Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Mouloud Mammeri, Tizi-Ouzou
Faculté de Génie Electrique et d'Informatique
Département Automatique

MEMOIRE

De fin d'études
en vue de l'obtention du diplôme d'ingénieur d'état
En Automatique

THEME

Evaluation des résultats de la segmentation d'images basée sur des indices de validité

Proposé et Dirigé par

Mr HAMMOUCHE KAMAL

Présenté par

Mr GUERBAS RABAH

Mr HAMMOUCHE SOFIANE

Promotion 2007-2008

Sommaire

Introduction générale....4

Chapitre I : Segmentation d'image
I.1- Introduction....6
I.2- Segmentation....6
I.3- Binarisation d'une image....7
I.4- Différentes approches des méthodes de segmentation....7
I.4.1- Approche par région....8
I.4.1.1- Méthode par division (partage)....8
I.4.1.2- Méthodes par fusion....9
I.4.1.3- Méthodes par division-fusion (Split and merge)....10
I.4.1.4 Méthodes par croissance de régions....11
I.4.2- Approche par contour....12
I.4.3- Approche par classification des pixels....13
I.4.3.1- Segmentation d'image par seuillage d'histogramme....13
I.4.3.1.a-Définition du seuillage....13
I.4.3.1.b -Définition de l'histogramme....14
I.4.3.2- Méthodes de calcul des seuils....15
I.4.3.2.1- Méthode basée sur l'analyse discriminante....16
I.4.3.2.2- Méthode de seuillage global basées sur l'entropie....17
I.4.3.2.2.a- Méthode de Kapur, Sahoo et Wang....17
I.4.3.2.2.b- Méthode de cross entropy....18
I.4.3.2.2.c- Méthode de corrélation entropique....18
I.4.3.3-Seuillage par minimisation de l'erreur quadratique....19
I.4.3.4- Multi seuillage des images....21
I.5- Algorithme itératif....22
I.6- Algorithme génétique....23
I.6.1- Codage chromosomes....23
I.6.2- Génération de la population initiale....24
I.6.3- Evaluation....24
I.6.4- Sélection....24
I.6.5- Reproduction....24
I.6.6- Croisement....24
I.6.7- Mutation....25
I.6.8- Remplacement....26
I.7- Conclusion....26

Chapitre II : Classification et indices de validités
II.1- Introduction....27
II.2- Définition de la classification....27
II.3- Définition d'une classe....29
II.4- Algorithme de C- Means....30
II.5- Classification floue....32
II.5.1-Introducion à la notion floue....32
II.5.2- L'algorithme de C-moyens flou (FCM)....33
II.6- Notion de validité en classification....36
II.7- Indices de validité....37
II.7.1- Indice de bezdek....38

II.7.1.a- Indice du coefficient de partition (PC) 38
II.7.1.b- Indice de l'entropie moyenne de la partition (PE) 38
II.7.2- Indices de Gath et Geva (*FH*) 38
II.7.3- Indice de Xie et Beni (XB) 39
II.7.4- Indice de Kwon (*KW*) 40
II.7.5- Indice de Maulik (*I*) 40
II.7.6- L'indice de Fukuyama et Sugeno (FS) 41
II.7.7- L'indice de Wu et Yang (PCAES) 41
II.7.8- Indice de Calinski Harabasz (CH) 42
II.7.9- Indice de Davies-Bouldin (DB) 42
II.7.10- Indice de Dunn (D) 43
II.7.11- Indice de Razae (R) 43
II.7.12- Indice de Boudraa (B) 44
II.7.13- Indice de De Franco (Icc) 45
II.7.14- Indice de Turi (V) 45
II.7.15-L'indice VCR 46
II.7.16-L'indice de Krzanowiski et lai (KL) 46
II.7.17-L' indices RMSSDT 46
II.7.18-L'indice de Chou et Sun (CS) 47
II.8- Conclusion 47

Chapitre III : Adaptation des indices de validités au seuillage d'histogramme
III.1- Introduction 49
III.2- Indices de Bezdek 49
III.2.a- Indice du coefficient de partition (PC) 49
III.2.b- Indice de l'entropie moyenne de la partition (PE) 49
III.3- Indice de Gath et Geva (FH) 50
III.4- Indice de Xie et Beni (XB) 51
III.5- Indice de Kwon (*KW*) 51
III.6- Indice de Maulik (*I*) 51
III.7- Indice de Fukuyama et Sugeno (FS) 52
III.8- Indice de WU et Yan (PCAES) 52
III.9- Indice de Calinski Harabasz (CH) 53
III.10- Indice de Davies-Bouldin (DB) 53
III.11- Indice de Dunn (D) 54
III.12- Indice de Razae (R) 54
III.13- Indice de Boudraa (B) 55
III.14- Indice de De Franco (Icc) 55
III.15- Indice de Turi (V) 56
III.16- Indice (VCR) 56
III.17 -Indice de Krzanowiski et lai (KL) 56
III.18- Indices de RMSSTD 57
III.19-Indice de Chou et Sun (CS) 57
III.20- Indice spécifiques en segmentation 58
III.20.1-Indice de Deng et al (J) 58
III.20.2- L'indice Yen de chang (*F*) 58
III.21-Conclusion 59

Chapitre IV : Tests & résultats

IV.1- Introduction....60
IV.2- Histogrammes artificiels....60
Exemple 1....60
Exemple 2....70
IV.3- Images réelles....78
Image 1....78
Image 2....87
IV.4- Conclusion....91
Conclusion générale....92

Introduction générale

Le traitement des images constitue actuellement l'une des grandes orientations de traitement de l'information. L'identification et la reconnaissance des éléments d'une scène ou d'une image donnée requièrent une extraction préalable des différent objets composant cette dernière. De ce besoin est née la segmentation d'image qui constitue une étape cruciale de l'analyse d'image. Son objectif consiste à décomposer des scènes plus complexes en des éléments individuellement identifiable ou plus aisément traitables que l'image entière. Une manière simple de segmenter une image consiste à seuiller son histogramme, celle-ci consiste à déterminer un ensemble de seuils à partir des quels les pixels de l'image sont affectés à un ensemble de classe. La recherche des seuils est réalisée on optimisant une fonction objective définie à partir des caractéristiques de l'image. La segmentation permet en d'autres termes de séparer les objets du fond.

Lorsque les objets sont caractérisés par une même luminance différente de celle du fond un seul seuil est exigé pour extraire ces objets. Mais lorsque on est en présence d'images contenant plusieurs objet différents leurs extraction nécessitent le calcule de plusieurs seuils.

La recherche d'un seul seuil peut être effectué par une méthode exhaustive alors que la recherche de plusieurs seuils nécessite un algorithme adapté.

Ainsi la reconnaissance du nombre de seuils au préalable est plus que nécessaire, car il conditionne ce résultat de la segmentation.

Nous proposons dans ce mémoire d'étudier un ensemble d'indices dans le but de déterminer le nombre optimal de seuils. Ces indices proviennent du domaine de la classification des données dans le quel il sont dénommés ''indices de validité''. La classification est définie comme un processus de regroupement d'un ensemble d'objets en groupes ou classes tel que les objets d'une même classe sont similaires alors que les objets de classes différentes sont dissimilaires.

Comme pour la segmentation la connaissance ou préalable du nombre de classes est primordial. Dans le cadre de la classification les indices de validités permettent de calculer le nombre optimal de classes dans un ensemble de données à classer. La classification et la segmentation sont des notions très similaires. Par conséquent les indices de validités utilisés en classification peuvent être adaptés pour le calcul du nombre de seuils.

Notre plan de travail a été réparti comme suit :

Dans le premier chapitre, nous présenterons les différentes approches des méthodes de segmentation d'image. Quelques méthodes de seuillage des images en niveau de gris, ainsi que deux méthodes de calcul de seuils, à savoir un algorithme génétique et un algorithme itératif seront présentés. Le deuxième chapitre, décrit quelques notions sur la classification des données, une panoplie d'indices de validité utilisés en classification sera également présentée.

Le troisième chapitre est consacré à l'adaptation des indices de validités définis dans le chapitre précédant pour le cas du seuillage d'histogramme d'une image. Les tests effectués et les résultats obtenus seront présentés et interprétés dans le quatrième chapitre. Nous terminerons notre travail par une conclusion générale.

I.1- Introduction

La segmentation est l'une des étapes les plus importantes dans la chaîne d'analyse automatique des images. Son but est de diviser l'image originale en plusieurs régions distinctes.

La segmentation et sans doute la tâche qui, en analyse d'images, mobilise le plus d'efforts. Certes, cette étape importante du traitement n'apparaît pas toujours de façon explicite, mais on peut affirmer qu'elle est toujours présente, même lorsque les images à analysée sont simples.

Nous présenterons dans ce chapitre quelques notions de bases et les différentes approches de la segmentation.

I.2- Segmentation

La segmentation d'image est une technique qui permet de diviser l'image originale en plusieurs zones homogène, tel que l'union des régions adjacentes ne donne pas une région homogène. Elle peut être considérée comme un problème de classement ou de classification automatique des points de l'image (pixels) ; basée sur les propriétés de ces points, leur voisinage et de leur arrangement spatial.

Définition mathématique de la segmentation

Segmenter une image I en n régions, revient à la partitionner en n sous ensembles $R_1, R_2, \ldots\ldots\ldots R_n$ tels que :

1. $I = \bigcup_i R_i$.
2. R_i est constituée de pixels connexes pour tout i.
3. $P(R_i)$ = Vrai pour tout i.
4. $P(R_i \cup R_j) = faux$ pour tous i, j, R_i et R_j étant adjacentes dans I.

La première condition implique que chaque pixel de l'image doit appartenir à une région R_i et l'union de toutes les régions correspond à l'image entière. La deuxième condition est relative à la structure des régions. Elle définit une région comme un ensemble de pixels qui doivent être connexes. La troisième condition exprime que chaque région doit respecter un

prédicat d'uniformité. La dernière condition implique la non réalisation de ce même prédicat pour la réunion de deux régions adjacentes.

Le résultat de la segmentation est une image dans laquelle une étiquette est attribuée à chaque pixel. L'étiquette d'un pixel correspond au numéro de la région à laquelle il appartient.

I.3- Binarisation d'une image

La binarisation est la segmentation d'image à plusieurs niveaux de gris en une image à deux niveaux de gris (noir et blanc). Le niveau de gris de tous les pixels est comparé à un seul choisi, à partir duquel on affecte à chacun d'eux une valeur : noir ou blanc, selon la position de son niveau de gris par rapport au seuil.

Soit I(x,y) la fonction donnant le niveau de gris du point (x,y) et T la valeur du seuil.

L'opération de binarisation donne une image ayant deux classes C_1 et C_2 en appliquant la règle de décision suivante :

Si I(x,y)>T **alors** I(x,y) $\in C_1$

Sinon I(x,y) $\in C_2$

Après binarisation, la classe C_1 (ou C_2) est la classe ayant le niveau de gris noir (ou blanc). Donc tous les problèmes se situent au niveau du calcul du seuil T, qui est déterminé à partir des méthodes de seuillage des images.

I.4- Différentes approches des méthodes de segmentation

Dans le but d'avoir une meilleure segmentation d'une image, de nombreuses méthodes ont été proposées suivant des paramètres caractérisant cette image. La différence entre ces méthodes réside dans leur façon de classer les différents pixels de l'image.

Généralement, la segmentation d'une image est effectuée par l'utilisation de l'une des deux grandes approches basée sur l'extraction de contours (frontières) ou la croissance des régions

1. L'approche par région.
2. L'approche par contour.
3. Approche par classification des pixels.

I.4.1- Approche par région

Ce type de méthodes consiste à regrouper itérativement des ensembles de points connexes pour constituer des régions plus importantes, vérifiant les conditions dépendantes du critère d'homogénéité.

D'un point de vue topologique, la segmentation par croissance de régions correspond à une partition de l'image I en K régions connexes R_i, $i=1,\ldots,K$ toute on respectant les condition suivantes :

- ✓ $\bigcup_{i=1}^{K} R_i = R, \quad R = I$ (*image originale*)
- ✓ Les R_i sont des partitions connectées
- ✓ $R_i \cap R_j = \phi,$ pour tous $i \neq j$.
- ✓ $P(R_i)$ = vrai pour i allant de i à K.
- ✓ $P(R_i \cup R_j)$ = faux, pour $i \neq j$.

P étant un prédicat qui définit un ensemble de critères d'homogénéité.

Généralement, il existe quatre types de méthodes dans l'approche par région :

I.4.1.1- Méthode par division (partage)

Ces méthodes s'attachent à diviser l'image en région d'une manière récursive.

Initialement, une mesure d'inhomogénéité $E(R)$ est calculée sur l'unique région que forme l'image originale au moyen de l'équation :

$$E(R) = \frac{1}{N(R)} \sum_{i-1}^{N(R)} (x_i - v(R))^2$$

Où $N(R)$: représente le nombre de pixels de la région R.

x_i : représente le niveau de gris du pixel i.

v(R) :valeur moyenne des niveaux de gris des pixels de la région *R* donnée par :

$$v(R) = \frac{\sum_{i=1}^{N(R)} x_i}{N(R)}$$

Cette mesure d'inhomogénéité *E(R)* représente le degré de fluctuation du niveau de gris de chaque pixel de la région. Le but du prédicat d'uniformité *P(R)* est de décider si la région *R* est homogène ou non. Cette décision binaire s'obtient généralement en comparant *E(R)* à un seuil *T* comme l'indique la relation suivante :

$$P(R) = \begin{cases} vrai \;\; si \;\; E(R) < T \\ faux \;\; \text{sin}\, on \end{cases}$$

Dans le cas homogène, la région *R* est laissée telle quelle, alors que dans le cas inhomogène, la région *R* est divisée en sous régions plus homogènes ; cette opération peut s'effectuer, par exemple en divisant la région *R* en quatre quadrants d'égales surfaces R_i, $i = 1,...,4$ comme l'illustre la figure (I.1),

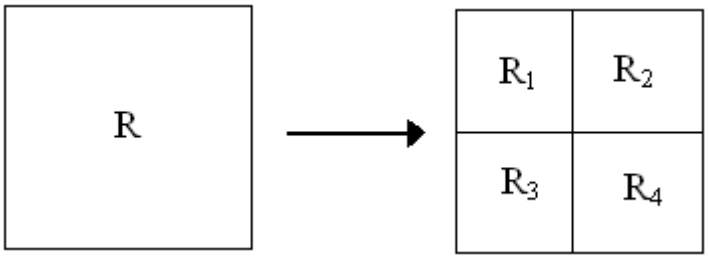


Figure (I.1)- Division d'une région inhomogène *R* en quatre quadrants R_K $k = 1,..., 4$

Le processus de calcul de la mesure d'inhomogénéité *E(R)* , de l'évaluation du prédicat d'uniformité *P(R)* et de la division de quatre quadrants, s'effectue récursivement pour chacune des régions trouvées quelle que soit sa taille, jusqu'à ce que toutes les régions soient homogènes.

I.4.1.2- Méthodes par fusion

Cette méthode réunit les différentes régions adjacentes en une nouvelle région, si la mesure d'inhomogénéité de cette dernière ne dépasse pas un certain seuil *T*. L'arrêt de cette opération est atteint lorsqu'un certain critère de fin de rassemblement est satisfait. Par exemple lorsque la somme des

erreurs quadratiques *SEQ* calculée entre l'image originale et l'image approchée, excède un certain seuil T . Ce critère est décrit formellement par :

$SEQ(K)=\sum_{i=1}^{K}\sum_{x_j\in R_i}\left|x_j-v(R_i)\right|^2$, avec K le nombre de régions ;

Cette méthode est considérée comme une opération de post traitement qui vient compléter un algorithme de segmentation. Elle peut être modélisée au moyen du graphe de contiguïtés des régions GCR (Region Adjacency Graph). Ce dernier se construit à partir de la segmentation obtenue par division de la manière suivante :

- A chaque région de l'image correspond à un nœud du graphe de contiguïté.
- Chaque couple de régions adjacentes dans l'image est représenté dans le graphe GCA par un lien qui connecte les deux nœuds correspondants.
- Chaque lien est associé à une valeur qui est une mesure de dissimilarité qui existe entre les deux régions qu'il relie.

