorithms, ANNGA 93, Vol. 1, pp. 80-85, Insbruck, Autriche, 1983.

Dakart S., Hamad D., Daoudi M., Postaire J.G., Application of neural networks to gradient search techniques in cluster analysis. International Conference on Neural Networks and and Genetic Algorithms, ANNGA 93, Vol 1, pp. 154-153, Insbruck, Autriche, 1993.

Daoudi M., Hamad D., Postaire J.G., A display oriented technique for interactive pattern recognition by multi-layer neural networks, IEEE International Conference on Neural Networks, Vol. 3, pp. 1633-1637, San Fransico, CA, USA, 1983.

Daoudi M., Ben Sliman R., Hamad D., Postaire J.G., A new interactive pattern recognition approach using multi-layer neural networks and mathematical morphology, IEEE International Conference Systems, Man and Cybernetics, le Touquet, France, Vol. SMC 4, pp. 125-130, 1993.

Diday E., Lemaire J., Pouget J., Testu F., Eléments d'analyse des données. Dunod, Paris, 1982.

Dubes R.C., How many clusters are best ? - An experiment. Pattern Recognition, Vol. 20, n° 6, pp. 645-663, 1987.

Dubuisson B., Malvache P., Grenier D., The use of Pattern Recognition in order to improve fast nuclear reactor monitoring. In Pattern Recognition in Pratice II, E.S. Gelsema, L.N. Kanal (Eds.), Elsevier Science (North-Holland), pp. 471-478, 1986.

Dubuisson B., Diagnostic et reconnaissance des formes. Traité des Nouvelles Technologies, séries Diagnostic et Maintenance, Hermès, 1990 a.

Duda R.O., Hart P.E., Pattern Classification and Scene Analysis. John Wiley & Sons, New York, 1973.

Eigen D.L., Fromm F.R., Northouse R.A., Cluster analysis based on dimensional information with applications to feature selection and classification, IEEE Transaction Systems, Man and Cybernatics, Vol. SMC Vol. 4, n° 3, pp. 284-294, 1974.

Everitt B.S., Cluster analysis. The analysis of survey data, Wiley and Son, Vol. 1, New York, 1977.

Everitt B., Hand D., Finite mixture distributions. Chapman and Hall, 1981.

Fehlauer J., Eisenstein B.A., A declustering criterium for feature estimation in pattern recognition, IEEE Transaction on Computers, Vol. C-27, n° 3, pp. 261-266, 1978.

Florek K., Lukerzewicz J., Perkal J., Steinhaus H., Zubrzycki S., Sur la liaison et la division des points d'un ensemble fini. Colloquium Math., 2, pp. 282-285, 1951.

Fromm F.R., Northouse R.A., Some results of nonparametric clustering of large data problems. First Conference Pattern Recognition, pp. 18-21, 1973.

Fukunaga K., Hostetler L.D., The estimation of the gradient of a density fuction, with applications in pattern recognition, IEEE Trans. Inf. Theory, Vol. 21, n° 1, pp. 32-40, 1975.

Fukunaga K., Ando S., The optimum nonlinear features for a scatter criterium in discriminant analysis, IEEE Information theory Vol. IT-23, n° 4, pp. 453-459, 1977.

Fukunaga K., Mantock J.M., Nonparametric discriminant analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 5, n° 6, pp. 671-678, 1983.

Fukunaga K., Hummels D.M., Bayes Error Estimation using Parzen and K-NN Procedures. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 9, n° 5, pp. 634-643, 1987.

Fukunaga K., Introduction to statistical pattern recognition. Academic Press, New York, 1990.

Gana K., Suivi d'évolution et aide au diagnostic en maintenance de système industriel. Thèse de Docteur-Ingénieur, Université de Valenciennes, 1987.

Gordon A.D., A review of hierachical classification. J.R. Statist. Soc. A, 150, part 2, pp. 119-137, 1987.

Grenier D., Methode de détection d'évolution. Application à l'instrumentation nucléaire. Thèse de Docteur-Ingénieur de l'Université de Technologie de Compiègne, 1984.

Jain A.K., Waller W.G., On the optimal number of features in the classification of multivariate gaussian data, Fourth international conference on Pattern Recognition, Kyoto, pp. 265-269, 1978.

Jain A.K., Dubes R.C., Chen C.C, Bootstrap Techniques for Error Estimation. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 9, n° 5, pp. 628-633, 1987.

Jain A.K., Moreau J.V., Bootstrap technique in cluster analysis. Pattern Recognition, Vol. 20, n° 5 , pp. 547-568, 1987.

Jozwik A., A learning scheme for a fuzzy K-NN rule. Pattern Recognition Letters, Vol. 1, pp. 287-289, 1983.

Keller J.M., Gray M.R., Givens J.A., A fuzzy K-Nearest Neigbor Algorithm. IEEE Trans. on Systems, Man and Cybernetics, Vol. 15, n° 4, pp. 580-585, 1985.

King B.F., Market and industry factors in stock price behaviour, Journal of business, Vol. 39, pp. 139-190, 1966.

