

50376  
1991  
284

N° d'ordre : 821



68255

50376  
1991  
284

# THESE

*présentée à*

L'UNIVERSITÉ

DE LILLE

SCIENTIFIQUE

*pour l'obtention du titre de*

# DOCTEUR

*en Productique : Automatique et Informatique Industrielle*

*par*

**Claudine BOTTE-LECOCQ**

L'ANALYSE DE DONNEES MULTIDIMENSIONNELLES  
PAR  
TRANSFORMATIONS MORPHOLOGIQUES BINAIRES.

*Soutenue le 17 décembre 1991 devant la commission d'examen :*

MM.

<b>P. VIDAL</b>	<i>Président</i>	<i>Professeur à l'U.S.T.L.F.A.</i>
<b>A. GORDON</b>	<i>Rapporteur</i>	<i>Professeur à St Andrews (Ecosse).</i>
<b>G. GOVAERT</b>	<i>Rapporteur</i>	<i>Professeur à l'U.T.C. Compiègne.</i>
<b>J.G. POSTAIRE</b>	<i>Directeur de thèse</i>	<i>Professeur à l'U.S.T.L.F.A.</i>
<b>S. BEUCHER</b>	<i>Examineur</i>	<i>Professeur à l'Ecole des Mines de PARIS.</i>
<b>R. LAURENT</b>	<i>Examineur</i>	<i>Professeur à ORLEANS.</i>
<b>L. POVY</b>	<i>Examineur</i>	<i>Professeur à l'U.S.T.L.F.A.</i>

## AVANT-PROPOS.

Le travail présenté dans ce mémoire a été effectué au Centre d'Automatique de l'Université des Sciences et Techniques de Lille Flandre Artois, dirigé par Monsieur le Professeur Pierre Vidal. Je le remercie de l'accueil qu'il m'a réservé au sein de son laboratoire et de l'honneur qu'il me fait en acceptant la présidence du jury de Thèse.

J'adresse mes remerciements les plus sincères à Monsieur Jack-Gérard Postaire, Professeur à l'Université des Sciences et Techniques de Lille Flandres Artois pour son parfait encadrement tout au long de mes travaux. Qu'il trouve ici l'expression de ma gratitude pour ses précieux conseils et son dynamisme.

Que Monsieur A. D. Gordon, Professeur à l'Université de St Andrews, trouve ici l'expression de ma sincère reconnaissance pour avoir bien voulu être rapporteur de cette thèse.

Je remercie également sincèrement Monsieur G. Govaert, Professeur à l'Université de Technologie de Compiègne, pour avoir accepté d'être rapporteur de cette thèse.

Je tiens à exprimer ma reconnaissance à Monsieur Beucher, Professeur à l'Ecole des Mines de Paris, pour m'avoir fait l'honneur de juger mon travail.

Que Monsieur Laurent, Professeur à l'Université d'Orléans, trouve ici l'expression de toute ma gratitude pour sa participation au Jury de Thèse.

J'exprime ma sincère reconnaissance à Monsieur L. Povy, Professeur à l'Ecole Universitaire D'Ingénieurs de Lille, pour avoir accepté de juger ce travail.

Je tiens également à remercier très sincèrement Monsieur R. D. Zhang pour ses nombreux et très précieux conseils, ainsi que Monsieur B. Ceurstemont pour son aide logistique.

Enfin, mes remerciements vont à tous ceux qui, de loin ou de près, ont permis la réalisation de ce mémoire de thèse, et notamment les membres de l'équipe Image et Décision du Centre d'Automatique. Je remercie tout particulièrement mon mari pour l'aide qu'il m'a apportée, aussi bien pour la réalisation pratique de ce mémoire que pour son soutien moral tout au long de mes recherches. Que mon fils trouve en ce mémoire les raisons de mes nombreuses absences.

# SOMMAIRE

## - PREMIERE PARTIE -

### CLASSIFICATION AUTOMATIQUE ET MORPHOLOGIE MATHEMATIQUE

#### *I - LA CLASSIFICATION AUTOMATIQUE.*

<b>I.1. INTRODUCTION.....</b>	<b>1</b>
<b>I.2. PROCEDURES GLOBALES.</b>	
I.2.1. Approche statistique paramétrique.....	2
I.2.2. Approche métrique.	
I.2.2.a. Optimisation d'un critère.....	3
I.2.2.b. Classification hiérarchique.....	4
I.2.3. Réduction de l'espace de représentation.....	4
<b>I.3. PROCEDURES LOCALES.</b>	
I.3.1. Détection des modes par recherche des maxima locaux.....	6
I.3.2. Analyse de la convexité.....	7
I.3.3. Extraction des contours des modes.....	7
<b>I.4. CONCLUSION.....</b>	<b>8</b>

***II - ADAPTATION DES TRANSFORMATIONS MORPHOLOGIQUES A  
L'ANALYSE DE DONNEES MULTIDIMENSIONNELLES.***

<b>II.1. INTRODUCTION.....</b>	<b>11</b>
<b>II.2. TRANSFORMATION DES DONNEES INITIALES EN UN ENSEMBLE DISCRET D'ELEMENTS A VALEURS BINAIRES.....</b>	<b>15</b>
<b>II.3. TRANSFORMATIONS MORPHOLOGIQUES BINAIRES ELEMENTAIRES.</b>	
II.3.1. La dilatation.....	22
II.3.2. L'érosion.....	23
II.3.3. L'ouverture et la fermeture.....	27
<b>II.4. TRANSFORMATIONS MORPHOLOGIQUES A ELEMENTS STRUCTURANTS MULTIPLES.</b>	
II.4.1. Notion de transformation en tout ou rien.....	30
II.4.2. Notion de transformation morphologique à éléments structurants multiples.....	32
II.4.3. Exemple d'application.....	34
<b>II.5. CONCLUSION.....</b>	<b>37</b>

**- DEUXIEME PARTIE -**

**DETECTION DES MODES PAR FILTRAGE  
MORPHOLOGIQUE.**

*III - EXTRACTION DES MODES PAR TRANSFORMATIONS  
MORPHOLOGIQUES ELEMENTAIRES.*

<b>III.1. INTRODUCTION.....</b>	<b>38</b>
<b>III.2. EXTRACTION DES DOMAINES MODAUX PAR TRANSFORMATIONS MORPHOLOGIQUES A ELEMENT STRUCTURANT UNIQUE.</b>	
III.2.1. Principe de détection des domaines modaux.....	39
III.2.2. Algorithme associé.....	41
III.2.3. Identification des modes détectés.....	42
III.2.4. Exemple d'application.....	43
III.2.5. Conclusion.....	48
<b>III.3. EXTRACTION DES FRONTIERES MODALES PAR TRANSFORMATIONS MORPHOLOGIQUES A ELEMENT STRUCTURANT UNIQUE.</b>	
III.3.1. Introduction.....	48
III.3.2. Principe de détermination des frontières modales.....	48
III.3.3. Algorithme associé.....	50
III.3.4. Amélioration de la méthode de détermination des frontières modales.....	51
III.3.5. Exemple d'application.....	53
<b>III.4. CONCLUSION.....</b>	<b>56</b>

**IV - EXTRACTION DES MODES PAR ANALYSE MORPHOLOGIQUE  
DE LA CONNEXITE.**

<b>IV.1. INTRODUCTION.....</b>	<b>57</b>
<b>IV.2. DETERMINATION DES DOMAINES MODAUX PAR ANALYSE MORPHOLOGIQUE DE LA CONNEXITE DANS LES VALLEES.</b>	
IV.2.1. Principe général de la méthode.....	58
IV.2.2. Définition de la famille structurante <i>CV</i> .....	59
IV.2.3. Transformation morphologique associée à la famille structurante <i>CV</i> .....	61
IV.2.4. Exemple d'application.....	63
IV.2.5. Conclusion.....	65
<b>IV.3. DETERMINATION DES DOMAINES MODAUX PAR ANALYSE MORPHOLOGIQUE DE LA CONNEXITE DANS LES REGIONS MODALES.</b>	
IV.3.1. Principe général de la méthode.....	66
IV.3.2. Définition de la famille structurante <i>CM</i> .....	67
IV.3.3. Définition de la transformation morphologique associée.....	68
IV.3.4. Exemple d'application.....	69
IV.3.5. Conclusion.....	71
<b>IV.4. COMPLEXITE DES ALGORITHMES D'EXTRACTION DES MODES.....</b>	<b>72</b>
<b>IV.5. CONCLUSION.....</b>	<b>73</b>

**- TROISIEME PARTIE -**

**APPLICATION A L'ANALYSE DES DONNEES.**

*V - CLASSIFICATION DES DONNEES A PARTIR DES MODES  
MORPHOLOGIQUES.*

<b>V.1. INTRODUCTION.....</b>	<b>75</b>
<b>V.2. CLASSIFICATION OPTIMALE DE BAYES.....</b>	<b>77</b>
<b>V.3. PRESENTATION DES DIFFERENTES TECHNIQUES DE CLASSIFICATION ETUDIEES.</b>	
V.3.1. Assignation des observations au domaine modal le plus proche.....	79
V.3.2. Assignation des observations au centre de gravité du domaine modal le plus proche.....	80
V.3.3. Assignation des observations au noyau le plus proche.....	82
V.3.4. Assignation ordonnée des observations au noyau le plus proche.....	85
<b>V.4. EXEMPLES D'APPLICATION.</b>	
V.4.1. Exemple 1.....	85
V.4.2. Exemple 2.....	91
V.4.3. Conclusion.....	96
<b>V.5. COMPARAISON DES DIFFERENTES METHODES DE CLASSIFICATION.</b>	
V.5.1. Variations du taux d'erreur en fonction du pas de discrétisation.....	97
V.5.2. Comparaison en temps d'exécution.....	99
V.5.3. Comparaison par rapport à la méthode de détection.....	100
V.5.4. Récapitulatif des résultats obtenus pour les exemples 1 et 2.....	102
<b>V.6. CONCLUSION.....</b>	<b>105</b>

## *VI - RESULTATS EXPERIMENTAUX.*

<b>VI.1. INTRODUCTION.....</b>	<b>106</b>
<b>VI.2. PRESENTATION GENERALE DE DEUX METHODES CLASSIQUES EN ANALYSE DE DONNEES.</b>	
VI.2.1. L'algorithme Isodata.....	106
VI.2.2. L'algorithme des K-means.....	109
<b>VI.3. EXEMPLE BIDIMENSIONNEL COMPOSE DE TROIS CLASSES NORMALES NON SPHERIQUES.....</b>	<b>110</b>
<b>VI.4. EXEMPLE BIDIMENSIONNEL CONSTITUE DE TROIS CLASSES NON SPHERIQUES ET NON NORMALES.....</b>	<b>129</b>
<b>VI.5. EXEMPLE MULTIDIMENSIONNEL.....</b>	<b>136</b>
<b>VI.6. CONCLUSION.....</b>	<b>145</b>

## *VII - CONCLUSION GENERALE.*

## NOTATIONS UTILISEES

- $N$  : Dimension de l'espace de représentation des données
- $Q$  : Taille de l'échantillon
- $K$  : Nombre de classes de l'échantillon
- $C_k$  : Classe  $k$ ,  $k = 1, 2, \dots, K$
- $D_k$  : Domaine modal associé à la classe  $C_k$
- $d_k$  : Fonction de décision associée à la classe  $C_k$
- $E^N$  : Espace euclidien de dimension  $N$
- $Z^+$  : Ensemble des entiers positifs ou nuls
- $(Z^+)^N$  : Espace discrétisé de dimension  $N$
- $Y^*$  : Echantillon d'observations
- $Y_q$  : Observation définie dans l'espace euclidien
- $Y'_q$  : Observation définie dans l'espace euclidien normalisée
- $Y_q^n$  :  $n$ -ième coordonnée de l'observation  $Y_q$
- $H_q$  : Hypercube contenant l'observation  $Y'_q$
- $R$  : Pas de discrétisation de l'espace de représentation des données
- $B$  : Fonction binaire définie sur  $(Z^+)^N$
- $\underline{X}$  : Ensemble discret à valeur binaire 1
- $\underline{X}^C$  : Ensemble discret à valeur binaire 0
- $\underline{X}^-$  : Ensemble symétrique de  $\underline{X}$  par rapport à l'origine
- $X$  : Point du réseau d'échantillonnage, à valeur binaire 1 ou 0
- $x_n$  :  $n$ -ième coordonnée du point  $X$
- $\underline{S}$  : Élément structurant
- $S$  : Point de l'élément structurant  $\underline{S}$
- $-S$  : Point du symétrique de  $\underline{S}$  par rapport à l'origine
- $G$  : Point Origine de l'élément structurant
- $\underline{S}^1$  : Ensemble des points de  $\underline{S}$  à valeur binaire 1
- $\underline{S}^0$  : Ensemble des points de  $\underline{S}$  à valeur binaire 0

- $M$**  : Ensemble des points des modes morphologiques détectés
- $\underline{X} \odot \underline{S}$**  : Dilatation de  $\underline{X}$  par  $\underline{S}$
- $\underline{X} \ominus \underline{S}$**  : Erosion de  $\underline{X}$  par  $\underline{S}$
- $\underline{X} \circ \underline{S}$**  : Amincissement de  $\underline{X}$  par  $\underline{S}$
- $\underline{X} \otimes \underline{S}$**  : Transformation en tout ou rien de  $\underline{X}$  par  $\underline{S}$
- $(\underline{X})_S$**  : Résultat de la translation de  $\underline{X}$  par  $S$ , élément de  $\underline{S}$
- $(\underline{X})_S$**  : Ouverture de  $\underline{X}$  par  $\underline{S}$
- $(\underline{X})^S$**  : Fermeture de  $\underline{X}$  par  $\underline{S}$
- $C_S(X)$**  : Configuration locale au point  $X$  dans le domaine délimité par  $\underline{S}$
- $S$**  : Famille structurante, ensemble d'éléments structurants
- $\underline{X} \circ S$**  : Amincissement de  $\underline{X}$  par la famille  $S$
- $\underline{X} \odot S$**  : Epaissement de  $\underline{X}$  par la famille  $S$
- $\underline{X} \otimes S$**  : Transformation en tout ou rien de  $\underline{X}$  par la famille  $S$
- $/$**  : Différence symétrique de deux ensembles :
- $$A/B = (A \cup B) - (A \cap B)$$
- $\cup$**  : Union de deux ensembles
- $\cap$**  : Intersection de deux ensembles
- $\subset$**  : Inclusion de deux ensembles au sens strict
- $\subseteq$**  : Inclusion de deux ensembles au sens large

## ABREVIATIONS UTILISEES

- FMEU** : Méthode de détermination des **Frontières Modales** par transformation morphologique à **Elément structurant Unique**.
- DMEU** : Méthode de détermination des **Domaines Modaux** par transformation morphologique à **Elément structurant Unique**.
- CVEM** : Méthode de détermination des domaines modaux par analyse de la **Connexité dans les Vallées** par transformation morphologique à **Eléments structurants Multiples**.
- CMEM** : Méthode de détermination des domaines modaux par analyse de la **Connexité dans les régions Modales** par transformation morphologique à **Eléments structurants Multiples**.
- ADMP** : Technique de Classification par **Assignment des observations au Domaine Modal le plus Proche**.
- ACDP** : Technique de Classification par **Assignment des observations au Centre de gravité du Domaine modal le Plus proche**.
- ASNP** : Technique de Classification par **Assignment Simple des observations au Noyau le plus Proche**.
- AONP** : Technique de Classification par **Assignment Ordonnée des observations au Noyau le plus Proche**.

**- PREMIERE PARTIE -**

**CLASSIFICATION AUTOMATIQUE  
ET  
MORPHOLOGIE MATHEMATIQUE**

## **I - LA CLASSIFICATION AUTOMATIQUE.**

## LA CLASSIFICATION AUTOMATIQUE.

### I.1. INTRODUCTION.

La classification est une démarche naturelle car, de tout temps, l'homme a cherché à découvrir un ordre sous-jacent à la multitude des objets constituant son environnement. En fait, dans de nombreux domaines de l'activité humaine, cette démarche devient une méthodologie dont les premiers qui en prirent conscience furent les naturalistes du 18<sup>e</sup> siècle. Aujourd'hui, dans de nombreuses disciplines telles que la biologie, la botanique, la géologie, la zoologie, l'économie, ..., le spécialiste se trouve confronté à de grands nombres d'objets comparables, qui peuvent être décrits par des paramètres observables en nombre fini. Un élément important dans la compréhension de ces données est la définition de groupements, ou classes, au sein de l'ensemble de ces objets, ce qui est le but de la **classification automatique**.

Il s'agit en effet de classer, de manière automatique, un ensemble fini d'objets de telle sorte que les objets appartenant à une même classe soient plus semblables que ceux appartenant à des classes différentes. Cette démarche est relativement difficile à formaliser, surtout quand on se place dans un **contexte non supervisé**, c'est-à-dire quand on ne dispose d'aucune information a priori sur la structure de l'ensemble des objets à classer.

Dans les années 60, les techniques de classification automatique ont connu un essor important dû à l'utilisation des calculateurs numériques. En effet, ceux-ci ont non seulement permis d'augmenter les volumes de données pouvant être traités, mais ont

également fait évoluer la philosophie de classification qui s'est adaptée à ces nouveaux outils.

Les procédures de classification automatique non supervisées peuvent être schématiquement regroupées en deux grandes catégories selon que toutes les observations disponibles sont prises en compte simultanément pour découvrir l'existence de classes ou, qu'au contraire, on s'attache à ne considérer que les relations entre les observations et leurs voisines, pour découvrir la structure de leur distribution.

Les procédures de la première catégorie, que nous qualifierons de procédures globales par opposition à celles de la deuxième catégorie, que nous qualifierons de locales, sont certainement les plus populaires.

## **I.2. PROCEDURES GLOBALES.**

### **I.2.1. Approche statistique paramétrique.**

Dans ce type d'approche, les modèles de distribution des observations sont supposés connus a priori. Sous cette hypothèse paramétrique, le problème de l'analyse de données peut être ramené à celui de la détermination des paramètres d'un mélange de fonction de densité représentant les distributions des observations provenant de chacune des classes en présence dans l'échantillon analysé.

Le problème de l'estimation des paramètres d'un mélange de fonctions de densité à partir d'un échantillon d'observations représentatif de ce mélange a simultanément été abordé par des techniques d'apprentissage Bayessien et par des procédures d'estimation par maximum de vraisemblance.

Dans la première catégorie, on attribue généralement à Daly [DAL - 62] la formulation Bayessienne de l'apprentissage des paramètres d'un mélange, mais Hillborn et Lainiotis donnent une formulation beaucoup plus complète en 1968 [HIL - 68].

Sensiblement à la même époque, Hasselbald [HAS - 66] puis Day [DAY - 69] utilisent les techniques d'estimation par maximum de vraisemblance pour aboutir à des résultats très semblables. Des techniques basées sur la détermination des moments des distributions ont également été proposées par D. Cooper et P. Cooper [COO - 64] [COO - 67].

Cependant, toutes ces techniques statistiques d'apprentissage non supervisé nécessitent, outre la possibilité d'utiliser un modèle paramétrique pour décrire les fonctions de densité sous-jacentes, des hypothèses souvent restrictives. Ainsi la connaissance du nombre de classes en présence est souvent exigée [SCH - 76], celui-ci pouvant même être limité à deux dans certains cas [MAK - 77] [MIZ - 75]. D'autres hypothèses restrictives telles que l'égalité des matrices de covariance ou la connaissance des probabilités a priori des différentes classes sont parfois exigées [WOL - 70].

### **I.2.2. Approche métrique.**

Pour éviter d'avoir recours à des modèles statistiques paramétriques qui peuvent conduire à imposer une structure aux données plutôt qu'à découvrir leur organisation véritable, tout un courant de l'analyse de données fait appel à des notions métriques de similarité plutôt qu'à des notions de fonctions de densité paramétrables [SOK - 63] [BAL - 65].

#### **I.2.2.a. Optimisation d'un critère.**

De nombreux critères globaux indiquant la cohésion des classes ont été proposés [FRI - 67] [JON - 68] [FUK - 70]. La recherche des extrema de ces critères conduisent en général à des classifications qui maximisent la dispersion inter-classe tout en minimisant la dispersion intra-classe [BAL - 67] [MAC - 67] [JAI - 88]. Diday a proposé la méthode des nuées dynamiques qui est basée sur la notion de noyaux. Ceux-ci sont initialement décrits par des représentants tirés au hasard dans la population

disponible [DID - 71]. La partition obtenue dépendant du choix initial de ces noyaux, la notion de formes fortes a été introduite pour améliorer les performances de la procédure [DID - 79].

### **I.2.2.b. Classification hiérarchique.**

Les partitions de l'ensemble des objets à classer peuvent être représentées par des structures arborescentes en adoptant une démarche hiérarchique ascendante ou descendante [JAM - 78]. La démarche ascendante consiste à considérer initialement chaque objet comme une classe, et à regrouper, à chaque étape, les classes deux à deux en maximisant un critère de similarité. La démarche descendante consiste à grouper initialement tous les objets à classer dans une seule classe et, à chaque étape, à maximiser un critère de dissimilarité permettant de diviser chacune d'elles en deux. Dans les deux cas, une hiérarchie des classes est ainsi réalisée et la partition est obtenue en respectant soit un nombre de classes préfixé, soit un critère prédéfini [BAY - 80] [LAN - 67] [LUK - 79]. Une autre méthode consiste à réduire la hiérarchie des parties ou l'arbre des classifications sous forme d'un ensemble de noeuds significatifs, avant d'effectuer la partition de l'ensemble des objets à classer [GOR - 87] [LER - 91].

### **I.2.3. Réduction de l'espace de représentation.**

De nombreuses techniques d'analyse de données multidimensionnelles font appel aux capacités de classification de l'opérateur humain dans le plan. Il s'agit de trouver une technique qui permette de représenter les données multidimensionnelles par des points dans un plan, de telle sorte que les observations appartenant à la même classe dans l'espace d'origine restent regroupées dans le plan, et que les observations provenant de classes différentes soient nettement séparées dans ce plan.

La technique la plus couramment utilisée pour réduire ainsi la dimension des données est certainement l'analyse en composantes principales [COO - 71] [SEB - 84] qui est basée sur la transformation de Karhunen-Loeve [AHM - 75].

Lorsque cette technique ne permet pas de réduire suffisamment la dimension, du fait que trop d'informations seraient perdues si on ne conservait que deux axes principaux, certains auteurs proposent d'utiliser cette technique comme point de départ pour une technique plus élaborée [GOW - 66] [KRU - 77], appelée, en anglais, "multidimensional scaling". Il s'agit de minimiser des critères qui indiquent comment la répartition des points image dans le plan est le reflet de la répartition des observations dans l'espace d'origine. On peut, par exemple, tenter de conserver la relation d'ordre qui existe entre les distances inter-observations au niveau des distances entre les points image du plan [KRU - 64] [SHE - 62] [SAM - 69].

Plutôt que de minimiser les distorsions entre les distances inter-observations et les distances inter-points image, d'autres auteurs cherchent à préserver la séparabilité entre les classes en représentant les données en deux dimensions [SAM - 70] [FUK - 71] [FUK - 82].

Cependant, ces techniques ne sont pas à proprement parler des méthodes de classification automatique. Leur rôle se limite à réduire la dimension de l'espace de représentation des données, laissant au système visuel humain le soin d'achever la classification.

### **I.3. PROCEDURES LOCALES.**

Les méthodes globales prennent en considération l'ensemble des observations disponibles pour les séparer en classes, soit par des techniques d'estimation des paramètres, soit par des techniques d'optimisation de critères indiquant la qualité de la répartition entre différentes classes, soit encore en faisant appel aux facultés de perception et d'analyse bidimensionnelle du système visuel humain.

A l'opposé, les méthodes locales analysent de manière plus ponctuelle la répartition des observations, avec pour objectif de mettre en évidence soit des concentrations locales d'observations qui correspondent à des classes, soit des zones circonscrites de l'espace relativement vides d'observations qui correspondent aux vallées séparant ces classes.

### **I.3.1. Détection des modes par recherche des maxima locaux.**

Dans ce type d'approche, il s'agit d'analyser la fonction de densité sous-jacente à la distribution des observations disponibles pour extraire l'information nécessaire à leur classification. En effet, on peut estimer cette fonction de densité par une méthode non paramétrique [PAR - 62] [LOF - 65] pour mettre ses maxima locaux en évidence.

Dans la plupart des méthodes basées sur l'estimation des fonctions de densité, il est admis que chaque maximum local correspond à une concentration locale d'observations et la recherche des classes peut alors être assimilée à la recherche des modes de cette fonction [DEV - 82] [ASS - 89].

Les modes peuvent être détectés en remontant les pentes de la fonction de densité de probabilité selon la direction de son gradient [KOO - 76] ou en déplaçant progressivement les observations jusqu'à ce que chaque observation atteigne le voisinage d'un mode de cette fonction [BOC - 79]. Une variante de cette approche consiste à calculer le gradient directement à partir des observations [FUK - 75a].

En adoptant une technique de sériation, Kittler a construit une séquence de points de telle sorte que la majorité des points voisins de chaque mode deviennent des éléments successifs de cette séquence [KIT - 76].

Toutes ces techniques sont connues pour être sensibles aux irrégularités des distributions des observations, et tendent à générer de nombreux modes parasites, difficiles à différencier des véritables modes de la fonction de densité.

Touzani a développé une technique de relaxation directement appliquée à la fonction de densité estimée, permettant de diminuer l'effet de ces irrégularités [TOU - 88].

### **I.3.2. Analyse de la convexité.**

Au lieu de considérer les modes comme des extrema locaux de la fonction de densité de probabilité, C. Vasseur et J.-G. Postaire [VAS - 80] les assimilent à des régions de l'espace où cette fonction est concave. Dans cette approche, l'analyse de la convexité de la fonction de densité est effectuée en intégrant cette dernière sur des domaines d'observation de taille variable [POS - 82b]. Cette analyse améliore considérablement la robustesse de la méthode par rapport aux techniques faisant appel aux notions de gradient, mais elle reste encore sensible aux irrégularités de la distribution des données. Des techniques d'étiquetage probabiliste ou "relaxation" ont été développées pour améliorer la robustesse de la méthode [OLE - 88].

### **I.3.3. Extraction des contours des modes.**

A. Touzani et J.-G. Postaire considèrent les modes comme des régions délimitées par leur contour. Après réalisation d'un filtrage médian multidimensionnel de la fonction de densité, des opérateurs différentiels multidimensionnels permettent d'extraire les contours des modes [TOU - 89]. Des techniques d'étiquetage probabiliste itératif [POS - 89] ont par ailleurs été mises en oeuvre pour augmenter la robustesse de ces méthodes.

En résumé, toutes ces méthodes de classification font appel, directement ou indirectement, à des estimateurs non paramétriques de la fonction sous-jacente, ce qui nécessite un nombre d'observations disponibles très important surtout lorsque la dimension des observations est élevée. De plus, de nombreuses procédures utilisent des opérateurs différentiels locaux qui sont sensibles au bruit et aux irrégularités de la

distribution des observations, ce qui nécessite un filtrage préalable et/ou des techniques de traitement post-opérateur afin d'améliorer la robustesse des méthodes proposées.

De nombreux auteurs ont mis en évidence l'intérêt des méthodes locales dans le cas de classes de formes et de tailles très variées, les méthodes globales étant surtout adaptées aux classes de formes sphériques, ou tout du moins globulaires.

#### **I.4. CONCLUSION.**

La plupart des méthodes locales que nous venons de décrire sont essentiellement basées sur des considérations statistiques, et il serait intéressant d'analyser localement la distribution des observations de l'échantillon disponible en utilisant des critères géométriques.

Nous proposons, dans ce mémoire, une nouvelle approche locale pour la classification automatique, l'analyse des données étant fondée sur l'utilisation de critères géométriques et structuraux. La base théorique de cette approche est celle de la morphologie mathématique. Notre objectif n'est pas de présenter une méthode universelle, mais de montrer l'intérêt d'une telle approche dans les domaines de l'analyse de données et de la reconnaissance des formes.

Conscients des difficultés qu'éprouve en général l'analyste pour choisir une méthode particulière parmi toutes celles disponibles, nous nous attacherons à définir clairement les intérêts et les limites d'applications des méthodes proposées. A cet effet, nous précisons les hypothèses de travail que nous nous sommes fixés. Nous nous sommes placés dans un contexte non supervisé, sans aucune hypothèse paramétrique, afin de traiter des observations multidimensionnelles à valeurs réelles, représentées dans l'espace euclidien  $E^N$ . Enfin, nous admettrons qu'il est possible de réaliser une partition de l'ensemble des données en classes, deux à deux disjointes.

En règle générale, les méthodes locales aboutissent à la classification des données par l'intermédiaire de deux étapes successives : une étape de détection des modes et une étape de classification des observations s'appuyant sur les modes détectés. En général, on s'accorde à définir un mode comme une région de forte concentration locale d'observations. Chaque mode peut être décrit soit par un extremum local de la fonction de densité, soit par un domaine de l'espace  $E^N$  entourant cet extremum, soit encore par son noyau, que nous définirons comme l'ensemble des observations situées à l'intérieur de ce domaine. Le domaine modal associé à la classe  $C_k$ ,  $k = 1, 2, \dots, K$  sera noté  $D_k$ ,  $k = 1, 2, \dots, K$ . L'étape finale de classification des observations consiste à assigner les observations aux différentes classes associées aux domaines modaux mis en évidence au cours de la procédure de détection des modes. Cette affectation est réalisée en respectant une règle de décision, généralement définie par un ensemble de fonctions de décision associées aux classes détectées.

La nouvelle approche pour la classification automatique que nous présentons dans ce mémoire peut également être décrite par ces deux étapes successives. La détection des modes est basée sur des critères géométriques relevant de la morphologie mathématique dont l'adaptation aux problèmes de classification automatique est présentée dans la première partie de ce mémoire. Dans le second chapitre, nous montrons comment un ensemble d'observations multivariées représenté dans l'espace euclidien  $E^N$  peut être transformé en un ensemble "discret" dans l'espace  $(Z^+)^N$ , espace dans lequel nous définissons les notions de base de la morphologie mathématique binaire.

La deuxième partie de ce mémoire est consacrée à l'extraction des modes qui est réalisée par une analyse morphologique des données à classer. Les modes sont alors définis comme un sous-ensemble de  $(Z^+)^N$ . En combinant deux opérateurs morphologiques binaires de base, nous définissons, dans le chapitre III, deux techniques de détection des modes, chaque mode étant défini soit par son domaine modal, soit par la frontière de ce domaine, appelée frontière modale. Le chapitre IV présente deux

autres méthodes d'extraction des modes basées sur l'analyse morphologique de la connexité des points de l'ensemble discret. Ces méthodes ont un pouvoir de discrimination entre les classes plus faible que celles du chapitre III, mais elles permettent de classer de plus petits échantillons avec succès.

La troisième partie de ce mémoire a pour but de montrer l'intérêt de l'approche morphologique en classification automatique. Dans le chapitre V, sont définies et comparées des procédures de classification permettant d'obtenir une partition de l'ensemble des observations à partir des modes extraits. Le chapitre VI illustre cette méthodologie de classification morphologique à l'aide d'ensembles de données générées artificiellement. Afin de préciser les domaines de validité de cette approche, nous comparons les résultats obtenus à ceux de deux méthodes bien établies en classification automatique.

**II - ADAPTATION DES TRANSFORMATIONS  
MORPHOLOGIQUES A L'ANALYSE DES  
DONNEES MULTIDIMENSIONNELLES.**

## ADAPTATION DES TRANSFORMATIONS MORPHOLOGIQUES A L'ANALYSE DE DONNEES MULTIDIMENSIONNELLES.

### II.1. INTRODUCTION.

Nous avons vu, dans le chapitre précédent, que nous pouvions aborder le problème de la classification automatique sous l'angle de la détection des modes, ceux-ci étant définis comme des régions de l'espace euclidien de représentation des données présentant de fortes concentrations locales d'observations. Ces modes peuvent être mis en évidence par différentes techniques. Nous avons choisi de les extraire par des techniques d'analyse structurale et morphologique, en nous basant sur la théorie de la morphologie mathématique.

La morphologie mathématique est née en 1964 d'une collaboration de G. Matheron et de J. Serra. De 1964 à 1968, ces auteurs ont essentiellement mis au point les notions de transformations en tout ou rien, les opérateurs d'ouverture, de fermeture et les modèles booléens, en même temps que se créait le Centre de Morphologie Mathématique à Fontainebleau.

En fait, la morphologie mathématique est fondée sur des concepts de la théorie des ensembles, et regroupe trois aspects fondamentaux : un aspect algébrique, un aspect probabiliste et un de géométrie intégrale. Elle est basée sur des transformations de l'ensemble à analyser qui est comparé à un ensemble prédéfini, appelé **élément structurant**, de structure connue et généralement plus simple. Le but de ces transformations ensemblistes est alors d'extraire de l'ensemble analysé des caractéristiques structurales et morphologiques. L'élément structurant est défini par son origine (appelée également son centre) et par sa structure spécifique. Aux propriétés

géométriques de chaque élément structurant (symétrie, isotropie, connexité) correspond une signification géométrique ou structurale des transformations associées.

Essentiellement appliquée aux images binaires lors de son introduction par G. Matheron [MAT - 75] et J. Serra [SER - 82] en France et par S. Sternberg [STE - 82] aux Etats-Unis, la Morphologie mathématique a été étendue aux images en niveaux de gris [STE - 86].

Il est possible de distinguer trois grands types d'applications des transformations morphologiques en analyse d'images. Tout d'abord, sont apparues les transformations de type tout ou rien comme la dilatation, l'érosion, l'ouverture et la fermeture [HAR - 87] dans le but de filtrer l'image à analyser [MAR - 87a] [MAR - 87b]. Ensuite, des outils de la morphologie mathématique permettant de réduire l'image sans en altérer ses caractéristiques géométriques, telles que des transformations de type squelettisation, amincissement et épaissement ont également été développés [MAR - 86]. Enfin, plus récemment, la morphologie a été utilisée pour réaliser des transformations assurant la décomposition des images binaires en éléments simples en vue de leur reconnaissance [PIT - 90]. En règle générale, à chaque type de transformation est associé un type d'élément structurant. Récemment, J. Song et E. Delp ont montré que l'utilisation de plusieurs éléments structurants permettait d'améliorer les performances des filtres morphologiques [SON - 90].

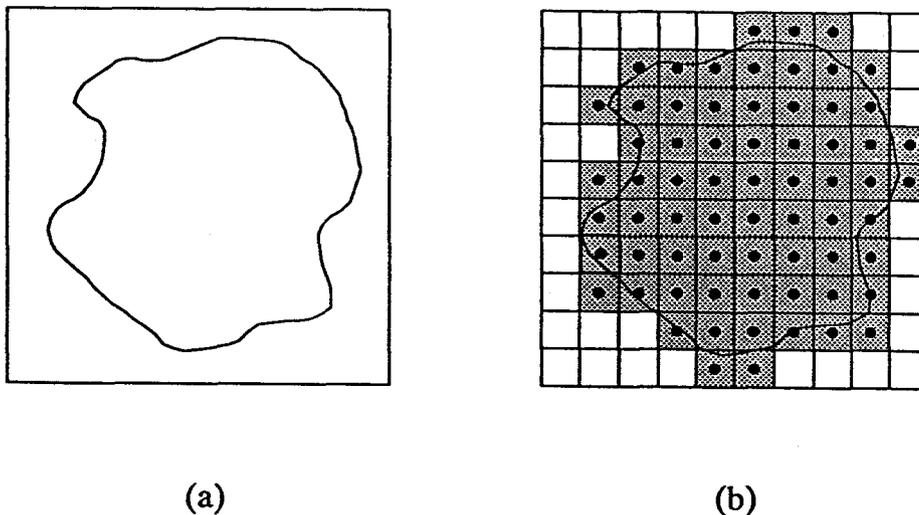
Pour répondre aux contraintes des systèmes numériques de traitement d'image, la morphologie mathématique a été développée sur des espaces numérisés. J. Serra a décrit la numérisation d'une image binaire, en trois phases principales [SER - 82].

Tout d'abord, l'espace continu  $E$  est remplacé par un réseau de points régulièrement espacés, les deux principaux réseaux d'échantillonnage utilisés étant le réseau carré et le réseau hexagonal. On considère alors l'intersection de l'image avec ce réseau de points d'échantillonnage, et on assigne la valeur 1 à tous les points du réseau ayant une intersection non nulle avec les objets de l'image, et la valeur 0 à tous les

autres points du réseau.

La notion de réseau étant insuffisante pour permettre une analyse des images, il est nécessaire de définir la notion de graphe. Soit  $\underline{X}$  l'ensemble des points du réseau à valeur 1, et  $\underline{X}^C$  l'ensemble des points à valeur 0. Un graphe, généralement noté  $(\underline{X}, D_n)$  est défini par deux ensembles : celui des sommets, qui sont les points du réseau d'échantillonnage appartenant à  $\underline{X}$ , et celui des arêtes qui joignent les couples de points de l'ensemble  $\underline{X}$ . A tout graphe  $(\underline{X}, D_n)$  est associé le graphe complémentaire  $(\underline{X}^C, D_n)$ ,  $n$  indiquant le type de graphe utilisé.

Sur le réseau carré, on définit le plus souvent, soit le graphe carré noté  $D_4$ , soit le graphe octogonal  $D_8$ , tandis que sur le réseau hexagonal est généralement défini le graphe hexagonal  $D_6$ . La figure II.1 illustre le passage d'un espace continu à un espace discrétisé et la figure II.2 montre les différents graphes possibles. En règle générale, le réseau et le graphe hexagonal associé sont les mieux adaptés à l'analyse morphologique des images du fait de leur isotropie.



**Figure II.1 :** Passage d'un espace continu à un espace discrétisé en réseau carré.

*a - Image continue.*

*b - Image discrétisée avec représentation du réseau carré.*

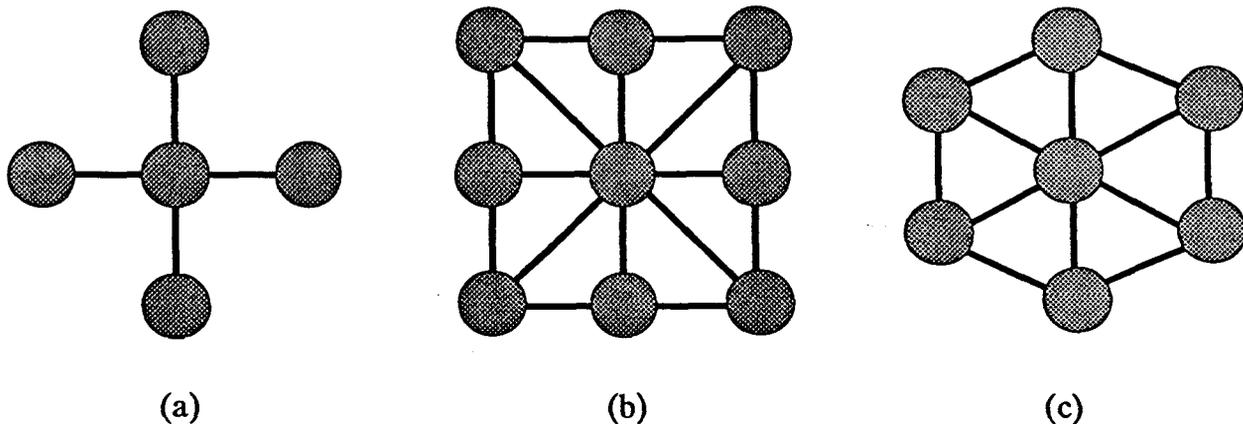


Figure II.2 : Les différents graphes utilisés en analyse d'images.

*a - Graphe carré associé au réseau carré.*

*b - Graphe octogonal associé au réseau carré.*

*c - Graphe hexagonal associé au réseau hexagonal.*

Dans le cadre de notre étude, les données à analyser sont des ensembles d'observations multivariées représentées par des points dans l'espace euclidien  $E^N$ . Afin d'appliquer la théorie de la morphologie mathématique à l'analyse de ces données, il est nécessaire de commencer par discrétiser leur espace de représentation. Il s'agit donc de transformer l'ensemble des observations disponibles en un ensemble discret de points à valeurs soit binaires, par analogie à la morphologie binaire du traitement d'image, soit multiniveaux par analogie à la morphologie en niveau de gris. Les transformations que nous décrivons dans ce mémoire sont de type tout ou rien, c'est-à-dire que la fonction mathématique associée à la transformation prend ses valeurs dans l'ensemble  $\{0, 1\}$ . D'autres travaux, actuellement en cours d'étude dans notre laboratoire, ont pour but d'adapter les concepts de la morphologie multiniveaux au domaine de la classification automatique.

La première phase de la discrétisation de l'espace de représentation des données consiste à définir, sur l'espace euclidien  $E^N$ , un réseau d'échantillonnage qu'il est nécessaire de préciser. Des travaux récents ont montré que l'on pouvait appliquer les outils morphologiques de base à l'analyse de données discrétisées sur un réseau

"pseudo-hexagonal", obtenu après translation d'un réseau carré [DOU - 91a] [DOU - 91b]. Néanmoins, ces transformations morphologiques n'ont été appliquées que dans l'espace bidimensionnel après une décomposition des données en composantes principales. Or, notre analyse des données a lieu directement dans l'espace multidimensionnel des observations et, dans ce contexte, ce réseau est beaucoup plus difficilement exploitable. Nous avons donc opté pour l'utilisation du réseau hypercubique, ce réseau étant l'équivalent du réseau carré dans l'espace bidimensionnel, ainsi que pour un graphe dont l'ensemble des arcs sera noté  $D_{3N}$ , ce graphe étant l'équivalent du graphe octogonal dans l'espace multidimensionnel.

Dans le second paragraphe de ce chapitre, nous montrons comment l'ensemble des observations est transformé en un ensemble discret de points à valeurs binaires, représenté sur un réseau de discrétisation hypercubique. Les paragraphes suivants ont pour but de présenter, dans un contexte multidimensionnel, les différentes transformations morphologiques binaires élémentaires, ainsi que le principe général des transformations à éléments structurants multiples encore nommées transformations de voisinage [COS - 85]. Nous exploiterons ces transformations pour la détection des modes morphologiques.

## II.2. TRANSFORMATION DES DONNEES INITIALES EN UN ENSEMBLE DISCRET D'ELEMENTS A VALEURS BINAIRES.

Soit  $Y^*$  un échantillon de  $Q$  observations multidimensionnelles  $Y_1, Y_2, \dots, Y_q, \dots, Y_Q$  avec :

$$Y_q = [y_q^1, y_q^2, \dots, y_q^n, \dots, y_q^N]^T,$$

où  $y_q^n$  désigne la  $n$ -ième coordonnée de l'observation  $Y_q$ .

Afin d'utiliser les outils de la morphologie mathématique pour la recherche des modes des différentes classes constituant l'échantillon, celui-ci doit subir, dans un premier temps, une transformation permettant de le représenter dans l'espace discret  $(Z^+)^N$ .

Pour effectuer cette transformation, l'origine  $O$  de l'espace euclidien  $E^N$  est d'abord translaturée au point  $O'$  (cf. figure II.3.a) de coordonnées :

$$O' = [ \min_q y_q^1, \min_q y_q^2, \dots, \min_q y_q^n, \dots, \min_q y_q^N ]^T.$$

Une transformation définie par une matrice diagonale telle que :

$$y'_q{}^n = (y_q^n - \min_q y_q^n) * R/L_n,$$

$$\text{où } L_n = \max_q y_q^n - \min_q y_q^n.$$

permet alors de travailler dans un espace où les valeurs extrêmes des  $N$  attributs définissant les observations sont identiques. Par conséquent, après cette normalisation, les observations sont situées à l'intérieur d'un hypercube de côté égal à  $R$  (cf. figure II.3.b).

Chaque axe du nouvel espace de représentation des données est alors découpé en  $R$  intervalles adjacents égaux de longueur unité. Cette discrétisation définit un ensemble de  $R^N$  hypercubes de côté unité, dont les centres constituent un réseau régulier de points d'échantillonnage (cf. figure II.3.c). Pour accélérer les calculs, chaque centre :

$$X = [x^1, \dots, x^n, \dots, x^N]^T$$

est repéré par les  $N$  parties entières de ses  $N$  coordonnées. On définit l'hypercube correspondant :

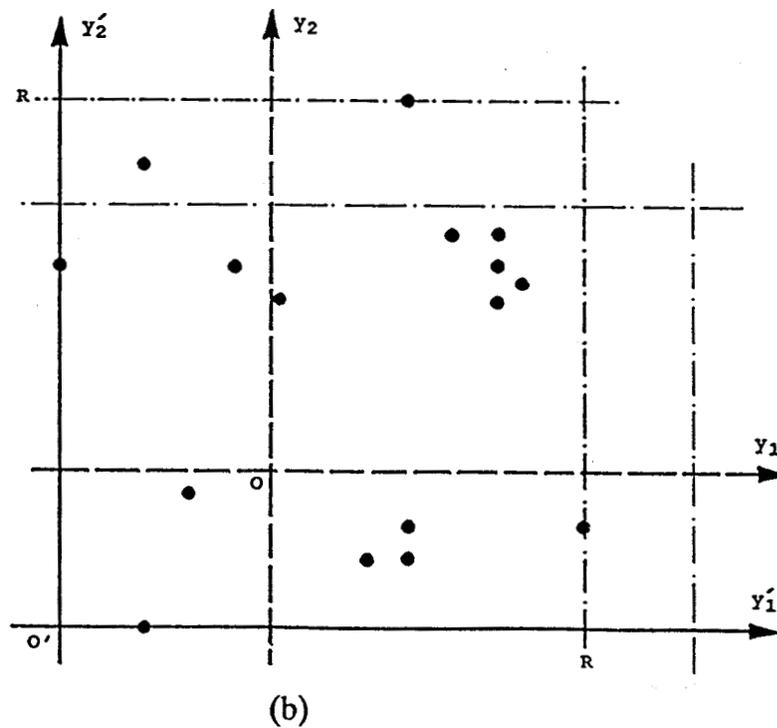
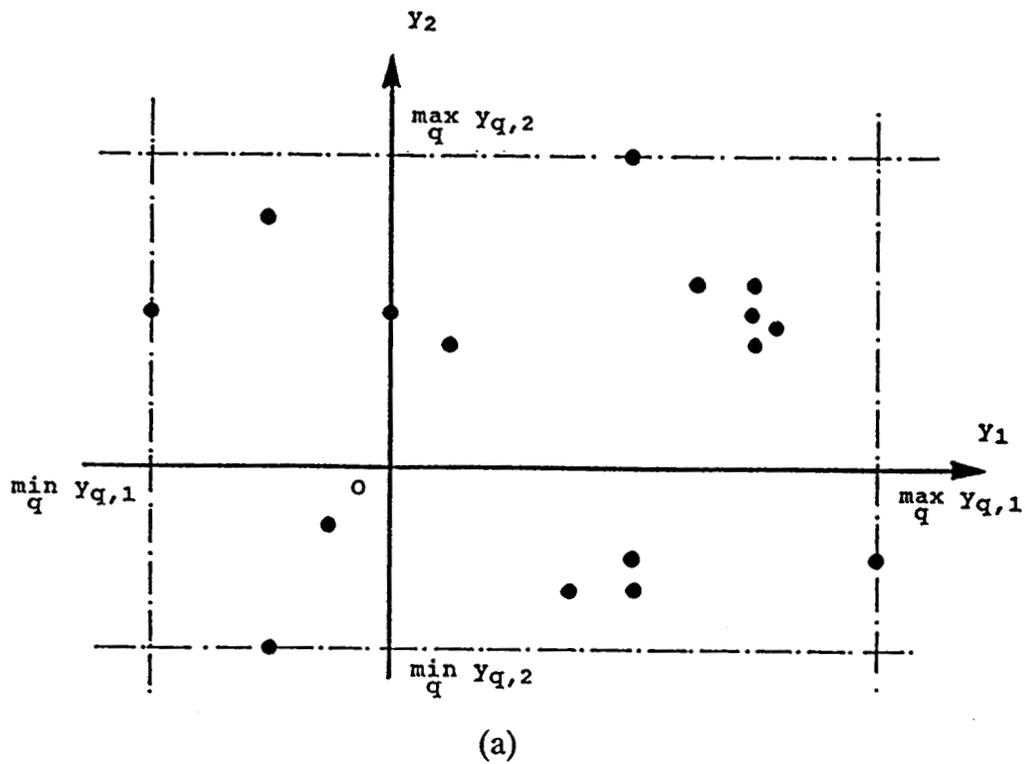
$$H = [h^1, \dots, h^n, \dots, h^N]^T$$

par  $N$  coordonnées entières, égales à celles de son centre.

Pour simplifier la description de la distribution des observations  $Y'_q$  dans l'espace normalisé, où  $Y'_q = [y'_q{}^1, \dots, y'_q{}^n, \dots, y'_q{}^N]^T$ , on détermine les hypercubes unitaires non vides. La liste de ces hypercubes non vides peut être déduite immédiatement de celle des observations  $Y'_q$ ,  $q = 1, 2, \dots, Q$ . En effet, on peut montrer que l'observation  $Y'_q$  est située dans l'hypercube de coordonnées :

$$H_q = [INT(y'_q{}^1), \dots, INT(y'_q{}^n), \dots, INT(y'_q{}^N)]^T$$

où  $\text{INT}(y'_q{}^n)$  représente la partie entière de la n-ième coordonnée de l'élément  $Y'_q$  [POS - 82a].



**Figure II.3 :** Représentation d'un ensemble de données multivariées par un ensemble discret.

- a - Ensemble des observations brutes.
- b - Normalisation des observations.

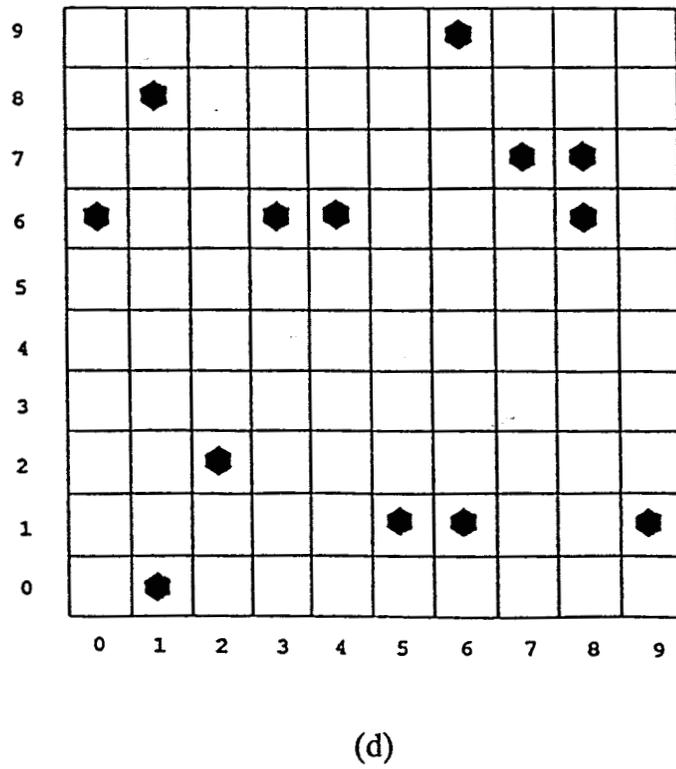
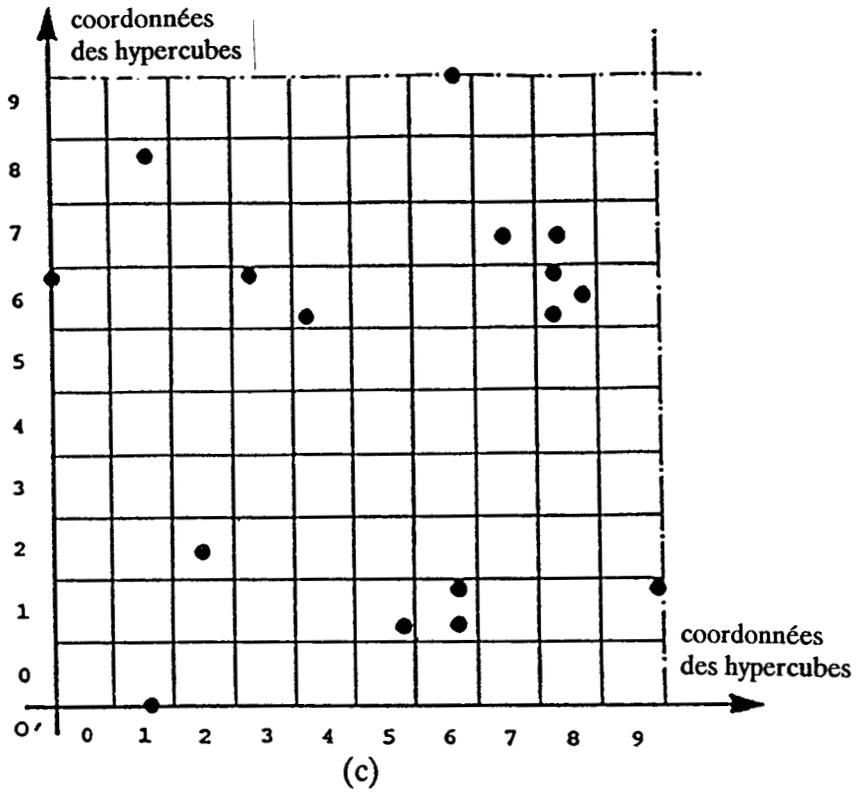


Figure II.3 : Représentation d'un ensemble de données multivariées par un ensemble discret (suite).

- c - Réseau de points d'échantillonnage ( $R = 10$ ).
- d - Ensemble discret de points à valeurs binaires.

En prenant la partie entière de toutes les coordonnées des  $Q$  observations disponibles, on obtient la liste des hypercubes non vides. Si plusieurs observations sont situées dans le même hypercube, celui-ci apparaît plusieurs fois dans la liste, mais il est aisé d'éliminer les répétitions. De plus, tous les hypercubes qui ne figurent pas dans cette liste sont vides de toute observation. Cet ensemble d'hypercubes non vides peut être considéré comme une version simplifiée des données initiales (cf. figure II.3.d). Une fonction binaire  $B$  est définie sur la base de cette discrétisation de sorte qu'à chaque hypercube non vide du réseau d'échantillonnage est associée la valeur 1, tandis qu'à chaque hypercube vide du réseau est associée la valeur 0.

Plus précisément, soit  $\underline{X}$  l'ensemble discret constitué des hypercubes non vides du réseau et soit  $\underline{X}^C$  l'ensemble des hypercubes vides de ce réseau.

La fonction binaire  $B$  est définie sur ce réseau par :

$$\begin{aligned} B : (\mathbb{Z}^+)^N &\longrightarrow \{0, 1\} \\ X &\longmapsto B(X) \end{aligned}$$

où  $B(X) = 0$  lorsque l'hypercube de centre  $X$  est vide de toute observation, et  $B(X) = 1$  lorsque l'hypercube de centre  $X$  est non vide.

Le résultat de cette procédure de discrétisation est donc l'ensemble discret  $\underline{X}$  constitué de tous les points  $X$  tels que  $B(X) = 1$ . Dans cette procédure, le seul paramètre à ajuster par l'opérateur est le pas de discrétisation  $R$  dont le choix est déterminant pour la suite des traitements. En effet, si le pas  $R$  est trop grand, la discrétisation est trop fine de telle sorte que les hypercubes non vides sont relativement isolés. Il est alors difficile de mettre en évidence les différents modes. Par contre, si  $R$  est trop petit, la résolution est trop faible, et il est alors difficile d'identifier les différents modes. En fait, le pas  $R$  dépend non seulement du nombre total  $Q$  d'observations de l'échantillon et de la dimension  $N$  de l'espace de représentation des données, mais aussi de la structure de la distribution de ces données. Or, comme nous nous sommes placés dans un contexte non supervisé, nous ne disposons d'aucune information concernant

cette structure. Nous proposons donc, dans la deuxième partie de ce mémoire, un moyen d'ajuster au mieux ce paramètre.