Cette méthode est considérée comme une opération de post-traitement qui vient compléter un algorithme de segmentation.

I.4.1.3- Méthodes par division-fusion (Split and merge)

En se basant sur les deux concepts qu'on vient de citer à savoir la division et la fusion, Horwitz et Pavlidis ont développé une méthode mixte qui divise et fusionne simultanément les régions. [HP77]

Cette méthode utilise un arbre quaternaire (Quad-tree) où chaque nœud représente une région. Si ce nœud n'est pas terminal, il possède au maximum quatre fils. Cette méthode se déroule comme suit :

1. Division d'une région R_i en plusieurs sous régions disjointes, Si $P(R_i)=faux$.
2. Fusion de deux régions adjacentes R_i et R_j si $P(R_i \cup R_j)= vrai$.
3. Arrêt si le critère de fin de rassemblement est atteint.

L'exemple de la Figure (I.2) illustre ce principe et celui de la Figure (I.3) donne l'arbre quaternaire correspondante.

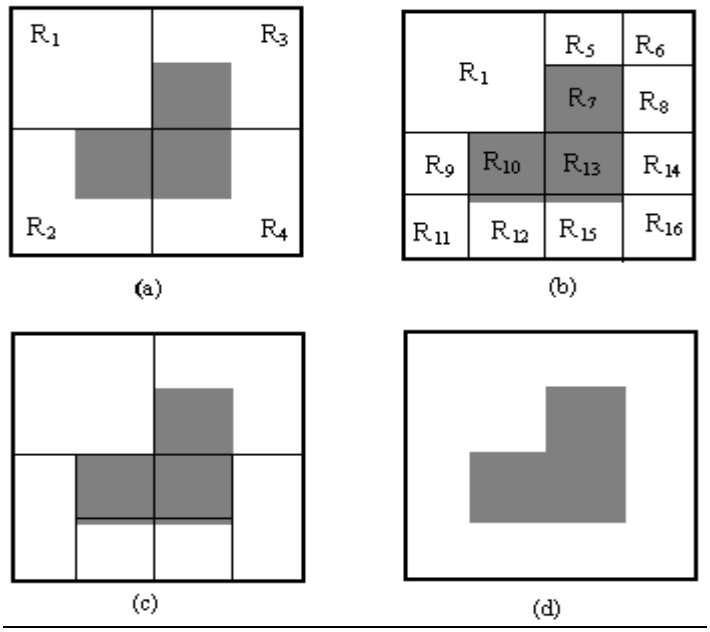


Figure (I.2)- Exemple d'application de l'algorithme de division –fusion
(a) Image divisée en quatre régions, (b) image divisée en 13 régions,
(c) fusion de quelque régions et division de deux autres, (d) Image segmentée.

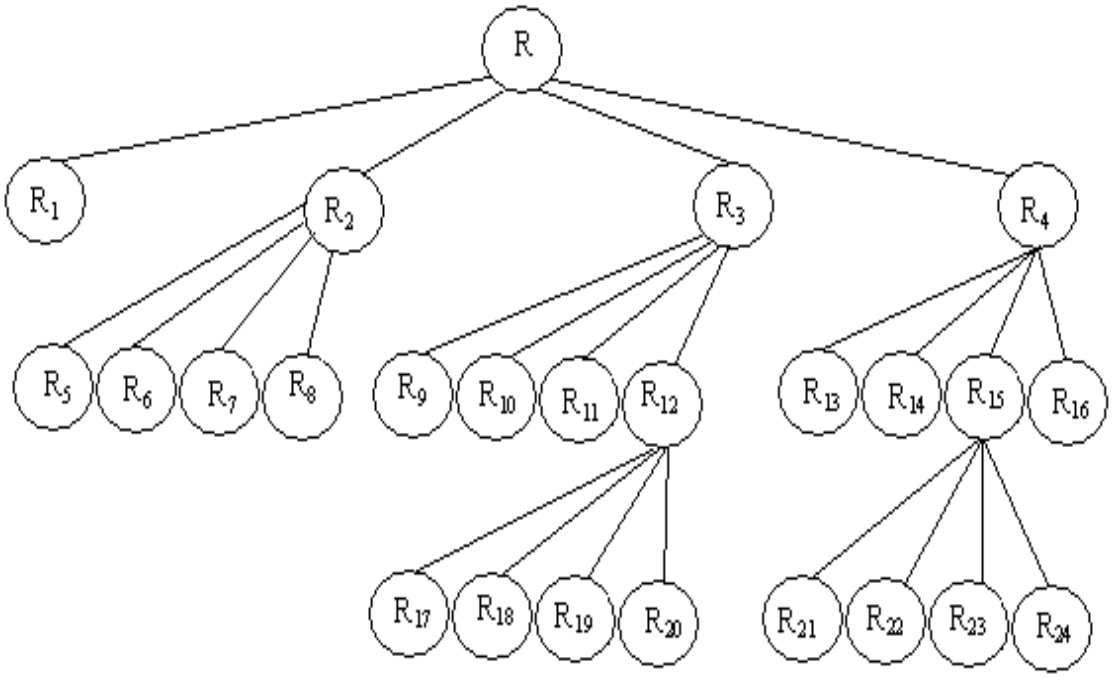


Figure (I.3)- Description du découpage de l'image au moyen d'un arbre quaternaire.

I.4.1.4 Méthodes par croissance de régions

Les méthodes de croissance de régions consistent à agréger aux régions les pixels adjacents en respectant un critère d'homogénéité.

Les régions sont initialisées à l'aide de germes qui correspondent généralement à de groupes de pixels. Puis les régions sont construites en y ajoutant successivement les pixels qui leur sont connexes et respectant un critère de similarité. La croissance s'arrête lorsque tous les pixels ont été traités.

Le critère de similarité consiste à comparer le niveau de gris du pixel examiné au niveau de gris de la région candidate. On peut aussi y tenir compte d'une contrainte sur la forme ou la taille de la région.

I.4.2- Approche par contour

L'approche contour consiste à rechercher les discontinuités locales, les transitions entre les régions. Remarquons qu'une discontinuité dans l'image n'est pas forcément liée à une variation géométrique ou physique de la surface observée : elle peut également être due à une différence d'éclairage, par exemple un effet d'ombre. L'approche contour n'aboutit pas directement à une segmentation, car les contours détectés ne sont pas toujours connexes. Il existe cependant des techniques permettant d'obtenir des contours fermés.

Dans ce cas, on observe une parfaite dualité entre les contours et les régions.

La détection de contours consiste à balayer l'image avec une fenêtre définissant la zone d'intérêt. A chaque position, un opérateur est appliqué sur les pixels de la fenêtre afin d'estimer s'il y a une transition significative au niveau de l'attribut choisi. A partir des pixels susceptibles d'appartenir à un contour, il faut ensuite extraire des contours fermés.

Un pixel contour est souvent défini comme un minimum local du module du gradient dans la direction du gradient, ou encore comme un passage par zéro de la dérivée seconde dans cette même direction. Comme les opérateurs de dérivation sont très sensibles au bruit, des images bruitées doivent être préalablement lissées. Un grand nombre d'opérateurs gradient ont été proposés. Ils se distinguent entre eux principalement par le choix du filtre de lissage.

Le lissage et la dérivation sont en pratique réunis dans un seul filtre. [Can86] [She96] [Der90].

I.4.3- Approche par classification des pixels

Dans cette approche, les pixels sont caractérisés par un ensemble d'attributs. Ces attributs correspondant au niveau de gris , deux attributs de texture calculés à partir des pixels voisins situé à l'intérieure d'une fenêtre de voisinage centrée sur chaque pixels ou encore, aux composantes couleur s'il s'agit d'une image couleur. Les pixels sont projetés dans l'espace des attributs et forment de nuages de points

La classification des pixels consiste alors à extraire des nuages de points compacts qui correspondent aux classes de pixels dans l'image. Les méthodes de classification construisent les classes de pixels et affectent une étiquette à chaque pixels. La formation des régions homogènes dans l'image n'est obtenue qu'après l'analyse de connexité des pixels dans l'image étiquetée.

Les méthodes exploitant plusieurs attributs sont qualifiées de multidimensionnelles. Quelques algorithmes appartenant à cette catégorie seront présentés dans le chapitre II. Les méthodes ne prenant en compte qu'un seul attribut (niveau de gris) sont qualifiées de monodimensionnelles. Elles consistent à extraire automatiquement des seuils puis affectent les pixels à une classe par comparaison de leur niveau de gris à ces seuils. On les appelle communément méthodes de seuillage. Elles sont présentées dans le paragraphe suivant

I.4.3.1- Segmentation d'image par seuillage d'histogramme

I.4.3.1.a-Définition du seuillage

Les méthodes par seuillage consistent à attribuer chaque point d'image à une certaine classe, par comparaison des valeurs des niveaux de gris de ces points à des seuils calculés à l'avance selon un certain critère. Cette définition est formalisée de la manière suivante :

Soient L={0,1,2,……..l-1} l'ensemble de niveau de gris d'une image et $l(x,y)$ la luminance (niveau de gris) d'un pixel de coordonnées (x,y). L-1 étant le niveau de gris maximal, souvent L= 256.

La segmentation par seuillage est une opération qui consiste à répartir les pixels en K classes (C_1 , C_2 ,…… C_k) à partir d'un ensemble de seuils T= {t_1, t_2, ….,t_{k-1}}.Par convenance on utilise deux autres seuils, t_0=0 et t_k=L-1

Un pixel de niveau de gris $l(x, y)$ est affecté à la classe C_k si :

$$t_k \langle l(x,y) \leq t_{k+1} \quad k= 0, 1,2,\ldots\ldots\ldots K\text{-}1$$

Le calcule des seuils t_k est généralement basé sur les propriétés de l'histogramme de l'image.

I.4.3.1.b -Définition de l'histogramme

L'histogramme représente la distribution des fréquences d'apparition des niveaux de gris dans une image. Il constitue une densité de probabilité de la variable liée à l'apparition du niveau de gris dans l'image. C'est un outil très privilégié en analyse de l'image. L'histogramme est une courbe monodimensionnelle décrite par une fonction discrète *h(i)* ou *p(i)* qui représentent respectivement la fréquence et la probabilité d'apparition du niveau de gris *i* , tel que :

$$p\ (i) = \frac{h(i)}{N}$$

h(i) est le nombre de pixels ayant le niveau de gris *i* et *N* le nombre total de pixels dans l'image,selon sa forme on peut distinguer:

- **Histogramme uni modal**

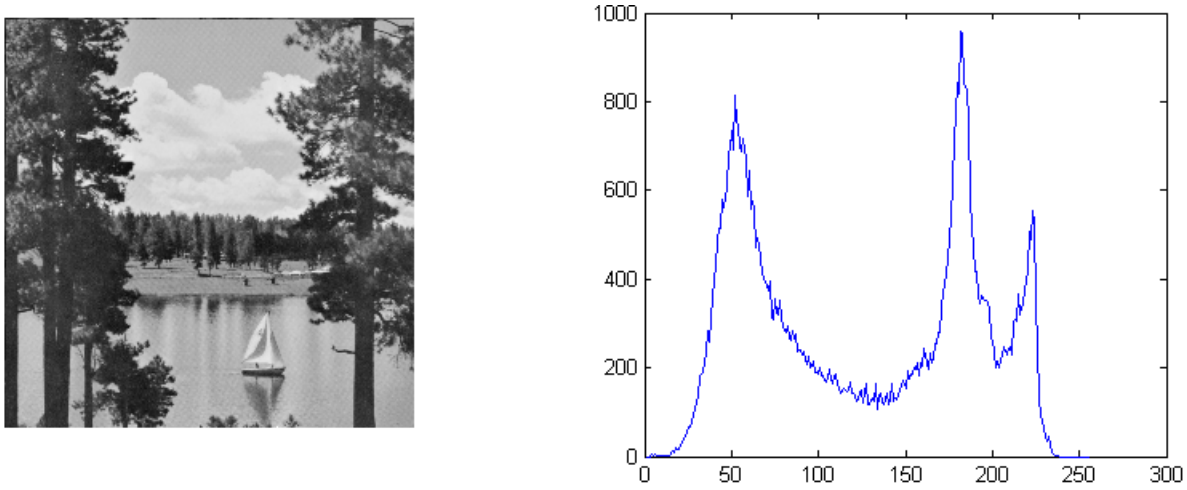
  Il est formé d'un seul pic représentant un objet ou bien un fond.

- **Histogramme bimodal**

  Il est caractérisé par deux modes séparés par une vallée, il indique l'existence d'un objet sur un fond.

- **Histogramme multimodal**

  Il comporte plus de deux modes séparés par des vallées, ce qui correspond à une image constituée de plus de deux régions.



(a) **(b)**

Figure (I.4) - image réelle (a) et son histogramme (b).

Remarque : Les seuils sont souvent localisés dans les vallées de l'histogramme.

I.4.3.2- Méthodes de calcul des seuils

Les méthodes de seuillage reposent sur l'exploitation de l'histogramme de toute l'image qui caractérise la distribution des niveaux de gris.

Un histogramme idéal est un histogramme représentant des modes qui traduisent, parfaitement les classes de l'image. Les seuils optimaux sont alors localisés dans les vallées situées entre les modes (Fig. I.5).

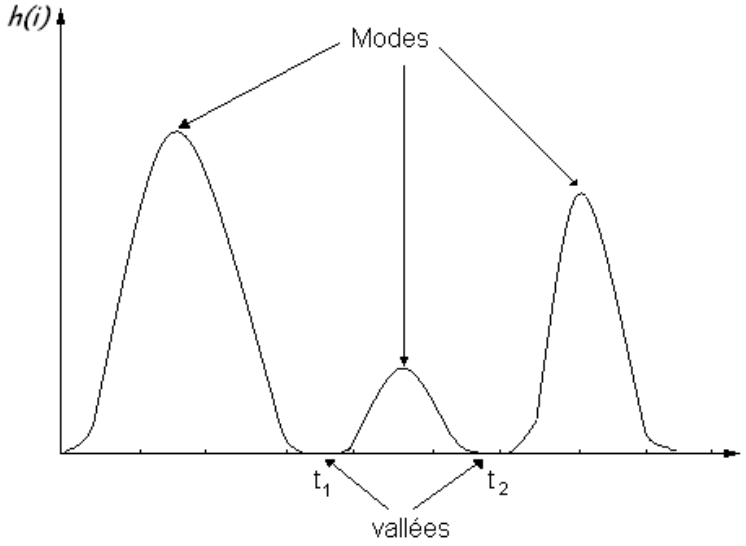


Figure (1.5)- Allure d'un histogramme idéal constitué de 3 modes

Or en pratique, il est rare de trouver un histogramme qui présente deux modes bien distincts. En effet, la plupart des images présentent des histogrammes bruités caractérisés par des modes non discernables.

Plusieurs méthodes des seuillages ont été développées. Ces méthodes ont été initialement proposées pour le calcul d'un seul seuil mais elles restent tout de même extensibles dans le cas du multi seuillage.

I.4.3.2.1- Méthode basée sur l'analyse discriminante [NS79]

Dans cette méthode, l'opération de seuillage est vue comme une séparation (un partitionnement) des pixels d'une image en deux classes C_0, C1 (objet et fond) à partir d'un seuil t. Ces deux classes sont désignées en fonction du seuil t.

C_0= {0,1,……t} et C_1= {t+1,…., L-1}

Soient σ_w^2 la variance d'une classe, σ_b^2 la variance interclasse et σ_T^2 la variance totale.

