Etude du codage et du classement de pièces mécaniques en technologie de groupe

Michel Costantini

▶ **To cite this version:**

Michel Costantini. Etude du codage et du classement de pièces mécaniques en technologie de groupe. Sciences de l'ingénieur [physics]. Université Paul Verlaine - Metz, 1987. Français. NNT : 1987METZ017S . tel-01775741

HAL Id: tel-01775741
https://hal.univ-lorraine.fr/tel-01775741

Submitted on 24 Apr 2018

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

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

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact : ddoc-theses-contact@univ-lorraine.fr

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

Année 1987

dre :

THESE

présentée à

La Faculté des Sciences de l'Université de METZ

en vue de l'obtention du

DIPLOME DE DOCTEUR 3e CYCLE

Option : Production Automatisée

par

M. COSTANTINI

ETUDE DU CODAGE ET DU CLASSEMENT DE PIECES MECANIQUES EN TECHNOLOGIE DE GROUPE

Soutenue le 24 septembre 1987 devant le jury composé de :

MM.	B.	MUTEL	Président
	C.	COUET	Professeur
	M.	POTIER-FERRY	Professeur
	J.M.	PROTH	Directeur de recherches à l'INRIA
	CH.	SAUVAIRE	Directeur de l'ADEPA

BIBLIOTHEQUE UNIVERSITAIRE DE METZ

022 373851 1

AVANT PROPOS

Ce travail a été réalisé au sein du Laboratoire d'Automatique et d'Electronique Industrielles (L.A.E.I.) de l'Université de Metz, sous la direction de Monsieur B. Mutel. Je lui exprime ma sincère gratitude pour les conseils et la bienveillance qu'il n'a cessé de me prodiguer au cours de ces années d'études.

Je suis très reconnaissant à Monsieur C. Couet, Professeur à l'Université de Reims, d'avoir accepté de juger ce travail en tant que membre du jury.

Je remercie vivement Monsieur Potier-Ferry, Professeur à l'Université de Metz, d'avoir bien voulu s'intéresser à ce travail et pour sa participation au jury.

J'exprime ma vive reconnaissance à Monsieur J.M. Proth, Directeur de recherches à l'INRIA, pour sa participation au jury, ainsi qu'à son équipe et en particulier Monsieur F. Bonneau, pour l'aide et le soutien qu'ils m'ont apportés par de nombreuses discussions fructueuses.

Je suis très honoré par la présence de Monsieur Ch. Sauvaire, Directeur de l'Adepa, qui a accepté de juger ce travail.

Je présente mes remerciements à tous mes camarades chercheurs et techniciens à l'Université de Metz pour leur aide, les échanges constructifs et l'atmosphère amicale qu'ils contribuent à faire régner.

SOMMAIRE

Pages

CHAPITRE I : INTRODUCTION 7

11 - Généralité 8

12 - Concept de technologie de groupe 8

13 - Mise en oeuvre de la technologie de groupe 12

14 - Implantation de la technologie de groupe 15

15 - Limitation des méthodes actuelles de classification 18

16 - Classification par mesures de proximités 20

17 - Objectifs de cette étude 24

18 - Plan du mémoire 26

Pages

CHAPITRE II : CODAGE DES DONNEES DE PRODUCTION ET METHODES D'ANALYSE DISCRIMINANTE 27

21 - Codage des données 28

211 - Introduction 28

212 - Les échelles de mesures 30

213 - Homogénéisation des variables 32

22 - Problème de classement en technologie de groupe 33

23 - Méthodes de discrimination 35

231 - Introduction 35

232 - Les méthodes 36

Pages

CHAPITRE III : ETUDE D'UNE METHODE DE CLASSEMENT D'UN NOUVEAU PRODUIT ET DE SIMPLIFICATION DU CODE DE REPRESENTATION 40

31 - Méthode de classement 41

311 - Modélisation 41

312 - Principe de la méthode 41

313 - Définitions 42

314 - Traitement initial 44

3141 - Partitionnement en régions 44

a - Introduction 44

b - Algorithme de partitionnement 46

c - Description de l'algorithme 46

3142 - Modélisation des régions 49

a - Introduction 49

b - Algorithme de reconnaissance des régions 51

315 - Affectation de nouveaux individus 52

Pages

3151 - Principe 52

3152 - Algorithme d'affectation 53

3153 - Procédure des k plus proches voisins 59

a - Principe 59

b - Algorithme de recherche des k plus proches voisins 61

c - Description de l'algorithme 62

32 - Simplification du code 62

321 - But de la simplification 62

322 - Elimination des variables 63

3221 - Variables qualitatives 63

a - Filtrage des modalités 63

b - Modalités peu significatives 63

3222 - Variables quantitatives 68

a - Corrélations 69

b - Variances 69

33 - Conclusion 71

Pages

CHAPITRE IV : APPLICATIONS 72

41 - Structure générale du logiciel 73

42 - Exemple académique 75

43 - Exemples industriels 76

431 - Exemple 1 76

432 - Exemple 2 94

CHAPITRE V : CONCLUSION 112

BIBLIOGRAPHIE 117

ANNEXE 125

1 - Différentes méthodes de discrimination 126

11 - Méthode bayésienne 126

12 - Méthode de Sebestyen 128

13 - Discrimination par voisinage 130

14 - Méthodes de réduction
Elimination de variables 131

141 - But des méthodes 131

142 - Méthode de pas à pas 132

Pages

1421 - Critère du pourcentage de bien classé 132

1422 - Critère trace ($V^{-1}B$) 133

143 - Analyse factorielle discriminante 133

144 - Discrimination par des méthodes de régression linéaire 135

15 - Méthodes d'analyse discriminante sur variables qualitatives 135

151 - Méthodes de réduction des variables 135

1511 - Méthode séquentielle des corrélations canoniques 136

1512 - Méthode des coefficients de Tschuprow 136

1513 - Méthode non paramétrique 137

1514 - Méthode par segmentation 138

152 - Méthode d'apprentissage de descriptions structurelles complexes 140

CHAPITRE I

INTRODUCTION

I INTRODUCTION

11) Généralités

Pour faire face aux évolutions des marchés des produits, les entreprises doivent intensifier les mesures de rationalisation à tous les échelons du processus de production. Ceci permet d'assurer une production compétitive malgré l'augmentation constante des coûts et de la concurrence. Le phénomène est particulièrement sensible dans les entreprises fabricant à l'unité ou en petites séries.

Les entreprises doivent allier des nécessités de productivité accrue, de flexibilité, d'un niveau de qualité supérieur et de réduction des stocks et des en-cours. On ne peut atteindre ces objectifs qu'avec une réorganisation globale de l'entreprise.

Dans ce contexte, le processus de production intégrée implique l'ensemble des activités ayant trait à l'élaboration du produit, depuis l'étude, la préparation du travail, la fabrication et jusqu'au montage.

Les mesures de rationalisation n'auront un effet que limité si elles ne concernent qu'un seul domaine de la production. La compétitivité de l'entreprise dépend de la bonne répartition des mesures de rationalisation dans tous les services de la production (fig. I_1).

L'utilisation du concept de la technologie de groupe permet d'améliorer de façon sensible la compétitivité des entreprises [17].

12) Concept de technologie de groupe

La technologie de groupe est un concept qui a pour but d'identifier et de rassembler des produits identiques ou similaires (fig. I_2) aux différents stades de leurs traitements dans l'entreprise (fig. I_{3a}-I_{3b}).

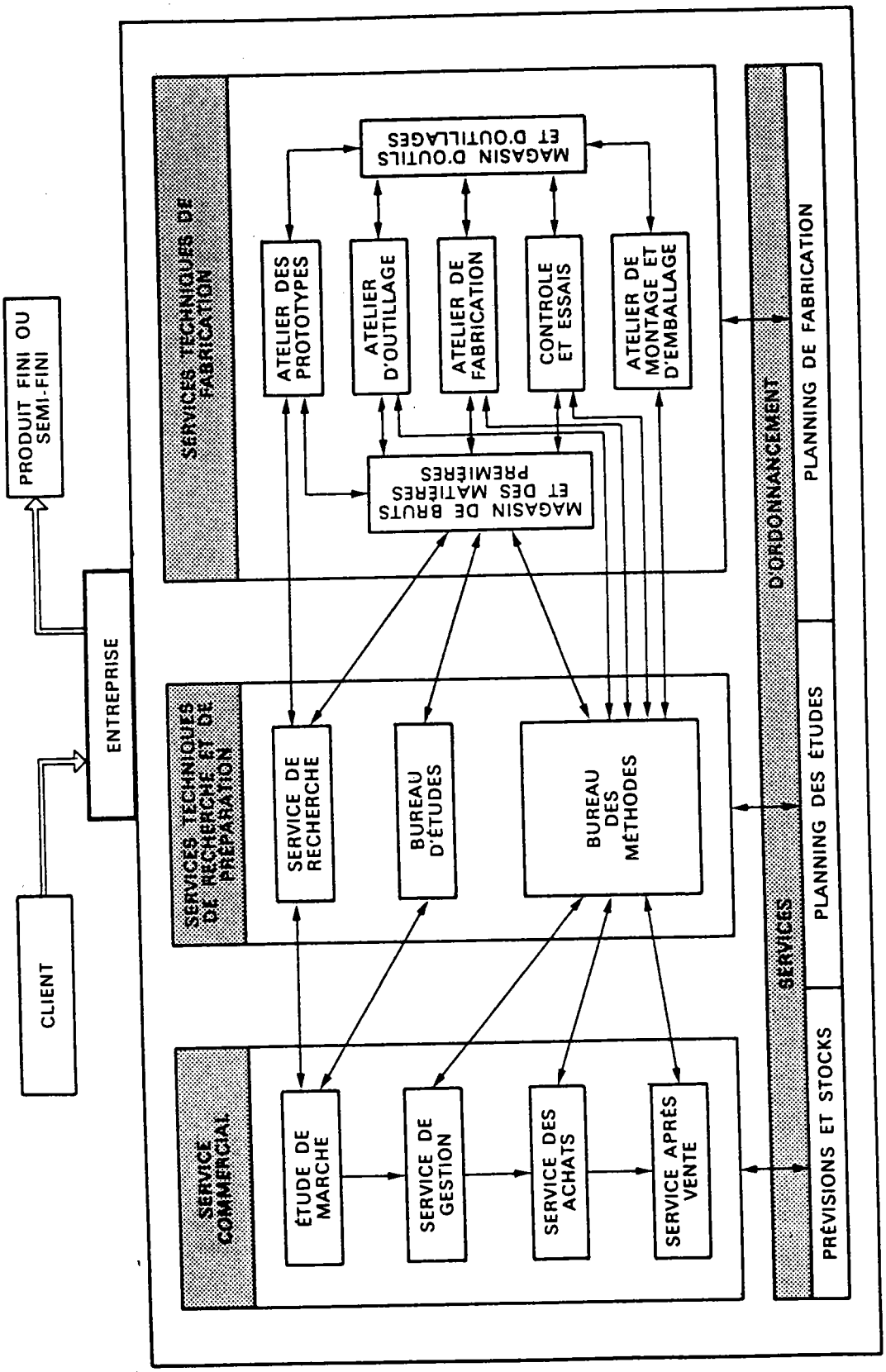


figure I, : Organigramme d'une entreprise [50].

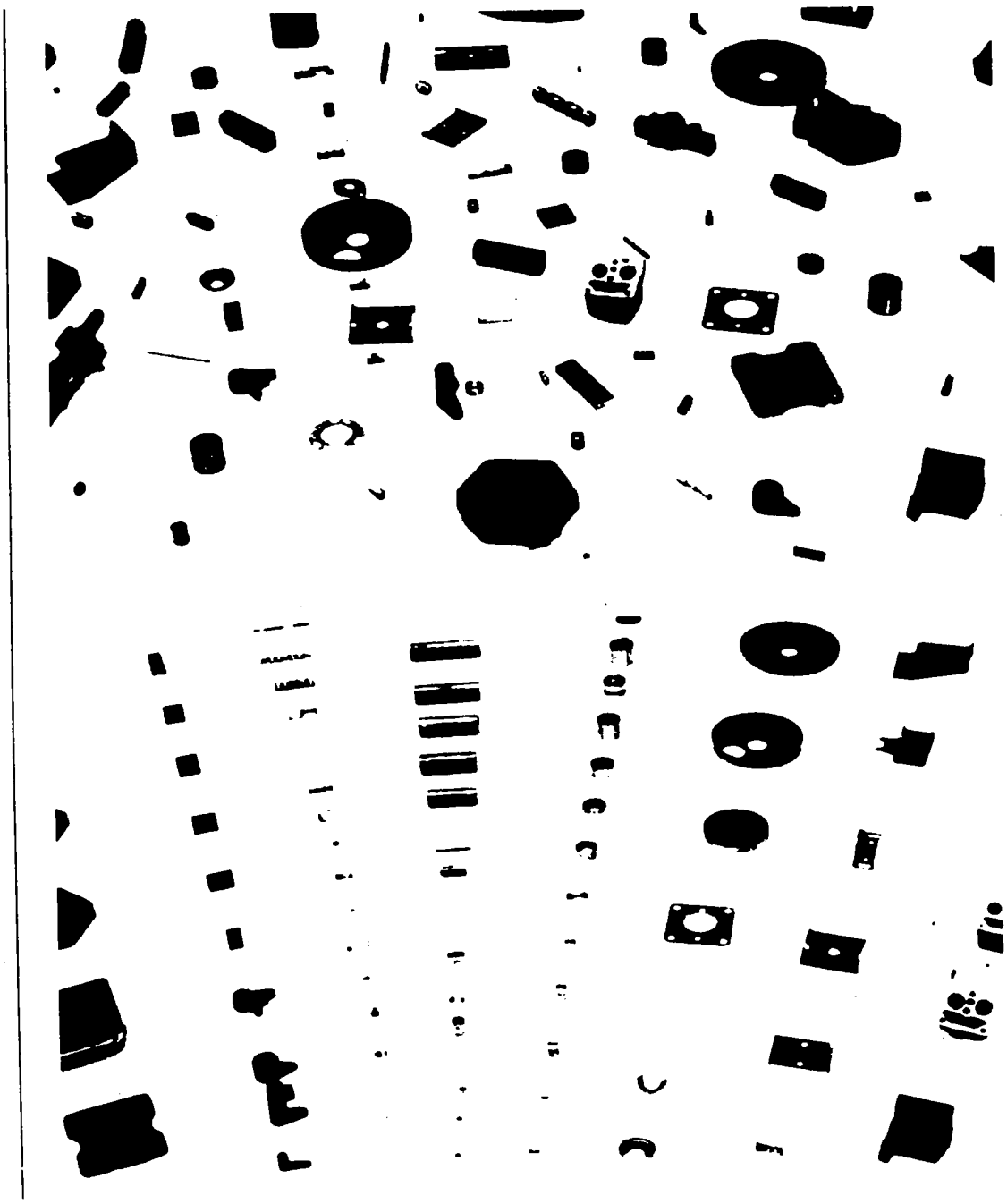


figure I_2 : Exemple de pièces hétérogènes et de leur regroupement en familles [49].

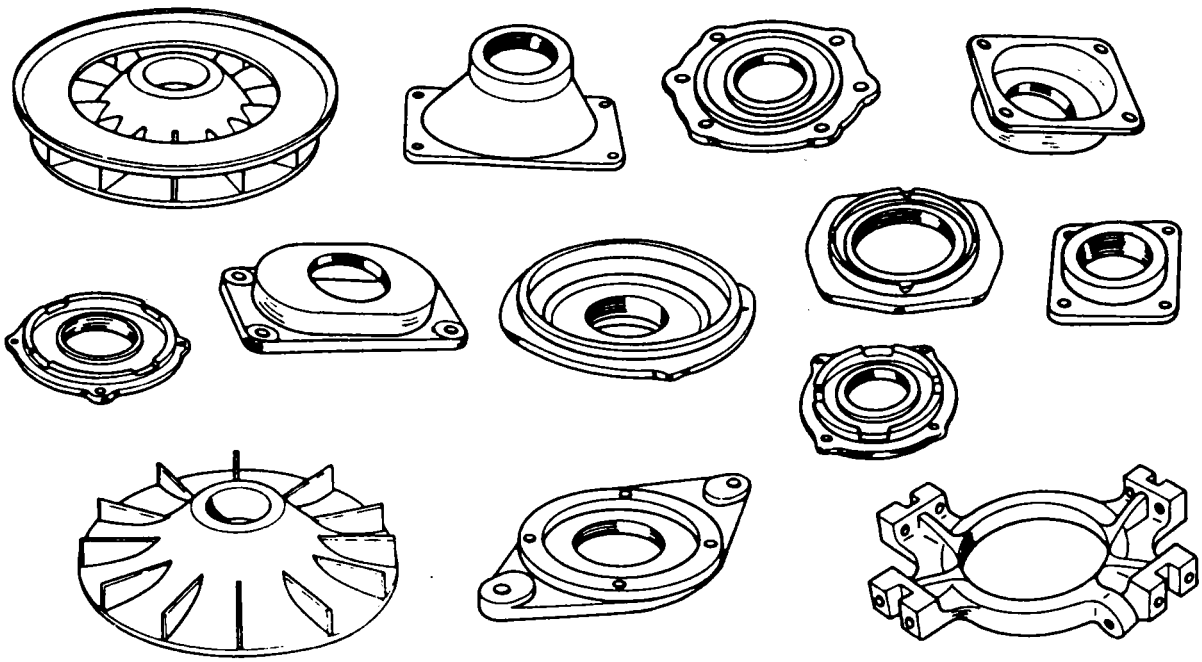
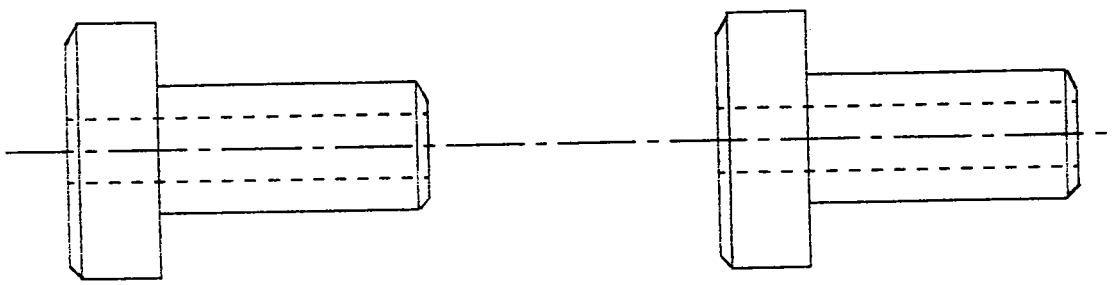


figure I,a : Pièces de conception différentes mais au processus de fabrication identique.



Tolérance : ± 0,15
Nombre : 200 000 / an
Matière : XC 38

Tolérance : ± 0,001
Nombre : 200 / an
Matière : 25 CD 4

figure I,b : Pièces de forme et de taille identiques mais de fabrication différente.

Les produits ayant des formes ou des opérations de fabrication semblables sont rassemblés en familles. Ceci permet, dans un premier temps, de tirer profit de leurs similitudes en créant des séries plus importantes, et donc de rapprocher leurs techniques de fabrication à celles des grandes séries. Dans un deuxième temps, la technologie de groupe permet une réorganisation différente des moyens de production en îlots ou cellules de fabrication (fig. I_4). Les produits d'une même familles sont fabriqués complètement dans un des îlots, ceux-ci étant très peu liés les uns aux autres. Ceci conduit ainsi à réduire les en-cours, les manutentions et facilite la gestion et le suivi de production.

Une organisation en technologie de groupe induit ainsi une standardisation des produits, des éléments de forme, des outillages et des procédés de fabrication. Ceci évite de créer un produit et de définir de nouveaux processus de fabrication alors qu'il en existe déjà d'identiques ou de similaires.

De même pour une organisation en îlots la gestion de production est décentralisée sur chaque cellule et permet une plus grande souplesse de conduite et de suivi.

13) Mise en oeuvre de la technologie de groupe

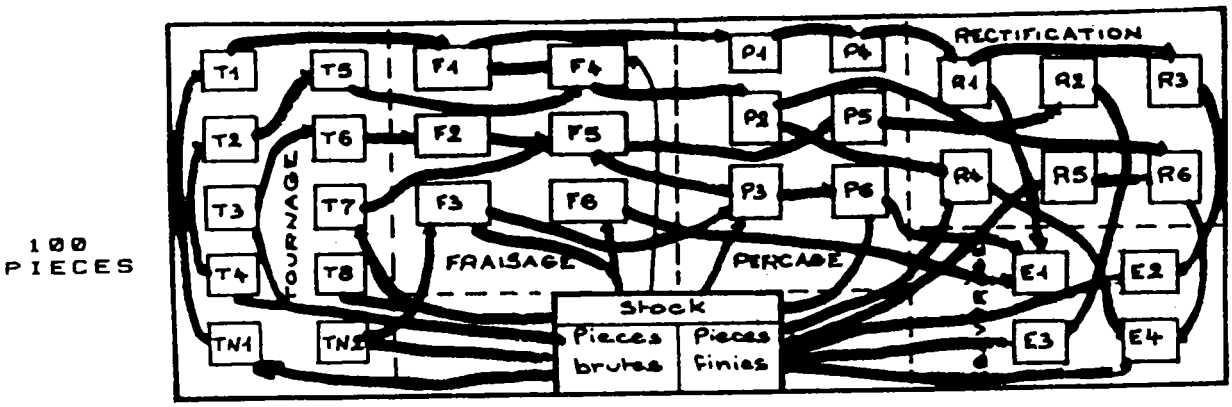
Un système de production est composé d'un nombre important d'informations techniques et économiques de natures diverses, inter-agissant entre elles de manière multiple, ce qui en fait un système complexe E [25].

Une solution, qui permet de contrôler un tel système, est de le décomposer en sous-ensembles d'homogènes et indépendants. La maitrise du système dans son intégralité en est ainsi simplifiée par cette décomposition [33] qui divise la complexité du système étudié.

Nous proposons une approche (fig.I_5) de la technologie de groupe qui consiste, dans une première étape d'apprentissage, à étudier le

AVANT IMPLANTATION
DE LA T.G

ATELIER CONVENTIONNEL
EN SECTIONS HOMOGENES



APRES IMPLANTATION
DE LA T.G.

ATELIER ORGANISE EN ILOTS
DE FABRICATION

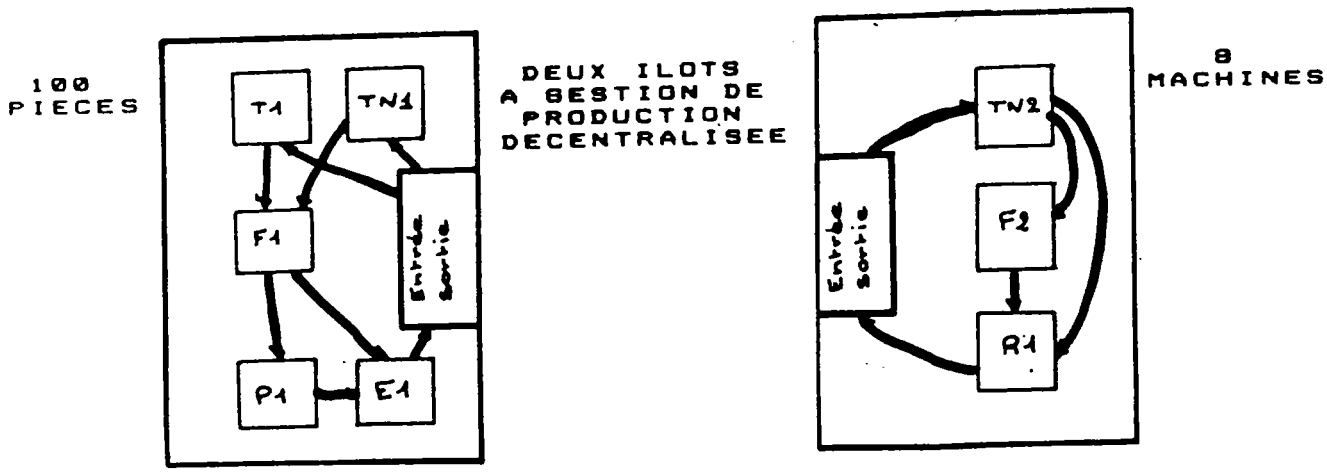


figure I$_4$: Exemple d'organisation en ilots de fabrication.

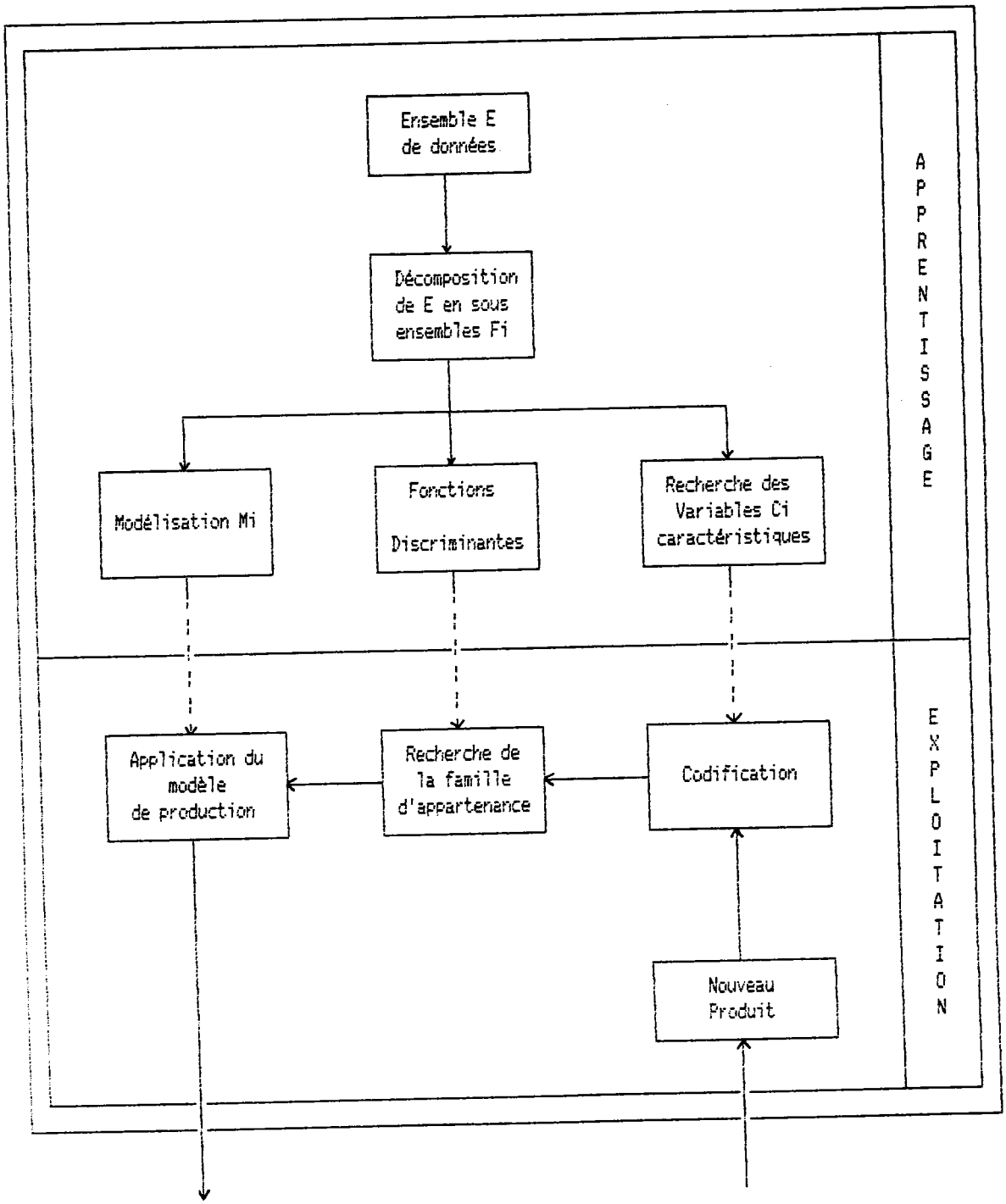


figure I_5 : Approche Technologie De Groupe.

système de production E défini sur un horizon T. Nous décomposons alors ce système en sous-ensembles (ou familles) indépendants Fi, dans lesquels les produits sont très fortement corrélés entre eux. Pour chacun de ces sous-ensemble Fi, nous recherchons alors un modèle de production Mi, une fonction discriminante pour affecter de nouveaux produits à ces familles, et les variables caractéristiques Ci les plus représentatives de chacune des familles. Cette première étape implique une analyse, une classification et une modélisation des données de production. A cette fin, il est utilisé des méthodes descriptives d'analyse de données [28-30] qui permettent de vérifier la pertinence des données initiales E de représentation du système de production. Puis il est utilisé des méthodes d'analyse typologique et de classification automatique pour reconnaître les familles.

Dans une seconde étape, l'exploitation, le système de production étant organisé en familles de produits, une fonction discriminante affecte un nouveau produit, défini seulement par quelques caractéristiques éxogènes, à une famille. Nous pouvons alors appliquer au nouveau produit, les lois de production correspendant au modèle de production Mi de la famille Fi.

Cette approche, basée sur le concept des groupements analogiques, peut en fait s'appliquer à divers services dans l'entreprise (gamme de fabrication, outillage, assemblage, calcul de devis, éléments de formes, etc...)

14) Implantation de la technologie de groupe

Lors de l'implantation de la technologie de groupe dans une entreprise, les différentes phases de l'approche, décrite ci-dessous, sont les suivantes (fig.I$_6$) :

a) Devant la multitude de données de production, nous choisissons un échantillon test E de la population qui répond au spectre de

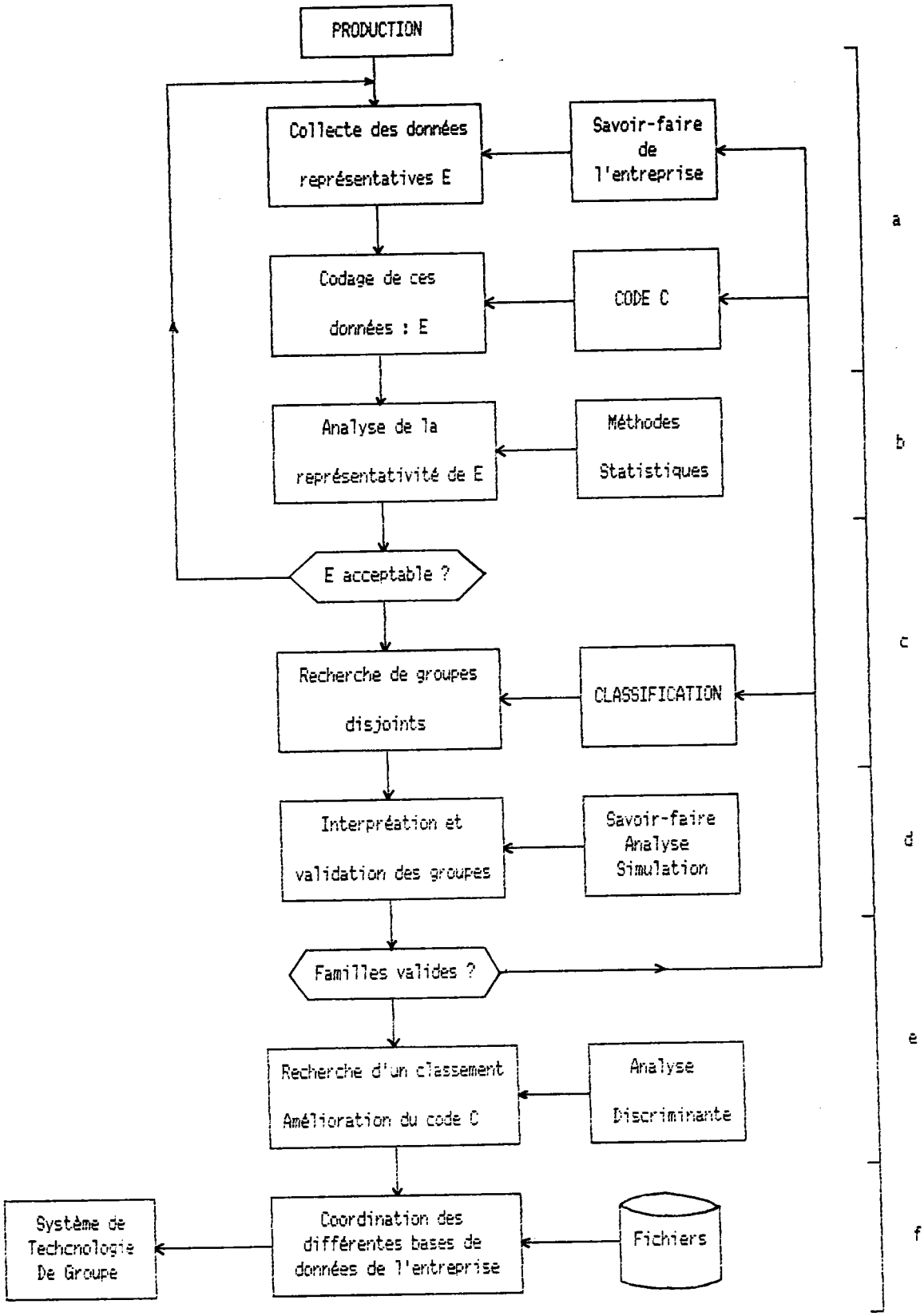


figure I_6 : Différents stades de l'implantation de la Technologie de Groupe dans une entreprise.

produits les plus représentatifs de la production. Nous codons ensuite l'échantillon de la population à l'aide d'un code existant, ou le plus adapté à l'application que nous voulons réaliser [29].

b) A cette étape, nous effectuons une analyse statistique de l'échantillon E, afin de vérifier sa représentativité ainsi que celle des variables du code. Nous utilisons des méthodes telles que l'analyse factorielle de correspondances [21-26]. Sur l'ensemble E, cette méthode permet en particulier d'avoir une représentation graphique simultanée de la distribution des produits et de leurs variables représentatives.

c) Si nous estimons que le mode de représentation défini par les variables et que l'échantillon E est convenablement choisi, nous procédons alors à la recherche d'une partition par des méthodes de classification [14-22-25-27-28-33-34].

d) Les familles étant obtenues, nous devons alors valider et interpréter les classes de la partition. Il est également intéressant de tester la stabilité de celles-ci en fonction des variations de la production par des simulations tenant compte de l'évolution des commandes.

e) Les classes ou familles étant définitivement validées, nous recherchons du point de vue de l'utilisateur final un code simple, nécessaire et suffisant pour réaliser le classement de nouveaux produits. Ceci permet d'accélérer le codage et de faire ressortir les principales caractéristiques des différentes classes.
Cette étape de classement et d'amélioration de code correspond à l'analyse discriminante [12-31].

f) La classification et le classement étant acceptés, nous devons étudier la coordination de la base de données de l'entreprise.