En résumé, dans l'espace euclidien  $E^N$ , les classes de l'ensemble à analyser sont caractérisées par une forte concentration locale d'observations séparées par des régions faiblement peuplées, sinon vides d'observations. Lorsque l'ensemble d'observations a subi la transformation précédemment décrite, ces différences de concentrations locales sont reflétées par la structure spatiale de l'ensemble discret  $\underline{X}$ . En effet, dans l'espace discrétisé, les modes associés aux classes en présence sont représentés par des sous-ensembles connexes de points de  $\underline{X}$  de sorte que chaque classe peut être identifiée à l'une de ces composantes connexes.

L'examen de la structure de l'ensemble discret montre qu'au coeur des classes, chaque point  $X$  de  $\underline{X}$  tend à avoir un grand nombre de voisins adjacents appartenant à  $\underline{X}$ , alors qu'à leur périphérie, où les observations sont relativement éloignées les unes des autres, les points de  $\underline{X}$  tendent à avoir moins de voisins de  $\underline{X}$  immédiats, sinon à n'être entourés que de points de  $\underline{X}^C$  (cf. figure II.4).

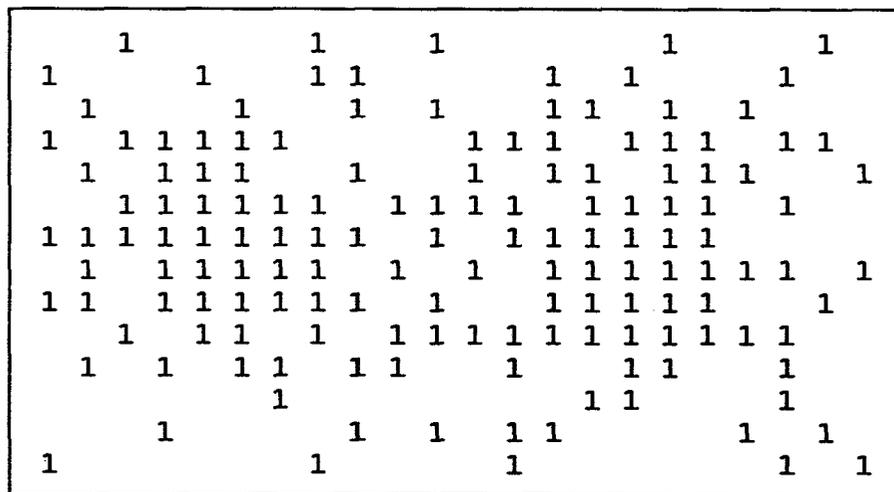


Figure II.4 : Structure spatiale de l'ensemble  $\underline{X}$ .

Pour détecter ces différences d'agencement spatial, nous analysons le voisinage

de chaque point  $X$  de  $\underline{X}$  en utilisant les outils de la morphologie mathématique dont nous rappelons les principes de base dans le cadre de la morphologie binaire.

### II.3. TRANSFORMATIONS MORPHOLOGIQUES BINAIRES ELEMENTAIRES.

Avant de définir les transformations morphologiques élémentaires, il est nécessaire de préciser les notations utilisées.

On notera :

$\underline{X}$  : un ensemble discret de points  $X$  tels que  $B(X) = 1$ .

$\underline{X}^C$  : l'ensemble complémentaire de  $\underline{X}$  de points  $X$  tels que  $B(X) = 0$ .

$\underline{X}^{\sim}$  : l'ensemble symétrique de l'ensemble  $\underline{X}$  par rapport à l'origine.

$X$  : un point du réseau d'échantillonnage, dont les coordonnées sont des entiers positifs ou nuls.

$x^n$  : n-ième coordonnée du point  $X$ , avec :

$$X = [x^1, x^2, \dots, x^n, \dots, x^N]^T.$$

Dans la suite de ce mémoire, nous utilisons indifféremment le terme de point ou d'hypercube. Un point  $X$  appartient à  $\underline{X}$  s'il est le centre d'un hypercube non vide. Par contre, un point de  $\underline{X}^C$  correspond à un hypercube vide de toute observation.

Comme nous l'avons mentionné précédemment, le principe fondamental de la morphologie consiste à comparer l'ensemble à analyser à un ensemble prédéfini appelé élément structurant, afin d'en extraire des caractéristiques structurales ou morphologiques. Le rôle de la morphologie est de transformer l'ensemble à étudier en un ensemble de structure généralement plus simple, sans en altérer ses principales caractéristiques géométriques.

Les transformations morphologiques les plus simples pouvant être appliquées à un ensemble discret de points à valeurs binaires sont l'érosion et la dilatation, issues des

opérations ensemblistes d'addition et de soustraction définies par Minkowski.

### **II.3.1. La dilatation.**

Soient  $\underline{X}$  et  $\underline{S}$  deux ensembles discrets de  $(\mathbb{Z}^+)^N$  dont les points sont respectivement définis par :

$$\underline{X} = [x^1, x^2, \dots, x^n, \dots, x^N]^T$$

et par :

$$\underline{S} = [s^1, s^2, \dots, s^n, \dots, s^N]^T$$

Soit  $(\underline{X})_{\underline{S}}$  le résultat de la translation de  $\underline{X}$  par  $\underline{S}$ , défini par :

$$(\underline{X})_{\underline{S}} = [T \in \mathbb{Z}^N \mid T = X + S, X \in \underline{X}]$$

La dilatation de  $\underline{X}$  par  $\underline{S}$  est définie à partir de l'addition de Minkowski, sous la forme :

$$\underline{X} \oplus \underline{S}^{\sim} = \cup_{S \in \underline{S}^{\sim}} (\underline{X})_S = [X \mid (\underline{S})_X \cap X \neq \emptyset, X \in \underline{X}]$$

La notation utilisée ici est celle adoptée par J. Serra [SER - 82] et G. Matheron [MAT - 75]. D'autres auteurs tels que S. Sternberg [STE - 82], R. Haralick [HAR - 87] et H. Heijmans [HEI - 90], ont utilisé une notation différente. Ces deux types de notations diffèrent dans le sens où, pour certaines opérations, il faut remplacer l'ensemble  $\underline{S}^{\sim}$  des définitions de J. Serra par l'ensemble  $\underline{S}$  pour les définitions de S. Sternberg. Néanmoins, les résultats fondamentaux restent vérifiés quelle que soit la notation adoptée.

Dans cette définition de la dilatation, les ensembles  $\underline{X}$  et  $\underline{S}$  jouent, théoriquement, des rôles symétriques. Néanmoins, dans la pratique,  $\underline{X}$  désigne l'ensemble discret associé à l'ensemble des observations à classer, tandis que  $\underline{S}$  désigne l'élément structurant utilisé pour réaliser la transformation désirée. Désignons par  $G$  l'origine de l'élément structurant  $\underline{S}$ .

Les propriétés essentielles de la dilatation sont les suivantes :

1) Invariance par translation

$$\forall T \in (\mathbb{Z}^+)^N, (\underline{X})_T \oplus \underline{S} = (\underline{X} \oplus \underline{S})_T$$

2) Commutativité

$$\underline{X} \oplus \underline{S} = \underline{S} \oplus \underline{X}$$

3) Associativité

$$(\underline{X} \oplus \underline{R}) \oplus \underline{S} = \underline{X} \oplus (\underline{R} \oplus \underline{S})$$

4) Distributivité par rapport à l'union

$$(\underline{X} \cup \underline{Y}) \oplus \underline{S} = (\underline{X} \oplus \underline{S}) \cup (\underline{Y} \oplus \underline{S})$$

5) Croissance

$$\text{Si } \underline{X} \subset \underline{Y}, \text{ alors } \underline{X} \oplus \underline{S} \subset \underline{Y} \oplus \underline{S}$$

6) Extensivité

$$\text{Si } G \in \underline{S}, \text{ alors } \underline{X} \subset \underline{X} \oplus \underline{S}$$

Cette dernière propriété explique le terme "Dilatation". En effet, lorsque l'origine  $G$  de l'élément structurant appartient à l'élément structurant, la version dilatée de l'ensemble discret contient l'ensemble discret original.

### II.3.2. L'érosion

L'érosion, qui est issue de la soustraction ensembliste de Minkowski, est l'opération duale de la dilatation par rapport à la complémentation.

Elle est définie, avec les notations de J. Serra et de G. Matheron, par :

$$\underline{X} \ominus \underline{S} = \bigcap_{S \subset \underline{S}} (\underline{X})_{-S} = \{ X \mid (\underline{S})_X \subset \underline{X}, X \in \underline{X} \}$$

où  $-S$  désigne le symétrique de  $S$  par rapport à l'origine de  $\underline{S}$ .

L'ensemble érodé est construit en translatant  $\underline{X}$  par tous les éléments du symétrique de  $\underline{S}$  par rapport à son origine et en prenant l'intersection de ces ensembles translétés.

Les propriétés essentielles de l'érosion sont les suivantes :

1) Invariance par translation

$$\forall T \in (\mathbb{Z}^+)^N, (\underline{X})_T \ominus \underline{S} = (\underline{X} \ominus \underline{S})_T$$

2) Composition en série

$$(\underline{X} \ominus \underline{R}) \ominus \underline{S} = \underline{X} \ominus (\underline{R} \oplus \underline{S})$$

3) Composition parallèle

$$(\underline{X} \cap \underline{Y}) \ominus \underline{S} = (\underline{X} \ominus \underline{S}) \cap (\underline{Y} \ominus \underline{S})$$

$$\underline{X} \ominus (\underline{R} \cup \underline{S}) = (\underline{X} \ominus \underline{R}) \cap (\underline{X} \ominus \underline{S})$$

4) Croissance

$$\text{Si } \underline{X} \subseteq \underline{Y}, \text{ alors } \underline{X} \ominus \underline{S} \subseteq \underline{Y} \ominus \underline{S}$$

5) Anti-extensivité

$$\text{Si } G \in \underline{S}, \text{ alors } \underline{X} \ominus \underline{S} \subseteq \underline{X}$$

Au regard de cette dernière propriété, il apparaît que si l'élément structurant contient son origine, l'ensemble transformé par l'érosion est contenu dans l'ensemble discret de départ. Cette remarque justifie la dénomination de la transformation.

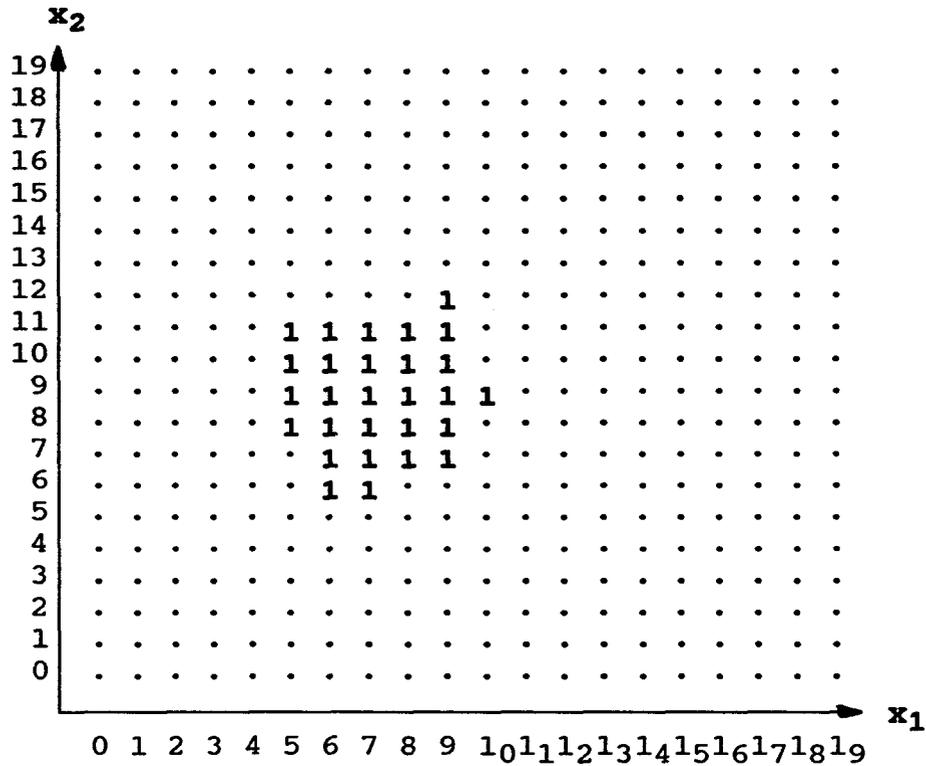
Afin de montrer l'intérêt des opérations de dilatation et d'érosion, considérons un échantillon  $Y^*$ , présenté figure II.5.a, composé de 200 observations bidimensionnelles réparties selon une loi normale et générées artificiellement. L'ensemble discret  $\underline{X}$  correspondant à cet échantillon est alors l'ensemble de tous les points repérés par la valeur 1 sur la figure II.5.b où le pas de discrétisation de l'espace a été fixé à  $R = 20$ .

La dilatation de l'ensemble  $\underline{X}$  par un élément structurant de taille  $3 \times 3$  est présentée sur la figure II.5.c tandis que l'effet d'une érosion par le même élément structurant est illustré par la figure II.5.d. Notons que plus l'élément structurant est grand, plus l'effet du filtrage est important.

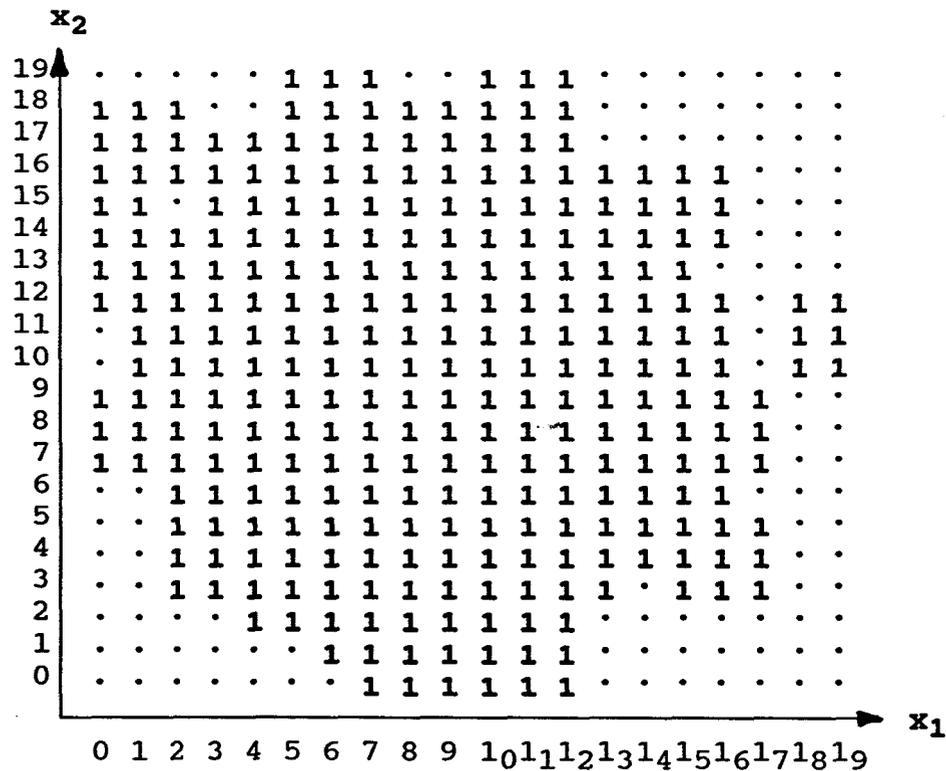
Nous rappelons que l'érosion est l'opération duale de la dilatation, c'est-à-dire :

$$(\underline{X} \ominus \underline{S})^C = (\underline{X})^C \oplus \underline{S}^{\sim}$$





(c)



(d)

Figure II.5 : Erosion et dilatation d'un échantillon de 200 observations normales (suite).

c - Erosion de l'ensemble  $\underline{X}$ .  
 d - Dilatation de l'ensemble  $\underline{X}$ .

### II.3.3. L'ouverture et la fermeture.

En pratique, l'érosion et la dilatation sont rarement utilisées seules. Leurs combinaisons donnent naissance à deux autres opérations morphologiques : l'ouverture et la fermeture.

L'ouverture de  $\underline{X}$  par  $\underline{S}$ , notée  $\underline{X}_S$ , est l'ensemble résultant d'une érosion de  $\underline{X}$  par  $\underline{S}$ , suivie d'une dilatation de l'ensemble érodé par le même élément structurant :

$$\underline{X}_S = (\underline{X} \ominus \underline{S}) \oplus \underline{S}$$

En règle générale, l'ensemble résultant de l'ouverture  $\underline{X}_S$  diffère sensiblement de l'ensemble de départ  $\underline{X}$ . L'opération d'ouverture supprime les petits détails se trouvant à la périphérie des sous-ensembles connexes, à condition toutefois que ces détails soient de taille inférieure à celle de l'élément structurant. L'ensemble ouvert est, en quelque sorte, une version simplifiée de l'ensemble  $\underline{X}$ .

L'opération duale, nommée fermeture et notée  $\underline{X}^S$ , est le résultat d'une dilatation suivie d'une érosion, en utilisant le même élément structurant :

$$\underline{X}^S = (\underline{X} \oplus \underline{S}) \ominus \underline{S}$$

Cette opération de fermeture permet de combler les "trous" dans l'ensemble discret  $\underline{X}$  résultant de l'irrégularité de la distribution des observations.

A partir des propriétés de l'érosion et de la dilatation, on peut déduire que l'ouverture et la fermeture sont toutes deux :

- invariantes par translation :

$$\forall T \in (\mathbb{Z}^+)^N, (\underline{X}_S)_T = (\underline{X})_T \ominus \underline{S}$$

$$\forall T \in (\mathbb{Z}^+)^N, (\underline{X}^S)_T = (\underline{X})_T \oplus \underline{S}$$

- croissantes :

$$\text{Si } \underline{X} \subset \underline{Y}, \text{ alors } \underline{X}_S \subset \underline{Y}_S$$

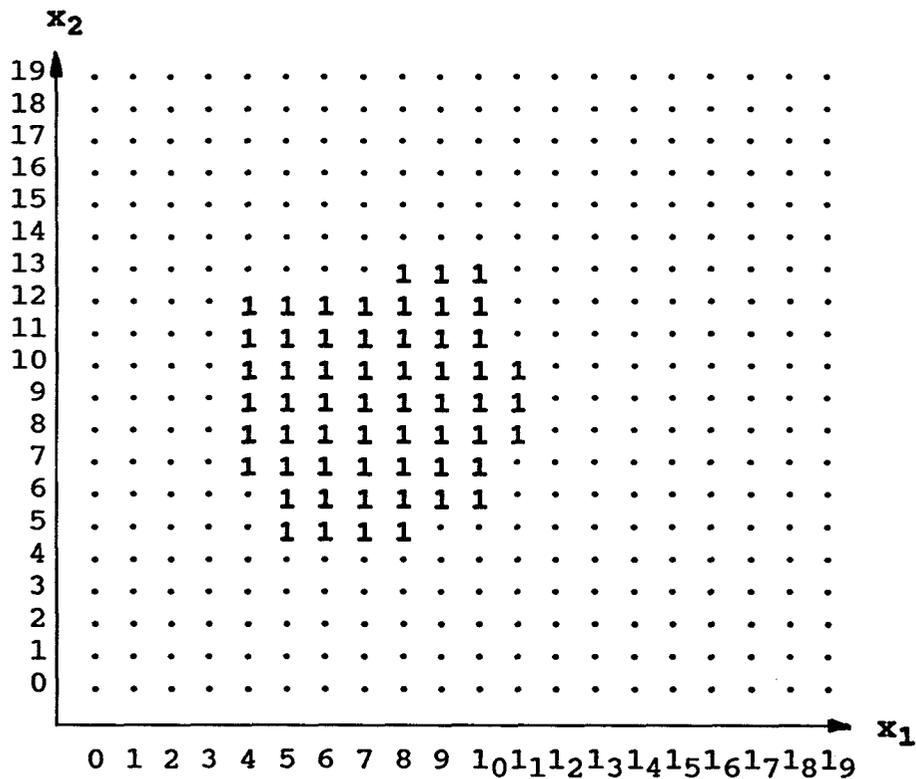
$$\text{Si } \underline{X} \subset \underline{Y}, \text{ alors } \underline{X}^S \subset \underline{Y}^S$$

De plus, ces transformations ont une propriété très intéressante qui est celle de l'idempotence, c'est-à-dire :

$$(\underline{X}_S)_S = \underline{X}_S$$

$$(\underline{X}^S)_S = \underline{X}^S$$

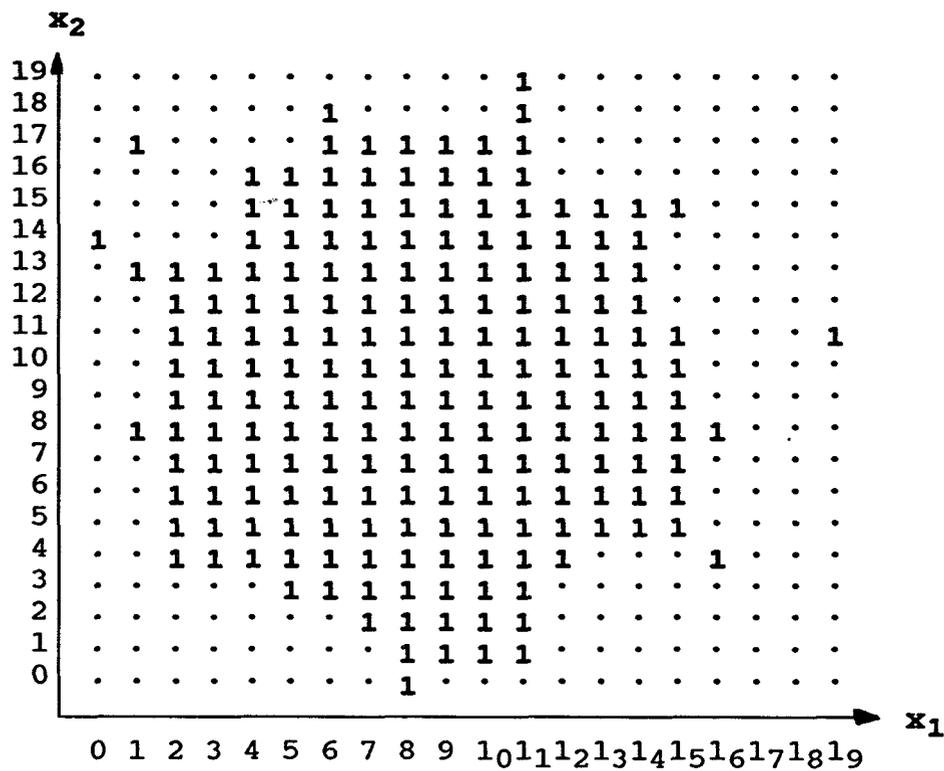
Si nous reprenons l'exemple présenté sur la figure II.5, ouvrir l'ensemble  $\underline{X}$  revient à dilater l'ensemble présenté à la figure II.5.c (cf. figure II.6.a). De même, fermer cet ensemble  $\underline{X}$  (cf. figure II.6.b) par un élément structurant 3 x 3 consiste à éroder l'ensemble présenté sur la figure II.5.d.



(a)

Figure II.6 : Ouverture et Fermeture d'un ensemble de 200 observations normales.

a - Ouverture de l'ensemble  $\underline{X}$ .



(b)

Figure II.6 : Ouverture et Fermeture d'un ensemble de 200 observations normales (suite).

b - Fermeture de l'ensemble  $X$ .

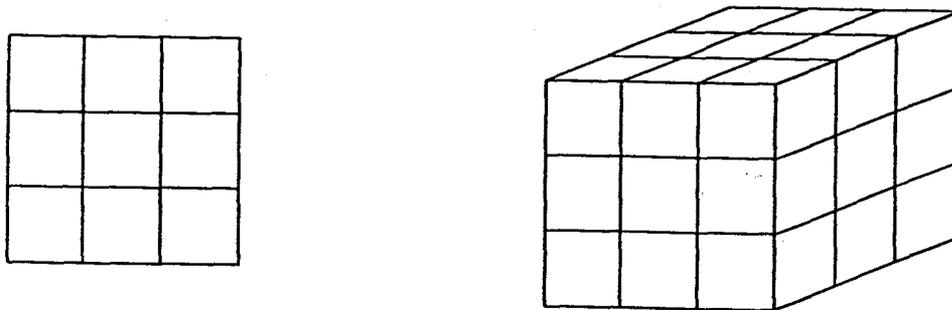
Ces résultats montrent que ces deux transformations produisent de nouveaux ensembles dont la forme générale est plus simple que celle de l'ensemble de départ. C'est pourquoi ces opérateurs morphologiques binaires de base vont être utilisés pour détecter les modes associés aux classes en présence dans l'ensemble à analyser. Nous verrons, dans la deuxième partie de cette thèse, que ce sont de très bons opérateurs de filtrage.

Néanmoins, il existe d'autres types de transformations, que nous dénommerons transformations morphologiques à éléments structurants multiples, qui permettent également de mettre en évidence les modes morphologiques et ce, en utilisant une famille d'éléments structurants plutôt qu'un élément structurant unique.

## II.4. TRANSFORMATIONS MORPHOLOGIQUES A ELEMENTS STRUCTURANTS MULTIPLES.

### II.4.1. Notion de transformation en tout ou rien.

Les opérateurs morphologiques élémentaires, comme ceux que nous venons de décrire, ont pour but de modifier l'ensemble discret  $X$  à l'aide d'un élément structurant unique  $S$  qui est décrit par son origine et par la liste des points qui le constituent. Cet élément structurant peut être défini dans une **fenêtre d'observation** multidimensionnelle. En effet, on peut translater l'origine de cette fenêtre en chaque point  $X$  de l'ensemble discret  $X$ , et analyser la structure locale de l'ensemble  $X$  dans le domaine délimité par cette fenêtre. Cette structure observée à travers la fenêtre est appelée **configuration locale** au point  $X$  et est notée  $C_S(X)$ . Dans une fenêtre d'observation multidimensionnelle de taille  $a \times a \times \dots \times a$  ( $N$  fois),  $a$  étant un entier positif impair, l'ensemble des  $a^N$  points constituant cette fenêtre est appelé le  **$a^N$ -voisinage** de  $X$  (cf. figure II.7).



*Figure II.7 : Fenêtres d'observation multidimensionnelles.*

On peut alors définir un élément structurant en faisant référence à cette fenêtre. Dans le  $a^N$ -voisinage du point  $G$ , centre de l'élément structurant, on peut considérer l'ensemble des points du réseau à valeur binaire 1 et ceux à valeur binaire 0. Les points à valeur binaire 1 sont ceux qui constituent l'élément structurant au sens que nous avons donné à ce terme au paragraphe II.3.

On peut ainsi introduire une nouvelle notation qui permet de définir l'élément structurant dans la fenêtre d'observation. Un élément structurant  $\underline{S}$  est défini par  $\underline{S} = (\underline{S}^1, \underline{S}^0)$ ,  $\underline{S}^1$  étant l'ensemble des points de  $\underline{S}$  à valeur 1 et  $\underline{S}^0$  l'ensemble des points de  $\underline{S}$  à valeur 0 (cf. figure II.8).

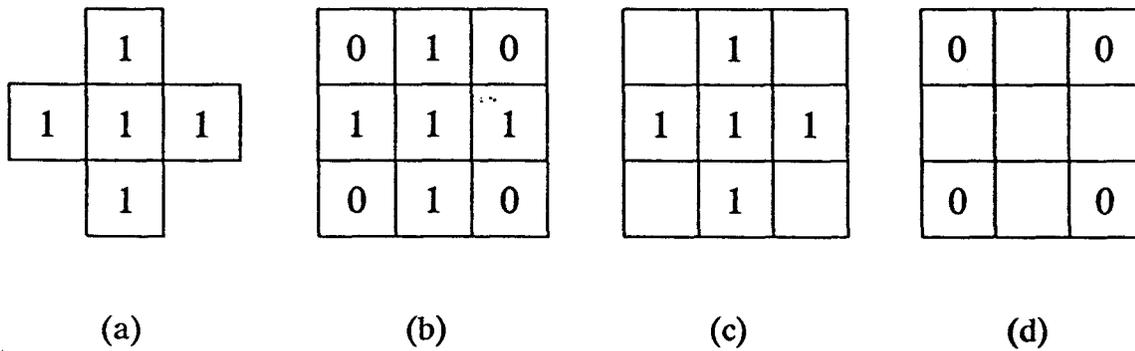


Figure II.8 : Définition d'un élément structurant.

a - Première notation.  
c - Sous-ensemble  $\underline{S}^1$ .

b - Nouvelle notation.  
d - Sous-ensemble  $\underline{S}^0$ .

J. Serra, en 1969, a défini la notion de **transformation en tout ou rien** (Hit or Miss Transformation), notée  $\otimes$ , de la façon suivante :

$$\underline{X} \otimes \underline{S} = \{ X \mid C_{\underline{S}^1}(X) \subset \underline{X}; C_{\underline{S}^0}(X) \subset \underline{X}^C, X \in \underline{X} \},$$

$C_{\underline{S}^1}(X)$  étant la configuration locale au point  $X$  dans le domaine défini par  $\underline{S}^1$  centré en  $X$ ,  $C_{\underline{S}^0}(X)$  étant la configuration locale au point  $X$  dans le domaine défini par  $\underline{S}^0$  centré en  $X$ .

Dans le cas particulier où  $\underline{S}^0 = \emptyset$ , la condition  $C_{\underline{S}^0}(X) \subset \underline{X}^C$  est toujours remplie. L'ensemble transformé  $\underline{X} \otimes \underline{S}$  se réduit alors à :

$$\underline{X} \otimes \underline{S} = \{ X \mid C_{\underline{S}^1}(X) \subset \underline{X}, X \in \underline{X} \}$$

et la transformation tout ou rien  $\otimes$  correspond à l'opération d'érosion que nous avons précédemment décrite.

Il est possible de définir plus simplement cette transformation en tout ou rien. En effet, la transformée de l'ensemble  $\underline{X}$  par l'élément structurant  $\underline{S}$  peut être

caractérisée par :

$$\underline{X} \otimes \underline{S} = \{ X \in \underline{X} \mid B'(X) = 1 \}$$

où  $B'$  est une fonction binaire, définie sur le réseau d'échantillonnage  $(\mathbb{Z}^+)^N$ , décrite par:

$$\begin{aligned} B'(X) &= 1 \text{ si } C_{\underline{S}}(X) = \underline{S} \\ B'(X) &= 0 \text{ si } C_{\underline{S}}(X) \neq \underline{S}. \end{aligned}$$

Dans ces conditions, l'ensemble transformé  $\underline{X} \otimes \underline{S}$  est composé de tous les points  $X$  tels que  $B'(X) = 1$ .

Ce type de transformation revient à reconnaître, dans l'ensemble  $\underline{X}$ , tous les points  $X$  tels que leur configuration locale  $C_{\underline{S}}(X)$  coïncide avec l'élément structurant  $\underline{S}$ .

Pour reconnaître simultanément plusieurs configurations locales particulières dans l'ensemble  $\underline{X}$ , on recherche la coïncidence entre la configuration locale de chaque point  $X$  avec l'un des éléments d'un ensemble d'éléments structurants. Cet ensemble, que nous appelons **famille structurante**, caractérise les différentes structures locales recherchées autour du point  $X$ .

#### II.4.2. Notion de transformation morphologique à éléments structurants multiples.

Soit  $S$  une famille de  $p$  éléments structurants correspondant aux configurations locales recherchées dans l'ensemble  $\underline{X}$  telle que :

$$S = \{ \underline{S}_1, \underline{S}_2, \dots, \underline{S}_j, \dots, \underline{S}_p \}.$$

La transformation en tout ou rien de l'ensemble discret  $\underline{X}$  par la famille structurante  $S$  est définie par :

$$\underline{X} \otimes S = \{ X \in \underline{X} \mid \exists j = 1, \dots, p, C_{\underline{S}_j}(X) \subset \underline{X}, C_{\underline{S}_j}(X) \subset \underline{X}^c \}.$$

Cette transformation en tout ou rien par éléments structurants multiples peut être également définie par :

$$\underline{X} \otimes S = \{ X \in \underline{X} \mid B'(X) = 1 \}$$

où  $B'$  est définie par :

$$B'(X) = 1 \text{ si } C_S(X) \in S$$

$$B'(X) = 0 \text{ si } C_S(X) \notin S.$$

A. Golay a établi une nomenclature des familles structurantes les plus usuelles dans le cas du réseau hexagonal bidimensionnel, appelée "alphabet de Golay" (voir Table T2.1).

L	M	D	C	E
1 1	1 *	0 *	1 *	* *
* 1 *	1 1 0	0 1 1	1 0 *	0 1 0
0 0	1 *	0 *	1 *	0 0

I	F	F'	H	R
0 0	1 0	1 1	1 1	* *
0 1 0	0 1 1	* 1 *	1 1 1	* 1 0
0 0	1 0	* *	1 1	* *

1 : Points de  $\underline{X}$

0 : Points de  $\underline{X}^C$

\* : Points indifférent.

*Table T2.1 : Alphabet de Golay.*

Deux transformations morphologiques importantes peuvent être définies à partir de cette notion de transformation en tout ou rien. Il s'agit de l'amincissement et de l'épaississement.

L'ensemble résultant d'un amincissement par la famille structurante  $S$  est défini par :

$$\underline{X} \circ S = \underline{X} / (\underline{X} \otimes S)$$

où / désigne la différence ensembliste, et  $S$  une famille d'éléments structurants.

L'ensemble résultant d'un épaississement par la famille structurante  $S$  est défini par :

$$\underline{X} \circ S = \underline{X} \cup (\underline{X} \otimes S).$$

L'amincissement par la famille  $S$  peut être considéré comme l'élimination de tous les points de l'ensemble  $\underline{X}$  dont la configuration locale coïncide avec l'un des éléments structurants de la famille structurante  $S$ .

On peut noter que l'amincissement et l'épaississement sont deux opérations duales vis à vis de la complémentation. Amincir un ensemble  $\underline{X}$  par une famille structurante  $S$  revient à épaissir son complémentaire par la famille structurante complémentaire :

$$(\underline{X} \circ S)^C = \underline{X}^C \circ S^C.$$

De plus, quelle que soit la famille structurante  $S$  définie, la relation de double inclusion ci-dessous est vérifiée :

$$\underline{X} \circ S \subset \underline{X} \subset \underline{X} \circ S^C.$$

Les transformations morphologiques en tout ou rien, l'amincissement et l'épaississement dépendent toutes d'une famille structurante  $S$ . Les propriétés vérifiées par ces transformations dépendent également des propriétés vérifiées par la famille structurante  $S$  et ne peuvent donc pas être présentées à ce niveau.

### II.4.3. Exemple d'application.

Voyons, par l'intermédiaire d'un exemple, comment se construit une famille structurante. Supposons que nous cherchions à éliminer les points isolés ainsi que les points extrêmes des ramifications d'un ensemble discret bidimensionnel  $\underline{X}$ . Dans une fenêtre d'observation de taille 3 x 3, un point isolé est un point de  $\underline{X}$  entouré uniquement de points appartenant à  $\underline{X}^C$ . L'élément structurant  $\underline{I}$  permettant de détecter un tel point est donc défini par :

0	0	0
0	<u>1</u>	0
0	0	0

dans lequel l'origine est repérée par une valeur en caractère gras et souligné.

D'autre part, un point extrémité d'une ramification est un point ne possédant, dans une fenêtre d'observation de taille 3 x 3, qu'un seul voisin appartenant à  $\underline{X}$ . Ceci nécessite donc, afin de définir une transformation isotropique, l'utilisation de huit éléments structurants définis par l'élément :

0	1	0
0	<u>1</u>	0
0	0	0

et les sept éléments structurants obtenus par rotations de  $\pi/4$ ,  $2\pi/4$ , ...,  $7\pi/4$  de celui-ci.

En repérant par une étoile "\*" un point d'une configuration locale dont la valeur binaire peut être indifféremment 1 ou 0, la famille structurante  $EI$  permettant de détecter ces points extrêmes ainsi que les points isolés est définie par :

$$EI = \{ \underline{EI}_1, \underline{EI}_2, \dots, \underline{EI}_8 \}.$$

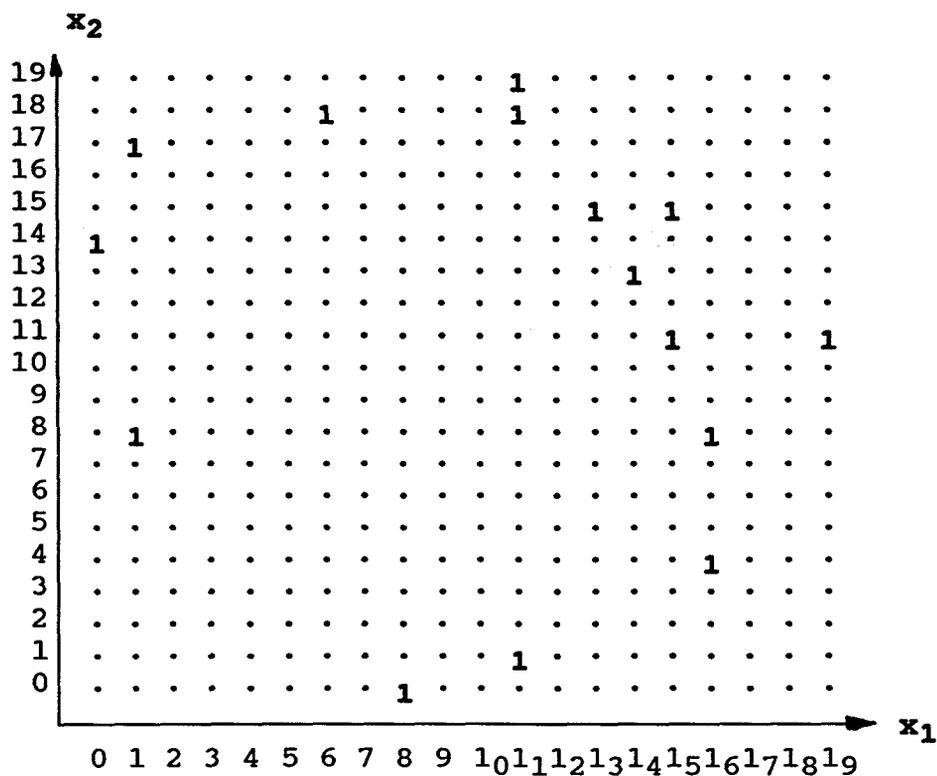
$\underline{EI}_1$  est de la forme :

0	*	0
0	<u>1</u>	0
0	0	0

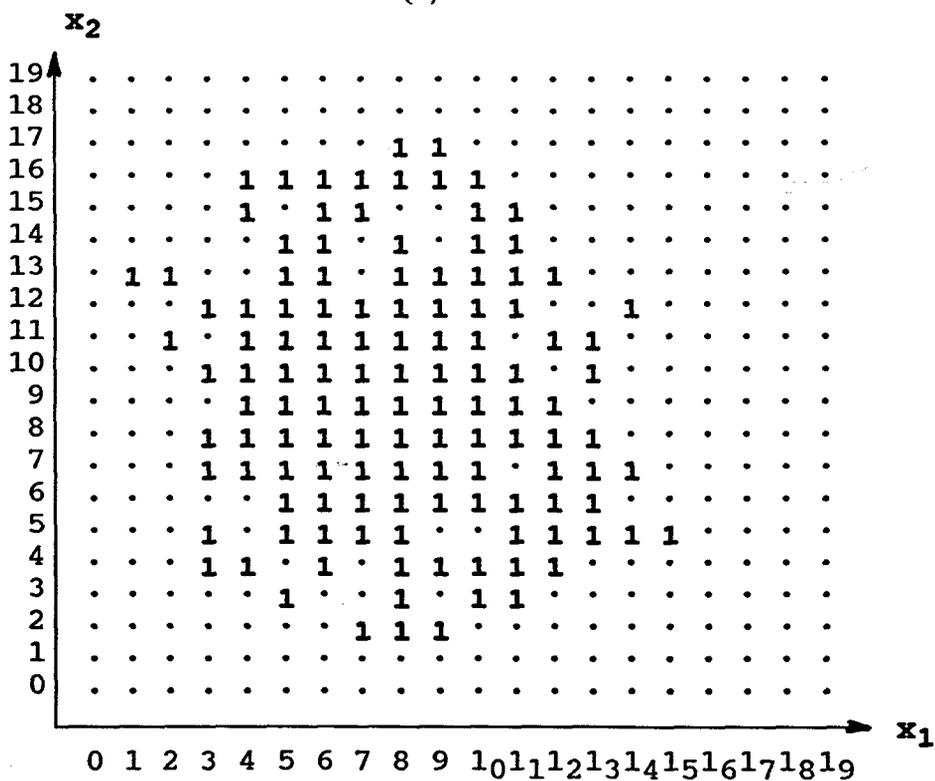
et les éléments  $\underline{EI}_2$ , ...,  $\underline{EI}_8$  sont obtenus par les 7 rotations successives de  $\pi/4$  de  $\underline{EI}_1$ .

La transformation morphologique permettant de détecter de tels points est la transformation en tout ou rien ( $\underline{X} \otimes EI$ ) tandis que leur élimination est obtenue par l'amincissement de l'ensemble  $\underline{X}$  par la famille structurante  $EI$  noté  $\underline{X} \circ EI$ .

Nous avons appliqué ces transformations à l'ensemble discret présenté sur la figure II.5.b. La figure II.9.a présente l'ensemble  $\underline{X} \otimes EI$ , définissant l'ensemble des points extrêmes détectés, tandis que la figure II.9.b présente l'ensemble aminci par la famille structurante  $EI$ .



(a)



(b)

Figure II.9 : Résultats des transformations morphologiques à éléments structurants multiples appliquées à l'ensemble discret de la figure II.5.b.

a - Ensemble  $\underline{X} \otimes EI$  constitué des points isolés devant être éliminés.

b - Ensemble  $\underline{X} \circ EI$ .

## II.5. CONCLUSION.

Dans ce chapitre, nous avons montré comment un ensemble d'observations multidimensionnelles pouvait être représenté sous forme d'un ensemble discret composé de points à coordonnées entières et positives de l'espace multidimensionnel, auxquels sont associées des valeurs binaires.

Nous avons ensuite défini deux grandes catégories de transformations morphologiques binaires, à savoir les transformations morphologiques binaires élémentaires et les transformations morphologiques binaires à éléments structurants multiples.

Nous avons vu que nous pouvions considérer l'ensemble discret comme une version simplifiée de l'ensemble des observations à classer, version dont la structure spatiale reflète les différences de concentration de ces observations dans l'espace de représentation des données. En effet, dans cet ensemble discret, les regroupements d'observations donnent naissance à des sous-ensembles connexes de points à valeurs 1, de sorte qu'une classe peut être identifiée à chacune des composantes connexes de cet ensemble discret. Nous avons également remarqué que ces sous-ensembles possèdent à leur périphérie des irrégularités ne pouvant que nuire à la mise en évidence de ces composantes.

La partie consacrée à l'extraction des modes que nous allons aborder maintenant va avoir pour but de régulariser la forme extérieure de ces composantes, permettant ainsi la mise en évidence des modes morphologiques. Ceux-ci seront détectés sous la forme de domaines modaux, définis soit par la liste explicite de tous les points qui les constituent, soit par leurs frontières extérieures.

**- DEUXIEME PARTIE -**

**DETECTION DES MODES  
PAR  
FILTRAGE MORPHOLOGIQUE**

**III - EXTRACTION DES MODES  
PAR TRANSFORMATIONS MORPHOLOGIQUES  
ELEMENTAIRES.**

## EXTRACTION DES MODES PAR TRANSFORMATIONS MORPHOLOGIQUES ELEMENTAIRES.

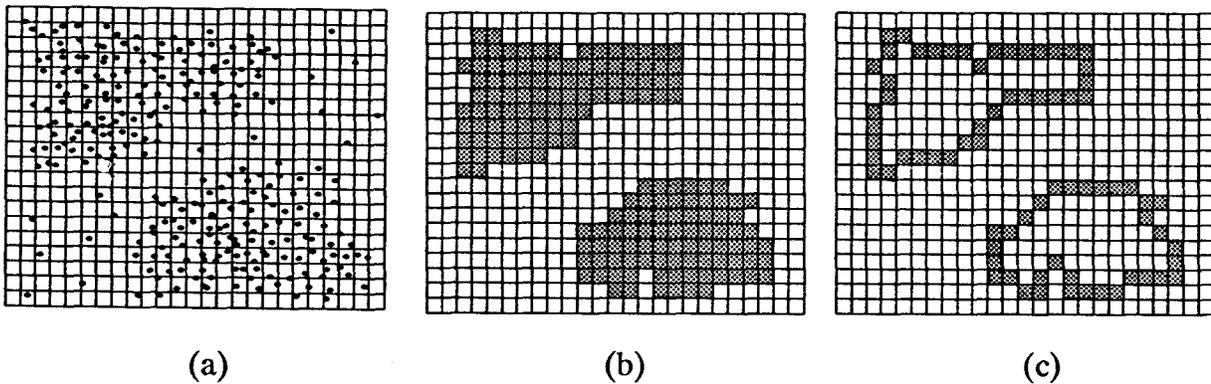
### III.1. INTRODUCTION.

Comme nous l'avons précisé dans le chapitre I, les données à analyser se présentent sous la forme d'un ensemble d'observations à N dimensions pouvant être facilement transformé en un ensemble discret  $X$  de points à valeurs binaires (voir chapitre II, paragraphe 2). A chaque classe de l'ensemble d'observations, on peut ainsi associer un sous-ensemble connexe de points à valeur binaire 1 de l'ensemble discret.

La deuxième partie de ce mémoire consiste à présenter différentes méthodes morphologiques permettant de régulariser la forme extérieure de ces sous-ensembles connexes de manière à les séparer facilement en composantes connexes bien individualisées. Les deux techniques de régularisation de ces composantes connexes présentées dans ce chapitre, procèdent par combinaisons des transformations morphologiques de base que sont l'érosion et la dilatation binaire. La première technique a pour but de caractériser les modes par des domaines modaux, c'est-à-dire par des sous-ensembles de points de  $X$  indiquant la présence de régions modales à fortes concentrations locales d'observations. Cette méthode de détection des modes, qui débouche sur une description point à point des domaines modaux, est suivie d'une autre technique visant à déterminer uniquement les frontières extérieures de ces domaines (cf. figure III.1).

La première méthode sera désignée sous le terme "Extraction des domaines modaux par transformation morphologique à élément structurant unique", et la seconde

par "Extraction des frontières modales par transformation morphologique à élément structurant unique".



*Figure III.1. : Extraction des modes morphologiques.*

- a - Ensemble d'observations à classer.*
- b - Les domaines modaux.*
- c - Les frontières modales.*

Dans les deux cas, l'analyse morphologique de l'ensemble discret  $X$  fournit directement un ensemble discret  $M$  de points définissant les modes morphologiques associés aux classes mises en évidence.

## III.2. EXTRACTION DES DOMAINES MODAUX PAR TRANSFORMATIONS MORPHOLOGIQUES A ELEMENT STRUCTURANT UNIQUE.

### III.2.1. Principe de détection des domaines modaux.

Dans le chapitre précédent, nous avons présenté les transformations morphologiques binaires de base que sont l'érosion, la dilatation ainsi que l'ouverture et la fermeture. Nous allons maintenant utiliser ces transformations pour mettre en évidence les domaines modaux associés aux différentes classes.

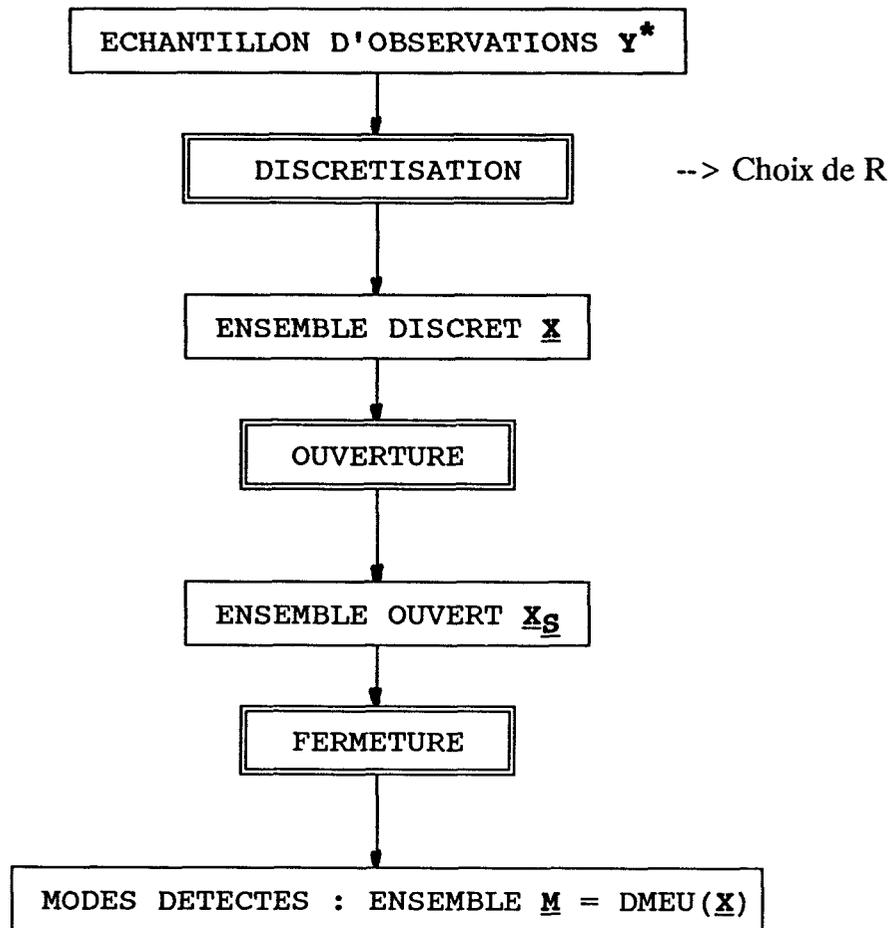
Comme nous l'avons déjà précisé, les domaines modaux sont constitués d'ensembles de points de  $\underline{X}$  situés au coeur des classes dans les régions modales à forte concentration locale d'observations.

Le principe de cette méthode est de sélectionner, dans l'ensemble discret  $\underline{X}$ , les points situés dans les régions modales afin de les retenir comme points des domaines modaux. En d'autres termes, il s'agit d'éliminer tous les points de  $\underline{X}$  situés à la périphérie des sous-ensembles connexes de  $\underline{X}$ , afin d'obtenir des composantes bien individualisées.

L'érosion transforme l'ensemble  $\underline{X}$  en un ensemble érodé ne contenant que les points dont la configuration locale coïncide avec l'élément structurant  $\underline{S}$  utilisé. En choisissant un élément structurant hypercubique de côté 3 unités, chacun des points de cet élément étant un point à valeur binaire 1, l'ensemble érodé contient alors tous les points de  $\underline{X}$  dont les  $3^N - 1$  voisins appartiennent à  $\underline{X}$ . De tels points caractérisent une forte concentration locale d'observations et doivent être retenus comme points constitutifs des domaines modaux. Néanmoins, cette opération d'érosion ne conserve que l'origine de l'élément structurant. Une opération de dilatation par le même élément structurant hypercubique permet de retrouver la forme initiale du domaine après élimination des détails périphériques.

La transformation à réaliser pour mettre en évidence les domaines modaux est donc une ouverture de l'ensemble discret  $\underline{X}$ . En reprenant une stratégie devenue classique en analyse morphologique des images, une fermeture de l'ensemble ouvert augmente sensiblement la robustesse de cette méthode de détermination des domaines modaux par rapport au procédé de discrétisation spatiale de l'espace de représentation des données. L'ensemble discret résultant de cette détection des Domaines Modaux par transformation morphologique à Élément structurant Unique sera noté  $DMEU(\underline{X})$ .

Le principe de détection des modes morphologiques que nous venons de proposer peut être résumé par le schéma synoptique suivant :



### III.2.2. Algorithme associé.

Les performances de cette procédure dépendent à la fois du choix de l'élément structurant  $\underline{\mathbf{S}}$  et du pas de discrétisation R.

Puisque nous ne possédons aucune connaissance a priori sur la structure de l'échantillon et qu'aucun critère analytique ne permet de choisir l'élément structurant  $\underline{\mathbf{S}}$ , nous nous sommes imposés la taille de cet élément structurant, laissant à l'opérateur le soin d'ajuster le pas de discrétisation R. Pour être plus précis, nous avons défini :

- un élément structurant (3 x 3) dans l'espace bidimensionnel,
- un élément structurant cubique (3 x 3 x 3) dans l'espace tridimensionnel,
- un élément structurant hypercubique (3 x 3 x 3 x ... x 3) dans les espaces de dimension supérieure.

L'ajustement du pas de discrétisation  $R$  est basé sur le concept de stabilité du nombre de modes détectés [EIG - 74]. Lorsque l'ensemble à analyser est composé de plusieurs classes distinctes, les modes associés doivent en effet apparaître pour un grand intervalle de valeurs de  $R$ .

Par conséquent, l'ajustement du pas  $R$  consiste à appliquer l'algorithme de détection des domaines modaux pour toutes les valeurs entières positives de  $R$  jusqu'à ce que tous les domaines modaux aient disparu. Ceci permet de déterminer le plus grand intervalle de valeurs de  $R$  pour lesquelles le nombre de modes détectés reste stable et différent de 1 (résultat trivial). Selon une technique heuristique éprouvée [POS - 81] [TOU - 88], le pas de discrétisation est finalement ajusté au milieu de cet intervalle.

### III.2.3. Identification des modes détectés.

L'identification des modes morphologiques consiste à extraire les différentes composantes connexes contenues dans l'ensemble discret  $\underline{M}$ .

Ces composantes connexes sont définies de la manière suivante :

- Tous les points d'une composante connexe  $\underline{C}$  appartiennent à  $\underline{M}$ .
- Entre tout couple de points  $A$  et  $B$  appartenant à  $\underline{C}$ , il existe un chaînage de points appartenant à  $\underline{C}$ .
- Tout ensemble de points de  $\underline{M}$ , contenant  $\underline{C}$ , ne satisfait pas aux deux premières conditions.

Les deux premières conditions assurent la connexité des points d'une composante  $\underline{C}$ , la dernière condition permet de s'assurer qu'on a trouvé le plus grand ensemble de points connexes.

Rappelons qu'il existe un chaînage entre deux points  $A$  et  $B$  lorsqu'on peut trouver une séquence de points  $A = X_0, X_1, X_2, \dots, X_L = B$  telle que  $X_i$  soit un

voisin de  $X_{l-1}$  pour tout  $l = 1, 2, \dots, L$  au sens du voisinage choisi. Nous avons opté pour un  $3^N$ -voisinage constitué des  $3^N - 1$  voisins situés dans l'hypercube centré sur  $X_l$  et de côté égal à 3 unités.

### III.2.4. Exemple d'application.

Cette deuxième partie du mémoire présente différentes méthodes de détection des modes morphologiques associés aux classes en présence dans un ensemble d'observations donné. Pour comparer ces méthodes entre-elles, il nous a semblé intéressant de les confronter en utilisant un même exemple d'évaluation. Nous avons choisi un échantillon de 700 observations bivariées générées artificiellement et réparties en deux classes distinctes de 350 observations chacune.