Le seuil optimum t^* peut être déterminé en maximisant un des trois critères suivant :

$$\lambda(t)=\frac{\sigma_b^2}{\sigma_w^2}\ ; \qquad \eta(t)=\frac{\sigma_b^2}{\sigma_T^2}\ ; \qquad \kappa(t)=\frac{\sigma_T^2}{\sigma_w^2}$$

avec $\sigma_T^2=\sum_{i=0}^{L-1}(i-\mu_t)^2 p_i$; $\mu_t=\sum_{i=0}^{L-1} i p_i$;

$$\sigma_b^2=p_t q(\mu_1-\mu_2)^2\ ; \qquad p_t=\sum_{i=0}^{t}\frac{h(i)}{N}\ ; \qquad q=1-p_t\ ;$$

$$\mu_1=\frac{\mu_t-\mu_s}{1-p_t}\ ; \qquad \mu_2=\frac{\mu_s}{p_t}\ ; \qquad \mu_s=\sum_{i=0}^{t} i p_i\ ;$$

Et $p_i=\frac{h(i)}{N}$: correspond à la probabilité d'apparition du niveau de gris *i*.

Les trois critères sont équivalents, mais le plus simple à utiliser est $\eta(t)$.

Le seuil optimum t^* est défini comme suit :

$$t^*=\text{Arg}\max_{t\in L}\eta(t)$$

Cette méthode peut s'étendre à un nombre de classes plus important. Le principe est de choisir un certain nombre de seuils qui vont servir de séparateurs entre les objets qui comportent l'image.

Dans ce cas, (K-1) seuils optimaux $T= \{t_1, t_2,.....,t_{k-1}\}$ peuvent être déterminés en minimisant la variance intra classes $\sigma_w^2(T)$ ou en maximisant la variance inter classes $\sigma_b^2(T)$ tel que :

$$\sigma_w^2(k) = \sum_{i=0}^{k-1} \sum_{j=t_i}^{t_{i+1}-1} (j-\mu_{i+1})^2 p_j \quad , \quad \sigma_b^2 = \sum_{i=1}^{k} p_i(\mu_i - \mu_{i+1})^2$$

$$\sigma_T^2 = \sum_{j=0}^{L-1} (j-\mu)^2 p_j \; , \quad \sigma_T^2 = \sigma_w^2(k) - \sigma_b^2(k) \; , \quad \mu = \frac{1}{N}\sum_{j=0}^{L-1} jh_j \; ,$$

$$\mu_i = \frac{1}{N}\sum_{j=t_i}^{t_{i+1}-1} jh_j \quad , \quad p_i = \frac{1}{N}\sum_{j=t_i}^{t_{i+1}-1} h_j$$

I.4.3.2.2- Méthode de seuillage global basées sur l'entropie

Les méthodes de seuillage utilisant la notion d'entropie sont basées sur la théorie de l'information. L'entropie d'une image peut être considérée comme une évaluation de l'information qu'elle contient. Elle permet également de déterminer le degré de désordre dans l'image. L'entropie est en général minimale pour une image homogène.

I.4.3.2.2.a- Méthode de Kapur, Sahoo et Wang [KSW85]

Dans cette méthode, deux distributions de probabilités, l'une relatif aux objets, l'autre au fond, découlent de la distribution des niveaux de gris de l'image originale, comme suit :

$$\frac{P_0}{P_t}, \frac{P_1}{P_t}, \ldots\ldots\ldots, \frac{P_t}{P_t}, \frac{P_{t+1}}{1-P_t}, \ldots\ldots\ldots, \frac{P_{l-1}}{1-P_t}$$

Où t est la valeur du seuil et $P_t = \sum_{i=0}^{t} p_i$

Soient l'entropie des deux distributions :

$$H_b = -\sum_{i=0}^{t} \frac{p_i}{p_t} \log\left(\frac{p_i}{p_t}\right) \; ; \qquad H_n = -\sum_{i=t+1}^{L-1} \frac{p_i}{1-p_t} \log\left(\frac{p_i}{1-p_t}\right)$$

Le seuil optimal t^* est définit comme étant le niveau de gris qui maximise la quantité $H_b + H_n$; C'est-à-dire que :

$$t^* = Arg \max_{t \in L} \{H_b(t) + H_n(t)\}$$

Cette méthode peut être également étendue au calcul de plusieurs seuils. Il s'agira de maximiser le critère suivant :

$$J(T) = \sum_{i=1}^{k} H_i$$

avec $H_i = \sum_{j=t_{i-1}}^{t_i - 1} \frac{p_j}{p_i} \log(\frac{p_j}{p_i})$; $p_i = \frac{1}{N} \sum_{j=t_{i-1}}^{t_i - 1} h_j$

I.4.3.2.2.b- Méthode de cross entropy

Cette méthode a été introduite par Li et Al [CL83], elle consiste à minimiser l'entropie entre l'image et sa version binaire. Sans faire d'hypothèses sur la répartition a priori des distributions des populations, elle fournie une estimation non biaisée d'une image binarisée.

Le seuil optimal t^* est tel que :

$$t^* = Arg \max_{t} J(t).$$

avec $$J(t) = \sum_{i=0}^{t} ih(i) \log(\frac{i}{\mu_0}) + \sum_{i=t+1}^{l-1} ih(i) \log(\frac{i}{\mu_1})$$

Après quelque modification, l'expression de *J(t)* devient :

$$J(t) = N^2 (w_0 m_0 \log(m_0) + w_1 m_1 \log(m_1))$$

avec $w_0 = \sum_{i=0}^{t} p_i$; $w_1 = \sum_{i=t+1}^{L-1} p_i$; $m_0 = \sum_{i=0}^{t} i \frac{p_i}{w_0}$; $m_1 = \sum_{i=t+1}^{L-1} i \frac{p_i}{w_1}$

Cette méthode peut être également étendue au calcul de plusieurs seuils. Il s'agira de maximiser le critère suivant :

$$J(T) = N^2 \sum_{i=1}^{K} w_i m_i \log(m_i)$$

avec $w_i = \sum_{j=t_{i-1}}^{t_i - 1} p_j$; $m_i = \sum_{j=t_{i-1}}^{t_i - 1} j \frac{p_j}{w_i}$

I.4.3.2.2.c- Méthode de corrélation entropique

Proposée par Chang et Al [CYC95], cette méthode maximise la corrélation entropique entre les classes objet et fond. Soit X une variable aléatoire discrète, avec un rang B= x_0, x_1,….fini ou infini et p_i la probabilité de X=x_i. La corrélation de X est définie comme suit :

$$C_x(t) = -\sum_{i\geq 0} p_i^2$$

En se basant sur cette définition, les corrélations associées aux classes C_0 (fond) et C_1 (objet) sont données par les relations suivantes :

$$C_{C_0} = -\ln \sum_{i=0}^{t} \left(\frac{p_i}{p(C_0)} \right)^2 \qquad \text{et} \qquad C_{C_1} = -\ln \sum_{i=t+1}^{L-1} \left(\frac{p_i}{1-p(C_0)} \right)^2$$

La corrélation total (TC) associée aux deux classes est alors :

$$TC(t) = C_{C_0}(t) + C_{C_1}(t) = -\ln \sum_{i=0}^{t} \left(\frac{p_i}{p(C_0)} \right)^2 - \ln \sum_{i=t+1}^{L-1} \left(\frac{p_i}{1-p(C_0)} \right)^2 = -\ln \left(\frac{G_{C_0}(t) G_{C_1}(t)}{p(C_0)^2 (1-p(C_0))^2} \right)$$

$$TC(t) = -\ln \left(G_{C_0}(t) G_{C_1}(t) \right) + 2 \ln \left(p(C_0)(1-p(C_0)) \right)$$

avec $G_{C_0}(t) = \sum_{i=0}^{t} p_i^2$; $G_{C_1} = \sum_{i=t+1}^{L-1} p_i^2$;

$p(C_0) = \sum_{i=0}^{t} p_i$; $p(C_1) = \sum_{i=t+1}^{L-1} p_i$;

Le seuil optimum t^* est déterminé en maximisant la fonction critère suivante :

$$J(t) = -\ln \left(G_{C_0}(t) G_{C_1}(t) \right) + 2 \ln \left(p(C_0)(1-p(C_0)) \right).$$

Une extension de cette méthode au cas du multi seuillage consiste à déterminer l'ensemble des seuils $T = \{t_1, t_2, \ldots . t_{k-1}\}$ qui maximise la fonction critère suivante :

$$J(T) = -\ln \prod_{i=1}^{k} G_{C_i} + 2 \ln \prod_{i=1}^{K} p(c_i)$$

avec $G_{C_i}(t) = \sum_{j=t_{i-1}}^{t_i - 1} p_j^2$; $p(C_i) = \sum_{j=t_{i-1}}^{t_i - 1} p_j$

I.4.3.3-Seuillage par minimisation de l'erreur quadratique

Proposée par Kittler et Illingworth [KI86], cette méthode considère l'histogramme des niveaux de gris comme une estimée de la fonction de

densité de probabilité *p(i)* d'un mélange de population formé des niveaux de gris des objets et du fond.

On peut supposer que chacune des deux composantes $p(i/C_k)$, $(k=1.2)$ du mélange est une distribution normale avec moyenne μ_k , un écart type σ_k et une probabilité a priori P_k. On a ainsi :

$$p(i/C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(i-\mu_k)^2}{2\sigma_k^2} \right) ; \qquad p(i) = \sum_{k=1}^{2} P_k \, p(i/C_k)$$

Le seuil optimal est obtenu en résolvant l'équation quadratique ci- dessous :

$$\frac{(i-\mu_1)^2}{\sigma_1^2} + \log\sigma_1^2 - 2\log P_1 = \frac{(i-\mu_2)^2}{\sigma_2^2} + \log\sigma_2^2 - 2\log P_2$$

Ou, en minimisant la fonction $J_1(t)$ suivante :

$$J_1(t) = 1 + 2\left[P_1(t)\log\sigma_1(t) + P_2\log\sigma_2(t)\right] - 2\left[P_1(t)\log P_1(t) + P_2(t)\log P_2(t)\right]$$

avec :

$$\sigma_1^2 = \frac{\sum_{n=o}^{t} (i-\mu_1(t))^2 h(i)}{p_1(t)} ; \qquad \sigma_2^2 = \frac{\sum_{i=t+1}^{L-1} (i-\mu_2(t))^2 h(i)}{p_2(t)} ;$$

$$p_1(t) = \sum_{i=0}^{t} h(i) ; \qquad p_2(t) = \sum_{i=t+1}^{L-1} h(i) ;$$

$$\mu_1(t) = \frac{\sum_{i=0}^{t} ih(i)}{p_0(t)} ; \qquad \mu_2(t) = \frac{\sum_{i=t+1}^{L-1} ih(i)}{p_1(t)} ;$$

Soit la fonction de critère $J_2(t)$:

$$J_2(t) = N^2\left(w_1 \log\frac{\sigma_1}{w_1 N^2} + w_2 \log\frac{\sigma_2}{w_2 N^2} \right) ;$$

Avec

$$\sigma_1^2 = \frac{1}{w_1 N^2} \sum_{i=0}^{t} (i-\mu_1)^2 h(i)$$

$$\sigma_2^2 = \frac{1}{w_2 N^2} \sum_{i=t}^{L-1} (i-\mu_2)^2 h(i)$$

Notons que N est le nombre de pixels dans l'image.

Les paramètres w_1, w_2, μ_1, et μ_2 sont calculés comme suit :

$$w_1 = \sum_{i=0}^{t} P_i \ ; \qquad \mu_1 = \sum_{i=0}^{t} i \frac{P_i}{w_1} \ ;$$

$$w_2 = \sum_{i=t}^{L-1} P_i \ ; \qquad \mu_2 = \sum_{i=t}^{L-1} i \frac{P_i}{w_2} \ ;$$

Dans le cas d'un mélange de *K* distributions gaussiennes, l'ensemble des seuils $T = \{t_1, t_2, ..., t_{k-1}\}$ peuvent être déterminés en maximisant le critère suivant :

$$J(T) = N^2 \sum_{j=1}^{K} w_j \log \frac{\sigma_j}{w_j N^2}$$

$$\sigma_j^2 = \frac{1}{w_j N^2} \sum_{i=t_j}^{t_j - 1} (i - \mu_j)^2 h(i)$$

$$w_j = \sum_{i=t_j}^{t_j - 1} p_i$$

I.4.3.4- Multi seuillage des images

Plusieurs méthodes de seuillage ont été développées pour déterminer un seul seuil. Ces méthodes peuvent être étendues au cas de multi seuillage comme cela a été décrit dans les paragraphes précédents.

Cependant, en pratique cette extension peut engendrée des temps de calculs prohibitifs. En effet la recherche rapide d'un seul seuil d'une manière exhaustive est tout à fait possible. Cependant cette recherche exhaustive devient prohibitive lorsque le nombre de seuils augmente.

Chang et Al notent que la complexité des calculs augmente exponentiellement lorsque le nombre de seuils augmente [JR95]. Pour *K* classes le nombre d'opérations *NO* nécessaires pour le calcul de *K-1* seuils est donné par :

$$NO = \frac{(L+K)!}{L!K!}$$

Où *L* représente le nombre de niveau de gris.

Ainsi, pour une image ayant *L=256* niveaux de gris :

NO=33153 Si K=2.

NO=2862209 Si K=3.
NO= 9711475137 Si K=5.

Pour résoudre ce problème plusieurs techniques ont été proposées. Parmi elles on peut citer l'algorithme itératif proposé par Yin et Chen [YC97] et les algorithmes génétiques.

I.5- Algorithme itératif

Supposant qu'on veut chercher (K-1) seuils en optimisant certaines fonctions objectives telles que celles proposées par Kapur et Otsu. L'algorithme itératif commence par choisir (K-1) seuils initiaux qui serons par la suite ajustés d'une manière itérative.

L'algorithme est détaillé comme suit :

Etape 1 : Choisir (K-1) seuils $[t_1,t_2,\ldots,t_{k-1}]$ .Par convenance, deux seuils supplémentaires : $t_0=0$ et $t_k=L-1$ sont définis.

Etape 2 : Calculer la valeur de la fonction objective *f* en utilisant les seuils $[t_1,t_2,\ldots,t_{k-1}]$.

Etape 3 : i=1.

Etape 4 : Soit l'intervalle des niveau de gris $[t_{i-1},t_{i+1}]$. Déterminer le seuil optimal t^* compris entre t_{i-1} et t_{i+1} en utilisant la fonction G et remplacer t_i par t^*.

Etape 5 : i=t+1, Aller a l'étape 4 si i<k-1.

Etape 6 : Calculer la valeur de la fonction objective *f* en utilisant les nouveaux seuils $[t_1,t_2,\ldots,t_{k-1}]$.