15) Limitation des méthodes actuelles de classification

Plusieurs systèmes de classification ont été créés depuis plusieurs année, afin de déterminer des familles de produits [32-35-36-37]. La plupart de ces systèmes utilisent simplement des techniques de tris sur les valeurs numériques du code de représentation des produits.

Cette technique de recherche de groupements technologiques par tris présente des limitations :

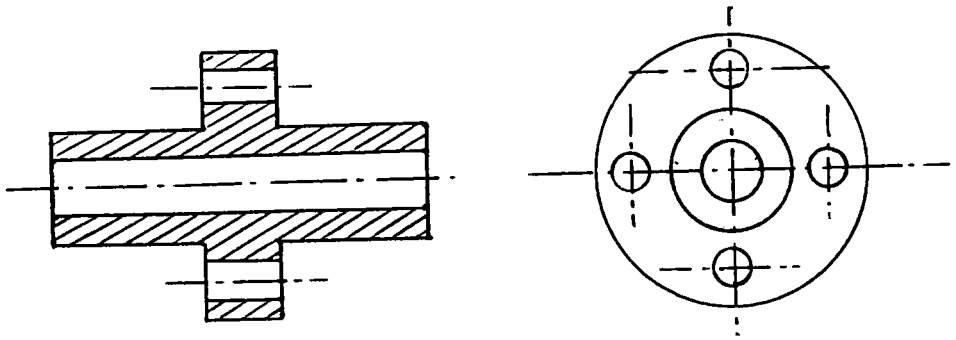
- Les familles formées dépendent uniquement du choix des valeurs des clés de tri. Le résultat dépend donc essentiellement de l'expérience de l'utilisateur ;

- Cette technique ne permet des regroupements que de produits très proches, au sens des valeurs numériques des clés sélectionnées. Avec cette méthode, deux produits sont dans la même classe si et seulement si toutes les clés sélectionnées par l'utilisateur ont des valeurs identiques. Cela est très contraignant. En effet dans certains codes, deux produits peuvent être très ressemblant tout en ayant des codes assez différents.

Le processus de tris direct donne des résultats satisfaisants lorsque le code est bien adapté aux objectifs de la classification, et lorsque les variantes d'un produit standard d'une famille recherchée possède sur ce code des variables dont les valeurs ont une distribution gaussienne.

Ainsi les logiciels de classification procédant par tris sont très bien adaptés aux familles monoproduits avec variantes simples (fig. I$_7$).

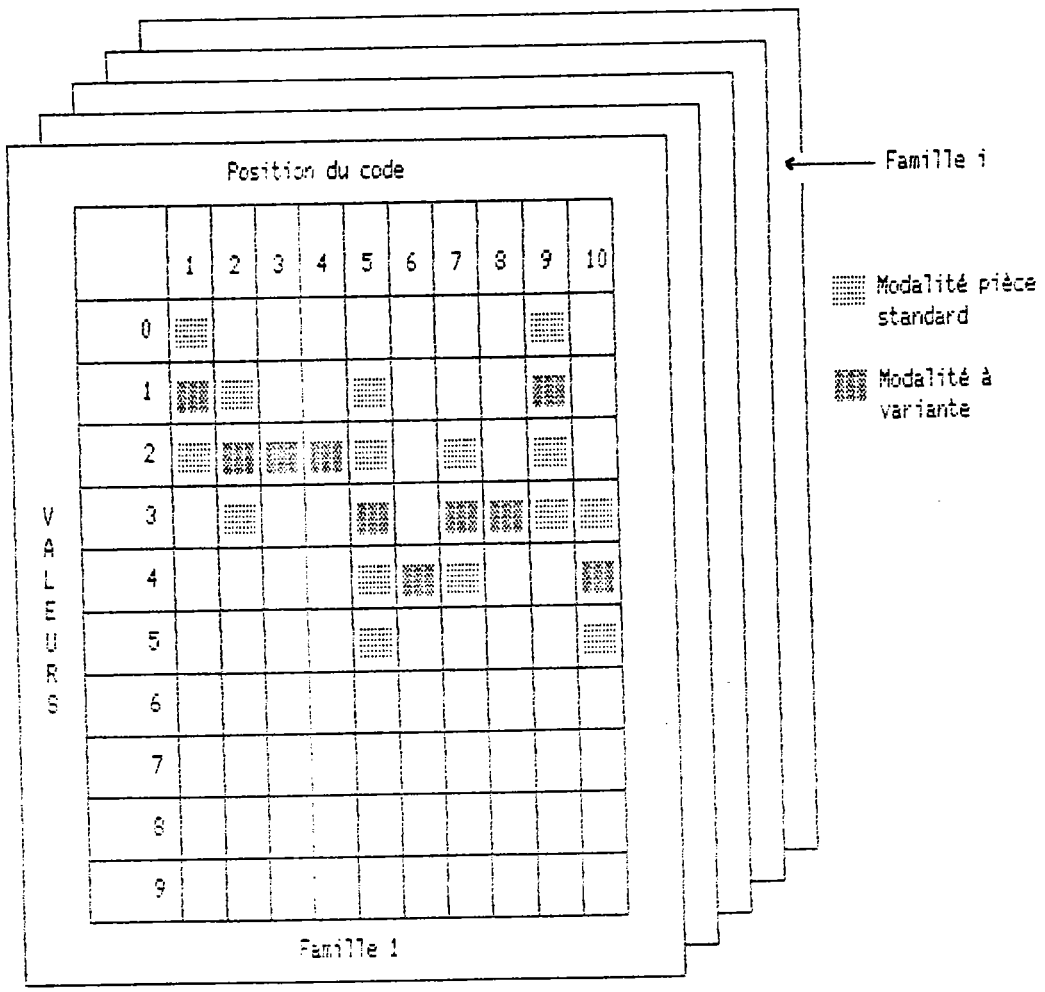
La tendance actuelle est à la recherche de l'augmentation de la flexibilité des systèmes de production en élargissant le spectre des produits appartenant à une même famille.



Nouvelle pièce

Code : 1 2 2 2 5 4 3 3 0 4

Recherche de la matrice englobant la nouvelle pièce.



Application à la nouvelle pièce des lois de production correspondant à la famille matrice trouvée.

figure I_3 : Exemple de classement d'une pièce à une famille monoproduits avec variantes.

Du fait de ses limitations, les méthodes de tris sont mal adaptées aux familles possèdant un large spectre de produits. Il est alors intéressant d'utiliser des techniques employant la notion de mesures de proximité entre produits basées sur la classification automatique, etc...

Les références [12-14-19-20-21-22-38] montrent quelques applications à la technologie de groupe de méthodes d'analyse typologique de données.

16) Classification par mesure de proximité

Par principe, la classification par mesure de proximité [23-34] consiste à définir une mesure de ressemblance (nombre scalaire) entre produits à comparer. Cette mesure de ressemblance est appelée indice de similarité s_{ij} ou indice de dissimilarité d_{ij}. Ces indices sont calculés en fonction des valeurs des paramètres physiques caractérisant les produits.

A priori, nous ne pouvons faire d'hypothèses statistiques sur les données de production, nous écarterons donc des méthodes de classification en analyse de données qui respectent des lois de distribution statistique. Nous utilisons ainsi des approches basées sur la minimisation de critères utilisateurs (classification hièrarchique, méthodes de réallocation,...).

Les méthodes hiérarchiques recherchent une famille de partitions telle que les regroupements ou les divisions successifs forment un arbre hièrarchique. Les méthodes hièrarchiques sont telles que l'affectation d'un produit à une classe à une étape n'est jamais remise en cause dans les étapes ultérieures. De plus, le nombre de calculs peut être important lorsque le nombre de produits à analyser est élevé. Pour chaque étape, les indices de ressemblance sont en effet calculés deux à deux entre tous les produits [39].

Dans les méthodes de réallocation, l'affectation d'un produit à une classe, obtenue à une étape, est remise en cause dans les étapes ultérieures. La figure I₈ montre l'algorithme général des méthodes de réallocation [23]. Dans ces méthodes, le nombre de classe est généralement fixé au départ et le nombre de calculs est moins important que dans la précédante, car les similarités sont calculées entre chaque produit (individu) et les représentants de chaque classe [40-41-42-43-44]. Ces méthodes recherchent des classes homogènes et bien séparées de produits, mais la solution obtenue dépend de la partition intitiale.

Quelque soit la méthode utilisée, son efficacité résulte principalement du choix judicieux de la mesure de proximité entre les produits. En général, les mesures les plus employées en analyse de données ne sont pas ou peu adaptées aux données de production. La figure I₉ [15] montre différentes hiérarchies obtenues sur des données réelles de production avec diverses mesures. Ces donnée représentent des gammes de fabrication caractérisées par une succession de phases d'usinage. Chaque gamme est représentée par un vecteur ligne x_i (Ø1ØØ...x_{ij}...1Ø) telque x_{ij} = 1 si la gamme x_i possède la phase j sinon x_{ij} = Ø. Les mesures essayées sont les suivantes :

- indice de Russel et Rao : P/T
- indice de Dice : 2P/2P+N
- indice de Jaccard : P/P+N
- indice de Kulczynski : P/N
- indice de Sokal et Sneath : P/P+2N.

avec P : nombre de coprésences.

N : nombre de non coïncidences.

T : nombre total de phases.

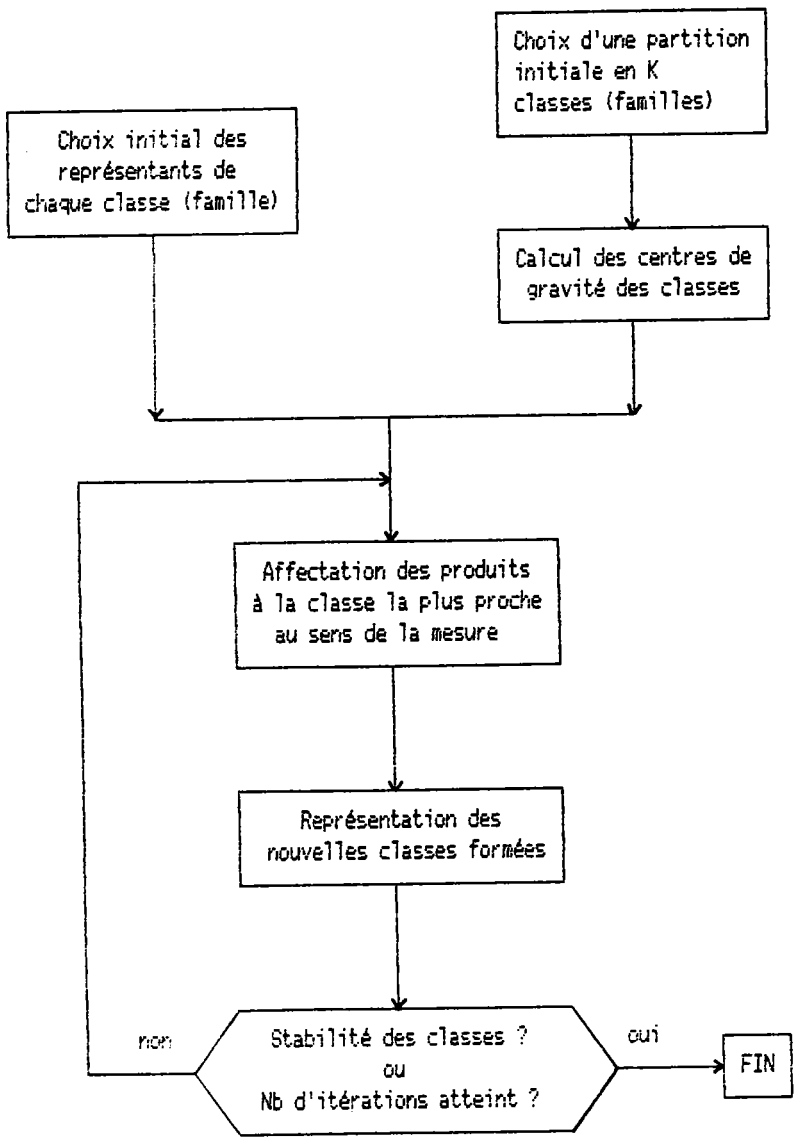
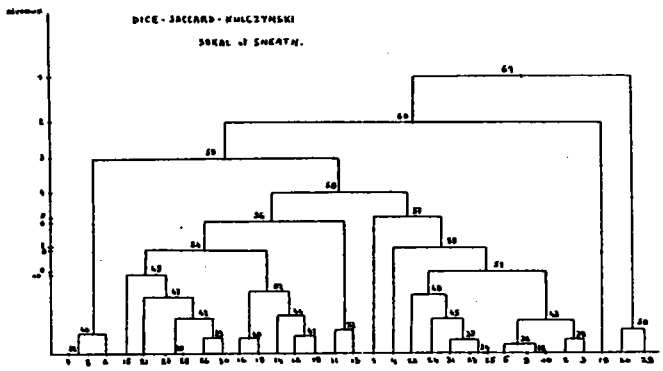
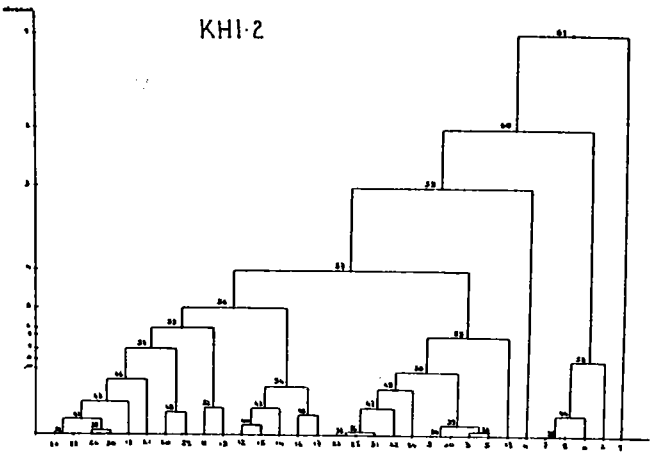
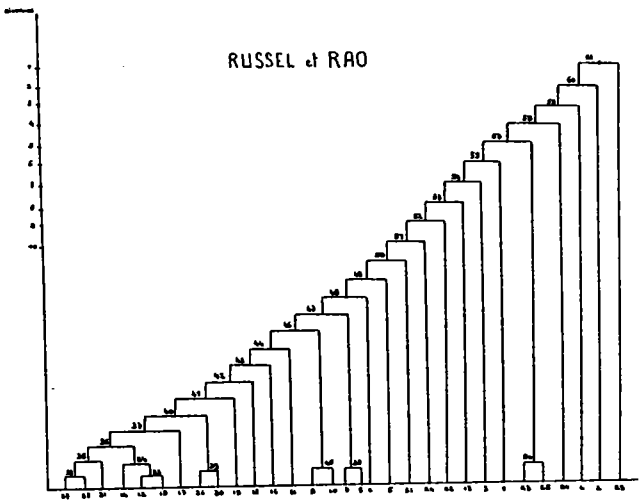


figure I_6 : Algorithme général des méthodes de réallocation.







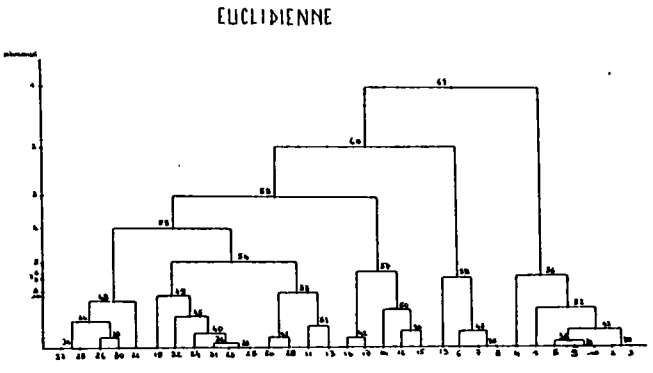


figure I_9 : Exemple de dendogrammes obtenus par une classification hiérarchique sur des gammes d'usinage avec différentes mesures de distance.

- distance euclidienne :

$$d^2(x_i, x_k) = \sum_{j=1}^{p} (x_{ij} - x_{kj})^2$$

- distance du Khi2 :

$$d^2(x_i, x_k) = \sum_{j=1}^{p} \frac{1}{a_{.j}} \left(\frac{x_{ij}}{a_{i.}} - \frac{x_{kj}}{a_{k.}} \right)^2$$

avec $a_{.j} = \sum_{i=1}^{n} X_{ij}$, $a_{i.} = \sum_{j=1}^{p} x_{ij}$, $a_{k.} = \sum_{k=1}^{p} x_{kj}$

Cet exemple de classification hiérarchique sur des gammes d'usinage montre bien l'importance du choix d'une mesure, du fait des différents dendogrammes obtenus sur les mêmes données de production. En coupant chaque dendogramme par une ligne horizontale à un même niveau, nous obtenons un partitionnement différent pour chaque mesure de proximité.

Le choix d'un indice de proximité est fonction de la nature des variables caractéristiques (code) représentant les produits. Ces variables dépendent de l'objectif à atteindre, c'est-à-dire rechercher des groupements de codes morphodimensionnels des gammes d'usinage, des îlots de fabrication, etc...

17) Objectifs de cette étude

On peut montrer que pour un code donné (variables caractéristiques), une partition optimale peut être trouvée, et pour une partition donnée un code optimal. Partitions et codes sont donc liés très étroitment et il faudrait effectuer des itérations sur l'un et l'autre pour

atteindre un optimum global. Ceci est irréalisable car trop long et trop coûteux. Aussi en pratique, on choisit un code de base C_B qui a déjà prouvé son efficacité dans l'application envisagée. A ce code C_B, on associe un code spécifique C_C qui est censé tenir compte des particularités des produits en fonction de l'objectif de la classification fixée. Le choix de ce code est heuristique, ainsi certaines variables pourront être redondantes et d'autres sans intérêt pour la classification recherchée.

De cette façon, chaque produit est représenté par un code C_B+C_C comportant grand nombre de variables. Au stade de la recherche des classes, il est préférable d'avoir trop d'informations que pas assez. Les variables non caractéristiques des familles introduisent simplement un "bruit de fond". Par contre, les familles de produits étant identifiées, du point de vue de l'utilisateur, l'emploi de codes plus simples est préférable pour plusieurs raisons. Ainsi, d'une part l'opération de codage de nouveaux produits sera plus rapide et les risques d'erreurs moindres. D'autre part, les variables retenues caractériseront véritablement chaque famille, les modèles de représentation Mi seront plus simples à déterminer et la compréhension du processus de production améliorée.

Les objectifs que nous nous fixons dans cette étude sont premièrement de trouver une méthode de classement efficace compte tenu de la nature des données de production, deuxièmement de rechercher un code "optimal", qui pour chaque application (gammes, outillages, devis, assemblages...) aura le rôle de :

- représenter au mieux le produit dans sa famille,
- activer la règle de classement de nouveaux produits,
- faciliter la recherche de modèles de production.

Ces objectifs relèvent principalement des méthodes d'analyse discriminante.

18) Plan du mémoire

Dans le chapitre suivant, nous définissons les problèmes rencontrés par le classement de produits et leurs codes de représentation. Nous analysons ensuite les différentes méthodes de discrimination éxistantes, ce qui nous conduit à proposer une nouvelle méthode mieux adaptée aux problèmes de classement de données de production.

Dans le troisième chapitre, nous présentons cette nouvelle méthode de classement. Celle-ci est basée sur un partitionnement de la production E dans un espace Euclidien, afin de trouver des zones homogènes caractéristiques des familles et une zone hétérogène trouble. Puis nous proposons une méthode d'amélioration du code qui ne retient que les variables les plus représentatives des produits aux familles.

Le chapitre quatre présente des applications de la méthode proposée à des données de fabrication mécanique. Un premier exemple traite du classement de pièces mécaniques de révolution représentées par le code OPITZ, tandis que le second exemple traite des pièces plus diverses codées à l'aide de Multi-M.

Dans la conclusion, nous analysons de façon critique les performances de la méthode étudiée et nous proposons des développements de celle-ci.

En annexes, nous présentons succinctement le principe des principales méthodes actuelles d'analyse discriminante.

CHAPITRE II

CODAGE DES DONNEES DE PRODUCTION ET METHODES D'ANALYSE DISCRIMINANTE

II CODAGE DES DONNEES DE PRODUCTION ET METHODES D'ANALYSE DISCRIMINANTE

21) Codage des données

211) Introduction

Le codage des données de production, c'est-à-dire la représentation des caractéristiques physiques, technologiques, économiques des données en être mathématique sans perte ni déformation de l'information est un problème général non résolu de façon satisfaisante.

Depuis de nombreuses années, on a essayé de déterminer les paramètres pouvant caractériser les produits à fabriquer ainsi que les moyens de fabrication. Ceci a donné lieu à la création de nombreux codes en fabrication mécanique dont quelques uns sont présentés à la figure II1. D'une manière générale, les paramètres caractérisant les produits dépendent des objectifs de la classification envisagée : standardisation des formes, familles de gammes de fabrication, d'outillages, devis, assemblages, réorganisation en îlots, etc...

A chaque objectif de la classification correspond un code particulier, et le code complet de représentation d'un produit sera l'union de ces différents codes associés à chaque objectif.

Le nombre de variables caractérisant un produit, doit trouver un juste milieu car :

- un faible nombre rendra plus aisé le codage et l'interprétation du code. Toutefois si ce nombre est trop limité, la représentation du produit sera trop imprécise et grossière et ne permettra pas de créer des familles

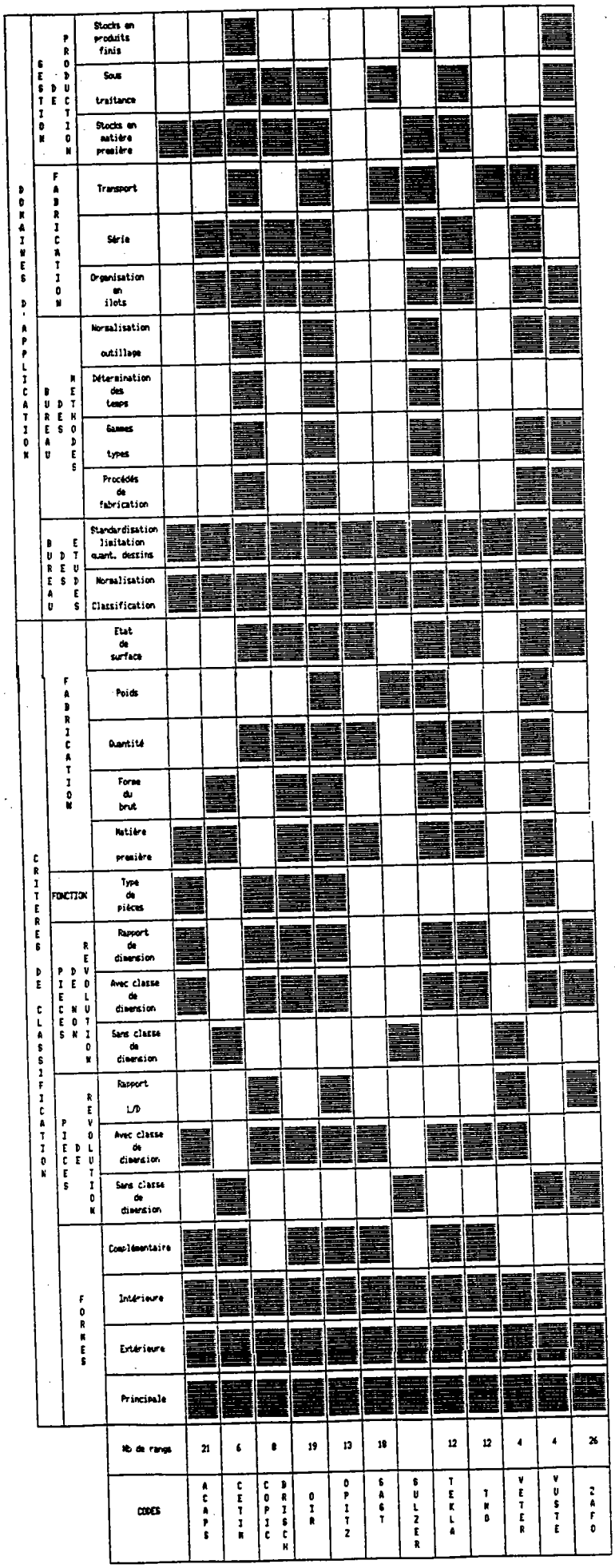


fig II$_1$: Tableau des codes de représentation des pièces mécaniques et de leurs domaines d'application.

intéressantes ;

- par contre un nombre important de paramètres risque de dissoudre les caractéristiques principales dans un amoncellement de caractéristiques secondaires qui va tendre à rassembler des produits.

En pratique, les codes industriels comportent de 7 à 31 variables comme indiqué figure II_1. Ces variables représentent différentes grandeurs physiques qui ont des qualités propres. Les variables ont ainsi des propriétés différentes.

212) Les échelles de mesure

Les variables des codes utilisées sont le plus souvent hétérogènes et appartiennent à plusieurs échelles de mesures (fig.II_2) selon les entités physiques auxquelles elles se rapportent.

- Echelles nominale :

La variable est représentée par une suite de modalités. Elle permet de séparer la population en diverses classes, pour lesquelles la variable garde la même valeur. Les modalités utilisées représentent divers états ou évènements possibles ;

- Echelle ordinale :

La variable est représentée par une suite de modalités respectant un ordre. Cette échelle est identique à la précédante mais une relation d'ordre est introduite entre les modalités de la variable, précisant ainsi

Type d'échelle	NOMINALE	ORDINALES	D'INTERVALLE	METRIQUE
Opérations mathématiques	- Tableau de fréquences - Comptage d'éléments	- Médiane - Quartile - Ecart type	- Moyenne - Variance - Ecart type	Toutes opérations mathématiques
Caractéristiques	- Qualitative discrète	- Qualitative discrète	- Quantitative discrète ou continue - Zéro arbitraire	- Quantitative discrète ou continue - Zéro arbitraire
Relations mathématiques	- Equivalence entre les membres d'une même classe	- Equivalence entre les membres d'un même rang - Ordre - Préordre	- Equivalence entre les objets ayant la même valeur. - Ordre - Rapport entre deux intervalles	- Equivalence - Ordre - Rapport entre deux intervalles - Rapport entre deux valeurs
Matrices de données associées	- D'occurence - De fréquences - Nominale	- De rang - Logique	- De similarités	- De mesures
Exemples	- Nature des matériaux - Fonction géométrique d'une pièce - Description de la fonction d'une pièce - Forme du brut d'usinage	- Tolérance d'usinage - Complexité d'usinage	- Quantité produite - Température	- Longueur - Largeur - Hauteur - Diamètre - Volume - Surface - Poids

figure II_2 : Tableau de caractéristiques des différents types d'échelle.

que deux modalités connexes se ressemblent plus que deux modalités extrêmes ;

- Echelle d'intervalle :

La variable est plongée dans R^+. Elle permet d'expliquer la différence entre deux variables. L'exemple courant utilisé pour représenter cette échelles est l'échelle des températures (Celsius, Farenheit). Le zéro ne représente pas l'absence de phénomène ;

- Echelle métrique :

La variable est plongée dans R. Cette échelle est la plus facile à manipuler car toutes les opérations mathématiques peuvent être définies sur ces variables.

On appelle communément variables qualitatives, des variables mesurées sur une échelle nominale et variables quantitatives, des variables mesurées sur une échelle d'intervalle ou métrique.

213) Homogénéisation des variables

Les variables hétérogènes sont définies sur des ensembles différents. Elles doivent être homogénéisées afin d'obtenir la même échelle de mesure lors du calcul des indices de proximité. Deux possibilités s'offrent pour obtenir des variables de même nature :

- soit, on dégrade les variables métriques dont l'échelle est la plus riche en variables dont l'échelle est plus pauvre (ex : transformation de variables métriques en variables ordinales) ;

- soit, on enrichit les variables dont l'échelle est la plus pauvre en variables dont l'échelle est supérieure(ex : transformation de variables nominales en variables ordinales).

Suivant la transformation utilisée, il y a une perte d'informations ou une introduction d'informations supplémentaires. Les conséquences qui découlent du fait des hypothèses supplémentaires engagées sont mal connues quant à la perturbation des résultats de l'analyse typologique. Grenn et Tull [45] proposent ainsi d'effectuer pour chaque application deux analyses typologiques, dans lesquelles l'une aura des variables dégradées et l'autre des variables enrichies et de comparer les résultats.

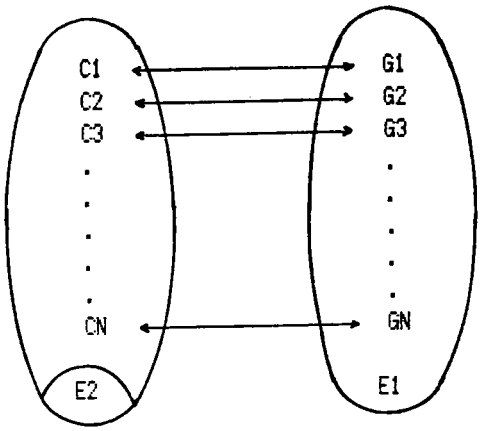
Si toutes les variables sont mesurées sur la même échelle de mesure, certaines peuvent avoir différentes unités de mesure(ex : mètre-centimètre). Dans le calcul des indices de proximités l'utilisation des variables brutes peut entrainer une pondération implicite de certaines d'entre elles par rapport à d'autres, ce qui biaise l'analyse.

On supprime cet effet de pondération en standardisant ou en normalisant les variables.

22) Problème de classement en technologie de groupe

Dans la plupart des applications de technologie de groupe, on est amené le plus souvent à considérer deux ensembles de variables pour représenter un produit [12]. Nous sommes conduit à effectuer, par exemple, une classification sur un ensemble de variables E_1 alors que le classement de nouveaux produits se fera sur un autre groupe de variables E_2.

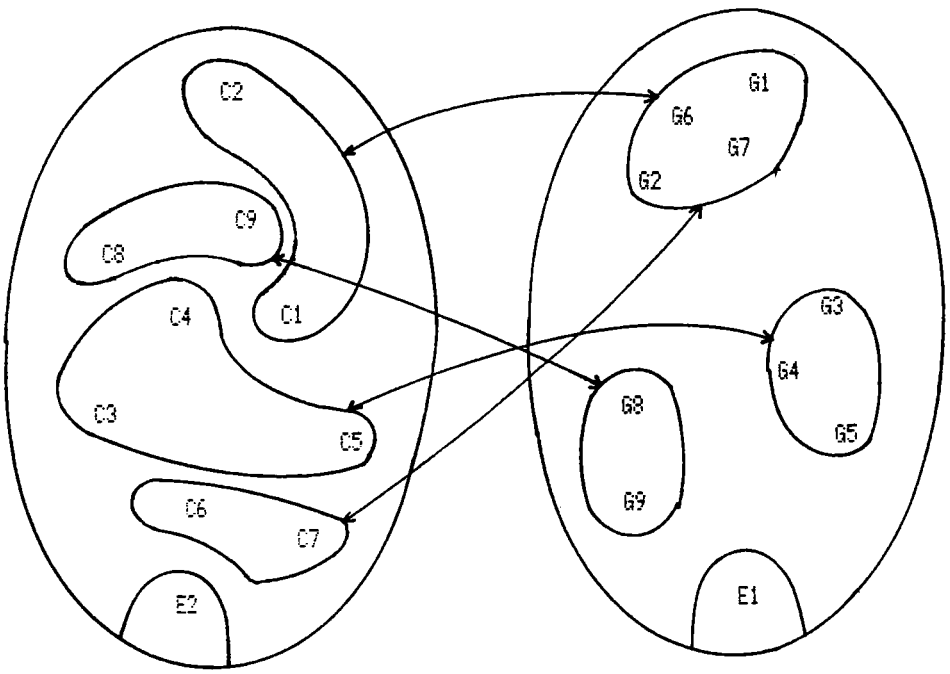
On recherche, par exemple (fig.II$_3$), des familles de gammes de fabrication (G_1, G_2...G_n) présentant des séquences d'opérations similaires (classification sur E_1, E_1 = (G_1, G_2,...G_n)). Par contre, on désire affecter



Ensemble de codes morphodimensionnels

Ensemble de gammes de fabrication

Par produit, il y a bijection entre éléments des deux ensembles, ce qui n'est plus le cas au niveau des familles.