Les deux attributs  $y^1, y^2$  décrivant ces observations, sont définis par :

$$y^1 = A_1 \cos \theta + B_1$$

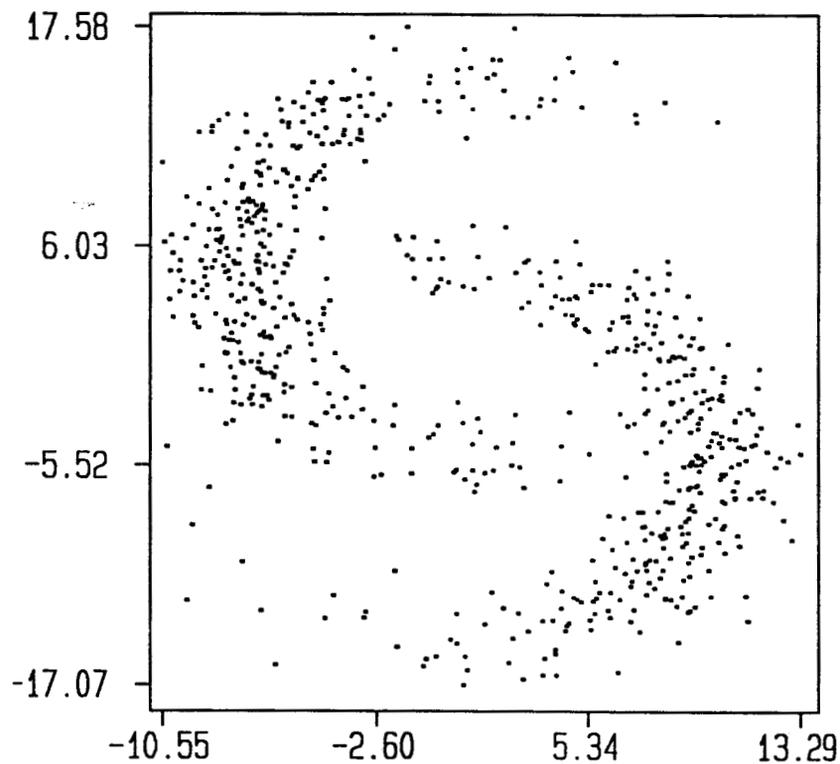
$$y^2 = A_2 \sin \theta + B_2$$

où  $\theta$  est une variable aléatoire normale de moyenne  $m$  et de variance  $s$ , et où  $B_1$  et  $B_2$  sont des variables aléatoires normales de moyennes  $\mu$  et de variance  $\sigma$  [FUK - 84]. Les valeurs de ces paramètres sont détaillées dans la table T3.1.

	$\theta$	$B_1$	$B_2$	$A_1$	$A_2$
Population 1	$m = 0^\circ$ $s = 50^\circ$	$\mu_1 = 0$ $\sigma_1 = 2$	$\mu_2 = -5$ $\sigma_2 = 2$	$A_1 = 10$	$A_2 = 10$
Population 2	$m = 180^\circ$ $s = 50^\circ$	$\mu_1 = 2$ $\sigma_1 = 2$	$\mu_2 = 5$ $\sigma_2 = 2$	$A_1 = 10$	$A_2 = 10$

*Table T3.1 : Paramètres statistiques des 700 données générées artificiellement.*

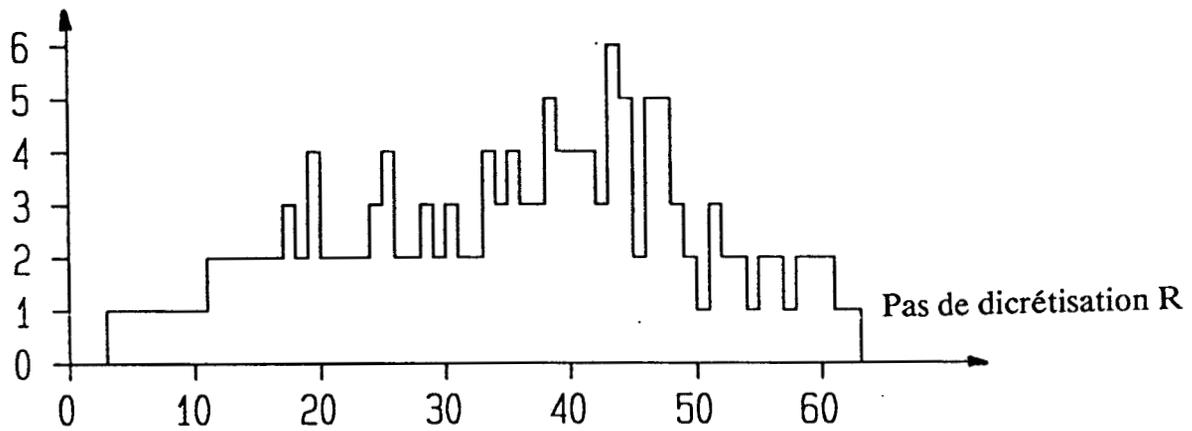
L'ensemble des données dont les paramètres sont explicités dans la table T3.1 est présenté par la figure III.2.



*Figure III.2 : Représentation de l'ensemble de 700 données générées artificiellement conformément aux paramètres de la table T3.1.*

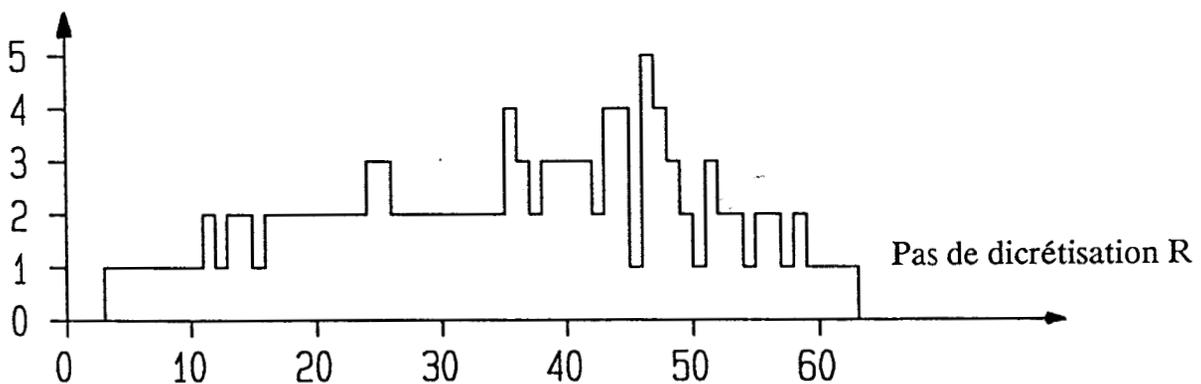
La procédure de détection des modes consiste à réaliser successivement l'ouverture puis la fermeture de  $X$  afin d'extraire les sous-ensembles connexes correspondant aux modes morphologiques. Afin d'analyser les performances de cette méthode, la procédure de détection a été effectuée pour différentes valeurs du pas  $R$  de discrétisation de l'espace de représentation des données. La figure III.3.a montre l'effet de ce paramètre  $R$  sur le nombre de composantes connexes détectées lorsque l'on ne réalise qu'une ouverture sur l'ensemble discret, tandis que la figure III.3.b montre l'effet de ce même paramètre sur le nombre de modes détectés en réalisant une ouverture suivie d'une fermeture.

nombre de modes détectés.



(a)

nombre de modes détectés.



(b)

**Figure III.3 :** Effet du paramètre  $R$  sur le nombre de modes détectés.

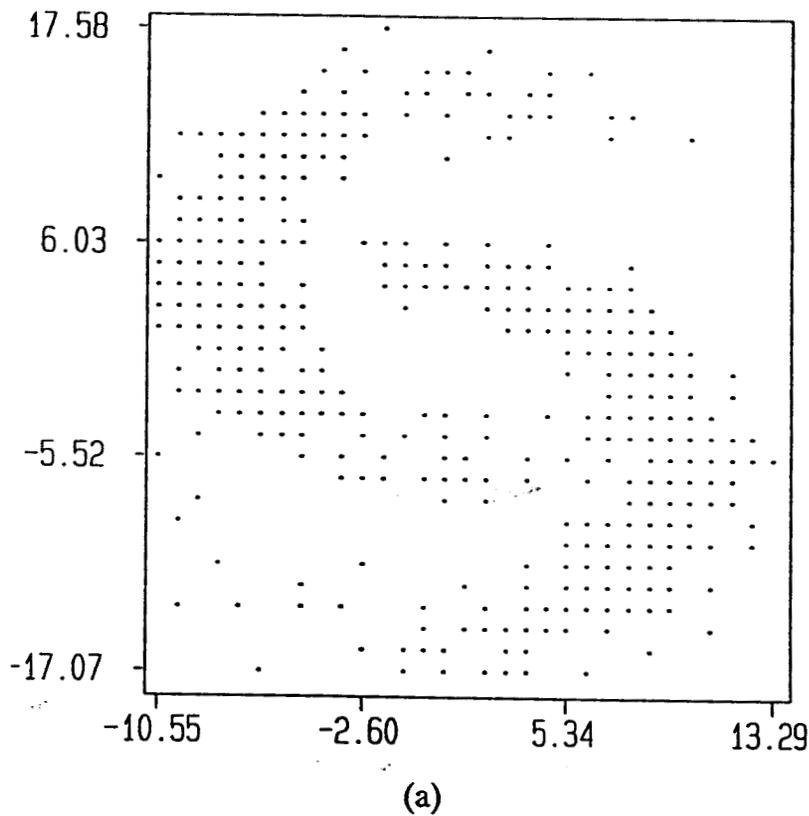
*a - En réalisant une simple ouverture sur  $X$ .*

*b - En réalisant une ouverture suivie d'une fermeture.*

Lorsque l'on réalise une simple ouverture, le plus grand intervalle de valeurs de  $R$  pour lesquelles les deux modes sont détectés est  $[11, 16]$  tandis qu'avec une opération supplémentaire de fermeture, l'intervalle de stabilité est  $[26, 34]$ . Ceci montre bien comment la transformation de fermeture permet d'élargir la plage de stabilité.

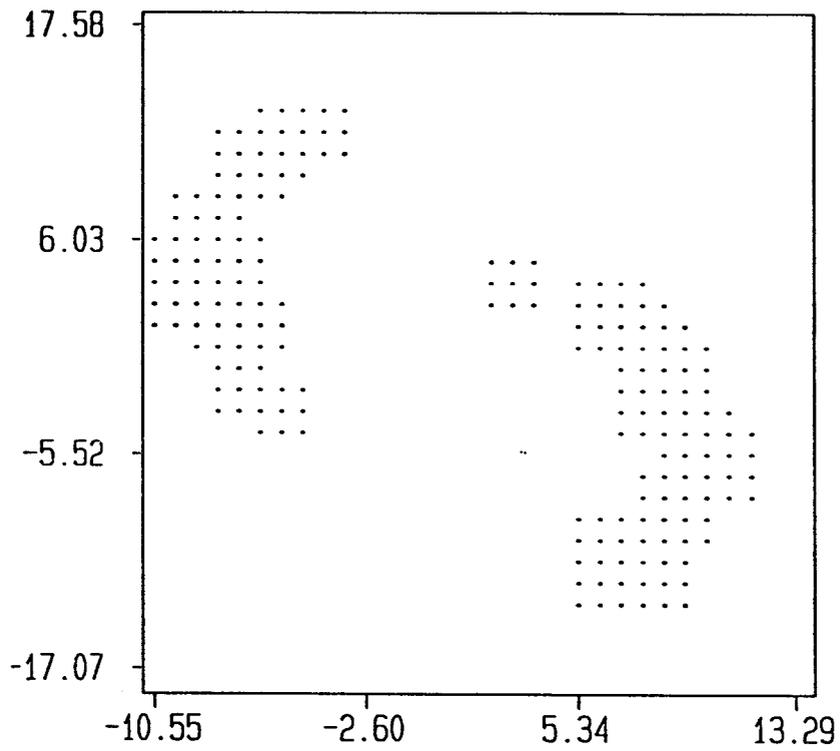
La figure III.4.a présente l'ensemble discret obtenu à partir de l'échantillon de

la figure III.2, avec un pas  $R = 30$  qui correspond au milieu du plus grand intervalle de valeurs de  $R$  pour lesquelles le nombre de modes détectés reste constant. La figure III.4.b montre la présence de trois modes morphologiques après une simple opération d'ouverture, alors que la figure III.4.c illustre le rôle de la fermeture qui, succédant à l'ouverture, permet de retrouver les deux modes auxquels on s'attendait.

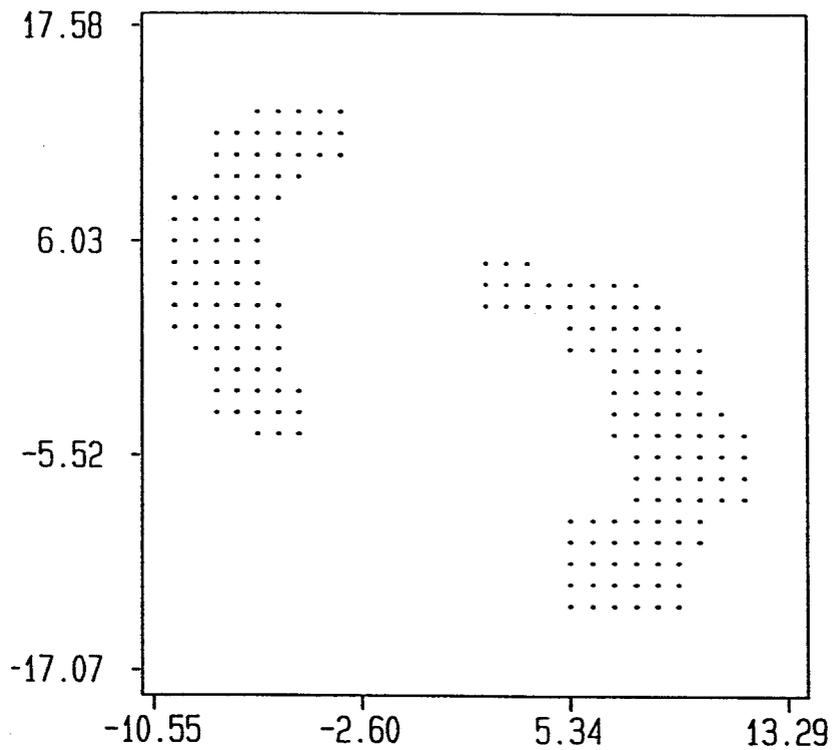


*Figure III.4 : Détection des domaines modaux par transformations morphologiques à éléments structurants uniques.*

*a - Ensemble discret de points à valeurs binaires ( $R = 30$ ).*



(b)



(c)

*Figure III.4 : Détection des domaines modaux par transformations morphologiques à éléments structurants uniques (suite).*

*b - Les trois modes détectés après une simple ouverture.*

*c - Les 2 domaines modaux obtenus après une ouverture suivie d'une fermeture.*

### **III.2.5. Conclusion.**

Nous venons de présenter une méthode de détection des modes morphologiques en utilisant successivement les transformations élémentaires d'ouverture et de fermeture, l'ouverture consistant en une érosion suivie d'une dilatation, la fermeture étant l'application d'une dilatation suivie d'une érosion.

Le fait de commencer cette détection par une opération d'érosion permet d'écartier, dès le début de la procédure, les points n'étant pas situés dans les régions à forte concentration locale d'observations. Cette première érosion contribue également à l'individualisation des modes morphologiques en creusant les vallées qui séparent ces différentes composantes.

## **III.3. EXTRACTION DES FRONTIERES MODALES PAR TRANSFORMATIONS MORPHOLOGIQUES A ELEMENT STRUCTURANT UNIQUE.**

### **III.3.1. Introduction.**

Les transformations morphologiques élémentaires décrites dans le paragraphe II.3 peuvent également être utilisées pour déterminer les frontières des domaines modaux associées aux classes en présence dans l'échantillon d'observations à analyser.

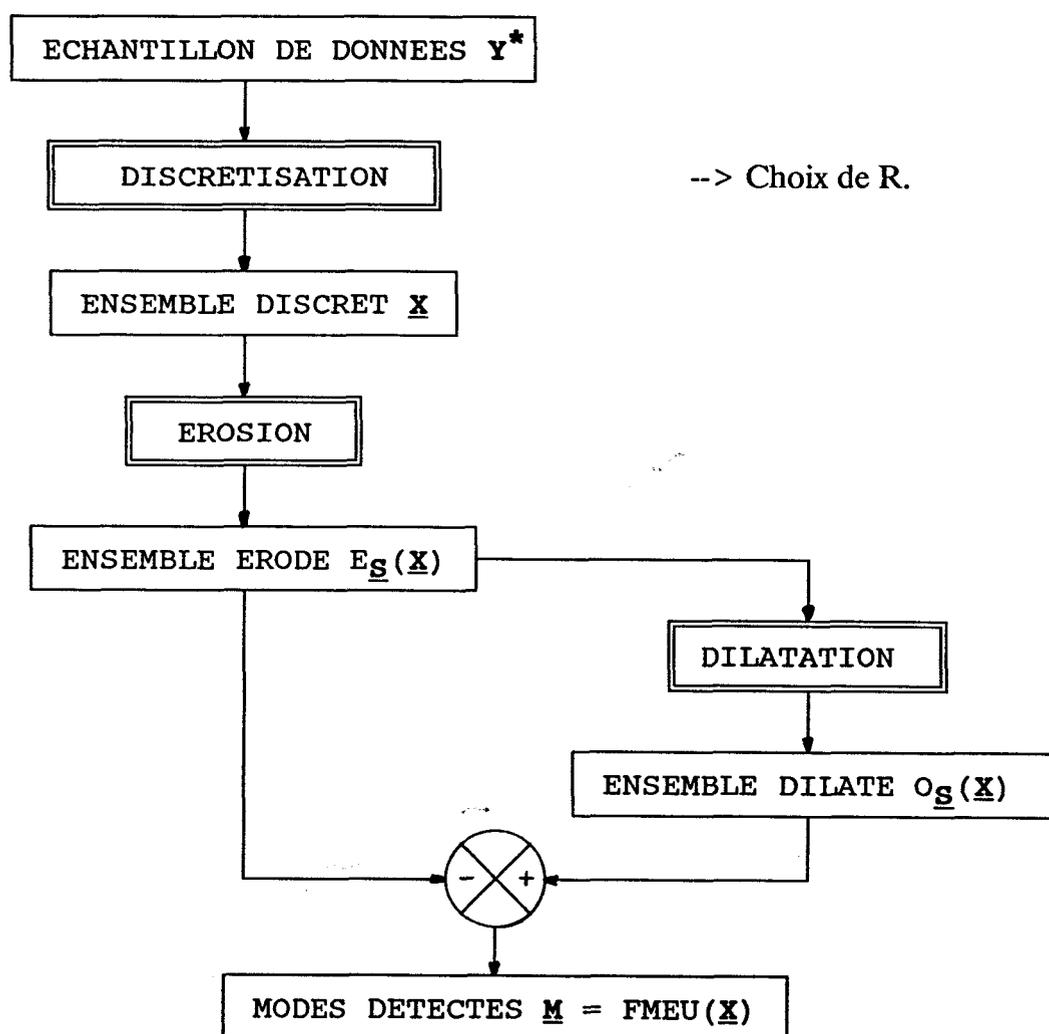
La méthode que nous proposons maintenant a pour but de déterminer les frontières modales par combinaison des transformations de dilatation et d'érosion de l'ensemble discret  $\underline{X}$ .

### **III.3.2. Principe de détermination des frontières modales.**

Après discrétisation de l'ensemble des observations à classer, l'ensemble discret  $\underline{X}$  est érodé avec un élément structurant  $\underline{S}$ . L'ensemble érodé ainsi obtenu, noté  $E_{\underline{S}}(\underline{X})$ , est alors dilaté par le même élément structurant  $\underline{S}$ , définissant ainsi une ouverture de

l'ensemble discret  $\underline{X}$ . L'ensemble résultant noté  $O_{\underline{g}}(\underline{X})$  est ensuite comparé à l'ensemble érodé  $E_{\underline{g}}(\underline{X})$ . Les frontières modales sont déterminées en réalisant une simple différence entre les ensembles  $O_{\underline{g}}(\underline{X})$  et  $E_{\underline{g}}(\underline{X})$ , donnant ainsi l'ensemble FMEU( $\underline{X}$ ) constitué des points frontières des domaines modaux.

Cette méthode de détection des Frontières Modales par transformation morphologique à Elément structurant Unique (FMEU) peut être résumée par le schéma synoptique suivant :



Comme pour la méthode précédente, l'érosion a pour but d'écartier tous les points se trouvant à la périphérie des composantes connexes qu'il s'agit d'individualiser. L'ensemble érodé  $E_{\underline{g}}(\underline{X})$  peut être considéré comme une version "amincie" de l'ensemble discret de départ. La dilatation permet d'ajouter des points à la périphérie

des composantes de l'ensemble érodé qui peuvent être isolées par la différence entre ces deux ensembles.

L'ensemble  $FMEU(\underline{X})$ , composé des points des frontières modales, est défini par :

$$FMEU(\underline{X}) = O_{\underline{S}}(\underline{X}) / E_{\underline{S}}(\underline{X}),$$

où / désigne l'opération de différence de deux ensembles.

En conservant les notations initialement définies au chapitre II, cette transformation est définie par :

$$FMEU(\underline{X}) = (\underline{X})_{\underline{S}} / (\underline{X} \ominus \underline{S}).$$

### III.3.3. Algorithme associé.

Les performances de cette procédure dépendent, comme pour la méthode précédente, à la fois du choix de l'élément structurant  $\underline{S}$  et du pas de discrétisation  $R$ .

Nous nous sommes imposés le même élément structurant  $\underline{S}$  que précédemment, soit :

- un élément structurant (3 x 3) dans l'espace bidimensionnel,
- un élément structurant cubique (3 x 3 x 3) dans l'espace tridimensionnel,
- un élément structurant hypercubique (3 x 3 x 3 x ... x 3) dans les espaces de dimension supérieure.

Le pas de discrétisation  $R$  est, ici encore, le seul paramètre de cette procédure de détection des modes à définir par l'opérateur. Son ajustement est, comme pour la méthode de détermination des domaines modaux mettant en oeuvre les opérateurs d'ouverture et de fermeture, basé sur le concept de la stabilité du nombre de modes détectés.

### III.3.4. Amélioration de la méthode de détermination des frontières modales.

Après l'opération d'érosion, il peut subsister certains points isolés dans l'ensemble érodé  $E_{\underline{S}}(\underline{X})$ . A titre d'exemple, nous présentons sur la figure III.5.b le résultat de l'érosion de l'ensemble discret de la figure III.5.a, obtenu à partir de l'exemple de la figure III.2. L'érosion fait apparaître un point isolé, encadré sur la figure III.5.b, qui va nuire à la détection correcte des deux modes. Nous allons donc, avant de réaliser l'opération de dilatation, chercher à éliminer de tels points isolés.

Notons que les  $3^N - 1$  points voisins d'un point isolé sont des points du réseau d'échantillonnage appartenant à  $\underline{X}^C$ . La transformation morphologique permettant d'éliminer ces points consiste à rechercher la coïncidence entre la configuration locale en chaque point  $X$  de l'ensemble érodé  $E_{\underline{S}}(\underline{X})$  et un élément structurant spécifique  $\underline{I}$  caractéristique de tels points.  $\underline{I}$  est défini sur une fenêtre hypercubique de taille  $3 \times 3 \times 3 \times \dots \times 3$  ( $N$  fois) dont l'origine, située au centre géométrique de  $\underline{I}$ , a pour valeur 1. Tous les autres points de cette fenêtre hypercubique ont pour valeur 0. Il s'agit donc de l'amincissement de l'ensemble  $E_{\underline{S}}(\underline{X})$  par l'élément structurant  $\underline{I}$ , noté  $E_{\underline{S}}(\underline{X}) \circ \underline{I}$ .

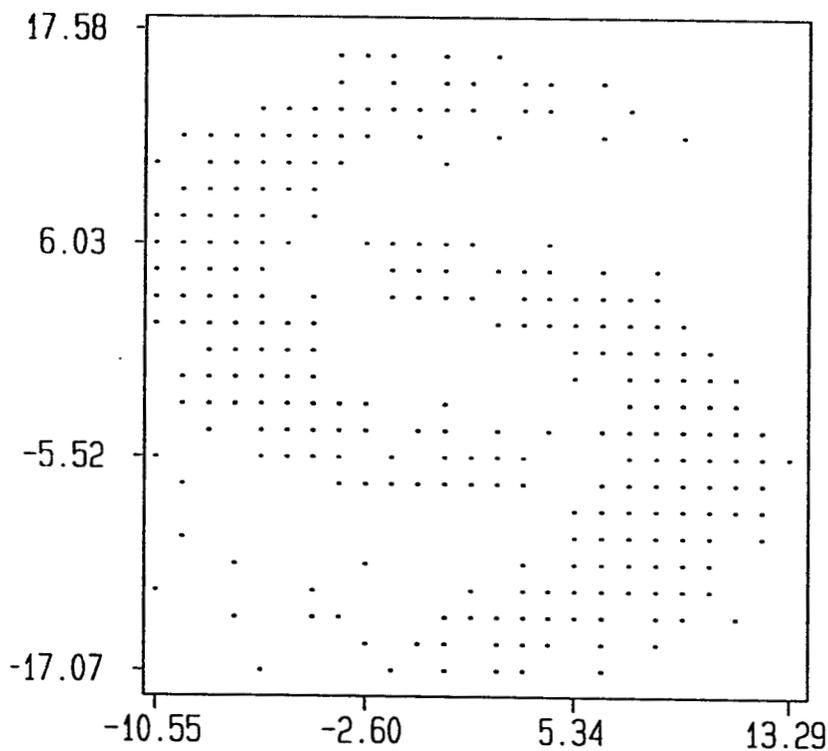
Lorsque tous les points isolés de l'ensemble érodé sont éliminés, la dilatation peut avoir lieu, ainsi que la comparaison aboutissant à la détection des modes morphologiques, ces modes étant décrits par leurs frontières modales toujours fermées.

La transformation morphologique qui permet de déterminer l'ensemble des points des frontières modales  $FMEU(\underline{X})$  est définie par :

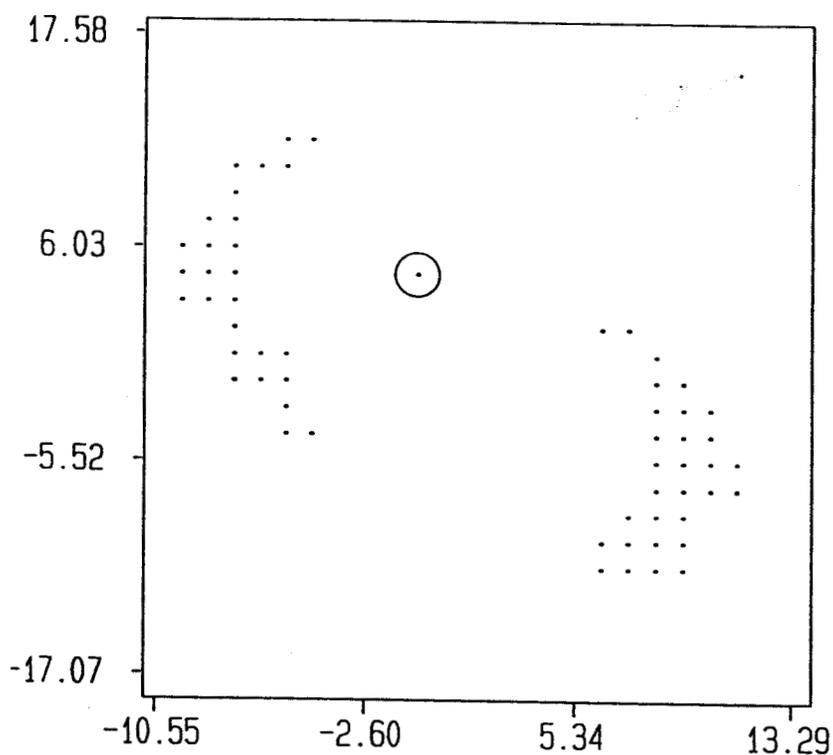
$$FMEU(\underline{X}) = D_{\underline{S}}(E_{\underline{S}}(\underline{X}) \circ \underline{I}) / (E_{\underline{S}}(\underline{X}) \circ \underline{I}).$$

En reprenant les notations définies au chapitre II, cette transformation peut être décrite par la relation :

$$FMEU(\underline{X}) = (((\underline{X} \ominus \underline{S}) \circ \underline{I}) \oplus \underline{S}) / ((\underline{X} \ominus \underline{S}) \circ \underline{I})$$



(a)



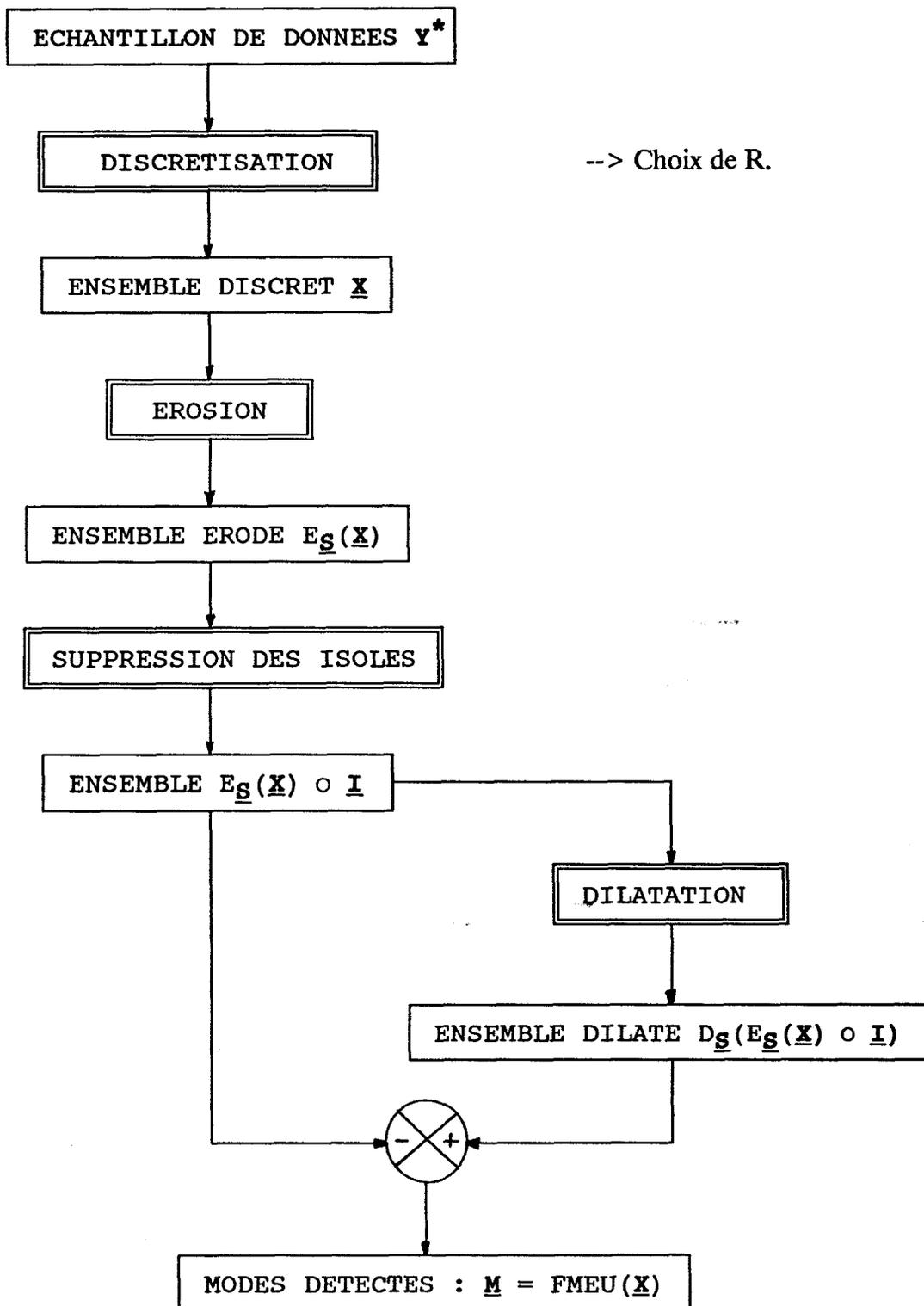
(b)

*Figure III.5 : Mise en évidence des points isolés après l'opération d'érosion.*

*a - Ensemble discret  $X$  ( $R = 24$ ).*

*b - Résultat de l'érosion de  $X$ .*

Le schéma synoptique de cette procédure de détection des frontières modales devient :



### III.3.5. Exemple d'application.

La procédure de détection des frontières modales, qui consiste à réaliser successivement les opérations d'érosion, d'amincissement par l'élément structurant **I**, de dilatation et de différence ensembliste, a été appliquée, à partir de l'exemple d'évaluation de la figure III.2, pour l'ensemble des valeurs entières positives de  $R$  pour lesquelles le nombre de modes détectés était non nul (voir figure III.6).

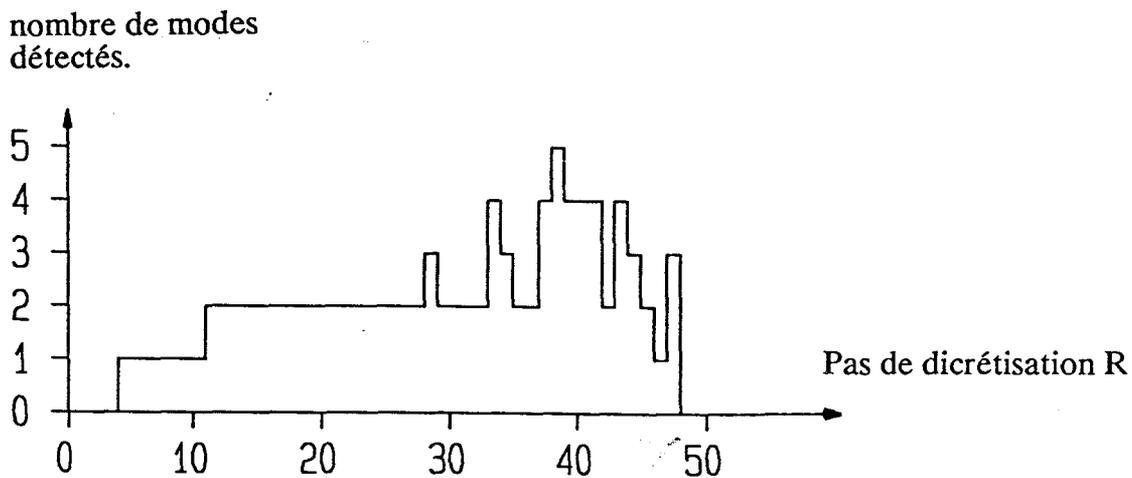
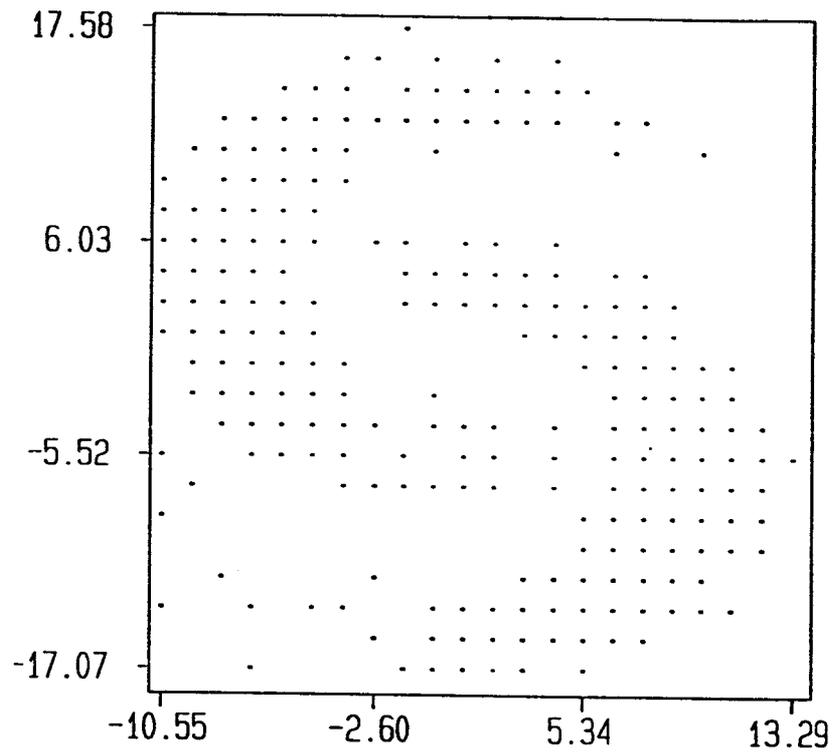


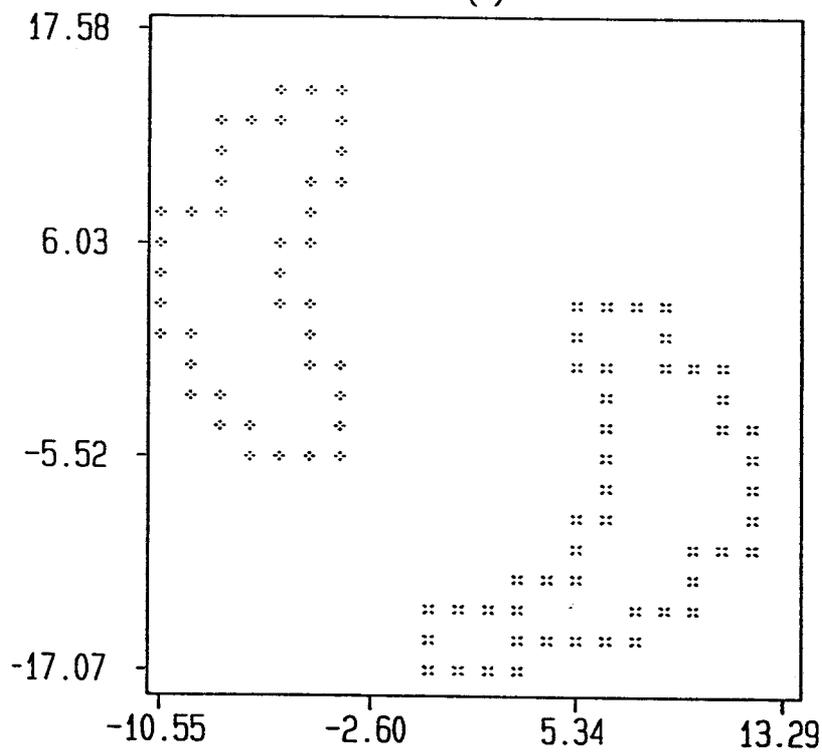
Figure III.6 : Effet du paramètre  $R$  sur le nombre de modes détectés.

La figure III.7.a présente l'ensemble discret obtenu avec un pas de discrétisation fixé à  $R = 21$ . La figure III.7.b montre l'ensemble des points constituant les frontières modales morphologiques individualisées.

Cet exemple montre comment cette méthode permet de détecter les modes morphologiques, en fournissant un ensemble discret des modes **M** constitué des points des différentes frontières modales. Afin de limiter l'effet du pas de discrétisation  $R$  sur le nombre de modes détectés, nous avons inclus un traitement intermédiaire d'amincissement par l'élément structurant **I** qui élimine les points isolés de l'ensemble érodé. Cette amélioration a permis d'augmenter la robustesse de la méthode puisque la plage de stabilité est passée, pour l'exemple présenté, de  $[11, 16]$  sans cet amincissement à  $[11, 27]$ .



(a)



(b)

**Figure III.7 :** Détection des modes par détermination des frontières modales.

*a - Ensemble discret ( $R = 21$ ).*

*b - Les deux modes détectés.*

### III.4. CONCLUSION.

Les méthodes que nous venons de présenter reposent essentiellement sur l'utilisation, simple ou composée, des opérations de dilatation et d'érosion. La première méthode fournit une description des modes morphologiques par leurs domaines modaux, la deuxième les définit par les frontières de ces domaines. Ces deux méthodes commencent toutes deux par une érosion de l'ensemble discret, ce qui permet l'analyse d'ensembles d'observations présentant un chevauchement des classes relativement important. Néanmoins, lorsque la dimension  $N$  de l'espace de représentation des données devient élevée, cette opération d'érosion tend à éliminer presque tous les points de l'ensemble discret puisqu'un point de  $\underline{X}$  ne sera conservé que s'il possède  $3^N - 1$  voisins appartenant eux-mêmes à  $\underline{X}$ . A titre d'exemple, pour  $N = 5$ , un point quelconque de l'ensemble discret est éliminé s'il ne possède pas ses 242 voisins dans l'ensemble discret! Ces deux méthodes sont donc fiables à condition que le nombre  $N$  d'attributs décrivant les observations ne soit pas trop grand.

Etant donnée cette restriction, nous abordons maintenant le problème de la recherche des modes morphologiques en adoptant une démarche différente. En effet, au lieu de sélectionner les points situés au coeur des classes par une transformation morphologique à élément structurant unique, nous allons analyser la connexité des points du réseau d'échantillonnage par des transformations morphologiques à éléments structurants multiples afin d'élargir les vallées ou de mettre en évidence les régions modales.

**IV - EXTRACTION DES MODES  
PAR ANALYSE MORPHOLOGIQUE  
DE LA CONNEXITE.**

## EXTRACTION DES MODES PAR ANALYSE MORPHOLOGIQUE DE LA CONNEXITE.

### IV.1. INTRODUCTION.

Nous avons vu, dans le chapitre précédent, deux techniques de mise en évidence des modes basées sur l'hypothèse qu'un point de l'ensemble discret  $\underline{X}$  peut être considéré comme un point d'un mode morphologique lorsque tous ses points voisins appartiennent à l'ensemble discret  $\underline{X}$ . Nous avons donc régularisé les composantes connexes représentant les domaines modaux en ne conservant que les points situés au coeur des classes.

Nous allons maintenant baser notre étude de recherche des modes morphologiques sur une analyse de connexité des points de l'ensemble discret. Dans l'espace euclidien  $E^N$ , les modes, caractérisés par des régions de forte concentration locale d'observations, sont séparés par des régions faiblement peuplées, sinon vides de toute observation, que l'on dénomme vallées. Dans l'espace discrétisé  $(Z^+)^N$ , ces différences de concentrations locales sont reflétées par la structure spatiale de l'ensemble discret  $\underline{X}$ . En effet, dans les vallées où les observations sont dispersées, les hypercubes non vides du réseau d'échantillonnage sont entourés d'un grand nombre d'hypercubes vides de toute observation. Par contre, au coeur des classes où les observations se regroupent, les hypercubes non vides du réseau sont entourés d'un grand nombre d'hypercubes non vides.

Comme nous l'avons défini dans le chapitre II, le centre d'un hypercube non vide du réseau d'échantillonnage est un point de l'ensemble discret  $\underline{X}$  à valeur binaire 1, tandis que le centre d'un hypercube vide est un point de l'ensemble  $\underline{X}^C$  à valeur

binaire 0.

L'extraction des modes morphologiques peut être ramenée à la détermination des domaines modaux par élimination des points de l'ensemble discret  $\underline{X}$  situés dans les vallées, qui tendent à avoir un grand nombre de voisins adjacents appartenant à  $\underline{X}^C$ .

Les domaines modaux peuvent également être caractérisés par l'ensemble des points de l'ensemble discret  $\underline{X}$ , situés au coeur des classes c'est-à-dire dans les régions modales, et qui tendent donc à avoir un grand nombre de voisins adjacents appartenant à  $\underline{X}$ .

Ces deux points de vue nous ont conduits à deux méthodes distinctes de détermination des domaines modaux par analyse morphologique de la connexité des points de  $\underline{X}$ , soit dans les vallées, soit dans les régions modales.

## IV.2. DETERMINATION DES DOMAINES MODAUX PAR ANALYSE MORPHOLOGIQUE DE LA CONNEXITE DANS LES VALLEES.

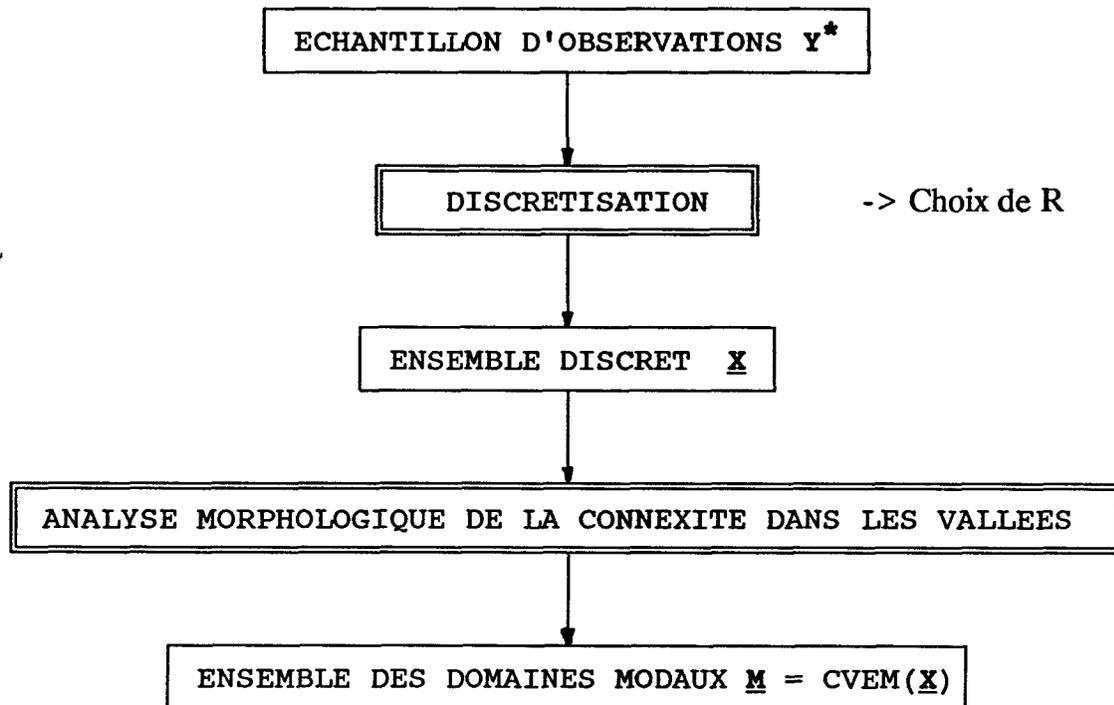
### IV.2.1. Principe général de la méthode.

Les domaines modaux, constitués de la liste explicite des points  $X$  de l'ensemble discret  $\underline{X}$  situés au coeur des classes, peuvent être déterminés en éliminant des vallées tous les "intrus" qui sont des points de l'ensemble  $\underline{X}$  plus ou moins isolés et donc entourés de nombreux points de  $\underline{X}^C$ . La connexité de ces voisins constitue, de plus, un indice intéressant de présence d'une vallée. Plus les voisins d'un point  $X$  appartenant à  $\underline{X}^C$  sont connexes entre-eux et plus le point considéré a de chance de se trouver au coeur d'une vallée.

L'élimination de ces points plus ou moins isolés de  $\underline{X}$  peut être effectuée à l'aide d'une transformation morphologique par une famille structurante  $CV$ , chaque élément structurant de cette famille caractérisant une configuration locale possible dans le

voisinage des points "intrus" qu'il s'agit de détecter. L'ensemble discret résultant de cette analyse de la Connexité dans les Vallées par une transformation morphologique à Elément structurant Multiple sera noté  $CVEM(\underline{X})$ .

Le principe général de cette méthode peut être résumé par le schéma synoptique suivant :



#### IV.2.2. Définition de la famille structurante CV.

Le but de ce paragraphe est de définir la famille structurante CV permettant de reconnaître, dans l'espace discrétisé, les configurations locales qui caractérisent un point de  $\underline{X}$  à éliminer dans une vallée. Afin de permettre une visualisation simple de la procédure proposée, celle-ci est présentée, dans un premier temps, dans un contexte bidimensionnel avant d'être généralisée aux espaces multidimensionnels.

Soit l'ensemble bidimensionnel discret  $\underline{X}$  illustré par la figure IV.1 qui résulte de la discrétisation d'un ensemble de 300 observations sur un réseau de  $20^2$  carrés élémentaires. Les carrés non vides, représentés par la valeur 1 et définissant l'ensemble

X, forment ici deux modes morphologiques qu'il s'agit d'individualiser.

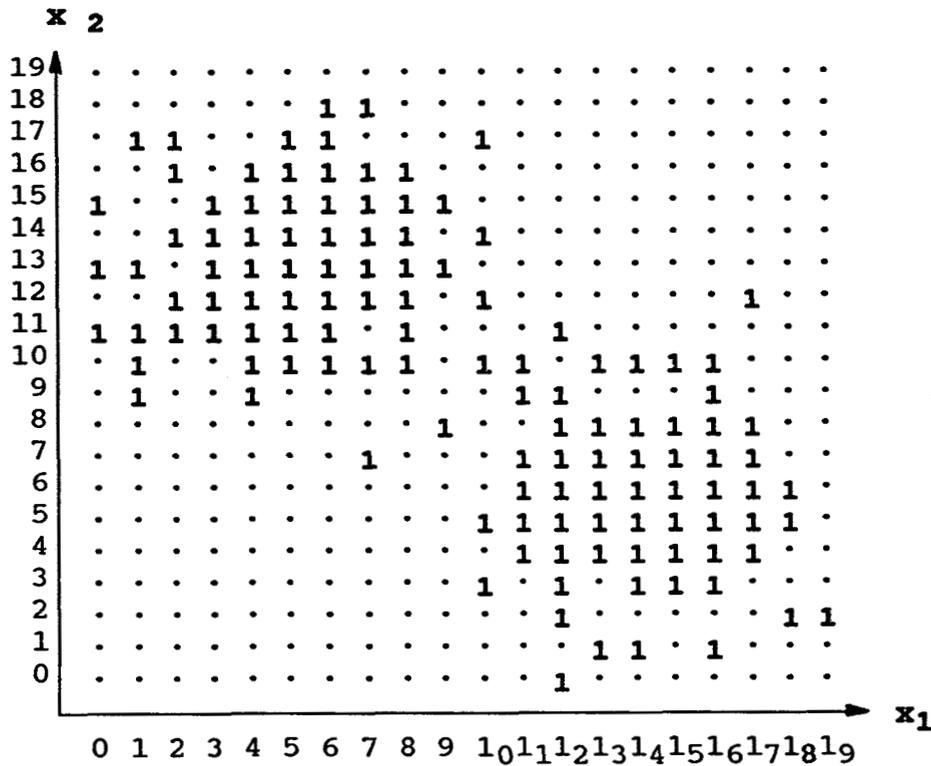


Figure IV.1 : Ensemble discret X à analyser.

Considérons un élément structurant E quelconque défini, dans l'espace bidimensionnel, sur une fenêtre d'observation de taille 3 x 3 et constitué de points à valeurs binaires 1 ou 0 dont l'origine est située au centre de la fenêtre.

L'examen de la structure spatiale de l'ensemble discret X permet d'affirmer que, lorsque la majorité des voisins d'un point X de l'ensemble X situés dans la fenêtre d'observation centrée en X, sont des points appartenant à X, le point considéré est certainement à l'intérieur d'un domaine modal, où la concentration des observations est élevée. A l'opposé, si la majorité des voisins du point X situés dans cette fenêtre appartiennent à X<sup>C</sup>, cela signifie que ce point est, soit situé à la périphérie d'un domaine modal, soit dans une vallée quasiment vide d'observation.

On pourrait admettre que si la majorité des points situés dans la fenêtre, c'est-à-dire cinq points sur neuf, appartiennent à X<sup>C</sup>, le point X appartient à une vallée. Mais cette condition doit être rendue plus forte car une vallée se caractérise par une certaine

connexité de points de  $\underline{X}^C$ . Nous imposons donc que le point  $X$  soit éliminé des vallées si et seulement si il a au moins cinq voisins appartenant à  $\underline{X}^C$  connexes entre-eux.

Cette contrainte de connexité conduit à la définition d'une famille structurante  $CV$  de huit éléments structurants caractérisant les points des vallées, telle que :

$$CV = \{\underline{CV}_1, \underline{CV}_2, \dots, \underline{CV}_8\}$$

L'élément structurant  $\underline{CV}_1$  est défini par la configuration :

0	0	*
0	<u>1</u>	*
0	0	*

dans laquelle les cinq "0" sont connexes et où les étoiles "\*" indiquent des points des configurations locales pouvant être indifféremment à 1 ou à 0.

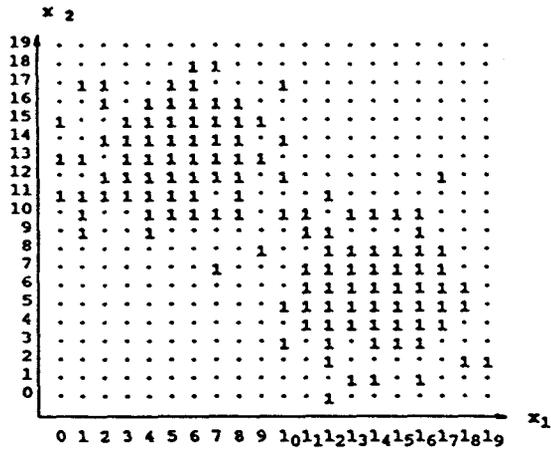
Les éléments structurants  $\underline{CV}_2, \dots, \underline{CV}_8$  sont obtenus par sept rotations successives de  $\pi/4$  de  $\underline{CV}_1$ .

### IV.2.3. Transformation morphologique associée à la famille structurante $CV$ .

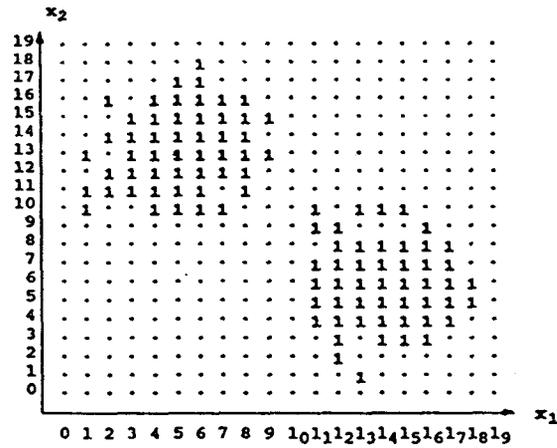
La transformation morphologique permettant d'éliminer les points de  $\underline{X}$  dont la configuration locale coïncide avec l'un des huit éléments structurants définis ci-dessus est un amincissement par la famille structurante  $CV$  (cf. paragraphe II.4). L'ensemble résultant de cette transformation est noté  $\underline{X} \circ CV$ .

Au cours de cette phase du traitement, l'ordre dans lequel les points de  $\underline{X}$  sont pris en compte n'influe pas sur le résultat obtenu. En effet, les points dont la configuration locale répond au critère morphologique ne sont définitivement éliminés que lorsque tous les points de  $\underline{X}$  ont été traités.

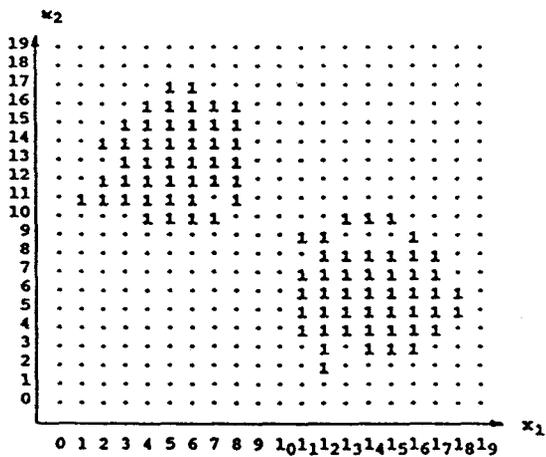
La procédure d'amincissement que nous venons de définir est implantée en mode itératif afin d'éliminer progressivement tous les points de  $\underline{X}$  appartenant aux vallées. Lorsqu'il y a stabilisation du nombre de points de l'ensemble résultant, noté  $CVEM(\underline{X})$ , tous les points de l'ensemble aminci possèdent moins de cinq voisins appartenant à  $\underline{X}^C$  et connexes entre-eux (cf. figure IV.2).