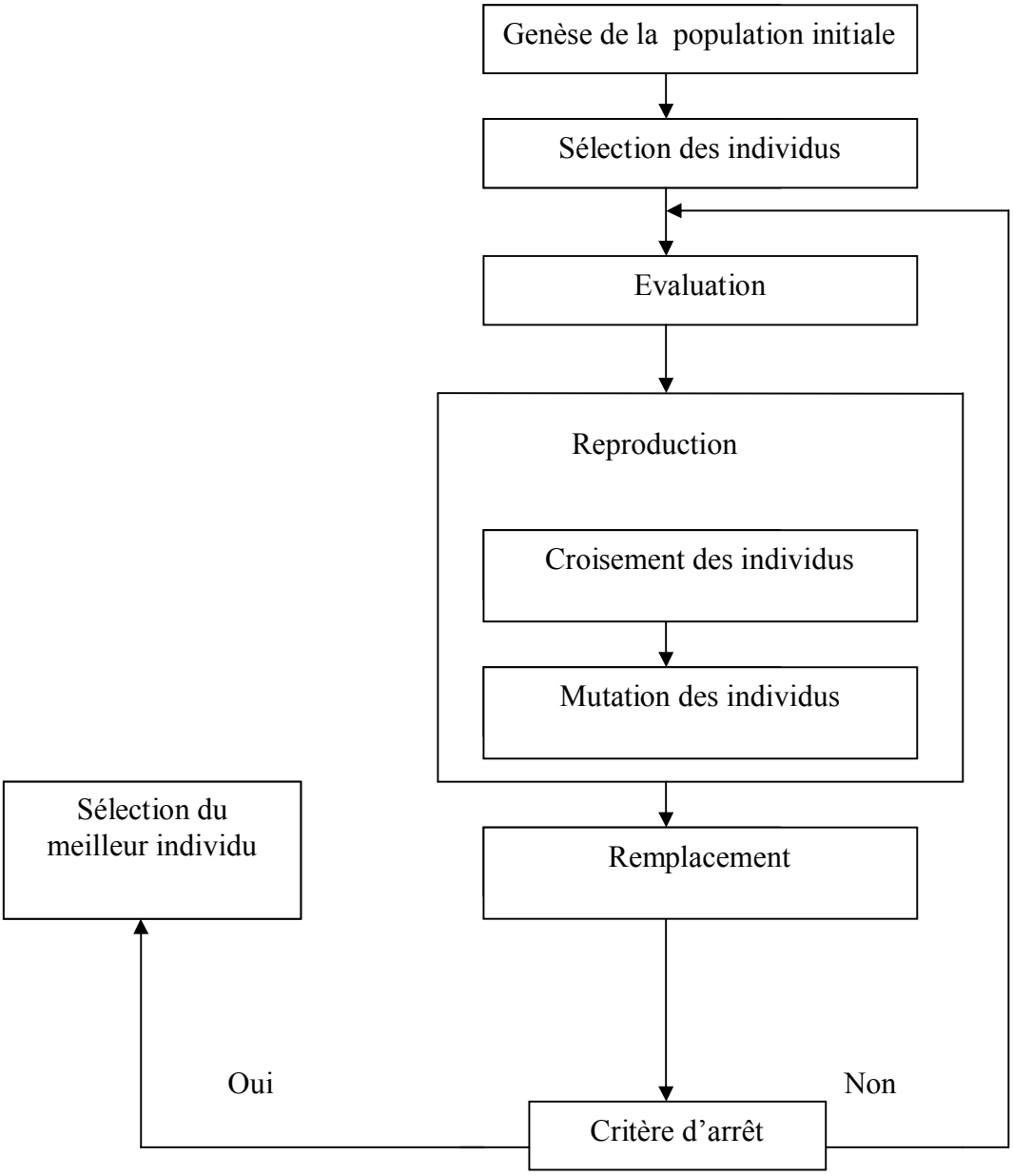
Etape 7 : Si *f* accroît alors aller à l'étape 3, sinon, arrêter.

Notons que G correspond à une fonction objective utilisée pour le calcul d'un seul seuil t* et que f correspond à la fonction objective utilisée pour mesure la performance des (K-1) seuils.

Les étapes 3 à 7 sont itérées jusqu'à ce que la valeur de la fonction objective cesse d'augmenter après deux itérations consécutives. Il a été prouvé que l'algorithme s'arrête après un nombre finis d'itérations.

I.6- Algorithme génétique

La recherche des seuils optimaux peut être également effectuée par un AG. Les étapes de cette méthode proposée sont décrites par l'organigramme suivant :



Figure(I.6)- Organigramme de la méthode

I.6.1- Codage chromosomes

Dans un algorithme génétique, un individu représente une solution du problème à traiter. Dans notre cas une solution représente les valeurs des seuils. Chaque solution est codée en une chaîne T tel que T= $\{t_1,t_2,\ldots,t_{k-1}\}$ avec $t_i \neq t_j \; pour \; i \neq j$. t_i indique la valeur du $i^{éme}$ seuil.

Les valeurs des seuils sont généralement comprit entre 0 et L-1 ; où L représente le nombre des niveaux de gris ; souvent L=256.

I.6.2- Génération de la population initiale

La population initiale est crée en générant aléatoirement *p individu* $A_1, A_2, ...A_p$ .Chaque individu A_i contient (k-1) seuils. La taille de la population est choisie par l'utilisateur, son choix dépend principalement du temps de convergence de l'AG.

I.6.3- Evaluation

L'objectif de cet algorithme est de trouver les valeurs des seuils (niveau de gris) qui optimisent une fonction objective donnée. On peut évidemment utiliser l'une des fonctions objectives présentées précédemment.

I.6.4- Sélection

La sélection est un mécanisme qui consiste à former une nouvelle génération par une sélection aléatoire des individus d'une population existante en fonction de leurs performances (fitness). Dans notre cas, on a utilisé la sélection par tournoi. Pour rappel, la sélection par tournoi consiste à choisir deux individus choisis aléatoirement dans la population. L'individu ayant la fitness la plus élevée sera sélectionné pour accéder à la génération intermédiaire.

Cette opération est répété jusqu'à remplir la population intermédiaire sélectionnée.

I.6.5- Reproduction

La reproduction est une étape très importante de l'AG. Dans cette phase, on applique deux opérateurs appelés respectivement croisement et mutation. Ces derniers ont pour but d'enrichir la diversité de la population en manipulant les composantes des individus (chromosomes). C'est le rôle du croisement.

L'opérateur de mutation apporte aux algorithmes génétiques l'aléa nécessaire à une exploration efficace de l'espace de recherche des solutions.

I.6.6- Croisement

Après avoir sélectionner les individus les mieux adaptés, ils vont subir maintenant l'opération du croisement qui consiste a échangé des Matériels Génétique entre deux individus (parents) pour produire deux nouveaux individus (enfants). Le processus du croisement est effectué de la manière suivante :

D'abord, on choisis *P/2* couples d'individus tirés d'une manière aléatoire à partir de la population sélectionnée pour former la population intermédiaire. Le Croisement est ensuite applique à chaque couple choisis aléatoirement avec une probabilité p_c.

Soient α et β un des couples, on génère un nombre aléatoire s entre [1, m], ou m= (k-1) est la longueur des individus α et β

Deux individus enfants $\alpha' et \beta'$ sont obtenus en échangeant tous les caractères de α et β après la position s. Notons que $\alpha' et \beta'$ gardent les caractéristiques communes de leurs parents.

Exemple :

Soient α et β deux individus constitués de 2 seuils t_1 et t_2 chacun codés sur 8 bits (m=2)

t_1=68 et t_2=190 pour α

t_1=95 et t_2=237 pour β

α = 68 190

β = 95 237

Soit s=1, un nombre générer aléatoirement entre 1 et m=2. En échangeant les gènes de α et β après la position s=1, on obtient deux nouveaux individus $\alpha' et \beta'$ tel que

α' =68 237

β' =95 190

Ces deux individus codent deux seuils t_1 et t_2 ayant les valeurs suivantes :

t_1 =68 et t_2=237 pour α'

t_1 =95 et t_2=190 pour β'

I.6.7- Mutation

La mutation est la modification aléatoire occasionnelle des gènes des individus de la population obtenus après la phase de croisement. L'opérateur de mutation consiste a modifier la valeur d'un gène avec une probabilité p_m de faible valeur. L'opérateur de la mutation donne naissance à une population dite enfants de même taille N_P que celle de la population parent.

Exemple :

Soit l'individu A constitué de deux seuils t_1 =20 et t_2=124 (m=2)

A = 20 124

En mutant le 2éme gène de A, on obtient l'individu A suivant

A = 20 120 qui correspond aux deux seuils t_1 =20 et t_2=124.

I.6.8- Remplacement

A chaque itération, on doit procéder au remplacement de la population des parents par la population des enfants. Dans notre travail, nous avons choisi un remplacement déterministe qui consiste à remplacer la population parente par la population des enfants.

I.7- Conclusion

Dans ce chapitre nous avons présenté quelques principes des différentes approches de la segmentation d'images. Parmi elle nous nous sommes particulièrement intéressé aux méthodes de segmentation par seuillage d'histogramme.

Le problème du calcul des seuils est posé comme un problème d'optimisation d'une fonction objective utilisant des informations issues de l'histogramme de l'image. Quelques méthodes de seuillage ont été ainsi décrites. Ces méthodes ont été développées initialement pour le calcul d'un seul seuil puis généralisées dans le cas du multi seuillage en utilisant par exemple soit un algorithme itératif ou un algorithme génétique.

Cependant, ces méthodes sont confrontées au choix initial des nombres de seuils. Pour résoudre ce problème, nous allons exploiter les indices de validité utilisés en classification pour déterminer le nombre de classes.

Le prochain chapitre sera ainsi consacré à la notion de classification de données et aux indices de validités.

II.1- Introduction

La classification est une activité mentale qui intervient fréquemment dans la vie courante. En effet, les objets, quelque soit leur nature, sont souvent répertoriés par rapport à des catégories ou des classes auxquelles ils sont censés appartenir.

La classification a pour but de classer en 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 à priori sur la structure de l'ensemble des objet à classer ni sur le nombre exacte de classes. Un certain nombre de critères ont été alors proposés afin d'évaluer et de valider la performance d'une classification. Ces critères appelés indices de validité permettent aussi de déterminer le nombre exacte de classes en présence.

Plusieurs indices ont été proposés dans la littérature. Ce foisonnement est du au fait qu'aucun indice universel n'est à présent connu. Dans ce chapitre, on se propose d'étudier certains d'entre eux.

II.2- Définition de la classification

En classification, les objets ou observations sont définis par un vecteur de mesures. Ce vecteur de dimension P correspond à un ensemble de P attributs ou variables (caractéristiques) définies a priori. Ces observations sont généralement représentées comme des points projetés dans un espace à P dimensions (Fig. II.1)

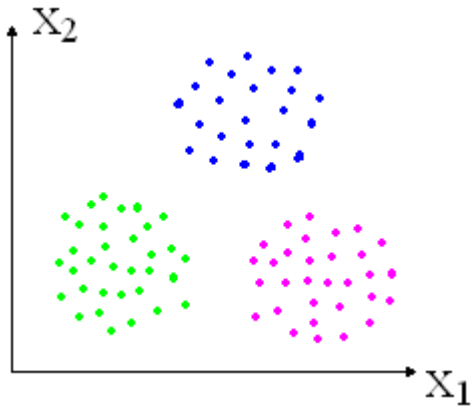


Figure (II.1)- Répartition des données dans un espace bidimensionnel.

Les observations ayant des attributs semblables appartiennent à la même classe et forment un nuage de point très dense dans l'espace P dimensions. Ainsi, la classification en générale est définie comme étant l'action de construire une collection d'objets similaires au

sein d'un même groupe, dissimilaires quand ils appartiennent à des groupes différents. Elle est communément connue sous le nom anglophone << Clustering >>.

Mathématiquement la classification est définie de la manière suivante :

Soit $X = \{X_1, X_2, \ldots\ldots\ldots X_N\}$ l'ensemble des *N* observations à classer.

Chaque observation X_i est caractérisée par *P* paramètres. $X_i = \{x_{i1}, x_{i2}, \ldots\ldots\ldots, x_{iP}\}$ tel que $X_i \in R^p$.

Soit $C = \{C_1, C_2, \ldots\ldots\ldots C_K\}$ l'ensembles des K classes.

La classification consiste à répartir l'ensemble des N observations en K classes tel que :

1. $C_i \neq \phi$ Pour *i=1,2,.....K*
2. $C_i \bigcap C_j = \phi \quad \forall\ i \neq j$
3. $\bigcup_{i=1}^{K} C_i = Q$

La première condition indique qu'une classe ne doit pas être vide, tandis que la deuxième condition stipule qu'aucun recouvrement entre les classes n'est toléré, alors que la troisième vérifie que l'union de toutes les classes doit aboutir au nombre d'observations disponibles.

Généralement, on exige de la classification de vérifier deux autres propriétés :

- Compacité : les points représentant une classe donnée sont plus proches entre eux que des points de toutes les autres classes.
- Séparabilité : les classes doivent être séparées le plus possible.

La figure (II.2) illustre un exemple graphique simple de classification :

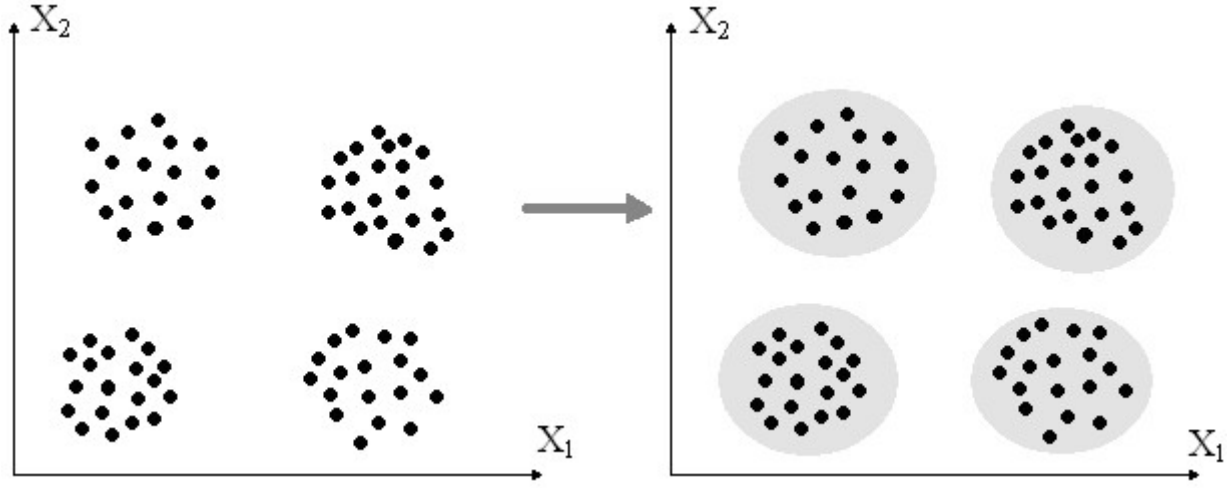


Figure (II.2)- Principe de la classification

Dans cet exemple, nous pouvons identifier facilement 4 classes où les données peuvent être regroupées selon un critère de similitude basé sur le calcule de *distance*. Deux objets peuvent appartenir à une même classe s'ils sont ''proches'' selon une distance donnée

Cette définition n'est cependant valable que dans le cas d'une *classification exclusive ou dure ''hard en anglais''* où un objet doit appartenir à une seule classe contrairement à la *classification non exclusive*, également appelée *douce ou de recouvrement ''soft en anglais'',*où un objet peut être assigné à plusieurs classes. La classification floue est un exemple de classification non exclusive où un objet peut appartenir à plusieurs classes avec des degrés d'appartenances différents. Par exemple, le regroupement des personnes en classes en fonction de leur âge ou sexe est exclusif, alors que leur regroupement par catégorie de maladies ne l'est pas puisqu'une personne peut avoir plusieurs maladies simultanément. Une autre catégorie de méthodes de classification est la *classification hiérarchique* qui tolère l'imbrication entre les classes.

II.3- Définition d'une classe

Une classe est une collection d'objets qui sont « *semblables* » entre eux et sont «différents » aux objets appartenant à d'autres classes**.**

D'après EVERRIT :

-Une classe est un ensemble d'entités qui sont semblables, alors que les entités provenant de classes différentes ne sont pas semblables.

-Une classe est un agrégat de points dans l'espace de représentation des données tels que la distance entre le centre et le point de cette classe est moins importante que celle entre le centre de cette classe et n'importe quel point d'une autre classe.

-Les classes peuvent être décrites comme des régions connexes de l'espace contenant relativement une grande densité de points.

Les classes peuvent ainsi avoir une existence '' naturelle'', il existe des groupes bien distincts que la plus part des méthodes de classification automatique devraient mettre en évidence., ou au contraire n'être que le résultat d'un découpage d'un nuage d'individus constituant un continuum. Dans ce cas, la classification présente une part d'arbitraire, et deux méthodes différentes donneront le plus souvent des résultats différents.

Nous proposerons de décrire deux méthodes de classification non supervisée basées sur l'approche métrique et qui sont dénommées C- Means et Fuzzy C-Means. Ces deux algorithmes sont très populaires et sont les plus utilisés en pratique. Ils présentent l'avantage d'être à la fois simples et efficaces.