Représentation des familles de codes correspondant à la classification

Ensemble de familles obtenues par classification

figure II_3 : Exemple montrant l'homogénéité des classes de gammes de fabrication obtenue par classification et leur hétérogénéité sur les codes morphodimensionnels.

un nouveau produit connu seulement par son code morphodimensionnel à une classe de E_1 (classement sur E_2 : $E_2 = (v_1, v_2, \ldots v_n)$) afin d'en déduire sa gamme de fabrication. Les groupes de variables des ensembles E_1 et E_2 sont très peu corrélés aussi, comme le montre la figure II_3, à chaque gamme peut correspondre un code, mais il n'y a aucune raison pour que les familles portant sur le code morphodimensionnel soient homogènes alors que les familles de gammes le sont naturellement par construction.

Dans ces conditions, dans l'espace de représentation des produits par leur code, il y a peu de chance de trouver des zones homogènes convexes de produits appartenant à une même classe. Aussi, la méthode de discrimination choisie devra tenir compte de cette configuration.

23) Méthodes de discrimination

231) Introduction

On rassemble sous le nom d'analyse discriminante, un ensemble de techniques explicatives, descriptives et prédictives qui ont pour objectif d'étudier une population d'individus* répartis en plusieurs classes (familles) et caractérisés par de nombreuses variables.

L'analyse discriminante s'efforce donc de résoudre le problème de l'affectation d'un individu à une classe parmi plusieurs connues a priori par une classification préalable.

Suivant le but poursuivi, plusieurs approches sont possibles :

a) Approche descriptive

A ce niveau, on cherche un nombre restreint de variables pouvant expliquer au mieux la séparation des classes.
On cherche donc à mettre en évidence le

pouvoir discriminant des variables ;

* Nota : Dans ce qui suit nous appellerons individus une représentation codée quelconque d'un produit.

b) Approche décisionnelle

On cherche à résoudre, dans ce cas, le problème suivant :

Un nouvel individu se présente, dont on ne connaît que les valeurs des variables explicatives mesurées sur lui. On suppose que cet individu appartient à l'une des classes définies sur un ensemble d'apprentissage, et on veut déterminer cette classe d'appartenance

Il est difficile de dissocier les deux approches, car toutes les méthodes mènent à la recherche d'une règle d'affectation basée sur l'apport d'informations des variables explicatives.

232) Les méthodes

Les principales méthodes d'analyses discriminantes ne s'appliquent qu'à des variables explicatives à caractères quantitatif.
En annexe I, nous présentons le principe de ces différentes méthodes en séparant les deux approches. Les méthodes exposées sont les suivantes :

- Méthode Bayésienne,

- Méthodes de Sebestien,

- Discrimination par voisinage.

Pour les méthodes utilisant la réduction des variables :

- Méthodes pas à pas basées sur plusieurs critères de sélection des variables,
- Analyse factorielle discriminante,
- Discrimination par méthode de régression linéaire.

Dans une deuxième partie de cette annexe I, nous avons énoncé des méthodes pouvant s'appliquer à des variables qualitatives :

- Méthode séquentielle de corrélation canonique,
- Méthode des coefficients de Tschuprow,
- Méthode non paramétrique,
- Méthode de discrimination par segmentation,
- Méthode d'apprentissage de descriptions structurelles complexes,
- Discrimination par voisinage.

Les différentes méthodes utilisant des variables qualitatives conduisent, dans leur majorité, à réaliser un codage quantitatif des variables qualitatives afin d'adopter une méthode classique de discrimination.

Certaines méthodes classiques sont très bien adaptées au problème de discrimination dans le cas de deux classes, telles les méthodes de régression linéaire.

La méthode Bayésienne impose la multinormalité des distributions de probabilité dans chacun des groupes. Dans cette méthode, le problème d'évaluation des coûts de mauvais classement est un problème délicat à résoudre. La règle d'affectation est équivalente à celle utilisée en analyse discriminante linéaire si certaines hypothèses sont vérifiées.

La méthode de Sebestyen fournit un pourcentage de bien classé supérieur aux autres méthodes car elle tient mieux compte de la forme des classes.

La méthode par voisinage, basée sur le principe empirique que deux individus proches ont une forte probabilité d'appartenir à la même classe, est une méthode locale qui ne dépend pas de la disposition géométrique des classes.

De toutes les méthodes énoncées, la discrimination par voisinage peut s'appliquer aussi bien à des variables qualitatives que quantitatives. Cette méthode ne dépend pas d'une métrique particulière pour mesurer les indices de proximités entre individus. On peut ainsi choisir la mieux adaptée en fonction des applications envisagées.

Par ailleurs, cette méthode demande des temps de calcul longs pour une population d'individus élevée.

Des méthodes, telles que celle de Sebestyen ou l'analyse factorielle sont de type géométrique, c'est à dire qu'elles affectent un individu à la classe la plus proche au sens d'une mesure de proximité entre cet individu et une fonction linéaire ou quadratique représentant cette classe. Les méthodes de type géométrique sont efficaces si les classes sont géométriquement discernables. Ces méthodes sont rapides car elles nécessitent uniquement des calculs d'indice de proximité entre le nouvel individu et chacune des classes.

Récemment F. Bonneau et J.M. Proth [13] ont proposé une méthode dont l'affectation est mixte de type géométrique et par voisinage. Cette méthode s'intéresse à l'aspect décisionnel de l'analyse discriminante.

La méthode, que nous proposons, exposée dans le chapitre suivant, est une variante de cette méthode mixte.

CHAPITRE III

ETUDE D'UNE METHODE DE CLASSEMENT D'UN NOUVEAU PRODUIT ET DE LA SIMPLIFICATION DU CODE DE REPRESENTATION

III ETUDE D'UNE METHODE DE CLASSEMENT D'UN NOUVEAU PRODUIT ET DE LA SIMPLIFICATION DU CODE DE REPRESENTATION

31) Méthode de classement

311) Modélisation

Soient :

- E : une population d'individus pouvant être infinie, chaque individu x est défini par un code à p variables ;

- w(x) : une variable qualitative à n modalités connue sur un sous ensemble fini F de E.
  F est appelé population de base.
  Cette variable décompose l'ensemble F en n classes $A_1 \ldots A_n$.

On appelle :

$A_r = \{x \in F \ / \ w(x) = r\}$ avec $r = 1 \ldots n$

La règle de décision D sera telle que D(x) = r représente la décision d'affecter l'individu x à la classe Ar.

312) Principe de la méthode

Nous réalisons tout d'abord un traitement initial des données sur l'ensemble F des individus qui permet d'estimer comment sont topologiquement distribuées les classes A_r.

L'ensemble F sera décomposé en plusieurs régions homogènes connexes C et, une région hétérogène appelée zone trouble (ZT).

Dans un deuxième temps, afin de classer un nouvel individu, nous lui appliquons une règle de décision mixte qui est tantôt de type géométrique, tantôt de type par voisinage. Dans ce dernier cas, les plus proches voisins ne sont pas recherchés sur tout l'ensemble F, mais uniquement sur une fenêtre, bien définie, de celui-ci.

313) Définitions

Nous définissons, dans ce paragraphe, les éléments utiles dans la suite.

- Boule : On appelle boule B et F, l'ensemble formé par l'individu x et ses s plus proches voisins.
  s est appelé rayon de la boule et x en sera son centre ;

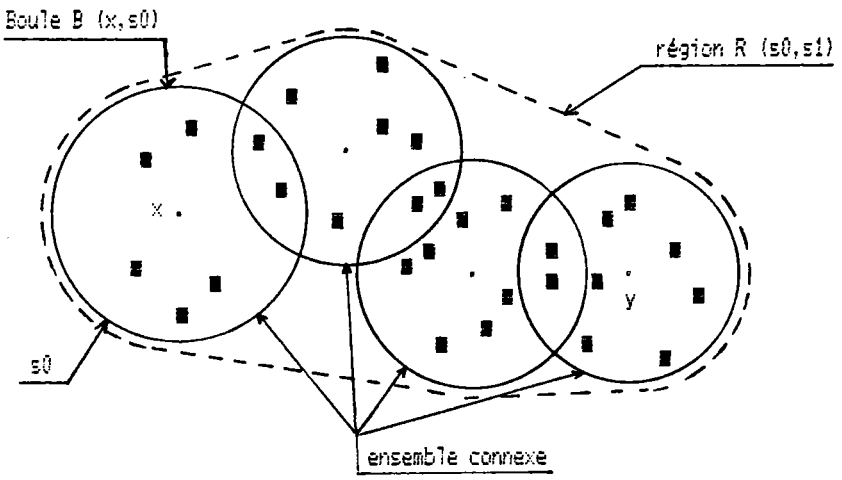
- Ensemble connexe : Nous appelons C ensemble connexe, un sous-ensemble de F vérifiant la propriété suivante :

  Pour tous (x,y) de C^2, il éxiste une suite de boules Bø...Bl incluses dans C, telles que pour tout i de Ø à l-1 :

  $B_i \quad B_{i-1} \neq \emptyset$ avec $x \in B_ø$ et $y \in B_l$

- Epaisseur et taille : Nous dirons qu'un sous-ensemble C de F est d'épaisseur au moins sø si il éxiste une boule de rayon sø incluse dans C. Il sera de taille sl si son cardinal est au moins égal à sl.

- <u>Région</u> : On appelle R(sø,s1) région (sø,s1) de F, un sous-ensemble connexe d'épaisseur au moins sø et de taille au moins s1

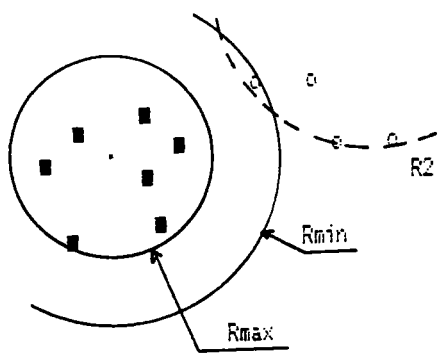


- <u>Eléments caractéristiques d'une boule</u> : On note d(x,y), la distance ou mesure de proximité entre des individus quelconques x et y. Nous définissons deux valeurs associées à chaque boule :

- rayon maximal : $r_{max} = \underset{x \in R}{\text{Max}} \{ d(x,G) \}$

G étant le centre de la boule

- rayon minimal : $r_{min} = \underset{y \in R}{\text{Min}} \{ d(y,G) \}$



- Remarques : Telle qu'est définie la notion de régions celles-ci peuvent avoir toutes les formes possible du moment qu'elles restent connexes (fig.III$_1$) [12-31].

314) Traitement initial

3141) Partitionnement en régions

a) Introduction

Etant donné un sous-ensemble A (classe) de E, on dira que $R_1 \ldots R_k$, ZT est un partitionnement en régions (sø, s1) si $R_1 \ldots R_k$, ZT vérifient les trois propriétés suivantes :

a) Pour tout i<k, Ri est une région d'épaisseur au moins sø et de taille au moins s1,

b) Pour tout couple (i,j), l'ensemble Ri U Rj n'est plus une région,

c) Le sous-ensemble Zt est tel que :

- Pour tout x de Zt et pour tout i, l'ensemble Ri U (x) n'est plus une région,

- Il n'éxiste pas de région (sø,s1) incluse dans ZT.

Quelque soit le sous-ensemble A et E, et pour sø et s_1 donnés, il éxiste un seul partitionnement en régions (sø,s1) de A. Une démonstration de l'éxistence et de l'unicité du partitionnement est donnée en [13].

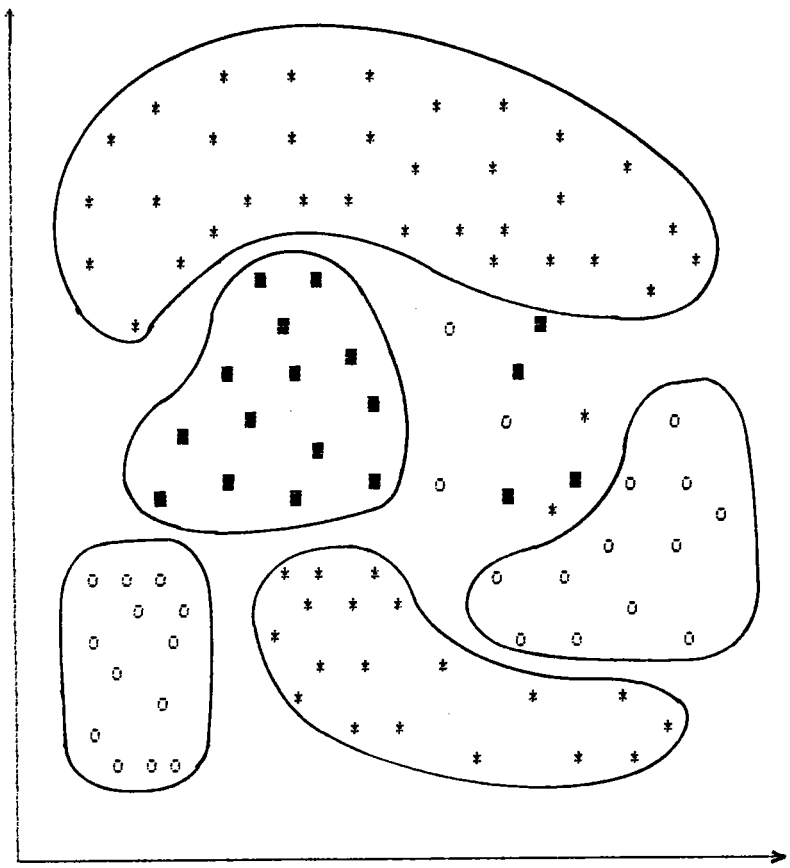


figure III_4 : Exemple de régions formées sur trois classes (rond,carré,croix).

La classe des ronds comporte deux régions

" " " carrés " une région

" " " croix " deux régions

b) Algorithme de partitionnement

Soient les seuils s_0, s_1 et la classe A_r à partitionner en régions (s_0, s_1).

1. Initialisation de A_r à la zone trouble Zt.
2. Pour chaque individu x de A_r.
   - 2.1 Calcul de la plus grande Boule B_x.
   - 2.2 Si rayon (B_x) < s_0
     - x reste dans Zt
     - retourner en 2.

     sinon
     - aller en 2.2.1.

   - 2.2.1 Si il éxiste une région R déjà formée et que B_x $R \neq \emptyset$
     - $R = B_x \cup R_1 .. \cup R_m$ est augmentée.

     sinon
     - $R = B_x$ (x, card(B_x)) est formée.

3. Si stabilisation du partitionnement.
   - aller en 4

   sinon
   - retourner en 2.

4. On met dans Zt les régions de cardinal < s_1.
5. Fin de procédure.

c) Description de l'algorithme

Nous plaçons, tout d'abord, tous les individus de la classe A_r à partitionner dans la zone trouble (étape 1). Nous prenons ensuite chaque individu x de A_r et nous calculons la plus grande boule de centre x incluse dans A_r (étape 2.2.1.). A ce stade trois cas peuvent se présenter :

1er cas : Si le rayon de la boule B_x est inférieur au seuil s_0, nous laissons cet individu dans la zone trouble et nous prenons le suivant.

2ème cas : Si il existe une région R déjà formée telle que l'intersection de cette région avec la boule est différente de l'ensemble vide, nous formons une nouvelle région qui est l'union de l'ancienne avec la boule.

3ème cas : Si il n'y a pas de région déjà formée nous en créons une qui est la boule de centre x et de cardinal card(B_x).

Nous itérons jusqu'à ce que le partitionnement se stabilise (étape 3). Puis nous mettons dans la zone trouble les régions dont le cardinal est inférieur au seuil s_1.

Remarque 1 : En pratique deux ou trois itérations suffisent pour que le partitionnement de chaque classe A_r se stabilise.

Remarque 2 : Pendant le déroulement de l'algorithme de partitionnement, nous pouvons nous trouver dans le cas où certains ensembles A_r ne possèdent aucune région, ceci pouvant être dû à :

- la taille de l'ensemble A_r inférieure au seuil s_1. En effet, si l'ensemble A_r ne compte pas assez d'individus, il n'y aura aucune région de créée.

- la taille de chaque région temporairement formée inférieure au seuil s_1. Effectivement l'algorithme construit des boules et des régions temporairement quelque soit leurs tailles (supérieure à s_0), mais si

celles-ci ont un nombre d'individus inférieur à s_1, les régions ne seront pas validées.

Dans ces deux cas, l'algorithme place les individus dans la zone trouble.

Un autre cas qui ne conduit pas à la création de régions est lorsque la taille de chaque boule est inférieure au seuil sø. On pourra remédier à cet inconvénient en diminuant la valeur de sø.

Remarque 3 : Compte tenu de l'hétérogénéité des variables citée dans le chapitre II, la distance entre deux individus est définie comme la somme pondérée d'une distance mesurée sur les variables qualitatives, et d'une distance mesurée sur les variables numériques. Nous avons choisi, une distance de type Khi2 pour les variables qualitatives et une distance euclidienne sur les variables quantitatives [15]. Ces deux distances sont pondérées par un coeficient afin de donner le même poids aux deux groupes de variables . La pondération s'établit comme ceci :

$$d(x_i, x_j) = \alpha\, d_1(x_i^1, x_j^1) + \beta\, d_2(x_i^2, x_j^2)$$

avec :

d_1 = distance du Khi2 sur les variables qualitatives

$$d_1^2(x_i, x_j) = \sum_{k=1}^{P} \frac{1}{a.k} \left(\frac{x_{ik}}{ai.} - \frac{x_{jk}}{aj.} \right)^2$$

avec : $a.k = \sum_{i=1}^{N} x_{ik}$, $ai. = \sum_{k=1}^{P} x_{ik}$, $aj. = \sum_{k=1}^{P} x_{jk}$

d_2 = distance euclidienne sur les variables quantitatives

$$d_2^2(x_i, x_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

$\beta = 1-\alpha$

$$\alpha = \frac{\overline{td_2}}{\overline{ld_1} + \overline{td_2}}$$

$\overline{d_1}$ et $\overline{d_2}$ = moyennes des distances

t = nombre de variables quantitatives

l = nombre de variables qualitatives

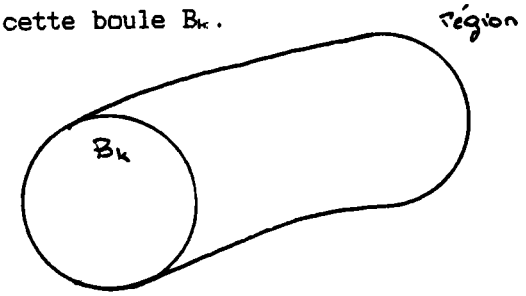
3142) Modélisation des régions

a) Introduction

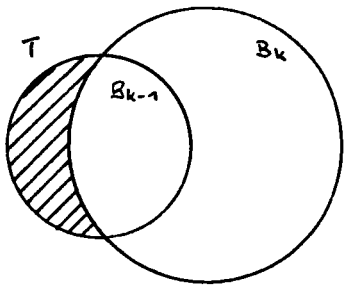
Nous modélisons chaque région par une succession de boules. Cette modélisation, nous permet de calculer rapidement les distances de nouveaux individus à chaque région, lors de l'étape d'affectation en ne considérant que le centre des boules et leurs rayons au lieu des individus inclus dans les boules. La modélisation des régions par des boules induit l'un des quatre cas de figures suivant :

1er cas : Première boule de la région.

Nous mémorisons cette boule B_k.



2ème cas : Si card (B_k) > card (B_{k-1}).

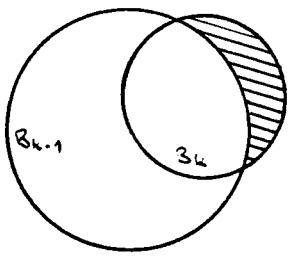


On définit l'ensemble T comme :

card (T) = tol1 * card (B_{k-1}).

Si l'ensemble (B_{k-1}-T) est inclu dans la boule B_k, l'ensemble T appartenant à la boule B_{k-1}, nous éliminons l'ancienne boule B_{k-1} et nous mémorisons la nouvelle : B_k

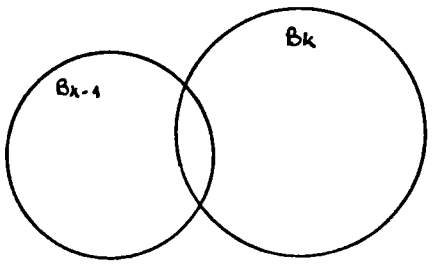
3ème cas : Si card (B_{k-1}-T ⩾ card (B_k).



L'ensemble T est défini par card (T) = tol1 * card (B_k).

Si l'ensemble (B_k-T) est inclu dans la boule B_{k-1}, l'ensemble T appartenant à la boule B_k, nous éliminons la nouvelle boule B_k et nous gardons l'ancienne : B_{k-1}.

4ème cas :



L'ensemble (B_{k-1}-T) n'est pas inclu dans la boule B_k et l'ensemble (B_k-T) n'est pas inclu dans la boule B_{k-1}. Nous gardons l'ancienne boule B_{k-1} et nous mémorisons la nouvelle : B_k.

b) Algorithme de reconnaissance des régions

Etant donnés le seuil sø, la tolérance tol1 et l'ensemble A_r que l'on a partitionné en régions (sø, s_1), l'algorithme de reconnaissance des régions par des boules est le suivant :

```
1.  Pour chaque région de la classe Ar courante
    1.1  Pour chaque élément x de la région
         1.1.1  Calcul du ppv de x à la région
         1.1.2  Calcul de la boule Bkx
         1.1.3  Si rayon de Bkx < sø
                    retourner en 1.1
                sinon
                    aller en 1.1.4
         1.1.4  Si k=1 (cas n°1)
                    mémorisation de B1
                    retourner en 1.1
                sinon
                    aller en 1.1.5
         1.1.5  Si card (Bk) > card (Bk-1) (cas n°2)
                        Si (Bk-1-T)    Bk
```

suppresion de B_{k-1}
mémorisation de B_k
retourner en 1.1
sinon
aller en 1.1.7
sinon
aller en 1.1.6

1.1.6 Si card (B_{k-1}) ≥ card (B_k) (cas n°3)

Si $(B_k - T)$ B_{k-1}
suppression de B_k
retourner en 1.1
sinon
aller en 1.1.7
sinon
aller en 1.1.7

1.1.7 Cas n°4

mémorisation de B_k
retourner en 1.1

2. On met dans la zone trouble tous les individus n'appartenant pas à une des boules.
3. Fin de procédure.

Chaque région de la classe Ar est ainsi représentée par une suite de boules caractérisées par leur centre, leur cardinal et leurs rayons minimum et maximum.

315) Affectation de nouveaux individus

3151) Principe

L'objectif de la méthode d'affectation est de réunir les avantages d'une méthode géométrique et d'une méthode par voisinage, afin d'obtenir une règle de décision D très rapide.

Nous recherchons, tout d'abord, à affecter le nouvel individu à une classe par une suite de tests de type géométrique (figIII$_2$). Si ces tests ne donnent pas de résultats concluants, nous appliquons alors une méthode de type plus proches voisins (figIII$_3$). Cette méthode est uniquement appliquée sur un ensemble réduit d'individus appartenant à une fenêtre ouverte autour du nouvel individu à classer. Dans cette fenêtre, nous recherchons les k plus proches voisins de l'individu à affecter pour k variant entre deux valeurs limites k_{min} et k_{max}. A chaque valeur de k,nous calculons la probabilité d'appartenance aux classes qui sera notre critère de décision.

L'affectation du nouvel individu n'est réalisée qu'à la condition que ce critère soit maximum. De plus, ce critère doit être supérieur à une certaine tolérance fixée au départ. Si cette tolérance n'est pas atteinte, on affectera l'individu à la région homogène la plus proche.

3152) Algorithme d'affectation

Soient deux valeurs k_{min} et k_{max} avec $k_{min} < k_{max}$ et un nombre tol telque $\emptyset \leq tol < 1$.

Soit X_{new}, un nouvel individu que l'on désire affecter à un ensemble Ar, l'algorithme d'affectation est le suivant :

1. Entrer des coordonnées du nouvel individu X_{new}.

2. Calcul de la boule la plus proche B_{min} (X_{new}).

   2.1 $r\emptyset = d(x_{new}, B_{min})$: rayon de la fenêtre

   2.2 $i\emptyset$ = numéro de la boule B_{min} (x_{new})

3. Si $r\emptyset < r_{min}$ (B_{min}) (test n°1)

   affectation directe à $i\emptyset$

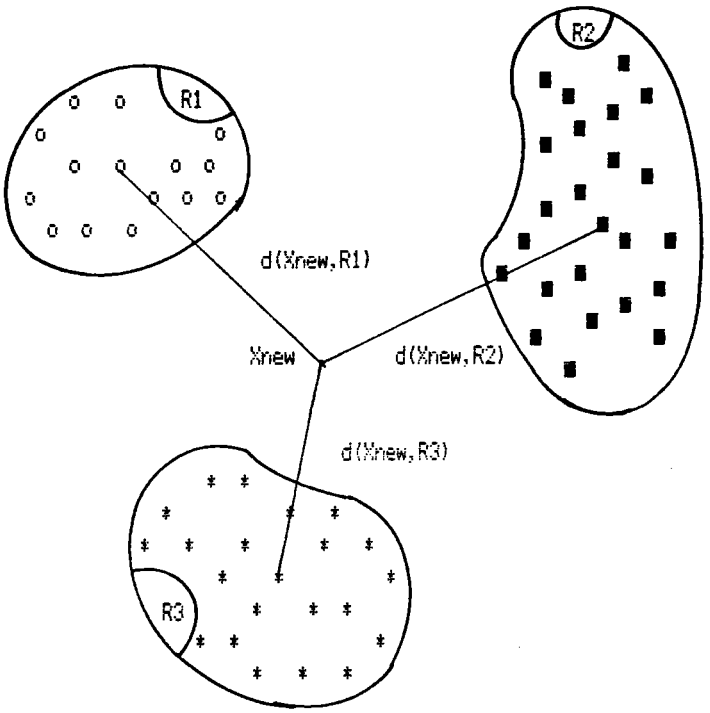


figure III_2 : Exemple d'affectation géométrique.
Xnew sera affecté à la région la plus proche (R3).

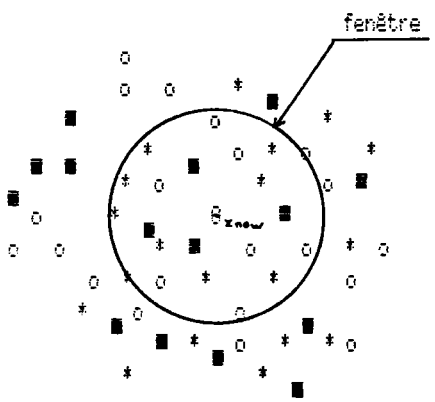


figure III_3 : Exemple d'affectation par voisinage sur une fenêtre.

effectif * = 9
■ = 4
o = 8

Xnew sera affecté à la classe dont l'effectif est maximum (*).

aller en 5

sinon

aller à 3.1.

3.1. Détermination du nombre de boules concernées par la fenêtre tel que :

$$d(x_{new}, B_i) < r_{min} + r\text{ø}$$

3.2. Détermination du nombre d'éléments de la zone trouble appartenant à la fenêtre.

3.3. Si une seule boule est concernée et que le nombre d'élément de Zt est inférieur à k_{min} (test n°2).

affectation directe à iø
aller en 5

sinon

aller en 3.3.1

3.3.1. Remplissage de la fenêtre par les éléments des boules concernées.

3.3.2. Remplissage de la fenêtre par les éléments concernés de la zone trouble.

3.3.3. Si le nombre d'éléments de la fenêtre $< k_{min}$ (test n°3)

affectation directe à iø
aller en 5

sinon

aller en 3.3.3.1.

3.3.3.1. Procédure des k plus proches voisins à l'intérieur de la fenêtre

4. Fin de Procédure

Description de l'algorithme :

Dans un premier temps, nous entrons les coordonnées du nouvel individu (étape 1), et nous recherchons la boule la plus proche (étape 2) (figIII$_4$). La distance, du nouvel individu à la boule la plus proche, nous fournit le rayon de la fenêtre dans laquelle nous pourrons, éventuellement, calculer les k plus proches voisins.

Le test qui suit (test n°1) permet de savoir si le nouvel individu se trouve dans une boule (étape 3). Dans ce cas, nous l'affectons directement à cette boule (figIII$_5$).

Dans le cas contraire, nous recherchons les boules ayant une intersection avec la fenêtre (étape 3.1) (figIII$_6$). Nous comptons le nombre de boules concernées. Nous dénombrons ensuite le nombre d'éléments de la zone trouble appartenant à la fenêtre. S'il n'existe qu'une seule boule et que le nombre d'éléments de la zone trouble est inférieur à k_{min}, nous affectons x_{new} à la boule la plus proche (test n°2) (étape 3.3) (figIII$_7$). La raison de cette décision est que nous estimons que les éléments de la zone trouble ne sont pas assez nombreux pour qu'une affectation par les plus proches voisins soit significative.

Si l'affectation n'est toujours pas réalisée, nous plaçons, dans la fenêtre, les individus des boules concernées (étapes 331 et 332). Si le nombre total d'éléments de la fenêtre est inférieur ou égal à k_{min}, nous affectons de même à la boule la plus proche (test n°3 - étape 3.3.3.) (figIII$_8$), sinon nous faisons appel à la procédure des k plus proches voisins.

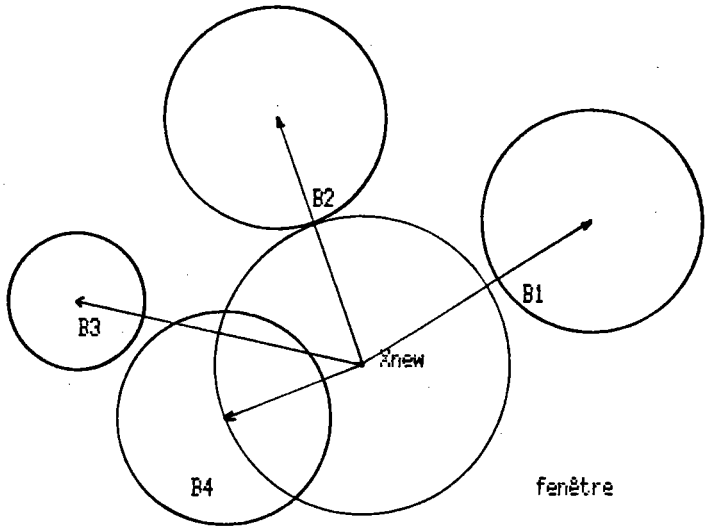


figure III_4 : Exemple d'une recherche de la boule la plus proche, dans le cas B4 et construction de la fenêtre (étape 2 de l'algorithme).

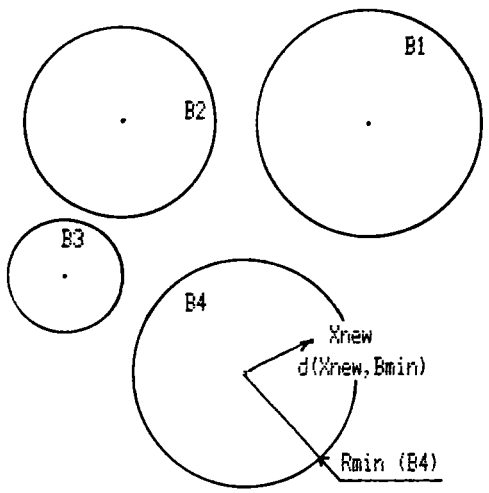


figure III_5 : Exemple d'affectation directe à la boule B4 Xnew se trouve dans B4 (d(Xnew,B4) < Rmin(B4)).