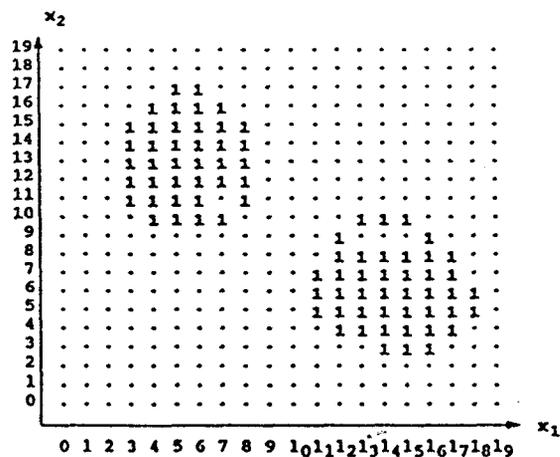
(a)



(b)



(c)



(d)

Figure IV.2 : Amincissement itératif par la famille structurante CV.

a - Ensemble  $\underline{X}$  à analyser.

c - Résultat de l'amincissement après 2 itérations.

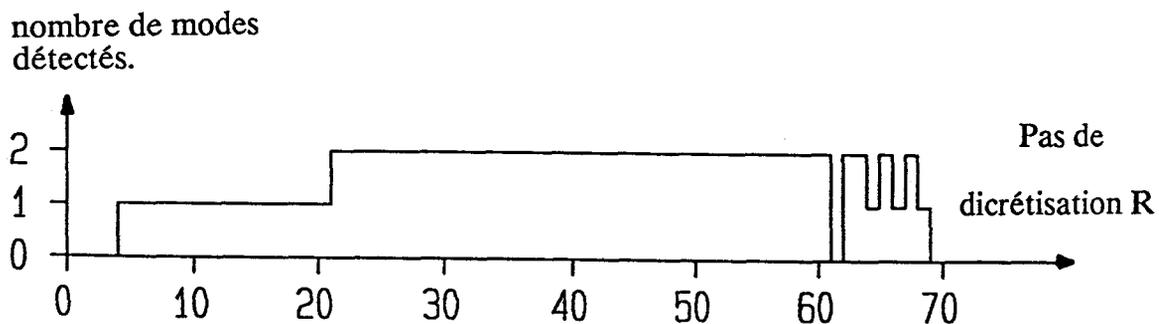
b - Résultat de l'amincissement après une itération.

d - Résultat stabilisé  $CVEM(\underline{X})$  après 5 itérations.

La transformation que nous venons de présenter dans un espace bidimensionnel peut être facilement généralisée dans un contexte multidimensionnel. Dans un espace à  $N$  dimensions, la procédure d'amincissement itératif permettant l'analyse morphologique des vallées consiste à tester la coïncidence entre l'un des éléments structurants de la famille  $CV$  et la configuration locale en chaque point  $X$  de l'ensemble discret  $\underline{X}$ , dans chacun des  $(N * (N - 1)) / 2$  plans définis par le point  $X$  et les axes de l'espace normalisé pris deux à deux. Lorsque la configuration locale  $C_{CV}(X)$ , définie dans l'un de ces plans, coïncide avec l'un des éléments structurants de la famille  $CV$ , le point est éliminé de l'ensemble  $\underline{X}$ . Après plusieurs itérations, lorsqu'il y a stabilisation du résultat, on obtient l'ensemble  $\underline{M} = CVEM(\underline{X})$  constituant les domaines modaux.

#### IV.2.4. Exemple d'Application.

Nous allons maintenant montrer les résultats obtenus par cette technique avec le même exemple d'évaluation que pour le chapitre précédent. La procédure de détermination des domaines modaux par analyse morphologique de la connexité dans les vallées a été appliquée pour toutes les valeurs du pas de discrétisation  $R$  jusqu'à ce que le nombre de modes détectés soit nul (cf. figure IV.3).



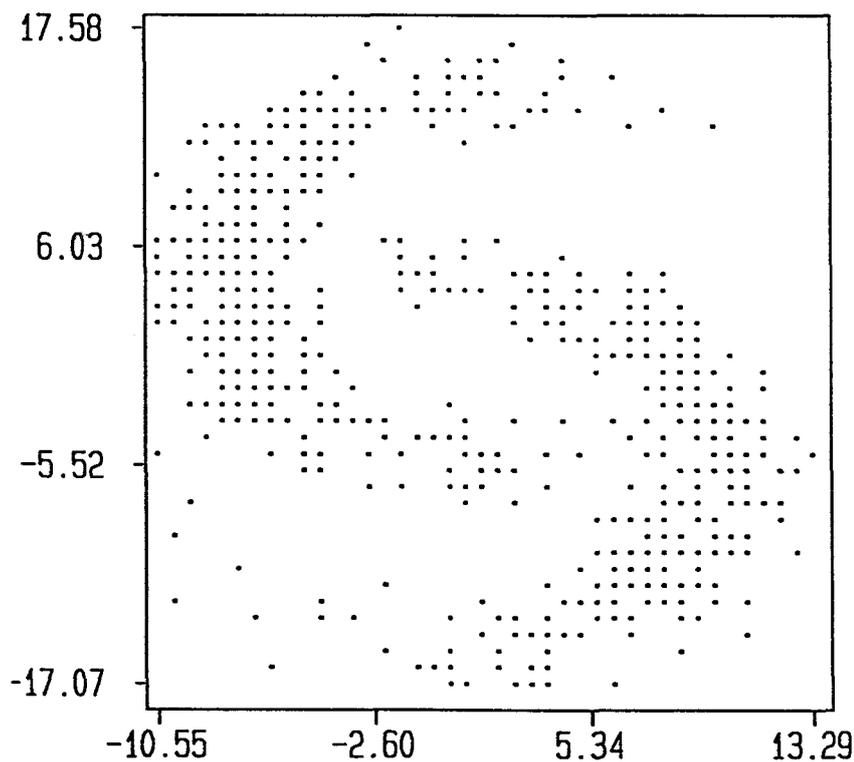
*Figure IV.3 : Effet du paramètre  $R$  sur le nombre de modes détectés par amincissement itératif.*

Contrairement aux courbes du même type relatives aux méthodes présentées au chapitre III, on peut remarquer que le nombre maximum de modes détectés n'excède

pas deux. Cette propriété est directement liée au principe général de la méthode, qui est de renforcer les vallées sans chercher à réduire les régions modales, qui, de ce fait, ne se scindent pas en plusieurs parties.

En choisissant le pas de discrétisation au milieu de la plage de stabilité [21, 60], l'amincissement itératif de l'ensemble discret de la figure IV.4.a par la famille CV permet de mettre en évidence deux modes bien individualisés comme le montre la figure IV.4.b.

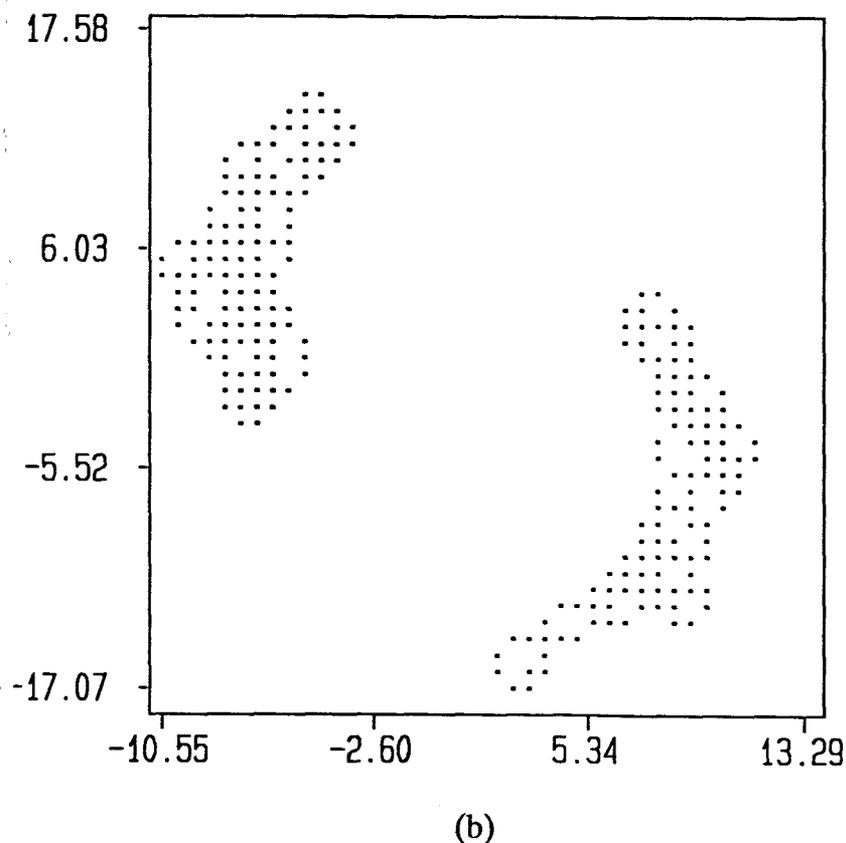
De plus, on peut voir sur cette figure que la forme des domaines modaux reflète fidèlement la forme des deux groupes d'observations.



(a)

**Figure IV.4 :** Détermination des domaines modaux par analyse morphologique de la connexité dans les vallées.

a - Ensemble discret de points à valeurs binaires ( $R = 40$ ).



*Figure IV.4 : Détermination des domaines modaux par analyse morphologique de la connexité dans les vallées (suite).*

*b - Modes morphologiques détectés.*

### **IV.2.5. Conclusion.**

Dans ce paragraphe, nous proposons une technique de détermination des domaines modaux par analyse morphologique de la connexité dans les vallées. Pour cela, le voisinage de chaque point de l'ensemble discret  $\underline{X}$  est comparé à chaque élément structurant de la famille  $\underline{CV}$ , ce qui débouche sur une procédure d'amincissement itératif. Cette méthode est simple à implanter puisqu'elle ne met en oeuvre que des opérateurs logiques tels que ET et OU. De plus, l'ordre de passage des éléments structurants et l'ordre dans lequel les points de  $\underline{X}$  sont pris en compte n'interviennent pas sur le résultat obtenu.

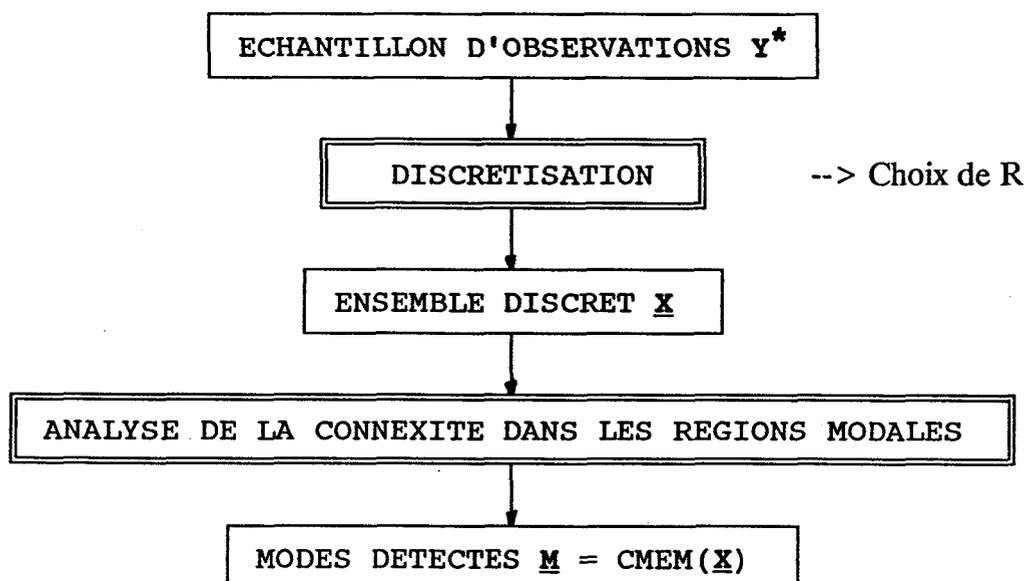
### IV.3. DETERMINATION DES DOMAINES MODAUX PAR ANALYSE MORPHOLOGIQUE DE LA CONNEXITE DANS LES REGIONS MODALES.

#### IV.3.1. Principe général de la méthode.

La détermination des domaines modaux vient d'être abordée par une analyse morphologique de la connexité des points de l'ensemble discret  $\underline{X}$  situés dans les vallées, afin d'éliminer tous les points qui répondent à certains critères morphologiques de connexité.

Les domaines modaux peuvent également être déterminés par une analyse morphologique de la connexité des points de l'ensemble discret  $\underline{X}$  situés dans les régions modales. Comme précédemment, la recherche des points situés dans ces régions nous a conduit à la définition d'une nouvelle famille structurante  $CM$ . L'ensemble résultant de la détermination des domaines modaux par analyse de la Connexité dans les régions Modales par une transformation morphologique à Eléments structurants Multiples est noté  $CMEM(\underline{X})$ .

Le principe général de détermination des domaines modaux que nous venons de présenter peut être résumé par le schéma synoptique suivant :



### **IV.3.2. Définition de la famille structurante CM.**

Pour permettre une visualisation simple de la méthode proposée, nous la présentons dans un contexte bidimensionnel avant de la généraliser aux espaces multidimensionnels.

Considérons l'ensemble discret  $\underline{X}$  présenté sur la figure IV.1 (cf. paragraphe IV.2.2). Le but de la méthode que nous proposons maintenant est de conserver tous les points de l'ensemble  $\underline{X}$  qui appartiennent aux régions modales.

Nous avons déjà précisé que les points  $X$  de l'ensemble discret  $\underline{X}$  dont la majorité des voisins situés dans une fenêtre d'observation  $3 \times 3$  centrée en  $X$ , appartiennent à l'ensemble  $\underline{X}$ , ont de grandes chances d'être situés dans des régions modales.

Cependant, l'analyse approfondie de la répartition spatiale des points de ces régions nécessite une condition supplémentaire de connexité. Ainsi, chaque point de l'ensemble discret  $\underline{X}$  sera conservé comme point constitutif des domaines modaux si et seulement si il possède, dans une fenêtre d'observation de taille  $3 \times 3$  centrée en ce point, au moins quatre voisins appartenant à  $\underline{X}$  et connexes entre-eux.

La famille structurante  $CM$  permettant de rechercher de tels points est composée de huit éléments structurants telle que :

$$CM = \{\underline{CM}_1, \underline{CM}_2, \dots, \underline{CM}_8\}.$$

L'élément structurant  $\underline{CM}_1$  est défini par :

1	1	1
1	<u>1</u>	*
*	*	*

les éléments structurants  $\underline{CM}_2, \dots, \underline{CM}_8$  sont obtenus par les sept rotations successives de  $\pi/4$  de  $\underline{CM}_1$ .

### IV.3.3. Définition de la transformation morphologique associée.

La transformation morphologique qui permet de conserver tous les points  $X$  dont la configuration locale  $C_{CM}(X)$  coïncide avec l'un des huit éléments structurants de la famille  $CM$  est la transformation en tout ou rien définie par  $X \otimes CM$ .

Afin d'écartier progressivement tous les points de l'ensemble discret qui ne répondent pas à ce critère morphologique, la transformation en tout ou rien  $X \otimes CM$  est itérée jusqu'à stabilisation du nombre de points de l'ensemble résultant  $CMEM(\underline{X})$  (cf. figure IV.5).

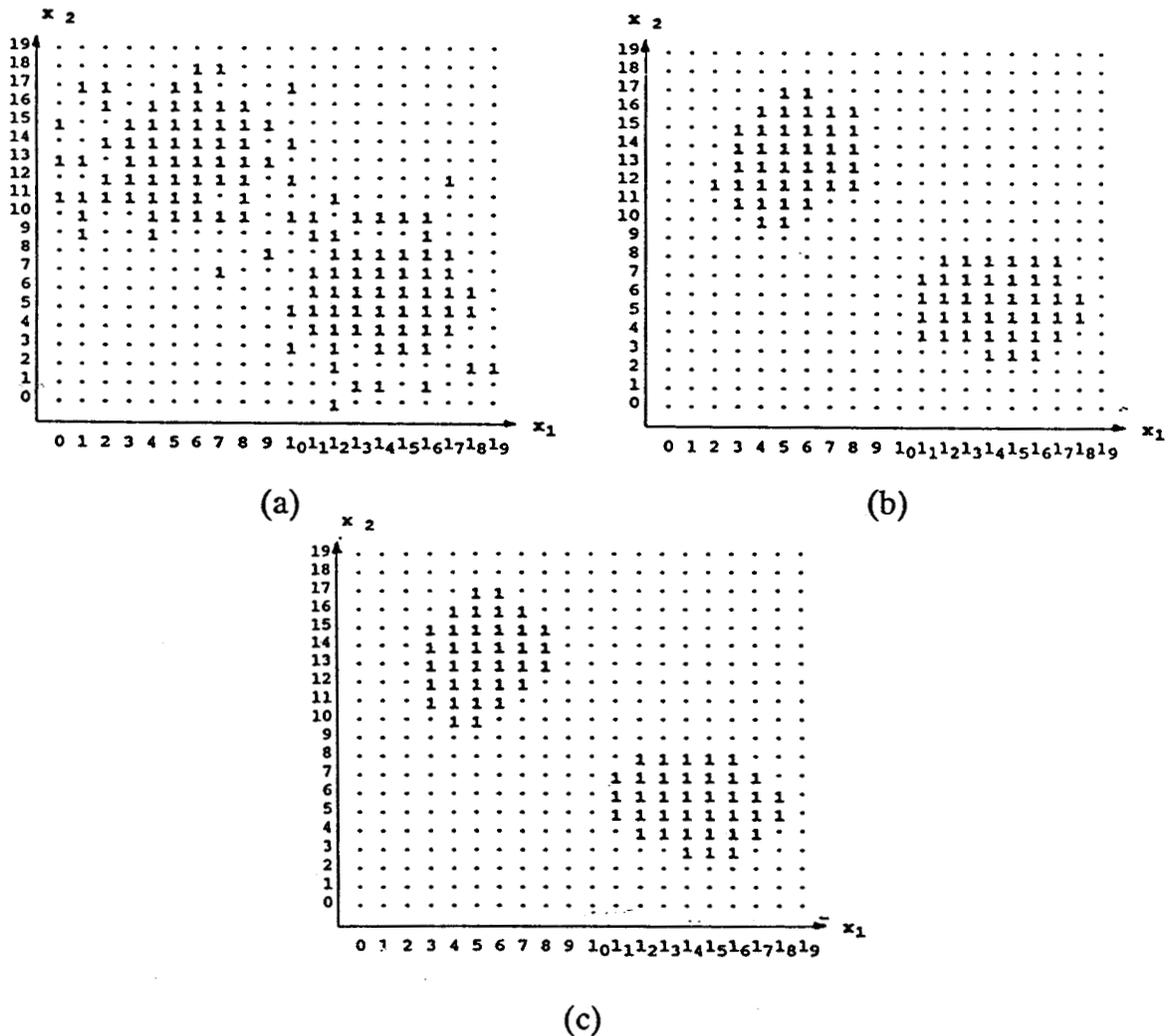


Figure IV.5 : Analyse morphologique itérative de la connexité dans les régions modales.

a - Ensemble  $\underline{X}$  à analyser.

b - Résultat de la transformation après une itération.

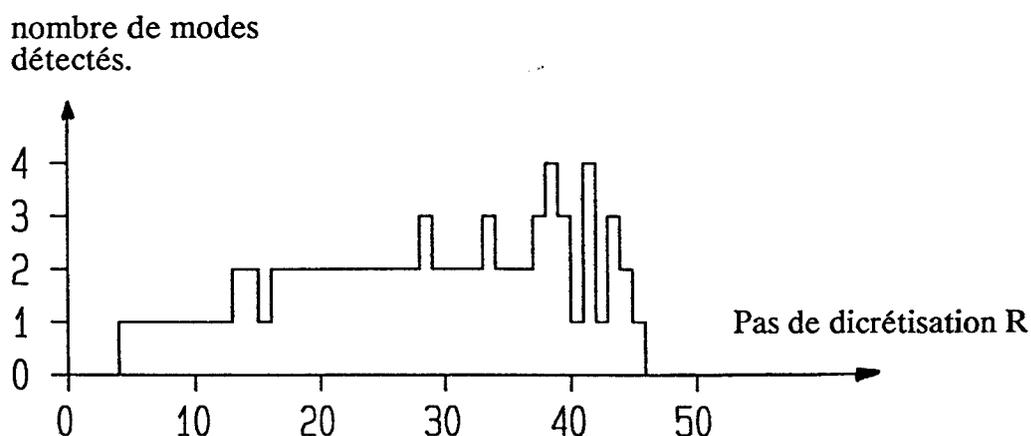
c - Résultat stabilisé  $CMEM(\underline{X})$  après 2 itérations.

Pour des données multidimensionnelles, on opère de la même manière que pour l'amincissement itératif, en filtrant plan par plan, c'est-à-dire en prenant les axes deux à deux. Il faut donc, dans un espace à  $N$  dimensions, considérer les  $(N * (N - 1))/2$  plans perpendiculaires aux axes de cet espace.

Si, dans chacun des plans de travail ainsi définis, le point  $X$  de  $\underline{X}$  possède au moins quatre voisins appartenant à  $\underline{X}$  et connexes entre-eux, il est conservé comme point de  $\text{CMEM}(\underline{X})$ , sinon il est éliminé.

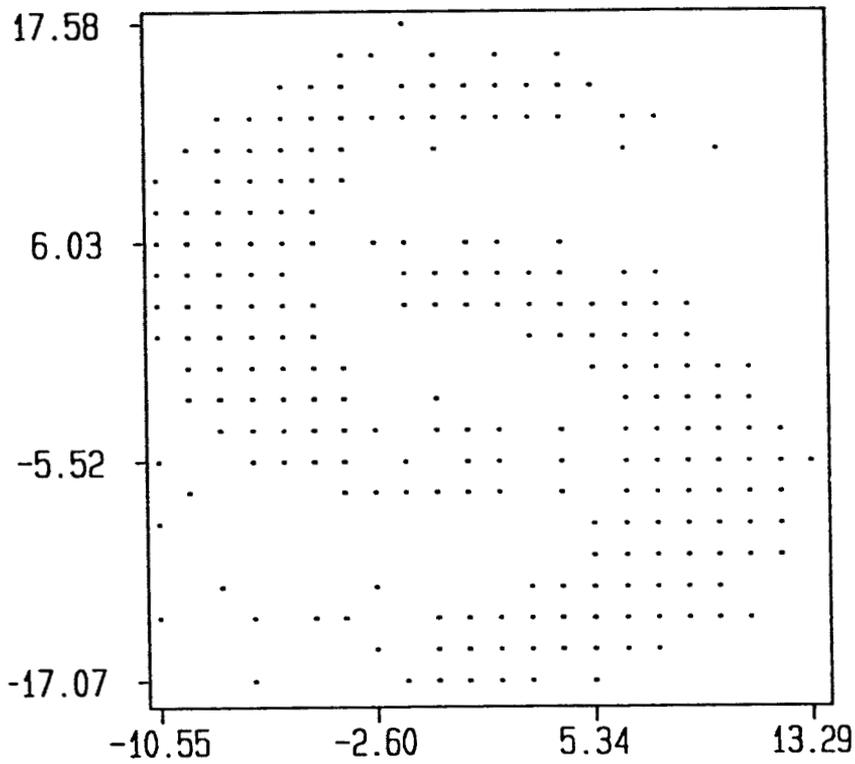
#### **IV.3.4. Exemple d'application.**

Afin de comparer les différentes techniques présentées dans cette deuxième partie, nous avons également fait varier le pas de discrétisation afin de relever le nombre de modes détectés pour chaque valeur de  $R$ , comme le montre la figure IV.6.

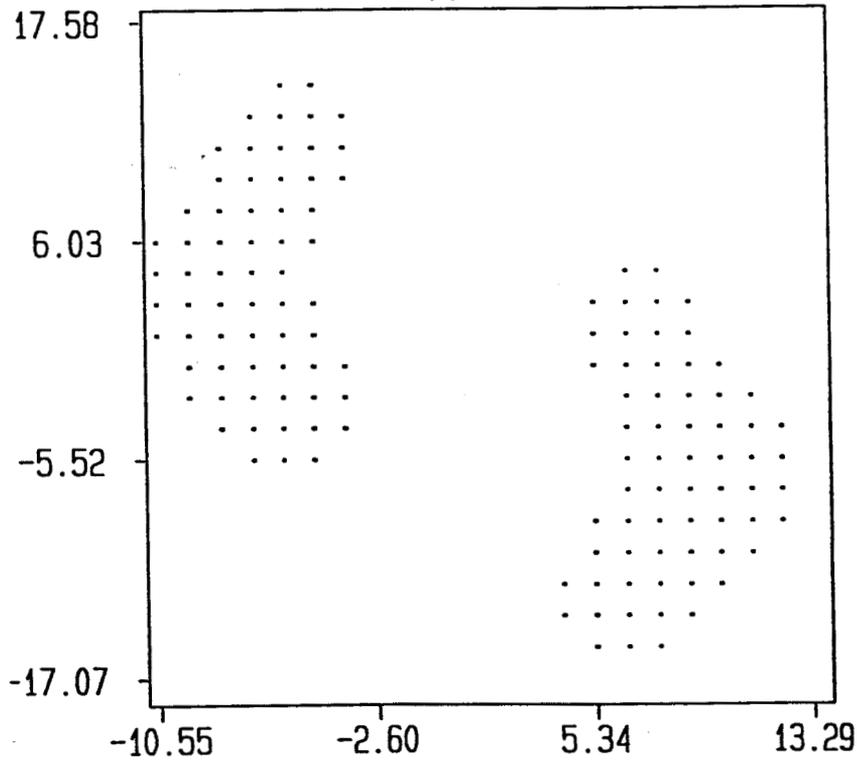


*Figure IV.6 : Effet du paramètre  $R$  sur le nombre de modes détectés par analyse morphologique de la connexité dans les régions modales.*

Les domaines modaux déterminés par analyse morphologique de la connexité dans les régions modales de l'ensemble discret  $\underline{X}$  de la figure IV.7.a sont présentés sur la figure IV.7.b.



(a)



(b)

Figure IV.7 : Détermination des domaines modaux par analyse morphologique de la connexité dans les régions modales.

a - Ensemble discret ( $R = 21$ ).

b - Les 2 modes détectés.

### **IV.3.5. Conclusion.**

Nous venons de proposer une méthode de détermination des domaines modaux par analyse morphologique de la connexité des régions modales.

Cette technique itérative présente les mêmes avantages que celle basée sur l'analyse morphologique de la connexité dans les vallées. Ainsi, elle est simple à mettre en oeuvre et non séquentielle, c'est-à-dire que l'ordre de prise en compte des points de l'ensemble discret  $X$  ainsi que l'ordre de passage des éléments structurants de la famille  $CM$  n'interviennent pas sur le résultat obtenu.

De plus, cette méthode permet l'analyse d'échantillons où les classes présentent un degré de chevauchement plus important que dans le cas de la détermination des domaines modaux par analyse morphologique des vallées. C'est la configuration des points, dans chaque plan d'étude, situés à la périphérie immédiate des composantes connexes que nous cherchons à individualiser, qui expliquent le plus clairement cette propriété.

En effet, ces régions périphériques sont à la fois trop peuplées pour qu'au moins cinq voisins adjacents aux points de ces régions appartiennent à  $X^C$ , et insuffisamment peuplées pour que ces mêmes points aient au moins quatre voisins adjacents appartenant à  $X$ . Ces points ne sont donc pas éliminés des vallées lors de l'analyse de leur connexité et ne contribuent donc pas à leur élargissement. Ils ont donc tendance à limiter le procédé d'individualisation des classes de la première méthode proposée dans ce chapitre. Par contre, ils sont éliminés par la méthode d'analyse de la connexité dans les régions modales et ne sont donc pas conservés pour définir les domaines modaux. Leur disparition dans la seconde méthode proposée dans ce chapitre, contribue donc à individualiser les classes.

#### IV.4. COMPLEXITE DES ALGORITHMES D'EXTRACTION DES MODES.

Le temps de calcul nécessaire à l'extraction des modes est un facteur important. En effet, pour toutes les techniques d'extraction des modes que nous présentons dans ce mémoire, la procédure de détermination des domaines modaux ou de leurs frontières doit être itérée jusqu'à ce que le nombre de modes détectés soit nul afin de déterminer le pas de discrétisation.

Nous avons donc évalué le temps d'exécution des procédures associées aux différentes méthodes d'extraction des modes proposées dans ce mémoire (cf. table T4.1), pour différentes tailles d'échantillons et différentes valeurs de la dimension des données.

CODE METHODE	INTITULE DE LA METHODE DE DETECTION DES MODES MORPHOLOGIQUES
DMEU	Détermination des <u>D</u> omaines <u>M</u> odaux par <u>E</u> lément structurant <u>U</u> nique.
FMEU	Détermination des <u>F</u> rontières <u>M</u> odales par <u>E</u> lément structurant <u>U</u> nique.
CVEM	Analyse de la <u>C</u> onnexité des <u>V</u> allées par transformation morphologique à <u>E</u> léments structurants <u>M</u> ultiples.
CMEM	Analyse de la <u>C</u> onnexité des régions <u>M</u> odales par transformation morphologique à <u>E</u> léments structurants <u>M</u> ultiples.

*Table T4.1 : Codification des différentes méthodes d'extraction des modes morphologiques présentées dans la deuxième partie de ce mémoire.*

Les données utilisées pour cette évaluation sont des observations normales réparties en deux classes distinctes. Les vecteurs moyennes associés sont définis par  $M_1 = [2, 2, \dots, 2]^T$  et  $M_2 = [-2, -2, \dots, -2]^T$ , les matrices de covariance étant des matrices unités.

Comme le but de ces simulations est de montrer les effets de la dimension de l'espace de représentation des données et de la taille des échantillons sur le temps de calcul nécessaire à l'extraction des modes, tous les essais ont été effectués avec le même pas de discrétisation  $R$  sur chacun des axes, ici fixé à  $R = 20$  (cf. figure IV. 8).

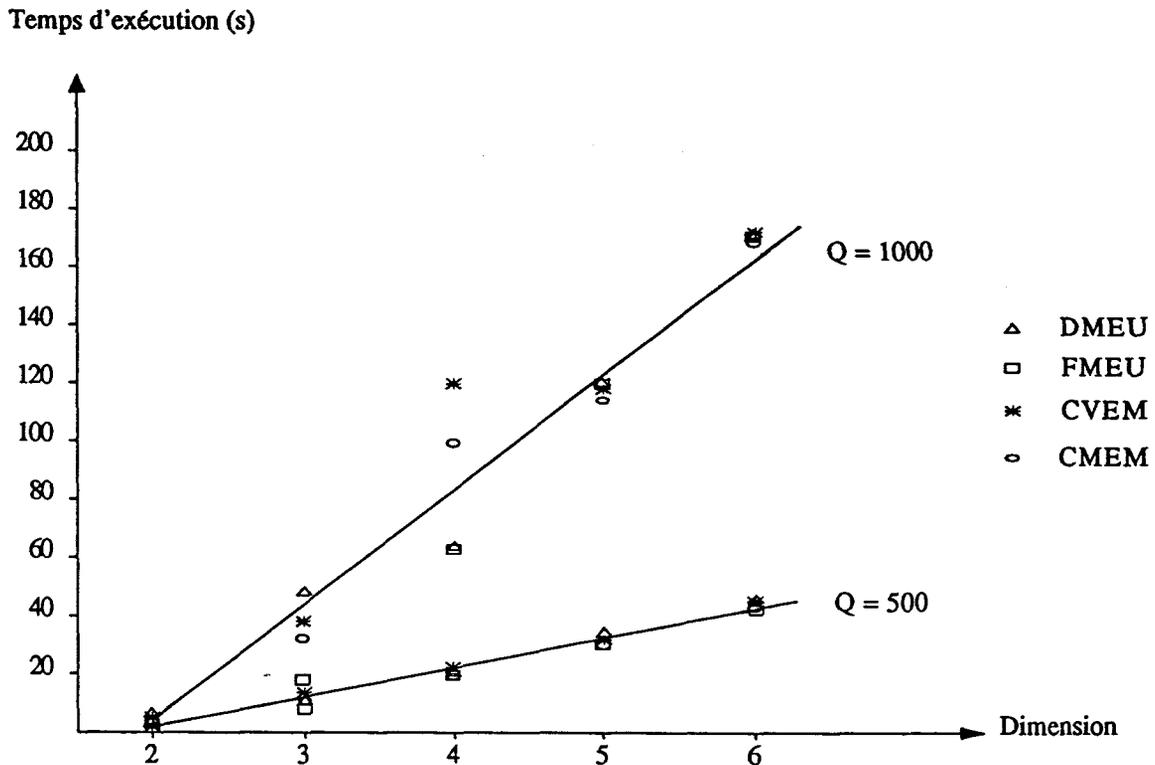


Figure IV.8 : Evolution du temps d'exécution des procédures de détection des modes en fonction de la dimension de l'espace et de la taille des échantillons.

Bien que ces temps d'exécution dépendent de l'environnement informatique utilisé pour effectuer les simulations, ils donnent une idée générale du temps de calcul nécessaire à l'extraction des modes.

#### IV.5. CONCLUSION.

Nous avons présenté, dans ce chapitre, deux méthodes de détermination des domaines modaux reposant sur l'analyse morphologique de la connexité des points de l'ensemble discret, soit dans les vallées, soit dans les régions modales, ce qui nous a conduit à la définition de deux transformations morphologiques à éléments structurants multiples distinctes.

L'illustration des procédures associées sur un exemple d'évaluation commun à toutes les méthodes, a montré que la forme géométrique des domaines modaux reflète bien la forme géométrique des classes de l'échantillon.

A l'issue de la deuxième partie de ce mémoire, nous disposons de quatre procédures distinctes d'extraction des modes morphologiques. Ces modes morphologiques sont décrits soit par les domaines modaux (méthodes DMEU, CMEM, CVEM) soit par les frontières de ces domaines (méthode FMEU).

Nous allons aborder maintenant la troisième partie de ce mémoire dont le but est, d'une part, de présenter les techniques de classification des observations pouvant être mises en oeuvre après détermination des domaines modaux, d'autre part, de montrer les résultats obtenus par l'ensemble de ces techniques à l'aide d'exemples de données simulés.

**- TROISIEME PARTIE -**

**APPLICATION A L'ANALYSE  
DE DONNEES**

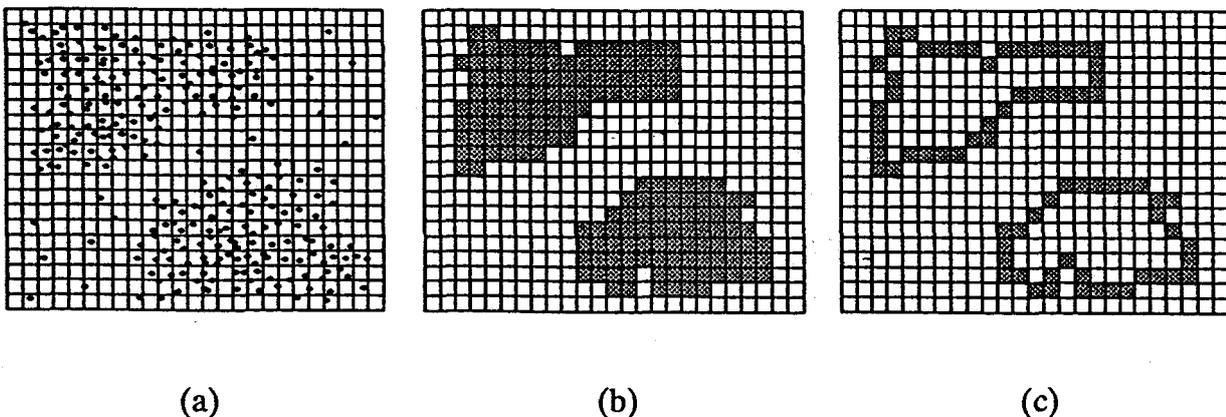
**V - CLASSIFICATION DES DONNEES A PARTIR  
DES MODES MORPHOLOGIQUES.**

## CLASSIFICATION DES DONNEES A PARTIR DES MODES MORPHOLOGIQUES.

### V.1. INTRODUCTION.

La deuxième partie de ce mémoire a été consacrée à la présentation de différentes techniques d'extraction des modes morphologiques en présence dans l'ensemble des observations à analyser. Une étude de l'effet du pas de discrétisation sur le nombre de modes détectés a permis de dégager, pour chaque technique, une plage de stabilité où les modes morphologiques peuvent être facilement identifiés.

L'ensemble des observations à analyser peut ainsi être partitionné en  $K$  classes  $C_k$ , chacune des classes  $C_k$  mises en évidence étant définies à partir d'un domaine modal  $D_k$ ,  $k = 1, 2, \dots, K$ . Dans le cadre de notre étude, les domaines  $D_k$  sont définis par l'ensemble discret  $\underline{M}$  constitué soit de tous les points des domaines modaux soit uniquement des points situés à la frontière de ces domaines (cf. figure V.1).



*Figure V.1 : Définition des domaines modaux permettant la classification des observations.*

- a - Ensemble des observations à classer.*
- b - Ensemble discret  $\underline{M}$  constitué des domaines modaux.*
- c - Ensemble discret  $\underline{M}$  constitué des frontières modales.*

De manière générale, l'objectif de la classification est d'affecter chacune des observations de l'échantillon à la "meilleure" classe possible. Plusieurs méthodes peuvent être mises en oeuvre pour effectuer cette opération. Les deux premières techniques que nous présentons dans ce chapitre utilisent directement les points de l'ensemble  $M$  pour affecter les observations. Les deux dernières techniques commencent par définir les observations situées à l'intérieur des domaines modaux  $D_k$  comme prototypes des classes  $C_k$  correspondantes, pour réaliser ensuite la classification des observations extérieures à ces domaines (cf. figure V.2). L'ensemble des prototypes de la classe  $C_k$  constitue le noyau de la classe  $C_k$ .

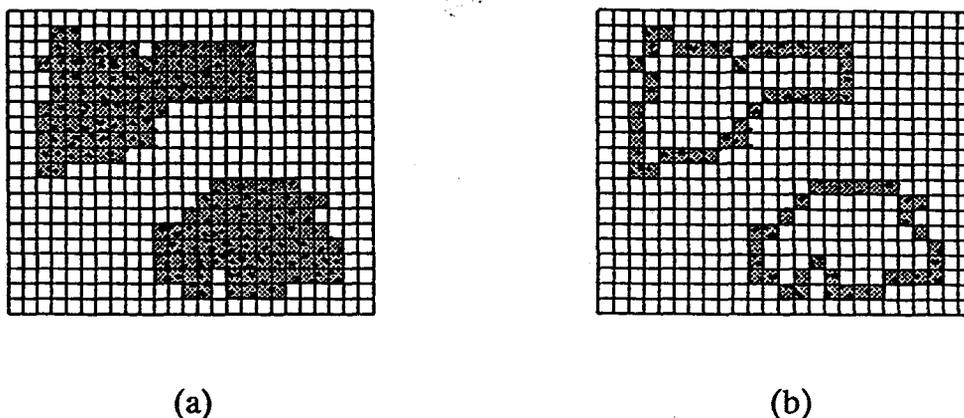


Figure V.2 : Définition des prototypes des classes situés à l'intérieur des domaines modaux.

- a - Prototypes situés à l'intérieur des domaines modaux.
- b - Prototypes situés à l'intérieur des frontières modales.

Nous précisons qu'un point désigne un élément de l'ensemble discret  $M$  constitué des domaines modaux ou de leurs frontières, tandis qu'un prototype désigne une des observations constitutives du noyau d'une classe.

Aux quatre méthodes d'extraction des modes que nous avons présentées, nous allons associer maintenant quatre techniques de classification. Afin de ne pas alourdir la présentation des résultats, nous avons analysé ces quatre techniques de classification en nous appuyant sur deux exemples. Les modes morphologiques sont extraits, pour chaque exemple, par l'analyse morphologique de la connexité dans les régions modales présentée au paragraphe IV.3.

Le seul moyen d'évaluer les performances d'une procédure de classification non supervisée est de l'appliquer à un ensemble d'observations générées artificiellement, et de comparer le résultat obtenu avec la véritable structure des données. Nous allons donc analyser les résultats obtenus par les quatre techniques de classification présentées dans ce chapitre à l'aide de deux exemples distincts. Le premier est l'exemple d'évaluation que nous avons utilisé dans les chapitres III et IV. Le deuxième exemple est un ensemble de 660 observations normales réparties en trois classes de 220 observations chacune. Dans ce dernier cas, nous comparerons les résultats obtenus à ceux de la classification optimale définie par la règle de Bayes, que nous rappelons succinctement.

## V.2. CLASSIFICATION OPTIMALE DE BAYES.

Dans le cas où l'échantillon est composé d'observations tirées indépendamment de  $K$  classes  $C_k$ ,  $k = 1, 2, \dots, K$ , selon des lois de probabilités conditionnelles  $P(Y/C_k)$  connues, il est possible de déterminer la probabilité pour qu'une observation  $Y$  appartienne à la classe  $C_k$ .

Par application du théorème de Bayes, cette probabilité, appelée probabilité a posteriori de la classe  $C_k$ , est donnée par la relation :

$$P(C_k/Y) = \frac{P(Y/C_k) * P(C_k)}{f(Y)}$$

où :

-  $f(Y)$  désigne la valeur de la fonction de densité de probabilité sous-jacente à la distribution des observations provenant de l'échantillon au point  $Y$ ,

-  $P(C_k)$  désigne la probabilité a priori de la classe  $C_k$  de telle sorte que :

$$\sum_{k=1}^K P(C_k) = 1$$

-  $P(Y/C_k)$  désigne la probabilité conditionnelle de la classe  $C_k$ , c'est-à-dire la fonction de densité de probabilité sous-jacente à la distribution des observations provenant de la classe  $C_k$ .

Dans ces conditions, le risque d'erreur lié à l'assignation de l'observation Y à la classe  $C_k$  est :

$$R(C_k/Y) = 1 - P(C_k/Y)$$

Pour obtenir un classement optimal, il faut affecter toute observation Y à la classe  $C_k$  qui minimise ce risque, donc qui maximise la probabilité a posteriori.

Le problème de la classification d'une observation Y peut être résolu en adoptant la règle de décision suivante :

$$\text{Décider } C_S \text{ si } P(C_S/Y) > P(C_k/Y), \forall k = 1, 2, \dots, K, k \neq S.$$

En appliquant la règle de Bayes, ceci revient à déterminer un ensemble de fonctions de décision  $d_k(Y)$ ,  $k = 1, 2, \dots, K$ , telles que l'observation Y est assignée à la classe  $C_k$  si et seulement si :

$$d_k(Y) > d_i(Y), \forall i = 1, 2, \dots, K, i \neq k.$$

où :

$$d_k(Y) = P(Y/C_k) * P(C_k).$$

Dans les conditions particulières où les échantillons ont la structure de mélanges gaussiens, la fonction de densité est donnée par :

$$f(Y) = \sum_{k=1}^K P(Y/C_k) * P(C_k)$$

où la probabilité conditionnelle de la classe  $C_k$  est définie par :

$$P(Y/C_k) = \frac{1}{(2\pi)^{N/2} |M_k|^{1/2}} \exp \left\{ -1/2 (Y - Y_k)^T M_k^{-1} (Y - Y_k) \right\},$$

$Y_k$  étant le vecteur moyenne de la classe  $C_k$ ,

et  $M_k$  la matrice de covariance de la classe  $C_k$ .

Ainsi, les fonctions de décision  $d_k$  peuvent être prises, après simplifications réalisées en prenant le logarithme népérien et en supprimant le coefficient multiplicateur constant, sous la forme :

$$d_k(Y) = -\frac{1}{2} \text{Log} |M_k| - \frac{1}{2} (Y - Y_k)^T M_k^{-1} (Y - Y_k) + \text{Log} (P(C_k))$$

la fonction Log étant une fonction croissante monotone.

Cette règle de classification est optimale, c'est-à-dire qu'aucune autre procédure ne peut conduire à un taux d'erreur plus petit [DUD - 73]. Nous utiliserons donc les résultats de cette règle comme références de classification dans le cas d'échantillons ayant la structure de mélanges gaussiens générés artificiellement et dont nous connaissons parfaitement les caractéristiques statistiques.

### V.3. PRESENTATION DES DIFFERENTES TECHNIQUES DE CLASSIFICATION ETUDIEES.

#### V.3.1. Assignation des observations au domaine modal le plus proche.

La première technique de classification considérée est simple. Les domaines modaux, qui sont définis soit par l'ensemble des points les constituant, soit par les points situés à leurs frontières, vont permettre de classer chaque observation de l'échantillon à la classe associée au domaine modal le plus proche. La notion de distance s'entend entre l'observation considérée et tous les points définissant les domaines ou leurs frontières. Le domaine modal le plus proche est celui contenant le point le plus proche de cette observation.

Les fonctions de décision ainsi que l'algorithme de classification peuvent être déduits directement de ce principe.

Soit  $N_k$  le nombre de points de  $\underline{M}$  caractérisant le domaine modal  $D_k$ , et soient  $X_{k,1}, X_{k,2}, \dots, X_{k,n}, \dots, X_{k,N_k}$  les points de  $\underline{M}$  constituant le domaine  $D_k$ , où  $X_{k,n}$  désigne le  $n$ -ième point du domaine  $D_k$ . Considérons un échantillon  $Y^*$  constitué de  $Q$  observations tel que :

$$Y^* = \{Y_1, Y_2, \dots, Y_q, \dots, Y_Q\}.$$

Pour toute observation  $Y_q$  de l'échantillon  $Y^*$ , les fonctions de décision  $d_k$  permettant la classification sont définies par :

$$d_k(Y_q) = - \underset{i=1}{\overset{N_k}{\text{Min}}} d(Y_q, X_{k,i}), \quad k = 1, 2, \dots, K,$$

où  $d(Y_q, X_{k,i})$  désigne la distance euclidienne de l'observation  $Y_q$  au point  $X_{k,i}$ .

La règle de décision est la même que celle que nous avons déjà définie précédemment, à savoir :

Décider  $C_S$  si  $d_S(Y_q) > d_k(Y_q), \forall k = 1, 2, \dots, K, k \neq S$ .

L'algorithme correspondant à ce principe est le suivant :

DEBUT

Pour chaque observation  $Y_q, q$  variant de 1 à  $Q$  Faire

$d_1(Y_q) = d(Y_q, X_{1,1})$

Pour chaque point  $X_{1,i}, i$  variant de 2 à  $N_1$  Faire

Si  $d_1(Y_q) > d(Y_q, X_{1,i})$  alors  $d_1(Y_q) = d(Y_q, X_{1,i})$  Fsi

Finpour

$d_1(Y_q) = -d_1(Y_q)$

$S = 1$

Pour chaque domaine  $D_k, k$  variant de 2 à  $K$  Faire

$d_k(Y_q) = d(Y_q, X_{k,1})$

Pour chaque point  $X_{k,i}, i$  variant de 2 à  $N_k$  Faire

Si  $d_k(Y_q) > d(Y_q, X_{k,i})$  alors  $d_k(Y_q) = d(Y_q, X_{k,i})$

Fsi

Finpour

$d_k(Y_q) = -d_k(Y_q)$

Si  $d_S(Y_q) < d_k(Y_q)$  alors  $S = k$  Finsi

Finpour

Affecter  $Y_q$  à la classe  $C_S$

Finpour

FIN

### V.3.2. Assignation des observations au centre de gravité du domaine modal le plus proche.

Une deuxième solution consiste à déterminer le centre de gravité de chaque domaine modal afin d'assigner chaque observation à la classe associée au centre de gravité le plus proche.

La définition des fonctions de décision nécessaires à l'implantation de la règle de décision précédemment décrite est alors quelque peu différente de celle de la méthode précédente.

Considérons le centre de gravité  $\bar{X}_k$  des  $N_k$  points constituant le domaine modal  $D_k$ . Chaque coordonnée  $\bar{X}_k^i$  du centre  $\bar{X}_k$  est définie par :

$$\bar{X}_k^i = (X_{k,1}^i + X_{k,2}^i + \dots + X_{k,N_k}^i) / N_k, i = 1, 2, \dots, N,$$

où  $X_{k,1}, X_{k,2}, \dots, X_{k,n}, \dots, X_{k,N_k}$  sont les points appartenant au domaine  $D_k$ .

Dans ces conditions, les fonctions de décision  $d_k$  permettant la classification de chaque observation  $Y_q$  de l'échantillon  $Y^*$  sont définies par :

$$d_k(Y_q) = - d(Y_q, \bar{X}_k), \quad k = 1, 2, \dots, K.$$

L'algorithme correspondant à ce principe est :

DEBUT

Pour chaque domaine  $D_k$ ,  $k$  variant de 1 à  $K$  Faire

    Déterminer  $\bar{X}_k$

Finpour

Pour chaque observation  $Y_q$ ,  $q$  variant de 1 à  $Q$  Faire

$S = 1$

$d_1(Y_q) = - d(Y_q, \bar{X}_1)$

Pour chaque centre de gravité  $\bar{X}_k$ ,  $k$  variant de 2 à  $K$  Faire

$d_k(Y_q) = - d(Y_q, \bar{X}_k)$

Si  $d_S(Y_q) < d_k(Y_q)$  alors  $S = k$  Finsi

Finpour

    Affecter  $Y_q$  à la classe  $C_S$

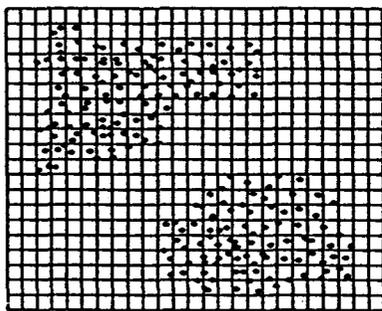
Finpour

FIN

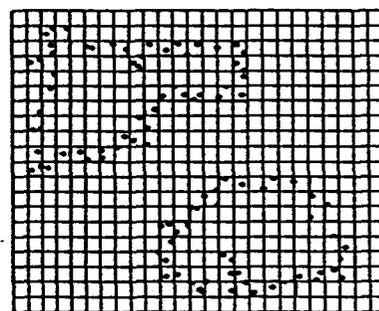
Les deux méthodes de classification que nous venons de présenter sont basées sur l'affectation des observations en fonction de leurs distances aux différents points des domaines modaux. Dans les paragraphes suivants, nous allons nous intéresser directement aux observations contenues dans les différents domaines modaux de l'ensemble discret qui constituent les noyaux des classes en présence.

### V.3.3. Assignment simple des observations au noyau le plus proche.

Cette technique de classification peut être décomposée en deux phases successives. La première étape consiste à affecter toutes les observations situées à l'intérieur de chaque domaine modal  $D_k$  à la classe  $C_k$  associée à ce domaine. Les prototypes des classes ainsi définis constituent les noyaux des classes en présence dans l'échantillon. Nous rappelons que les domaines modaux peuvent être décrits, soit par l'ensemble des points discrets les constituant, soit par leurs frontières. Le noyau  $N_k$  associé à la classe  $C_k$  est donc constitué, soit de l'ensemble des prototypes situés à l'intérieur du domaine modal  $D_k$ , soit de l'ensemble des prototypes inclus dans la frontière de ces domaines (cf. figure V.3).



(a)



(b)

*Figure V.3 : Définition des noyaux des classes en présence.*

*a - Noyaux définis par les domaines modaux.*

*b - Noyaux définis par les frontières modales.*

La deuxième étape de la procédure de classification consiste à assigner chacune des observations situées à l'extérieur des domaines modaux au noyau le plus proche. Notons que la distance d'une observation à un noyau  $N_k$  est définie par la distance euclidienne minimale de cette observation aux différents prototypes du noyau  $N_k$ .

La méthode de classification la plus connue permettant de classer les données à l'aide de prototypes est la méthode des  $k$  plus proches voisins [COV - 67]. Néanmoins, le temps de calcul nécessaire à la recherche des  $k$  plus proches voisins a conduit au

développement d'un ensemble d'algorithmes qui permettent soit de diminuer le nombre de prototypes nécessaires à la classification [CHA - 74] [ULL - 74] [YUN - 76], soit de supprimer les évaluations de distance inutiles [FUK - 75b] [NIE - 88].

Covert & Hart [COV - 67] ont montré que l'utilisation de la méthode du 1 plus proche voisin assure un taux d'erreur de classification ne pouvant pas excéder deux fois le taux d'erreur défini par la règle de Bayes. De plus, lors de l'utilisation de la méthode de classification par les k plus proches voisins, le choix de la valeur de k est souvent arbitraire [DUB - 90]. C'est pourquoi nous avons adopté la méthode du 1 plus proche voisin, la généralisation aux k plus proches voisins étant immédiate.

La technique de classification basée sur l'utilisation des prototypes des classes que nous présentons dans ce paragraphe a pour but d'affecter toute observation non encore classée à la classe dont le prototype est le plus proche. On considère l'observation concernée par chaque nouvelle affectation comme un nouveau prototype, et on itère la procédure jusqu'à ce que toutes les observations aient été affectées.

Plus précisément, soit  $NP_k$  le nombre de prototypes de la classe  $C_k$ , c'est-à-dire le nombre d'observations situées à l'intérieur du domaine  $D_k$  ou de sa frontière.

Les fonctions de décision à appliquer aux observations Y restant à classer sont définies par :

$$d_k(Y) = - \underset{i=1}{\text{Min}}^{NP_k} d(Y, Y_{k,i}), \quad i = 1, 2, \dots, K.$$

où  $Y_{k,i}$  désigne le i-ème prototype du noyau  $N_k$ .

La règle de décision à appliquer aux observations restant à classer est alors composée de deux assertions :

- Décider  $C_S$  si  $d_S(Y) > d_k(Y), \forall k = 1, 2, \dots, K, k \neq S$ .
- $NP_S = NP_S + 1$

L'algorithme associé à ce nouveau principe de classification est explicité ci-après.

DEBUT

Pour k variant de 1 à K Faire

Déterminer les  $NP_k$  prototypes de la classe  $C_k$

Finpour

$Nbre\_observations\_classees = NP_1 + NP_2 + \dots + NP_K$

Tantque ( $Nbre\_observations\_classees < Q$ ) Faire

Considérer une observation non classée  $Y$

$d_1(Y) = d(Y, Y_{1,1})$

Pour chaque point  $Y_{1,i}$ ,  $i$  variant de 2 à  $NP_1$  Faire

Si  $d_1(Y) > d(Y, Y_{1,i})$  alors  $d_1(Y) = d(Y, Y_{1,i})$  Fsi

Finpour

$d_1(Y) = -d_1(Y)$

$S = 1$

Pour chaque domaine  $D_k$ ,  $k$  variant de 2 à  $K$  Faire

$d_k(Y) = d(Y, Y_{k,1})$

Pour chaque point  $Y_{k,i}$ ,  $i$  variant de 2 à  $NP_k$  Faire

Si  $d_k(Y) > d(Y, Y_{k,i})$  alors  $d_k(Y) = d(Y, Y_{k,i})$

Fsi

Finpour

$d_k(Y) = -d_k(Y)$

Si  $d_S(Y_q) < d_k(Y_q)$  alors  $S = k$  Finsi

Finpour

Affecter  $Y_q$  à la classe  $C_S$

$NP_S = NP_S + 1$

$Nbre\_observations\_classees = Nbre\_observations\_classees + 1$

Fintantque

FIN

L'algorithme associé à cette méthode reste simple. Néanmoins, cette technique traite les observations restant à classer dans un ordre aléatoire, ce qui risque de provoquer des phénomènes de chaînage importants. La méthode que nous proposons maintenant permet de réduire considérablement ces phénomènes.

### **V.3.4. Assignment ordonnée des observations au noyau le plus proche.**

Cette technique est une amélioration de la technique précédente. En effet, nous commençons également par la définition des noyaux  $N_k$  associés à chacun des domaines  $D_k$ . Par contre, au lieu de prendre chacune des observations restant à classer dans un ordre aléatoire, nous allons chercher celle qu'il est préférable de classer en priorité. Il s'agit, à chaque nouvelle affectation, de déterminer l'observation la plus proche des prototypes des différents noyaux, de la classer dans la classe à laquelle appartient ce prototype, puis de la considérer comme un nouveau prototype, et ce, jusqu'à épuisement des observations restant à classer.