II.4- Algorithme de C- Means

Cet algorithme a été proposé par Macqueen en 1967 **[MAC67].** Son principe est de partitionner l'ensemble des points en un ensemble de classes prédéterminé. L'algorithme des C-means vise à minimiser la variance intra classes, qui se traduit par la minimisation de l'énergie suivante :

$$J(\pi)=\sum_{k=1}^{K}\sum_{i=1}^{N_K} d^2(x_i,v_k) \;\; avec \; x_i \in C_k$$

Où π est une partition composée de k classes.

$d(X_i,v_k)$ Représente la distance entre un individu X_i et une classe représentée par son centre v_k. Les coordonnées du centre de classe C_k sont données par l'expression suivante :

$$v_k=\frac{1}{N_K}\sum_{i=1}^{N_k} x_i \qquad x_i \in C_k$$ Où N_k est le nombre d'observations de la classe C_k

La partition π^* qui optimise le critère est définie par : $J(\pi^*) = \min J(\pi)$.

Plusieurs types de distances peuvent être utilisés dans R^N (euclidienne, Mahalanobis, Minkowsky, etc.…). Cependant en pratique, on utilise assez souvent la distance euclidienne définie comme suit :

$$d(x_i,v_k)=\|x_i-v_k\|=\sqrt{\sum_{j=1}^{D}(X_{ij}-v_{kj})^2}\ .$$

Pratiquement, l'algorithme C means se déroule de la manière suivante :
Des points initiaux sont choisis aléatoirement pour constituer les centres de classes. Les autres points sont alors assignés à la classe dont le centre est le plus proche. Ensuite, on calcule le centre des classes et on répète le même processus jusqu'à avoir une partition optimale. Géométriquement cela revient à partager l'espace des points en K zones définie par les plans médiateurs des segments de droite reliant deux centres de deux classes différentes (fig.II.1)

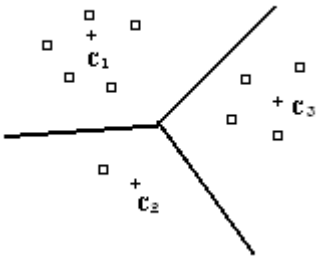


Figure (II.1)- Exemple de partition.

D'une manière plus générale, cet algorithme se déroule selon les étapes suivantes :
Soit un ensemble *X* de *N* individus caractérisés par *P* variables, à partitionner en *K* classes $\{C_1, ...C_k, ...C_K\}$.

Etape 0 : t=0 : Initialisation

On choisit aléatoirement *K* centres de classes $\{v_1^t,...,v_k^t,.....,v_K^t\}$. Ces *K* centres induisent une première partition π^t de l'ensemble des N individus en K classes $\{C_1^t,....C_k^t,....,C_K^t\}$.

Etape 1 : Chaque individu est affecté à la classe la plus proche. Par exemple un individu x_i est affecté à la classe C_k^t si il est plus proche de v_k^t que de tous les autres centres.

$$x_i \in C_k^t \text{ si } d(x_i, v_k^t) < d(x_i, v_{k'}^t)$$

avec $$k' = 1,...,k \quad \text{et} \quad k' \neq k$$

Etape 2 : On détermine les K nouveaux centres de classes $\{v_1^{t+1},...,v_k^{t+1},.....,v_K^{t+1}\}$.
Ces nouveaux centres induisent une nouvelles partition π^{t+1} de K classes $\{C_1^{t+1},....C_k^{t+1},....,C_K^{t+1}\}$.

Etape 3 : t=t+1 et aller vers l'étape 1 si le critère d'arrêt n'est pas respecté.

t : indique le numéro d'itération . Plusieurs critères peuvent être utilisés :

- Deux itérations successives conduisent à la même partition.
- Un nombre maximal d'itérations, fixé a priori, est atteint.
- Les centres de classe ne change pas entre deux itérations successives.
- Minimisation d'une fonction coût.

La figure (II.3) illustre le processus de partitionnement d'un ensemble d'individus en *K=3* classes. Les centre de classes sont indiquées par le symbole (+).

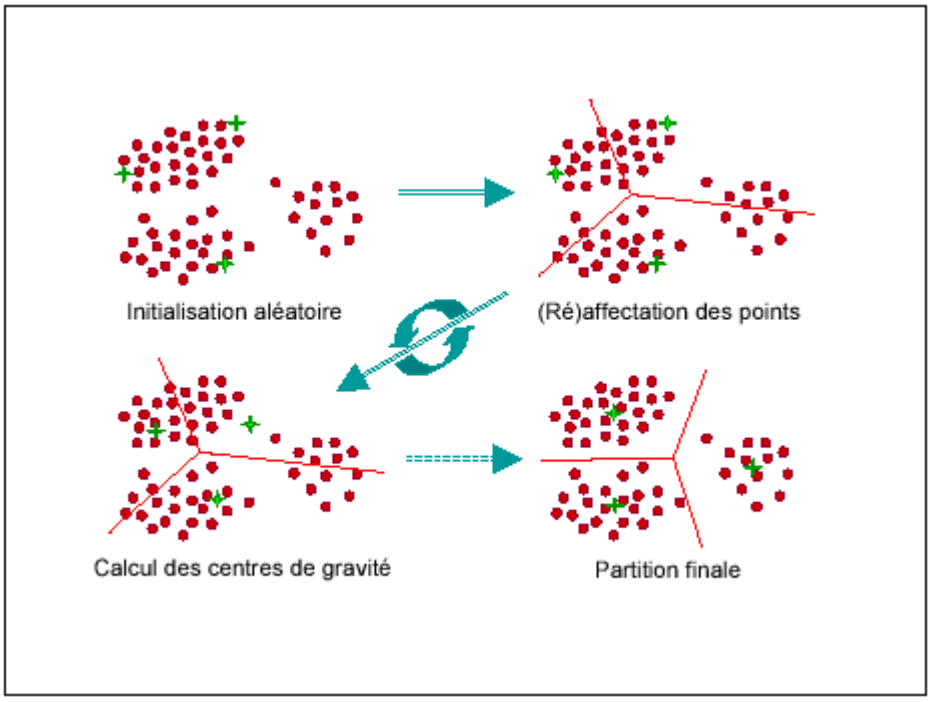


Figure (II.3)- Principe de l'algorithme de C-means.

II.5- Classification floue

II.5.1-Introducion à la notion floue

Un ensemble flou est un ensemble dont les bords sont mal définis. Cela se traduit par une fonction d'appartenance à valeur dans l'intervalle [0, 1] tout entier, par opposition au cas classique où la fonction d'appartenance prend deux valeur seulement : 0 ou 1. Cela permet de définir des ensembles d'une manière plus souple, tolérant à des informations imprécises, incomplètes et/ou incertaines. La figure II.4 montre un exemple d'ensemble flou.

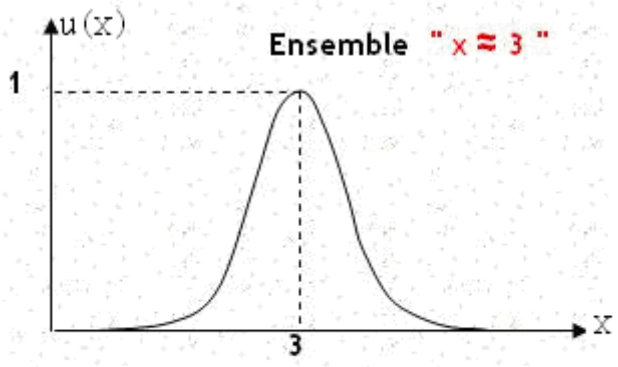


Figure II.4 Exemple d'ensemble flou

Cet ensemble contient les nombres « proches » de 3. Ainsi la valeur 3 appartient complètement à cet ensemble et il a par conséquent, un degré égal à 1. Plus on s'éloigne de la valeur 3, plus le degré d'appartenance diminue.

Cette propriété est exploitée en traitement d'image, et en classification des données où les classes, appelées aussi régions, sont représentées par des ensembles flous, Cela est fort utile lorsque les régions ne peuvent pas entre définies de manière nette et précise. La classification floue autorise le chevauchement des régions. La figure (II.5) montre un exemple de classification floue.

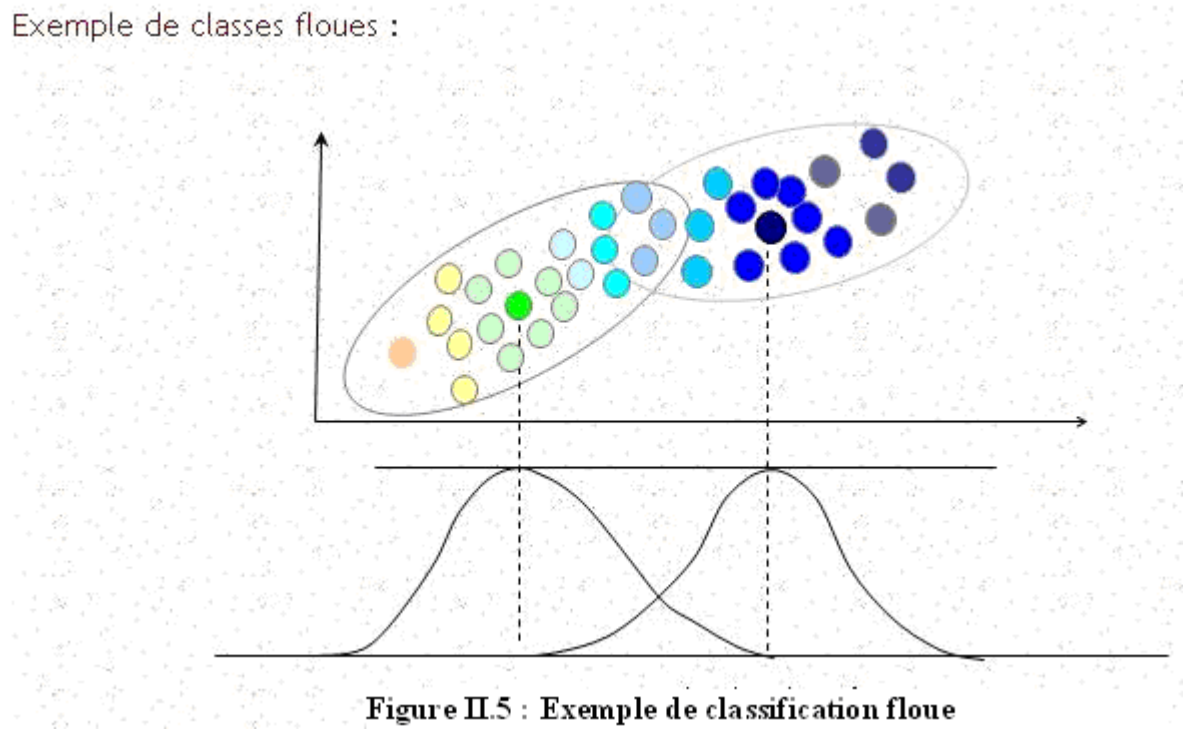


Figure II.5 : Exemple de classification floue

On présentera dans ce qui suit un algorithme de classification très populaire, basé sur la logique floue, connu pour son efficacité et sa robustesse.

II.5.2- L'algorithme de C-moyens flou (FCM)

L'algorithme de Fuzzy C-Means (FCM) est une extension directe de l'algorithme classique C-means, où la notion d'ensemble flou est introduite dans la définition des classes. Cet algorithme a été développé essentiellement par Bezdek, à partir des idées de Ruspini (clustering flou) et de Dunn (ISODATA flou). Contrairement à la classification exclusive où un individu peut appartenir à une seule classe, en classification floue un individu peut appartenir à plusieurs classes avec des degrés différents.

Le principe de base est de former à partir des individus, non étiquetés, K groupes qui soient les plus homogènes et naturels possible. '' Homogène ''et '' naturel '' signifient que les groupes obtenus doivent contenir des individus les plus semblables possible, tandis que des

individus des groupes différents doivent être le plus dissemblables possible. Le comportement flou en classification est alors présenté par une matrice qui quantifie le degré d'appartenance de chaque observation à chaque classe. Cette matrice notée u est de dimension ($N \times K$) où N est le nombre d'individus et K est le nombre de classes. Les éléments $u_{ij}\left(i = 1,...,N \; et \; j = 1,...,K\right)$ de cette matrice sont des nombres réels dont les valeurs sont comprises entre 0 et 1. La matrice u obéit aux conditions suivantes :

$$u_{ij} \in [0,1] \qquad \forall \; i, j.$$

$$\sum_{i=1}^{K} u_{ij} = 1 \qquad \forall \; j.$$

$$0 \prec \sum_{j=1}^{N} u_{ij} \prec N \; .$$

Par exemple, considérons un ensemble X de N individus à partir de deux classes.
La matrice d'appartenance suivante montre une répartition floue des N individus en 2 classes :

$$u_{K \times N} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .... & .... \\ 0.9 & 0.1 \end{bmatrix}$$

Le premier élément x_1 possède un degré d'appartenance à la première classe égale à $u_{11} = 0.8$ et un degré d'appartenance à la deuxième classe égale à $u_{12} = 0.2$. Cette matrice vérifie toutes les conditions requises.

Il est facile de passer d'une partition floue à une partition exclusive en prenant la valeur maximale des éléments de la matrice u sur chaque ligne. La matrice d'appartenance devient alors :

$$u_{K \times N} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ .... & .... \\ 0 & 1 \end{bmatrix}$$

Dans ce cas, chaque observation appartient à seulement à une seule classe indiqué par la valeur1.

L'algorithme de C-moyens flou (FCM) consiste à minimiser la fonction objective suivante :

$$J_m = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ik}^m d_{ik}^2 \qquad 1 \leq m \prec \infty$$

Où m est un nombre réel supérieur à 1, appelé facteur flou. u_{ik} est le degré d'appartenance de l'élément x_i à la classe C_k et d_{ik} est la distance séparant l'objet x_i et le centre v_k de la classe C_k.

Cette distance exprime la similarité entre l'observation et le centre d'une classe.

$d_{ik} = \|x_i - v_k\| \succ 0 \qquad \forall\ i\ et\ k$.

La classification floue s'effectue par le biais d'une optimisation itérative de la fonction objective cité ci-dessus. Celle-ci s'effectue par une mise à jour permanente du degré d'appartenance u_{ik} et du centre de classe $v_{k.}$

Les centres de classes sont mis à jour après chaque nouvelle itération des observations. Cette mise à jour est obtenue en fonction des degrés d'appartenances selon la formule suivante :

$$v_k = \frac{\sum_{i=1}^{N} u_{ik}^m x_i}{\sum_{i=1}^{N} u_{ik}^m} \qquad \forall i\,.$$

La mise à jour de ces centres de classes modifie la répartition des observations, par conséquent une mise à jour de la matrice u est nécessaire. Elle est effectuée en utilisant la formule suivante :

$$u_{ik} = \frac{1}{\sum_{j=1}^{K}\left(\frac{d_{ik}}{dij}\right)^{2/m-1}}$$

L'algorithme de Fuzzy C-means est résumé par les étapes suivantes :

1- Fixer le nombre k de classes et le facteur flou m.
2- Initialisation des centres de classes, t=0.
3- Déterminer la matrice d'appartenance $u^{(0)}$.
4- t=t+1, détermination des nouveaux centres des classes.
5- Déterminer la matrice d'appartenance $u^{(k)}$.
6- Si $\|u^{(t+1)} - u^{(t)}\| \leq \varepsilon$, s'arrêter ; sinon aller à l'étape 3.

Le critère d'arrêt est indiqué par la dernière étape de l'algorithme, il consiste à vérifier si la répartition entre deux itérations est la même ou non. ε est un nombre réel dont la valeur est comprise entre 0 et 1, alors que la variable t indique le numéro de l'itération.

L'algorithme de FCM soufre des mêmes inconvénients que l'algorithme de C-means à savoir le choix de la répartition initial et le choix de nombre de classes.
Différentes mesures ont été proposées pour évaluer la qualité de la classification et valider ses résultats. Dans ce cas, ces mesures sont dénommées *indices de validité*. Dans ce qui suit, nous présenterons quelques notions sur la validité des méthodes de classification. Quelques indices de validité seront également présentés.