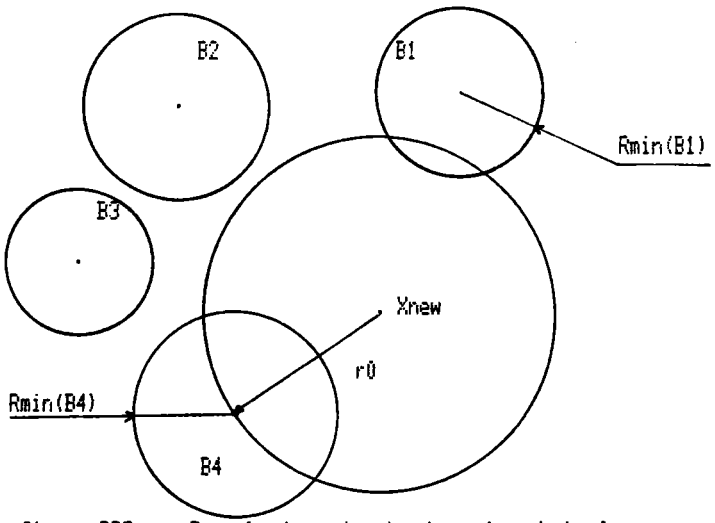


figure III_6 : Exemple de recherche du nombre de boules concernées par la fenêtre.Dans ce cas,il y en a deux (B1,B4) telque :
d(Xnew,Bi) < r0+Rmin(Bi)

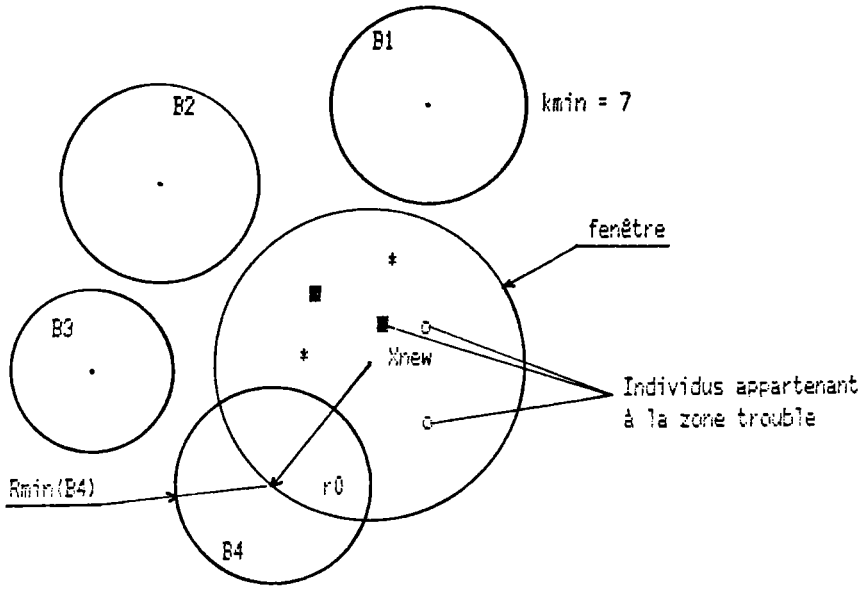


figure III_7 : Exemple d'affectation directe à la boule la plus proche (B4) car une seule boule est concernée par la fenêtre et le nombre d'éléments de la zone trouble (6) est inférieur à kmin.

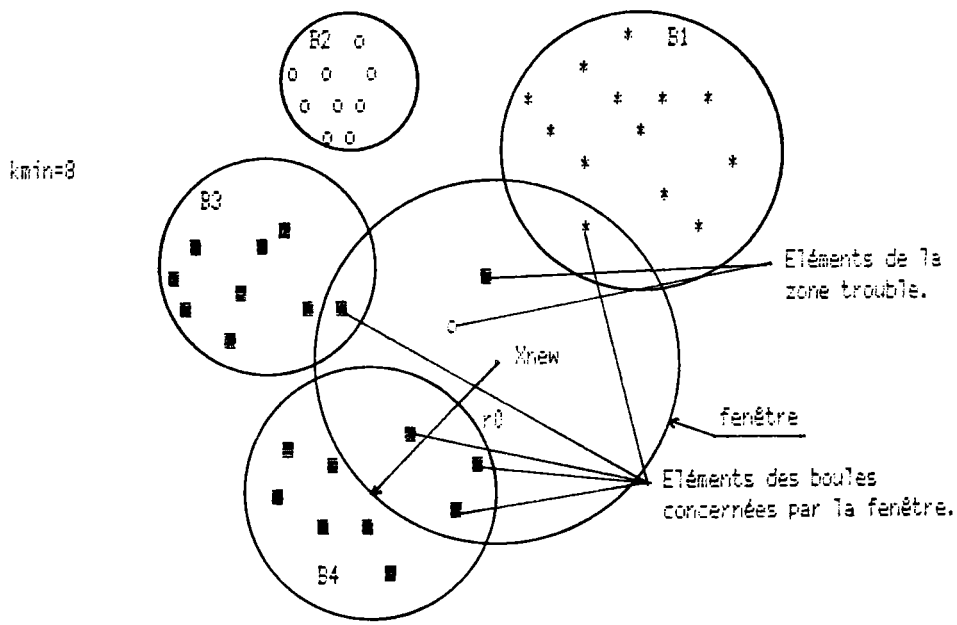


figure III_8 : Exemple d'affectation directe à la boule la plus proche (B4) car le nombre d'éléments de la zone trouble et le nombre d'éléments des boules concernées par la fenêtre est inférieure à kmin (nb total =7 < kmin).

3153) Procédure des k plus proches voisins

a) Principe

Nous appliquons cette méthode uniquement sur la fenêtre déterminée lors du déroulement de l'algorithme d'affectation précédant. Ceci permet d'éliminer, lors des calculs, tous les individus non directement concernés et ainsi accélérer l'éxécution de cette procédure.

Cette méthode par voisinage est basée sur le principe d'affecter, un nouvel individu x_{new} à la classe qui a obtenu la plus forte fréquence d'appartenance des individus de même nature au voisinage de x_{new}. Pour calculer ces fréquences, nous déterminons ses k plus proches voisins. Ceux-ci engendrent un ensemble V à l'intérieur de la fenêtre. Nous déterminons alors la probabilité de x_{new} d'appartenir à une classe A_r par :

$$P(x_{new}, A_r) = \text{card } (V \cap A_r)/k.$$

Nous affectons x_{new} à la classe qui rend maximum la probabilité $P(x_{new}, A_r)$, si elle est supérieure à une tolérance fixée au départ, sinon nous l'affectons à la classe la plus proche. Selon les valeurs de k, pour un ensemble donné les probabilités calculées peuvent être très différentes. Aussi le choix d'une bonne valeur de k est important. Pour ce choix, nous avons retenu la méthode proposée en [13]. A partir d'un ensemble test, nous traçons la courbe du pourcentage de bien classé en fonction de k (fig.III$_9$-III$_{10}$). A partir de cette courbe, nous retenons les deux valeurs limites, k_{min} et k_{max}, telque dans cet intervalle le pourcentage de bien classé reste stable. Nous déterminons alors les k plus proches voisins pour k variant entre ces deux valeurs. D'une façon imagée, nous regardons au voisinage proche de x_{new}, la tendance des individus. Puis au fur et à mesure que k grandit, nous étendons notre champ de vision tout en restant dans la fenêtre créée précédemment.

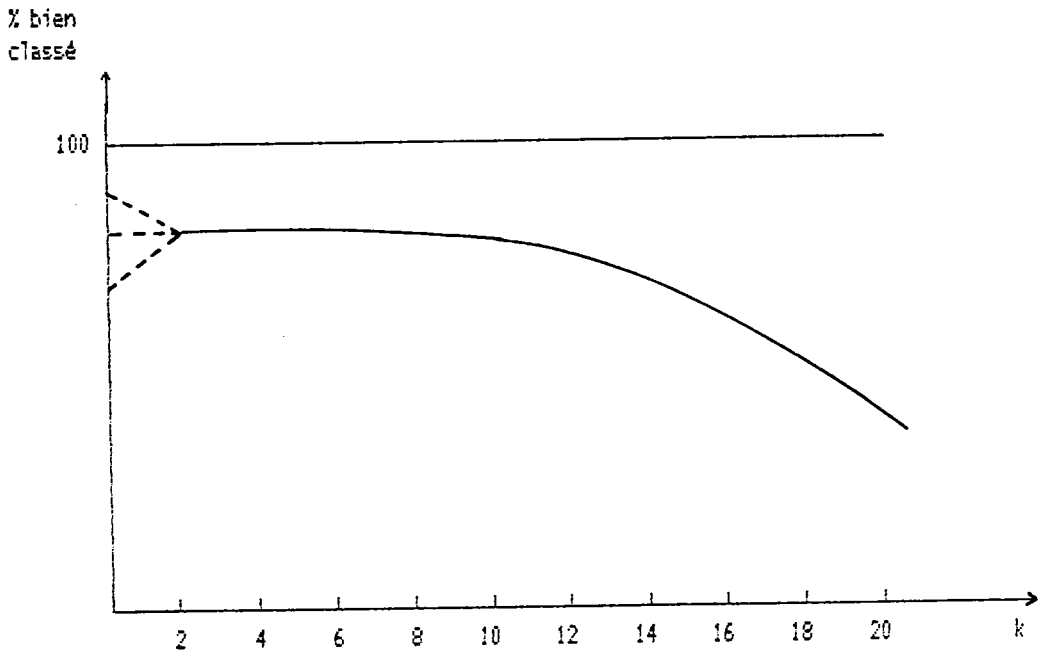


figure III_9 : Allure générale des courbes du pourcentage de bien classé en fonction du nombre de plus proches voisins.

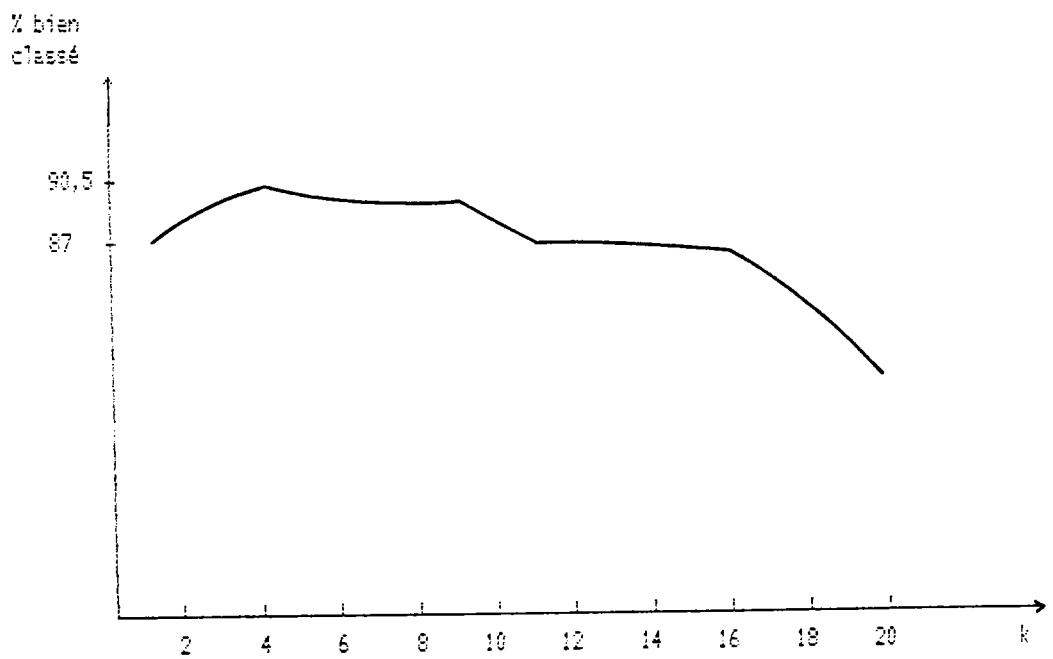


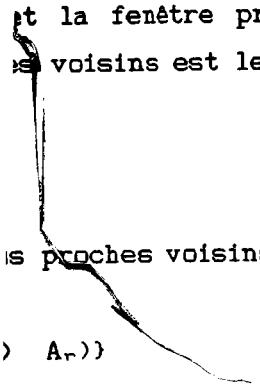
figure III_{10} : Courbe du pourcentage de bien classé en fonction du nombre de plus proches voisins pour l'application n°2.

rche des k plus proches

x valeurs entières k_{min} et

t la fenêtre précédemment

s voisins est le suivant :



s proches voisins de x_{new}

) A_r))

se ayant obtenu ce maximum

ilité de x_{new} d'appartenir à

test N°5)

sse la plus proche.

b) Algorithme de recherche des k plus proches voisins

Etant données deux valeurs entières k_{min} et k_{max} avec $1 < k_{min} < k_{max}$, la tolérance tol et la fenêtre précédemment créée, l'algorithme de recherche des k plus proches voisins est le suivant :

1. Pour k variant de k_{min} à k_{max}.

   1.1 Calcul de V(k) : l'ensemble des k plus proches voisins de x_{new}

   1.2 Calcul de $N_{max}(k) = \max \{(\text{card } (V(k) \quad A_r))$

   1.3 Recherche de im(k) : numéro de classe ayant obtenu ce maximum

   1.4 Calcul de $P(k) = N_{max}(k)/k$: probabilité de x_{new} d'appartenir à à im(k)

2. Recherche de $P(k\text{ø}) = \max(P(k))$.

3. Si $P(k\text{ø}) > tol$ (test n°4).

   affectation de x_{new} à im(kø)
   aller en 4

   sinon

   aller en 3.1

   3.1 Si de k_{min} à k_{max} $P(k) < tol$ (test N°5)

   affectation de x_{new} à la classe la plus proche.

4. Fin de procédure.

c) Description de l'algorithme

Dans un premier temps (étape 1 de l'algorithme) nous recherchons, pour chaque valeur de k, l'ensemble des k plus proches voisins de x_{new} (étape 1.1). Puis nous calculons les fréquences d'appartenance de ses voisins et nous retenons le maximum (étape 1.2).
Nous recherchons le numéro de la classe ayant obtenu ce maximum (étape 1.3.). Nous calculons la probabilité d'appartenance de x_{new} à la classe en question (étape 1.4.). A l'étape 2 de l'algorithme, nous recherchons la probabilité maximum. Si cette probabilité est supérieure à une tolérance donnée, nous affectons x_{new} à la classe qui a obtenu la majorité des fréquences d'appartenance (étape 3. test n°4). Dans le cas contraire, nous affectons le nouvel individu à la classe la plus proche. La figure III,1 montre un exemple d'affectation suivant cette procédure des k plus proches voisins.

32) Simplification du code

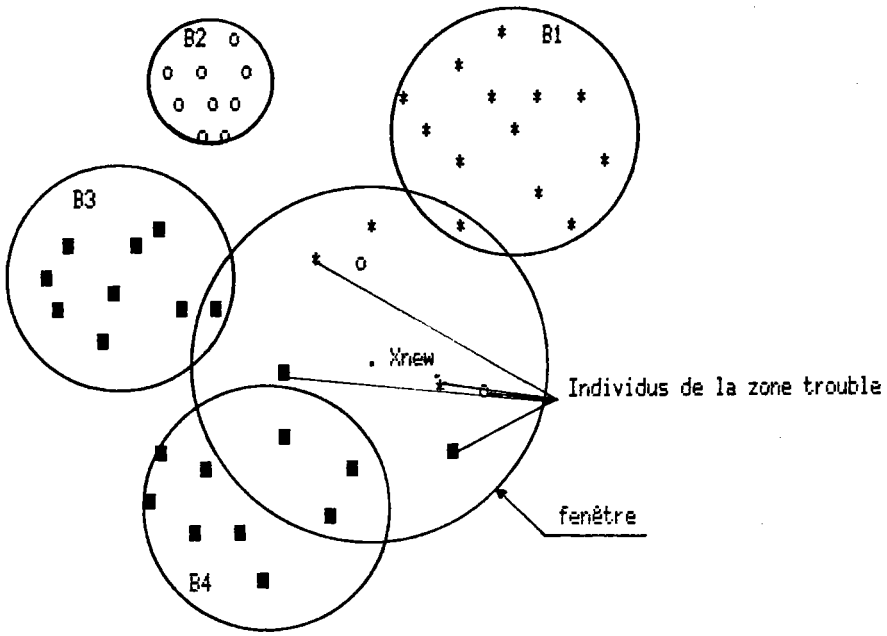
321) But de la simplification

Dans la pratique, les codes employés en technologie de groupe comportent de nombreuses variables . Un faible nombre de variables facilite l'écriture et l'interprétation du code.

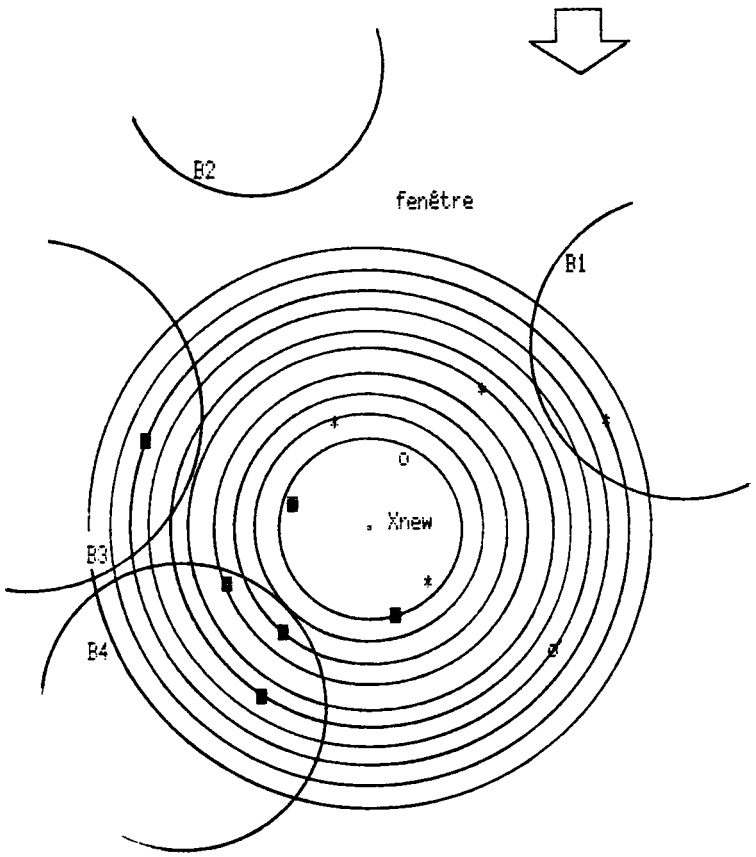
Le but de cette étape de simplification est de retenir le code nécessaire et suffisant qui respecte un bon pourcentage d'individus bien classés. Ceci permet de souligner le pouvoir explicatif de certaines variables du code et ainsi d'améliorer la compréhension des phénomènes étudiés.

La nature différente des variables, nous a conduit à séparer notre problème de simplification en deux parties. Nous appliquons des méthodes distinctes suivant que nous rencontrons des variables quantitatives ou qualitatives.

kmin = 4



Le test n°3 ne permet pas d'affecter Xnew car le nombre d'éléments de la fenêtre (13) est supérieur à kmin, nous faisons donc appel à la procédure des k plus proches voisins.



kmin = 4 kmax = 12

k	fréquences			P(k) max	tol = 0,5
	o	■	*		
4	1	2	1	0,50	=
5	1	2	2	0,40	<
6	1	3	2	0,50	=
7	1	4	2	0,57	>
8	1	4	3	0,50	=
9	1	5	3	0,55	>
10	2	5	3	0,50	=
11	2	6	3	0,66	>
12	2	6	4	0,50	=

k0 = 11 P(k0) = 0,66 > tol

figure III_{11} : Exemple d'affectation suivant la méthode des k plus proches voisins. Xnew est affecté à la classe des carrés "■".

322) Elimination des variables

3221) Variables qualitatives

3221-a) Filtrage des modalités

Dans les codes utilisés, certaines variables qualitatives peuvent comporter jusqu'à une centaine de modalités. De ce fait, sur un code entier, le nombre total de modalités est souvent important. Or, certaines modalités n'apportent pas ou peu d'informations si par exemple la segmentation de la variable est trop fine. Par ailleurs, il peut y avoir des modalités communes à tous les individus de la population F, d'autres absentes pour tout les individus et d'autres enfin complémentaires. Ces modalités, dont le pouvoir significatif est faible ou nul, ne sont d'aucune utilité pour la discrimination. Nous réalisons ainsi, dans une première phase, un filtrage qui permet de supprimer les modalités :

- communes à tous les individus,
- non utilisées,
- complémentaires.

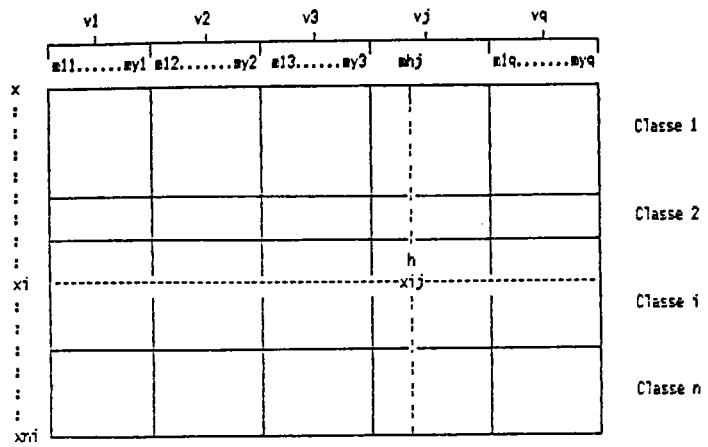
3221-b) Elimination des modalités peu significatives

La méthode que nous proposons permet de réaliser des regroupements de modalités les plus caractéristiques de chacune des classes.

Les données sont constituées de ni individus, chacun décrit par q variables qualitatives. Chaque variable qualitative v présente un ensemble de modalités Mj = (m.....my). Le nombre de modalités peut varier d'une variable à l'autre.

Nous réalisons un codage disjonctif complet qui transforme la matrice colonne de représentation des individus en matrice logique T (fig.III$_{1:2a}$). Cette matrice, d'élément :

$x_{ij}^{h} = 1$ si la variable v_j de l'individu x_i prend la modalité h sinon $x_{ij}^{h} = \emptyset$



Fif.III$_{1:2a}$: Codage disjonctif complet

Dans ces conditions, le nombre total de colonnes de la matrice logique T est :

$$n_T = \sum_j Mj$$

Nous recherchons ensuite une partition de ce tableau en n classes disjointes sur les colonnes telle que pour tout x_i d'une classe, il y ait un maximum d'éléments x_{ij}^{h} prenant la valeur 1. Ceci revient à rechercher n blocs diagonaux dans la matrice logique T (fig.III$_{1:2b}$).

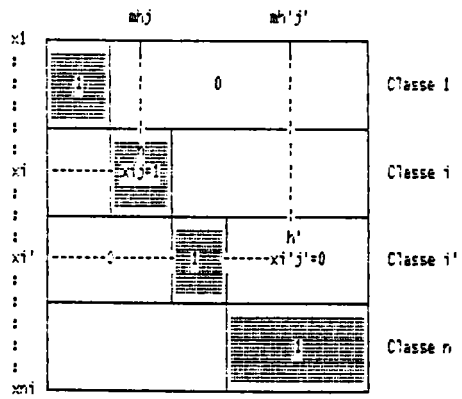


figure III$_{1:2b}$: Matrice de blocs diagonaux.

A l'intérieur de chaque bloc, nous trouvons un maximum de valeurs non nulles et un minimum de valeurs nulles et inversement à l'extérieur des blocs diagonaux (fig.III$_{12c}$).

algorithme de recherche de blocs diagonaux

Etant donné ϕ le pourcentage de valeurs nulles dans tout le tableau T, l'algorithme de recherche de blocs diagonaux est le suivant :

1. Rangement des individus x_i par classe.

2. Pour chaque modalité m_{hj}.

   2.1 Pour chaque classe z.

      2.1.1 Calcul du poids de la modalité.

$$Pz[m_{hj}] = \sum_{j \in z} x_{ij}^{h} \times \phi + \sum_{j \in z} (1-x_{ij}^{h}) \times (1-\phi)$$

   2.2 Recherche de l'affectation à la classe zø de la modalité m_{hj} tel que :

$$Pzø[m_{hj}] : \max \{Pz[m_{hj}]\}.$$

3. Arrangement des colonnes m_{hj} en fonction de l'affectation.

4. Fin de procédure.

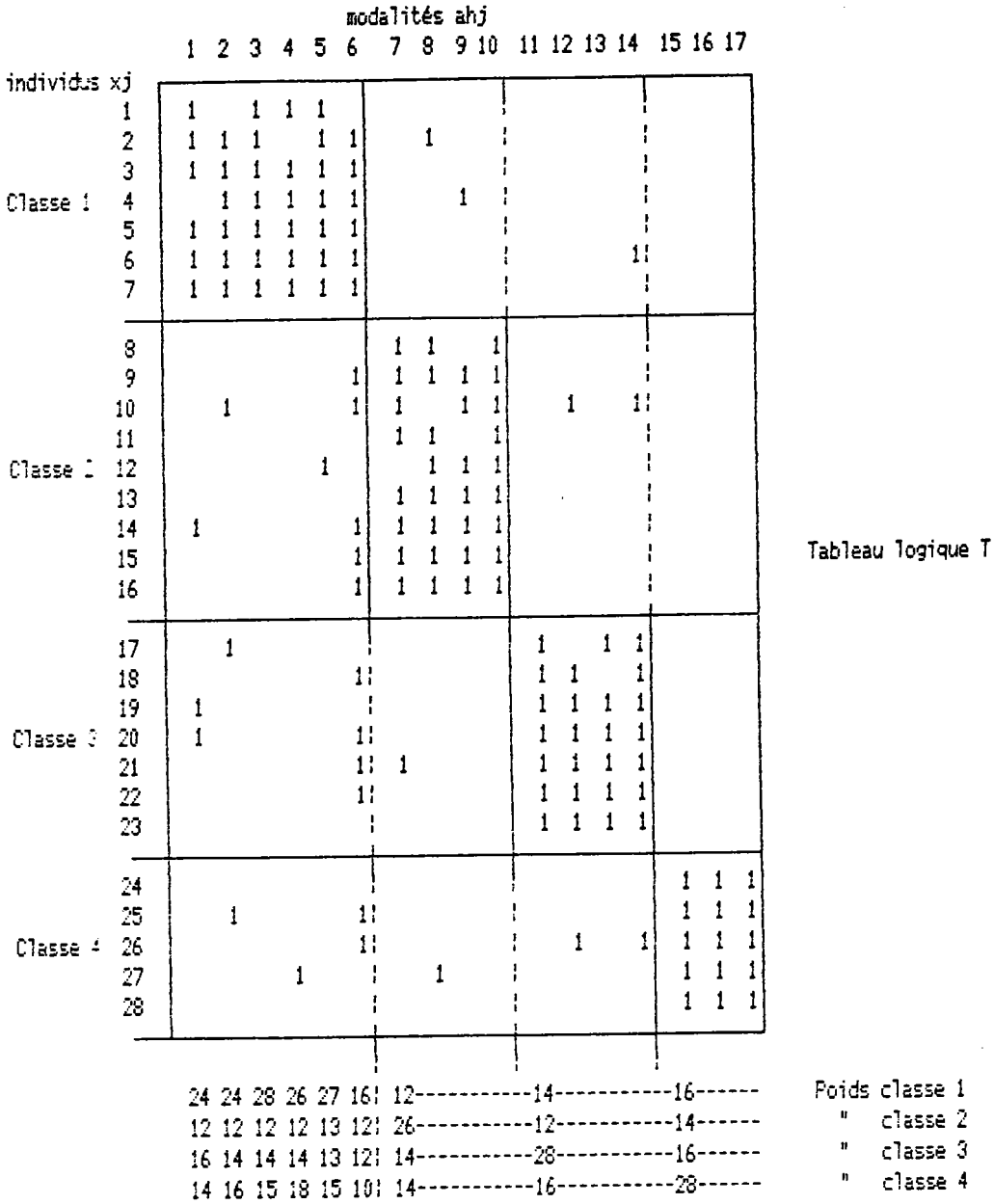


figure III$_{42c}$: Exemple de blocs diagonaux sur une matrice logique. Nous affectons une modalité à la classe dont le poids est maximum ,ex.: la modalité 1 est affectée à la classe 1 car le poids (24) est maximum.

Remarque : Les "0" n'ont pas été représentés dans le tableau pour plus de clarté.

Description de l'algorithme :

Dans un premier temps, nous réalisons un arrangement des individus x_i par ordre croissant des classes (étape 1). Nous calculons ensuite le poids de chaque modalité pour chaque classe (étape 2.1.1). Ce poids représente la somme de valeurs non nulles à l'intérieur de la classe et la somme de valeurs nulles à l'extérieur. Ces sommes sont pondérées par le coefficient f. Plus ce coefficient est grand, plus on favorise le nombre de valeurs non nulles à l'intérieur de la classe. Puis nous affectons la modalité à la classe ayant obtenue le poids maximum (étape 2.2). Nous réalisons ensuite un arrangement des modalités par ordre croissant des affectations aux classes. Ceci nous fournit la matrice de blocs diagonaux (fig.III_{12c}). Chaque modalité étant affectée à une classe, la matrice T réordonnée permet d'analyser le pouvoir discriminant des modalités et d'envisager leurs suppressions. Le critère de suppression est le poids de chaque modalité. Si certaines classes possèdent des modalités qui ont un poids faible par rapport aux autres, nous pouvons affirmer que ces modalités ne sont pas très caractéristiques d'une classe. C'est le cas de la modalité n°6 à la figure III_{12c}, dont le poids égal à 16 est très inférieur à celui des autres modalités de la classe 1.

Nous éxécutons l'algorithme de recherche de blocs diagonaux pour différentes valeurs du coefficient de pondération f. Pour des valeurs distinctes de f, certaines modalités peuvent être affectées à différents blocs. Nous concluons que ces modalités ne sont pas très caractéristiques d'une classe et peuvent aussi être supprimées.

3222) **Variables quantitatives**

Nous utilisons deux critères complémentaires pour supprimer des variables quantitatives peu discriminantes.

a) Corrélations

Nous vérifions tout d'abord, s'il éxiste des liens entre les différentes variables. Pour ceci nous calculons les coefficients de corrélation expérimentales [16].

Soient x_i et x_j deux variables, $cov(x_i,x_j)$ la covariance et σ_i et σ_j les écarts types respectivement des variables x_i et x_j. Le coefficient de corrélation entre x_i et x_j est défini de cette façon :

$$\rho(x_i, x_j) : \frac{cov(x_i, x_j)}{\sigma_i . \sigma_j}$$

$\rho(x_i,x_j)$ varie de -1 à +1 et $\rho(x_i, x_j) = \emptyset$ si les deux variables sont indépendantes. Nous éliminons donc les variables très corrélées qui expliquent de la même façon le phénomène : $|\rho| \# 1$

b) Variances

Dans un deuxième temps, nous déterminons les variables qui séparent au mieux les classes et qui caractérisent leur homogénéité (fig.III$_{1.4}$). Pour cela, nous calculons pour chaque variable quantitative sa variance intra-classe et sa variance inter-classe. Les variables les plus dicriminantes sont celles qui présentent une variance inter-classe maximum (classes bien différenciées) et une variance intra-classe minimum (classes homogènes). Nous éliminons donc celles qui ont un rapport

$$\frac{\text{Var inter}}{\text{Var intra}}$$ faible comparé aux autres.

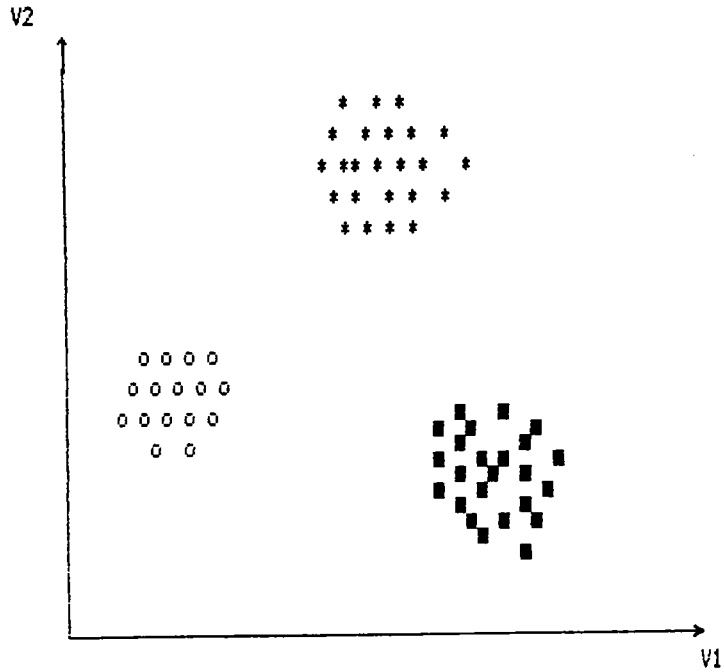


figure III$_{A4}$: Les variables V1,V2 séparent bien les classes et les rendent homogènes.

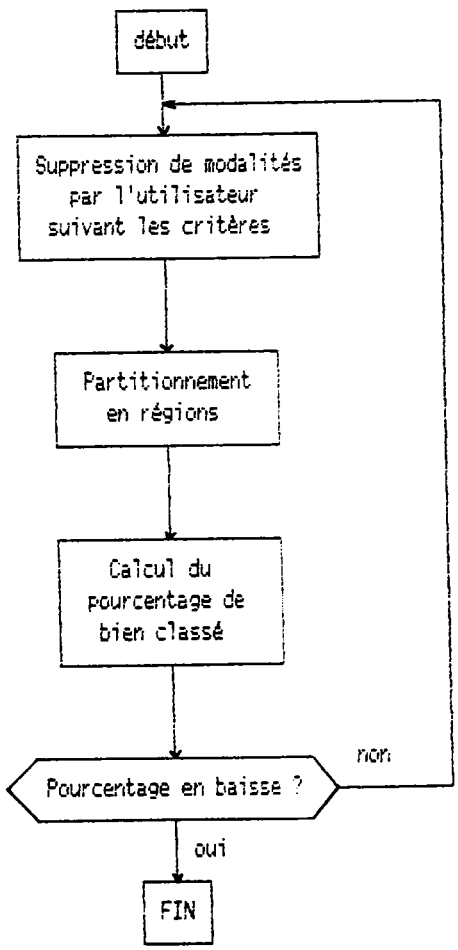


figure III$_{A3}$: Organigramme de suppression de modalités.