Compte tenu de ce principe, nous ne redécrivons, ni les fonctions de décision, ni les règles de décision associées, car ce sont les mêmes que pour la méthode d'assignation des observations au noyau le plus proche (paragraphe V.3.3).

De même, en ce qui concerne l'algorithme, il suffit de remplacer dans l'algorithme décrit au paragraphe V.3.3, l'action intitulée "Considérer une observation non classée Y" par l'action "Rechercher l'observation non classée la plus proche des prototypes", dans la boucle de classification des observations restant à classer.

## **V.4. EXEMPLES D'APPLICATION.**

Nous allons maintenant analyser les performances de ces quatre techniques de classification à l'aide de deux échantillons d'observations bidimensionnelles générées artificiellement.

### **V.4.1. Exemple 1.**

L'exemple 1 est l'exemple d'évaluation qui nous a permis d'illustrer les méthodes de détection des modes morphologiques présentées dans les chapitres III et IV. L'appartenance des observations à chacune des deux classes de cet exemple apparaît sur la figure V.4.

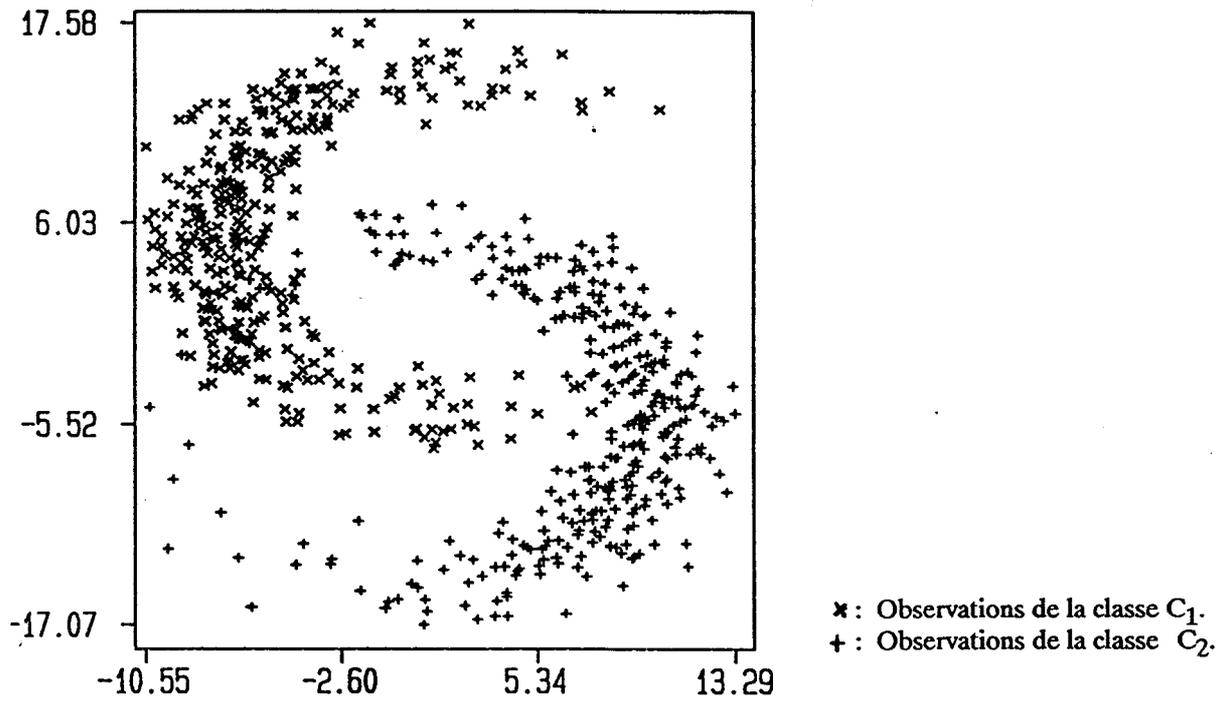


Figure V.4 : Appartenance des observations de l'exemple 1.

Les deux domaines modaux déterminés par une analyse morphologique de la connexité dans les régions modales (méthode CMEM) de l'ensemble discret résultant de la transformation de l'échantillon de l'exemple 1 sont présentés sur la figure V.5.

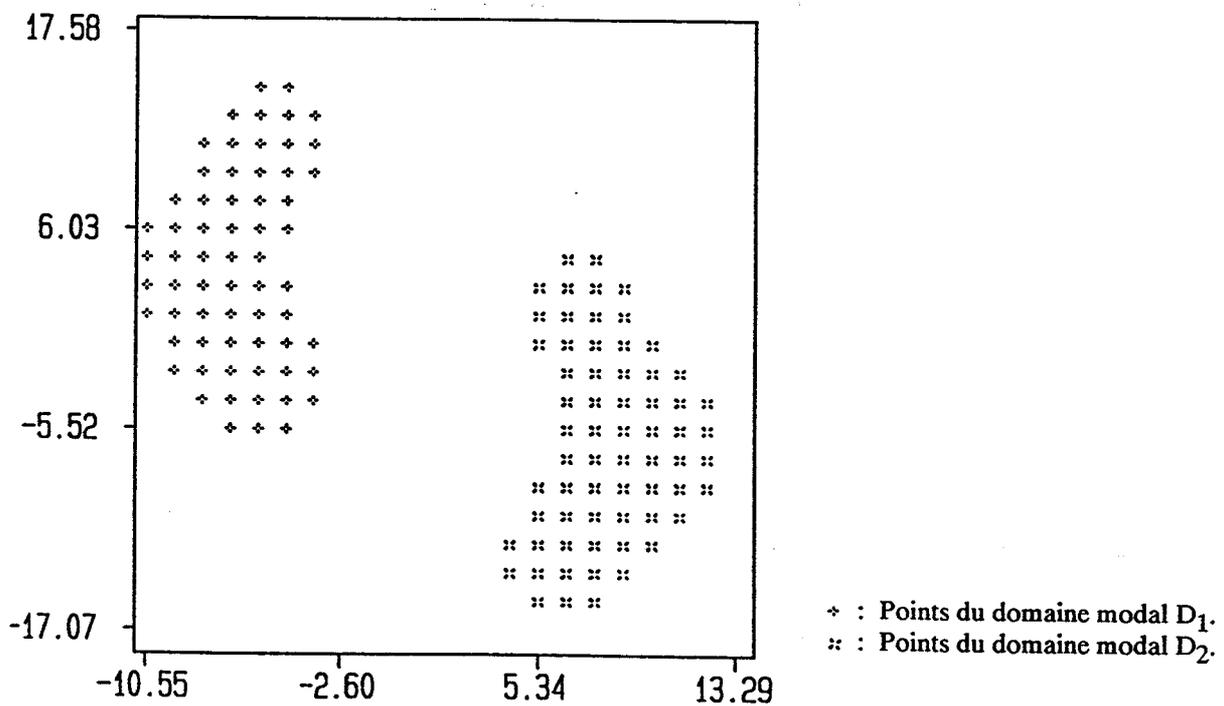


Figure V.5 : Les deux domaines modaux identifiés de l'exemple 1.

La codification utilisée pour distinguer les différentes techniques de classification est donnée dans la table ci-dessous.

CODE METHODE	INTITULE DE LA METHODE DE CLASSIFICATION
ADMP	<u>A</u> ssignation des observations au <u>D</u> omaine <u>M</u> odal le plus <u>P</u> roche.
ACDP	<u>A</u> ssignation des observations au <u>C</u> entre de gravité du <u>D</u> omaine modal le plus <u>P</u> roche.
ASNP	<u>A</u> ssignation <u>S</u> imple des observations au <u>N</u> oyau le plus <u>P</u> roche.
AONP	<u>A</u> ssignation <u>O</u> rdonnée des observations au <u>N</u> oyau le plus <u>P</u> roche.

*Table T5.1 : Codification des méthodes de classification.*

Les résultats de la classification des 700 observations de cet exemple illustrent bien les performances de chacune des quatre techniques utilisées (cf. figure V.6). En effet, lorsque les observations sont assignées au domaine modal le plus proche (technique ADMP), la forme géométrique des groupements d'observations obtenus reflète celle des domaines modaux. Notons que les domaines modaux ne recouvrent pas complètement la répartition spatiale des observations qui proviennent des deux classes en forme de croissants, surtout aux extrémités de ces derniers. Il en résulte des erreurs d'assignation notables au niveau de ces extrémités (cf. figure V.6.a).

Dans la figure V.6.b, le nombre d'observations mal classées situées aux extrémités des classes en forme de croissants est encore plus important. Ceci est dû au fait que le principe d'assignation des observations au centre de gravité du domaine modal le plus proche ne prend pas en compte la forme non sphérique des classes.

Dès qu'on affecte les observations à la classe du prototype le plus proche, en considérant chaque observation nouvellement classée comme un prototype, les observations sont mieux classées (figure V.6.c). Néanmoins, le résultat final dépend de

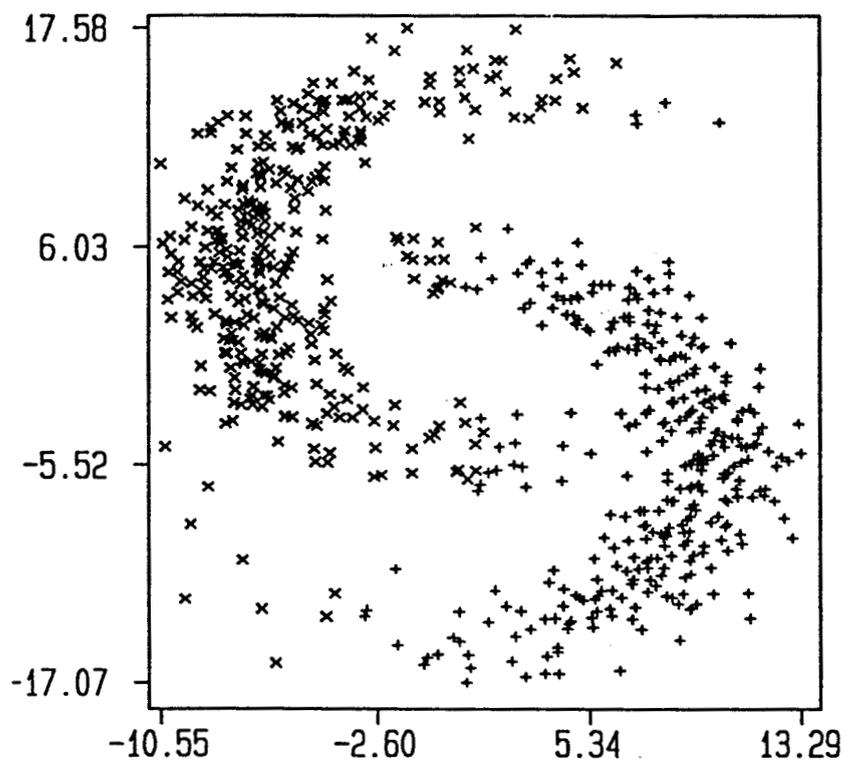
l'ordre dans lequel les observations sont prises en compte de telle sorte que les différentes régions de l'espace de représentation des données associées à chacune des classes ne peuvent pas être clairement délimitées. Il suffit de changer l'ordre de prise en compte des observations pour que la définition des régions soit différente. Par contre, si on ordonne l'assignation des observations, le résultat est nettement amélioré, comme l'illustre la figure V.6.d.

Toutes ces constatations sont confirmées par les taux d'erreur de classification associées à chacune de ces quatre techniques, ainsi que par l'analyse des matrices de confusion (cf. table T5.2). Dans ces matrices, les colonnes représentent les classes telles qu'elles ont été générées, les lignes indiquant la taille des classes obtenues après classification. Par exemple, l'élément d'une matrice se trouvant en ligne 2, colonne 1, représente le nombre d'observations assignées à la classe 2 alors qu'elles ont été générées dans la classe 1.

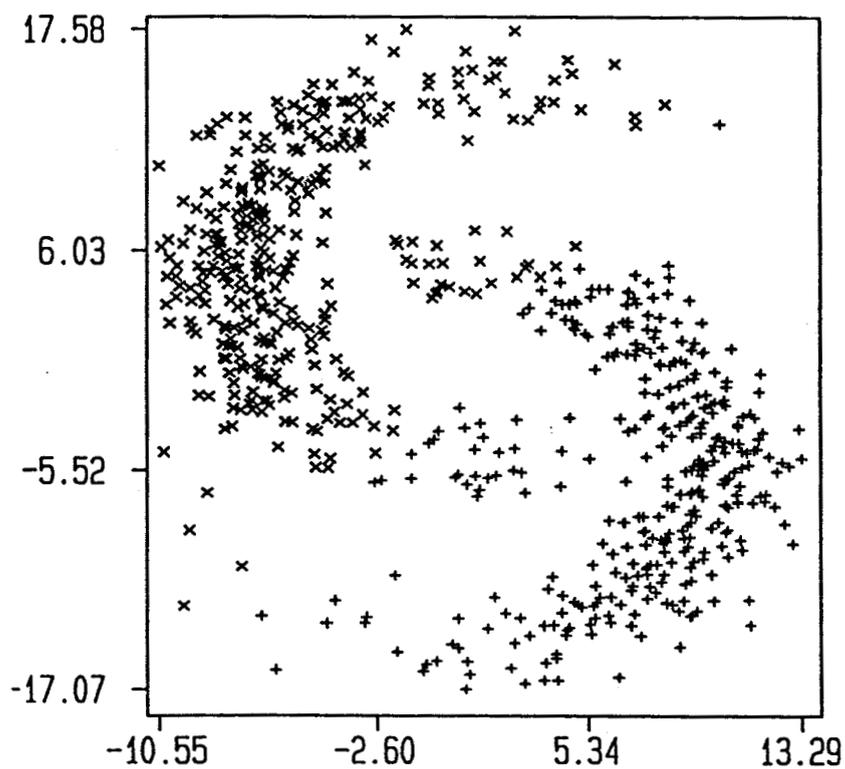
Le taux d'erreur le plus faible est effectivement obtenu, pour cet exemple, à l'aide de la méthode AONP assignant chaque observation au noyau le plus proche selon un ordre prédéfini.

METHODE DE CLASSIFICATION	TAUX D'ERREUR	MATRICE DE CONFUSION	
ADMP	7,57 %	319 31	22 328
ACDP	10,28 %	312 38	34 316
ASNP	5,00 %	319 31	4 346
AONP	2,14 %	340 10	5 345

*Table T5.2 : Résultats de la classification de l'exemple 1 par les quatre techniques de classification.*



(a)



(b)

*Figure V.6 : Observations classées de l'exemple 1 par les différentes techniques de classification.*

*b - Méthode ADMP.*

*a - Méthode ACDP.*

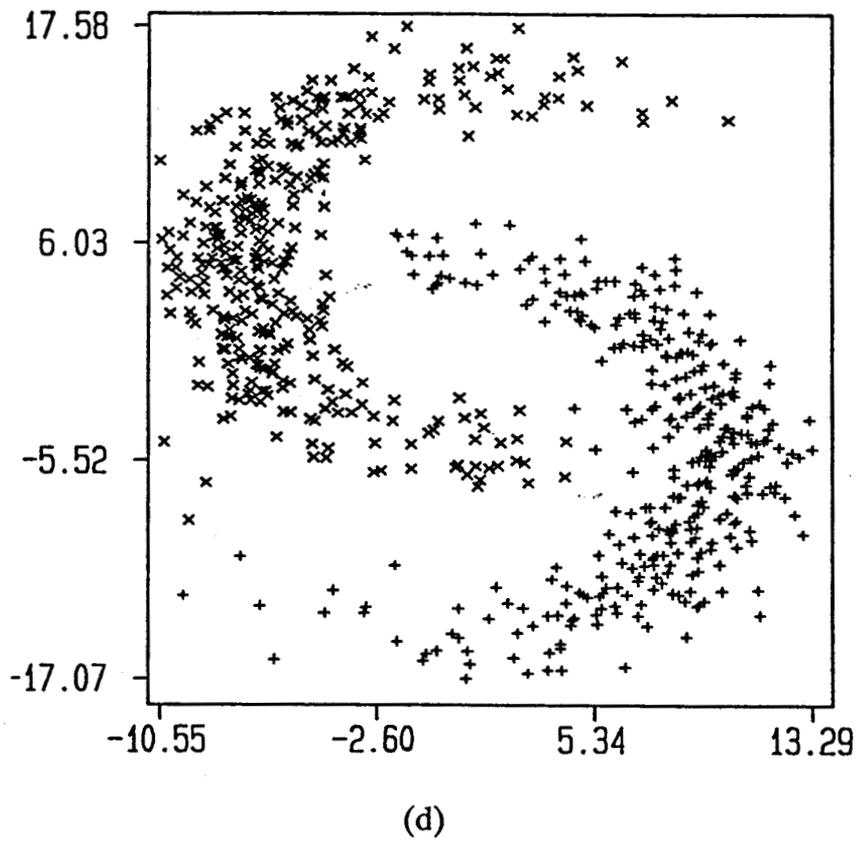
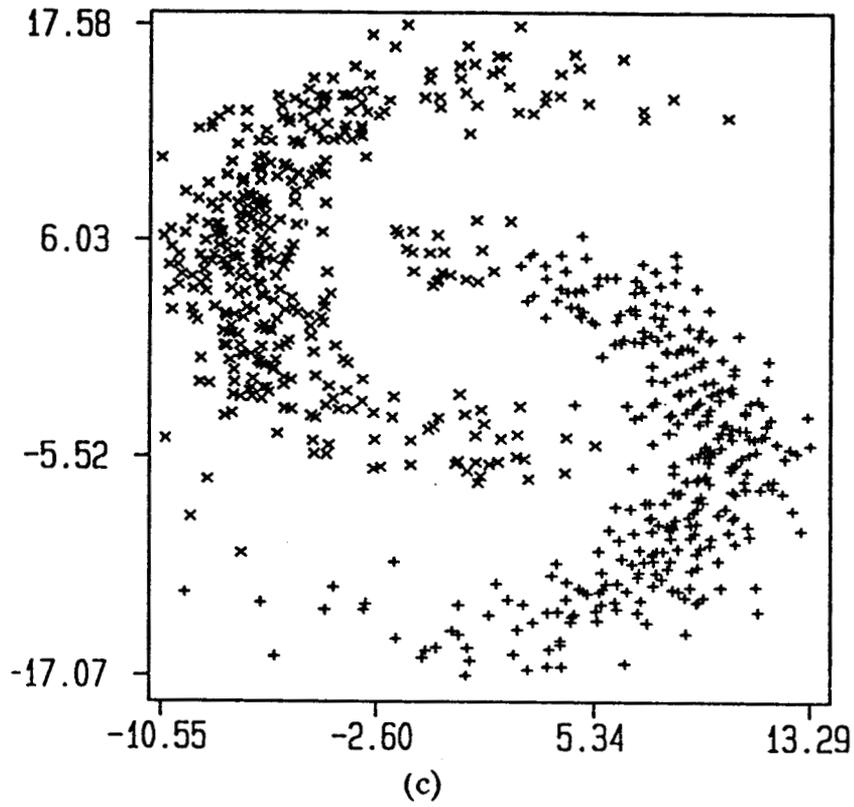


Figure V.6 : Observations classées de l'exemple 1 par les différentes techniques de classification (suite).

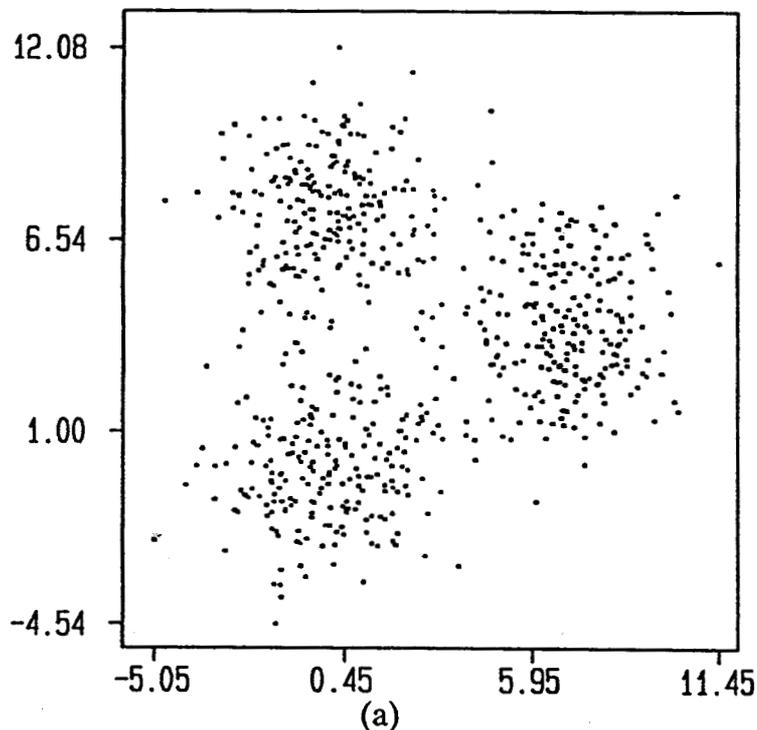
c - Méthode ASN.  
d - Méthode AON.

**V.4.2. Exemple 2.**

L'exemple 2 est composé de 660 observations réparties en trois classes de 220 observations normales chacune dont la représentation est illustrée figure V.7.a et dont les paramètres statistiques sont explicités dans la table T5.3. Les observations classées apparaissent sur la figure V.7.b.

	PROBABILITE A PRIORI	VECTEUR MOYENNE	MATRICE DE COVARIANCE	
POPULATION 1	0.33333	0.06147 -0.00983	3.24444 0.48030	0.48030 2.59250
POPULATION 2	0.33333	6.9796 4.0597	2.28603 -0.00609	-0.00609 3.18517
POPULATION 3	0.33333	0.00507 7.43925	2.54679 0.09868	0.09868 2.56584

*Table T5.3 : Paramètres statistiques de l'exemple 2.*



*Figure V.7 : Exemple 2 - Echantillon de 660 observations normales.*

*a - Représentation des données à analyser.*

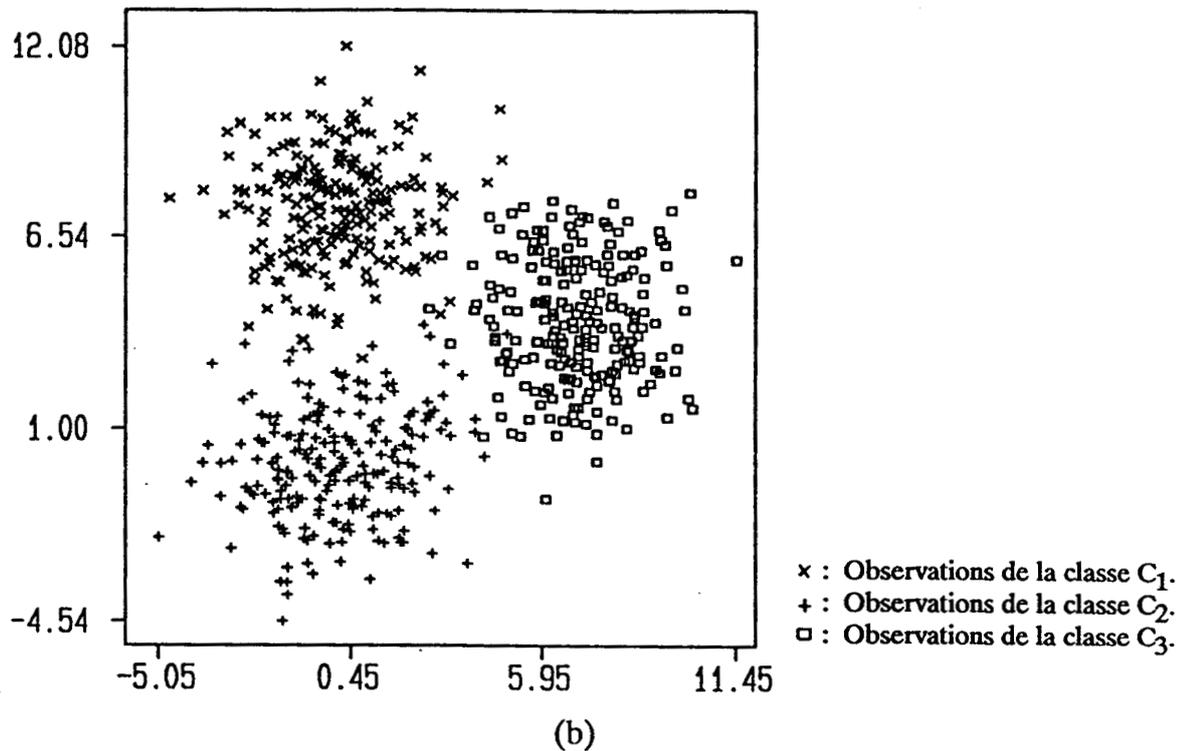


Figure V.7 : Exemple 2 - Echantillon de 660 observations normales (suite).

b - Appartenance des observations de l'exemple 2.

Pour ce second exemple, les trois domaines modaux déterminés par analyse morphologique de la connexité des points de l'ensemble discret situés dans les régions modales (méthode CMEM) sont présentés figure V.8.

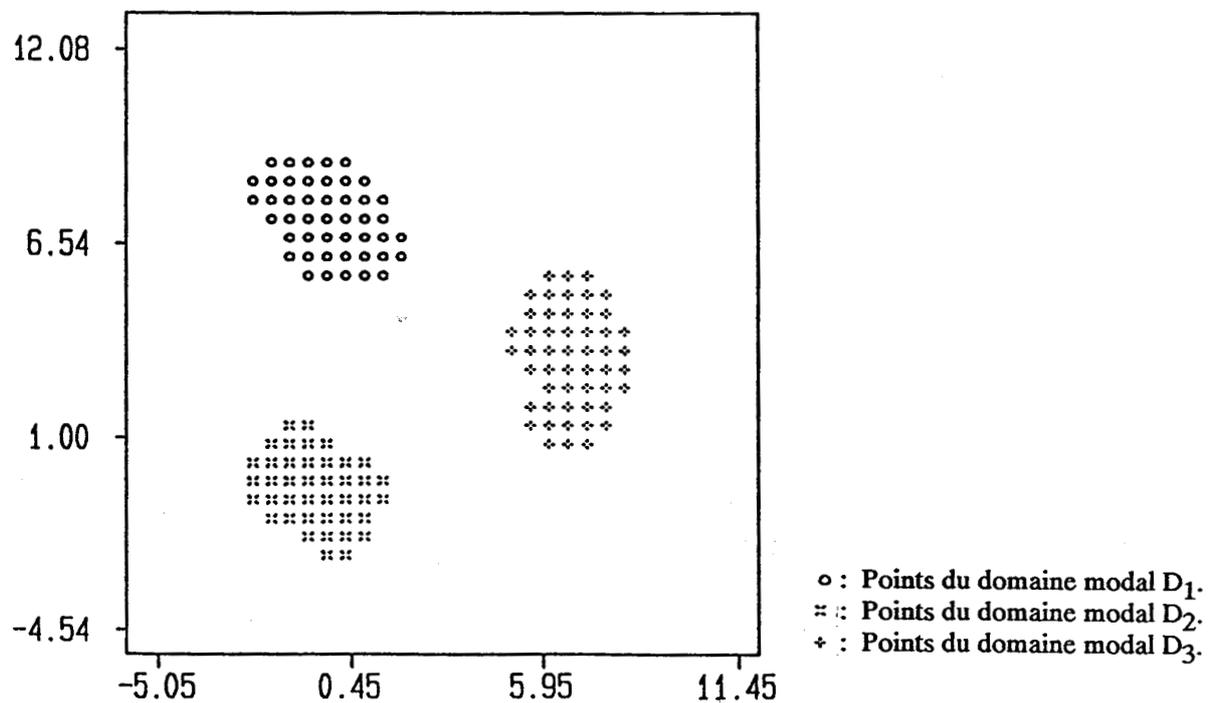


Figure V.8 : Ensemble  $M$  des modes morphologiques permettant la classification des observations de l'exemple 2 ( $R = 31$ ).

Les classes en présence dans cet échantillon étant de formes sphériques, les résultats de la méthode de classification basée sur l'affectation d'une observation au centre de gravité du domaine modal le plus proche (ACDP) sont relativement satisfaisants, comme le montre la figure V.9.b.

La méthode d'affectation des observations au domaine modal le plus proche (ADMP) donne, dans cet exemple, des résultats comparables à ceux de la technique ACDP (cf. figure V.9.a). Ces bonnes performances s'expliquent par le fait que, dans cet exemple particulier, la forme des domaines modaux reflète bien la forme sphérique des classes de l'échantillon.

La méthode ASNP n'assigne pas les observations situées à la frontière des deux classes représentées respectivement par des '+' et des '□', aux mêmes classes que les méthodes ADMP et ACDP (cf. figure V.9.c). Ceci est essentiellement dû au fait que les observations restant à classer sont prises en compte, dans cette technique ASNP, dans un ordre aléatoire. Il suffit donc qu'une première observation située à la frontière de ces deux classes soit affectée à la classe '+' pour que les observations voisines soient également affectées à cette classe par un effet de chaînage intempestif. Cet effet de chaînage est fortement réduit lorsqu'on utilise la technique AONP (cf. figure V.9.d).

L'examen de la table T5.4, qui indique les différents taux d'erreur ainsi que les matrices de confusion associés à chacune des quatre techniques de classification, montre que la méthode ADMP est la moins performante dans le cas de l'analyse de classes de formes sphériques. Nous précisons que les taux d'erreur obtenus par ces techniques sont à comparer au taux optimal de 1,67 % défini par la règle optimale de Bayes.

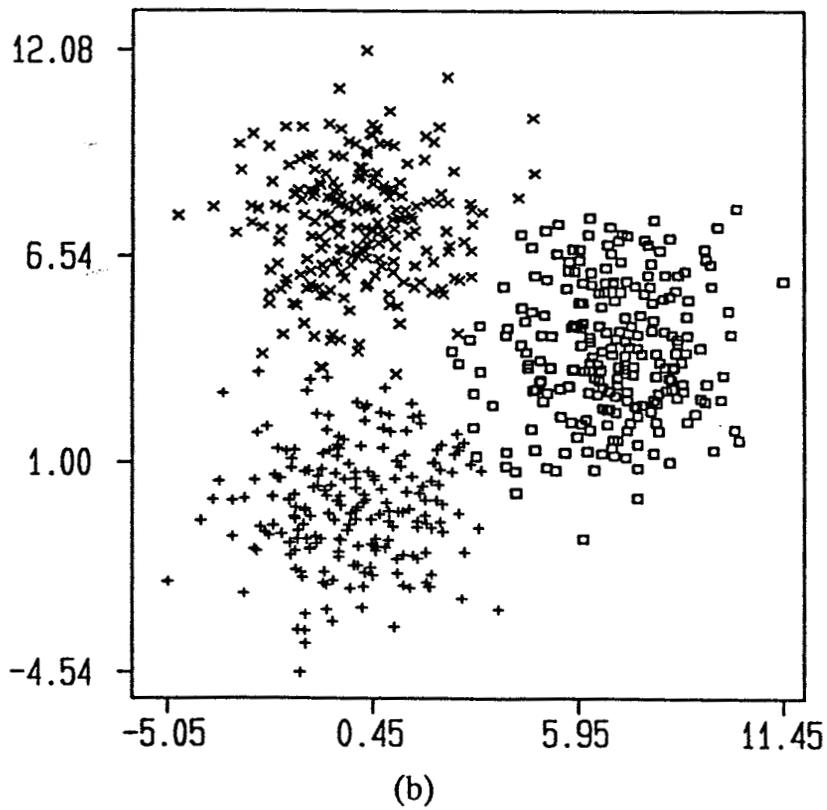
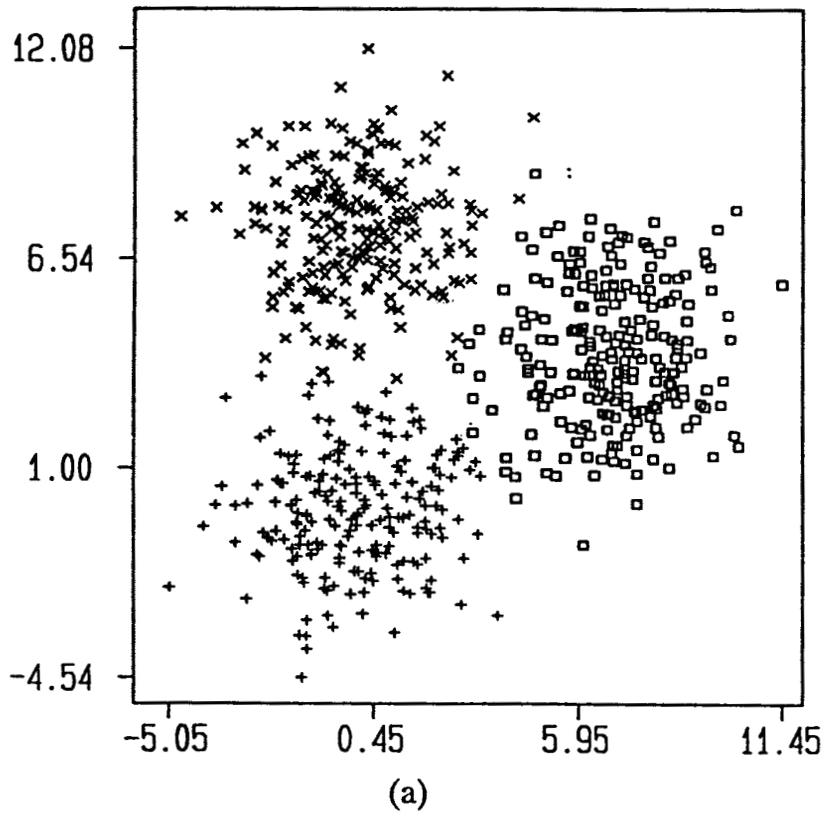
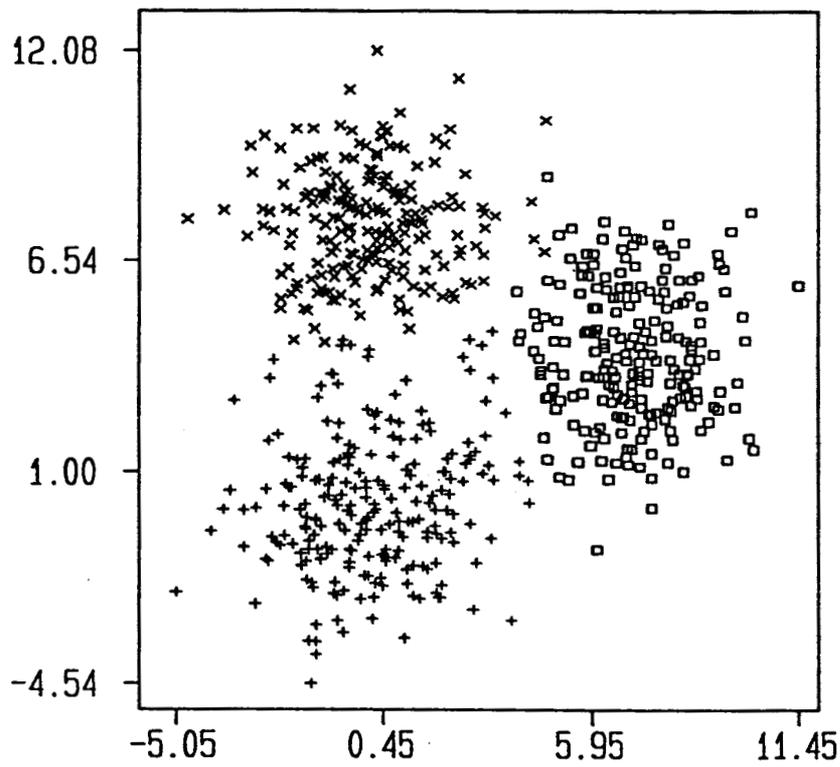
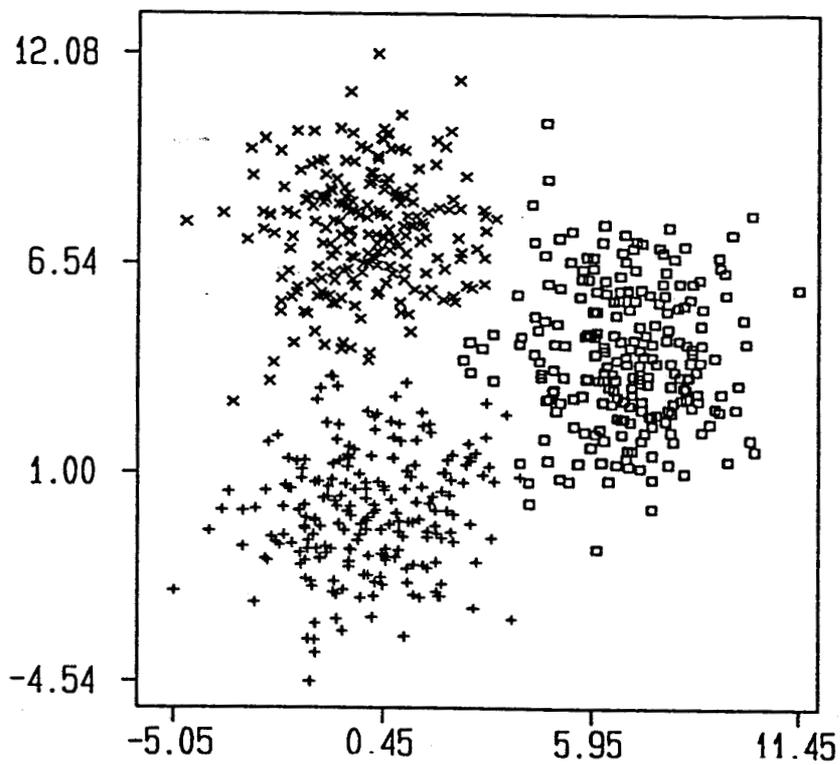


Figure V.9 : Observations classées de l'exemple 2 par les différentes techniques de classification.

a - Méthode ADMP.  
b - Méthode ACDP.



(c)



(d)

Figure V.9 : Observations classées de l'exemple 2 par les différentes techniques de classification (suite).

c - Méthode ASN.

d - Méthode AONP.

METHODE DE CLASSIFICATION	TAUX D'ERREUR	MATRICE DE CONFUSION		
ADMP	2,58 %	209 8 3	0 218 2	1 3 216
ACDP	2,58 %	208 10 2	0 210 2	1 2 217
ASNP	2,72 %	219 1 0	3 214 3	10 1 209
AONP	2,58 %	212 5 3	0 219 1	3 5 212

*Table T5.4 : Résultats de la classification de l'exemple 2 par les quatre techniques de classification.*

### **V.4.3. Conclusion.**

Dans ce paragraphe, nous avons présenté les résultats obtenus par les quatre techniques de classification proposées dans ce chapitre, en nous basant sur deux échantillons distincts. Nous avons ainsi mis en évidence, pour chacun des exemples, les différents taux d'erreur ainsi que les matrices de confusion obtenus à l'aide de ces quatre techniques.

Néanmoins, le choix d'une méthode de classification ne peut se faire sur le seul critère du taux d'erreur. En effet, toutes les procédures de classification que nous venons de décrire dans ce chapitre se basent sur l'ensemble discret  $\mathbf{M}$  constituant les domaines modaux associés aux différentes classes en présence dans l'échantillon à analyser. Or, ces domaines ont été déterminés après avoir choisi une méthode d'extraction des modes morphologiques parmi les quatre que nous avons proposées dans les chapitres III et IV, et après avoir fixé un pas de discrétisation unique. Il est important d'analyser le comportement des techniques de classification de ce chapitre vis à vis du choix du pas de discrétisation  $R$  et vis à vis du choix de la méthode de détermination des domaines modaux. De plus, les temps d'exécution nécessaires à la classification ont été évalués pour chaque technique afin d'évaluer la complexité des procédures proposées.

## V.5. COMPARAISON DES DIFFERENTES METHODES DE CLASSIFICATION.

### V.5.1. Variations du taux d'erreur en fonction du pas de discrétisation.

Le seul paramètre à ajuster par l'opérateur est le pas de discrétisation R. Nous allons montrer l'effet du choix de la valeur de ce paramètre sur le résultat final de la classification des observations. Dans ce but, nous avons relevé les différents taux d'erreur obtenus pour chacune des quatre techniques de classification, pour toutes les valeurs du pas de discrétisation R de la plage de stabilité, en nous appuyant sur les deux exemples précédemment utilisés dans ce chapitre. Le graphe de la figure V.10.a reflète les variations du taux d'erreur de classification en fonction du pas R pour l'exemple 1, tandis que celui de la figure V.10.b montre les différents taux d'erreurs en fonction de R pour l'exemple 2.

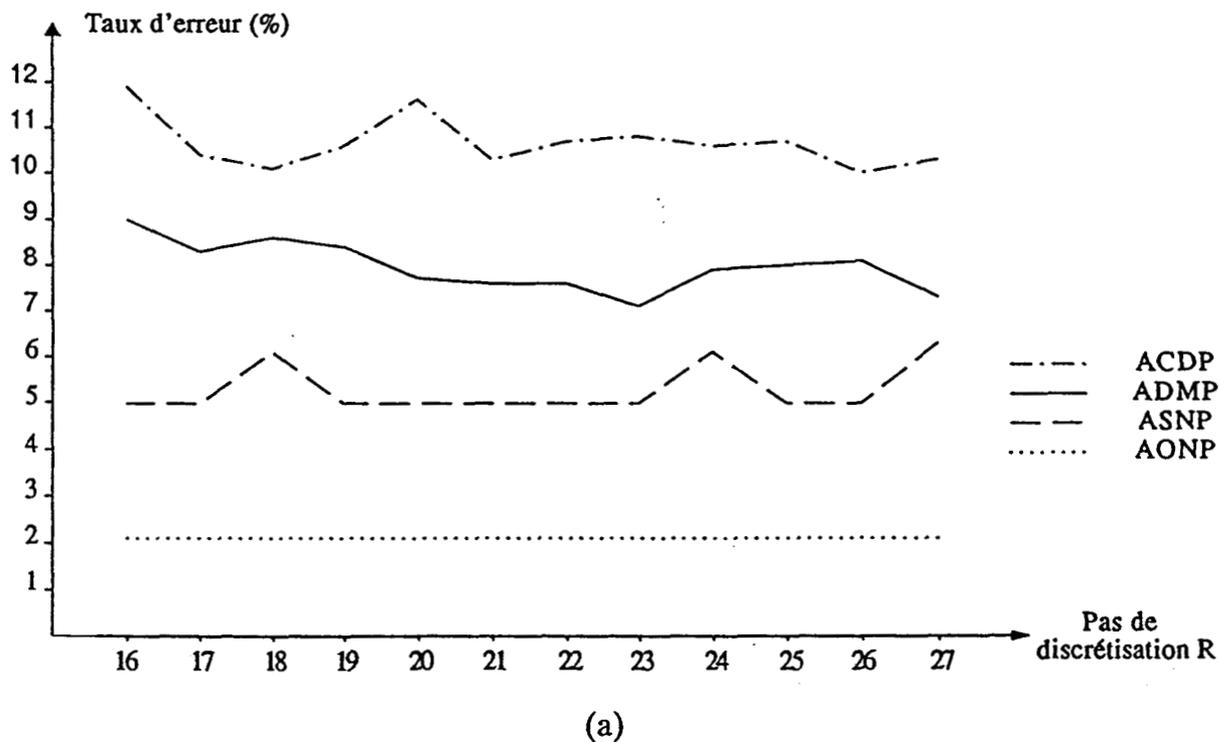
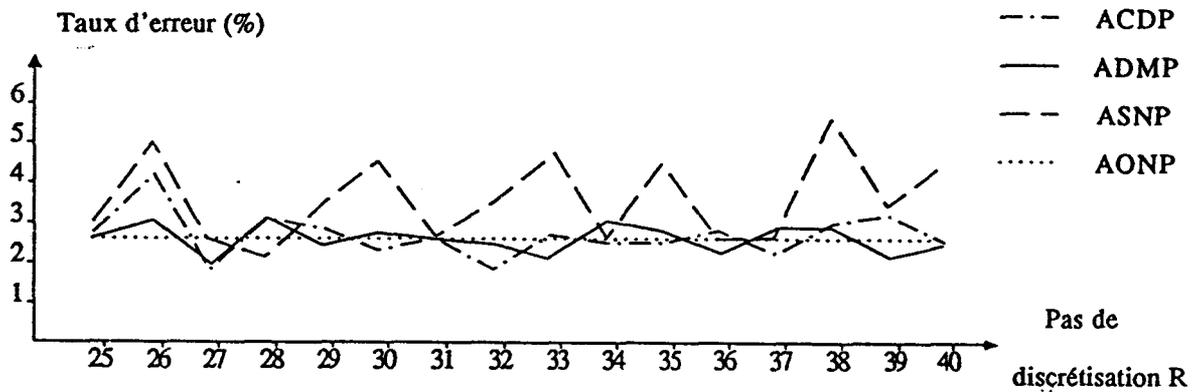


Figure V.10 : Variations du taux d'erreur de classification en fonction du pas de discrétisation R.

a - Résultats pour l'exemple 1.



(b)

Figure V.10 : Variations du taux d'erreur de classification en fonction du pas de discrétisation R (suite).

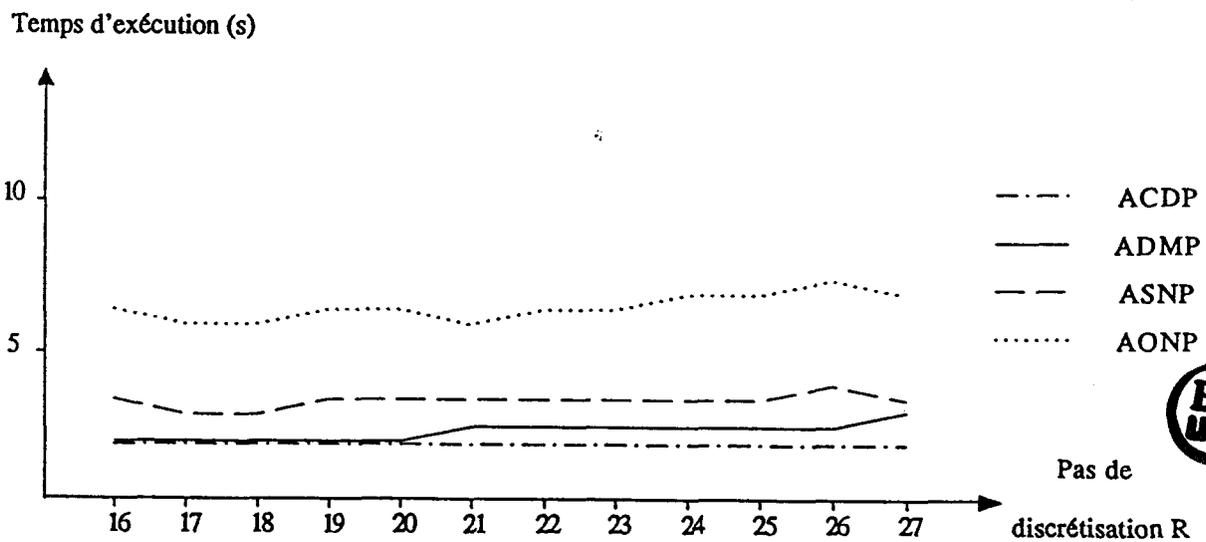
b - Résultats pour l'exemple 2.

Ces graphes montrent clairement que les résultats des techniques ADMP et ACDP sont très sensibles aux variations du pas de discrétisation R. En effet, pour deux valeurs de R distinctes, il peut exister plus de 2 % d'écart entre les taux d'erreur résultants. En ce qui concerne la technique ASNP, l'importance des variations du taux d'erreur dépend de la forme de l'échantillon. En effet, les résultats de cette technique sont relativement peu sensibles à l'ajustement de R dans le cas de l'exemple 1, mais très sensibles dans le cas de l'exemple 2. Enfin, la méthode AONP donne des résultats constants, indépendamment de la valeur du pas R et ceci pour l'exemple 1 comme pour l'exemple 2.

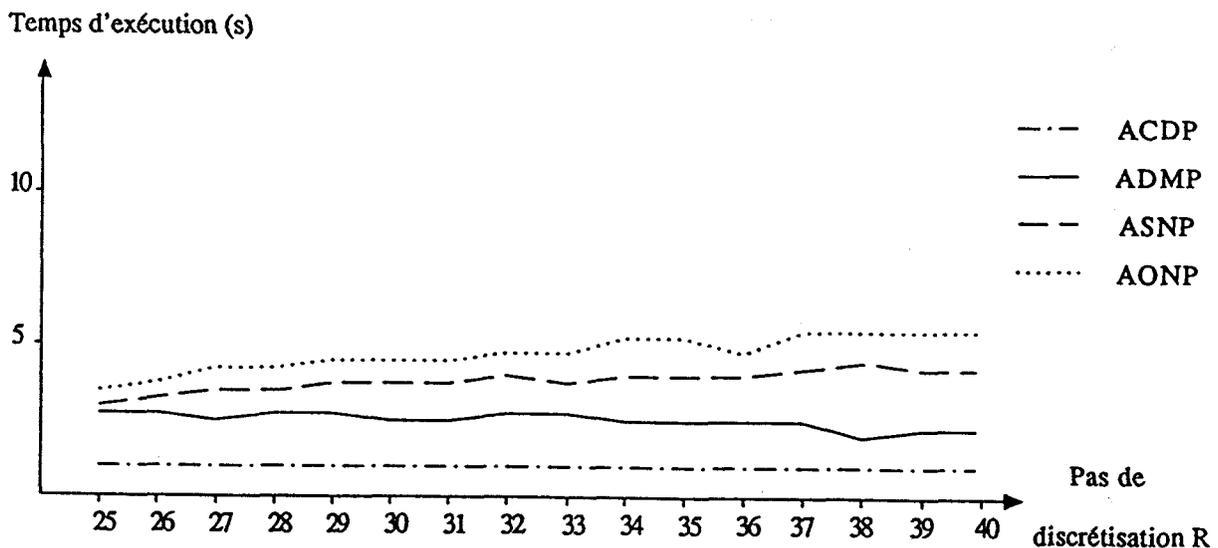
Par conséquent, si nous nous basons sur ce seul critère, nous opterons pour la méthode AONP. Nous allons maintenant voir s'il est nécessaire de nuancer ces résultats au regard des temps d'exécution.

**V.5.2. Comparaison en temps d'exécution.**

Un deuxième critère permettant d'évaluer ces méthodes est celui des temps d'exécution. Nous avons donc mesuré, pour tous les pas de discrétisation de la plage de stabilité, les temps d'exécution de la classification de l'échantillon avec chacune des quatre techniques. Le graphe de la figure V.11.a présente les temps de calculs nécessaires à la classification des observations de l'exemple 1, tandis que le graphe présenté figure en V.11.b illustre ces temps dans le cas de l'exemple 2.



(a)



(b)

Figure V.11 : Variation du temps d'exécution des différentes méthodes de classification en fonction du pas de discrétisation R.

a - Résultats pour l'exemple 1.

b - Résultats pour l'exemple 2.

Ces graphes illustrent bien le fait que la technique ACDP est la plus rapide et que la méthode AONP est la plus lente. De plus, comme la méthode ACDP est basée sur l'affectation des observations à la classe associée au centre de gravité du domaine modal le plus proche, le temps de calcul, pour un échantillon donné, reste insensible aux variations du pas de discrétisation. Par contre, les trois autres techniques (ADMP, ASNP et AONP) étant basées sur l'assignation des observations au domaine modal ou au noyau le plus proche, le temps d'exécution de chacune de ces techniques dépend soit du nombre de points constituant les domaines modaux (ou les frontières modales), soit du nombre de prototypes constituant les noyaux.

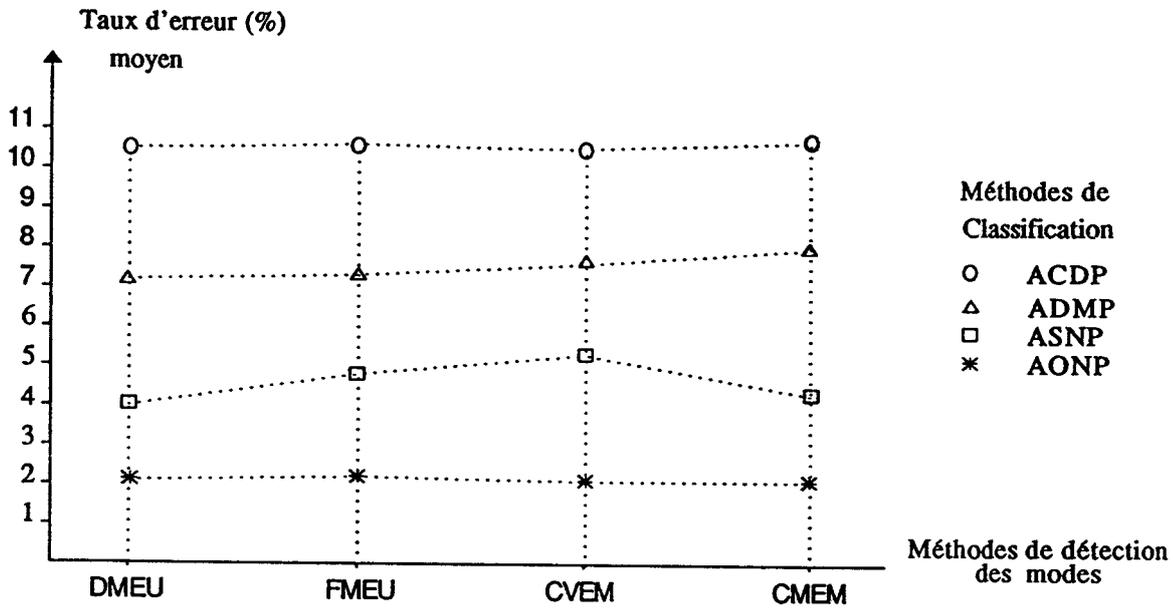
Malgré ces légères différences de temps d'exécution, aucune des méthodes de classification ne conduit à des volumes de calcul prohibitifs, de telle sorte que le critère "temps d'exécution" n'est pas déterminant pour le choix d'une méthode particulière.

### **V.5.3. Comparaison par rapport à la méthode de détection.**

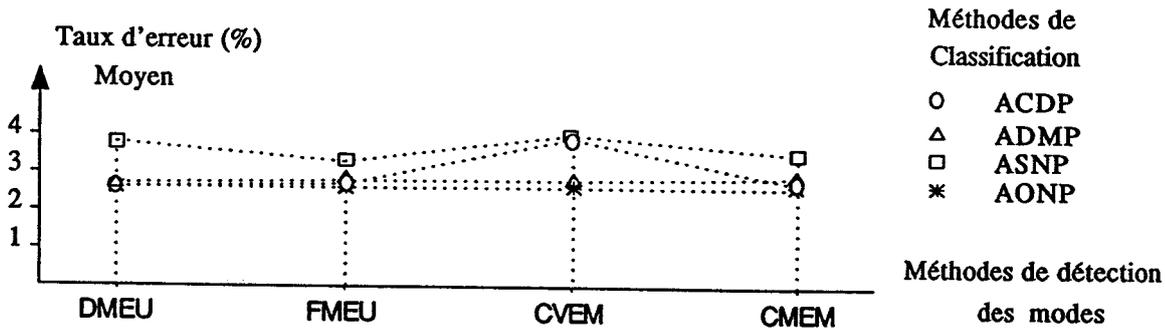
Nous allons maintenant examiner comment varient les taux d'erreur de classification en fonction de la méthode de détection des modes utilisée. Pour cela, nous avons, pour les deux exemples, réalisé la classification des observations par les quatre techniques, pour les quatre méthodes d'extraction des modes. La codification des méthodes d'extraction utilisée sur la figure V.12 est celle que nous avons déjà explicitée dans la table T4.1 du chapitre IV. Ces graphes indiquent les valeurs moyennes des taux d'erreur de classification obtenues pour chacune des techniques de classification dans l'intervalle de stabilité. Nous avons choisi de ne représenter que la valeur moyenne pour minimiser l'effet des variations du taux d'erreur en fonction de la valeur du pas de discrétisation.