II.6- Notion de validité en classification [HAL01]

Lorsqu'on est confronté à un problème de classification non supervisée, on est amené à faire des suppositions sur le nombre de classes présentes dans l'ensemble des données. Cependant, on ne dispose pas toujours d'informations à priori sur la structure interne de ces données pour nous aider à choisir le nombre optimal de classes correspondant. Dans ce cas, l'utilisateur doit appliquer un algorithme de classification non supervisée avec les différentes valeurs de ce qu'il estime plausibles (il est toujours possible de limiter le domaine des valeurs éventuelles que peut prendre ce paramètre), et de choisir la partition optimale correspondant à son problème. On est alors obligé de définir un critère, ou une fonction de validité, mesurant la performance de la classification pour choisir la partition optimale parmi toutes celles obtenues avec les différentes valeurs plausibles, et testées, du nombre de classes recherché.
Bien que cette étape de validation puisse manifestement paraître cruciale, le problème de la validité des partitions obtenues par des méthodes de classification non supervisées n'est toutefois pas facile. La difficulté provient du fait qu'il n'existe aucun critère universel qui puisse décider de ce qu'un algorithme donné soit adapté à un ensemble de données quelconques, et c'est souvent sur la base de constatations empiriques que l'on se fait une idée sur la distribution réelle des données traitées.

L'étude de la validité d'une répartition présente deux aspects :
- D'une part, il s'agit d'étudier l'existence ou non d'une structure quelconque au sein des données, c'est-à-dire, voir si les données sont distribuées d'une façon aléatoire ou peuvent être regroupées dans des groupes bien définis,
- Il faudrait, d'autre part, étudier si les classes identifiées sont bien réelles, en ce sens qu'elles doivent être liées aux propriétés intrinsèques des données, et non pas être juste un artefact de calcul ou un pur produit de l'algorithme utilisé.

Les fonctions de validité s'attaquent au premier problème. Quant au deuxième aspect, c'est l'utilisateur qui, par le choix de l'algorithme qu'il utilise, suppose ce que sont les propriétés des classes qu'il recherche. Par exemple, le choix de l'algorithme C-means laisse entendre que sont recherchées des classes compactes, sphériques et séparées. Une fonction de validité ne peut pas décider si un algorithme donné est bien adapté aux données analysées. Elle ne fait que refléter à quel degré la partition obtenue avec cet algorithme obéisse aux critères inhérents à ce dernier.

Ainsi une fonction de validité a pour but d'attribuer, à une partition donnée, un coefficient qui reflète la qualité de la classification obtenue à l'aide de l'algorithme utilisé. Dans le cas du FCM, par exemple, en évaluant cette fonction pour différents choix de valeurs de k et de m, on peut espérer identifier les valeurs optimales de ces deux paramètres qui correspondent à une partition reproduisant au mieux la structure des données traitées. Dans le cas général, une partition est d'autant meilleure que les éléments attribués à une classe donnée sont plus proches du centre de cette classe. Or, les degrés de similitude entre ces points et un centre quelconque sont mesurés par leurs degrés d'appartenance à la classe correspondant à ce centre. Si, pour un élément donné x_i, l'un des k degrés d'appartenance u_{ij}, est très largement supérieur aux (k - 1) autres degrés, alors ce point a toutes les chances d'être un bon représentant de la classe C_j correspondante. Si, au contraire, tous ses k degrés d'appartenance ont des valeurs voisines, alors la classe de ce point est indéterminée.

La validation en classification vérifie :
1- S'il y a réellement des regroupements « naturels » dans les données.
2- Si les groupes identifiés sont en accord avec nos (éventuelles) connaissances a priori du problème, on parle dans ce cas de validation externe. Ces connaissances peuvent être confirmatoires et non prescriptives (pas directement exploitables dans la fonctionnelle de coût).
3- Si les groupes identifiés sont bien « ajustés » aux données. Il s'agit alors de validation interne.

II.7- Indices de validité

On s'intéresse dans ce mémoire à l'étude de quelques indices de validité proposés dans le cadre de la validation interne. Dans cette section, les indices les plus souvent cités

dans la littérature seront brièvement présentés. Chaque indice sera référencié par son auteur. Un indice de validité est composé d'une mesure de la compacité des classes et d'une mesure sur leur séparabilité. En effet, une bonne classification se manifeste généralement par des classes très compactes et bien séparées.

II.7.1- Indice de bezdek

II.7.1.a- Indice du coefficient de partition (PC)

Initialement proposé par Bezdek, cet indice représente la moyenne de tous les éléments de matrice d'appartenance *U*. **[BEZ75]**, Il a pour expression la fonction suivante :

$$PC(k) = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N}(u_{ij})^2}{N}$$

tel que PC (k) $\in$ [1/k, 1].

Théoriquement, la classification est d'autant plus satisfaisante que ce coefficient est élevé, et donc, plus proche de 1. En pratique, ce coefficient s'est avéré être sensible aux valeurs de *k* et de *m*, et la valeur maximale de PC ne correspond pas forcément à la partition optimale.

II.7.1.b- Indice de l'entropie moyenne de la partition (PE)

Egalement proposé par Bezdek, cet indice prend en considération l'entropie du degré d'appartenance de chaque individu appartenant à chaque classe **[BEZ75]**. Il est donné par la formule suivante.

$$PE(k) = -\frac{\sum_{j=1}^{k}\sum_{i=1}^{N}u_{ij} \times \ln u_{ij}}{N}$$

tel que PE (k) $\in$ [0, ln k].

Cet indice demeure aussi sensible aux paramètres *k* et *m*. Une valeur proche de 0 indique une bonne classification.

II.7.2- Indices de Gath et Geva (*FH*)

D'autres indices de validité basés sur la matrice *U* ont été proposés dans **[GAT89]**. Soit $\sum_j$ la matrice covariance floue de la classe C_j telle que :

$$\sum_j = \frac{\sum_{i=1}^{N} u_{ij}^m (x_i - v_j)(x_i - v_j)^T}{\sum_{i=1}^{N} u_{ij}^m}$$

L'hyper volume flou de la classe C_j est donné par l'équation suivante :

$V_j = |\Sigma_j|^{1/2}$ où $|\Sigma_j|$ est le déterminant de la matrice Σ_j, il correspond à une mesure de la compacité de la classe C_j.

L'hyper volume flou de toutes les classes est alors :

$$FH(k) = \sum_{j=1}^{k} V_j$$

Une faible valeur de FH (k) indique l'existence de classes compactes.

D'autres facteurs comme : « Average Partition Dansity (PA (k)) »

$$PA(k) = \frac{1}{k} \sum_{j=1}^{K} \frac{S_j}{V_j} \qquad \text{tel que} \qquad S_j = \sum_{x_i \in C_j} u_{ij}$$

ou « Density Partition (DP (k)) » :

$$DP(k) = \frac{S}{FH(k)} \qquad \text{tel que} \qquad S = \sum_{j=1}^{k} S_j$$

peuvent être utilisés comme indices de validité pour la classification.

La valeur k^* optimale est obtenue en minimisant le critère DP (k).

II.7.3- Indice de Xie et Beni (XB)

Xie et Beni ont défini un indice qui traduit le rapport entre les mesures de compacité et de séparation des classes **[XIE91]**.

La compacité est représentée comme suit :

$$J(U,v) = \sum_{j=1}^{k} \sum_{i=1}^{N} (u_{ij})^2 \left\| (x_i - v_j) \right\|^2$$

alors que la séparation entre les classes est évaluée par le carré de la distance minimale entre les centres de deux classes.

$$(d_{\min})^2 = \left\{ \min_{i,j,i \neq j} \left\| v_i - v_j \right\| \right\}^2$$

L'indice de Xie et Beni se formulera donc comme suit :

$$XB(k) = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N}(u_{ij})^2 \left\| x_i - v_j \right\|^2}{N * \min_{i,j=1\ldots k}^{k} \left\| v_i - v_j \right\|^2}$$

La valeur de k optimale est obtenue en minimisant le critère XB(k)

II.7.4- Indice de Kwon (*KW*)

L'indice proposé par Kwon modifie celui de Xie et Beni afin d'éliminer sa nature monotone décroissante. Il diffère de celui de Xie et Beni par l'ajout d'un terme au numérateur. Il est donné par la formule suivante **[KWO98]** :

$$KW(k) = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N} u_{ij}{}^2 \left\| x_i - v_j \right\|^2 + \frac{1}{k}\sum_{j=1}^{k} \left\| v_j - v \right\|^2}{\min_{i \neq j, j=1,\ldots,k} \left\| v_i - v_j \right\|^2}$$

Avec $v = \frac{1}{N}\sum_{i=1}^{N} x_i$ le vecteur moyen de l'ensemble des données à classer.

La valeur k optimale est obtenue en minimisant le critère *KW (k*).

II.7.5- Indice de Maulik (*I*)

Maulik a proposé un indice que l'on notera I (k) et qui est défini comme suit **[MAU04]** :

$$I(k) = \left(\frac{1}{k} \times \frac{E_1}{E_k} \times D_k \right)^q$$

Avec :

$$E_k = \sum_{j=1}^{k}\sum_{i=1}^{N} \mu_{ij} \left\| x_i - v_j \right\|$$

et :

$$D_k = \max_{i \neq j, j=1}^{k} \left\| v_i - v_j \right\|$$

Le terme *1/k* a pour effet de réduire l'indice *I(k)* quand *k* augmente alors que le terme E_k, qui est lié à la compacité des classes, décroît lorsque *k* augmente et par conséquent *I(k)* augmente. E_1 est le facteur de normalisation employé de façon à éviter une la valeur minimale de l'indice en question. Le troisième terme D_k permet de mesurer la séparabilité des classes, tend à augmenter avec l'augmentation de *k*. Ces trois termes se complémentent en contre balançant l'effet de l'un par rapport à l'autre. La puissance *q* est utilisée afin de contrôler le contraste entre les différentes configurations des classes. La valeur *k* optimale est obtenue en maximisant l'indice *I(k)*.

II.7.6- L'indice de Fukuyama et Sugeno (FS)

L'indice de Fukuyama et Sugeno est défini comme suit **[MYM 01]** :

$$FS(k) = \sum_{j=1}^{K}\sum_{i=1}^{N} u_{ij}^{m} \left\| x_i - v_j \right\|^2 - \sum_{j=1}^{K}\sum_{i=1}^{N} u_{ij}^{m} \left\| v_j - v \right\|^2$$

Où $$v = \sum_{i=1}^{K} v_j / N$$

Le premier terme mesure la compacité des classes, alors que le deuxième mesure la séparabilité des centres de classes (v_j) avec le centre moyen de toutes les classes (v).

Une faible valeur de FS indique l'existence de classes compactes et bien séparées.
La valeur K optimale est obtenue en minimisant le critère FS(k).

II.7.7- L'indice de Wu et Yang (PCAES)

Récemment Wu et Yang **[WY 05]** ont proposé un indice de validité qui est défini comme suit :

$$PCAES(k) = \sum_{j=1}^{k}\sum_{i=1}^{N} u_{ij}^{2} / u_M - \sum_{j=1}^{k} \exp\left(-\min_{\substack{j=1 \\ j\neq i}}^{k} \left\{ \left\| v_i - v_j \right\|^2 \right\} / \beta_T \right)$$

avec $u_M = \max_{1\le j\le k} \sum_{i=1}^{N} u_{ij}^2$ et $\beta_T = \frac{1}{k}\sum_{j=1}^{k} \left\| v_j - v \right\|^2$

L'indice *PCAES (k)* contient deux termes, le premier terme est le coefficient de partition normalisé qui mesure la compacité, le deuxième terme est de type exponentiel, il mesure la séparabilité en effectuant la somme des distances entre les pairs des centres de classe les plus proches. La valeur K optimale est obtenue en maximisant le critère *PCAES(k)*.

II.7.8- Indice de Calinski Harabasz (CH)

Appelé aussi indice VRC (Variance Ratio Criterion), cet indice est basé sur la variance intra classes, c'est-à-dire la dispersion des points dans chaque classe et la variance interclasses **[CAH74]**. Il est donné par l'expression suivante :

$$CH(k) = \frac{dispertion\ \text{int}\ er\ classes}{dispertion\ \text{int}\ ra\ classe} = \frac{B}{W}$$

avec

$$B = \frac{1}{k-1}.\sum_{j=1}^{k} N_j \left\| v_j - v \right\|^2$$

et

$$W = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{N_j} \left\| x_i - v_j \right\|^2 \quad , \ x_i \in C_j$$

N_j correspond au nombre d'observations de la classe C_j.

W est une mesure de compacité des classes alors que B est lié à la séparabilité de celles-ci. La valeur maximale de l'indice CH(k) indique la meilleure classification et le nombre k correspondant représente le nombre correct de classes.

II.7.9- Indice de Davies-Bouldin (DB)

C'est un indice basé sur la minimisation du rapport des dispersions intra classes S_i et de la séparation inter classes R_i **[DAV79]**. La dispersion de la i^éme^ classe est calculée comme suit :

$$S_{i,q} = \frac{1}{|N_i|}\sum_{j=1}^{n_i} (\left\| x_j - v_i \right\|_2^q)^{\frac{1}{q}}$$

La séparation inter classe est :

$$R_{i,qp} = \max_{i,j=1,i\neq j}^{k} (\frac{S_{i,q} + S_{j,q}}{d_{i,j,p}})$$

$d_{i,j,p}$ est la distance entre les centres des deux classes C_i et C_j. tel que :

$$d_{i,j,p} = \left\| v_i - v_j \right\|_p$$

$$DB(k) = \frac{1}{k}\sum_{i=1}^{k} R_{i,qp}$$

q et p étant des nombres entiers dont les valeurs sont fixées par l'utilisateur. L'indice DB est petit si les classes sont compactes et éloignées les unes des autres. Par conséquent, l'indice de Davies-Bouldin aura une petite valeur quand la classification en k^* classe sera de bonne qualité.

II.7.10- Indice de Dunn (D)

Dunn **[DUN74]** a défini un indice basé sur l'identification d'ensemble de classes compactes et bien séparées. Soient S et T deux sous ensembles non vides de R^p. Le diamètre Δ de S et la distance δ sont alors :

$$\Delta(S) = \max_{x,y \in S}(d(x,y))$$

$$\delta(S,T) = \min_{x \in S, y \in T}(d(x,y))$$

d(x, y) étant une distance entre les éléments x et y.

L'indice de Dunn est donné par la formule suivante :

$$D(k) = \min_{1 \le i \le k-1} \left\{ \min_{i+1 \le j \le k, i \neq j} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \le k \le K}(\Delta(C_k))} \right\} \right\}$$

δ (C_i, $C_{j)}$ correspond à la distance entre deux classes C_i et C_j. Elle donne une mesure sur la séparabilité des deux classes. Tandis que max ($\Delta(C_k)$) donne une mesure sur le diamètre de la classe C_k et correspond à la compacité. Une grande valeur de D(k) indique la présence de classes compactes et bien séparées.

II.7.11- Indice de Razae (R)

L'expression de cet indice noté R (k) est **[RAZ89]** :

$$R(k) = \alpha * Scat(k) + Disc(k)$$

Où *Scat (k)* correspond à la compacité de la répartition définie comme suit:

$$Scat(k) = \frac{\frac{1}{k}\sum_{j=1}^{k} \left\| \sigma(v_j) \right\|}{\left\| \sigma(X) \right\|}$$

Le terme $\sigma(X)$ est la variance totale de l'ensemble d'individus X.

$$\sigma(X)=\left\{\sigma(X)^1,\sigma(X)^2,...,\sigma(X)^p,...,\sigma(X)^P\right\} \qquad \text{avec} \qquad \sigma(X)^p=\frac{1}{N}\sum_{i=1}^{N}\left(x_i^p-v^p\right)^2$$

tel que

$$v^p=\frac{1}{N}\sum_{i=1}^{N}x_i^p$$

le terme $\sigma(v_j)$ indique la variance de la classe C_j.

$$\sigma(v_i)=\left\{\sigma(v_j)^1,\sigma(v_j)^2,...,\sigma(v_j)^p,...,\sigma(v_j)^P\right\}$$

avec

$$\sigma(v_j)^p=\frac{1}{N_j}\sum_{i=1}^{N_j}(x_i^p-v_j^p)^2 \quad avec\ x_i\in C_j$$

La fonction distance *Dis(k)* indique la séparabilité entre les classes. Elle est définie par :

$$Dis(k)=\frac{D_{\max}}{D_{\min}}\sum_{i=1}^{k}\left(\sum_{j=1}^{k}\left\|v_j-v_i\right\|\right)^{-1}$$

où

$$D_{max}=\max_{\substack{i,j=1\\ i\neq j}}^{k}\|\ v_j-v_i\ \| \qquad \text{et} \qquad D_{min}=\min_{\substack{i,j=1\\ i\neq j}}^{k}\|\ v_j-v_i\ \|$$

Le facteur $\alpha=Dis(k_{max})$ est introduit pour compenser la différence d'échelle entre *Dis(k)* et *Scat(k)*. La valeur de *k* pour laquelle *R (k)* est maximal correspond à un nombre correct de classes.