King B.F., Stepwice clustering procedures, J. Am. Statist. Ass., Vol. 62, pp. 86-101, 1967.

Kittler J., A locally sensitive method for cluster analysis, Pattern Recognition, Vol. 8, pp. 23-33, 1976.

Khotanzad A., Bouarfa A., Image segmentation by a parallel, non-parametric histogram based clustering algorithm, Pattern Recognition, Vol. 23, n° 9, pp. 961-973, 1990.

Koontz W.L., Narendra P.M., Fukunaga K., A graph-theoretic approach to nonparametric cluster analysis, IEEE Trans. Comput. Vol. 25, pp. 936-944, 1976.

Safian S.R., Landgrebe D., A survey of decision tree classifier methodology. IEEE Trans. on Systems Man and Cybernetics, Vol. 21, n° 3, pp. 660-674, 1991.

Smith S.P., Jain A.K., A test to determine the multivariate normality of a data set. IEEE Trans. on Pattern Analysis and Machine Intellignece, Vol. 10, n° 5, pp. 757-761, 1988.

Smolarz A., Un algorithme de discrimination adaptatif avec rejet. R.A.I.R.O. APII, Vol. 21, n° 5, pp. 449-474, 1987.

Sneath P.H.A., The application of computers to taxonomy, J. Gen. Microbiol, Vol. 17, pp. 201-226, 1957.

Sneath P.H., Sokal R.R., Numerical Taxonomy, W.H. Freeman and Co., San Francisco, 1973.

Sokal. R.R., Michener C.D., A statistical method for evaluating systematic relationship. Univ. Kansas Sci. Bull., Vol. 38, pp. 1409-1438, 1958.

Sonquist J.A., Morgan J.N., Problems in the analysis of survey data and a proposal. J. Am. Statist. Ass., Vol. 58, pp. 415-435, 1963.

Sonquist J.A., Morgan J.N., The determination of Interaction Effects, Survey Research Centre, Institute of social Research, University of Michigan, 1964.

Sorsa T., Koivo H.N., Application of artificial neural networks in process fault diagnosis. IFAC Symposium SAFEPROCESS'91, Baden-Baden, pp. 133-138, 1991.

Thorndike R.F., Who belongs in a familly ? Psychometrika, Vol. 18, pp. 267-276, 1953.

Titterington D., Smith A., Makov U., Statistical analysis of finite mixture distribution. Wiley. 1985.

Touzani A., Postaire J.G., Mode detection by relaxation, IEEE Transaction Pattern Analysis and Machine Intelligence, Vol. PAMI - 10 - n° 6, pp. 970-978, 1988

Touzani A., Postaire J.G., La notion de contour des fonctions de densité de probabilité en classification automatique. II) Classification non paramétrique par extraction des contours des modes des fonctions de densité. APII. Vol. 23, n° 5, pp. 423-440, 1989 a.

Touzani A., Postaire J.G., III) Classification paramétrique par identification des mélanges gaussiens. APII, Vol. 25, n° 2, pp. 53-70, 1991.

Touzani A., Postaire J.G., Clustering by mode boundary detection, Pattern Recogn. Letters, Vol. 9, pp. 1-12, 1989 b.

Touzani A., Postaire J.G., La notion de contour des fonctions de densité de probabilité en classification automatique. III) Classification paramétrique par identification des mélanges gaussiens. APII, Vol. 25, n° 2, pp. 141-164, 1991.

Trauweart E., Kaufman L., Rousseeuw P., Fuzzy clustering algorithms based on the maximum likelihood principle. Fuzzy sets and Systems, Vol. 42, n° 2, pp. 213-227, 1991.

Usai M., Dubuisson B., An adaptative non parametric classification algorithm with an incomplete learning set. 7th International Conference on Pattern Recognition, Montreal, August 1984.

Vannoorenberghe P., Wachowski K., Barsoum B., Postaire J.G., Multilevel Thresholding for image segmentation. Third International Conference on Computer Graphics and Image Processing, Spale Poland, 1994.

Vasseur C., Postaire J.G. A convexity testing method for cluster analysis. IEEE Trans. Syst. Man Cybern., Vol. 10, pp. 145-149, 1980.

Vassiliadis C.A., Twelve learning algorithms. Proc. 22nd Southeasten Symposium on System Theory, Cookeville, pp. 449-454, 1990.

Venkatasubramanian V., Vaidyanathan R., Diagnosing noisy process data using neural networks. IFAC Symposium SAFEPROCESS' 91, Baden-Baden, pp. 375-378, 1991.

Ward J.H., Hierarchical grouping to optimize an objective function. J. Am. Statist. Ass., Vol. 58, pp. 236-244, 1963.

Wharton S.W., A generalized histogram clustering scheme for multidimensional image data, Pattern Recognition, Vol. 16, n° 2, pp. 193-199, 1983.

Windham M.P., Analysis of Continuous Assignment Clustering Results. In Classification and Related Methods of Data Analysis, Bock H.H. (Ed.), Elsevier Science Publishers, pp. 173-178, 1987.