33) Conclusion

Nous appliquons un processus itératif (fig.III$_{13}$) qui propose à l'utilisateur de supprimer les variables ou modalités non discriminantes suivant les critères énoncés. Puis nous réalisons le prétraitement des données compte tenu du code simplifié. Nous calculons ensuite le pourcentage d'individus bien classés. Ceci permet de vérifier a posteriori le bien fondé d'une suppression. Ce processus de suppression est exécuté pas à pas pour chaque modalité ou variable jusqu'à ce que le pourcentage d'individus bien classés diminue de façon significative. Cette étape de simplification est coûteuse en temps, mais elle n'est réalisée qu'une seule fois lors de l'implantation du code et de la fonction d'affectation.

CHAPITRE IV

APPLICATIONS

IV APPLICATIONS

Afin de pouvoir appliquer les méthodes décrites dans le chapitre précédent, nous avons développé un logiciel de classement et d'amélioration de codes.

41) Structure générale du logiciel

Le logiciel est composé de trois parties distinctes et relativement indépendantes (fig.IV,). Une description succinte est donnée ci dessous.

Partie initialisation :

Ce module permet de configurer le logiciel aux données de production à traiter. Les données à fournir à ce module sont :

- la structure du code utilisée c'est-à-dire :
  - le nombre total de variables,
  - le nombre de variables qualitatives,
  - le rang dans le code et le nombre de modalités des variables qualitatives,
- le nom du fichier contenant les données.

Les données sont introduites dans un fichier par l'intermédiaire d'un éditeur de textes classique, chaque ligne du fichier représente un individu composé de sa classe d'appartenance et de son code.

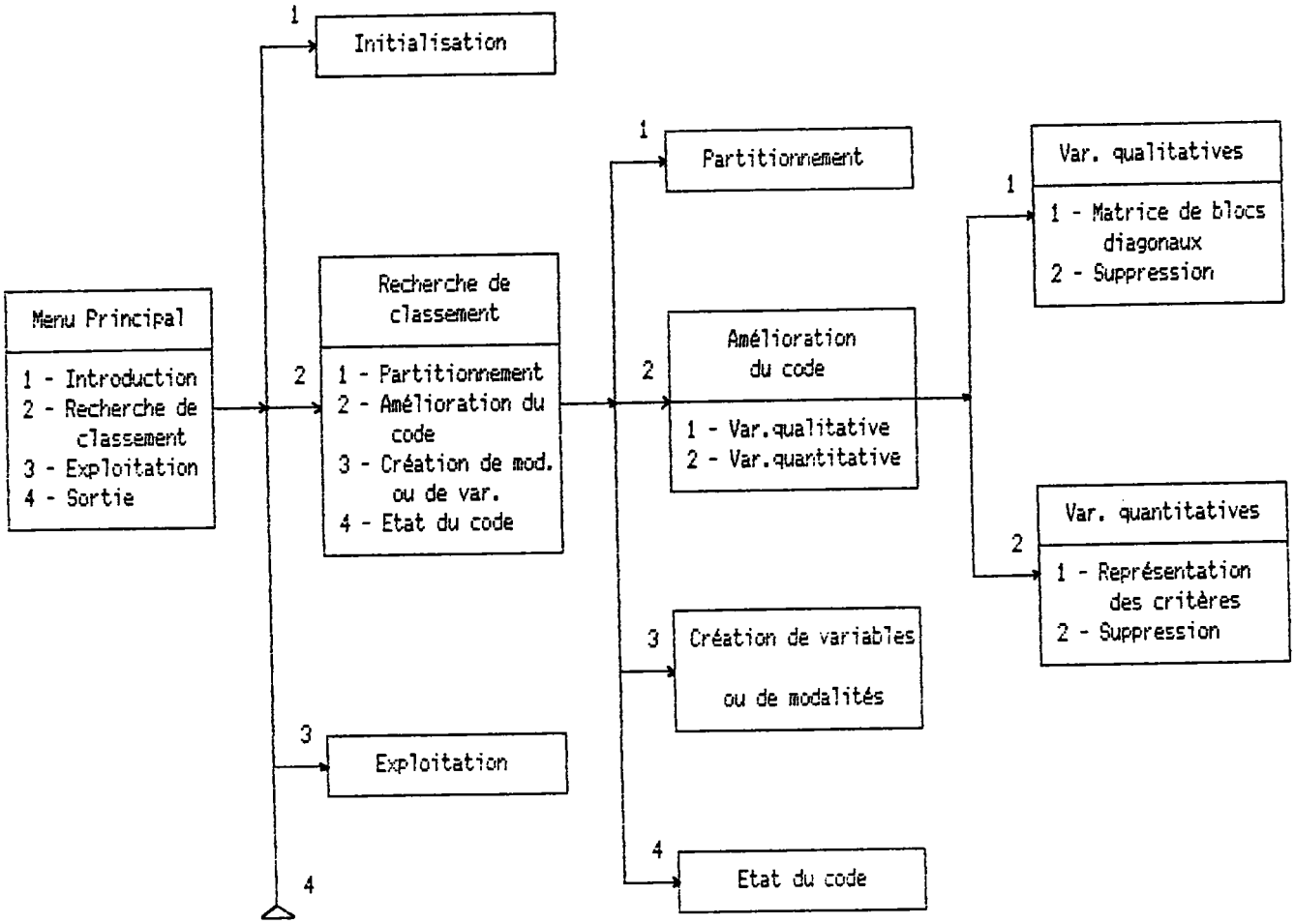


Figure IV$_4$: Structure générale du logiciel.

A la suite de ceci, le logiciel nous fournit le profil de chaque classe, et nous présente graphiquement la répartition des modalités par variable. Nous choisissons ensuite les mesures de distance que nous voulons adopter pour les variables qualitatives et quantitatives. Le logiciel crée alors divers fichiers de travail utilisés par les autres modules.

Partie Recherche de classement :

Ce module permet :

- de réaliser un prétraitement des données qui réalise le partitionnement en régions des classes,

- de simplifier le code de représentation des individus par suppression de modalités et de variables redondantes,

- de créer des variables ou modalités qui ont été supprimées à diverses étapes précédentes,

- d'avoir une représentation du code dans l'état précis.

Partie exploitation :

Cette partie affecte un nouvel individu, très rapidement, à une classe.

Ce logiciel a été implanté sur Micral 9050 et IBM PC, en turbo Pascal. Il permet de traiter des codes de 2 à 25 variables pour une population maximum d'individus de 180.

42) Exemple académique

Ce premier exemple est académique et montre les possibilités de la méthode de classement. Chaque individu est défini par deux variables métriques V_1 et V_2. Nous utilisons une distance euclidienne classique pour mesurer la similarité entre deux individus.

La figure IV_2 montre la représentation des individus par des points de l'espace $V_1 \times V_2$. Ces points sont répartis en trois familles (carrés, croix, cercles).

Dans une première étape, l'algorithme a partitionné chacune des familles de croix et de cercles en deux régions homogènes. Les figures IV_3, IV_4, IV_5 représentent la modélisation des régions homogènes par des boules.

Les caractéristiques des 10 boules (centre, rayon maximal, rayon minimal, composition) et la zone trouble représentent les règles de décision.

43) Exemples industriels

Les deux exemples suivant se rapportent à l'application de la technologie de groupe au bureau des méthodes dans le cadre de l'aide à l'écriture de gammes de fabrication.

Après avoir choisi une population test, chaque individu est défini par : référence, dessin de la pièce mécanique, code, gamme de fabrication. Il a été procédé à une classification afin de regrouper les gammes en classes homogènes.

Le problème est alors de déterminer une fonction de classement qui associera, au code d'un nouveau dessin de pièce, la gamme de fabrication correspondante.

431) Exemple 1 :

a) Les données :

Les données se rapportent à une population de pièces mécaniques de révolution regroupant des pignons, clabots, fourreaux

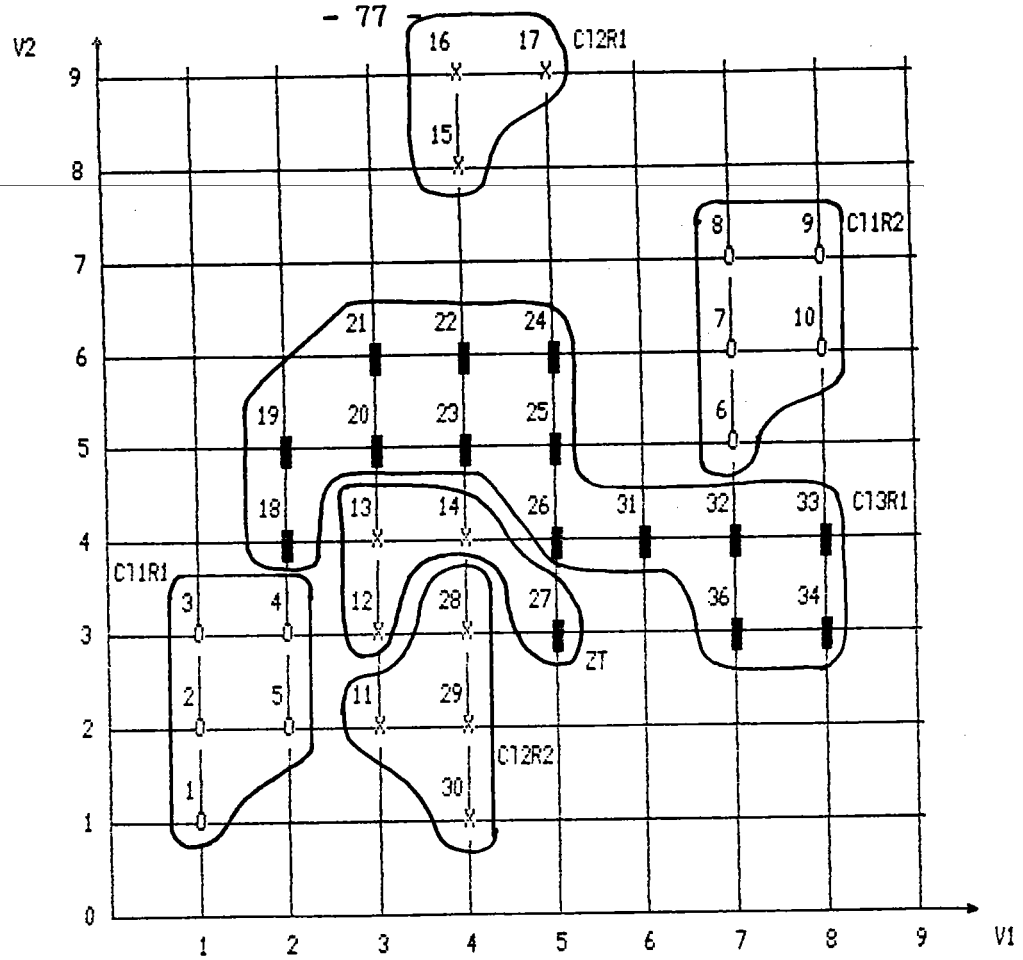


figure IV_2 : Partitionnement en régions des classes.

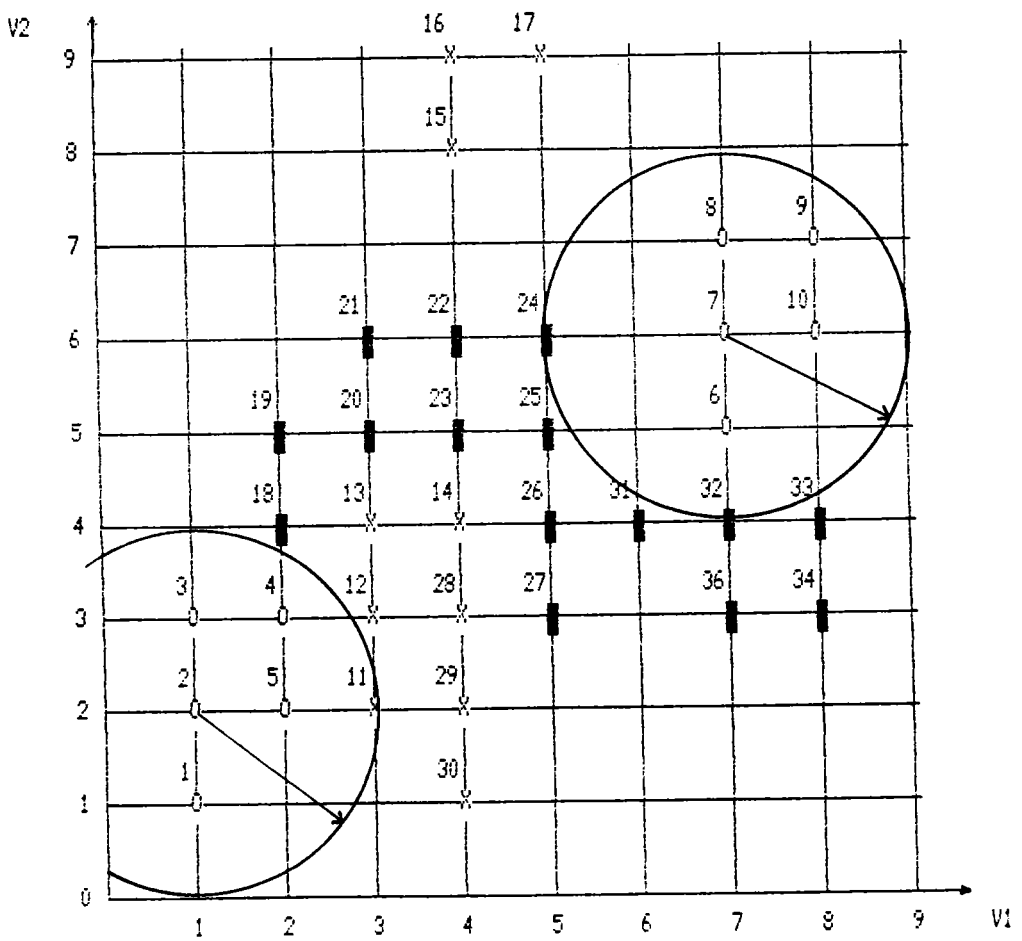


figure IV_3 : Modélisation des régions de la classe des cercles par des boules.
Les boules sont représentées par des cercles de rayon Rmin.

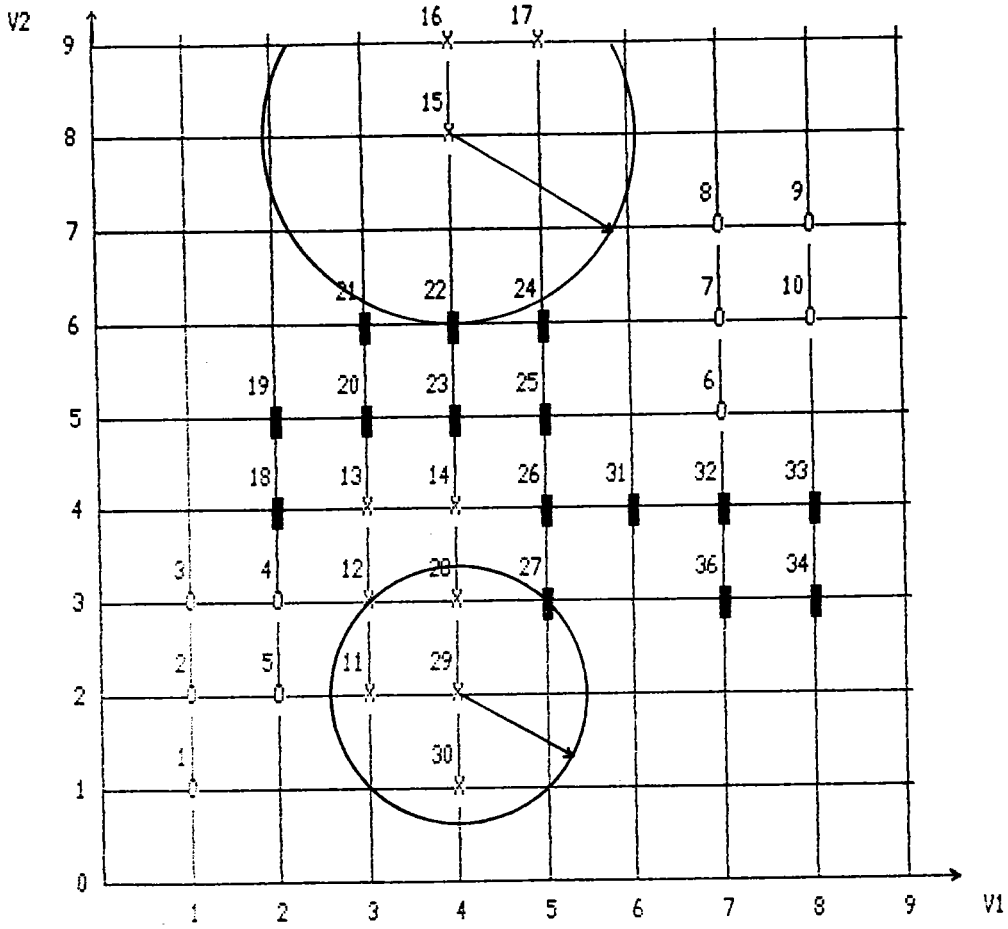


figure IV_4 : Modélisation des régions de croix par des boules.

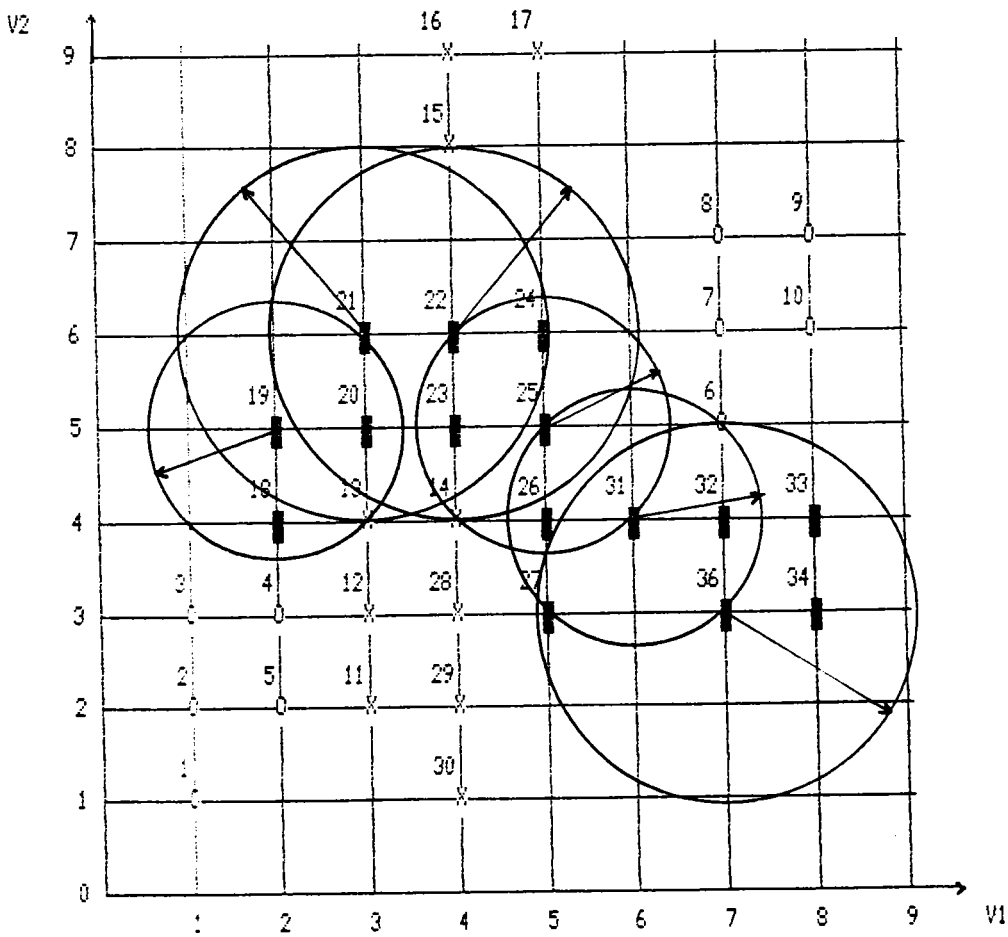


figure IV_5 : Modélisation de la région de la classe des carrés par des boules.

et diverses pièces où les plans de ces pièces ont été codés à l'aide du code OPITZ [46] (fig.IV$_6$) :

variables 1 : Type de pièces,
" 2 : Forme extérieure,
" 3 : Forme intérieure,
" 4 : Etat de surface,
" 5 : Perçage auxiliaire,
" 6 : Diamètre extérieur,
" 7 : Rugosité,
" 8 : Tolérance,
" 9 : Longueur,
" 10 : Série de fabrication.

De ces dix variables, six sont qualitatives (1-2-3-4-5-7) et quatre quantitatives. Chaque variable qualitative possède dix modalités, soit au total soixante modalités.

Un échantillon représentatif de 84 pièces a été prélevé sur la production annuelle de l'entreprise et une classification [14] a donné une partition en quatre classes (fig.IV$_7$) composées respectivement de 32, 14, 11 et 27 pièces.

b) Traitement des données

Nous filtrons, dans un premier temps, les variables qualitatives afin d'éliminer les modalités qui ne contiennent pas d'information (fig.IV$_8$), 29 modalités sont ainsi éliminées.

Nous appliquons ensuite l'algorithme de prétraitement qui fournit un partitionnement de chaque classe en régions (fig.IV$_9$).

La modélisation de ces régions en boules est montrée à la figure IV$_{10}$ ainsi que les caractéristiques de chacune d'entre

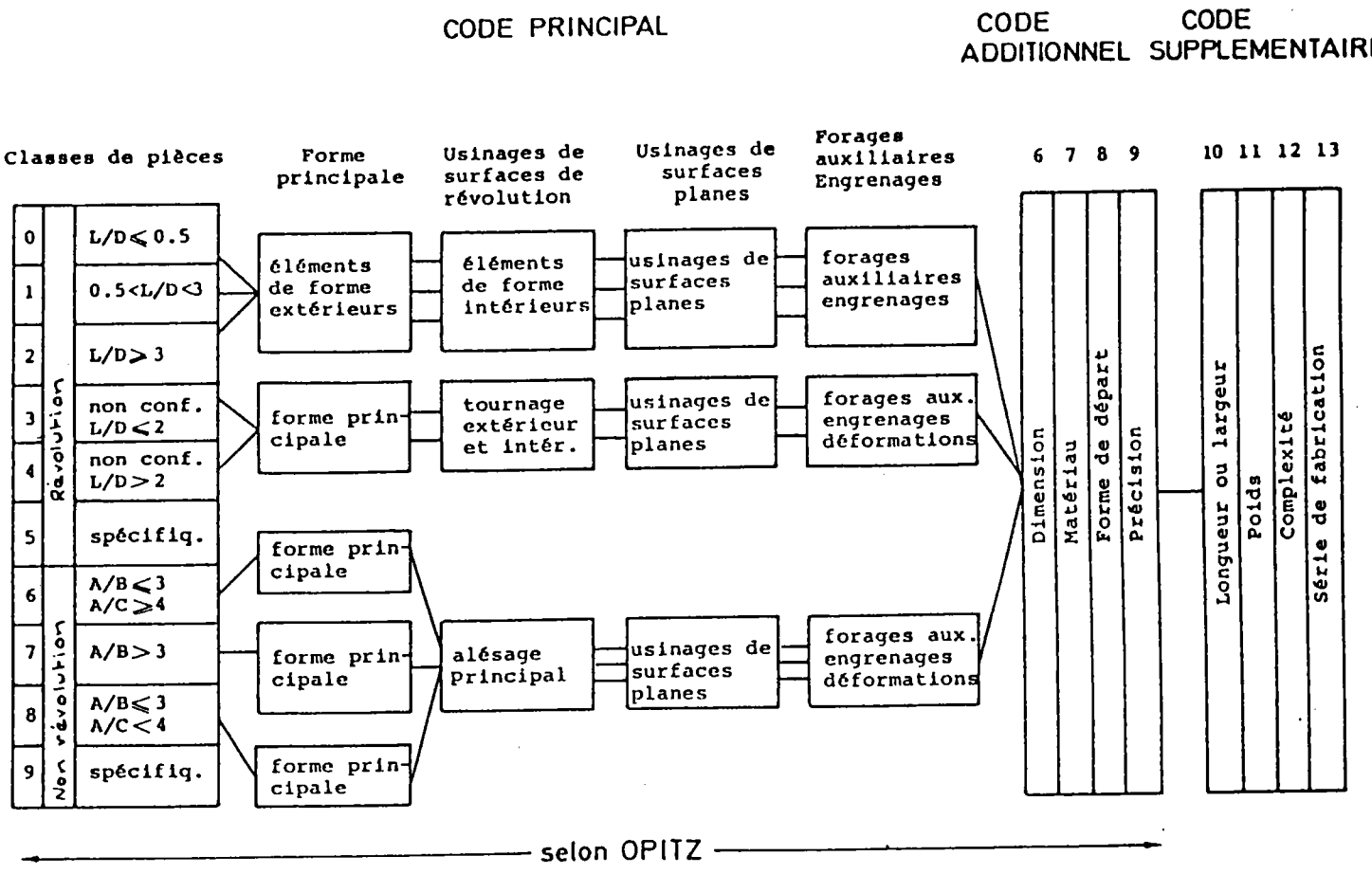


figure IV.4 : Description du code OPITZ.

*** PROFIL DES CLASSES ***

Fichier : data.dat

Classe : 1 Nb d'individus : 32

Composition : 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

Classe : 2 Nb d'individus : 14

Composition : 56 61 62 65 66 68 69 70 71 75 78 82 83 84

Classe : 3 Nb d'individus : 11

Composition : 63 64 67 72 73 74 76 77 79 80 81

Classe : 4 Nb d'individus : 27

Composition : 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 57 58 59 60

figure IV_7 : Répartition des individus par classe (Exemple 1).

**** FILTRAGE ****

sur variables qualitatives

Fichier : data.dat

---

1 - Modalites non utilisees

Var 1 : 3 4 5 6 7 8 9
Var 2 : 2 7 8 9
Var 3 : 3 5 6 7 8 9
Var 4 : 5 8
Var 5 : 1 4 8 9
Var 7 : 3 4 5 7 8 9

2 - Modalites identiques

Var 1 :
Var 2 :
Var 3 :
Var 4 :
Var 5 :
Var 7 :

3 - Modalites complementaires

figure IV_8 : Filtrage des modalités (Exemple 1).

```
*** PARTITIONNEMENT DES INDIVIDUS ***

 la classe numero 1 est composee des individus :
Region 1 : 1 2 3 5 6 9 10 12 13 16 17 18 19 20 22 30 31 15 21 28 29
Region 2 : 7 8 14 24 25 23 26 27

 la classe numero 2 est composee des individus :
Region 1 : 61 68 69 70 66 78
Region 2 : 71 75 82 83 56

 la classe numero 3 est composee des individus :
Region 1 : 63 64 67 72 74 79 73 80 81

 la classe numero 4 est composee des individus :
Region 1 : 33 34 35 46 47 48 54 53
Region 2 : 36 38 41 45 37 49 50 51 52 55
Region 3 : 39 40 42 43 44

la zone trouble est composee des individus :
 4 11 13 32 56 62 65 84 76 77 53 57 58 59 60

          Il y a au total 10 boules
       La zone trouble possede 15 individus
```

figure IV_9 : Partitionnement des classes en régions après filtrage (Exemple 1).

```
*** CARACTERISTIQUES DES 10 BOULES ***

Classe :   1    boule :    2        rmax=1.4E+000  rmin=1.4E+000
     composition :  1  2  3  5  6  9 10 12 16 17 18 19 20 22 30 31 15 21 28 29

Classe :   1    boule :    7        rmax=2.0E-001  rmin=2.5E-001
     composition :  7  8 14 24 25

Classe :   1    boule :   27        rmax=6.2E-001  rmin=7.0E-001
     composition : 24 25 23 26 27

Classe :   2    boule :   78        rmax=2.9E-001  rmin=4.5E-001
     composition : 61 68 69 70 66 78

Classe :   2    boule :   71        rmax=6.8E-001  rmin=1.3E+000
     composition : 71 75 82 83

Classe :   3    boule :   72        rmax=5.9E-001  rmin=6.8E-001
     composition : 63 64 67 72 74 79 73

Classe :   3    boule :   79        rmax=4.1E-001  rmin=5.3E-001
     composition : 63 64 67 74 79 80 81

Classe :   4    boule :   34        rmax=3.0E-001  rmin=3.4E-001
     composition : 33 34 35 46 47 48 54

Classe :   4    boule :   51        rmax=6.7E-001  rmin=6.7E-001
     composition : 36 38 41 45 37 49 50 51 52 55

Classe :   4    boule :   40        rmax=2.8E-001  rmin=4.7E-001
     composition : 39 40 42 43 44
```

figure IV_{10} : Caractéristiques des boules des régions (Exemples 1).

elles (centre, composition, rayon maximum, rayon minimum). Il y a ainsi au total 10 boules et 15 individus en zone trouble.

Afin d'appliquer l'algorithme d'affectation, dans de bonnes conditions, nous recherchons les valeurs limites k_{min} et k_{max} des plus proches voisins (fig.IV$_{11}$-IV$_{12}$) et nous choisissons ainsi les valeurs $k_{min} = 4$ et $k_{max} = 7$.

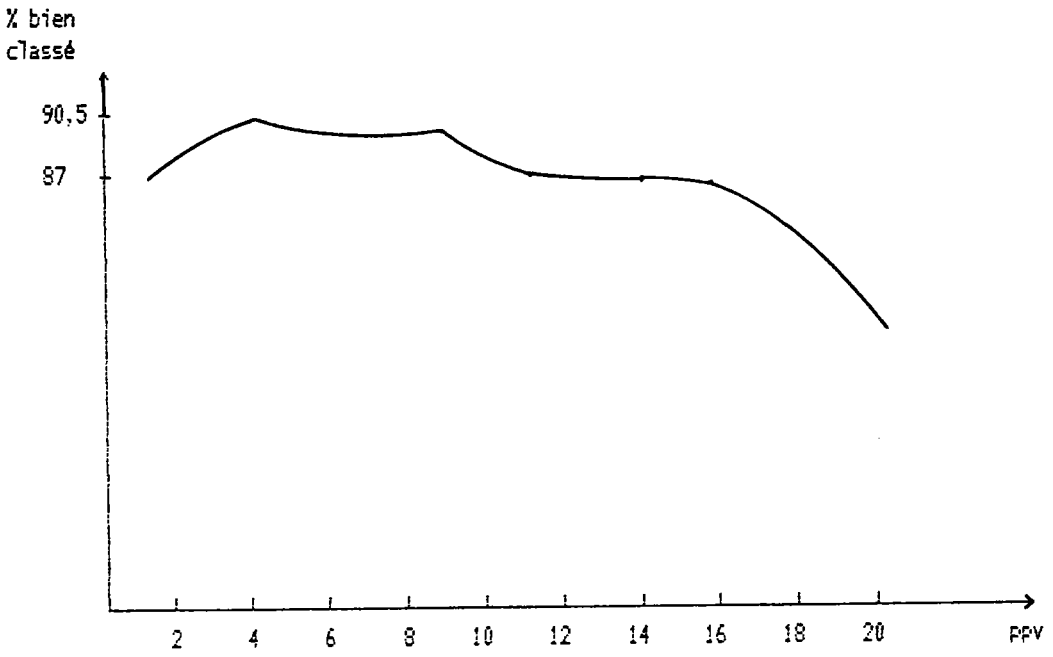
A cette étape, l'algorithme donne un pourcentage de bien classé égal à 88,10 %.

La matrice de corrélation montre que peu de lien existe entre les variables quantitatives.

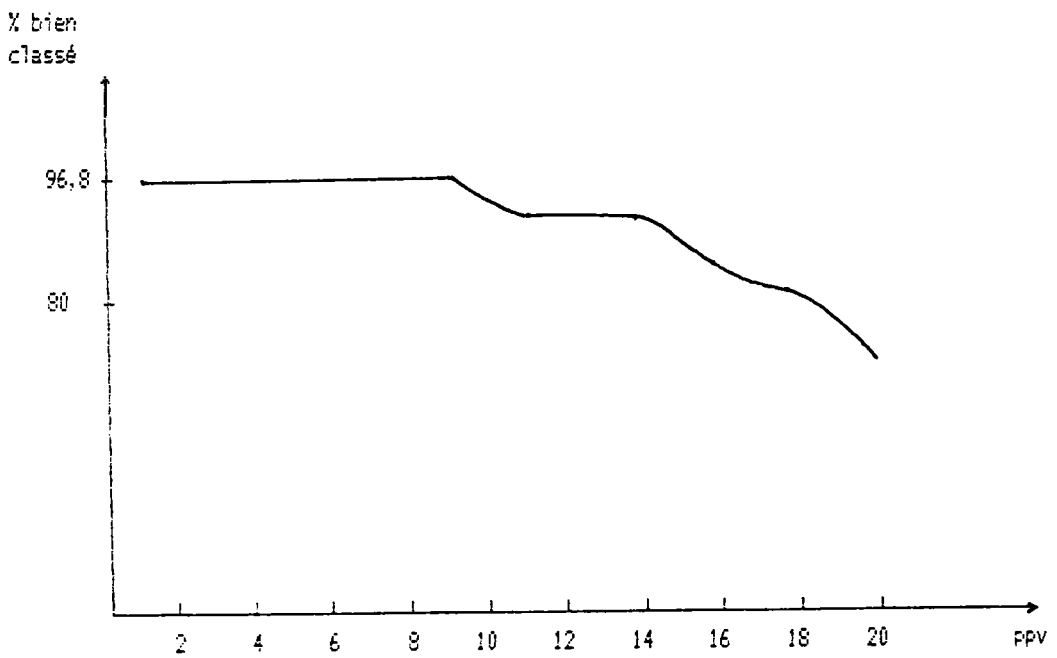
Par contre la variable quantitative 8 possède la plus faible variance intra-classe associée à une grande variance inter-classe (fig.IV$_{13\ b}$).