II.7.12- Indice de Boudraa (B)

L'indice de validité proposé par Boudraa dans **[BOU01]** est donné par l'équation suivante:

$$B(k)=\frac{\max\limits_{\substack{i,j=1\\ i\neq j}}^{k}\delta(v_i,v_j)}{\min\limits_{\substack{i,j=1\\ i\neq j}}^{k}\delta(v_i,v_j)}+\alpha\frac{1}{k}\frac{\sum_{p=1}^{P}\sum_{j=1}^{k}\sigma(v_j)^p}{\sum_{p=1}^{P}\sigma(X)^p}$$

où $\delta(v_i,v_j)$ représente la distance entre deux centres v_i et v_j de deux classes C_i et C_j. $\sigma(v_j)^p$ et $\sigma(X)^p$ correspondent respectivement à la variance du p^{ème} attribut de la classe

C_j et de l'ensemble X total des individus dont les expressions sont présentées dans le paragraphe précédent. α permet de compenser la différence d'échelle entre les deux termes de *B(k)*.

Le nombre k^* correct de classes est obtenu en minimisant *B(k)*.

II.7.13- Indice de De Franco (Icc)

Egalement appelé *Inter Class Contrast* (Icc) **[FRA02]**, cet indice de validité est défini comme suit:

$$Icc(k) = \frac{traceB}{N} D_{\min} \sqrt{k}$$

où *trace B* est un indice de compacité des classes, il est donné par l'expression de l'équation :

$$trace\ B = \sum_{j=1}^{k} N_j \left\| v_j - v \right\|^2$$

et

$$D_{min} = \min_{1 \le j \le k} \left\{ \min_{\substack{1 \le i \le k-1 \\ i \ne j}} \| \ v_j - v_i \ \| \right\}$$

une mesure de séparabilité des classes.

La valeur de k^* qui maximise *Icc*(k) correspond au nombre optimal de classes.

II.7.14- Indice de Turi (V)

Cet indice est formulé comme suit **[TU 05]** :

$$V(k) = (c * N(2,1) + 1) * \frac{W}{B}$$

N(2,1) est une distribution gaussienne de moyenne μ=2 et d'écart type égal à 1 et *c* est un paramètre à spécifier.

W est la variance intra classes donnant une mesure sur la compacité d'une partition :

$$W = \frac{1}{N} \sum_{j=1}^{k} \sum_{i \in C_j} \left\| x_i - v_j \right\|^2$$

et

$$B = \min_{\substack{j=1 \\ i=j+1}}^{k-1} \left\{ \| v_j - v_i \|^2 \right\}$$

B mesure la séparation des classes. A une valeur maximale de V (k) correspond un nombre optimal de classes.

II.7.15-L'indice VCR

Cet indice mesure le rapport entre le moment interclasse et le moment intra classe. Il est défini par l'expression :

$$R = \frac{\sum_{j=1}^{K} \sum_{i=1}^{N_j} \left\| x_i - v_j \right\|^2}{\sum_{k=1}^{K} \left\| v_j - v \right\|^2}$$

où *K* est le nombre de classes,

$$v_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i \quad x_i \in C_j$$

La valeur minimale de cet indice indique la meilleure partition de l'ensemble de données

II.7.16-L'indice de Krzanowiski et lai (KL)

Cet indice est défini par l'expression **[KL85]** :

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right|$$

où $DIFF(k) = (k-1)^{2/p} W(k-1) - (k)^{2/p} W(k)$, p, le nombre d'attributs de l'ensemble de données. La valeur de k qui maximise *KL(k)* donne la valeur optimale de nombre de classes.

II.7.17-L' indices RMSSDT

Ces indices sont employés habituellement pour évaluer les résultats des algorithmes de classification hiérarchique. Cependant, ils peuvent être employés pour évaluer les résultats de n'importe quel algorithme de classification. L'indice de RMSSTD (Root-Mean-Square Sandard Deviation) est la racine carrée de la variance des attributs de l'ensemble de données. Il est défini par l'expression **[MYM 01]** :

$$RMSSTD = \left[\frac{\sum_{j=1}^{K} \sum_{j=1}^{N_j} (X_i - v_j)}{\sum_{j=1}^{K} (N_j - 1)} \right]^{\frac{1}{2}}$$

Cet indice mesure l'homogénéité des classes. Comme le but du processus de classification est d'identifier les classes homogènes, la valeur la plus basse de RMSSTD signifie que le résultat de classification donne la meilleure partition.

II.7.18-L'indice de Chou et Sun (CS)

Chou et Sun ont proposé un nouveau indice de validité de la classification , CS (k) peut être defini comme suit **[CS 03]:**

$$CS(k) = \frac{\sum_{i=1}^{k} \left[\frac{1}{N_i} \sum_{x_j \in C_i} \max_{x_q \in C_i} \{d(x_i, x_q)\} \right]}{\sum_{i=1}^{k} \left[\min_{\substack{j=1,\dots k \\ i \neq j}} \{d(v_i, v_j)\} \right]}$$

$d(v_i, v_j)$ correspond à la distance entre deux centres de classes C_i et C_j respectivement ,elle donne une mesure sur la séparabilité de deux centres de classes La valeur de *k* pour laquelle *CS (k)* est minimal correspond à un nombre correct de classes.

II.8- Conclusion

Dans ce chapitre, nous avons présenté quelques notions de classification. Deux algorithmes (C-means et Fuzzy C-means), très populaires qui sont de nos jours les plus utilisés en pratique ont été décrits. Ils ont l'avantage d'être à la fois simples et efficaces. Cependant ces deux algorithmes souffrent principalement du problème du choix du nombre de classes. Si aucune information n'est disponible, le choix du nombre de classes reste délicat, surtout lorsque la dimension des attributs relatifs à ces données est très élevée et le résultat de la classification risque d'être erroné

Pour s'acquérir de cette information importante, on doit faire appel au calcul d'indices de validité. Plusieurs indices de validité utilisés pour la classification non supervisée des données ont été proposés dans la littérature. Dans ce chapitre, nous avons présenté juste quelques un d'entre eux, ceux qui sont les plus populaires. Certains d'entre eux peuvent être

utilisés avec les algorithmes de classification floue alors que les autres ne peuvent être utilisés que pour une classification exclusive. Quelques indices seront utilisés dans le chapitre suivant dans le cadre de la segmentation d'image.

III.1- Introduction

Dans le premier chapitre, nous avons vu que pour la segmentation d'une image par seuillage d'histogramme nécessite la connaissance au préalable du nombre de seuils ou de classes. Par conséquent, nous formulons ce problème d'une manière identique à celui rencontré en classification des données pour la recherche du nombre de classes.
Pour résoudre ce problème, plusieurs indices de validités ont été proposés dans le cadre de la classification (chapitre II).

Dans ce chapitre notre travail consiste a adapter ces derniers, afin de pouvoir les exploiter dans le cadre de la segmentation d'image par seuillage.

III.2- Indices de Bezdek

III.2.a- Indice du coefficient de partition (PC)

Rappelons qu'en classification, cet indice est défini comme suit :

$$PC(k) = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N}(u_{ij})^2}{N}$$

où μ_{ij} représente le degré d'appartenance de l'objet ''i'' à la classe C_j .

Or dans le cadre de notre travail, on s'intéresse aux méthodes de seuillage ayant un caractère exclusif, c'est-à-dire qu'un pixel peut appartenir qu'à une seul classe. Dans ce cas le degré d'appartenance des pixels μ_{ij} est égale soit à 0 ou soit à 1. Nous pouvons ainsi remarquer que :

$$\sum_{j=1}^{k}\sum_{i=1}^{N}\mu_{ij} = N$$

Par conséquent l'indice de PC(k) s'écrira de la manière suivante **:** $PC = \frac{N}{N} = 1 = Cste$

Ainsi cet indice ne peut pas être utilisé pour le cas du seuillage d'histogramme.

III.2.b- Indice de l'entropie moyenne de la partition (PE)

Comme les μ_{ij} prennent soit la valeur 0 soit la valeur 1, on peut facilement voir que l'indice PE devient indéfini.

III.3- Indice de Gath et Geva (FH)

Puisque $\mu_{ij} = \begin{cases} 1 & x_i \in C_j \\ 0 & \sin on \end{cases}$, on déduit que: $\sum_{i=1}^{N} u_{ij} = N_j$

Où N_j est le nombre de point dans la classe C_j.

On peut aussi écrire :

$$\sum_{i=1}^{N} u_{ij}^m (x_i - v_j)(x_i - v_j)^T = \begin{cases} \sum_{i=1}^{N} (x_i - v_j)(x_i - v_j)^T \; si \; x_i \in C_j \\ 0 \qquad\qquad \sin on \end{cases}$$

$$= \begin{cases} \sum_{i=1}^{N} (x_i - v_j)^2 \; si \; x_i \in C_j \\ 0 \qquad\qquad \sin on \end{cases}$$ si x_i est un niveau de gris.

Sachant qu'une classe C_j est délimitée par les seuils [t_j t_{j+1}-*1*], on peut alors facilement déduire que :

$$\sum_{i=1}^{N} u_{ij}^m (x_i - v_j)(x_i - v_j)^T = \sum_{i=t_j}^{t_{j+1}-1} (i - \mu_j)^2 * h(i)$$

Où μ_j est le niveau de gris moyen de la classe C_j : $\mu_j = \frac{1}{N_j} \sum_{i=t_j}^{t_{j+1}-1} ih(i)$.

h(i) est le nombre de pixels ayant le niveau de gris ''*i*''.

Cette expression n'est rien d'autre que la variance de la classe C_j. Dans ce cas, la matrice de covariance de la classe C_j s'écrira :

$$\sum_j = \frac{1}{N_j} \sum_{i=t_j}^{t_{j+1}-1} (i - \mu_j)^2 * h(i) = \sigma_j$$

La forme simplifiée de l'indice *FH* est : $FH(k) = \sum_{j=1}^{k} \sqrt{\sigma_j}$

Finalement l'indice *FH* représente la somme des racines carrés de la variance de chaque classes C_j

En effectuant les mêmes changements pour l'indice*'' Average Partition Dansity PA''* on obtient la formule suivante : $PA(k) = \frac{1}{k} \sum_{j=1}^{k} \frac{N_j}{\mu_j}$

où N_j représente le nombre de pixels de la classe C_j et μ_j le centre de la classe C_j.

Pour l'indice ''Density Partition (DP)'', s'écrira alors :

$$DP(k) = \frac{N}{FH(k)}$$

N étant le nombre de pixels total.

On remarque que l'indice de *DP* a presque la même forme que l'indice de *FH*, sauf qu'au lieu de chercher le minimum pour l'indice *FH* , on aura a chercher le maximum pour l'indice *DP* .

III.4- Indice de Xie et Beni (XB)

Comme pour les indices précédents, le degré d'appartenance μ_{ij} de chaque pixel prend soit la valeur 0, soit la valeur 1. En procédant aux mêmes modifications, on déduit que l'indice de Xie et Beni est de la forme suivante :

$$XB(k) = \frac{\sum_{j=1}^{k} \sum_{i=t_j}^{t_{j+1}-1} (i-\mu_j)^2 h(i)}{N * \min_{\substack{j=1 \\ i=j+1}}^{k} \|\mu_i - \mu_j\|^2}$$

III.5- Indice de Kwon (*KW*)

Comme pour l'indice de Xie et Beni, l'indice de Kwon s'écrira après quelques modification de la manière suivante :

$$KW(k) = \frac{\sum_{j=1}^{k} \sum_{i=t_j}^{t_{j+1}-1} (i-\mu_j)^2 h(i) + \frac{1}{k}\sum_{j=1}^{k} (\mu_j - \mu)^2}{\min_{\substack{j=1 \\ i=j+1}}^{k} \|\mu_i - \mu_j\|^2}$$

$\mu = \frac{1}{N}\sum_{j=1}^{L} jh(j)$; représente le niveau de gris moyen de l'image et $\mu_j = \frac{1}{N_j}\sum_{i=t_j}^{t_{j+1}-1} ih(i)$ le centre de la classe C_j.

III.6- Indice de Maulik (*I*)

En réspectant les modifications précedantes, il convient que le terme E_k , qui est lie à la compacité des classes, est donnée par la formule:

$$E_k = \sum_{j=1}^{k} \sum_{i=t_j}^{t_{j+1}-1} (i-\mu_j) * h(i)$$

Sachant que $\mu_1 < \mu_2 < .... < \mu_k$, le terme D_k qui représente la distance maximale entre deux centres de classe est: $D_k = \max_{\substack{j=1 \\ j \neq i}}^{k} (\mu_i - \mu_j) = (\mu_k - \mu_1)$

Par conséquent, l'indice de maulik, transformé, sera représenté par la formule ci dessous :

$$I(k) = \frac{1}{k} * \frac{\sum_{i=t_1}^{t_2-1} (i - \mu_1) * h(i)}{\sum_{j=1}^{k} \sum_{i=t_j}^{t_{j+1}-1} (i - \mu_j) * h(i)} \times (\mu_k - \mu_1)$$

III.7- Indice de Fukuyama et Sugeno (FS)

Compte tenu des nouvelles modifications, l'indice de Fukuyama et Sugeno en segmentation peut être réécrit comme suit :

$$FS(k) = \sum_{j=1}^{k} \sum_{i=t_j}^{t_{j+1}-1} \left((i - \mu_j)^2 * h(i) - (\mu_j - \mu)^2 \right)$$

$\mu = \sum_{j=1}^{k} \frac{\mu_j}{k}$; le centre moyen de toutes les classes.

Ces changements ne modifient pas la signification des termes de cet indice. Le premier terme désigne toujours la compacité des classes, alors que le deuxième mesure la séparabilité des centres de classes (μ_j) avec le centre moyen de toutes les classes (μ).

III.8- Indice de WU et Yan (PCAES)

Le premier terme de cet indice représente le coefficient de partition normalisé qui mesure la compacité, il s'écrira : $\frac{N}{N_M}$ car $\sum_{j=1}^{k} \sum_{i=1}^{N} \mu_{ij} = N$. N_M représente le nombre de pixel de la classe majoritaire $N_M = \max_{j=1}^{k} N_j$. Tandis que le deuxième terme qui mesure la séparabilité s'écrira comme suit :

$$\sum_{j=1}^{k} \exp\left(- \min_{\substack{j=1 \\ i=j+1}}^{k} \left\{ (\mu_j - \mu_i)^2 \right\} \Big/ \beta_T \right)$$

Finalement l'indice de PCAES prend la forme suivante :

$$PCAES(k) = \frac{N}{N_M} - \sum_{j=1}^{k} \exp\left(- \min_{\substack{j=1 \\ i=j+1}}^{k} \left\{ (\mu_i - \mu_j)^2 \right\} \Big/ \beta_T \right)$$

tel que $$\beta_T = \frac{1}{k}\sum_{j=1}^{k}\left(\mu_j - \mu\right)^2$$

III.9- Indice de Calinski Harabasz (CH)

L'indice de Calinski Harabasz donné par la formule suivante : $CH(k) = \frac{B}{W}$

où W est une mesure de compacité des classes tel que $W = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}\left(i - \mu_j\right)^2 * h(i)$

Le terme B est lié à la séparabilité : $B = \frac{1}{k-1}\sum_{j=1}^{k} N_j\left(\mu_j - \mu\right)^2$

L'indice CH aura l'expression générale suivante :

$$CH(k) = \frac{(N-k)}{(k-1)}\frac{\sum_{j=1}^{k} N_j(\mu_j - \mu)^2}{\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}\left(i - \mu_j\right)^2 * h(i)}$$

La valeur maximale de l'indice CH indique une bonne segmentation.