Zhang R.D., Botte-Lecocq C., Postaire J.G., Mode boundary extraction by binary morphology for cluster analysis, 3rd conf. of Classif. Edinburgh, Scotland, 1993.

Zhang R.D., Postaire J.G., Transformation morphologique des fonctions de densité de probabilité. Application à la détection des contours des modes en classification automatique, RFIA 1994, Paris.

Lakshminarasimhan A.L., Dasarathy B.U., A unified approach to feature selection and learning in unsupervised environments, IEEE Transaction on Computers, Vol. C-20, pp. 948-952, 1975.

Lambert J.M., Williams W.T., Multivariate methods in plant ecology, IV. Nodal Analysis, J. Ecol., Vol. 50, pp. 775-802, 1962.

Lambert J.M., Williams W.T., Multivariate methods in plant ecology, IV. Comparaison of information analysis and association analysis, J. Ecol., Vol. 54, pp. 635-664, 1966.

Lance G.N., Williams W.T., Computer programs for hierarchical polythetic classification. Comp. J., Vol. 9, pp. 60-64, 1966.

Lau C., Neural Networks. Theoretical foundation and Analysis, IEEE press, 1992.

Leonard M.S., Kilpatrick K.E., A minimum cost feature selection algorithm for binary-valued features, IEEE Systems, Man and Cybernatics, Vol. SMC-4, n° 6, pp. 536-542, 1974.

Linné C., Carl Linnearus., Species Plantarum, a fascimile of the fisrt edition 1753, Vol. 2, J. Ray Society London, rééd. 1959.

Ling R.F., Cluster Analysis, Unpublished Ph. D. Thesis, pp. 126, 1971.

Mac Naughton-Smith P., Some Statistical and other numerical techniques for classifying individuals. Home office Research Unit Report n° 6, London : H.M.S.O. 1965.

Mac Queen J., Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp., Vol. 1, pp. 281-297, 1967.

Masson M.H., Contribution à l'élaboration d'une méthode de décision avec rejet par réseaux de neurones. Application au diagnostic de systèmes. Thèse de l'Université de Technologie de Compiègne, 1992.

Matheron G., Randon sets and integral geometry, John Wiley, New York, 1985.

Mc Quitty L.L., Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. Educ. Psychol. Measmt., Vol. 17, pp. 207-229, 1957.

Milligan G.W., Cooper M.C., An examination of procedures for determining the number of clusters in a data set. Psychometrika, Vol. 50, pp. 159-179, 1985.

Mizouguchi R., Shimura M., Non parametric learning without a teacher based on mode estimation, IEEE transaction on Computers, Vol. C-25, pp. 1109-1117, 1976.

Mourot G., Contribution au diagnostic des sytèmes industriels par reconnaissances des formes, Thèse de doctorat à l'Institut National Polytechnique de Lorraine, 1993.

Narendra P.M., A non parametric clustering scheme for landsat, Pattern Recognition, Vol. 9, pp. 207-215, 1977.

Ouladhaj Thamir., Asselin de Beauville J.P., Détection des contours d'une image numérique par analyse anti-modes d'un histogramme local construit à partir d'un indice de similarité. Proceeding of the International Meeting on Distance Analysis, Distancia 1992, Rennes, 22-26 Juin 1992.

Postaire J.G., Vasseur C., An approximate solution to normal mixture identification with application to unsupervised pattern classification. I.E.E.E Transaction Pattern Analysis and Machine Intelligence, Vol. PAMI-3, Vol 2, pp. 163-179, 1981.

Postaire J.G., Fonctions convexes et optimisation du processus de classification automatique, RAIRO Automat., Vol. 16, pp. 357-379, 1982.

Postaire J.G., Fonctions convexes et optimisation du processus de classification automatique, RAIRO Automat., Vol. 17, n° 1, pp. 39-59, 1983.

Postaire J.G., De l'image à la décision, analyse des images numériques et théorie de la décision, Dunod Informatique, 1987.

Postaire J.G., Touzani A., Mode boundary detection by relaxation for cluster analysis, Pattern Recognition, Vol. 22, n° 5, pp. 477-490, 1989.

Postaire J.G., Touzani A., La notion de contour des fonctions de densité de probabilité en classification automatique. I) Estimation, filtrage et détection des contours des fonctions de densité multivariables. APII, Vol. 23, n° 2, pp. 139-160, 1989.

Postaire J.G., Touzani A., A sample-based approximation of normal mixtures by mode boundary extraction for pattern classification, Pattern Recogn. Letters, Vol. 11, 1990.

Postaire J.G., Zhang R.D., Botte-Lecocq C., Application of binary morphology to cluster analysis, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 15, pp. 170-180, 1993.

Pudil P., Blaha S., Evaluation of the effectiveness of features selected by discriminant analysis methods, Pattern Recognition, Vol. 14, N° 1, pp. 81-85, 1981.

Raudys S.J., Jain A.K., Small Sample Size Effects in Statistical Pattern Recognition : Recommendations for Practitioners. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 13, n° 3, pp. 222-234, 1991.