Après suppression de cette variable, le nouveau partitionnement induit une modélisation des régions en 14 boules et 8 individus en zone trouble. Le pourcentage de bien classé est alors égal à 85,71 %.

Afin de rechercher les modalités peu significatives des variables qualitatives, nous traçons la matrice de blocs diagonaux. La figure IV$_{13}$ représente la matrice avant filtrage et la figure IV$_{14}$ la matrice après filtrage. Les modalités de chaque classe ayant de faibles taux d'affectation sont éliminés (Exemple classe 1 : var.2 mod.ø et var.7 mod.1 qui ont respectivement des taux d'affectation de 64 et 60 sont éliminées). La nouvelle matrice de blocs diagonaux est montrée à la figure IV$_{15}$. Le partitionnement relatif à la structure du code résultant (fig.IV$_{16}$) fournit une modélisation des régions en 15 boules et 5 individus en zone trouble. Ce code résultant comporte 3 variables quantitatives et 12 modalités (fig.IV$_{16a}$). Le pourcentage de bien classé est alors de 92,86 %. Ceci montre que nous avons éliminé l'information bruitée non pertinente.

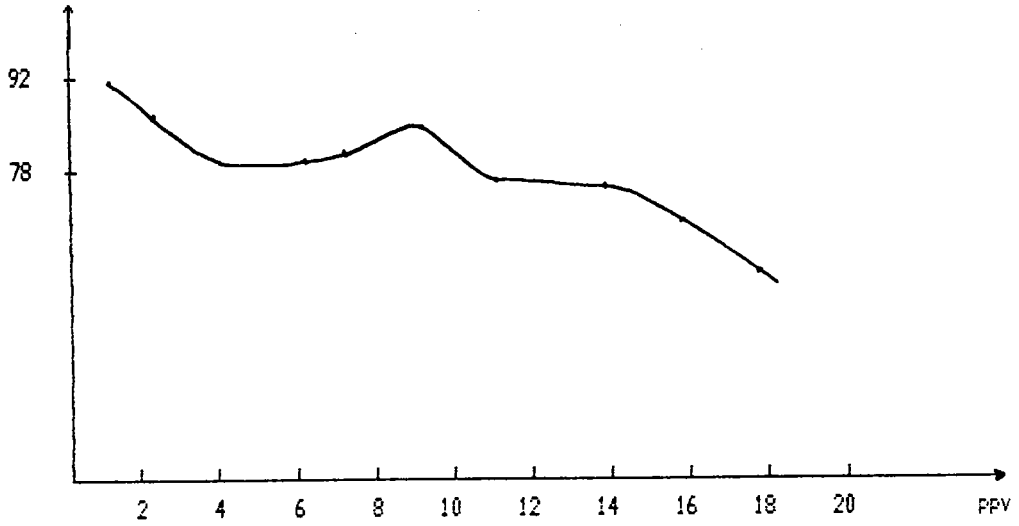


Pourcentage total de bien classé.

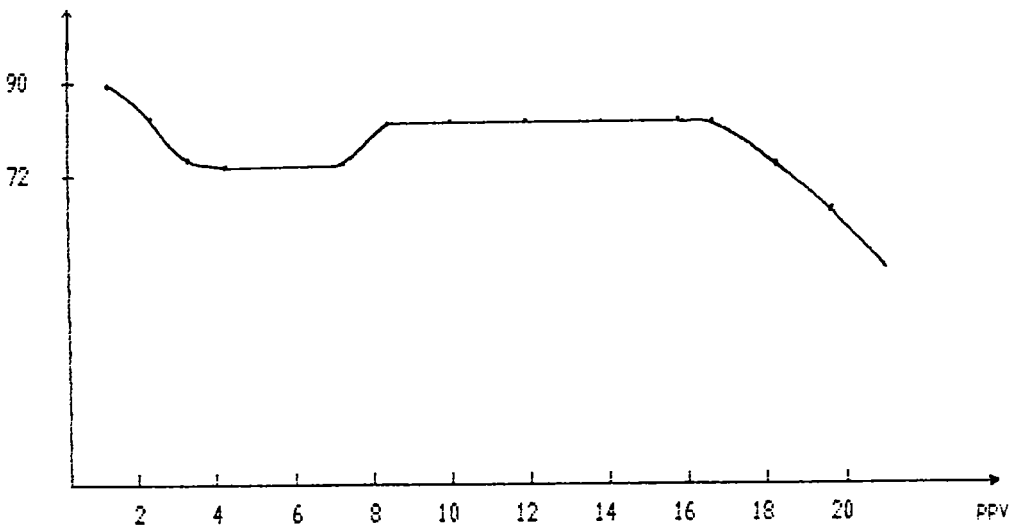


Pourcentage de bien classé de la classe 1.

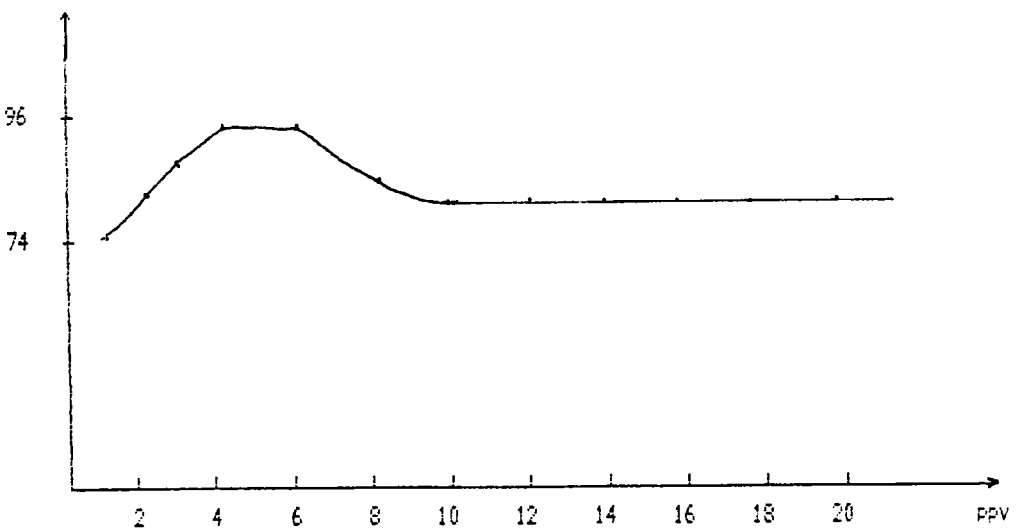
figure IV_{44} : Pourcentage de bien classé en fonction du nombre de plus proches voisins. (Exemple 1)



Pourcentage de bien classé de la classe 2.



Pourcentage de bien classé de la classe 3.



Pourcentage de bien classé de la classe 4.

figure IV_{42} : Pourcentage de bien classé en fonction du nombre de plus proches voisins.

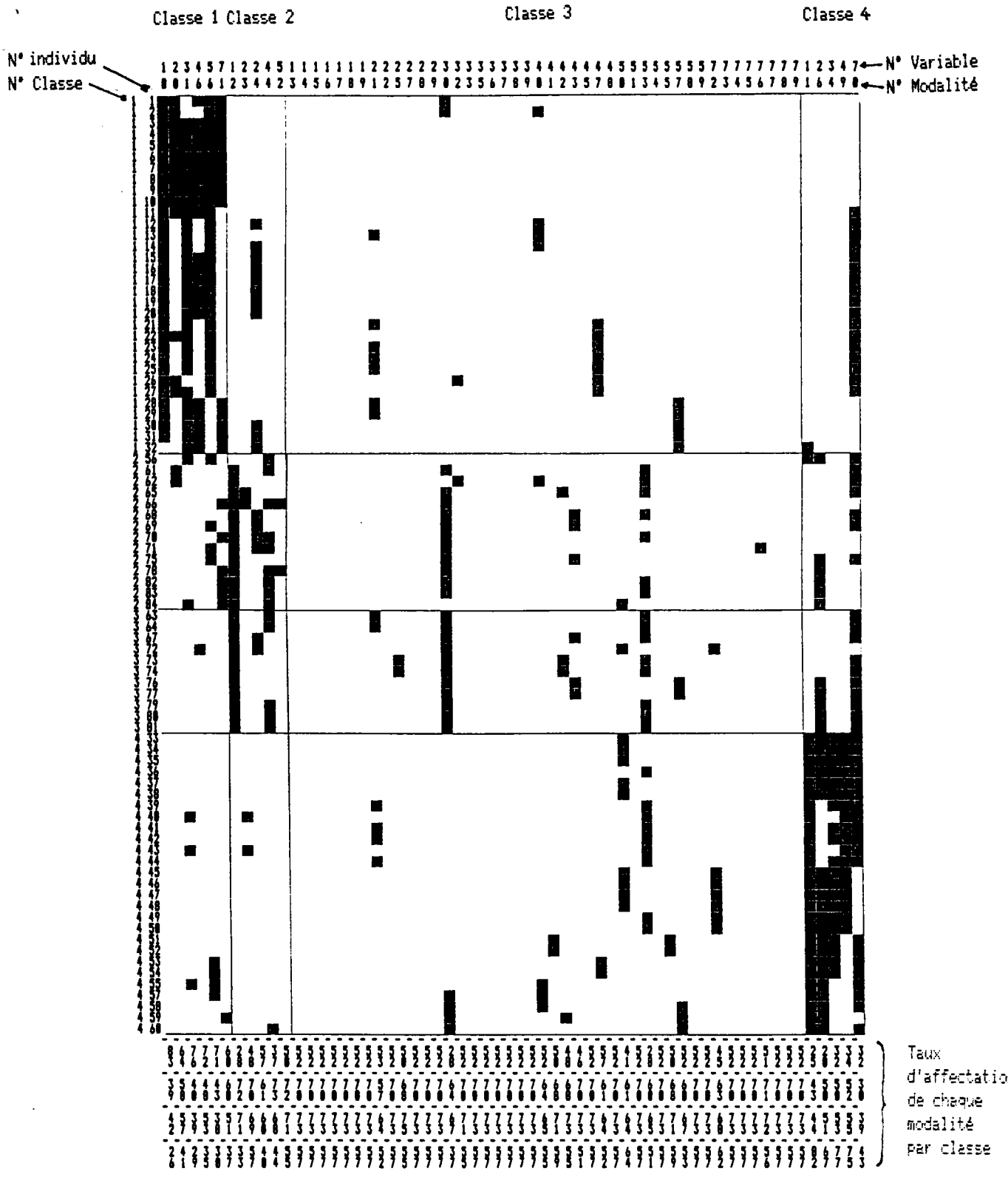


figure IV.$_3$: Matrice de blocs diagonaux avant suppression des modalités (Exemple 1).

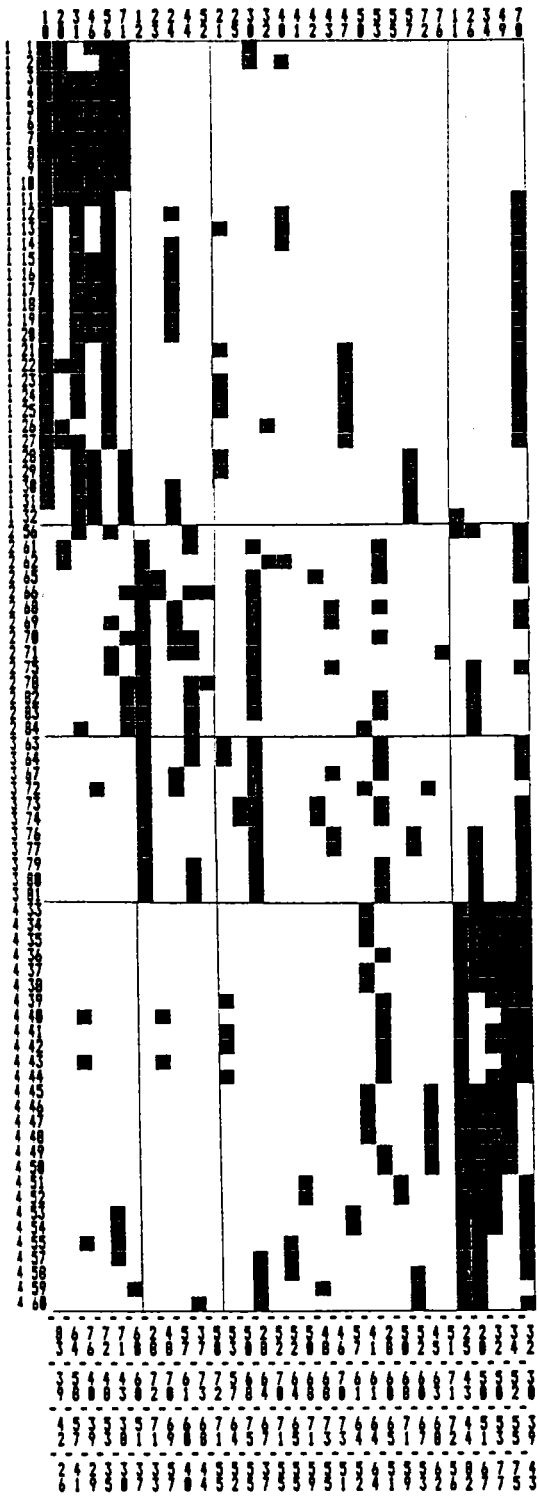


figure IV_{14} : Matrice de blocs diagonaux après filtrage (Exemple 1).

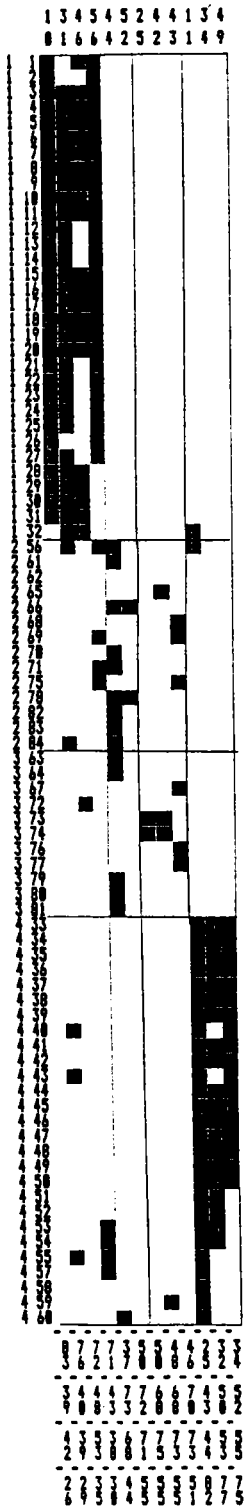


figure IV_{AS} : Matrice de blocs diagonaux des modalités les plus significatives (Exemple 1).

*** ETAT DU CODE ***

---

1 - Variables qualitatives (modalites restantes) :

Var 1 : 0 1
Var 2 : 5
Var 3 : 1 4
Var 4 : 2 3 4 6 9
Var 5 : 2 6
Var 7 :

2 - Variables quantitatives restantes :

Var 6
Var 9
Var 10

figure IV_{16a}: Etat final du code (Exemple 1).

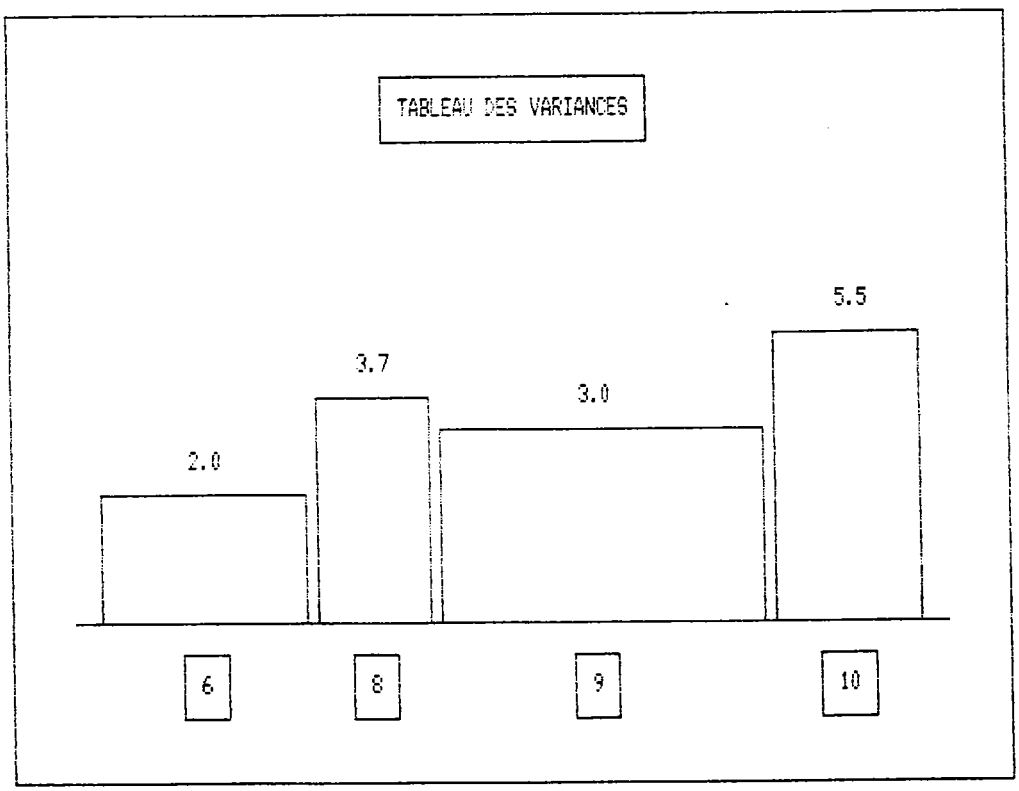


fig IV_{16}b : Tableau des variances (Exemple 1)
En ordonnée est représentée la moyenne et
en abscisse le rapport Var inter .
Var intra

Le tableau récapitulatif du code est le suivant :

Variable	Désignation	Nombre de modalités initiales	Modalités restantes
1	Type de pièces	10	0,1
2	Forme extérieure	10	5
3	Forme intérieure	10	1,4
4	Etat de surface	10	2,3,4,6,9
5	Perçage auxiliaire	10	2,6
6	Diamètre extérieur	Quantitative restante	
7	Rugosité	10	supprimée
8	Tolérance	Quantitative supprimée	
9	Longueur	Quantitative restante	
10	Série de fabrication	Quantitative restante	

Les modalités représentatives pour chaque classe sont ainsi :

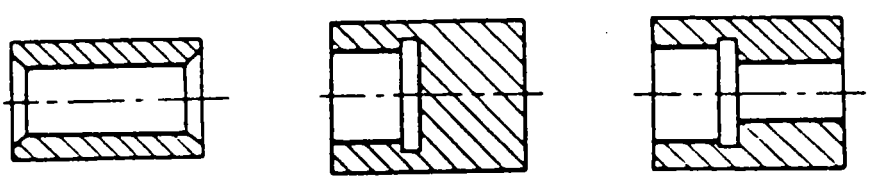
classe 1

- Variable 1 - modalité Ø :

Pièces de révolution dont le rapport longueur sur diamètre est inférieur ou égal à 0,5.

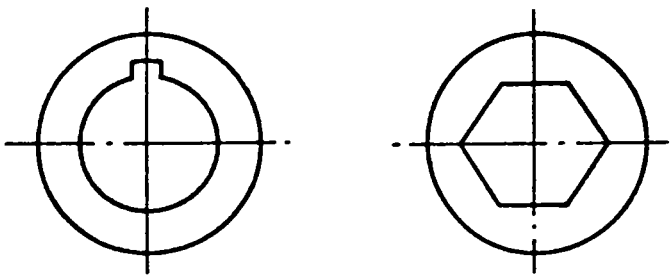
- Variable 3 - modalité 1 :

Forme intérieure de ce type =

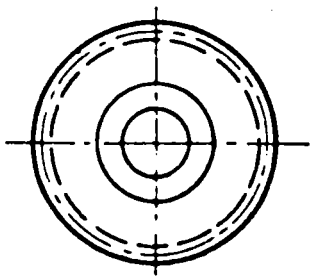


- Variable 4 - modalité 6 :

Surface intérieure avec plat et ou rainure.



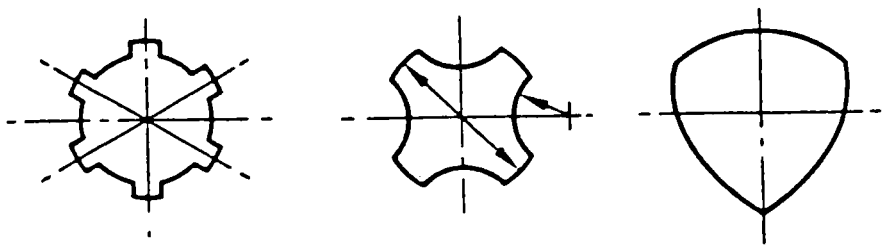
- Variable 5 - modalité 6 :



classe 2

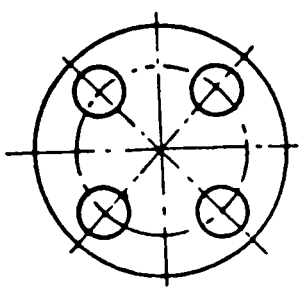
- Variable 4 - modalité 4 :

Surface extérieure courbe ou polygonale.



- Variable 5 - modalité 2 :

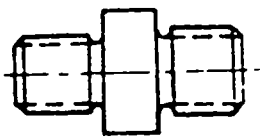
Trous usinés dans l'axe avec symétrique.



classe 3

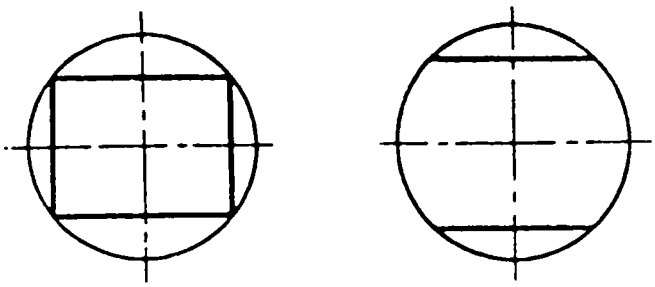
- variable 2 - modalité 5 :

Forme extérieure avec filetage.



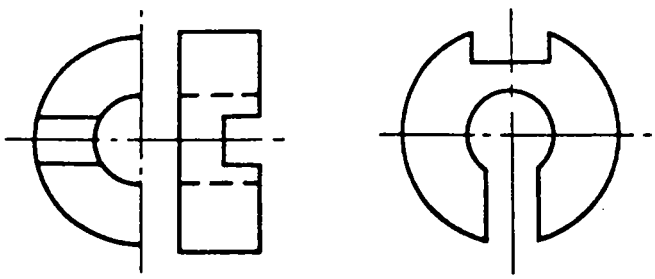
- Variable 4 - modalité 2 :

Surface extérieure plane avec symétrie radiale.



- Variable 4 - modalité 3 :

Rainure extérieure droite.



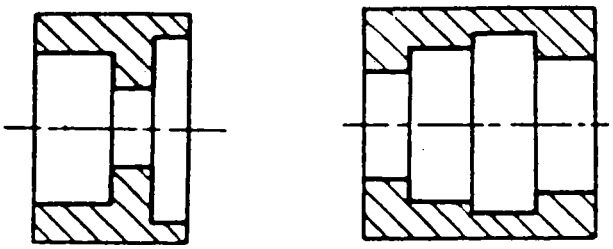
classe 4

- Variable 1 - modalité 1 :

Pièces de révolution avec rapport longueur sur diamètre compris entre 0,5 et 3.

- Variable 3 - modalité 4 :

Tous les diamètres intérieurs non visibles d'un seul côté avec possibilité de gorge.



- Variable 4 - modalité 9 :

Surface plane usinée intérieure ou extérieure obliques et sans symétrie.

Le tableau récapitulatif du code et du pourcentage de bien classé est le suivant :

Code	Etat du code	Pourcentage de bien classé
initial	60 modalités 4 var. quantitatives	88,10 %
filtré	31 modalités 4 var. quantitatives	88,10 %
Suppression de la variable 8	31 modalités 3 var. quantitatives	85,71 %
Elimination des modalités peu pertinentes	12 modalités 3 var. quantitatives	92,86 %

432) Exemple 2 :

a) Les données

Ce deuxième exemple concerne un échantillon de 70 pièces mécaniques de formes parallélépipèdiques. Les dessins de définition de ces pièces ont été codés à l'aide du code Multi-M du système Multiclass [47]. Le code Multi-M est composé de 31 variables réparties en deux codes, un code universel de 18 variables et un code spécifique de 13 variables. Les pièces sont codées à l'aide des 19 premières variables définies comme suit :

Variable 1 : module		
" 2		
" 3		
		Forme
" 4		
" 5		
" 6		
		Description de la fonction
" 7		
" 8		
" 9		
" 10		
		Dimensions
" 11		
" 12		
" 13		
" 14		Tolérance
" 15		
		Matière
" 16		
" 17 :		Forme du brut
" 18 :		Quantité
" 19 :		Complexité

Chacune des variables comporte 10 modalités. Certaines modalités (exemple Var : 15-16) sont couplées et la variable correspondante comporte en fait 100 modalités.

Variable	Nature	Nb modalités
1	Module	10
2	Forme	10
3		10
4		10
5		10
6	Fonction	100
7	Dimensions	Quantitative
8		Quantitative
9		Quantitative
10	Tolérance	10
11	Matière	100
12	Forme de brut	10
13	Quantité	Quantitative
14	Complexité	10

En résumé, le code comporte 14 variables, dont 10 sont qualitatives (280 modalités) et 4 quantitatives.

Une classification a été exécutée sur ces 70 pièces [48]. Elle a donnée 3 classes de 21,6 et 43 individus respectivement (fig.IV$_{17}$).

b) traitement des données

Le filtrage (fig.IV$_{18}$) permet de supprimer dans un premier temps :

- 229 modalités non utilisées,

```
*** PROFIL DES CLASSES ***

Fichier :  dat2.dat

Classe :  1     Nb d'individus :  21

     Composition :  1  2  5  6 14 27 28 31 32 33 34 35 38 39 46 55 56 57 58 61 62

Classe :  2     Nb d'individus :  6

     Composition :  16 22 24 25 37 47

Classe :  3     Nb d'individus :  43

     Composition :  3  4  7  8  9 10 11 12 13 15 17 18 19 20 21 23 26 29 30 36 40 41 42 43 44 45 48 49 50 51
                    52 53 54 59 60 63 64 65 66 67 68 69 70
```

figure IV$_{16}$: Répartition des individus par classe (Exemple 2).

```
**** FILTRAGE ****
sur variables qualitatives

Fichier :  dat2.dat

1 - Modalites non utilisees

     Var  1 : 0 2 3 4 5 6 7 8 9
     Var  2 : 0 1 2 3 4 5 6 9
     Var  3 : 3 4
     Var  4 : 0 5 7 8
     Var  5 : 0 4 5 6 7 8 9
     Var  6 : 0 2 3 4 5 6 7 8 9 12 14 15 19 20 21 22 23 24 25 26
              27 28 29 30 31 32 33 34 35 36 37 38 39 44 46 47 49 50 51 52 53
              54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 73 74 75
              76 77 78 79 80 81 82 85 86 87 88 89 90 91 92 93 94 95 96 97 98
              99
     Var 10 : 0 1 2 3 4 5 6 7 8
     Var 11 : 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
              20 21 22 23 24 25 26 27 28 29 35 36 37 38 39 40 41 42 43 44 45
              46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 63 64 65 66 67
              68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88
              89 90 91 92 93 94 95 96 97 98 99
     Var 12 : 0 2 3 5 6 7 8 9
     Var 14 : 0 1 2 4

2 - Modalites identiques

     Var  1 : 1
     Var  2 :
     Var  3 :
     Var  4 :
     Var  5 :
     Var  6 :
     Var 10 : 9
     Var 11 :
     Var 12 :
     Var 14 :

3 - Modalites complementaires

     Var  2 mod  7 et Var  2 mod  8
     Var  2 mod  7 et Var  3 mod  0
     Var  6 mod 83 et Var 12 mod  4
     Var 11 mod 34 et Var 12 mod  4
     Var 12 mod  1 et Var 12 mod  4
```

figure IV$_{17}$: Filtrage des modalités (Exemple 2).

- 2 modalités communes à tous les individus,

- 7 modalités complémentaires.

Le code ne comporte plus que 42 modalités et 4 variables quantitatives (fig.IV$_{19}$).

Un premier traitement fournit un partitionnement et une modélisation des régions en 11 boules et 30 individus en zone trouble (fig.IV$_{20}$-figIV$_{21}$).

Afin d'utiliser l'algorithme d'affectation, nous traçons les différentes courbes de pourcentage de bien classé en fonction du nombre de plus proches voisins. Nous retenons les valeurs k_{min}=2 et k_{max}=5 (fig.IV$_{22}$ - IV$_{23}$). Après filtrage des variables qualitatives le pourcentage de bien classé est de 72,86 %.

La matrice des blocs diagonaux (fig.IV$_{24}$) nous permet d'éliminer les modalités ayant pour chaque classe un faible taux d'affectation. De plus les matrices de blocs diagonaux calculées avec différents coefficients de pondération (fig.IV$_{25}$) montrent que les blocs sont modifiés. Aussi nous éliminons ces modalités qui passent d'un bloc à l'autre (exemple variable 3 modalité 1 - variable 4 modalité 6 - variable 4 modalité 6 ...etc...). Le code simplifié comporte maintenant 11 modalités et 4 variables quantitatives (figIV$_{26}$). Un nouveau partitionnement (fig.IV$_{29}$-IV$_{29}$) fournit une modélisation en 8 boules et 33 individus en zone trouble. Le pourcentage de bien classé est alors de 71,43 %. La matrice de corrélation (fig.IV$_{27}$) sur les variables quantitatives indique que la variable 9 est très corrélée. Après suppression de cette variable et de la variable 8 qui a des coefficients de corrélation relativement élevés, le partitionnement éxécuté, la modélisation des régions est réalisée en 12 boules et 25 individus en zone trouble (fig.IV$_{30}$-IV$_{32}$). L'affectation des individus est obtenue alors avec 78,57 pourcent de bien classé.

Le code comporte alors 11 modalités et 2 variables quantitiatives (fig.IV$_{31}$).

```
*** PARTITIONNEMENT DES INDIVIDUS ***

 la classe numero 1 est composee des individus :
Region 1 : 1 5 14

 la classe numero 2 est composee des individus :
Region 1 : 16 24 25 22

 la classe numero 3 est composee des individus :
Region 1 : 3 30 51 29 43
Region 2 : 9 10 15 17 19 20 36 44 63 64 68 70 13 21 59 60
Region 3 : 48 49 50
Region 4 : 11 53 54 65 66 67 7 8 52

la zone trouble est composee des individus :
 2 6 27 28 31 32 33 34 35 38 39 46 55 56 57 58 61 62 37 47 4 12 18 23 26 40 41 4
2 45 69

              Il y a au total 11 boules
              La zone trouble possede 30 individus
```

figure IV_{20} : Partitionnement des classes en régions après filtrage (Exemple 2).

```
*** ETAT DU CODE ***

1 - Variables qualitatives (modalites restantes) :

      Var  1 :
      Var  2 :
      Var  3 :  1 2 5 6 7 8 9
      Var  4 :  1 2 3 4 6 9
      Var  5 :  1 2 3
      Var  6 :  1 10 11 13 16 17 18 40 41 42 43 45 48 72 84
      Var 10 :
      Var 11 :  30 31 32 33 62
      Var 12 :
      Var 14 :  3 5 6 7 8 9

2 - Variables quantitatives restantes :

      Var 7
      Var 8
      Var 9
      Var 13
```

figure IV_{21} : Structure du code après filtrage (42 modalités et 4 variables quantitatives) (Exemple 2).