Sur les deux graphiques, on peut remarquer que c'est la technique de classification AONP qui fournit le taux d'erreur le plus constant et le plus faible en valeur moyenne. De plus, on peut voir que les performances des quatre techniques de

classification se classent de la même manière au regard du taux d'erreur obtenu, et ce, quelle que soit la méthode de détection utilisée.



(a)



(b)

**Figure V.12 :** Variations de la valeur moyenne du taux d'erreur en fonction de la méthode d'extraction des modes morphologiques utilisée.

a - Résultats pour l'exemple 1.

b - Résultats pour l'exemple 2.

#### **V.5.4. Récapitulatif des résultats obtenus pour les exemples 1 et 2.**

Dans ce mémoire, nous présentons quatre méthodes de détection des modes ainsi que quatre techniques de classification. Nous avons illustré nos propos à l'aide de deux exemples distincts. Les plages de stabilité obtenues par les différentes méthodes de détection ainsi que les taux d'erreurs obtenus par les techniques de classification sont réunis dans la figure V.13 pour l'exemple 1 et dans la figure V.14 pour l'exemple 2.

On constate la cohérence des résultats qui mettent en évidence l'incidence des méthodes de classification et les performances comparables des quatre méthodes de détection des modes.

De plus, l'analyse des résultats montre que la qualité de la classification est d'autant plus difficile à assurer que les classes s'éloignent de la forme sphérique.

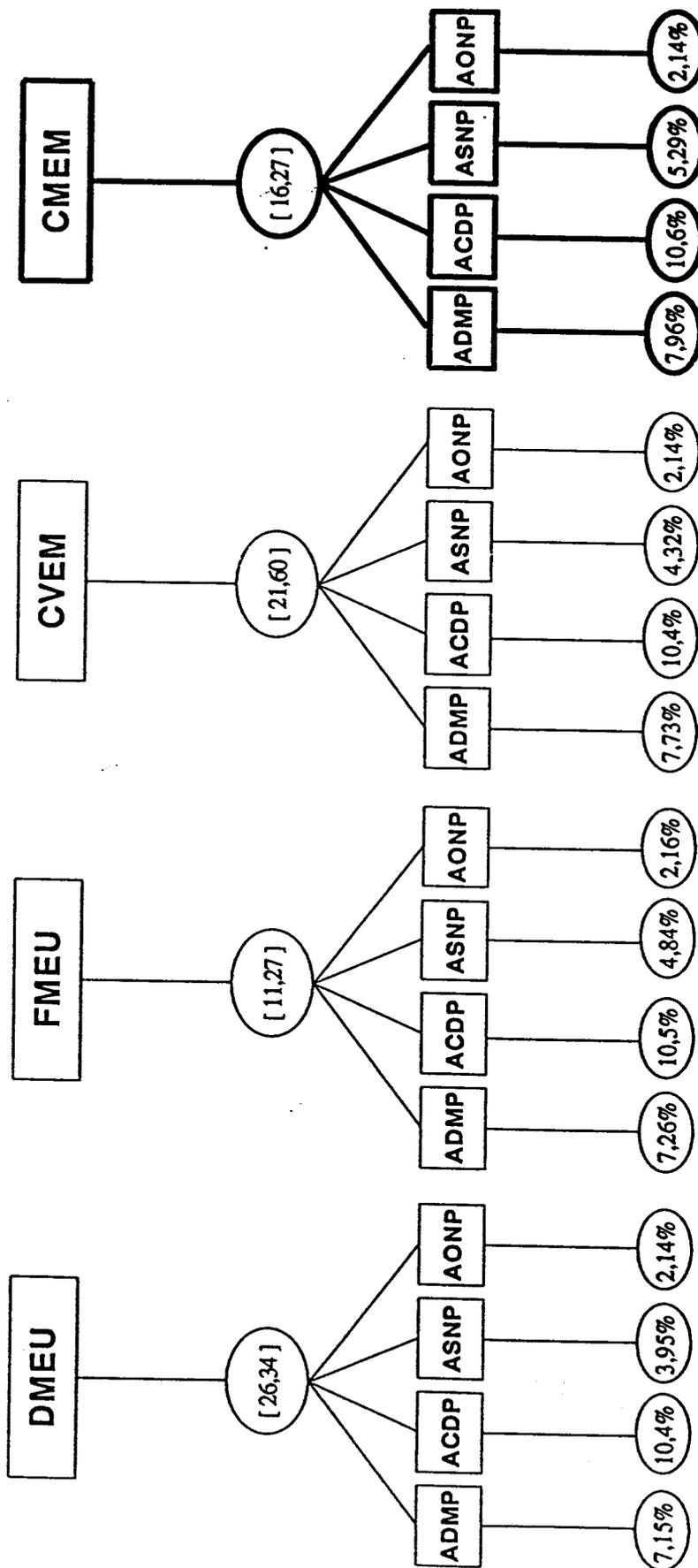


Figure V.13 : Récapitulatif des résultats obtenus pour l'exemple 1.

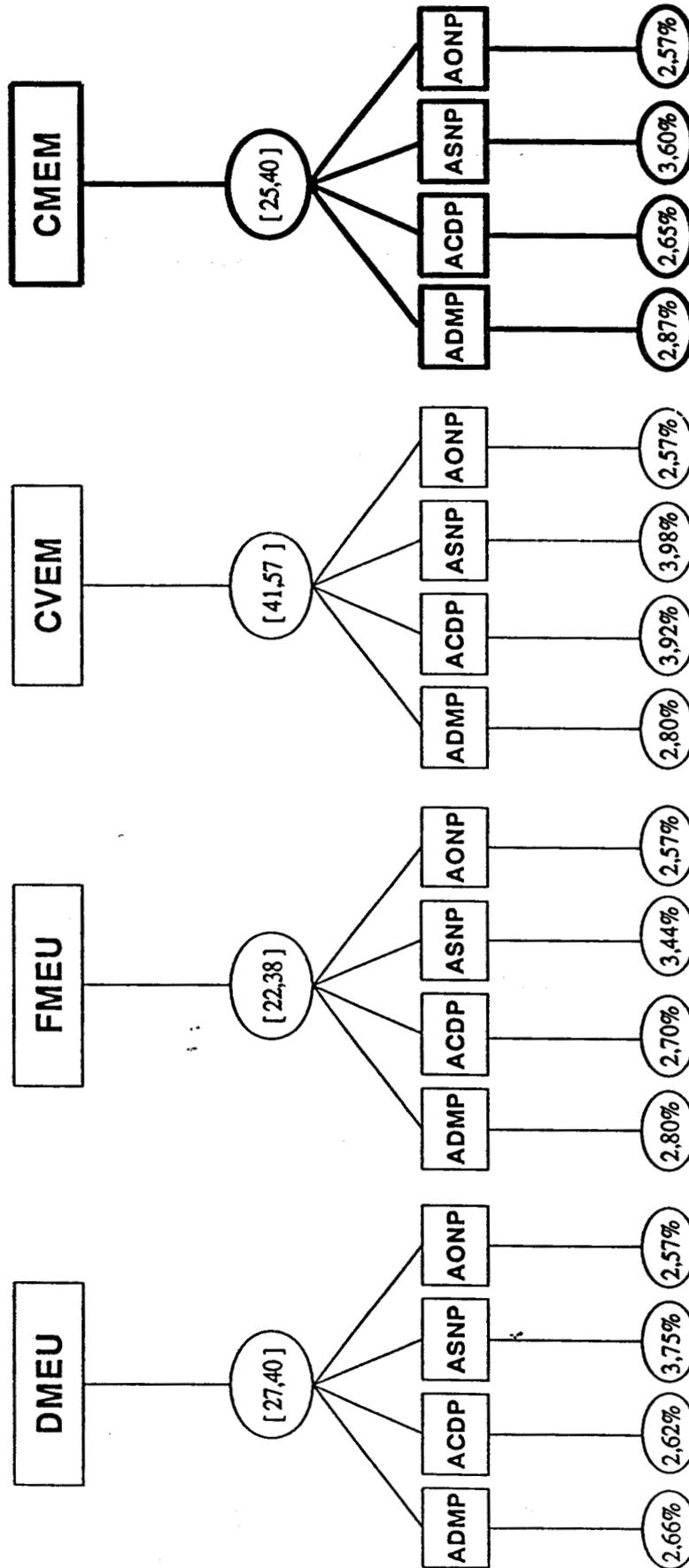


Figure V.14 : Récapitulatif des résultats obtenus pour l'exemple 2.

## V.6. CONCLUSION.

Dans ce chapitre, nous venons de présenter quatre techniques de classification, basées sur l'affectation de chaque observation à une classe donnée après détermination des domaines modaux par extraction morphologique. En nous appuyant sur deux exemples dont les observations définissent des classes de formes différentes, nous avons montré que les techniques ACDP et ADMP basées sur l'assignation des observations au domaine modal le plus proche permettaient d'obtenir des résultats satisfaisants dans le cas de classes de formes sphériques mais assez médiocres dans le cas de classes de formes quelconques. Par ailleurs, les résultats de la technique ASNP sont meilleurs dans le cas d'échantillons de formes non sphériques que dans le cas d'échantillons de formes sphériques.

La technique ASNP qui assigne de manière ordonnée chacune des observations au noyau le plus proche en réactualisant les noyaux à chaque nouvelle affectation, est la moins sensible aux variations du pas de discrétisation. De plus, le taux d'erreur résultant de la classification par cette technique ASNP est quasiment indépendant de la méthode de détection des modes employée. C'est pourquoi nous préconisons l'utilisation de cette technique bien que ce soit la plus lente, le but de l'analyse de données n'étant pas de fournir le résultat de la classification le plus rapidement possible mais le plus justement possible.

## **VI - RESULTATS EXPERIMENTAUX.**

## RESULTATS EXPERIMENTAUX.

### VI.1. INTRODUCTION.

Afin de montrer l'intérêt de l'approche morphologique dans le domaine de la classification automatique, nous considérons maintenant quelques exemples de données générés artificiellement. La simulation d'exemples bidimensionnels et multidimensionnels a pour but de faciliter l'analyse des résultats obtenus, le nombre de classes à mettre en évidence ainsi que leurs caractéristiques statistiques étant parfaitement connus.

De plus, afin de préciser les domaines de validité de cette méthodologie morphologique, nous allons comparer les résultats de cette approche à l'aide de deux méthodes classiques en analyse de données, à savoir les procédures dites "ISODATA" et "des K-MEANS" dont nous rappelons succinctement les principes généraux.

### VI.2. PRESENTATION GENERALE DE DEUX METHODES CLASSIQUES EN ANALYSE DE DONNEES.

#### VI.2.1. L'algorithme Isodata.

Le terme Isodata est le résultat de la contraction de "Iterative Self-Organizing Data Analysis Techniques A", la dernière lettre A ayant pour but de faciliter la prononciation du terme ISODATA [BAL - 67].

La méthode Isodata entre dans le cadre d'un ensemble de techniques métriques de partitions dites adaptatives. Dans l'ensemble de ces méthodes, on commence par

choisir arbitrairement  $K$  représentants  $R_1, R_2, \dots, R_k, \dots, R_K$  de  $K$  classes  $C_1, C_2, \dots, C_k, \dots, C_K$  définissant une représentation  $R$  et une partition  $C$  telles que :

$$R = \{ R_1, R_2, \dots, R_k, \dots, R_K \}$$

$$C = \{ C_1, C_2, \dots, C_k, \dots, C_K \}$$

Chacune des observations à classer est assignée à l'une des classes  $C_k$  selon une mesure de similarité entre chaque représentant  $R_k$  et l'observation considérée. La partition  $C$  ainsi obtenue est utilisée pour redéfinir une nouvelle représentation  $R$  [DID - 82].

Ce processus itératif, qui inclut les deux phases de définition de la partition  $C$  et la réactualisation de la représentation  $R$ , prend fin lorsqu'un critère mesurant l'adéquation de la représentation  $R$  à la partition  $C$  est optimisé.

Dans la méthode Isodata que nous présentons dans ce paragraphe, le représentant de chaque classe  $C_k$  est défini par son centre  $\bar{Y}_k$ , et la mesure de similarité utilisée est la distance euclidienne de chacune des observations aux différents centres  $\bar{Y}_k$ ,  $k = 1, 2, \dots, K$ . L'algorithme général de cette technique peut être résumé en six points principaux :

- 1 - Choisir arbitrairement le nombre  $K$  de classes, ainsi que les centres  $\bar{Y}_k$  de ces  $K$  classes,  $k = 1, 2, \dots, K$ .
- 2 - Assigner chaque observation à la classe associée au centre le plus proche au sens de la distance euclidienne.
- 3 - Si une classe ne contient pas suffisamment d'observations, la supprimer, aller en 2.
- 4 - Recalculer les centres  $\bar{Y}_k$  des classes  $C_k$  compte tenu de la nouvelle partition définie en 2.
- 5 - Calculer la norme de la matrice de covariance de chaque classe et le

comparer à la valeur autorisée. Si la norme de la matrice de covariance d'une classe  $C_k$  est trop grand et que la distance moyenne des observations de cette classe à son centre  $\bar{Y}_k$  est supérieure à la valeur moyenne des distances des observations des différentes classes à leurs centres respectifs, scinder la classe  $C_k$  considérée en deux classes distinctes et aller en 2.

- 6 - Calculer toutes les distances séparant les différents centres les uns des autres. Si certaines de ces distances sont plus faibles que la distance minimum autorisée séparant deux centres, regrouper les deux classes considérées en une seule, et aller en 2.

Cet algorithme, que nous venons de décrire dans sa version de base, nécessite la spécification de nombreux paramètres tels que le nombre de classes, les valeurs des différents seuils de scission et de regroupement, les coordonnées des différents centres initiaux. Les performances de cette procédure dépendent fortement des valeurs de ces paramètres prédéfinies par l'opérateur. C'est pourquoi de nombreux auteurs ont cherché à minimiser le nombre de paramètres nécessaires à la mise en oeuvre de cet algorithme en exploitant des informations pouvant être extraites des observations elles-mêmes [FOR - 74] [DAV - 79] [TOU - 79].

En ce qui nous concerne, nous avons utilisé cette technique dans une version simplifiée, le but de cette utilisation n'étant pas de scinder les classes ni de les regrouper mais seulement d'obtenir une partition  $C$  de l'échantillon en différentes classes afin de fournir une base d'évaluation universellement reconnue pour analyser les résultats obtenus par l'approche morphologique.

Nous avons donc donné au paramètre  $K$  la valeur correspondant au nombre exact de classes constituant chacun des échantillons soumis à l'analyse. Les  $K$  centres ont également été placés très près des véritables centres des classes. Nous verrons que, même dans ces conditions privilégiées, les résultats obtenus par cette technique peuvent être moins bons que ceux résultant de l'approche morphologique.

La version simplifiée de l'algorithme Isodata que nous avons utilisée peut être résumée de la façon suivante :

- 1 - Donner le nombre réel  $K$  de classes constituant l'échantillon et indiquer  $K$  centres  $\bar{Y}_k$  proches de ceux des  $K$  classes,  $k = 1, 2, \dots, K$ .
- 2 - Classifier les observations de l'échantillon en les assignant à la classe associée au centre le plus proche.
- 3 - Recalculer les centres  $\bar{Y}_k$  des classes  $C_k$  compte tenu de la nouvelle partition définie en 2.
- 4 - Si l'un des centres a changé, aller en 2, sinon arrêter.

### VI.2.2. L'algorithme des K-Means.

La deuxième technique que nous avons utilisée pour évaluer les différents résultats de l'approche morphologique est également une technique métrique de partitionnement adaptatif basée sur l'optimisation de l'adéquation d'une représentation  $R$  à une partition  $C$  composée de  $K$  classes.

Dans cette technique dite des K-Means qui est aussi connue sous le nom de procédure de minimisation de l'erreur quadratique [DUD - 73], la représentation  $R_k$  de chaque classe  $C_k$  est définie par son centre  $\bar{Y}_k$ , et la mesure de similarité retenue est la distance euclidienne  $d(Y_q, \bar{Y}_k)$  de chacune des observations  $Y_q$  de l'échantillon aux différents centres  $\bar{Y}_k$ . Ainsi, chaque coordonnée  $i$  du centre  $\bar{Y}_k$  des  $Q_k$  observations de la classe  $C_k$  est définie par :

$$Y_k^i = \frac{1}{Q_k} * \sum_{q=1}^{Q_k} Y_q^i$$

Le critère à optimiser est défini par :

$$J(C) = \sum_{k=1}^K \sum_{q=1}^{Q_k} d(Y_q, Y_k)^2 = \sum_{k=1}^K J_k(C)$$

La partition  $C^*$  qui optimise ce critère est définie par :

$$J(C^*) = \min_C J(C) .$$

L'algorithme général de la méthode que nous avons utilisée est la suivante :

- 1 - Donner le nombre réel  $K$  de classes, ainsi qu'une partition  $C$  des  $Q$  observations de l'échantillon en  $K$  classes.
- 2 - Calculer  $J(C)$ , et  $\bar{Y}_k$ ,  $k = 1, 2, \dots, K$ .
- 3 - Pour chaque observation  $Y_q$  affectée initialement à la classe  $C_i$ ,  $q$  variant de 1 à  $Q$ , faire :
  - \* Calculer  $d_k(Y_q)$ ,  $\forall k = 1, 2, \dots, K$ .
  - \* Affecter  $Y_q$  à la classe  $C_S$  telle que :
$$d_S(Y_q) \leq d_k(Y_q), \forall k = 1, 2, \dots, K.$$
  - \* Calculer  $\bar{Y}_i$  et  $\bar{Y}_S$ .
  - \* Calculer  $J(C) = J(C) - d_i(Y_q) + d_S(Y_q)$

Finpour.

- 4 - Si  $J(C)$  a changé durant la phase 3 retour au 3, sinon arrêter.

Dans cet algorithme, les fonctions de décision  $d_k$  permettant l'assignation des observations aux différentes classes sont définies par :

$$d_k(Y_q) = \frac{Q_k}{Q_k + 1} \left\| Y_q - \bar{Y}_k \right\|^2, \quad k \neq i$$

$$d_k(Y_q) = \frac{Q_i}{Q_i - 1} \left\| Y_q - \bar{Y}_i \right\|^2, \quad k = i$$

Nous avons également utilisé cet algorithme dans cette version de base, son exploitation se limitant à la comparaison de la partition obtenue  $C$  avec celle définie par la technique morphologique.

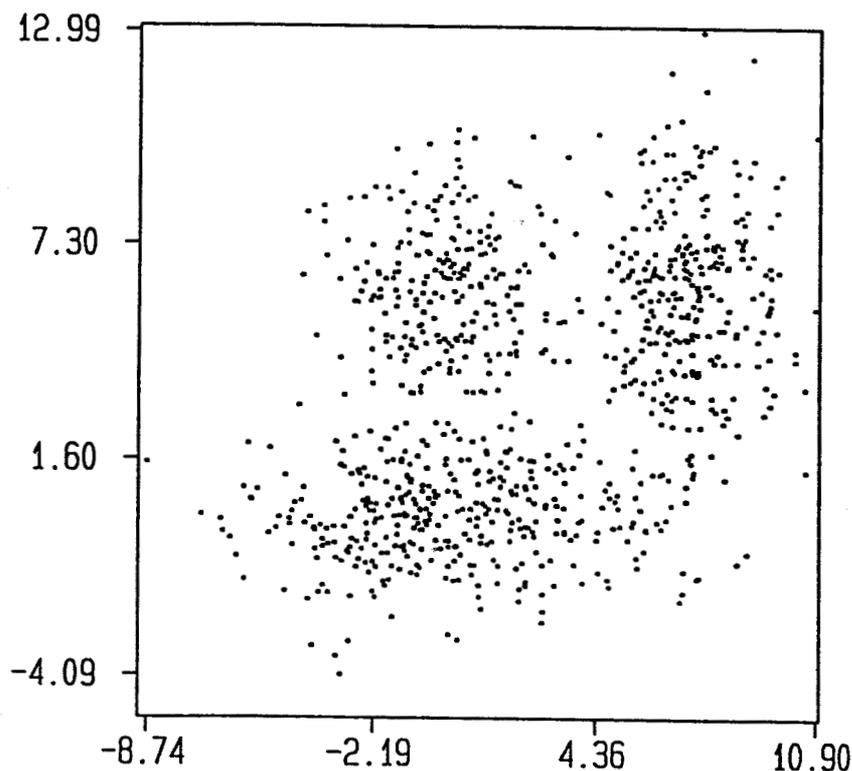
### VI.3. EXEMPLE BIDIMENSIONNEL COMPOSE DE TROIS CLASSES NORMALES NON SPHERIQUES.

Les données utilisées dans ce premier exemple sont des observations issues de

trois classes normales non sphériques et non équiprobables dont les paramètres statistiques ainsi que le nombre d'observations par classe sont précisés dans la table T6.1. La figure VI.1 présente l'échantillon tel qu'il a été simulé (exemple VI.1.a).

	NOMBRE D'OBSERVATIONS	VECTEUR MOYENNE	MATRICE DE COVARIANCE	
POPULATION 1	350	0.0277 0.0073	9.1639 0.0717	0.0717 1.4571
POPULATION 2	300	7.0393 5.9115	1.9541 0.0670	0.0670 5.7348
POPULATION 3	250	-0.0711 6.0252	2.6771 -0.0052	-0.0052 3.5141

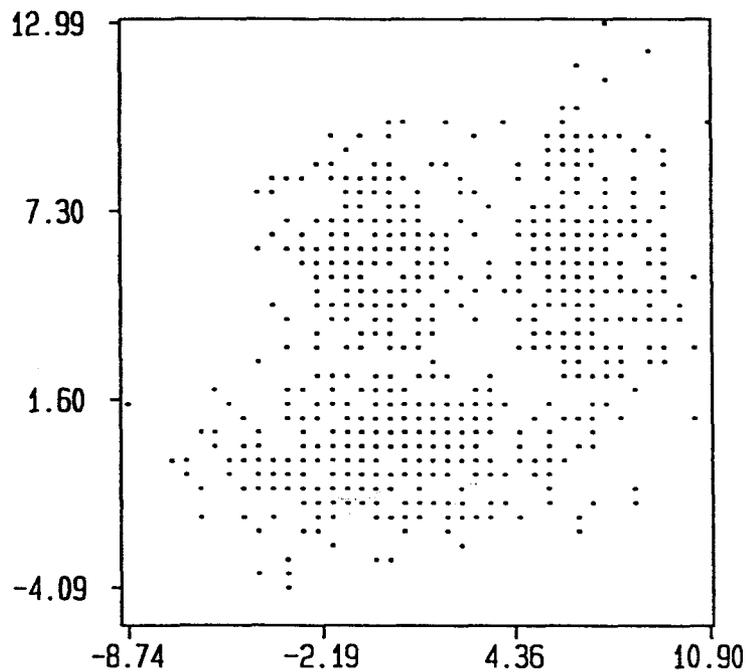
*Table T6.1 : Paramètres statistiques de l'exemple VI.1.  
(Paramètres recalculés après génération des données)*



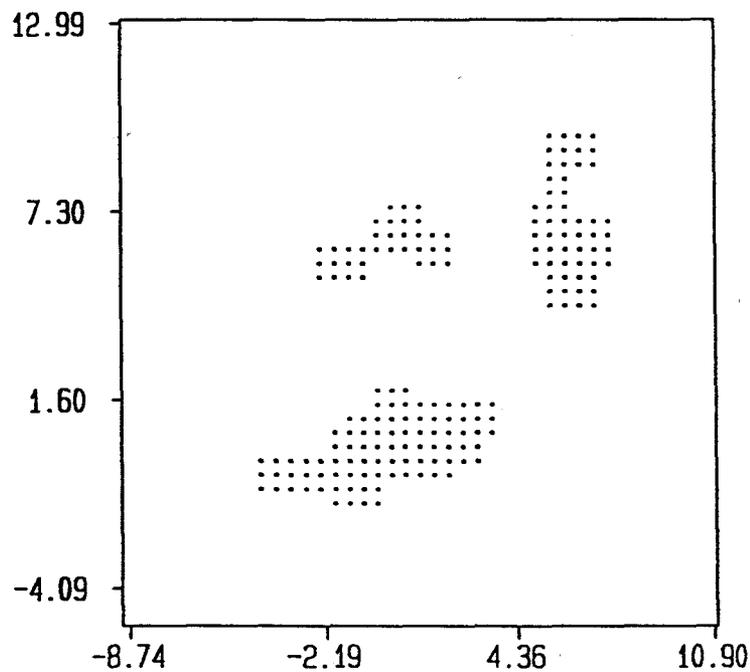
*Figure VI.1 : Représentation graphique de l'échantillon de l'exemple VI.1.a.*

Nous avons, pour cet exemple, déterminé les différents modes morphologiques

obtenus par chacune des quatre méthodes d'extraction morphologique. La figure VI.2 présente l'ensemble discret ainsi que les domaines modaux déterminés par la méthode DMEU, le pas retenu étant situé au milieu de la plage de stabilité.



(a)



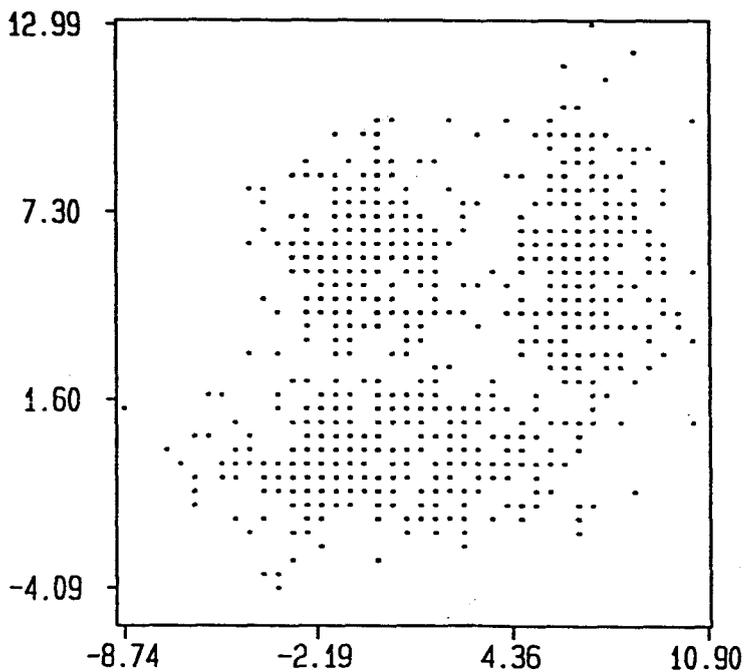
(b)

*Figure VI.2 : Détermination des domaines modaux par la méthode DMEU pour l'exemple VI.1.a.*

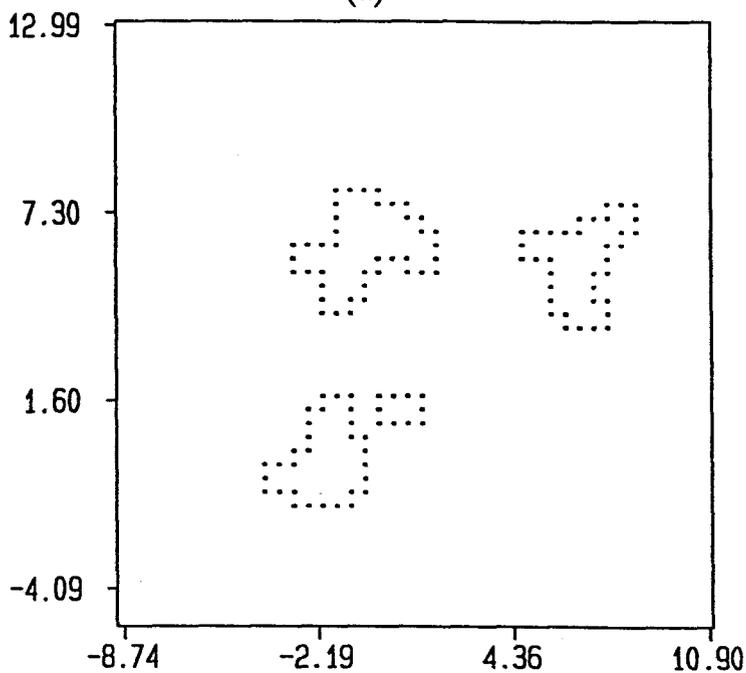
*a - Ensemble discret ( $R = 40$ ).*

*b - Les trois domaines modaux.*

La figure VI.3 présente l'ensemble discret ainsi que les frontières modales déterminées par la méthode FMEU, la figure VI.4 présente l'ensemble discret et les domaines modaux obtenus par la méthode CVEM, tandis que la technique CMEM est illustrée en figure VI.5.



(a)

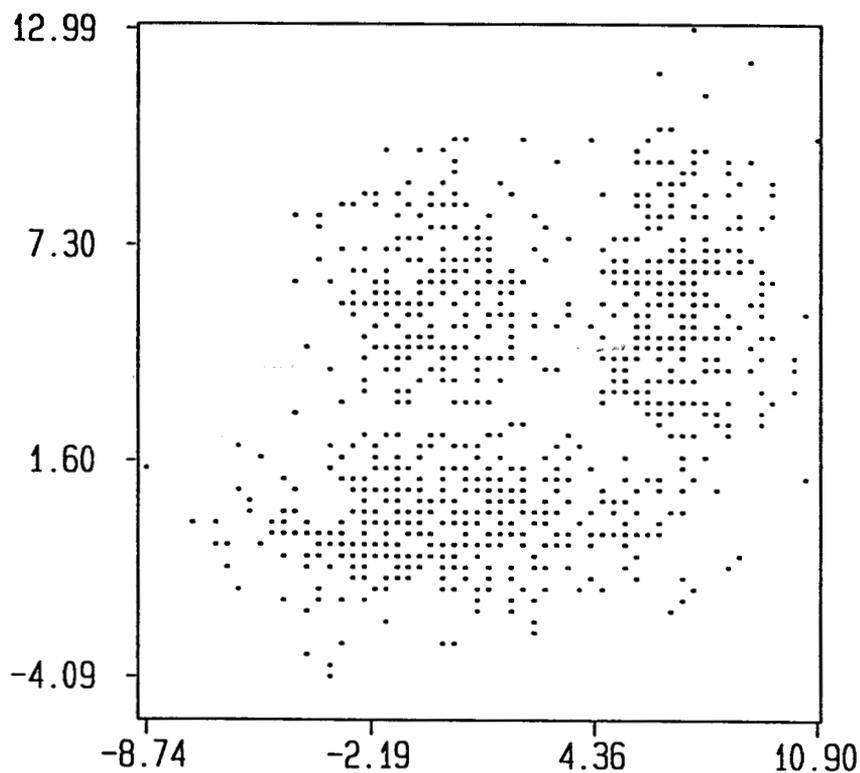


(b)

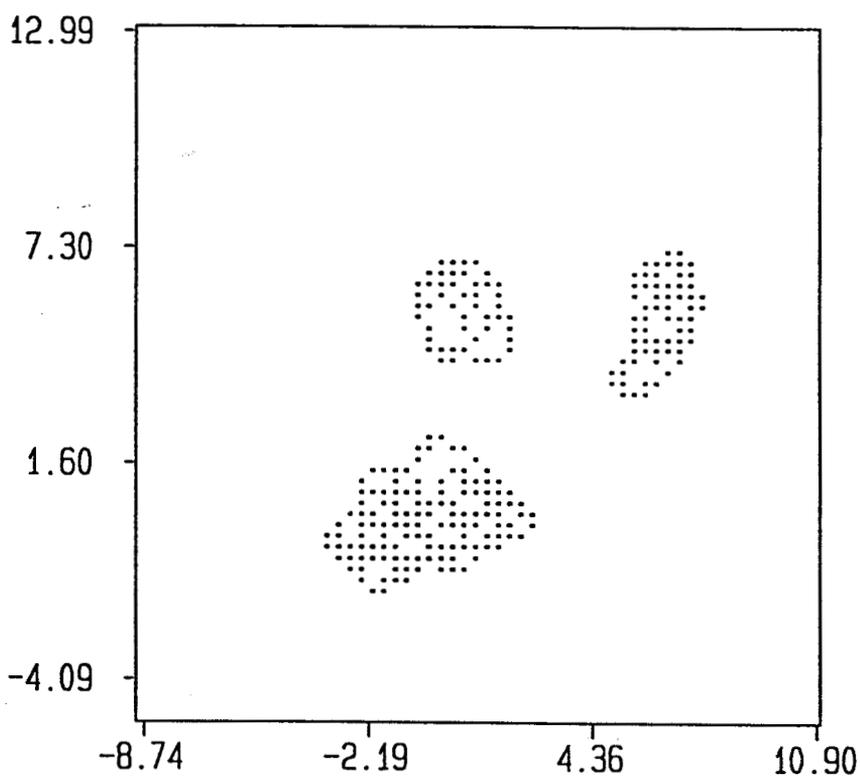
**Figure VI.3 :** Détermination des frontières modales par la méthode FMEU pour l'exemple VI.1.a.

*a - Ensemble discret ( $R = 41$ ).*

*b - Les frontières modales.*



(a)

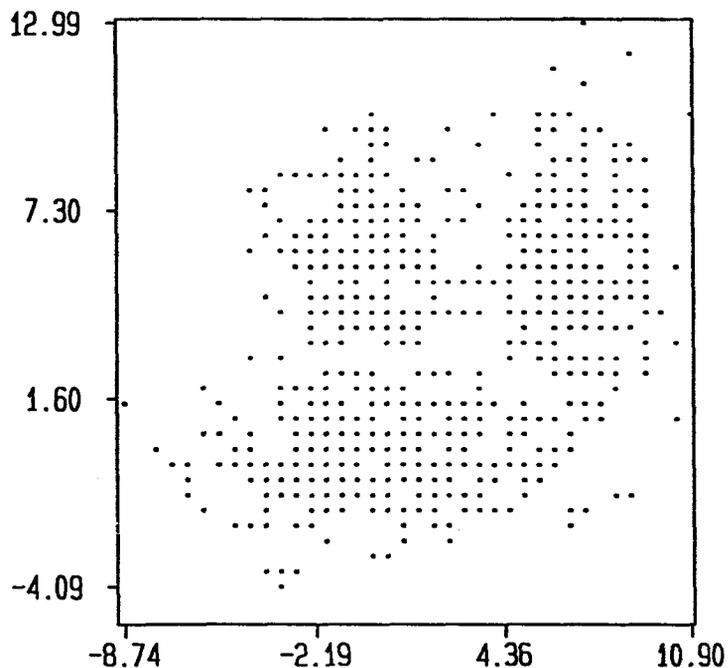


(b)

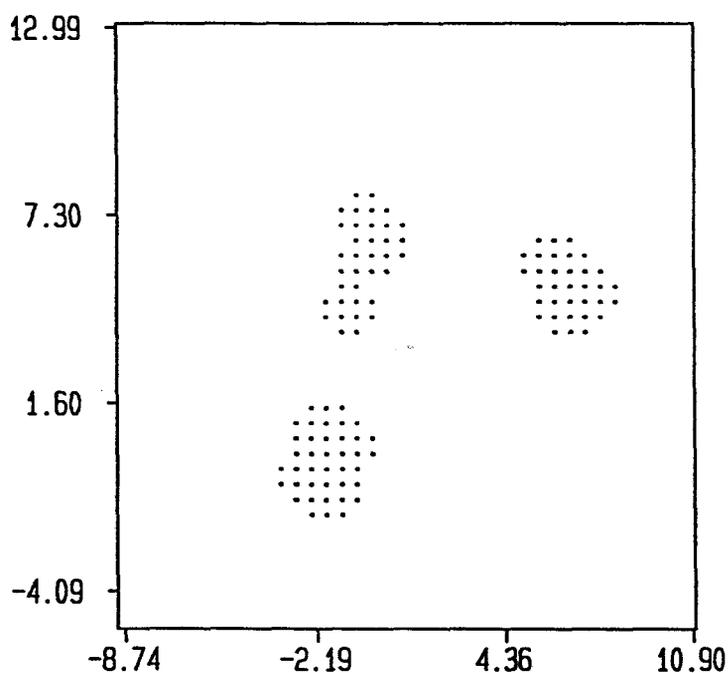
*Figure VI.4 : Détermination des domaines modaux par la méthode CVEM pour l'exemple VI.1.a.*

*a - Ensemble discret ( $R = 59$ ).*

*b - Les trois domaines modaux.*



(a)



(b)

**Figure VI.5 :** Détermination des domaines modaux par la méthode CMEM pour l'exemple VI.1.a.

*a - Ensemble discret ( $R = 37$ ).*

*b - Les trois domaines modaux.*

Après recherche de l'intervalle de stabilité du nombre de modes détectés pour chacune des quatre méthodes de détection des modes morphologiques, nous avons, pour chaque valeur du pas de discrétisation  $R$  de l'intervalle de stabilité, déterminé les domaines modaux ou leurs frontières. Nous avons ensuite effectué la classification des

observations de cet échantillon à l'aide des quatre méthodes de classification présentées dans le chapitre V à partir de ces différents domaines. Nous avons enfin calculé la valeur moyenne des taux d'erreur obtenus par chacune de ces méthodes sur la plage de stabilité. Le récapitulatif des divers résultats de cette approche morphologique est présenté en figure VI.6.

La technique de classification AONP est, une fois de plus, celle qui donne les résultats les plus stables vis à vis du pas de discrétisation, quelle que soit la méthode de détection des modes utilisée pour déterminer les domaines modaux ou leurs frontières. C'est pourquoi nous n'avons comparé les résultats des méthodes Isodata et K-means qu'à ceux obtenus à l'aide de cette technique de classification AONP.

La figure VI.7 présente, d'une part l'appartenance des observations aux différentes classes générées (cf. figure VI.7.a), et d'autre part les observations classées par la technique Isodata (cf. figure VI.7.b), par la technique des K-means (cf. figure VI.7.c) et par la technique morphologique composée de la méthode de détection des modes CVEM et de la technique de classification AONP (cf. figure VI.7.d). Le choix de la technique d'extraction des modes n'est pas déterminant, puisque le même taux d'erreur est en fait obtenu avec les quatre méthodes.

La table T6.2 reflète les différents taux d'erreur obtenus par ces trois techniques, qui sont à rapprocher du taux optimal de 3,04 % défini par la règle de Bayes. De plus, cette table présente les différentes matrices de confusion ainsi que les paramètres statistiques des groupements obtenus par chacune des trois techniques.

La différence d'assignation des observations entre les deux techniques classiques et la méthode morphologique CVEM + AONP concerne essentiellement les observations situées à la frontière des classes schématisées par des '+' et des '□' sur la figure VI.7. Cette différence s'explique par le principe même de l'approche morphologique, telle que nous la présentons dans ce mémoire, qui est de renforcer les vallées. Une fois les vallées creusées, les observations situées de part et d'autre de chaque vallée sont assignées différemment comme le montre la figure VI.7.d.

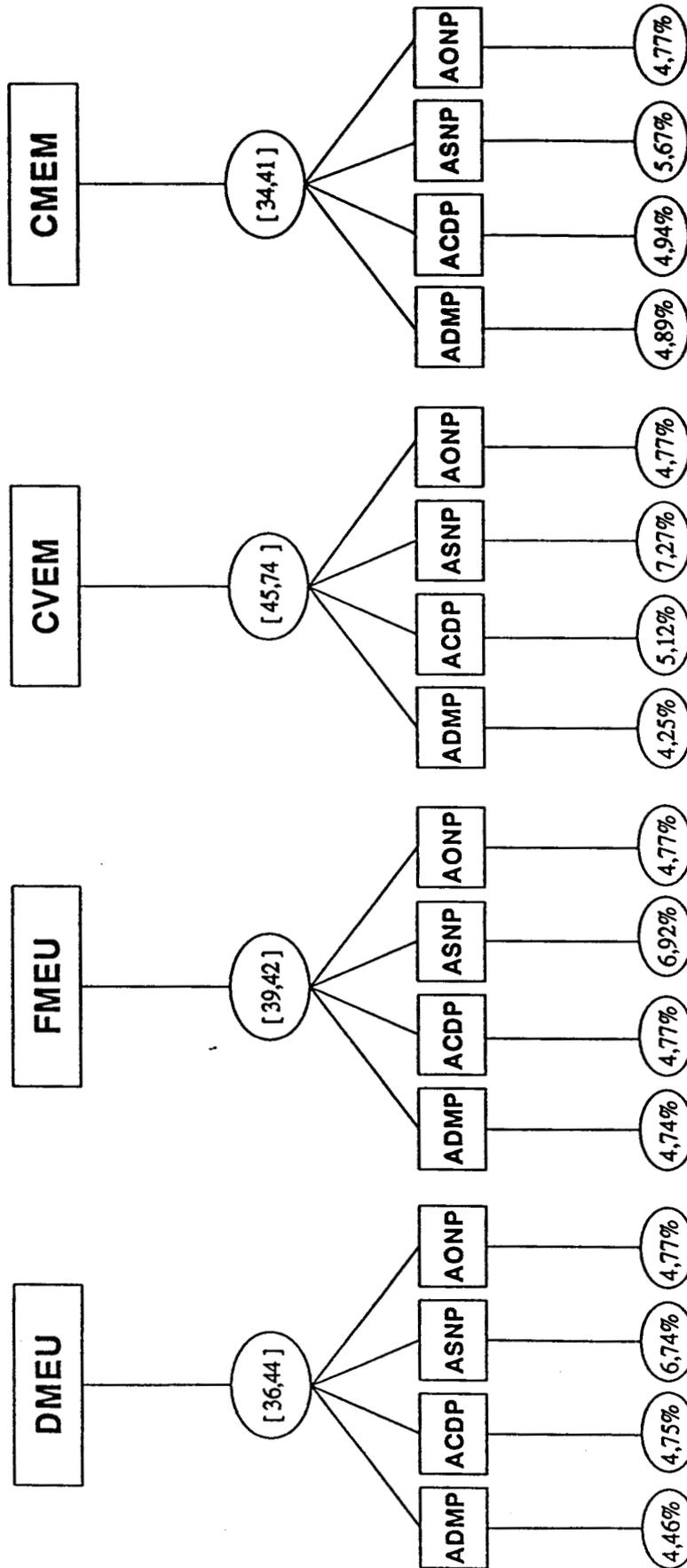
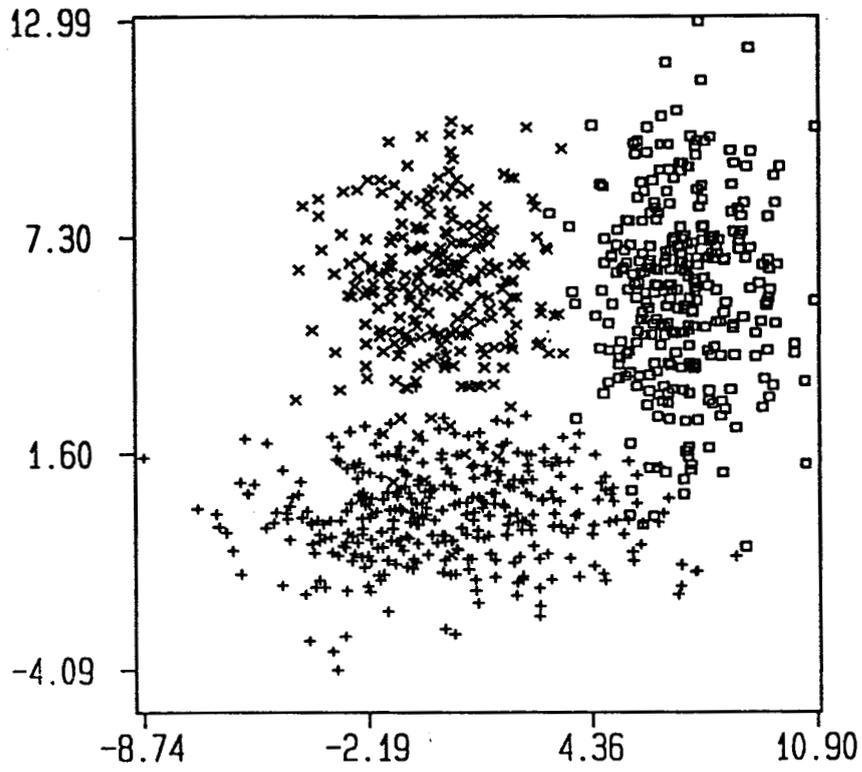
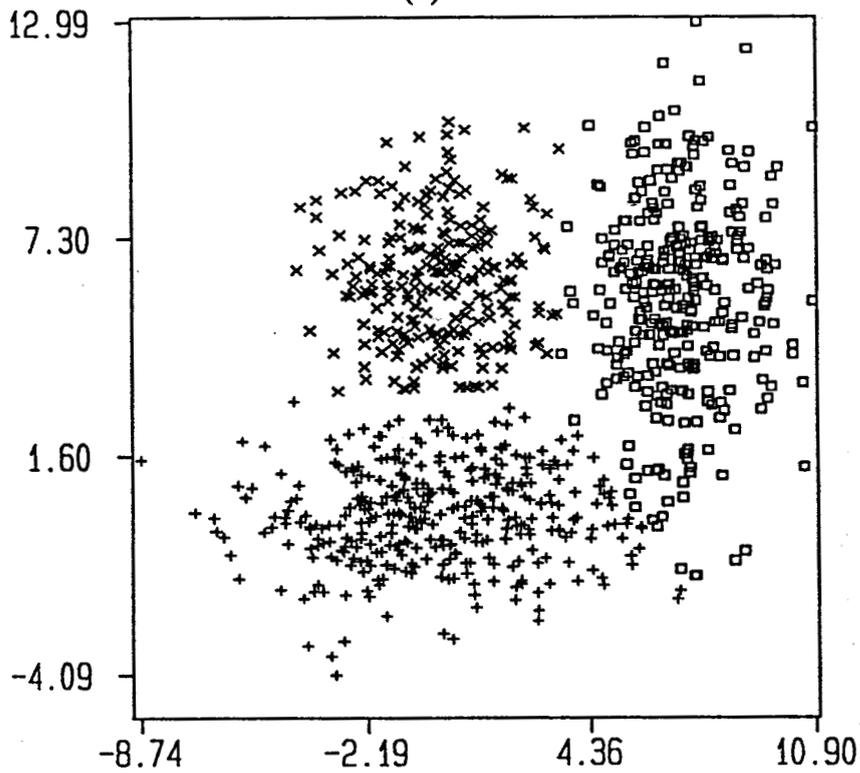


Figure VI.6 : Récapitulatif des résultats obtenus par l'approche morphologique pour l'exemple VI.1.a.



(a)

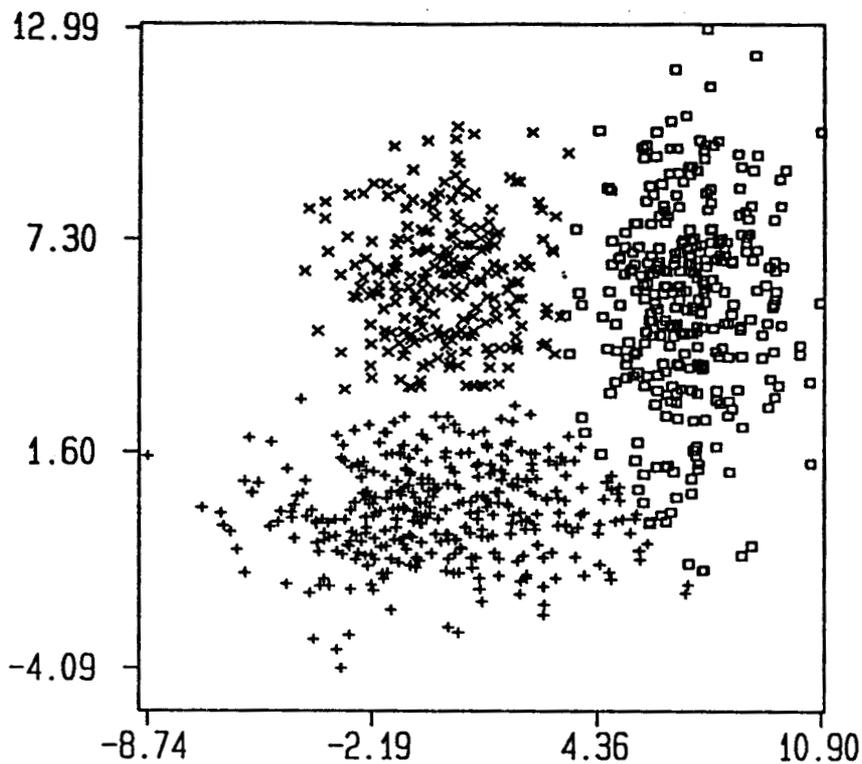


(b)

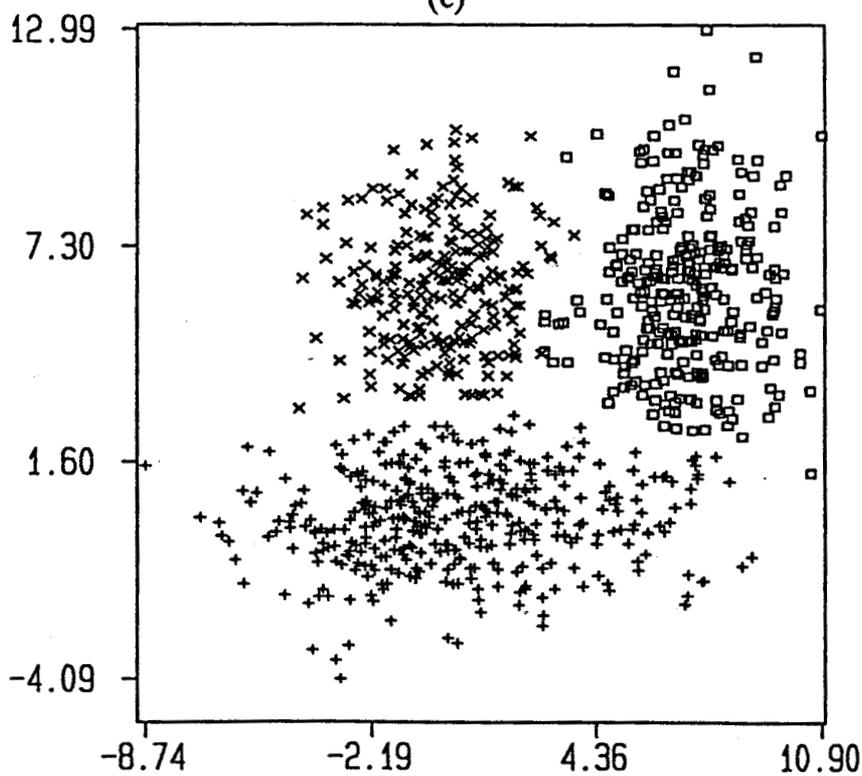
**Figure VI.7 :** Résultats de la classification des observations de l'exemple VI.1.a.

*a - Appartenance des observations de l'exemple VI.1.a.*

*b - Résultats de la classification des observations par la technique Isodata.*



(c)



(d)

**Figure VI.7 :** Résultats de la classification des observations de l'exemple VI.1.a (suite).

*c - Résultats de la classification des observations par la technique K-Means.*

*d - Résultats de la classification des observations par la technique morphologique (CVEM + AONP).*

	Données générées			Résultats d'Isodata			Résultats de K-Means			Résultats Morphologiques		
	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance
Population 1	350	0.0277	9.1639 0.0717	353	-0.1579	7.9446 0.0966	349	-0.2174	7.7152 0.0641	380	0.3592	10.623 0.3589
		0.0073	0.0717 1.4571		0.0659	0.0966 1.5294		0.0576	0.0641 1.5277		0.1005	0.3589 1.5070
Population 2	300	7.0393	1.9541 0.0670	310	7.0314	1.9059 0.1181	315	6.9945	1.9753 0.2328	288	6.9788	2.2709 0.0748
		5.9115	0.0670 5.7348		5.7220	0.1181 6.6053		5.6580	0.2328 6.8200		6.2085	0.0748 4.3064
Population 3	250	-0.0711	2.6771 -0.0052	237	-0.0854	2.5213 0.0339	236	-0.1002	2.4802 0.0488	232	-0.1840	2.3934 0.1658
		6.0252	-0.0052 3.5141		6.2667	0.0339 2.6252		6.2709	0.0488 2.6321		6.2761	0.1658 2.6392
Taux d'erreur				3.33 %			3.66 %			4.77 %		
Matrices de Confusion				338	2	13	335	1	13	349	19	12
				11	297	2	14	298	3	0	279	9
				1	1	235	1	1	234	1	2	229

*Table T6.2 : Valeurs des paramètres statistiques de la distribution bidimensionnelle de l'exemple VI.1.a. et paramètres statistiques des classes obtenues par les techniques Isodata, K-Means et (CVEM + AONP).*

Nous avons cherché à rapprocher les classes les unes des autres afin d'augmenter leur chevauchement pour définir les limites de la méthode morphologique proposée. Lorsque les données générées (cf. figure VI.8) sont définies par les différents paramètres de la table T6.3, les résultats de la classification des observations de l'échantillon par les techniques Isodata, K-means et (CVEM + AONP) sont ceux présentés en figure VI.9.

Dans la table T6.3 figurent également les caractéristiques des groupements obtenus par l'ensemble des trois techniques Isodata, K-means et (CVEM + AONP). Cette table contient aussi les matrices de confusion ainsi que les différents taux d'erreur relatifs à chacune de ces trois techniques. Précisons que ces taux d'erreur sont à rapprocher du taux optimal de 4,47 % fourni par la règle de Bayes.

On peut remarquer que l'approche morphologique donne des résultats plus proches de ceux obtenus par les deux autres techniques que dans le cas de l'exemple VI.1.a.