*** CARACTERISTIQUES DES 11 BOULES ***

Classe : 1 boule : 1 rmax=6.0E-001 rmin=6.8E-001
composition : 1 5 14

Classe : 2 boule : 16 rmax=6.5E-001 rmin=8.9E-001
composition : 16 24 25

Classe : 2 boule : 24 rmax=5.2E-001 rmin=6.3E-001
composition : 24 25 22

Classe : 3 boule : 3 rmax=5.3E-001 rmin=5.5E-001
composition : 3 30 51

Classe : 3 boule : 30 rmax=5.5E-001 rmin=5.6E-001
composition : 3 30 29 43

Classe : 3 boule : 9 rmax=6.4E-001 rmin=6.4E-001
composition : 9 10 15 17 19 20 36 44 63 64 68 70

Classe : 3 boule : 20 rmax=4.4E-001 rmin=5.0E-001
composition : 17 20 44 21

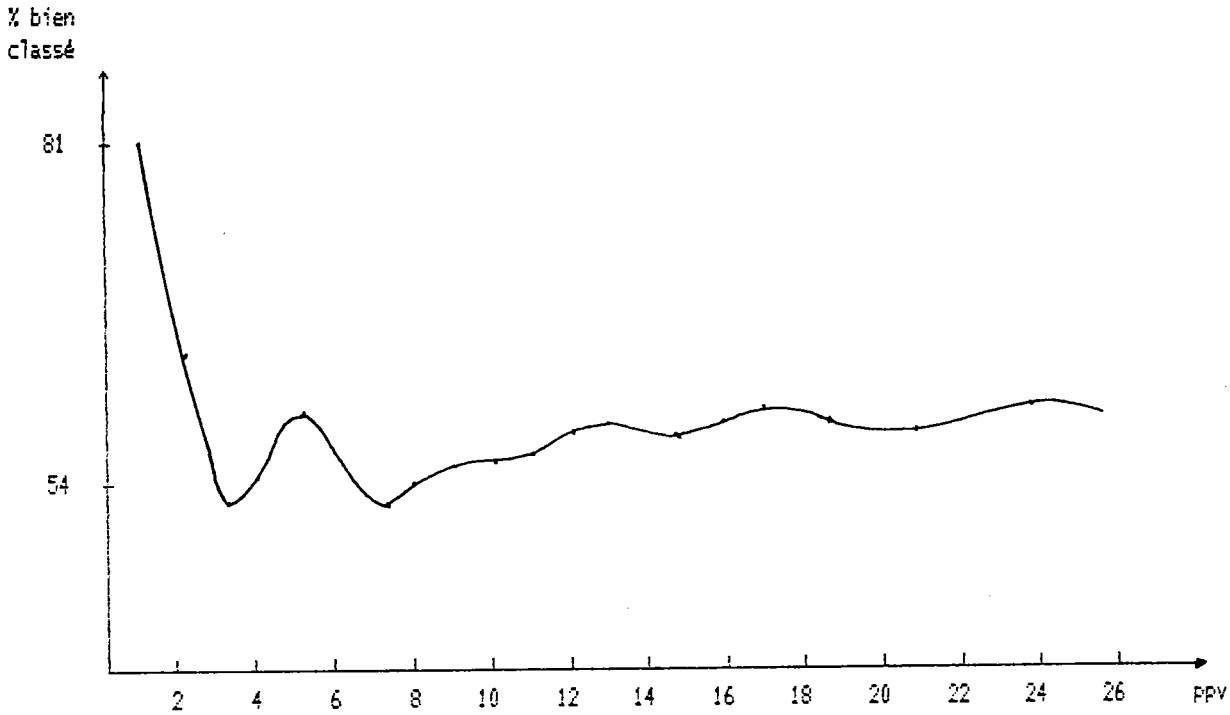
Classe : 3 boule : 13 rmax=1.9E-001 rmin=4.8E-001
composition : 15 13

Classe : 3 boule : 59 rmax=1.4E-001 rmin=3.6E-001
composition : 59 60

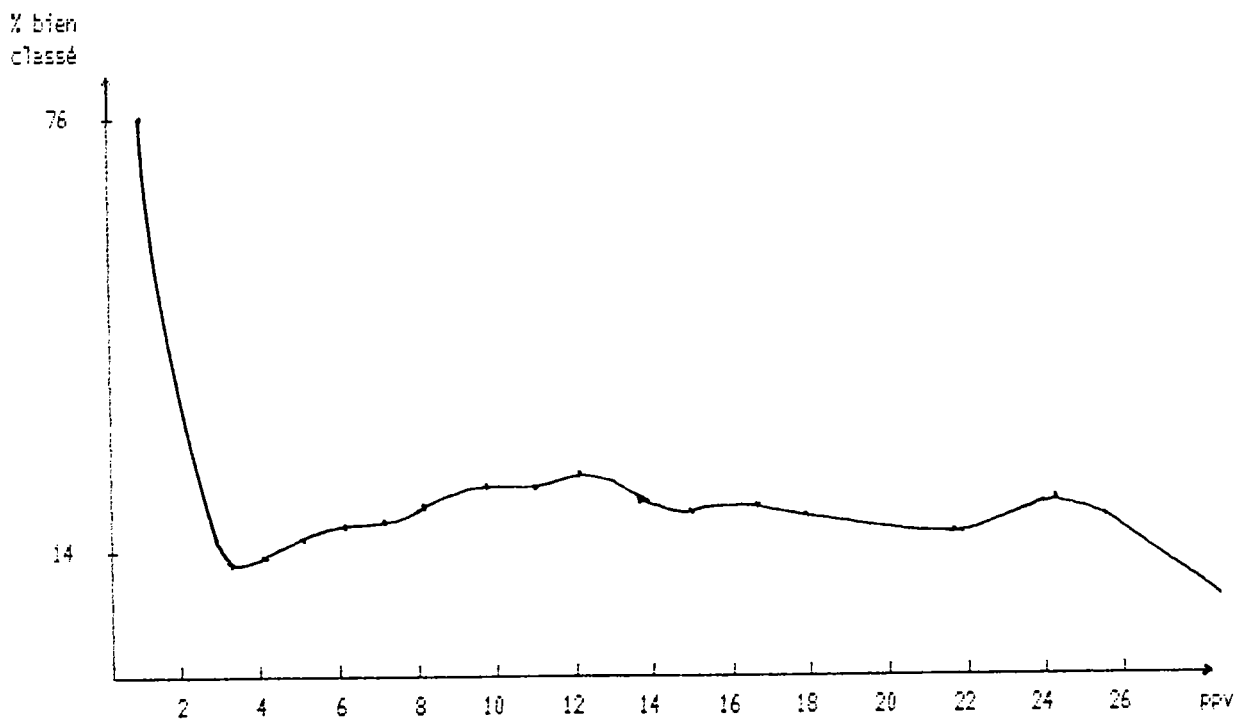
Classe : 3 boule : 48 rmax=3.1E-001 rmin=3.9E-001
composition : 48 49 50

Classe : 3 boule : 11 rmax=6.7E-001 rmin=6.8E-001
composition : 11 53 54 65 66 67 7 8 52

figure IV_{21} : Caractéristiques des boules après filtrage (Exemple 2).

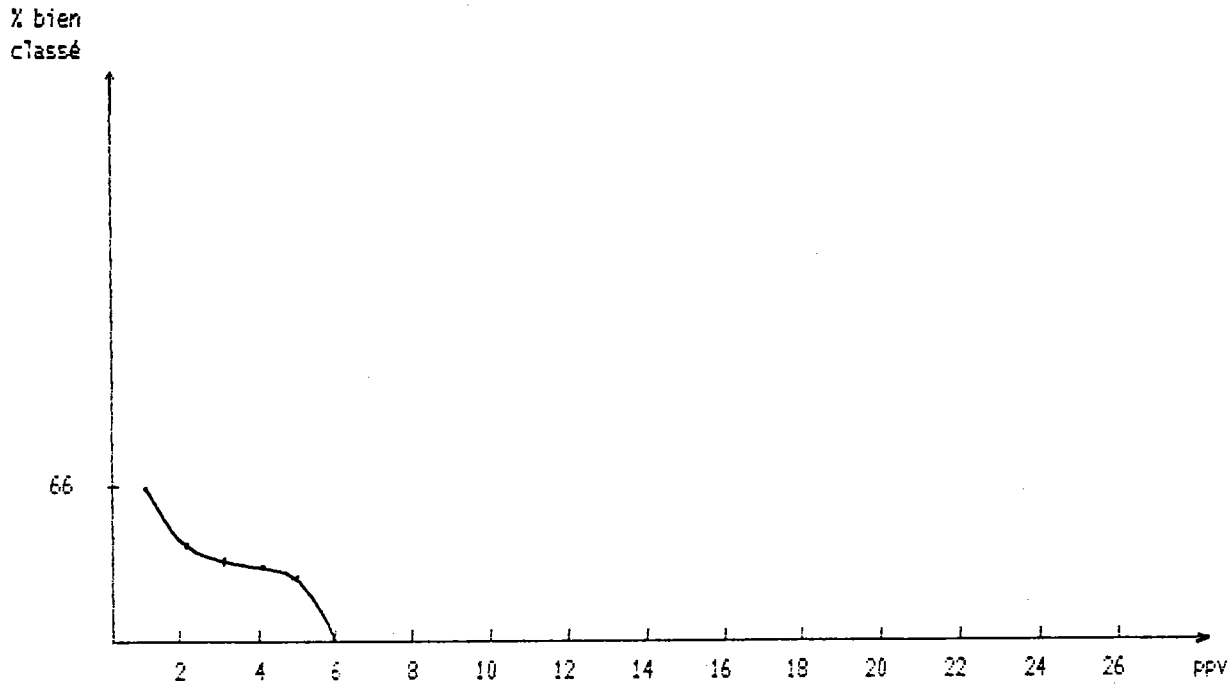


Pourcentage total de bien classé.

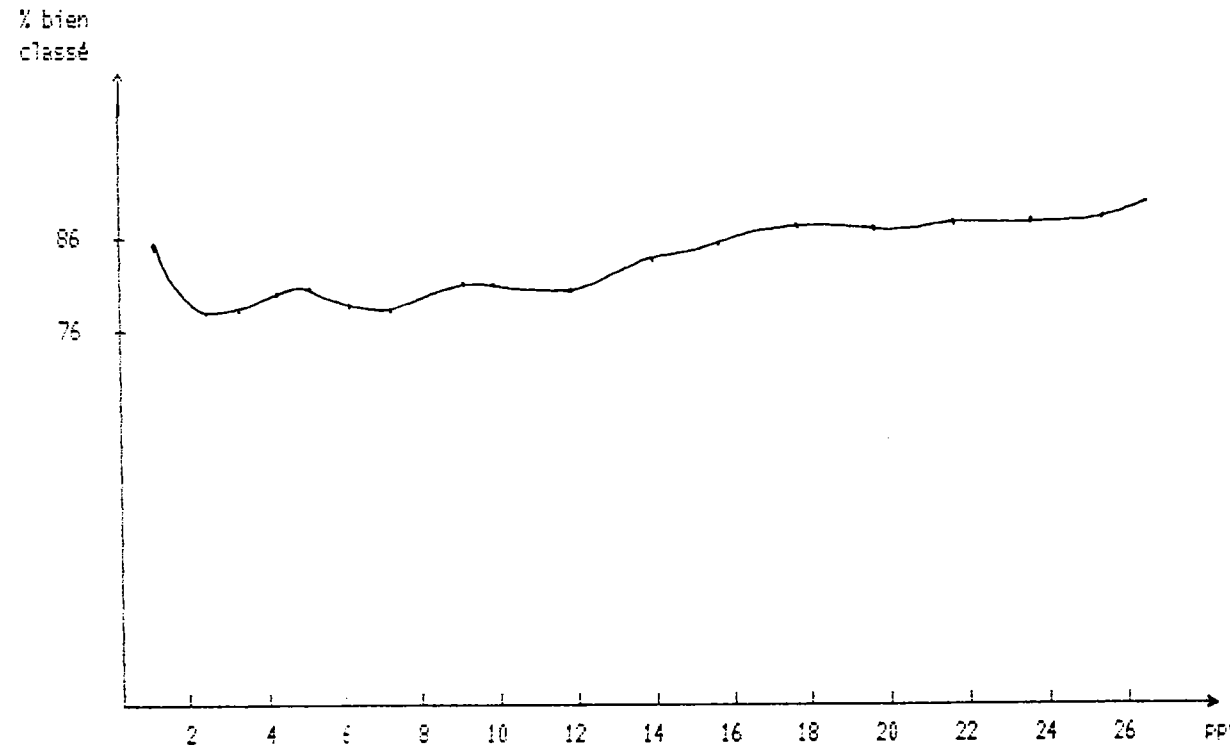


Pourcentage de bien classé de la classe 1.

figure IV_{21} : Courbes du pourcentage de bien classé en fonction du nombre de plus proches voisins. (Exemple 2)



Pourcentage de bien classé de la classe 2.



Pourcentage de bien classé de la classe 3.

figure IV_{28} : Courbes du pourcentage de bien classé en fonction du nombre de plus proches voisins. (Exemple 2)

B.U. METZ

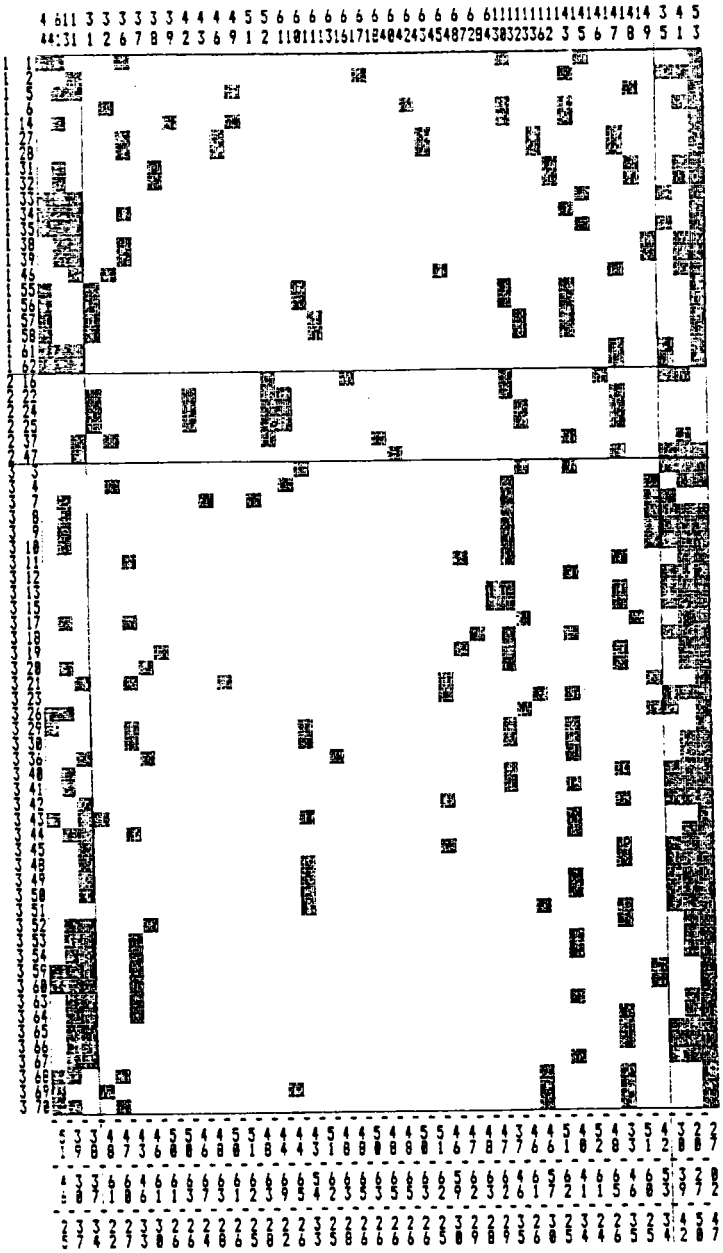


figure IV_{24} : Matrice de blocs diagonaux après filtrage (Exemple 2).

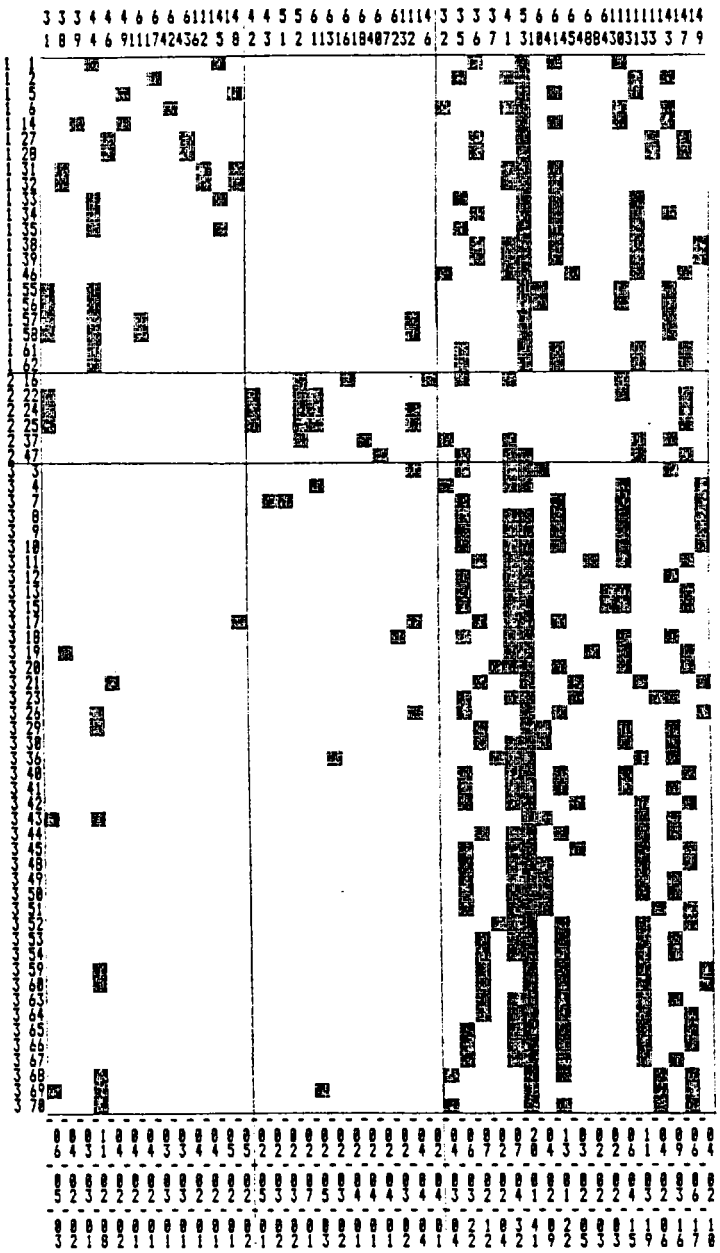


figure IV_{25} : Matrice de blocs diagonaux après filtrage et modification du coefficient de pondération (Exemple 2).

*** MATRICE DE CORRELATION ***

	7	8	9	13
7:	1.00	0.34	0.23	-0.26
8:	0.34	1.00	0.73	-0.43
9:	0.23	0.73	1.00	-0.49
13:	-0.26	-0.43	-0.49	1.00

figure IV_{17} : Matrice de corrélation sur variables quantitatives (Exemple 2).

*** ETAT DU CODE ***

1 - Variables qualitatives (modalites restantes) :

Var 1 :
Var 2 :
Var 3 : 2
Var 4 : 1 2 4
Var 5 : 2 3
Var 6 : 16 18 40 41
Var 10 :
Var 11 :
Var 12 :
Var 14 : 6

2 - Variables quantitatives restantes :

Var 7
Var 8
Var 9
Var 13

figure IV_{16} : Etat du code après filtrage et suppression des modalités peu pertinentes (Exemple 2).

*** PARTITIONNEMENT DES INDIVIDUS ***

```
 la classe numero 1 est composee des individus :
Region 1 : 1 5 14

 la classe numero 2 est composee des individus :
Region 1 : 16 24 25

 la classe numero 3 est composee des individus :
Region 1 : 9 10 13 15 17 19 20 36 44 48 49 50 63 64 68 70 21 59 60
Region 2 : 29 30 43
Region 3 : 11 53 54 65 66 67 7 8 52

la zone trouble est composee des individus :
 2 6 27 28 31 32 33 34 35 38 39 46 55 56 57 58 61 62 22 37 47 3 4 12 18 23 26 40
 41 42 45 51 69

              Il y a au total   8 boules
              La zone trouble possede 33 individus
```

figure IV_{29} : Partitionnement après suppression de modalités (Exemple 2).

*** CARACTERISTIQUES DES 8 BOULES ***

```
Classe :   1     boule :     1        rmax=4.8E-001  rmin=5.8E-001
      composition :  1 5 14

Classe :   2     boule :    16        rmax=4.8E-001  rmin=7.5E-001
      composition :  16 24 25

Classe :   3     boule :     9        rmax=5.3E-001  rmin=5.5E-001
      composition :  9 10 13 15 17 19 20 36 44 48 49 50 63 64 68 70

Classe :   3     boule :    20        rmax=3.5E-001  rmin=3.6E-001
      composition :  17 20 44 21

Classe :   3     boule :    59        rmax=1.4E-001  rmin=2.9E-001
      composition :  59 68

Classe :   3     boule :    29        rmax=3.9E-001  rmin=3.9E-001
      composition :  29 30 43

Classe :   3     boule :    11        rmax=5.0E-001  rmin=5.1E-001
      composition :  11 53 54 65 66  7  8 52

Classe :   3     boule :    67        rmax=8.6E-002  rmin=1.0E-001
      composition :  65 66 67
```

figure IV_{28} : Caractéristiques des boules après suppression de modalités (Exemple 2).

*** PARTITIONNEMENT DES INDIVIDUS ***

```
 la classe numero 1 est composee des individus :
Region 1 : 1 5 14
Region 2 : 27 28 34 57 58 31 32
Region 3 : 2 38 39

 la classe numero 2 est composee des individus :
Region 1 : 16 24 25 22

 la classe numero 3 est composee des individus :
Region 1 : 3 49 50
Region 2 : 7 8 11 12 23 26 52 53 54
Region 3 : 13 15 20 21
Region 4 : 19 40 41 48
Region 5 : 59 60 63 64 36 51 4
Region 6 : 65 66 67

la zone trouble est composee des individus :
 6 33 35 46 55 56 61 62 37 47 4 4 9 10 17 18 29 30 42 43 44 45 68 69 70

              Il y a au total 12 boules
              La zone trouble possede 25 individus
```

figure IV_{30} : Partitionnement en régions des classes après suppression des variables 8 et 9 (Exemple 2).

*** ETAT DU CODE ***

```
1 - Variables qualitatives (modalites restantes) :

      Var  1 :
      Var  2 :
      Var  3 :  2
      Var  4 :  1 2 4
      Var  5 :  2 3
      Var  6 :  16 18 40 41
      Var 10 :
      Var 11 :
      Var 12 :
      Var 14 :  6

2 - Variables quantitatives restantes :

      Var 7
      Var 13
```

figure IV_{31} : Structure du code final (Exemple 2).

*** CARACTERISTIQUES DES 12 BOULES ***

Classe : 1 boule : 14 rmax=3.0E-001 rmin=3.7E-001
composition : 1 5 14

Classe : 1 boule : 27 rmax=1.8E-001 rmin=2.3E-001
composition : 27 28 34 57 58

Classe : 1 boule : 31 rmax=1.6E-001 rmin=1.8E-001
composition : 34 31 32

Classe : 1 boule : 39 rmax=1.8E-001 rmin=1.9E-001
composition : 2 38 39

Classe : 2 boule : 16 rmax=3.7E-001 rmin=5.4E-001
composition : 16 24 25

Classe : 2 boule : 24 rmax=2.6E-001 rmin=3.5E-001
composition : 24 25 22

Classe : 3 boule : 3 rmax=0.0E+000 rmin=1.6E-001
composition : 3 49 50

Classe : 3 boule : 7 rmax=3.7E-001 rmin=4.1E-001
composition : 7 8 11 12 23 26 52 53 54

Classe : 3 boule : 13 rmax=1.9E-001 rmin=2.2E-001
composition : 13 15 20 21

Classe : 3 boule : 19 rmax=2.5E-001 rmin=3.2E-001
composition : 19 40 41 48

Classe : 3 boule : 64 rmax=1.8E-001 rmin=1.8E-001
composition : 59 60 63 64 36 51

Classe : 3 boule : 67 rmax=8.6E-002 rmin=1.0E-001
composition : 65 66 67

figure IV_{32} : Caractéristiques des boules après suppression des variables 8 et 9 (Exemple 2).

Les deux variables quantitatives les plus significatives sont :

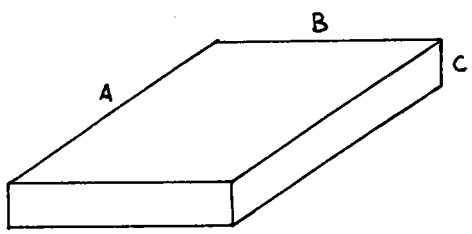
- variable 13 : Quantité produite

- variable 7 : Dimension de la plus grande face pour les pièces parallélépipèdiques.

Les modalités les plus caractéristiques pour chaque classe sont les suivantes (fig.IV$_{33}$) :

classe 1

- Variable 4 modalité 4 :
  Symétrie d'usinage dans les vues BC seul, BC et AB ou BC et AC.



classe 2

- Variable 4 modalité 2 :
  Symétrie d'usinage dans AB seulement;

- Variable 5 modalité 2 :
  Usinage secondaire de plats et ou de rainures.

- Variable 6 :
  Description de la fonction.

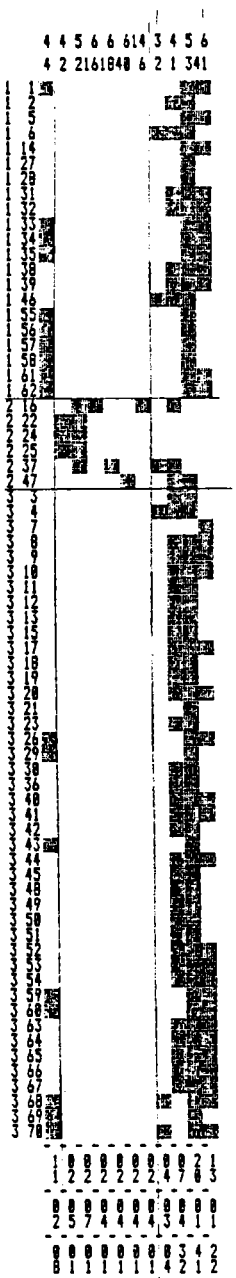


figure IV$_{33}$: Matrice des blocs diagonaux du code final (Exemple 2).

- Variable 14 modalité 6 :
  La variable 14 définit la complexité des usinages secondaires. La modalité 6 indique un axe d'usinage perpendiculaire à AC et ou BC et 2 axes maximum non perpendiculaires.

classe 3

- Variable 3 modalité 2 :
  Le plan AC a une forme de T, de + ou de ×

- Variable 4 modalité 1 :
  Il existe une symétrie dans AB et AC.

- Variable 5 modalité 3 :
  Usinage secondaire de trous plus de plats et ou de rainures.

Le tableau récapitulatif du code est le suivant :

Variable	Désignation	Nombre de modalités initiales	Modalités restantes
1	Module	10	supprimée
2	Forme	10	supprimée
3	Forme	10	2
4	Forme	10	1,2,4
5	Forme	10	2,3
6	Fonction	100	16,18,40,41
7	Dimensions	Quantitative restante	
8	Dimensions	Quantitative supprimée	
9	Dimensions	Quantitative supprimée	
10	Tolérance	10	supprimée
11	Matière	100	supprimée
12	Forme de brut	10	supprimée
13	Quantité	Quantitative restante	
14	Complexité	10	6

L'évolution du pourcentage de bien classé est résumé dans ce tableau :

Code	Etat du code	Pourcentage de bien classé
initial	280 modalités 4 var. quantitatives	72,86 %
filtré	42 modalités 4 var. quantitatives	72,86 %
Elimination des modalités peu pertinentes	11 modalités 4 var. quantitatives	71,43 %
Supression des variables 3 et 9	11 modalités 2 var. quantitatives	78,57 %

Dans cet exemple, M. Nadif [48] en utilisant une méthode d'analyse discriminante linéaire de type géométrique a trouvé un pourcentage de bien classé de 40 %. Ceci permet de juger de l'efficacité de notre méthode pour le classement de ce type de données de production.

CHAPITRE V

CONCLUSION

V CONCLUSION

La représentation des caractéristiques physiques, technologiques des pièces mécaniques, appelée codage, sans distorsion ou perte de l'information est un problème général qui n'a pas trouvé actuellement de solutions satisfaisantes. De nombreux codes de représentation de pièces ont été créés comportant des caractéristiques redondantes.

Au stade de la recherche des classes, il est préférable d'avoir trop d'informations que pas assez. Par contre, les familles de pièces étant identifiées, du point de vue de l'utilisateur, l'emploi de codes plus simples est préférable car, d'une part l'opération de codage de nouvelles pièces sera plus rapide et les risques d'erreurs moindres, d'autre part l'opération de classement sera accélérée, celle-ci permettant de déterminer la classe d'appartenance d'une nouvelle pièce.

Dans les applications de technologie de groupe, on est amené, le plus souvent, à considérer deux ensembles de variables pour représenter une pièce. Par exemple, nous sommes conduit à effectuer une classification sur un ensemble de variables E_1, alors que le classement de nouvelles pièces se fera sur un autre groupe de variables E_2. Les groupes de variables des ensembles E_1 et E_2 sont très peu corrélés. Dans ces conditions, dans l'espace de représentation E_2 des pièces, il y a peu de chances de trouver des zones homogènes convexes alors qu'elles existent naturellement par construction dans l'ensemble E_1. Aussi, les méthodes classiques de classement par discrimination se montrent peu efficace.

Cette étude présente une nouvelle méthode mieux adaptée à ces problèmes de classement en fabrication mécanique, et à la recherche de codes adaptés à chaque application (pièces, gammes, outillage, devis, etc...). Ces codes auront pour rôle de :

- représenter au mieux le produit dans sa famille,

- activer la règle de classement de nouveaux produits,

- faciliter la recherche de modèles de production.

La méthode de classement est basée sur un prétraitement des données de production afin de trouver des zones homogènes et une zone trouble. L'affectation de nouvelles pièces à une famille est mixte, de type géométrique et de type par voisinage.

Lors de la procédure des k plus proches voisins, nous avons été confronté au problème du choix du meilleur k . Ce choix est un problème très délicat qui dépend d'un nombre important de paramètres dont l'influence est difficile à évaluer. Nous avons résolu ce problème en choisissant, à partir d'une courbe du pourcentage de bien classé en fonction du nombre de plus proches voisins, une fourchette de deux valeurs limites entre lesquelles le pourcentage de bien classé reste stable.

Nous avons présenté ensuite, une série de critères permettant la sélection des variables. Pour ce faire, nous utilisons un processus itératif qui permet de supprimer des variables ou modalités en s'arrêtant lorsque le pourcentage de bien classé décroît de façon importante.

Cette étude a donné lieu à l'écriture d'un logiciel de classement et d'amélioration de codes. Ce Logiciel représente le module 4 du système GAAD (Groupement Analogique par l'Analyse de Données) [12] développé au laboratoire L.A.E.I. La structure du système GAAD est représentée à la figure C1 et comprend :

module 1 : ce module est composé d'une base de données relationnelle, dans laquelle on analyse les données de production ;

module 2 : celui-ci comprend un système de classification automatique [48] ;

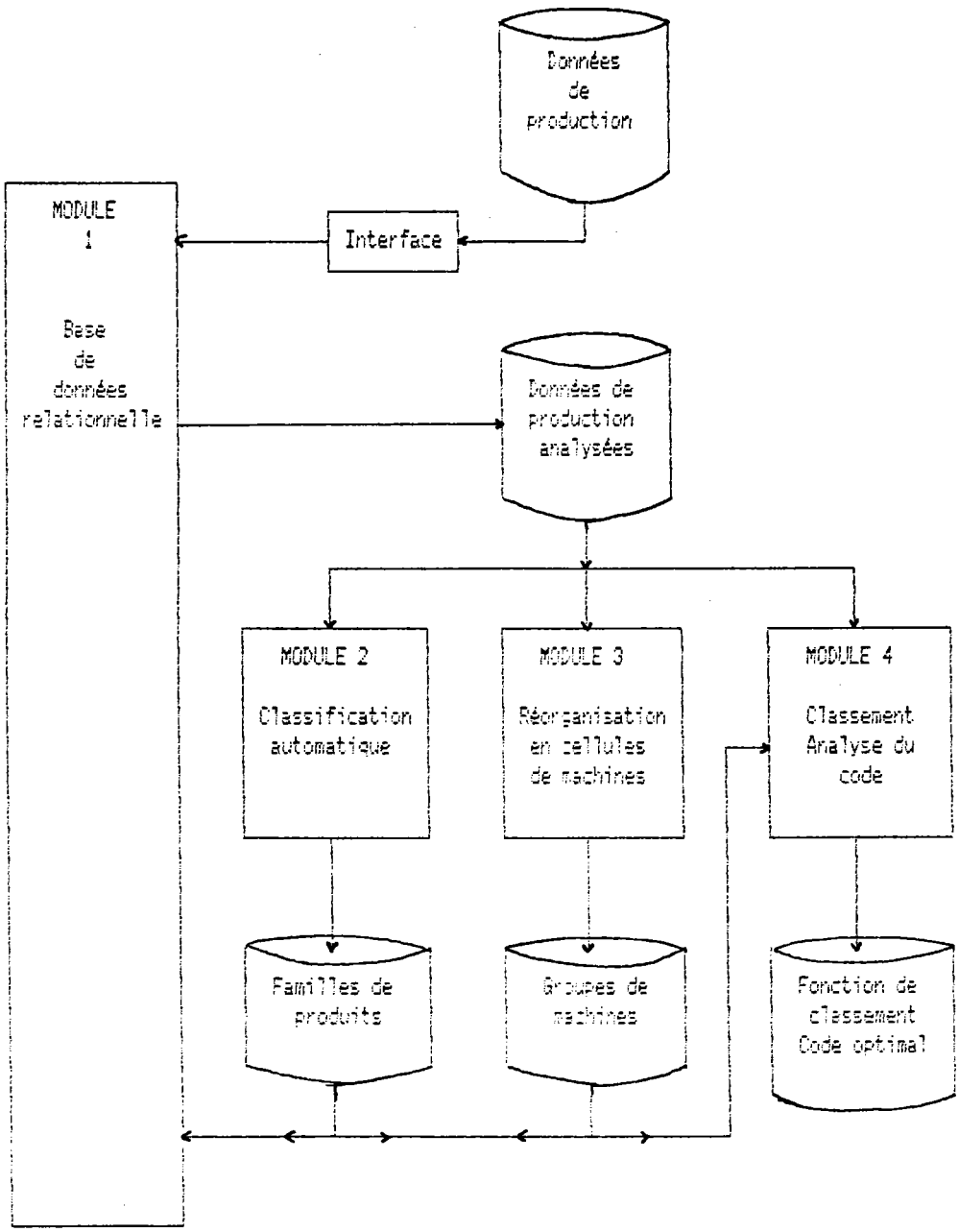


figure C1 : Structure du système GAAD.

module 3 : ce module permet de regrouper les machines en cellules de fabrication [22] ;

module 4 : ce module est l'objet de cette étude qui réalise le classement et l'amélioration de code.

Ce logiciel de classement et d'amélioration de code a été réalisée en diverses parties très indépendantes (prétraitement - affectation - suppression variables quantitatives - suppression de modalités). Ceci permet de l'utiliser avec très peu de modifications sur des micro ordinateurs de taille mémoire réduite.

Les exemples industriels traités ont montré l'efficacité de la méthode de classement retenue et l'intérêt d'optimiser les codes pièces pour une meilleure interprétation des familles de produits.

BIBLIOGRAPHIE

Bibliographie

[1] - DIDAY
Eléments d'analyse des données - Dunod

[2] - SAPORTA G.
Liaisons entre plusieurs ensembles de variables et codage de données qualitatives.
Thèse de 3è cycle Mathématiques appliquées.
Université Pierre et Marie Curie Paris VI (1975)

[3] - SAPORTA G.
Une méthode et un programme d'analyse discriminante pas à pas sur variables qualitatives.
Actes 1ères journées d'analyse de données et Informatiques IRIA 201-210 (1977).

[4] - SAPORTA G. - CAPPE DE BAILLON
DISQUAL manuel d'utilisation (1977).
Note de travail n°16 COREF 11 rue Couderet 92140 CLAMART

[5] - SAPORTA G.
Le traitement des variables qualitatives par codage.
Note de travail n°11 (1976) COREF.

[6] - GAUTIER J.M. - SAPORTA G.
Méthodes non paramétriques en analyse discriminante.
Quelques propositions nouvelles.
Note de travail COREF (1976)

[7] - NAKACHE J.P.
Comparaisons des règles d'affectation associées à différentes méthodes de discrimination.
Groupe de recherches U 88 INSERM
Méthodologie Informatique et statistique en médecine (1976)

[8] - DEHEDIN J.
Discrimination sur variables qualitatives.
thèse 3e cycle statistique Mathématiques (1975)
Université de Paris VI.

[9] - BACCINI A.
Aspect synthétique de la ségmentation et traitement de variables qualitatives à modalités ordonnées.
Thèse Université Paul Sabatier Toulouse (1975).

[10] - LEMERLE LOISEL R.
Méthode d'apprentissage de description structurelles complexes
Un traitement logique de l'image (1981)
Thèse 3e cycle Université Pierre et Marie Curie Paris VI.

[11] - ROMEDER J.M.
Méthodes et programmes d'analyse discriminante.
Dunod.

[12] - MUTEL B. - COSTANTINI M. - PROTH J.M.
Group technology using automatic classification : the ranking problem for new products.
Proposé à APMS Canada (1987).

[13] - BONNEAU F. - PROTH J.M.
Analyse discriminante : Méthode du type plus proches voisins utilisant un prétraitement des données.
Rapport INRIA n°440 Sept. 1985.

[14] - GARCIA H. - MUTEL B. - PROTH J.M.
Classifiction automatique des données techniques de production.
Rapport INRIA n°545 Juillet 1986.

[15] - NADIF A. - COSTANTINI M. - MUTEL B.
Mesures de ressemblance de gammes de fabrication.
Revue APII, vol. 19, n°5 (1985).

[16] - LEBART L. - FENELON J.P.
Statistique et Informatique appliquées - Dunod

[17] - COSTANTINI M.
Apports de la Technologie de groupe dans différents services d'une entreprise.
Etude bibliographique. DEA production automatisée (1984).
Université Bordeaux I.

[18] - FURTH M. - BEGELLA M.
Les caractéristiques essentielles de la notion de groupements analogiques.
USINICA Paris Juin 1985.

[19] - NADIF A. - COSTANTINI M. - MUTEL B.
Reconnaissance de gammes de fabrication.
INRIA Système de production Paris Avril 1985.

[20] - BRUYAND A. - MUTEL B.
Implantation of group technology by data analysis methods
Prolama Paris Juin 1985.

[21] - MINOT J.
Contribution à l'étude de la reconnaissance de familles de pièces en fabrication mécanique
Thèse de Docteur Ingénieur Metz 1983.

[22] - GARCIA H. - MUTEL B. - PROTH J.M.
Familles de produits et îlots de fabrication : le cas de machines multiples.
Rapport INRIA n°469 Décembre 1985.

[23] - CHANDON J.L. - PINSON S.
Analyse typologique - Théorèmes et applications.
MASSON Paris 1980.

[24] - MUTEL B. - COSTANTINI M. - SPADONI M. - MEIER K. - NADIF A.
Classification automatique et technologie de groupe.
Contrat ADEPA - LAEI Juin 1985 rapport n°2.

[25] - MUTEL B.
Classification automatique des données de production.
LAEI Université de Metz.

[26] - MINOT I. - MUTEL B. - LEMOINE Y.
Implantation assistée par ordinateur de la technologie de groupe.
Congrès AFCET Automatique, Besançon, Novembre 1983.

[27] - MUTEL B. - GARCIA H. - PROTH J.M.
Automatic classification of production data.
18th CIRP Stuttgart Juin 1986.

[28] - MUTEL B.
Reconnaissance de groupements technologiques par des méthodes d'analyse de données.
Colloque Productique et Robotique - Bordeaux Mars 1984.

[29] - NADIF A. - BRUYAND A. - MUTEL B.
Technologie de groupe : Problème d'implantation et d'utilisation
Matériaux Mécanique Electricité n°410 Janvier-Février 1985.

[30] - BRUYAND A. - MUTEL B.
Contribution of data analysis methods to computer aided group technology implantation.
Prolamat Paris Juin 1983.

[31] - MUTEL B. - COSTANTINI M. - PROTH J.M.
A general algorithm for the choice of decision rules.
7e Conférence cybernétique et système Londres Septembre 1987.

[32] - HOUTZEL A.
The many faces of group technology.
American Machinist Janvier 1973.

[33] - LEMOINE Y. - MUTEL B.
Automatic recognition of production cells and part family.
Prolamat Paris 1983.

[34] - MUTEL B.
Recherche automatique de groupements technologiques par mesures de ressemblances entre produits.
Division production automatisée du Gami Janvier 1987 Paris.
Les outils de la productique 2e congrès Tome 1.

[35] - PEKLENIK J. - GRUM J.
Investigation of the computer aided classification of Parts.
Annale du CIRP - Volume 29/1/1980.

[36] - AIZERMAN - BRAVERMAN - ROZONER.
The method of potential function.
Automation and Remote Control 26/Nr 11/1965

[37] - ACAPS.
9th North American Manufacturing research conference.
Published by the society of Manufacturing Enginneers May 81.

[38] - MARCOTORCHINO F.
Block seriation problem : A unified approach.
Applied stochastic Models and Data Analysis Vol.3 1987.

[39] - Mc AULEY J.
Machine Grouping for efficient production.
The Production Engineers Février 1972.

[40] - KING J.R.
Machine-component group formation in Group-Technology
OMEGA - The Int Jl of Mgmt. Sci. Vol.8 n°2 (1980)

[41] - KING J.R.
Machine-component grouping in production flew analysis an approach using a rank oder clustering algorithm.
The Int. J. Prod. Mes. 1980 Vol.18 n°2.

[42] - MC QUEEN J.
Some Methods-for classification and analysis of multivariate observations.
Proceeding 5th Berkeley Symposium, 1, 281-297, 1967.

[43] - DIDAY E.
Une nouvelle méthode en classification automatique et reconnaissance de formes.
Note scientifique n°6 Supplément du Bulletin de IRIA N°12 Mai-Juin 1972.

[44] - FORGY E.W.
Cluster analysis of multivariate data : efficiency versus interpretability of classification.
Biometrics, 21, 768-780, 1965.

[45] - GREEN P.E. - TULL D.S.
Research for Marketing Decisions.
Prentice Hall - Englewood Cliffs, N.J, 1975.

[46] - OPITZ H.
A classification system to describe workpieces.
Pergamon Press 1970.

[47] - ADEPA.
Code Multi-M. Multiclass 1982.
ADEPA - 13- 17 Rue Perier B.P. N°54 92123 Montronge

[48] - NADIF A.
Contribution à la classification automatique de données de production
Thèse de doctorat Université de Metz 1987.

[49] - ADEPA.
Dossier : Organisation et gestion de la production
ADEPA Actualités n°16 Décembre 1983

[50] - FRANCOIS A.R.
Manuel d'organisation
Edition Les éditions d'organisation

ANNEXE 1

ANNEXE 1

I Différentes méthodes de discrimination

11) Méthode Bayésienne [1] - [7].

La méthode de discrimination de Bayes présume que certaines valeurs doivent être connues ou données a priori :

- les probabilités Π_k a priori d'appartenance d'un individu à un groupe ;

- les densités f_k de probabilités d'appartenance des individus x pour chacun des r groupes ;

- les différents coûts C_{ij} associés à chaque éventualité E_{ij} de mauvais classement, c'est-à-dire d'affecter un individu à un groupe j alors qu'il appartient à un groupe i ;

- les différentes matrices de variances-covariances des groupes.

Afin de définir une règle de décision, il faut définir une quantité :

- le coût moyen a posteriori d'appartenance d'un individu x au groupe i avec une probabilité a posteriori d'appartenir au groupe i : P(x,i).

$$P(x,i) = \frac{\Pi i \; fi(x)}{\sum_{i}^{r} \Pi i \; fi(x)}$$

$$\text{Coût moyen : } \gamma_m(i) = \frac{\sum_{i} (C_{ij} \; \Pi j \; f_j(x))}{\sum_{j} \Pi_j \; f_j(x)} = \frac{\gamma^i(x)}{\sum_{j} \Pi_j \; f_j(x)}$$

Il apparaît donc normal d'affecter un individu x au groupe i pour $\gamma_i(x)$ minimum.

La règle de décision peut alors s'énoncer de cette manière :

On affecte un individu x_α au groupe i_α pour :

$$\gamma_{i\alpha}(x_\alpha) = \min (\gamma_i(x_\alpha)) \text{ avec } i = 1.....r$$

Remarques : le problème d'évaluation des coûts est un problème difficile à résoudre voir impossible.

La matrice carrée des différents coûts est une matrice à diagonale nulle et non symétrique, c'est-à-dire que le coût d'un bon classement est égal à zéro, et que les quantités C_{ij} et C_{ji} sont différentes.

Toutefois, la règle de décision bayésienne peut se rendre équivalente à la règle de décision de l'analyse discriminante linéaire sous certaines hypothèses :

- Egalité des différents coûts de mauvais classement,
- Egalité des matrices de variance-covariance des groupes,
- Egalité des probabilités a priori d'appartenance à un groupe.

12) Méthode de Sebestyen [11]

Cette méthode s'intègre dans le cadre du problème de la discrimination à but décisionnel. Il faudra rechercher, pour chacune des classes d'individus, une fonction mesurant la similitude d'un individu quelconque avec la classe en question. La règle de décision, qui permet d'affecter un individu non classé, consiste à trouver pour quel groupe, sa similitude est la plus grande.

L'idée de Sebestyen est d'accorder, dans le calcul des distances aux différents groupes, un poids d'autant plus faible que le groupe est fortement dispersé. Ceci mène à effectuer une transformation des variables et cela suppose donc de trouver une métrique. Deux méthodes sont proposées suivant que l'on a affaire à une métrique définie par une matrice diagonale ou bien à une matrice symétrique quelconque.

Le problème se résume de la façon suivante :

Déterminer la matrice Q de telle sorte que :

$$D^2_k = \frac{1}{N_k(N_k-1)} \sum_{x_m^k} \sum_{x_n^k} d^2(x_m^k, x_n^k) \quad \text{soit minimum}$$

avec $d^2(x_m^k, x_n^k) = (x_m^k - x_n^k)' Q (x_m^k, x_n^k)$ et $|Q|=1$

$(x_m^k - x_n^k)'$ désignant le transposé de $(x_m^k - x_n^k)$ noté comme vecteur ligne.

Sebestyen détermine D^2_k comme étant la distance moyenne entre individus du groupe y_k de cardinal N_k, cette quantité étant synonyme d'une mesure de l'agrégation de groupe. Pour indiquer la similitude d'un individu quelconque "a" avec un groupe, Sebestyen fixe une quantité $S(a,y_k)$ comme étant la moyenne des carrés des distances de "a" à chacun des individus du groupe.

$$S(a,y_k) = \frac{1}{N_k} \sum_{x_n^k \in y_k} d^2(a,x_n^k)$$

Règle de décision :

La règle de décision peut s'interpréter de cette façon :
On affecte le nouvel individu $a_\varnothing$ au groupe $y_\varnothing$ si :

$$S(a_\varnothing,y_\varnothing) = \min \{S(a_\varnothing,y_k) \ / \ k=1...k\}$$

k étant le nombre de groupe.

Résumé des 2 méthodes :

a) 1ère méthode : Pondération simple des variables

La métrique est définie par une matrice diagonale. On peut montrer que :

$$D^2_k = \frac{2\,N_k\,P}{N_k-1} \prod_{p=1}^{P} \sigma_p^2 \quad {}^{1/P}$$

avec p=1...P P=nombre de variables

$$\sigma_p^2 = \frac{1}{N_k} \sum_{n=1}^{N_k} (x_{np}-x_p)^2$$

$$S(a,y) = (\prod_{p=1}^{P} \sigma_p^2) \sum_{p=1}^{P} \frac{a_p - x_p}{\sigma_p} + P$$

b) 2ème méthode : Rotation et Pondération des variables

La métrique est définie par une matrice symétrique quelconque :

On démontre aussi que :

$$D^2_k = \frac{2N_k\ P}{N_k - 1}\ |\Sigma|^{1/P}$$

avec $|\Sigma|$ = déterminant de la matrice de covariance défini par :

$$\sigma_{pq} = \frac{1}{N_k} \sum_{n=1}^{N_k} (x_{np} - x_p)\ (x_{nq} - x_q) \quad \text{avec } p,q = 1 \ldots P$$

$$S(a,y) = |\Sigma|^{1/P}\ [(a-x)'\ \Sigma^{-1}\ (a-x) + P]$$

On remarque que la deuxième méthode donne les mêmes résultats que la première si Σ est une matrice diagonale (toutes les covariances des variables prises 2 à 2 sont nulles). La première méthode a l'avantage de ne nécessiter que des calculs simples.

13) Discrimination par voisinage

a) Méthode

L'idée est de calculer dans un premier temps les k voisins les plus proches de chacun des individus où k étant un entier donné a priori. A partir de ceci, on peut construire un tableau de contingence individu-classe dans lequel chaque case (i,k) indique le nombre d'individus voisins de i appartenant à la classe k. On peut tirer de ceci un tableau carré T(i,j) dont chaque élément (i,j) donne le nombre de voisins des individus de la classe i appartenant à la classe j.

On peut ainsi définir un indice de séparabilité des classes qui peut nous donner une indication sur la qualité de l'affectation.

b) Règle de décision

Pour classer un nouvel individu x_* se présentant, il faut déterminer ses k voisins les plus proches, on obtient ainsi les fréquences d'appartenance des voisisns de x_* aux différentes classes. La régle de décision peut s'énoncer de cette façon :

On décide d'affecter le nouvel individu x_* à la classe qui a obtenu la fréquence d'appartenance la plus forte.

Remarque : La méthode est valable quelque soit la nature des données, on peut prendre ainsi la mesure de distance ou de proximité la plus adéquate en fonction des données.

14) Méthodes de réduction - Elimination des variables

141) But des méthodes

Ces méthodes ne s'attacheront qu'au caractère descriptif de l'analyse discriminante. Elles permettront de rechercher un nombre plus restreint de variables qui expliqueront au mieux la séparation des classes. Ces variables auront donc un pouvoir discriminant plus important que les autres.

D'autre part, le nombre plus réduit de variables, permettra une mise en oeuvre plus simple des algorithmes, et une exécution beaucoup plus rapide.

On peut rapprocher le principe de ces méthodes aux techniques de transmission de données en présence de bruit, auquel cas, on essaye de filtrer ce bruit parasite.

142) Méthode pas à pas [11]

La technique de pas à pas consiste, sur un ensemble de p variables mesurées sur une population d'individus, à se restreindre successivement à la meilleure, puis aux 2 meilleures,...variables pour le problème considéré, en utilisant au pas q un critère de sélection d'une variable parmi les p-q+1 restantes. On arrête la procédure au moment où l'on estime que le nombre de variables discriminantes retenues devient trop grand.

Remarque : En règle générale, au pas q, on ne remet pas en cause le choix des variables déjà sélectionnées, ce qui ne conduit pas nécessairement au meilleur sous-ensemble de variables.

1421) Critère du pourcentage de bien classé

Ce critère permet de tester la validité d'une méthode de discrimination pour laquelle une procédure de classement a été définie.

La partition étant effectuée, on est en droit de se demander dans quelle mesure les individus des différents groupes se retrouvent bien dans les régions de classement correspondantes.

Pour ceci, on établit un tableau carré dont chaque élément n_{ij} représente le nombre d'individus du groupe j classé en i, et le rapport $n_{jj}/n_{.j}$ (avec $n_{.j} = \Sigma n_{ij}$) nous donne le pourcentage de bien classé. Le pourcentage global de bien classé étant :

$$\sum_i n_{ii} / \sum_j n_{.j}$$

Afin de pronostiquer sur l'appartenance d'un nouvel individu à l'un des groupes, on construit un nouveau tableau n'_{ij} qui fournit les estimations des probabilités a posteriori d'appartenance à chacun des groupes.

Le pourcentage de bien classé est donc utilisé comme critère de pas à pas pour une méthode.

1422) Critère trace($V^{-1}B$)

Ce critère contrairement au précédant ne nécessite pas la définition d'une procédure de classement. On recherche à chaque pas quel est l'ensemble de variables qui maximise l'inertie du nuage Y des individus, calculée avec la métrique V^{-1}, relativement à son centre de gravité. Donc au pas q, on cherche quel est le meilleur sous-ensemble de q variables qui maximise :

$$\sum_{y} \frac{Ny}{N} (y-y)' V^{-1} (y-y) \;/\; y \epsilon Y$$

Ceci revient à maximiser trace $V^{-1}{}_{q}B_{q}$ (B matrice de covariance inter-classe - V matrice de covariance totale).

On ne dispose pas de test d'arrêt naturel, comme dans le cas du critère de pourcentage de bien classé, en effet la quantité $(V^{-1}{}_{q}B_{q})$ pourra croître lors du passage du pas q au pas q+1 sans que la discrimination en soit pour autant améliorée.

143) Analyse factorielle discriminante

Le principe est de rechercher un nombre de variables s inférieur au nombre initial de variables qui soient des combinaisons linéaires de celles-ci et les plus discriminantes au sens de certains critères.

Le choix des variables se faisant de sorte à ce qu'elle maximise la variance inter-classe et minimise la variance intra-

classe, c'est-à-dire celles qui séparent au mieux les classes tout en rendant les classes les plus homogènes.

On peut utiliser ensuite, une règle d'affectation du type géométrique.

1431) Exposé du problème

On veut donc trouver l'axe factoriel ou la forme linéaire u qui discrimine au mieux non pas l'ensemble des individus x mais l'ensemble y des classes d'individus. Il faut donc trouver l'axe passant par le centre de gravité du nuage d'individus engendré par un vecteur u telque, le long de cet axe, la variance inter-classe uBu' soit maximum et la variance intra-classe uWu' soit minimum. W et B étant respectivement les matrices de covariances intra-classe et inter-classe liées à la matrice de covariance totale T par T=W+B (théorème d'Huygens). Le problème revient donc à trouver u qui maximise la relation :

$$\frac{uBu'}{uWu'} \quad \text{qui est identique à maximiser} \quad \frac{uBu'}{uTu'}$$

le premier axe factoriel discriminant u_1 est le vecteur propre de $T^{-1}B$ correspondant à la plus grande valeur propre λ_1.

1432) Règle de décision

On décide d'affecter un nouvel individu "a" à une classe y_φ pour :

$$d^2(a,y_\varphi) = \min \{d^2(a,y) \;/\; y \epsilon Y\}$$

avec $d^2(a,y) = (a-y)'\, T^{-1}\, (a-y)$.

144) Discrimination par des méthodes de regression linéaire [1]

Il s'agit de trouver un codage optimal de la variable qualitative à expliquer y, en lui associant la nouvelle variable quantitative obtenue z qui prend les valeurs $C_1 \ldots C_r$. On affecte à un individu i la valeur $z_i = C_k$ si i appartient à la classe C_k. Cette nouvelle variable doit avoir alors la meilleure régression linéaire en fonction des variables $x^1 \ldots x^q$. Cette régression linéaire fournit alors une fonction linéaire discriminante de type :

$$a(x_i) = \sum_{j=1}^{q} \alpha^j \ (x_j{}^j - x^j)$$

La règle de décision peut s'énoncer de cette façon : Un individu i est affecté à la classe C_k si :

$$| a(x_i) - C_k | \ < \ | a(x_i) - C_j | \quad \forall j \in 1 \ldots r$$

Cette méthode est surtout adaptée à l'analyse discriminante dans le cas de deux classes.

15) Méthodes d'analyse discriminante sur variables qualitatives.

Les principales méthodes s'appliquant à des variables explicatives qualitatives sont basées sur le principe de réduction des variables.

151) Méthodes de réduction des variables

Ces méthodes proposent une réduction pas à pas et recherchent ensuite un codage quantitatif associé aux variables "les plus discriminantes".

On utilise ensuite des méthodes de discrimination classique.

1511) Méthode séquentielles des corrélations canonique (Masson) [1]

Masson propose une méthode pas à pas qui consiste à inclure progressivement des variables. On cherche dans un premier temps, la variable qualitative, parmi les q au total, qui a le plus fort pouvoir discriminant, puis on code cette variable sous forme numérique. A chaque pas p ensuite, on cherche, parmi les q-p variables qualitatives restantes, celles qui, associées aux variables numériques précédantes, donne la meilleure discrimination ; ce qui revient à faire q analyses canoniques. On arrête le processus de sélection des variables lorsque l'on juge la discrimination satisfaisante. On utilise ensuite l'analyse discriminante classique avec les variables retenues.

Remarque : L'affectation d'un nouvel individu pose un problème pour trouver les variables quantitatives en utilisant le codage précédant [1]. Comme le fait remarquer Saporta [2], si le nombre de modalités des variables n'est pas identique, rechercher la variable qualitative qui a la plus forte corrélation canonique n'est pas synonyme de meilleure discrimination. A chaque étape, les variables précédantes ne sont pas remises en cause.

1512) Méthode des coefficients de Tschuprow (Saporta) [3] - [4]

Cette méthode a été développé par Saporta et la programme résultant porte le nom de DISQUAL [3-4] (DIScrimination sur variables QUALitatives).

La méthode permet de réaliser, une sélection progressive des variables explicatives et une analyse discriminante sur les variables retenues par la méthode des "facteurs z" [2].

La sélection pas à pas des variables est basée sur le calcul des coefficients de Tschuprow. Au premier pas, on retient la variable qui a le plus fort coefficient de Tschuprow avec la variable à expliquer. Ensuite, successivement, on sélectionne les variables qui ont le plus grand coefficient de Tschuprow partiel avec la variable à expliquer en tenant compte de la variable précédemment introduite. On arrête le processus lorsque le nombre de variables sélectionnées fixées à l'avance est atteint ou lorsque le coefficient de Tschuprow multiple calculé à chaque pas n'augmente plus suffisament. On réalise ensuite une analyse factorielle discriminante par la méthode des "facteurs z". L'affectation d'un nouvel individu se fait par une procédure classique de distance au centre de gravité des groupes.

1513) Méthode non paramétrique (Gautier-Saporta)

Gautier et Saporta proposent une démarche [6] à suivre en analyse discriminante, s'appuyant sur une méthode non paramétrique, et s'appliquant aussi bien à des variables qualitatives que quantitatives. La méthode repose sur l'utilisation de trois techniques :

- une méthode non paramétrique de sélection de variables, pour chaque classe, appelée méthode des fenêtres, basée sur un critère permettant de rechercher, à chaque pas, la variable qui élimine le plus possible d'individus des autres classes, et le moins possible du groupe en question, lors de l'intersection de la fenêtre avec les précédantes ;

- une analyse factorielle par groupe par l'utilisation d'une métrique associée à la nature des variables, qui peut être mixte dans le cas de variables explicatives qualitatives et quantitatives ;

- la méthode de noyaux produits gaussiens pour l'estimation de densité.

Une estimation de la probabilité d'appartenance d'un individu à un groupe, en utilisant les densités estimées, fourni par la formule de Bayes, permet de donner une idée de l'affectation de nouveaux individus, mais l'affectation sera d'autant meilleure que les régions de l'espace seront nombreuses donc denses [fig. A_1].

1514) Méthode de discrimination par segmentation

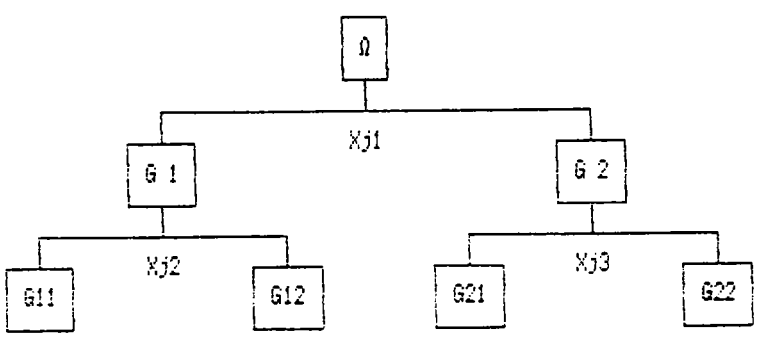
Dehedin a réalisé une étude [8] sur diverses méthodes de discrimination utilisables sur des variables qualitatives. Il en a retenu trois types :

- Il montre comment par codage des variables qualitatives, on résoud le problème de discrimination sur variables qualitatives par une analyse discriminante à l'aide de l'analyse factorielle.

- Il présente ensuite des méthodes probabilistes qui cherchent à reconstruire dans chaque groupe la distribution de probabilité.

- Puis il a énoncé les méthodes de segmentation telles que la méthode ELISEE et d'autres [9] qui cherchent à reconstruire les modalités de la variable à expliquer par croisement de certaines variables explicatives.

La méthode de segmentation vise à reconstruire une hiérarchie de groupes de la population de départ Ω de plus en plus homogène en fonction de la variable à expliquer Y. Elle permet de décrire aussi la liaison qui existe entre les modalités de la variable Y et les différentes variables x_j explicatives. La méthode procède pas à pas, par dichotomies successives issues à chaque pas d'une des variables x_j la plus discrétisante.



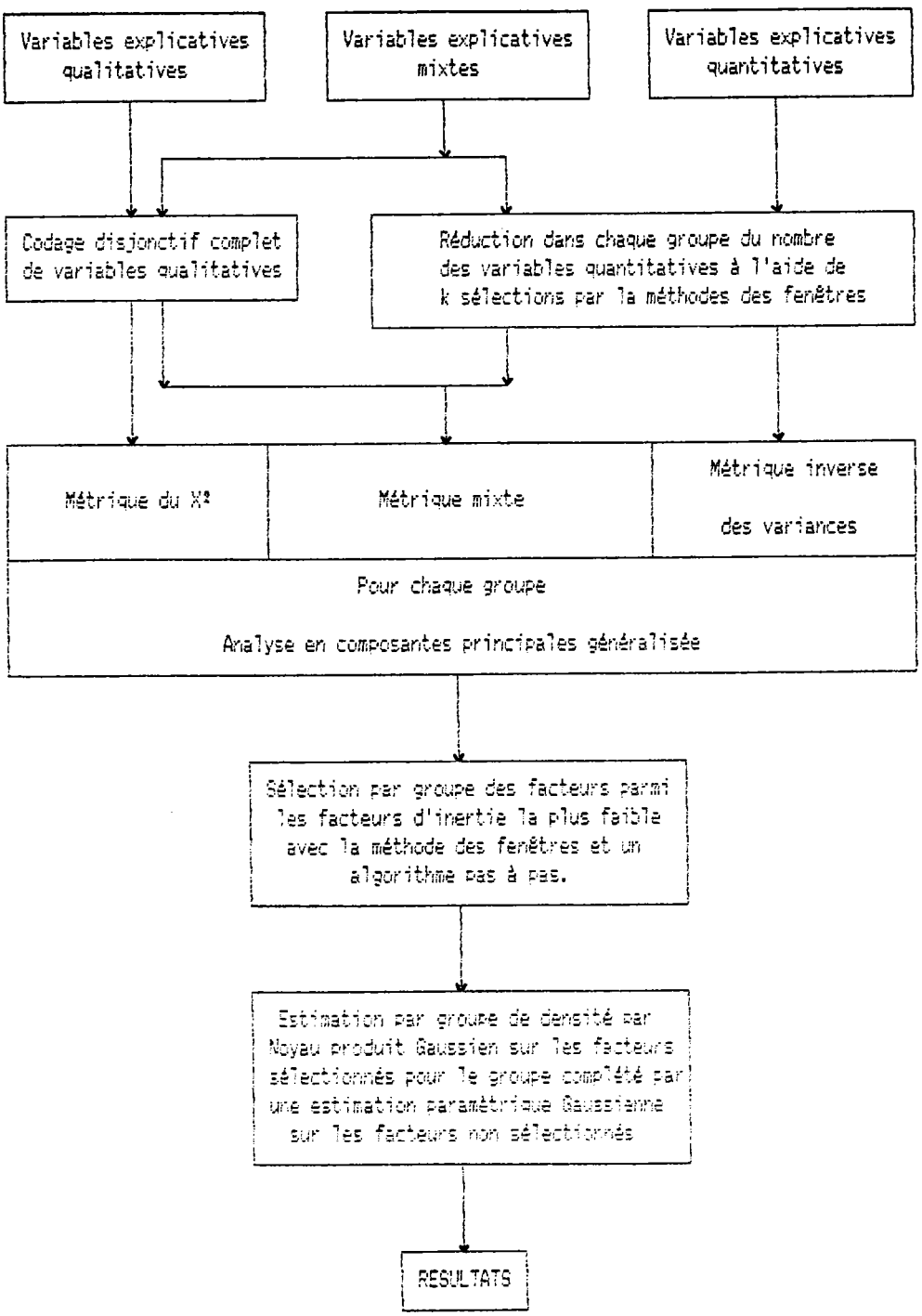


figure A1 : Méthode non paramétrique d'analyse discriminante à k groupes.

On choisit donc à chaque pas la variable, parmi toutes les autres, qui offre la dichotomie la plus homogène compte tenu de la variable à expliquer Y. La segmentation peut être arrêtée par différents tests :

- test du nombre de niveaux : on se limite à un certain nombre de variables descriptives ce qui revient à arrêter la segmentation au niveau pour lequel le nombre de variables est atteint.

- test d'un effectif minimum du segment :

La segmentation n'est poursuivie sur un segment que dans le cas où l'effectif est supérieur à un seuil fixé.

Un des inconvénients de cette méthode est qu'elle nécessite une population importante afin qu'elle ait une bonne efficacité.

1515) Méthode d'apprentissage de descriptions structurelles complexes.

Mme Lemerle Loisel a développé une méthode [10] dont l'objectif consiste à structurer un ensemble d'apprentissage (population d'individus) afin de hiérarchiser les exemples (individus) et de dire si un nouvel individu peut ou non appartenir à une famille, par la notion de nuance critique. La méthode permet d'extraire les différences les plus communes entre individus avant d'utiliser celles qui séparent effectivement les familles. Le choix du langage de description est indépendant des algorithmes proposés, il suffit de pourvoir caractériser certains objets constituant un individu et de décrire à l'aide de n variables ces objets.