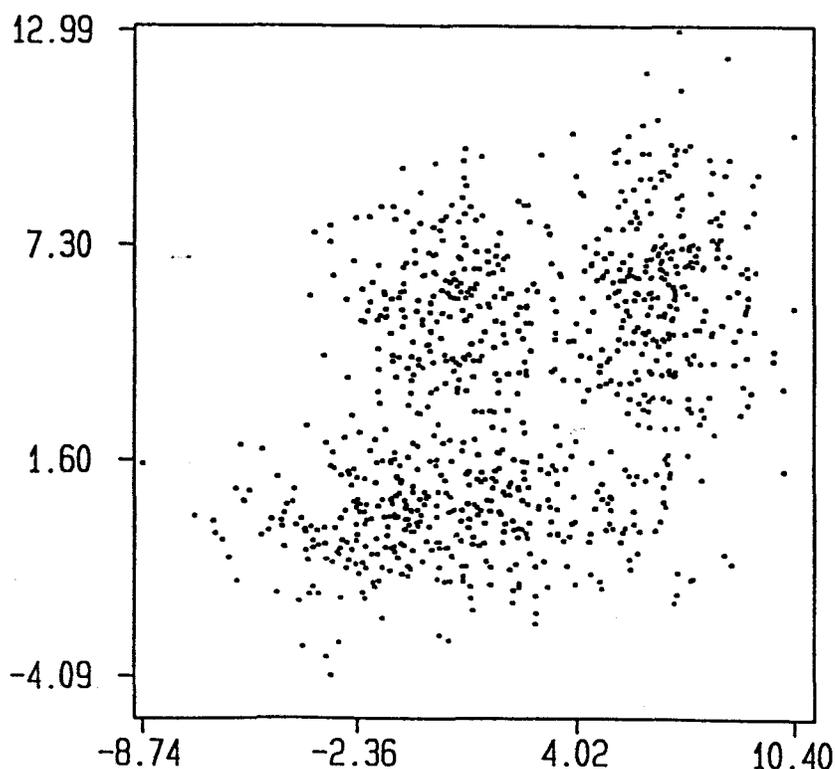
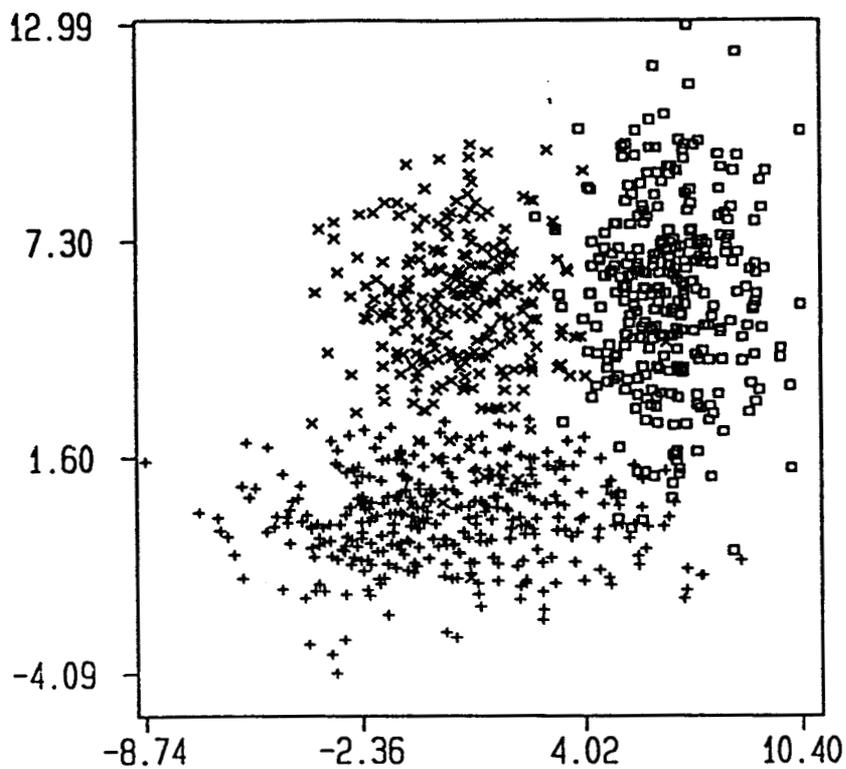


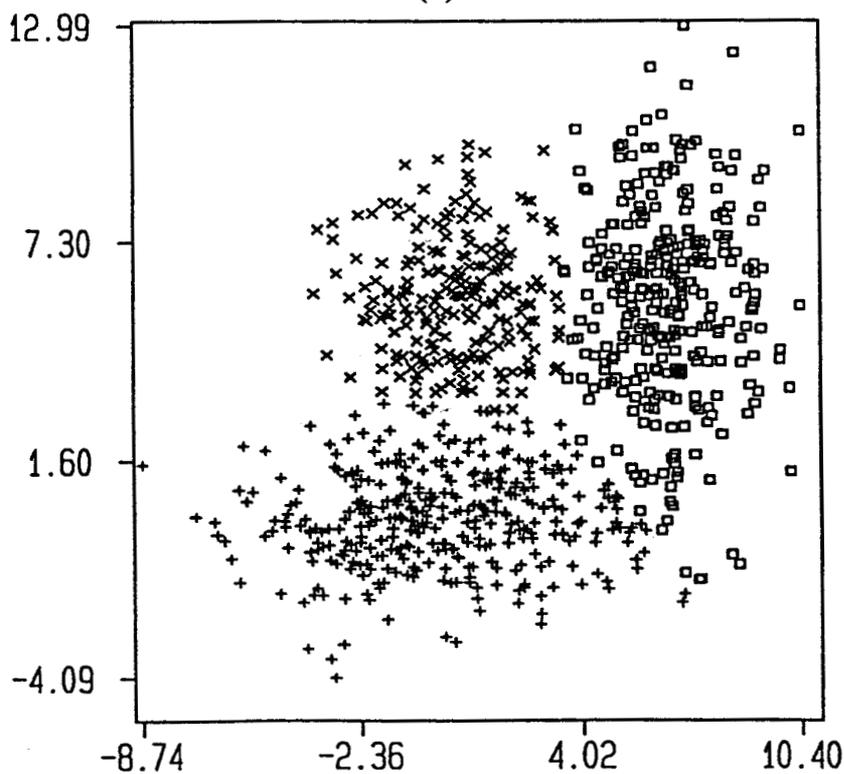
Figure VI.8 : Représentation graphique de l'échantillon de l'exemple VI.1.b.

	Données générées			Résultats d'Isodata			Résultats de K-Means			Résultats Morphologiques		
	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance
Population 1	350	0.0277 0.0073	9.1639 0.0717 0.0717 1.4571	363	-0.1300 0.1168	7.8349 0.0960 0.0960 1.6338	354	-0.2177 0.0812	7.6965 0.0445 0.0445 1.5816	381	0.3175 0.0932	10.352 0.2743 0.2743 1.4877
Population 2	300	6.5393 5.9115	1.9541 0.0670 0.0670 5.7348	312	6.5352 5.7368	1.9105 0.0080 0.0080 6.4937	317	6.5150 5.6526	1.9087 0.1108 0.1108 6.8368	289	6.4976 6.1984	2.1194 0.0691 0.0691 4.3824
Population 3	250	0.4288 5.525	2.6771 -0.0052 -0.0052 3.5141	225	0.3862 5.8890	2.3667 -0.0332 -0.0332 2.3530	229	0.3952 5.8370	2.3389 -0.0596 -0.0596 2.4643	230	0.3473 5.7846	2.2330 -0.0171 -0.0171 2.5582
Taux d'erreur				5.88 %			5.55 %			5.66 %		
Matrices de Confusion				336	5	22	333	3	18	347	19	15
				13	291	8	16	293	8	0	278	11
				1	4	220	1	4	224	3	3	224

*Table T6.3 : Valeurs des paramètres statistiques de la distribution bidimensionnelle de l'exemple VI.1.b et paramètres statistiques des classes obtenues par les techniques Isodata, K-Means et (CVEM + AONP).*



(a)

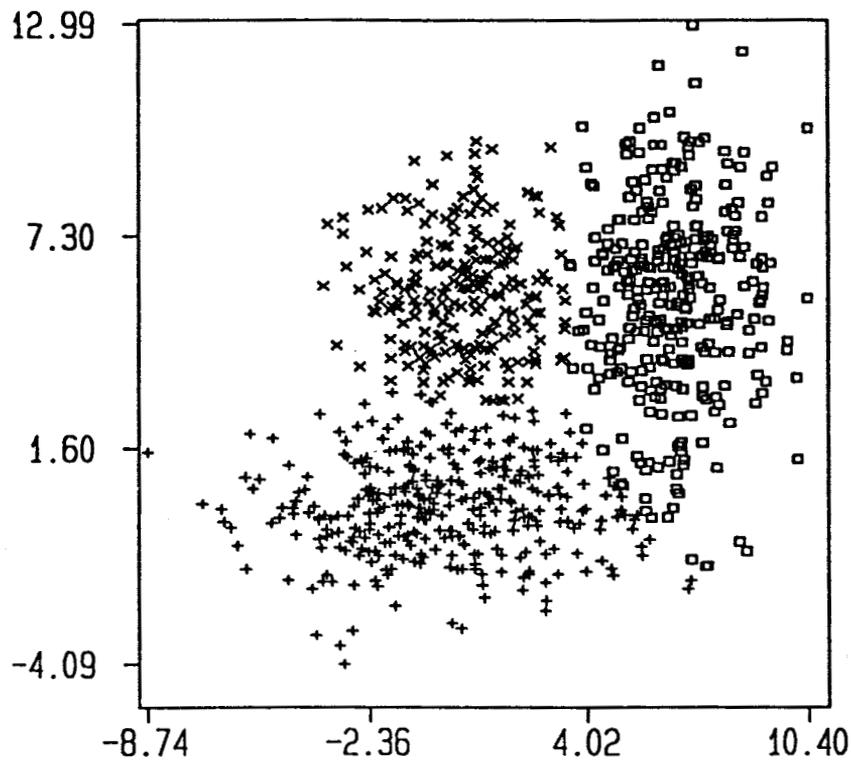


(b)

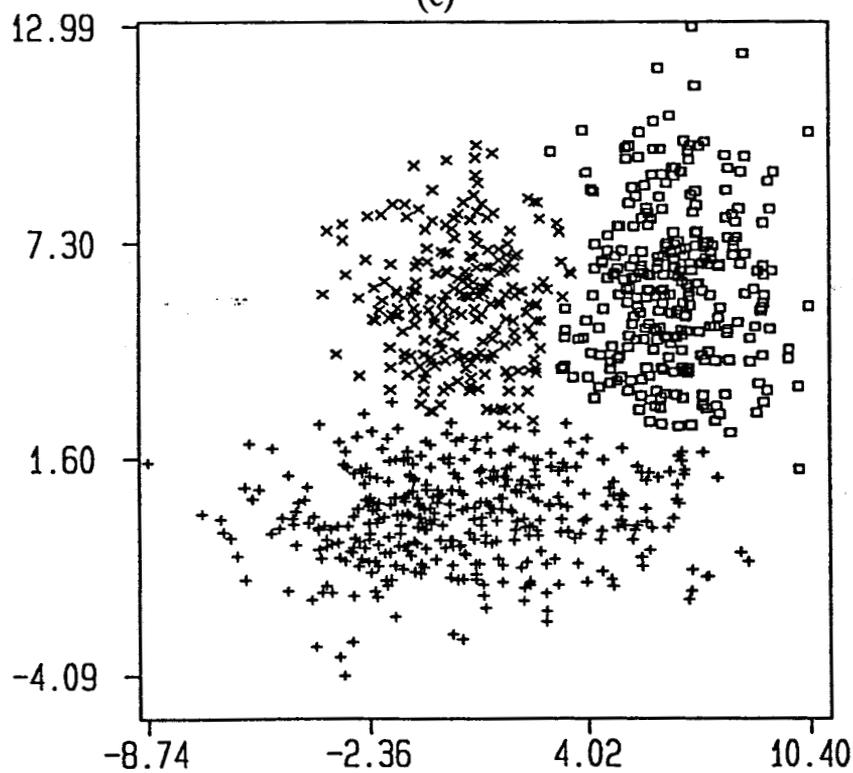
Figure VI.9 : Résultats de la classification des observations de l'exemple VI.1.b.

a - Appartenance des observations de l'exemple VI.1.b.

b - Résultats de la classification des observations par la technique Isodata.



(c)



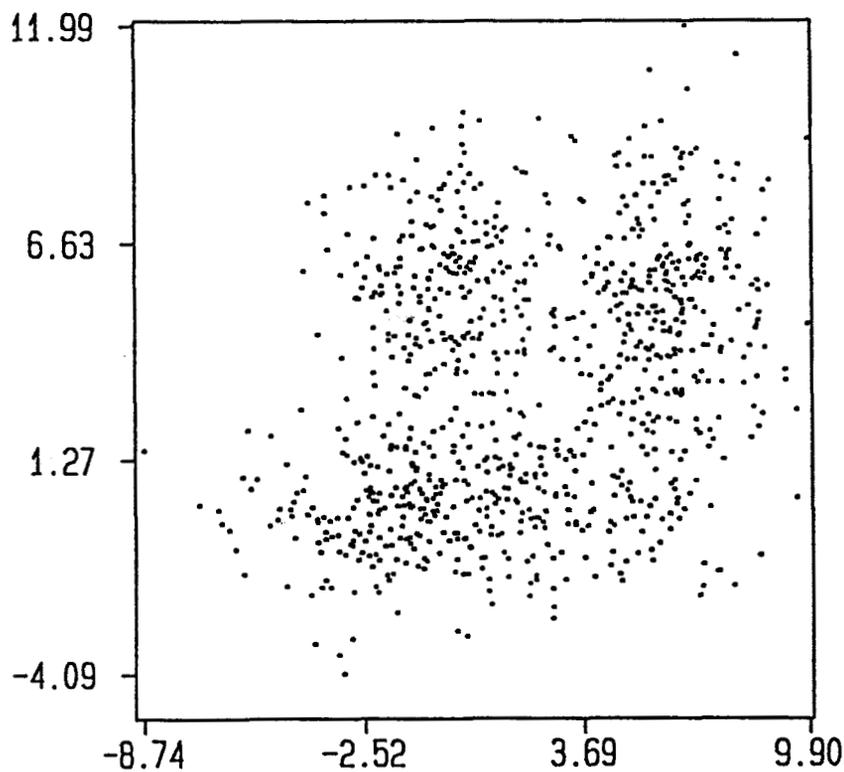
(d)

**Figure VI.9 :** Résultats de la classification des observations de l'exemple VI.1.b (suite).

*c - Résultats de la classification des observations par la technique K-Means.*

*d - Résultats de la classification des observations par la technique morphologique (CVEM + AONP).*

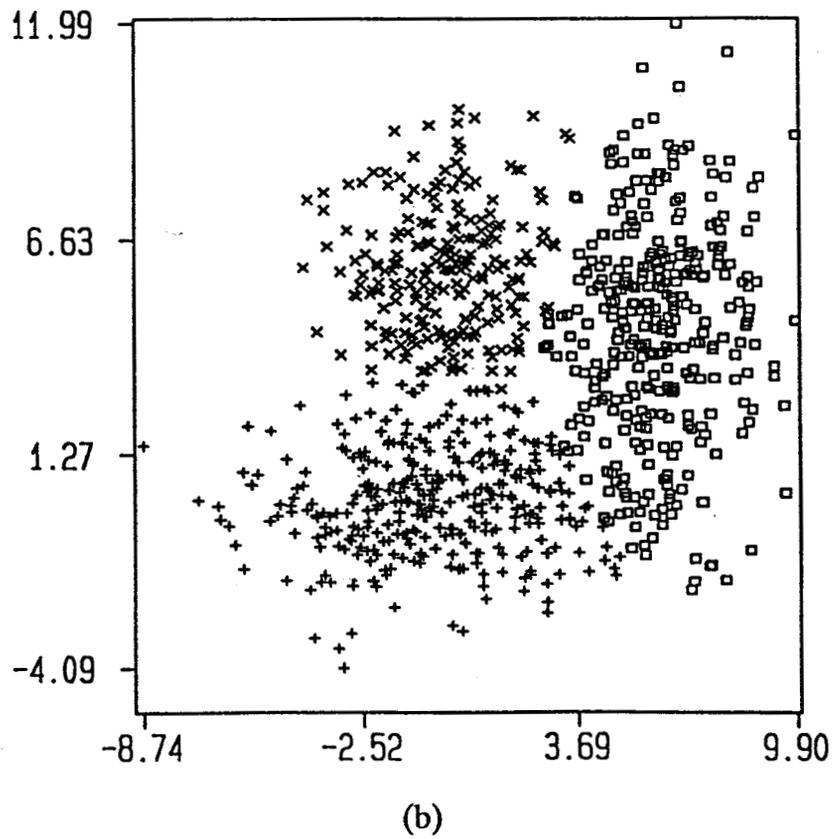
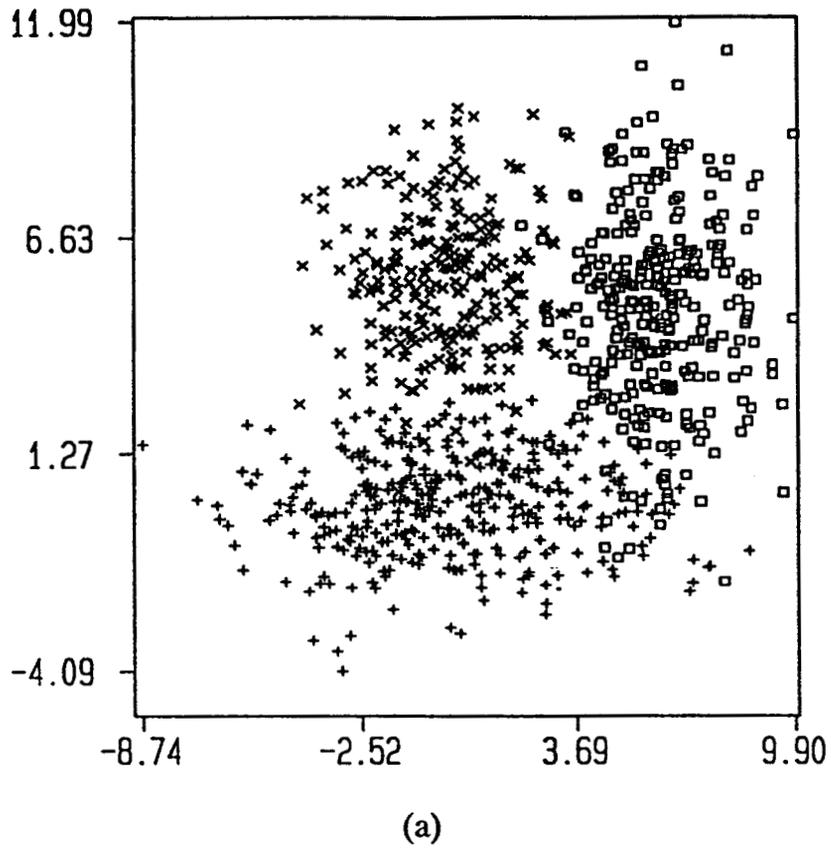
Lorsque les classes sont encore plus proches (cf. figure VI.10), les vallées sont alors si peu marquées que la méthode de détection des modes FMEU ne permet plus d'individualiser les différents modes morphologiques de manière stable. Par contre, les résultats morphologiques obtenus à partir des modes extraits par les méthodes CVEM, CMEM ou DMEU sont, dans cet exemple, supérieurs à ceux fournis par les techniques Isodata et K-Means (cf. table T6.4). Le taux d'erreur défini par la règle optimale de Bayes est de 5,58 %. Ce taux fournit, comme pour les résultats précédents, une valeur de référence qui permet d'évaluer les résultats obtenus. L'appartenance des observations aux différentes classes en présence ainsi que les observations classées par chacune des trois techniques d'évaluation sont présentées en figure VI.11.



*Figure VI.10 : Représentation graphique des observations de l'exemple VI.1.c.*

	Données générées			Résultats d'Isodata			Résultats de K-Means			Résultats Morphologiques		
	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance	Nombre d'observations	Vecteur moyenne	Matrice de covariance
Population 1	350	0.0277 0.0073	9.1639 0.0717 0.0717 1.4571	338	-0.5511 0.1110	6.1922 0.0512 0.0512 1.6932	338	-0.5511 0.1110	6.1922 0.0512 0.0512 1.6932	329	-0.5906 0.0960	6.1299 0.2742 0.2742 1.5330
Population 2	300	6.0393 4.9115	1.9541 0.0670 0.0670 5.7348	338	5.9600 4.4009	1.9552 0.2907 0.2907 7.1044	338	5.9600 4.409	1.9552 0.2907 0.2907 7.1044	342	5.9388 4.3650	1.9897 0.2484 0.2484 7.6838
Population 3	250	-0.0711 5.5252	2.6771 -0.0052 -0.0052 3.5141	224	-0.1093 5.9476	2.4129 0.1313 0.1313 2.3654	224	-0.1093 5.9476	2.4129 0.1313 0.1313 2.3654	229	-0.1442 5.8205	2.3031 0.0633 0.0633 2.4820
Taux d'erreur				8.44 %			8.44 %			7.55 %		
Matrices de Confusion				312	3	23	312	3	23	313	1	15
				37	293	8	37	293	8	36	295	11
				1	4	219	1	4	219	1	4	224

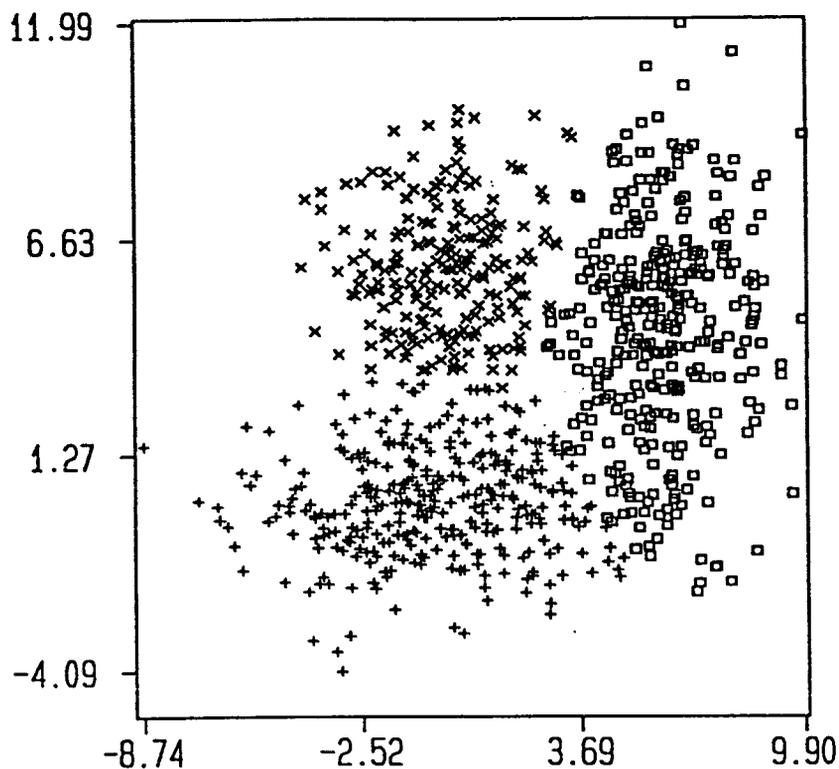
*Table T6.4 : Valeurs des paramètres statistiques de la distribution bidimensionnelle de l'exemple VI.1.c, et paramètres statistiques des classes obtenues par les techniques Isodata, K-Means et (CVEM + AONP).*



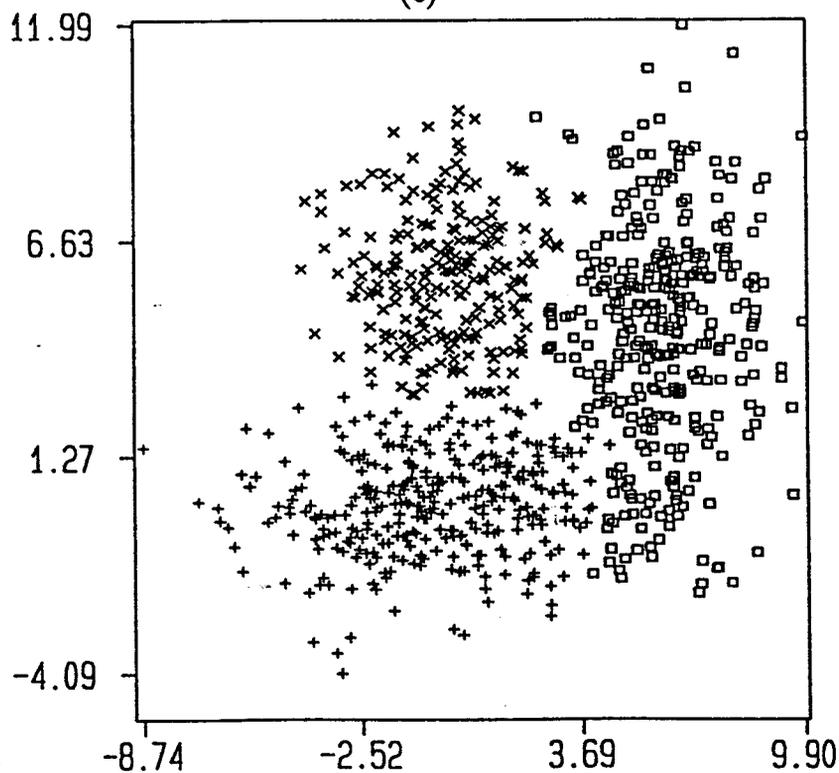
**Figure VI.11** : Résultats de la classification des observations de l'exemple VI.1.c.

*a - Appartenance des observations de l'exemple VI.1.c.*

*b - Résultats de la classification des observations par la technique Isodata.*



(c)



(d)

**Figure VI.11 :** Résultats de la classification des observations de l'exemple VI.1.c (suite).

*c - Résultats de la classification des observations par la technique K-Means.*

*d - Résultats de la classification des observations par la technique morphologique (CVEM + AONP).*

#### VI.4. EXEMPLE BIDIMENSIONNEL COMPOSE DE TROIS CLASSES NON SPHERIQUES NON NORMALES.

L'ensemble des données bidimensionnelles considéré dans ce deuxième exemple est constitué d'une classe sphérique et de deux classes en forme de croissants (cf. figure VI.12).

La première population est constituée de 250 observations normales dont les paramètres statistiques sont définis par :

$$Y = \begin{bmatrix} 0.0563 \\ 5.0819 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 12.17 & 0 \\ 0 & 11.15 \end{bmatrix}$$

Les deuxième et troisième populations sont composées de 450 observations chacune. Les attributs  $y^1, y^2$  décrivant ces observations sont définis par :

$$y^1 = A_1 \cos \theta + B_1$$

$$y^2 = A_2 \sin \theta + B_2$$

où  $\theta$  est une variable aléatoire normale de moyenne  $m$  et de variance  $s$ , et où  $B_1$  et  $B_2$  sont des variables aléatoires normales de moyennes  $\mu$  et de variance  $\sigma$ . Les valeurs de ces différents paramètres sont détaillées dans la table T6.5.

	$\theta$	$B_1$	$B_2$	$A_1$	$A_2$
POPULATION 2	$m = 0^\circ$ $s = 60^\circ$	$\mu_1 = 4.91$ $\sigma_1 = 2.52$	$\mu_2 = 10.00$ $\sigma_2 = 2.52$	$A_1 = 15$	$A_2 = 15$
POPULATION 3	$m = 180^\circ$ $s = 60^\circ$	$\mu_1 = -5.05$ $\sigma_1 = 2.54$	$\mu_2 = -0.00$ $\sigma_2 = 2.54$	$A_1 = 15$	$A_2 = 15$

*Table T6.5 : Paramètres statistiques des populations 2 et 3 de l'exemple VI.2.*

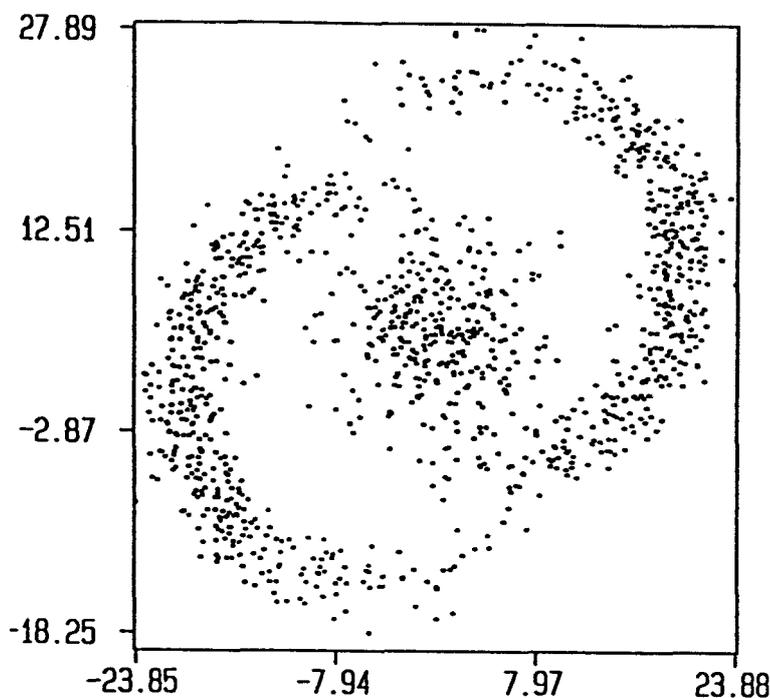
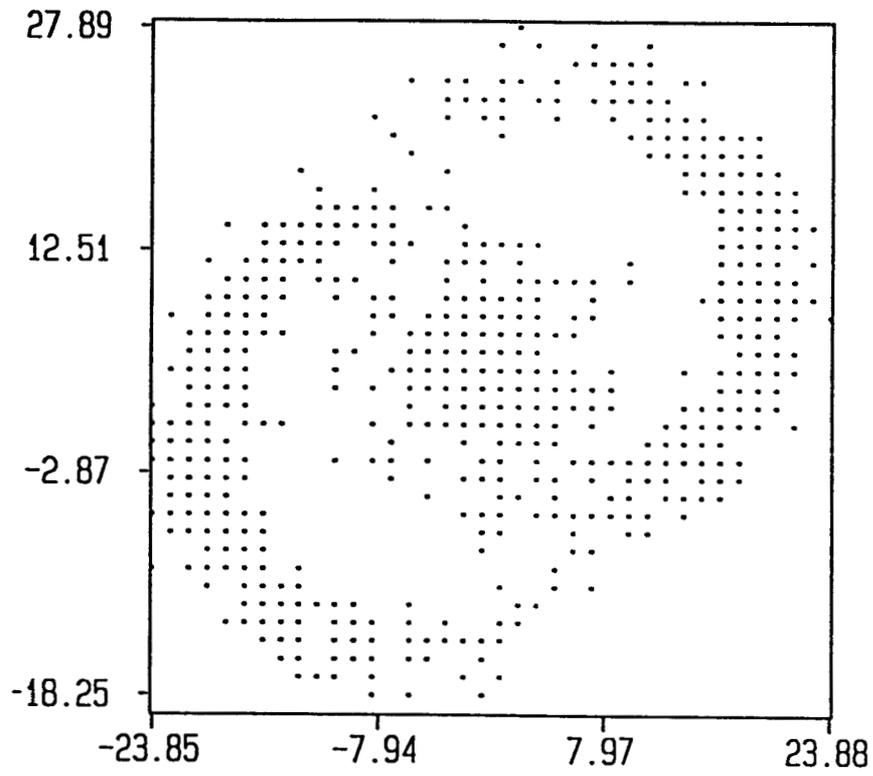


Figure VI.12 : Représentation graphique des observations de l'exemple VI.2.

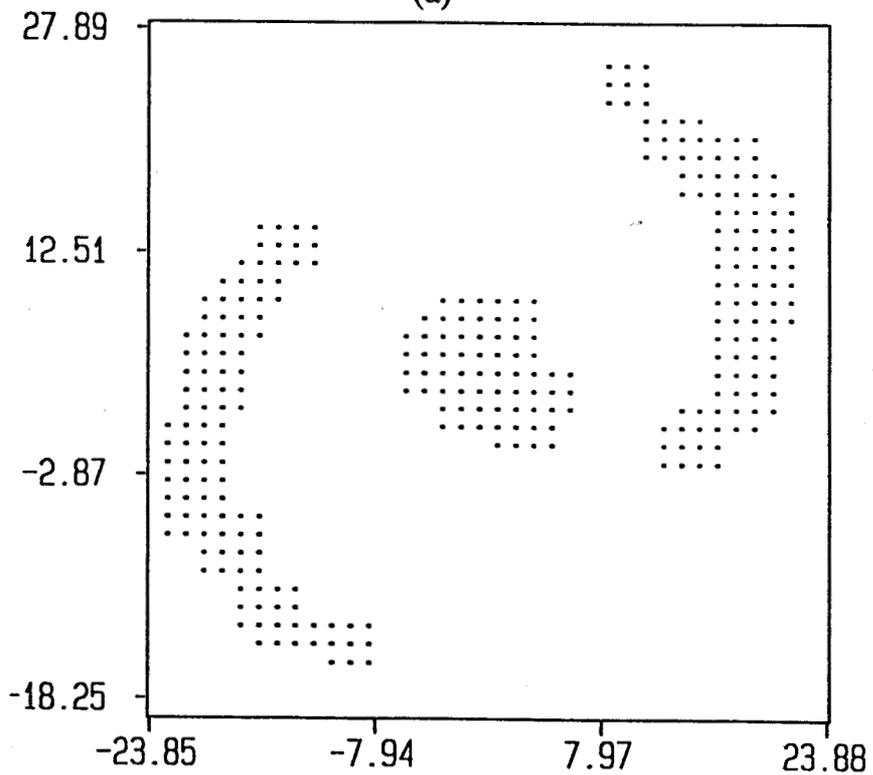
La recherche des domaines modaux par les différentes méthodes d'extraction des modes a abouti aux résultats présentés en figure VI.13 pour les méthodes DMEU, FMEU, et CMEM et en figure VI.14 pour la méthode CVEM. Notons que la forme géométrique des différents domaines obtenus reflètent correctement la forme des classes constituant l'échantillon analysé.

Après détermination des domaines modaux par la technique CVEM, nous avons effectué la classification des observations de l'échantillon par les techniques Isodata, K-Means et AONP (cf. figure VI.15). Ces résultats montrent bien l'intérêt de l'approche morphologique dans ce cas de figure. En effet, les deux techniques classiques Isodata et K-Means ne prennent pas en compte la forme géométrique des classes. Les observations situées aux extrémités des classes en forme de croissants ont tendance à être assignées à la classe sphérique gaussienne centrale schématisée par des '+' (cf. figure VI.15.b et VI.15.c). Par contre, comme l'approche morphologique prend en considération ces différentes formes, l'assignation est plus correcte (cf. figure VI.15.d).

Les résultats numériques relatifs aux différentes classifications obtenues sont fournis dans la table T6.6 par l'intermédiaire des matrices de confusion et des différents taux d'erreur.



(a)

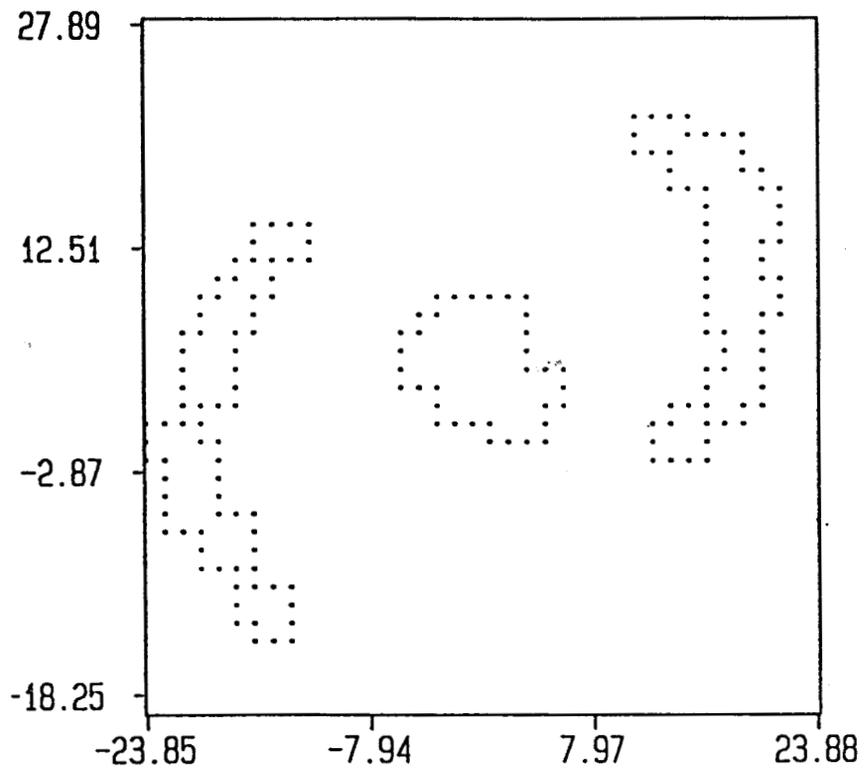


(b)

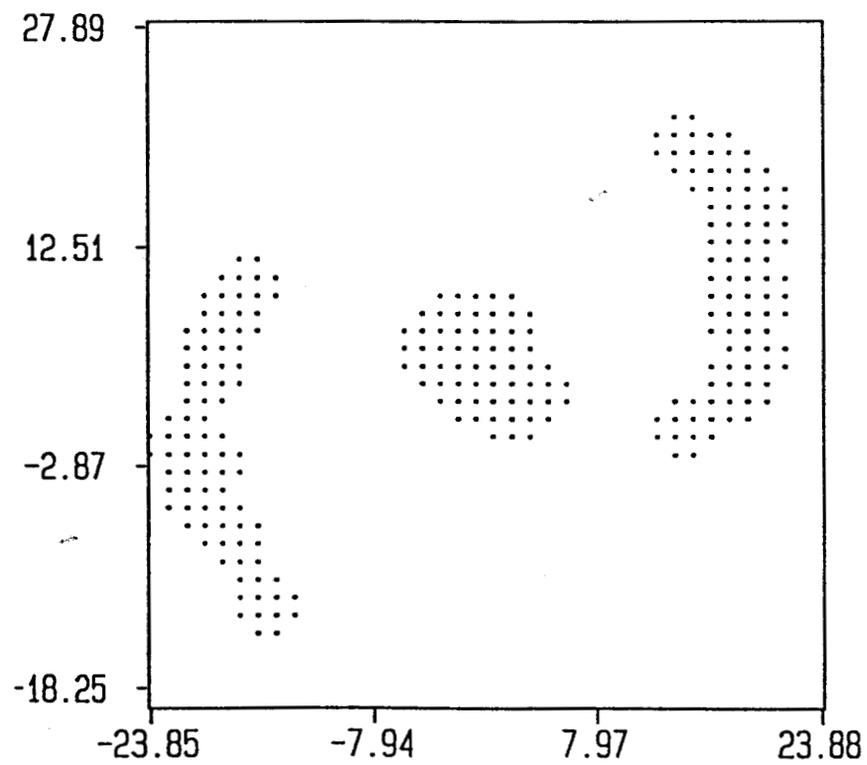
*Figure VI.13 : Détermination des domaines modaux de l'exemple VI.2.*

*a - Ensemble discret ( $R = 37$ ).*

*b - Domaines modaux déterminés par la méthode DMEU.*



(c)

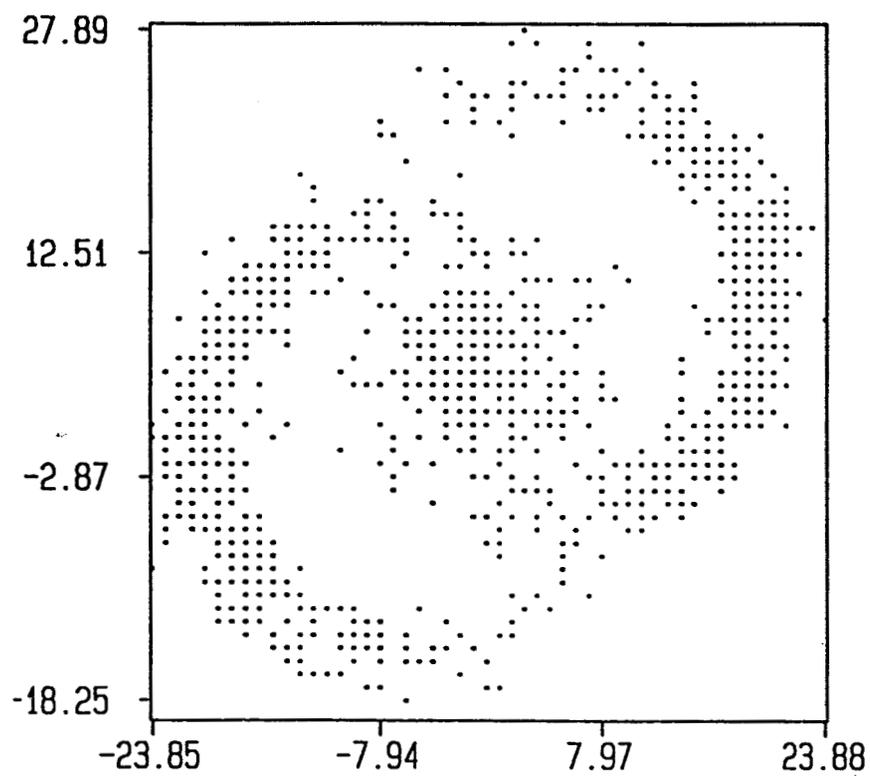


(d)

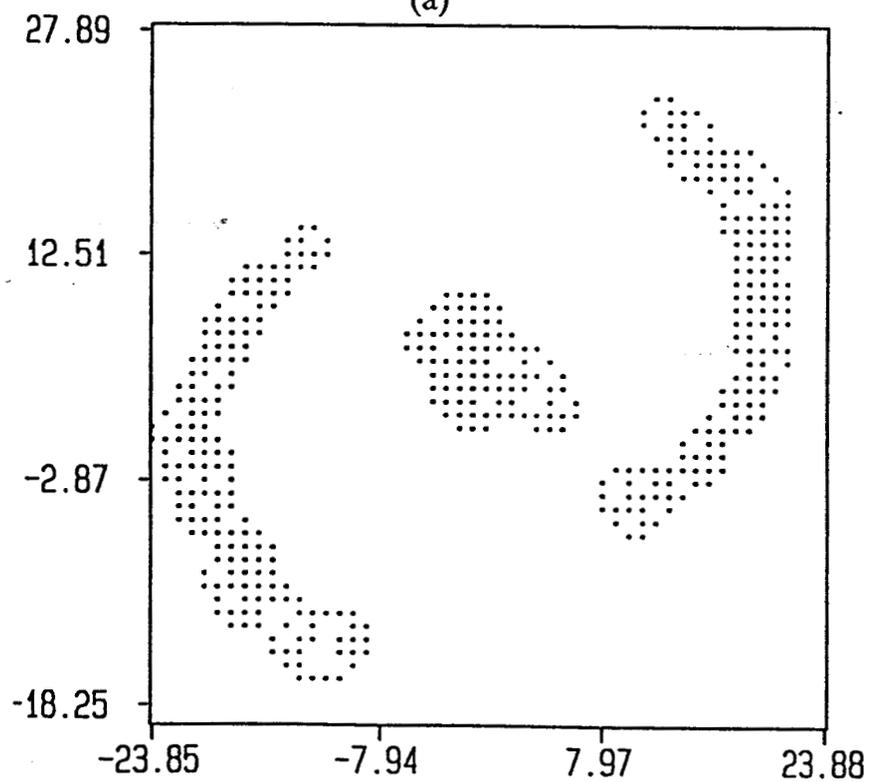
**Figure VI.13 :** Détermination des domaines modaux de l'exemple VI.2 (suite).

*c - Frontières modales déterminées par la méthode FMEU.*

*d - Domaines modaux déterminés par la méthode CMEM.*



(a)

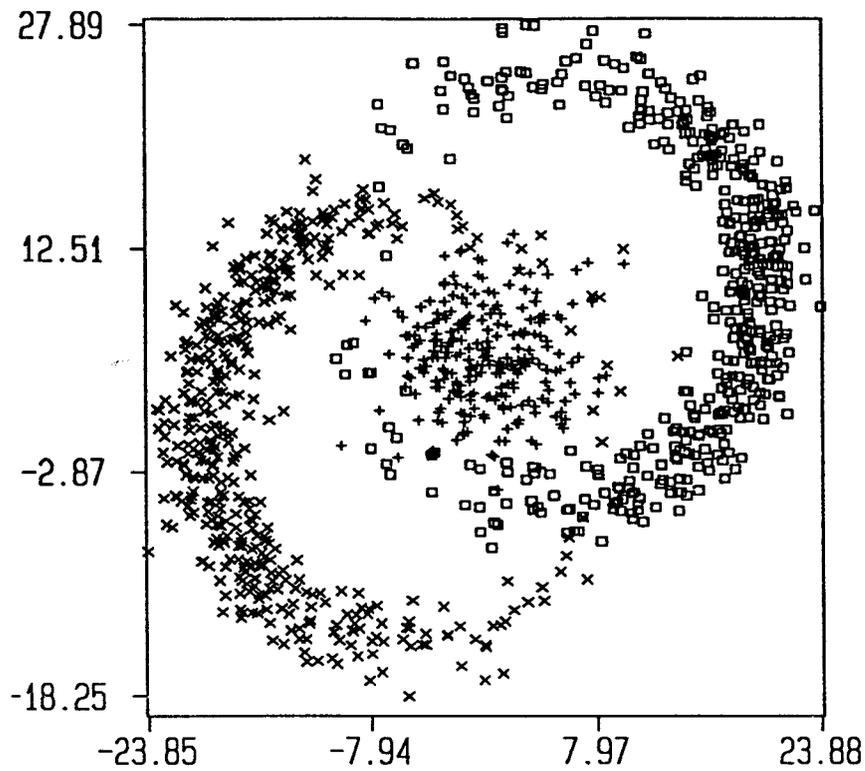


(b)

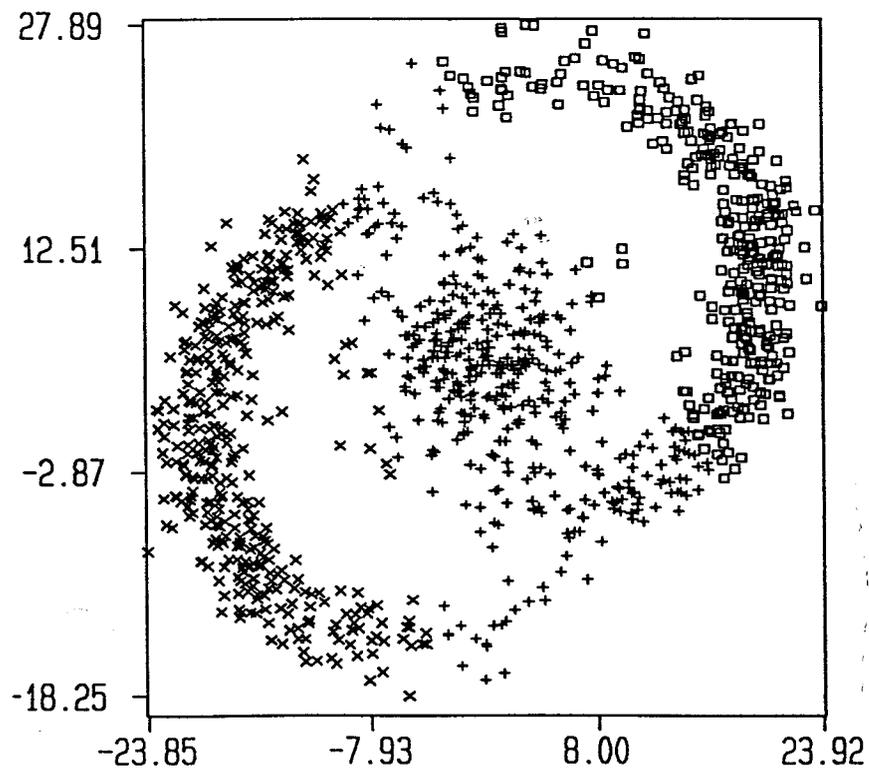
*Figure VI.14 : Détermination des domaines modaux de l'exemple VI.2.*

*a - Ensemble discret ( $R = 51$ ).*

*b - Domaines modaux déterminés par la méthode CVEM.*



(a)

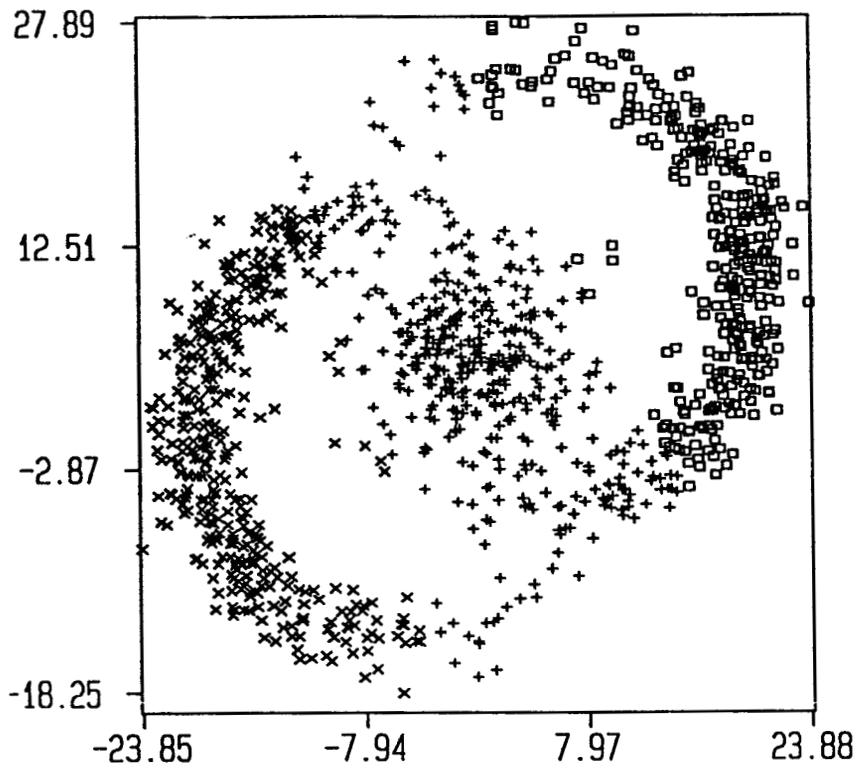


(b)

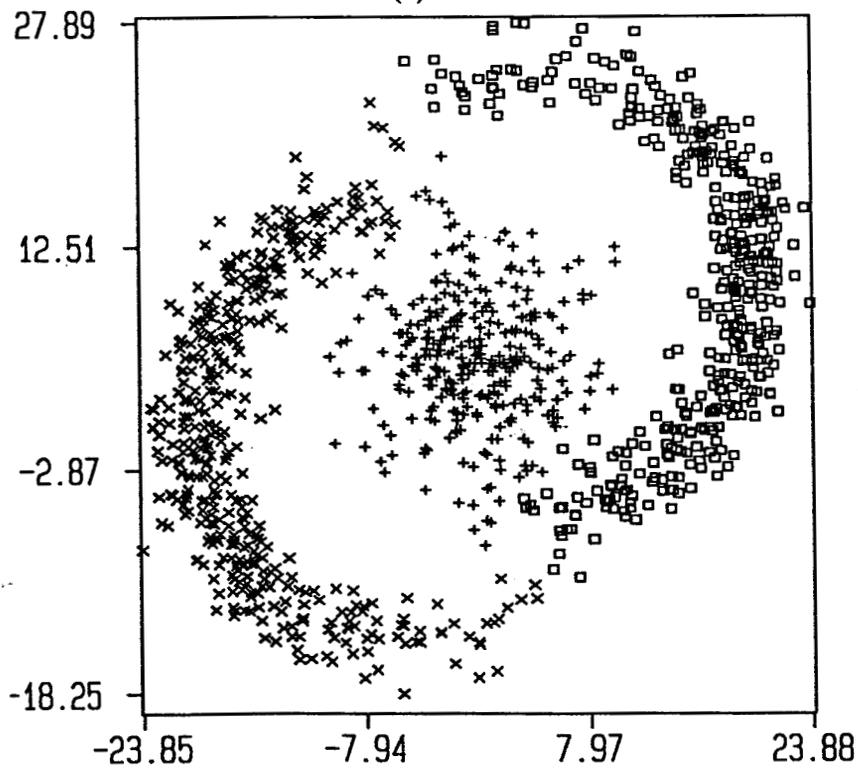
Figure VI.15 : Résultats de la classification des observations de l'exemple VI.2.

a - Appartenance des observations de l'exemple VI.2.

b - Résultats de la classification des observations par la technique Isodata.



(c)



(d)

**Figure VI.15 :** Résultats de la classification des observations de l'exemple VI.2 (suite).

*c - Résultats de la classification des observations par la technique K-Means.*

*d - Résultats de la classification des observations par la technique morphologique (CVEM + AONP).*

	Résultats d' Isodata			Résultats de K-Means			Résultats Morphologiques		
Taux d'erreur	1 6.34 %			16.69 %			5.91 %		
Matrices de confusion	246	110	62	247	104	76	250	32	21
	2	331	3	2	340	3	0	411	8
	2	9	385	1	6	371	0	7	421

Table T6.6 : Résultats de la classification de l'échantillon de l'exemple VI.2 par les trois procédures.

### VI.5. EXEMPLE MULTIDIMENSIONNEL.

L'échantillon que nous allons étudier maintenant est composé de deux classes de 800 observations normales chacune, caractérisées par trois attributs (cf. figure VI.16). Les caractéristiques statistiques de chacune des classes générées sont définies dans la table T6.7.

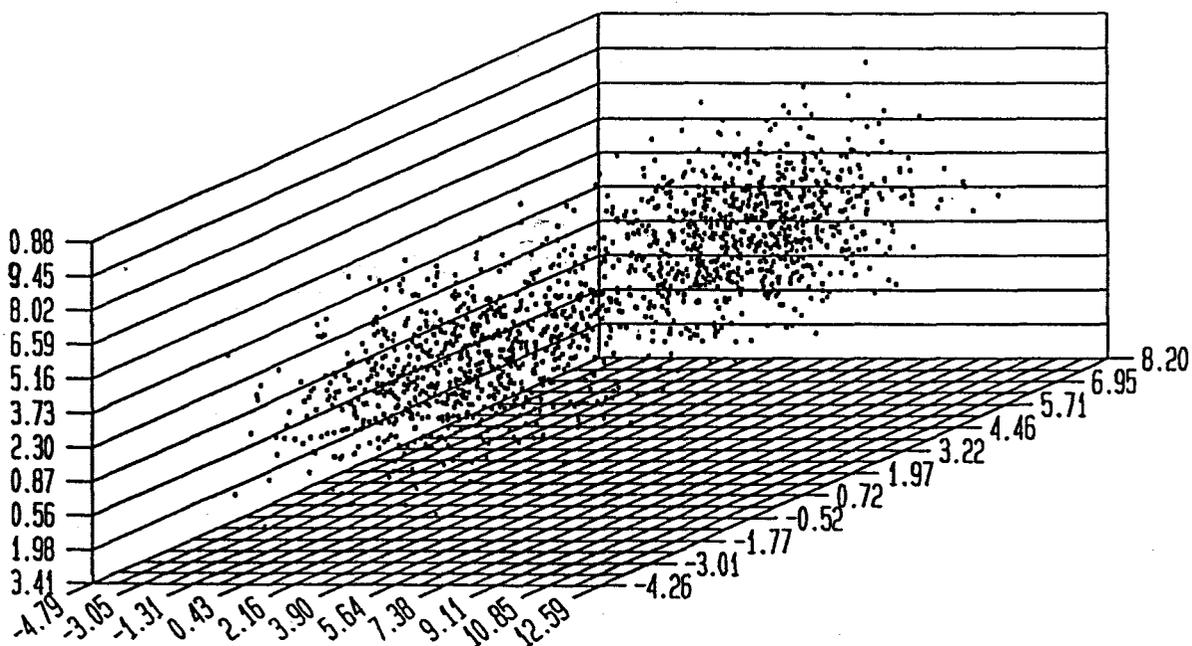


Figure VI.16 : Représentation des observations de l'exemple VI.3 composé de 1600 observations multidimensionnelles.

	Données générées				Résultats d'Isodata			
	Vecteur Moyenne	Matrice de covariance			Vecteur Moyenne	Matrice de covariance		
Population 1	1.0138	2.9465	-0.1299	-0.1987	0.9734	2.7749	-0.2119	-0.2761
	0.9545	-0.1299	3.0983	0.1024	0.9404	-0.2119	3.0880	0.0507
	1.9488	-0.1987	0.1024	2.9211	1.9289	-0.2761	0.0507	2.8870
Population 2	6.0754	2.9723	-0.0390	-0.1523	6.0904	2.8618	-0.0758	-0.2225
	4.0872	-0.0390	1.9843	-0.0090	4.0855	-0.0758	1.9616	-0.0141
	5.0390	-0.1523	-0.0090	3.0858	5.0433	-0.2225	-0.0141	3.0441
Taux d'erreur					2.25 %			
Matrices de confusion					780	16		
					20	784		

	Résultats de K-Means				Résultats Morphologiques			
	Vecteur Moyenne	Matrice de covariance			Vecteur Moyenne	Matrice de covariance		
Population 1	0.9734	2.7749	-0.2119	-0.2761	0.9542	2.7228	-0.2129	-0.2882
	0.9404	-0.2119	3.0880	0.0507	0.9543	-0.2139	3.1876	0.0856
	1.9289	-0.2761	0.0507	2.8870	1.9301	-0.2882	0.0856	2.9063
Population 2	6.0904	2.8618	-0.0758	-0.2225	6.0902	2.8168	-0.0040	-0.2039
	4.0855	-0.0758	1.9616	-0.0141	4.0602	-0.0040	1.9893	0.0341
	5.0433	-0.2225	-0.0141	3.0441	5.0305	-0.2039	0.0341	3.0681
Taux d'erreur	2.25 %				2.31 %			
Matrices de Confusion	780				778			
	20				22			
	16				15			
	784				785			

*Table T6.7 : Valeurs des paramètres statistiques de la distribution multidimensionnelle de l'exemple VI.3 telle qu'elle a été simulée, paramètres statistiques des classes obtenues par les techniques Isodata, K-Means et (CVEM + AONP).*

Les différents domaines ou frontières déterminés par chacune des quatre techniques d'extraction des modes morphologiques sont présentés sur la figure VI.17 pour les méthodes DMEU et FMEU, sur la figure VI.18 pour la méthode CVEM et sur la figure VI.19 pour la méthode CMEM.

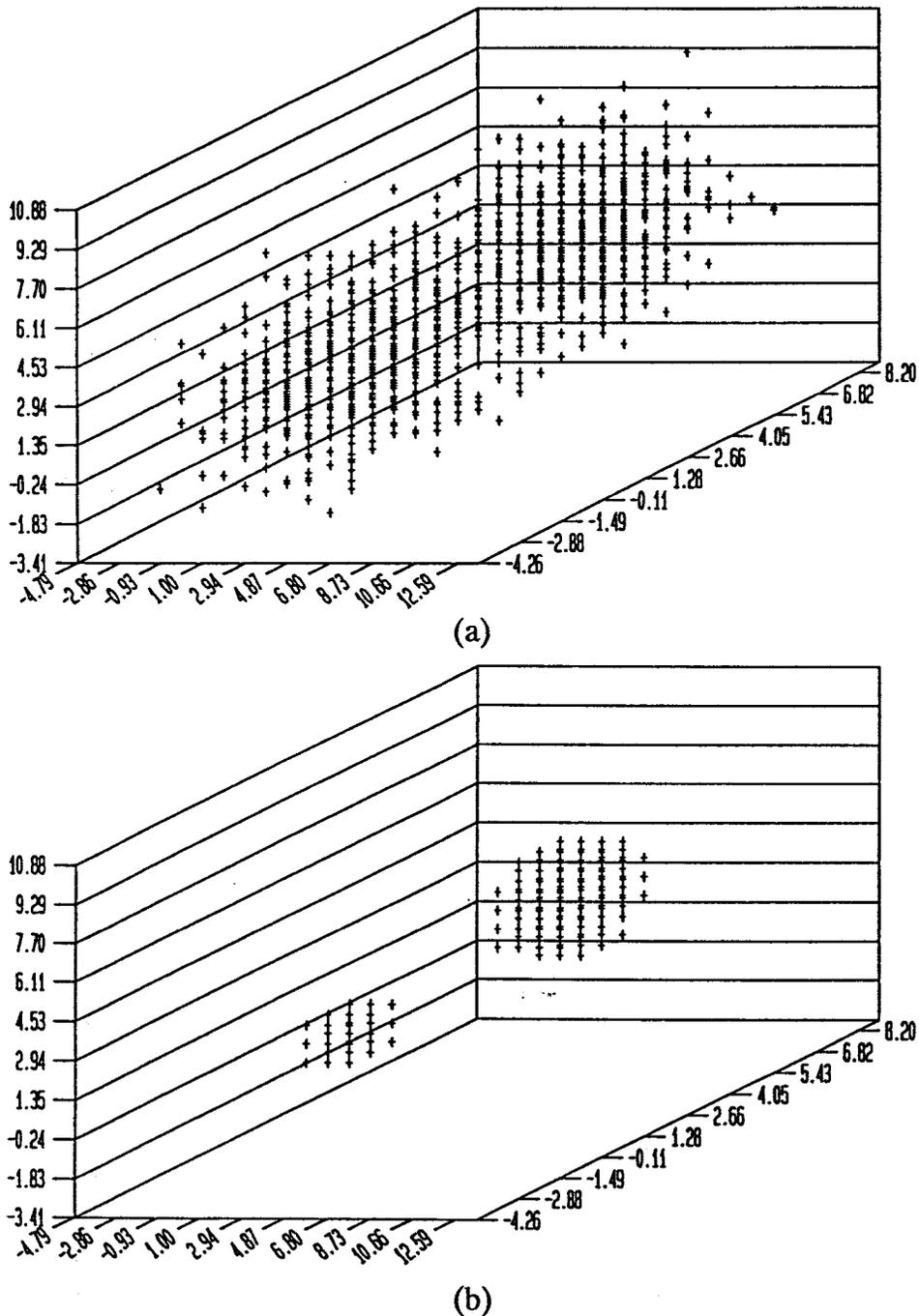


Figure VI.17 : Détermination des modes morphologiques par les méthodes DMEU et FMEU pour l'exemple VI.3.

a - Ensemble discret ( $R = 19$ ).

b - Détermination des domaines modaux par la méthode DMEU.

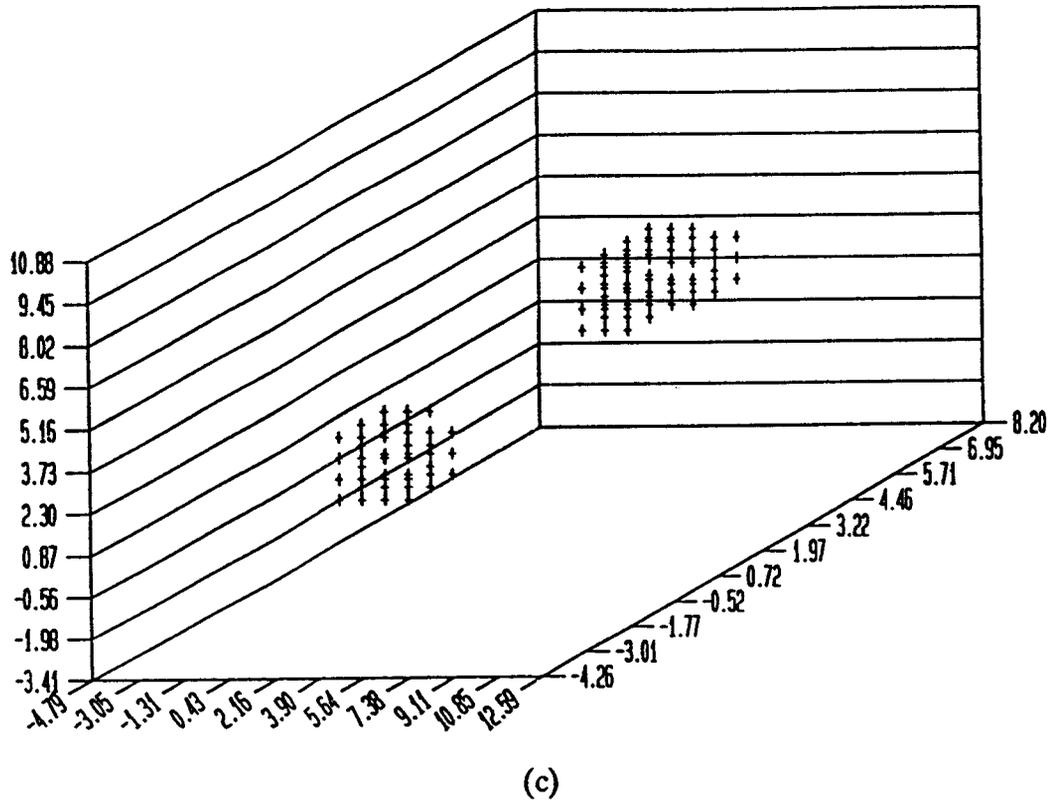


Figure VI.17 : Détermination des modes morphologiques par les méthodes DMEU et FMEU pour l'exemple VI.3 (suite).

c - Détermination des frontières modales par la méthode FMEU.

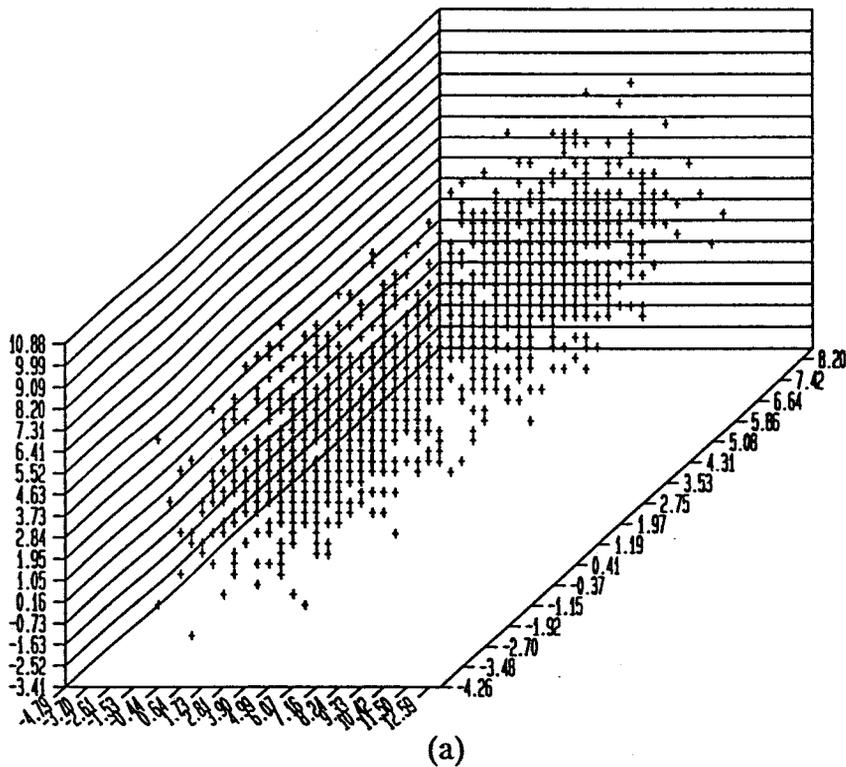


Figure VI.18 : Détermination des modes morphologiques par la méthode CVEM pour l'exemple VI.3.

a - Ensemble discret ( $R = 33$ ).

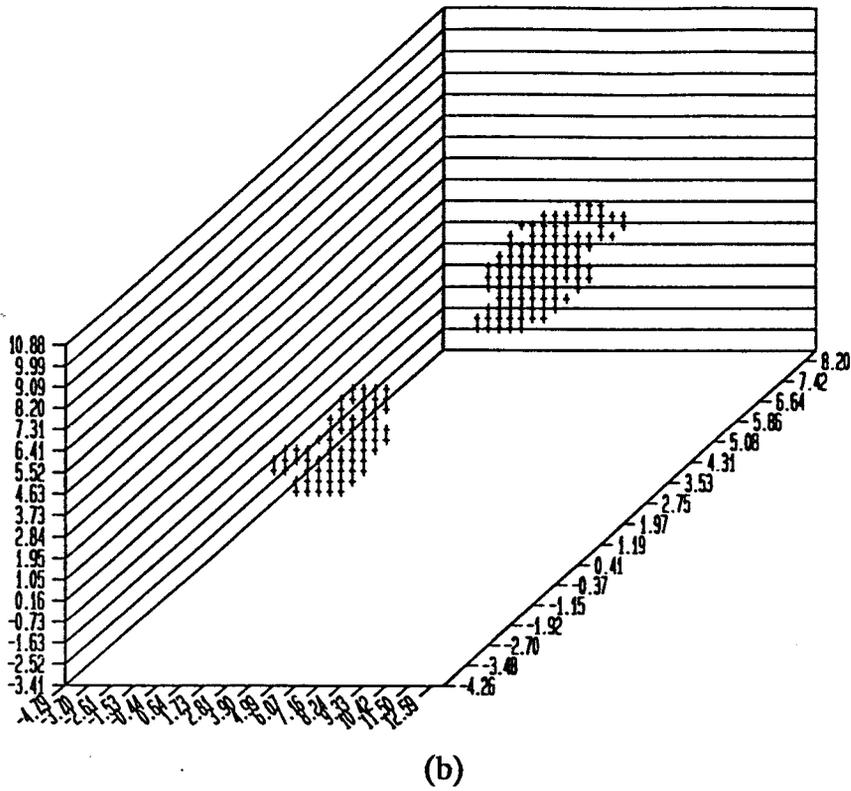


Figure VI.18 : Détermination des modes morphologiques par la méthode CVEM pour l'exemple VI.3 (suite).

b - Détermination des domaines modaux par la méthode CVEM.

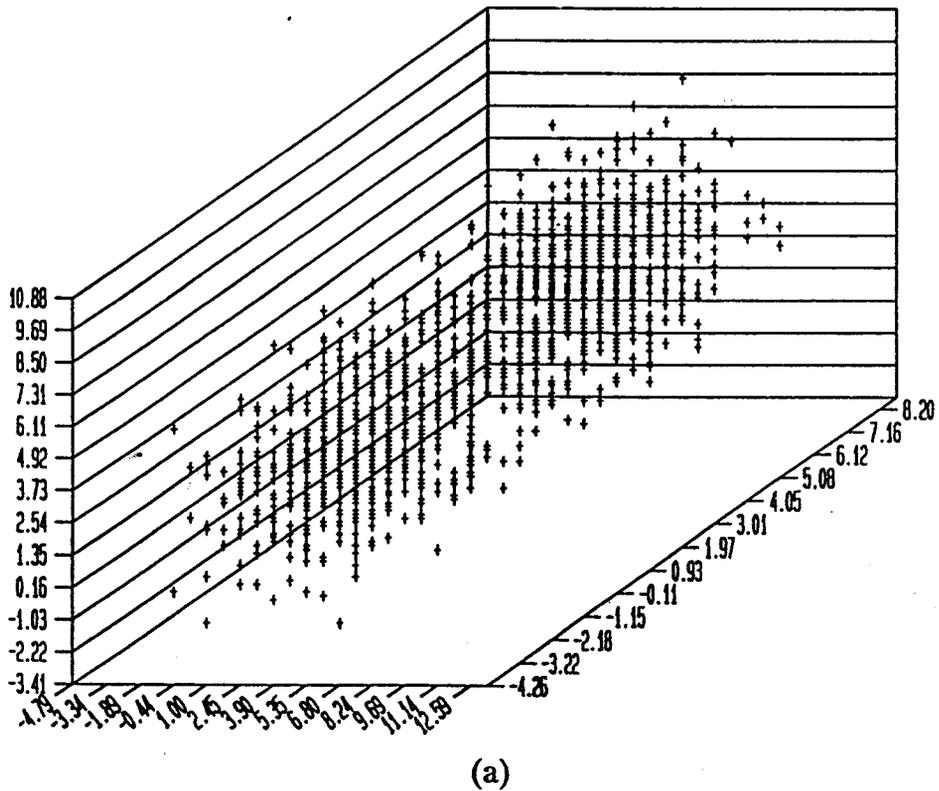
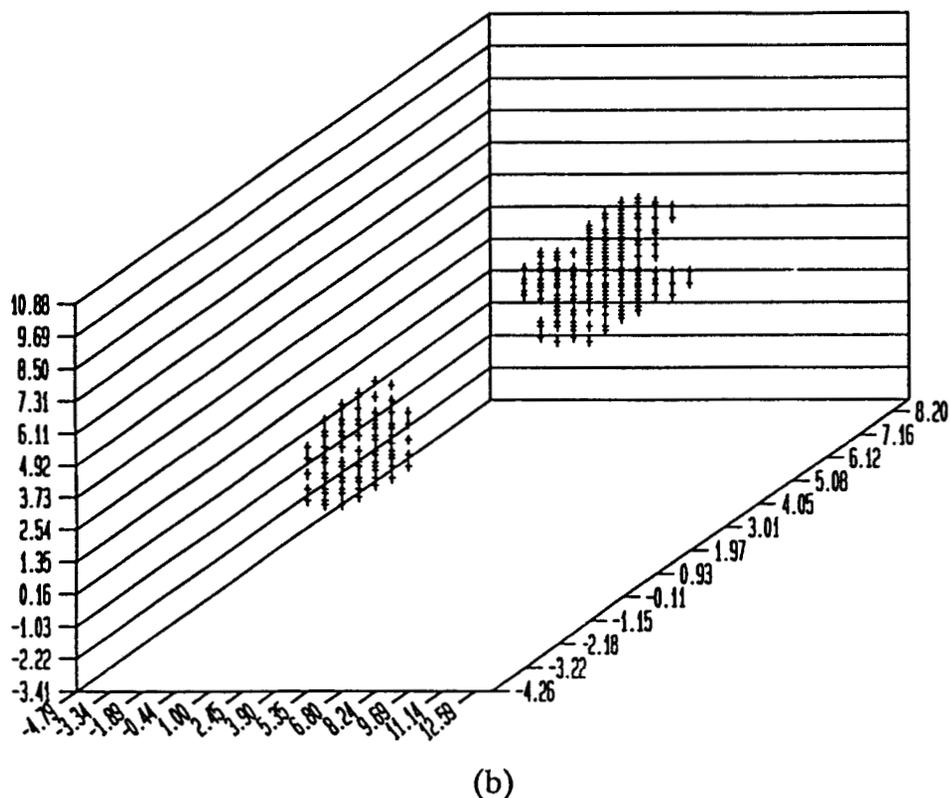


Figure VI.19 : Détermination des modes morphologiques par la méthode CMEM pour l'exemple VI.3.

a - Ensemble discret ( $R = 25$ ).



*Figure VI.19 : Détermination des modes morphologiques par la méthode CMEM pour l'exemple VI.3 (suite).  
b - Détermination des domaines modaux par la méthode CMEM.*

Nous avons associé à chacune des quatre méthodes de détection des modes morphologiques les quatre techniques de classification du chapitre V. Les résultats correspondants sont répertoriés sur la figure VI.20.

Nous avons également réalisé la classification des observations par les techniques Isodata et K-Means. La figure VI.21 présente l'appartenance des observations de l'échantillon à analyser (cf. figure VI.21.a), les résultats de la classification des observations par la technique Isodata (cf. figure VI.21.b), par la technique K-Means (cf. figure VI.21.c) et par la technique morphologique (DMEU + AONP) (cf. figure VI.21.d).

La table T6.7 montre que le taux d'erreur obtenu par la méthodologie morphologique est assez proche de ceux fournis par les méthodes Isodata et K-Means.

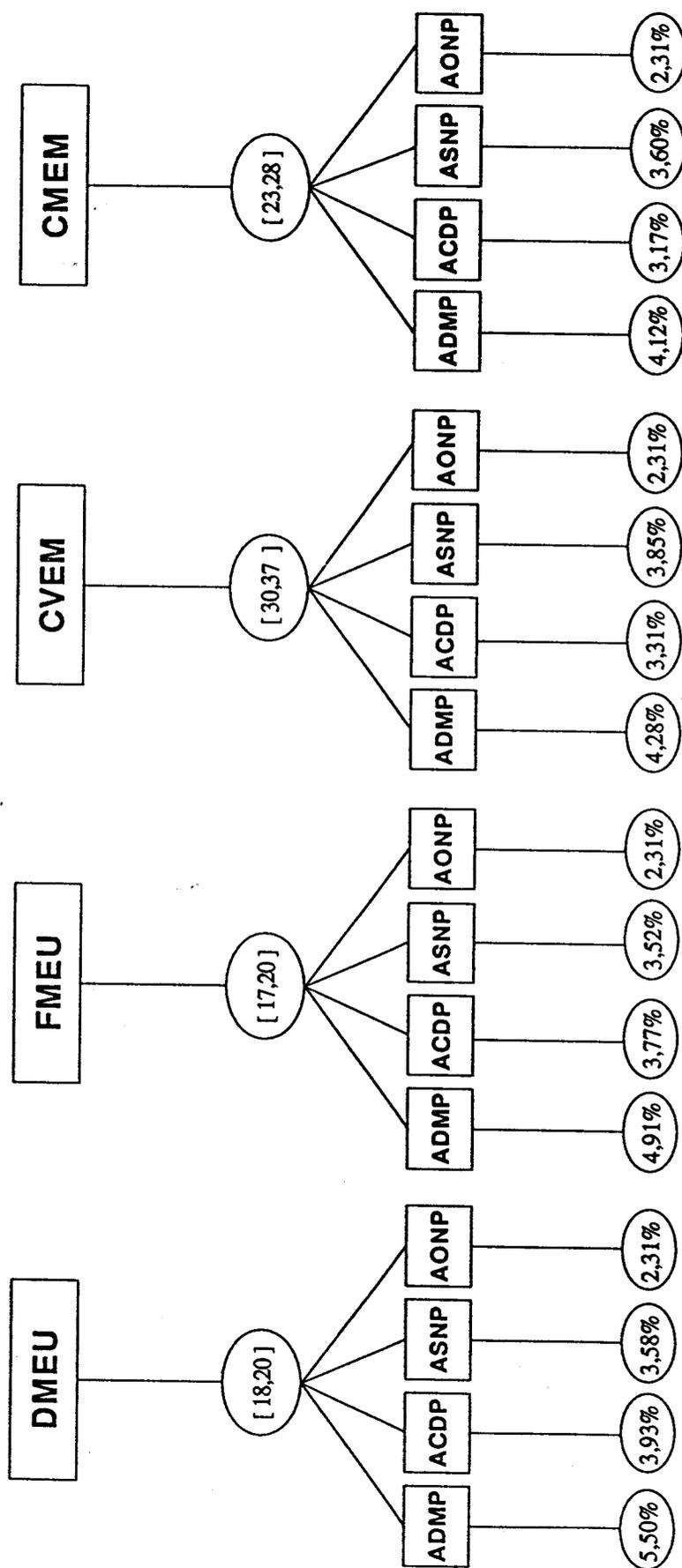
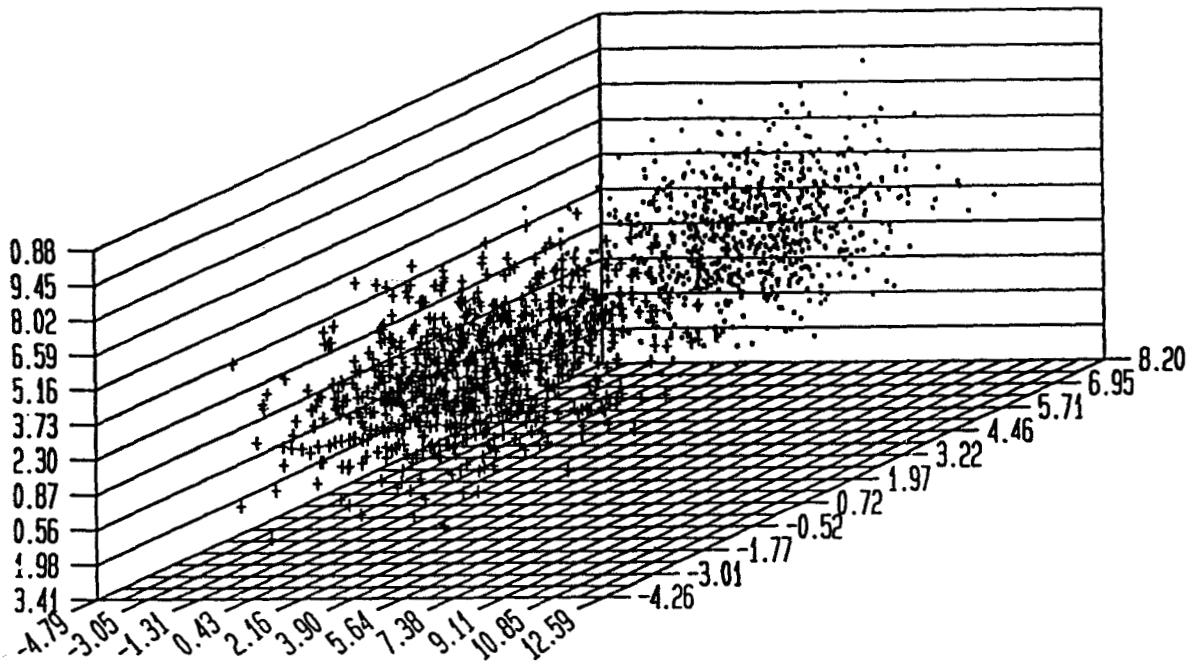
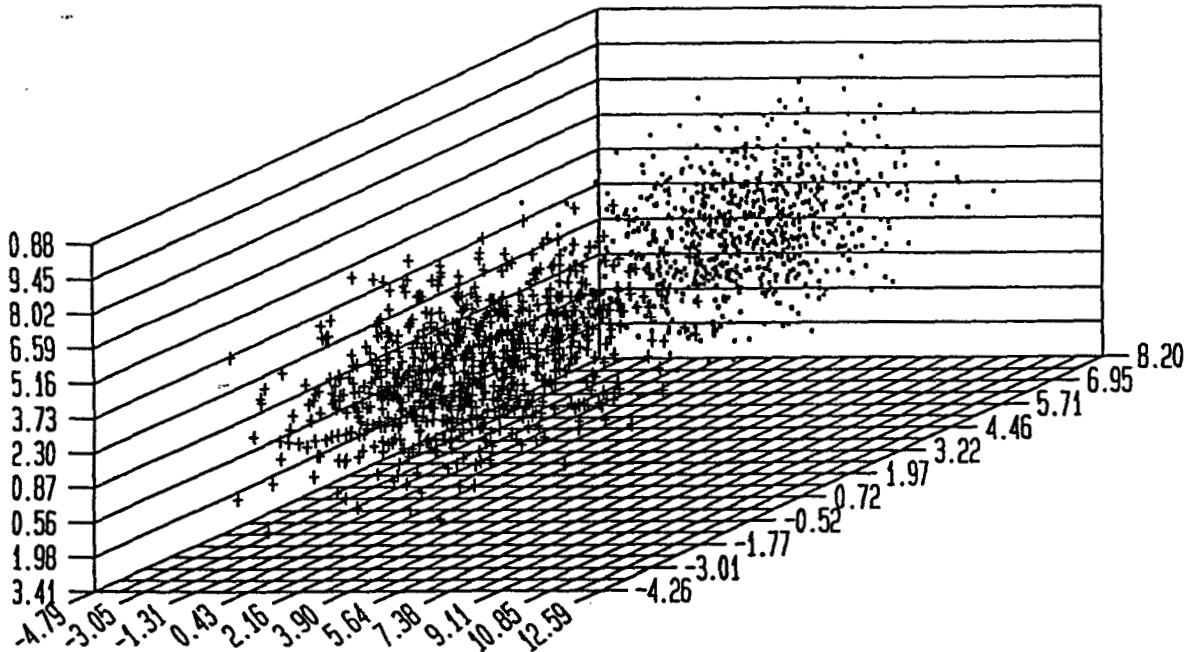


Figure VI.20 : Récapitulatif des résultats morphologiques de l'exemple VI.3.



(a)

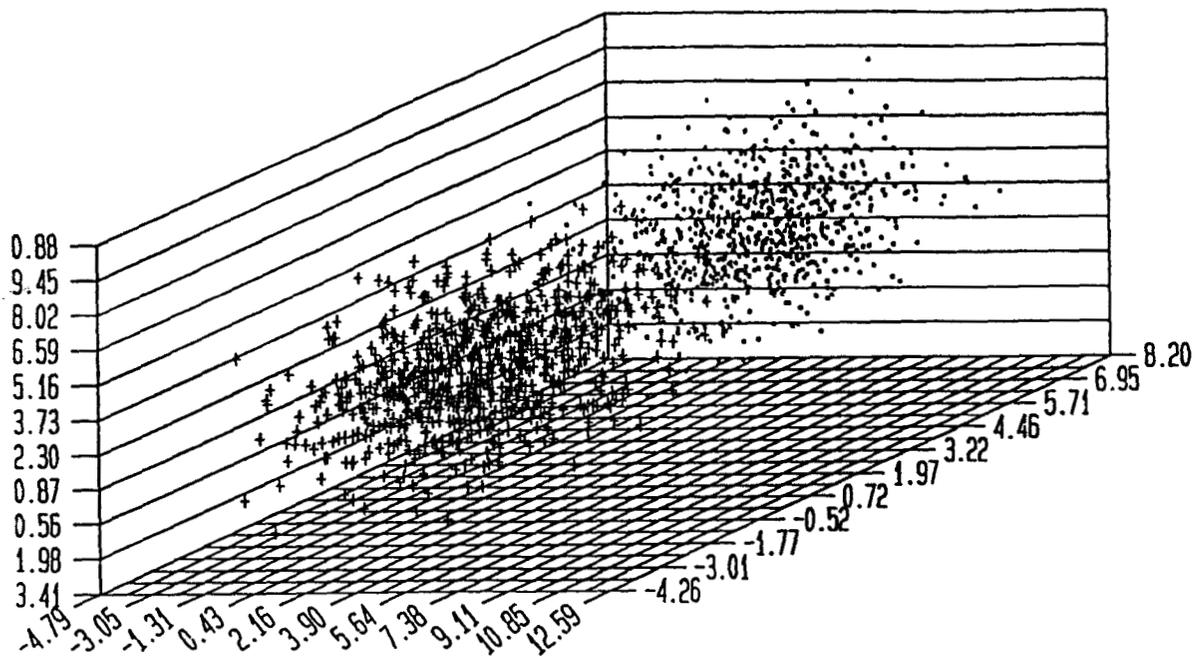


(b)

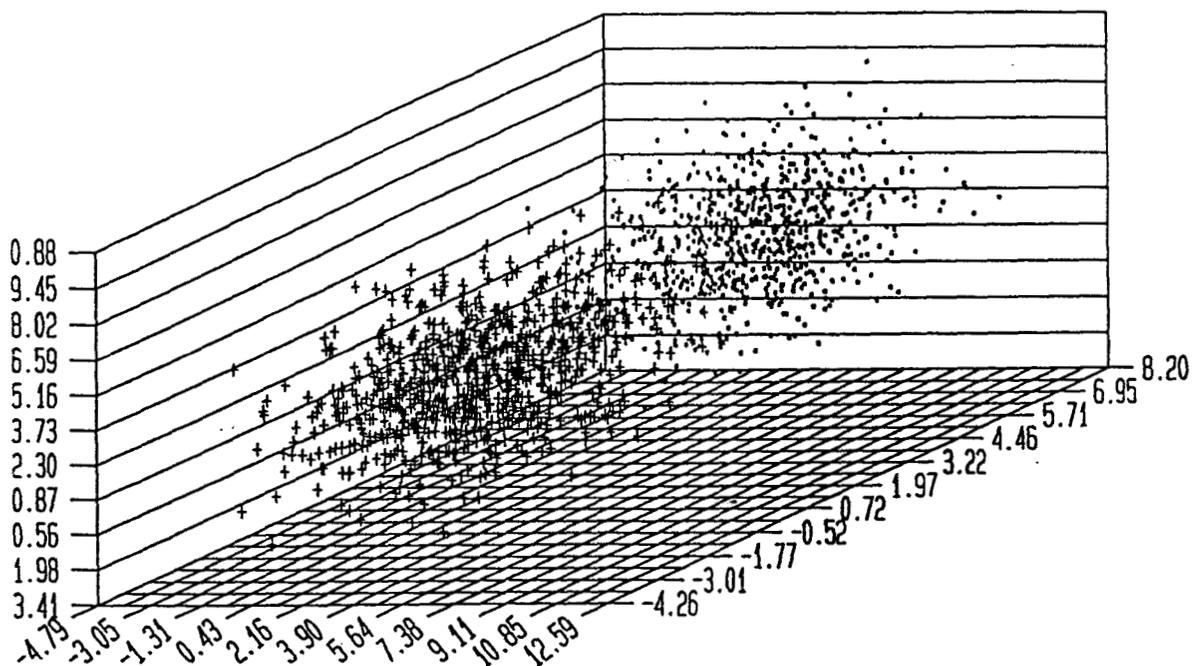
Figure VI.21 : Résultats de la classification des observations de l'exemple VI.3.

a - Appartenance des observations de l'exemple VI.3.

b - Résultats de la classification des observations par la technique Isodata.



(c)



(d)

**Figure VI.21 :** Résultats de la classification des observations de l'exemple VI.3 (suite).

c - Résultats de la classification des observations par la technique K-Means.

d - Résultats de la classification des observations par la technique morphologique (DMEU + AONP).

## VI.6. CONCLUSION.

Dans ce chapitre, nous avons présenté les résultats obtenus par la méthodologie morphologique que nous proposons dans ce mémoire à l'aide d'exemples bidimensionnels et multidimensionnels constitués de classes de formes variées.

Une comparaison de ces résultats à ceux fournis par deux méthodes classiques en analyse de données a permis de montrer l'intérêt d'une telle approche, essentiellement dans le cas de classes non sphériques. Enfin, dans le cas d'échantillons composés de classes sphériques, les résultats obtenus par les techniques classiques restent comparables à ceux résultant de l'emploi des techniques morphologiques.

**VII - CONCLUSION GENERALE.**

## CONCLUSION GENERALE.

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

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

A cet effet, nous avons, dans la deuxième partie de ce mémoire, présenté quatre méthodes distinctes d'extraction des modes morphologiques. Les deux premières (DMEU et FMEU), basées sur l'utilisation d'un élément structurant unique, sont bien adaptées aux classes de formes sphériques, présentant un degré de chevauchement assez important. Elles nécessitent néanmoins toutes deux un nombre d'observations d'autant plus important que la dimension des données est élevée. Elles seront donc plutôt réservées aux données de dimension réduite.

Les deux autres méthodes que nous avons développées (CVEM et CMEM) sont basées sur l'utilisation d'éléments structurants multiples constituant des familles structurantes. Ces méthodes sont bien adaptées à la détection de classes de formes non sphériques, mais à condition qu'elles présentent un degré de chevauchement plus faible que celui autorisé par l'emploi des deux méthodes précédentes. Par contre, elles

permettent l'analyse d'échantillons de plus petite taille et de dimension plus élevée que les méthodes basées sur l'utilisation d'un élément structurant unique.

Chacune de ces quatre méthodes d'extraction est précédée d'une phase de discrétisation de l'espace de représentation des données. Le réseau hypercubique ainsi introduit facilite l'utilisation des outils morphologiques binaires. Le seul paramètre à ajuster par l'opérateur durant cette phase d'extraction des modes est le pas de discrétisation. Nous avons proposé, dans ce mémoire, un moyen d'ajuster au mieux ce paramètre, mais une autre solution consisterait à prendre l'intersection de tous les domaines modaux obtenus pour la valeur du nombre de modes la plus fréquente. Par ailleurs, cette discrétisation de l'espace de représentation des données avec un pas fixe limite la détection correcte des modes morphologiques lorsqu'on se trouve en présence de groupements d'observations de densités très différentes. Il serait intéressant d'améliorer cette étape de discrétisation de l'espace de représentation des données en tenant compte des variations de densité dans la distribution des observations.

Au cours de cette étape de discrétisation de l'espace de représentation des données, nous avons uniquement pris en compte la présence ou l'absence d'au moins une observation dans chaque hypercube du réseau d'échantillonnage. Nous avons ainsi montré que les outils de la morphologie mathématique binaire pouvaient être favorablement adaptés à l'analyse de données, les différents modes morphologiques étant facilement individualisés. Ces résultats encourageants nous poussent à croire que la prise en compte de la densité locale sous-jacente à la distribution des observations, estimée en chaque point d'échantillonnage, ne pourrait que favoriser l'adaptation des outils morphologiques multiniveaux à l'analyse de données.

En ce qui concerne la deuxième phase de cette méthodologie morphologique présentée dans la troisième partie de ce mémoire, nous avons analysé quatre techniques de classification des observations après extraction des modes morphologiques. Les deux dernières techniques sont basées sur le principe classique de la classification par les  $k$  plus proches voisins. Tous les résultats que nous avons présentés dans ce mémoire

obtenus par cette technique ont été déterminés par la méthode du 1 plus proche voisin, ce qui correspond aux conditions les plus défavorables de l'utilisation de cette méthode. Nous avons fait ce choix pour montrer que même dans de telles conditions, les résultats sont satisfaisants.

Comme nous nous sommes situés dans un contexte non supervisé, le seul moyen d'évaluer les résultats obtenus par la méthodologie morphologique était de les comparer à ceux obtenus par des méthodes bien établies en analyse de données. Une confrontation avec les méthodes Isodata et des K-means a montré l'intérêt de l'approche morphologique, surtout dans le cas d'échantillons composés de classes non sphériques. Dans le cas d'échantillons composés de classes sphériques, nous ne pouvons que recommander l'utilisation de ces techniques classiques, qui donnent d'excellents résultats difficiles à surpasser.

Un dernier point important est celui du nombre d'attributs décrivant chacune des observations de l'échantillon à analyser. Nous avons précisé, au long de ce mémoire, que certaines techniques d'extraction des modes étaient mal adaptées à l'analyse de données de dimension très élevée. Dans ce cas, il serait peut-être intéressant de diminuer la dimension de l'espace de représentation des données, soit par une technique classique d'analyse en composantes principales, soit mettant en oeuvre des réseaux neuronaux multicouches avant d'utiliser ces techniques morphologiques.

Ce travail, ainsi que ses prolongements en cours de développement, montre que la détection des modes morphologiques associés aux différentes classes constitue une approche intéressante aux problèmes de la classification automatique non supervisée.

**REFERENCES BIBLIOGRAPHIQUES.**

## REFERENCES BIBLIOGRAPHIQUES

- [AHM - 75] N. AHMED & K. R. RAO.  
**Orthogonal transforms for digital signal processing.**  
*Springer, Berlin, 1975.*
- [ASS - 89] J. C. ASSELIN DE BEAUVILLE.  
**Panorama de l'utilisation du mode en classification automatique.**  
*RAIRO-APII, AFCET, N° 23, pp. 113-137, 1989.*
- [BAL - 65] G. H. BALL.  
**Data analysis in the social sciences : what about the details ?**  
*Proc. F.J.C.C., pp. 533-560, Spartan Books, Washington, D.C., 1965.*
- [BAL - 67] G. H. BALL & D. J. HALL.  
**A clustering technique for summarising multivariate data.**  
*Behavioral Science, Vol. 12, pp. 153-155, 1967.*
- [BAY - 80] C. K. BAYNE, J. J. BEAUCHAMP, C. L. BEGOVITCH, & V. E. KANE.  
**Monte Carlo comparison of selected clustering procedures.**  
*Pattern Recognition, Vol. 12, pp. 51-62, 1980.*
- [BOC - 79] H. H. BOCK.  
**Clustering by density estimation.**  
*Analyse de Données et Informatique, INRIA, pp. 173-186, 1979.*
- [CHA - 74] C. L. CHANG.  
**Finding prototypes for nearest neighbor classifiers.**  
*I.E.E.E. Trans. Comput., Vol. C-23, pp. 1179-1184, 1974.*
- [COO - 64] D. B. COOPER & P.W. COOPER.  
**Nonsupervised adaptive signal detection and pattern recognition.**  
*Information & Control, Vol. 7, pp. 416-444, 1964.*
- [COO - 67] P. W. COOPER.  
**Some topics on nonsupervised adaptive signal detection for multivariate normal distributions.**  
*Computer & Informations Sciences - II, pp. 123-146, Academic Press, New York, 1967.*
- [COO - 71] W. W. COOLEY & P. LOHNES.  
**Multivariate data analysis.**  
*Wiley, New York, 1971.*
- [COS - 85] M. COSTER & J.-L. CHERMANT.  
**Précis d'Analyse d'Images.**  
*Edition du C.N.R.S., 1985.*

- [COV - 67] T. M. COVER & P.E. HART.  
**Nearest neighbour pattern classification.**  
*I.E.E.E. Trans. Info. Theory, Vol. IT - 13, N° 1, pp. 21-27, 1967.*
- [DAL - 62] R. F. DALY.  
**The adaptative binary-detection problem on the real line.**  
*Technical Report 2003-3, Stanford University, Stanford, Calif., 1962.*
- [DAV - 79] D. L. DAVIES & D. W. BOULDIN.  
**A cluster separation measure.**  
*I.E.E.E. Trans. Pattern Anal. Machine Intell., Vol. PAMI-1, N° 2, pp. 224-227, 1979.*
- [DAY - 69] N. E. DAY.  
**Estimating the components of a mixture of normal distributions.**  
*Biometrika, Vol. 56, pp. 463-474, 1969.*
- [DEV - 82] P. A. DEVIJVER & J. KITTLER.  
**Pattern recognition : a statistical approach.**  
*Prentice-Hall, Englewood Cliffs, New Jersey, 1982.*
- [DID - 71] E. DIDAY.  
**Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques.**  
*Rev. Stat. Appl., Vol. 19, N° 2, pp. 20-33, 1971.*
- [DID - 79] E. DIDAY & Coll.  
**Optimisation en classification automatique.**  
*Tomes 1 & 2, INRIA, 1979.*
- [DID - 82] E. DIDAY, J. LEMAN, J. POUGET & F. TESTU.  
**Eléments d'analyse de données.**  
*Bordas, Paris, 1982.*
- [DOU - 91a] M.-F. DOUTRE.  
**Approche scientifique du contrôle qualité dans l'industrie intégrant les techniques de MSP, d'analyse de données et de reconnaissance des formes.**  
*Thèse de Doctorat, Université de Compiègne, 1991.*
- [DOU - 91b] M.-F. DOUTRE & B. DUBUISSON.  
**Classification by morphological boundaries. Application in the industrial domain.**  
*Symbolic-Numeric Data Analysis & Learning, INRIA, pp. 103-116, 1991.*

- [DUB - 90] B. DUBUISSON.  
**Diagnostic et reconnaissance des formes.**  
*Hermés, Paris, 1990.*
- [DUD - 73] R. O. DUDA & P. E. HART.  
**Pattern classification and scene analysis.**  
*Wiley, New York, 1973.*
- [EIG - 74] D. J. EIGEN, F. R. FROMM & R. A. NORTHOUSE.  
**Cluster analysis based on dimensionnal information with applications to feature selection and classification.**  
*I.E.E.E. Trans. Syst., Man, & Cybern., Vol. SMC-4, N° 3, pp. 284-294, 1974.*
- [FOR - 74] F. R. FROMM & R. A. NORTHOUSE.  
**Class : A non parametric clustering algorithm.**  
*Pattern Recognition, Vol.8, pp. 107-114,1974.*
- [FRI - 67] H. P. FRIEDMAN & J. RUBIN.  
**On some invariant criteria for grouping data.**  
*J. American Statistical Assn, Vol. 62, pp. 1159-1178, 1967.*
- [FUK - 70] K. FUKUNAGA & W. L. G. KOONTZ.  
**A criterion and an algorithm for grouping data.**  
*I.E.E.E. Trans. Comput., Vol. C-19, pp. 917-923, 1970.*
- [FUK - 71] K. FUKUNAGA & D. R. OLSEN.  
**A two-dimensional display for the classification of multivariate data.**  
*I.E.E.E. Trans. Comput., Vol. C-20, pp. 917-923, 1971.*
- [FUK - 75a] K. FUKUNAGA & L. D. HOSTETLER.  
**The estimation of the gradient of a density function with applications in pattern recognition.**  
*I.E.E.E. Trans. Info. Theory, Vol. IT-21, N° 1, pp. 32-40, 1975.*
- [FUK - 75b] K. FUKUNAGA & P. M. NARENDRA.  
**A branch and bound algorithms for computing k-nearest neighbors.**  
*I.E.E.E. Trans. Comput., pp. 750-753, 1975.*
- [FUK - 82] K. FUKUNAGA & J. M. MANTOCK.  
**A nonparametric two-dimensional display for classification.**  
*I.E.E.E. Trans. Pattern Anal. Machine Intell., Vol. PAMI-4, pp. 427-436, 1982.*
- [FUK - 84] K. FUKUNAGA & J. M. MANTOCK.  
**Nonparametric Data Reduction.**  
*I.E.E.E. Trans. Pattern Anal. Machine Intell., Vol. PAMI-6, N° 1, pp. 115-118, 1984.*

- [GOR - 87] A. D. GORDON.  
**Parsimonious trees.**  
*Journal of Classification*, pp. 85-101, 1987.
- [GOW - 66] J. C. GOWER.  
**Some distance properties of latent root and vector methods in multivariate analysis.**  
*Biometrika*, Vol. 53, pp. 325-338, 1966.
- [HAR - 87] R. M. HARALICK, S. R. STERNBERG, & X. ZHUANG.  
**Image analysis using mathematical morphology.**  
*I.E.E.E. Trans. Pattern Anal. Machine Intell.*, Vol. PAMI-9, N° 4, pp. 532-550, 1987.
- [HAS - 66] V. HASSELBALD.  
**Estimation of parameters for a mixture of normal distributions.**  
*Technometrics*, Vol. 8, pp. 431-444, 1966.
- [HEI - 90] H. J. A. M. HEIJMANS & C. RONSE.  
**The Algebraic Basis of Mathematical Morphology - I : Dilations and Erosions.**  
*Computer Vision, Graphics, & Signal Processing*, Vol. 50, pp. 245-295, 1990.
- [HIL - 68] C. G. HILLBORN & D. G. LAINIOTIS.  
**Optimal unsupervised learning multicategory dependent hypotheses pattern recognition.**  
*I.E.E.E. Trans. Info. Theory*, Vol. IT - 14, pp. 468-470, 1968.
- [JAI - 88] A. K. JAIN & R. C. DUBES.  
**Algorithms for clustering data.**  
*Prentice-Hall, Englewood Cliffs, New Jersey*, 1988.
- [JAM - 78] M. JAMBU.  
**Classification automatique pour l'analyse de données .**  
**1 - Méthodes et algorithmes.**  
*Dunod, Paris*, 1978.
- [JON - 68] K. L. JONES.  
**Problems of grouping individuals and the method of modality.**  
*Behavioral Science*, Vol. 13, pp. 496-511, 1968.
- [KIT - 76] J. KITTLER.  
**A locally sensitive method for cluster analysis.**  
*Pattern Recognition*, Vol. 8, pp. 23-33, 1976.
- [KOO - 76] W. L. G. KOONTZ, P. M. NARENDRA, K. FUKUNAGA.  
**A graph theoretic approach to nonparametric cluster analysis.**  
*I.E.E.E. Trans. Comp.*, Vol. C - 25, N° 9, pp. 936-944, 1976.

- [KRU - 64] J. B. KRUSKAL.  
**Nonmetric multidimensional scaling : a numerical method.**  
*Psychometrika, Vol. 29, pp. 115-129, 1964.*
- [KRU - 77] J. B. KRUSKAL.  
**Multidimensional scaling and other methods for discovering structures.**  
*Statistic methods for Digital Computers, Vol. 3, pp. 296-339, Wiley, New York, 1977.*
- [LAN - 67] G. N. LANE & W. T. WILLIAMS.  
**A general theory of classificatory sorting strategies 1 - Hierarchical systems.**  
*Computer J., Vol. 9, pp. 973-980, 1967.*
- [LER - 91] I. C. LERMAN & N. GHAZZALI.  
**What do we retain from a classification tree?  
 An experiment in image coding.**  
*Symbolic-Numeric Data Analysis & Learning, INRIA, pp. 27-42, 1991.*
- [LOF - 65] D. O. LOFTSGAARDEN & C. P. QUESENBERNY.  
**A nonparametric estimate of a multivariate density function.**  
*Ann. Math. Stat. Vol. 36, pp. 1049-1051, 1965.*
- [LUK - 79] A. LUKASOVA.  
**Hierarchical agglomerative clustering procedure.**  
*Pattern Recognition, Vol. 11, pp. 365-381, 1979.*
- [MAC - 67] J. MACQUEEN.  
**Some methods for classification and analysis of multivariate observations.**  
*Proc. 5th, Berkeley Symposium on Math. Stat. and Prob., Vol. I, pp. 281-297, University of California Press, Berkeley & Los Angeles, Calif., 1967.*
- [MAK - 77] U. E. MAKOV & A. F. M. SMITH.  
**A quasi-Bayes unsupervised learning procedure for priors.**  
*I.E.E.E. Trans. Info. Theory, Vol. IT- 24, N° 6, pp. 761-764, 1977.*
- [MAR - 86] P. MARAGOS & R. W. SCHAFER.  
**Morphological Skeleton Representation and Coding of Binary Images.**  
*I.E.E.E. Trans. Ac., Speech, Signal Proc., Vol. ASSP-34, N° 5, pp. 1228-1244, 1987.*

- [MAR - 87a] P. MARAGOS & R. W. SCHAFER.  
**Morphological Filters - Part 1 : Their Set-Theoretic Analysis and Relations to Linear Shift-Invariant Filters.**  
*I.E.E.E. Trans. Ac., Speech, Signal Proc., Vol. ASSP-35, N° 8, pp. 1153-1169, 1987.*
- [MAR - 87b] P. MARAGOS & R. W. SCHAFER.  
**Morphological Filters - Part 2 : Their Relations to Median, Order-Statistic, and Stack Filters.**  
*I.E.E.E. Trans. Ac., Speech, Signal Proc., Vol. ASSP-35, N° 8, pp. 1170-1184, 1987.*
- [MAT - 75] G. MATHERON.  
**Random sets and integral geometry.**  
*John Wiley, New York, 1975.*
- [MIZ - 75] R. MIZOGUCHI & M. SHIRUMA.  
**An approach to unsupervised learning classification.**  
*I.E.E.E. Trans. Comput., Vol. C-24, N° 10, pp. 979-983, 1975.*
- [NIE - 88] H. NIEMANN & R. GOPPERT.  
**An efficient branch-and-bound nearest neighbour classifier.**  
*Pattern Recognition Letters, Vol. 7, pp. 67-72, 1988.*
- [OLE - 88] S. OLEJNIK.  
**Analyse de la convexité d'une fonction de densité de probabilité par étiquetage probabiliste : application à la classification automatique non supervisée.**  
*Thèse de 3<sup>e</sup> cycle, 1988.*
- [PAR - 62] E. PARZEN.  
**On estimation of a probability density function and mode.**  
*Ann. Math. Stat., Vol. 33, pp. 1065-1076, 1962.*
- [PIT - 90] I. PITAS & A. N. VENETSANOPOULOS.  
**Morphological Shape Decomposition.**  
*I.E.E.E. Trans. Pattern Anal. Machine Intell., Vol. PAMI-12, N° 1, pp. 38-45, 1990.*
- [POS - 81] J.-G. POSTAIRE & C. P. A. VASSEUR.  
**An approximate solution to normal mixture identification with application to unsupervised pattern classification.**  
*I.E.E.E. Trans. Pattern Anal. Machine Intell., Vol. PAMI-3, N° 2, pp. 163-179, 1981.*
- [POS - 82a] J.-G. POSTAIRE & C. VASSEUR.  
**A Fast Algorithm for Non Parametric Probability Density Function.**  
*I.E.E.E. Trans. Pattern Anal. Machine Intell., Vol. PAMI-4, N° 6, pp. 663-666, 1982.*

- [POS - 82b] J.-G. POSTAIRE.  
**Optimisation du processus de classification automatique par analyse de la convexité des fonctions de densité.**  
*Thèse d'Etat, Université de Lille, 1981.*
- [POS - 89] J.-G. POSTAIRE & A. TOUZANI.  
**Mode boundary detection by relaxation for cluster analysis.**  
*Pattern Recognition letters, Vol. 9, pp. 1-12, 1989.*
- [SAM - 69] J. W. SAMMON.  
**A non-linear mapping for data structure analysis.**  
*I.E.E.E. Trans. Comput., Vol. C-18, pp. 401-409, 1969.*
- [SAM - 70] J. W. SAMMON.  
**An optimal discriminant plane.**  
*I.E.E.E. Trans. Comput., Vol. C-19, pp. 826-829, 1970.*
- [SCH - 76] A. SCHROEDER.  
**Analyse d'un mélange de distributions de probabilité de même type.**  
*Rev. Statist. App., Vol. 24, N° 1, pp. 39-62, 1976.*
- [SEB - 84] G. A. F. SEBER.  
**Multivariate observations.**  
*Wiley, New York, 1971.*
- [SER - 82] J. SERRA.  
**Image analysis and mathematical morphology.**  
*Academic Press, New York, 1982.*
- [SHE - 62] R. N. SHEPARD.  
**The analysis of proximities : multidimensional scaling with an unknown distance function.**  
*Psychometrika, Vol. 27, pp. 125-140, 1962.*
- [SOK - 63] R. R. SOKAL & P. H. A. SNEATH.  
**Principles of numerical taxonomy.**  
*W. H. Freeman, San Francisco, Cali., 1963.*
- [SON - 90] J. SONG & E. J. DELP.  
**The Analysis of Morphological Filters with Multiple Structuring Elements.**  
*Computer Vision, Graphics, & Image Processing, Vol. 50, pp. 308-328, 1990.*
- [STE - 82] S. R. STERNBERG.  
**Cellular computers and biomedical image processing.**  
*Biomedical Images & Computers, Vol. 17, pp. 294-319, Springer-Verlag, Berlin, 1982.*

- [STE - 86] S. R. STERNBERG.  
**Grayscale Morphology.**  
*Computer Vision, Graphics, & Image Processing, Vol. 35,*  
*pp. 335-355, 1986.*
- [TOU - 79] J.T. TOU.  
**Dynoc, a dynamic optimal cluster-seeking technique.**  
*Int. J. ; Comput. Inf. Sci., Vol. 8, N° 6, pp. 541-547, 1979.*
- [TOU - 87] A. TOUZANI.  
**Classification automatique par détection des contours des modes des fonctions de densité de probabilité multivariées et étiquetage probabiliste.**  
*Thèse d'Etat, Université de Lille, 1987.*
- [TOU - 88] A. TOUZANI & J.-G. POSTAIRE.  
**Mode detection by relaxation.**  
*I.E.E.E. Trans. Pattern Anal. Machine Intell., Vol. PAMI-10,*  
*pp. 970-978, 1988.*
- [TOU - 89] A. TOUZANI & J.-G. POSTAIRE.  
**Clustering by mode boundary detection.**  
*Pattern Recognition letters, Vol. 9, pp. 1-12, 1989.*
- [ULL -74] J. R. ULLMANN.  
**Automatic selection of reference data for use in a nearest-neighbor method of pattern classification.**  
*I.E.E.E. Trans. Info. Theory, pp. 541-543, 1974.*
- [VAS - 80] C. VASSEUR & J.-G. POSTAIRE.  
**A convexity testing method for cluster analysis.**  
*I.E.E.E. Trans. Syst. Man. & Cybern., Vol. SCM - 10, N°3,*  
*pp. 145-149, 1980.*
- [YUN - 76] T. P. YUNCK.  
**A technique to identify nearest neighbors.**  
*I.E.E.E. Trans. Syst., Man, & Cybern., Vol. SMC-6, N° 10,*  
*pp. 284-294, 1974.*
- [WOL - 70] J. H. WOLFE.  
**Pattern clustering by multivariate mixture analysis.**  
*Multi. Behav. Res., Vol. 5, pp. 329-350, 1970.*

## REFERENCES LIEES A CE TRAVAIL :

C. BOTTE-LECOCQ & J.-G. POSTAIRE.

**Classification automatique par extraction des composantes connexes d'un ensemble discret binaire représentatif des données.**

*8<sup>e</sup> Congrès Reconnaissance des Formes et Intelligence Artificielle, A.F.C.E.T., Lyon, 1991.*

R. D. ZHANG, C. BOTTE-LECOCQ & J.-G. POSTAIRE.

**Mode boundary extraction by binary morphology for cluster analysis.**

*3<sup>rd</sup> Conference of International Federation of Classification Societies, Edinburgh, 1991.*

J.-G. POSTAIRE, R. D. ZHANG & C. BOTTE-LECOCQ.

**Cluster analysis by binary morphology.**

*I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence, a paraître.*

C. BOTTE-LECOCQ & J.-G. POSTAIRE.

**Iterations of morphological transformations for cluster separation in pattern classification.**

*Symbolic Numeric Data Analysis and Learning, I.N.R.I.A., 1991.*