III.10- Indice de Davies-Bouldin (DB)

Il est aisé de voir dans l'indice de Davies-Bouldin que pour p=1 et q=1, la dispersion intra classe de la $i^{éme}$ classe est : $S_{i,1} = \frac{1}{N_i}\sum_{j=t_i}^{t_{i+1}-1}\left(\left(j - \mu_i\right)^2 * h(j)\right)$

est que la séparation inter classe R_i sera :

$$R_{i.1} = \max_{\substack{i.j=1 \\ i\neq j}}^{k}\left(\frac{S_{i1} + S_{j1}}{d_{i.j.1}}\right)$$

Avec $d_{i,j,1} = \left\|\mu_i - \mu_j\right\|$ qui est la distance entre les centres de classe $\mu_i \; et \; \mu_j$.

Finalement l'indice $DB(k) = \frac{1}{K}\sum_{i=1}^{k} R_{i,11}$, prend la forme suivante :

$$DB(k) = \frac{1}{k}\sum_{i=1}^{k}\max_{\substack{i.j=1 \\ i\neq j}}^{k}\frac{\frac{1}{N_i}\sum_{j=t_i}^{t_{i+1}-1}\left(j - \mu_i\right)^2 * h(j) + \frac{1}{N_j}\sum_{i=t_j}^{t_{j+1}-1}\left(i - \mu_j\right)^2 * h(i)}{\left|\mu_i - \mu_j\right|}$$

III.11- Indice de Dunn (D)

Nous avons vu dans le chapitre précédent que $\Delta(C_k) = \max_{x,y \in C_k} d(x,y)$ représente la distance maximale entre deux points appartenant à la même classe. Sachant qu'une classe C_j est délimitée par les seuils $[t_j \;\; t_{j+1}\text{-}1]$, cette expression devient :

$$\Delta(C_z) = \left| t_{z+1} - 1 - t_z \right|$$

On rappelle aussi que $\delta(S,T)$ représente la distance minimale entre deux points appartenant à deux classes différentes. On déduit alors que $\delta(C_j, C_i) = \left| t_j - 1 - t_i \right|$

Par conséquent, l'indice de Dunn est représenté par la formule suivante :

$$D(k) = \min_{i=1}^{k} \left\{ \min_{j=i+1}^{k} \left\{ \frac{\left| t_j - 1 - t_i \right|}{\max_{z=1}^{k} (t_{z+1} - 1 - t_z)} \right\} \right\}$$

Nous remarquons aussi que $\min_{i=1}^{k} \left\{ \min_{j=i+1}^{k} \left\{ t_j - 1 - t_i \right\} \right\} = 1$. Finalement cet indice s'écrira sous la forme simplifiée ci dessous :

$$D(k) = \frac{1}{\max_{z=1}^{k-1} (t_{z+1} - 1 - t_z)}$$

III.12- Indice de Razae (R)

En reprenant les différentes expressions de l'indice de Razae présenté dans le chapitre précédent et en procédant aux changements de notation adoptés pour la segmentation on a :

$$Scat(k) = \frac{\frac{1}{k} \sum_{j=1}^{k} \left\| \sigma(C_j) \right\|}{\left\| \sigma(X) \right\|}$$

Avec : $\sigma(X) = \frac{1}{N} \sum_{i=1}^{L} (i - \mu)^2 * h(i)$; $\sigma(C_j) = \frac{1}{N_j} \sum_{i=t_j}^{t_{j+1}-1} (i - \mu_j)^2 * h(i)$ et $\mu = \frac{1}{N} \sum_{i=1}^{L} i h(i)$

Comme $D_{\max} = (\mu_k - \mu_1)$ et $D_{\min} = \min_{\substack{j=1 \\ i=j+1}}^{k-1} (\mu_j - \mu_i)$, la séparabilité entre les classes est :

$$Dis(k) = \frac{(\mu_k - \mu_1)}{\left(\min_{\substack{j=1 \\ i=j+1}}^{k-1} (\mu_j - \mu_i) \right)} \sum_{j=1}^{k} \left(\sum_{\substack{i=1 \\ i \neq j}}^{k} \left\| \mu_i - \mu_j \right\| \right)^{-1}$$

L'indice de Razae donné par la formule suivante :

$$R(k) = \alpha * Scat(k) + Disc(k)$$

aura l'expression générale suivante :

$$R(k) = \alpha * \frac{\frac{1}{k}\sum_{j=1}^{k}\left\|\sigma(c_j)\right\|}{\left\|\sigma(x)\right\|} + \frac{(\mu_k - \mu_1)}{\left(\min_{\substack{j=1 \\ i=j+1}}^{k-1}(\mu_j - \mu_i)\right)}\sum_{j=1}^{k}\left(\sum_{\substack{i=1 \\ i\neq j}}^{k}\left\|\mu_j - \mu_i\right\|\right)^{-1}$$

Le facteur $\alpha = Dis(k_{max})$ permet de compenser la différence d'échelle entre *Dis(k)* et *Scat(k)*.

III.13- Indice de Boudraa (B)

En segmentation par seuillage, on considère le niveau de gris comme étant le seul attribut d'un pixel ; Dans ce cas, le facteur p prend la valeur 1. Et contenu des nouvelles notations, l'indice de Boudraa est donné par l'expression suivante

$$B(k) = \frac{(\mu_k - \mu_1)}{\left(\min_{\substack{j=1 \\ i=j+1}}^{k-1}(\mu_j - \mu_i)\right)} + \frac{\alpha}{k} \times \frac{\sum_{j=1}^{k}\sigma(C_j)}{\sigma(X)}$$

Où $\sigma(C_j)$ et $\sigma(X)$ sont défini de la même manière dans le paragraphe précédant.

III.14- Indice de De Franco (Icc)

L'indice de compacité des classes *trace B* est donné par l'expression suivante :

$traceB = \sum_{j=1}^{k} N_j(\mu_j - \mu)^2$;

et la mesure de séparabilité des classes, D_{min} est définit comme suit : $D_{\min} = \min_{\substack{j=1 \\ i=j+1}}^{k-1}(\mu_j - \mu_i)$

Sachant que $Icc(k) = \frac{traceB}{N} D_{\min}\sqrt{k}$

on déduit : $Icc(k) = \frac{\sum_{j=1}^{k} N_j(\mu_j - \mu)^2}{N} \min_{\substack{j=1 \\ i=j+1}}^{k-1}(\mu_j - \mu_i)\sqrt{k}$

III.15- Indice de Turi (V)

La variance intra classes donnant une mesure sur la compacité d'une partition s'écrira comme suit :

$$W = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}\left(i-\mu_j\right)^2 * h(i)$$

alors que la séparabilité des classes est donnée par :

$$B = \min_{\substack{j=1 \\ i=j+1}}^{K-1}(\mu_j - \mu_i)^2$$

Puisque : $V(k) = (c * N(2,1)+1) * \frac{W}{B}$,V (k) prendra la forme suivante :

$$V(k) = \left(c * N(2,1)+1\right) * \frac{\frac{1}{N}\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}\left(i-\mu_j\right)^2 * h(i)}{\min_{\substack{j=1 \\ i=j+1}}^{K-1}(\mu_j - \mu_i)^2}$$

c : est un paramètre à spécifier.

III.16- Indice (VCR)

En notant le moment intra classe par : $\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}\left(i-\mu_j\right)^2 * h(i)$ et le moment interclasse par l'expression suivante : $\sum_{j=1}^{k}\left(\mu_j-\mu\right)^2$, l'indice de VCR, qui mesure le rapport entre le moment intra classe et le moment inter classe, est défini par l'expression :

$$VCR(k) = \frac{\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}\left(i-\mu_j\right)^2 * h(i)}{\sum_{j=1}^{k}\left(\mu_j-\mu\right)^2}$$

III.17 -Indice de Krzanowiski et lai (KL)

Pour rappel, cet indice est donné par l'expression suivante :

$$KL(k) = \left|\frac{DIFF(k)}{DIFF(k+1)}\right|$$

Dans cette expression, la dispersion intra classe est représentée par la formule suivante :

$$W(k) = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}\left(i-\mu_j\right)^2 * h(i)$$

et *DIFF (k)= (k-1)^2 W(k-1)-(k)^2 W(k),*

III.18- Indices de RMSSTD

L'expression de l'indice RMSSTD définie dans le chapitre précèdent peut être réécrite comme suit :

$$RMSSTD(k) = \left[\frac{\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}(i-\mu_j)^2 * h(i)}{\sum_{j=1}^{k}\left(N_j-1\right)}\right]^{\frac{1}{2}}$$

En approximant la valeur $N_j - 1$ par N_j, et sachant que $\sum_{j=1}^{k} N_j = N$ on aura :

$$RMSSTD(k) = \left[\frac{\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}(i-\mu_j)^2 * h(i)}{N}\right]^{\frac{1}{2}}$$

III.19-Indice de Chou et Sun (CS)

En respectant les notations établies précédemment, l'indice de CS peut être reformulé comme suit:

$$CS(k) = \frac{\sum_{j=1}^{k}\frac{1}{N_j}\sum_{i=t_j}^{t_{j+1}-1}\max_{q=t_j}^{t_{j+1}-1}(d(j,q))}{\sum_{j=1}^{k-1}\left[\min_{i=j+1}^{k}(\|\mu_j-\mu_i\|)\right]}$$

$\mu_j = \frac{1}{N_j}\sum_{i=t_j}^{t_{j+1}-1} ih(i)$ qui représente la moyenne de la classe C_j.

d(j,q) représente la distance entre le niveau de gris ''*j''* et le niveau de gris ''*q''* dans l'histogramme de l'image.

III.20- Indice spécifiques en segmentation

En plus des indices de validités cités dans le chapitre précédent qui sont utilisés pour la classification des données, très peu d'indices de validité ont été proposé pour déterminer le nombre de seuils en segmentation d'images. Parmi eux nous citerons deux indices :

III.20.1-Indice de Deng et al (J)

Cet indice a été proposé par Deng Y .,et al [DA99], dans le cadre de la segmentation d'images couleurs. Il est donnée par la formule suivante:

$$J = \frac{S_B}{S_W} = \frac{S_T - S_W}{S_W}$$

Où S_B est la variance interclasse, la variance intra classe est $S_W = \sum_{j=1}^{k}\sum_{i=1}^{N_j}\left\|x_i - v_j\right\|^2$, la variance totale est donnée par $S_T = \sum_{i=1}^{N}\left\|x_i - v\right\|^2$

Pour une image en niveau de gris :

$$S_T = \sum_{i=1}^{L}(i-\mu)^2 * h(i); \quad \mu = \frac{1}{N}\sum_{i=1}^{L} ih(i) \ ; \ \mu_j = \frac{1}{N_j}\sum_{i=t_j}^{t_{j+1}-1} ih(i)$$

$S_W = \sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}(i-\mu_j)^2 * h(i)$; Par conséquent cet indice devient :

$$J = \frac{\sum_{i=1}^{L}(i-\mu)^2 * h(i) - \sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}(i-\mu_j)^2 * h(i)}{\sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}(i-\mu_j)^2 * h(i)} \ ;$$

En général, pour une segmentation d'image convenable il faut minimiser la valeur de J.

III.20.2- L'indice Yen de chang (*F*)

Pour déterminer le nombre de seuil optimal, Yen et Chang ont proposé l'indice suivant :

$$F(k) = \rho * (Disk(k))^{\frac{1}{2}} + (\log_2(k))^2$$

$Disk(k) = \sigma_w^2(k) = \sigma_T^2 - \sigma_B^2(k) = \sum_{j=1}^{k}\sum_{i=t_j}^{t_{j+1}-1}(i-\mu_i)* p_i$ Avec $p_i = \frac{h(i)}{N}$ et ρ une constante réelle positive. Le nombre de classe optimale est déterminé en minimisant l'indice *F*

III.21-Conclusion

Dans ce chapitre nous avons reformulé certains indices de validité, initialement proposé dans le cadre de la classification des données, dans le but de les utilisés pour déterminer le nombre de seuils en segmentation d'image.

Certains indices, comme l'indice du coefficient de partition PC(k) et l'indice d'entropie moyenne PE(k) ne peuvent pas être adaptés pour la segmentation d'image par seuillage d'histogramme, alors que d'autres ont été adaptés facilement. Ces indices seront testés et évalués dans le chapitre suivant.

IV.1- Introduction

Dans le chapitre précédent, nous avons présenté plusieurs indices de validité pour la segmentation d'images par seuillage d'histogramme. Dans ce chapitre nous allons évaluer leurs performances sur des histogrammes artificiels et sur des images réelles. Notons que l'utilisation des histogrammes artificiels permet d'évaluer objectivement les performances des différents indices de validités puisque la structure des histogrammes, ainsi que le nombre de classes est parfaitement contrôlé donc connue.

Les résultas des méthodes de calcul des seuils basés sur les Algorithmes Génétiques et l'Algorithme Itératif seront également présentés. Les fonctions d'Otsu et de Kapur présentées dans le premier chapitre sont utilisées comme des fonctions objectives pour le calcul des seuils. Notons que tous les programmes utilisés pour réalisé se travail ont été implémentés sous l'environnement MATLAB.

IV.2- Histogrammes artificiels

Dans ces exemples nous utiliserons des histogrammes générés artificiellement. Le nombre de modes ou de classes est fixé à priori.

Ces histogrammes sont constitués d'un mélange de k distributions gaussiennes selon la formule suivante :

$$h(i) = \sum_{j=i}^{k} \frac{P_j}{\sigma_j * sqrt(2\pi)} \exp\left(- \frac{\left(i - m_j\right)^2}{2 * (\sigma_j)^2} \right) \quad i = 1,...,L$$

où P_j, m_j et σ_j représentent respectivement la probabilité a priori, la valeur moyenne et l'écart type de la classe C_j.

k est le nombre de classes et L est le niveau de gris maximal.

Exemple 1

Pour cet exemple l'histogramme généré est composé de trois distributions gaussiennes dont les paramètres statistiques sont présentés dans le tableau (IV.1). L'allure de cet histogramme est représentée sur la figure (IV.1)

m_1	m_2	m_3	P_1	P_2	P_3	σ_1	σ_2	σ_3
80	175	200	12000	11000	11000	20	10	10

Tableau (IV.1)-Paramètres statistiques réels de l'histogramme

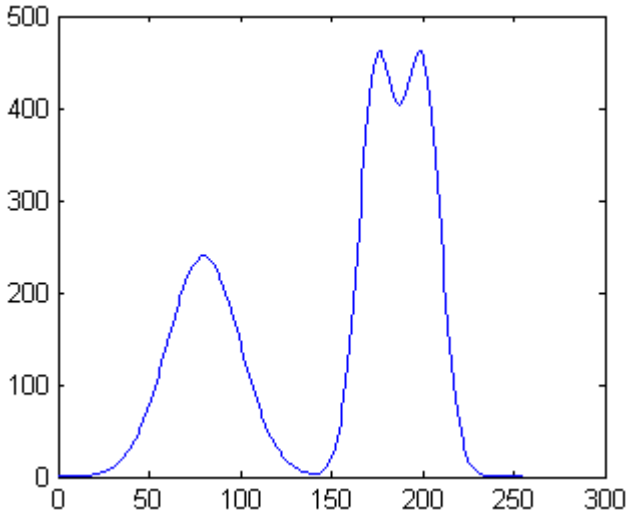


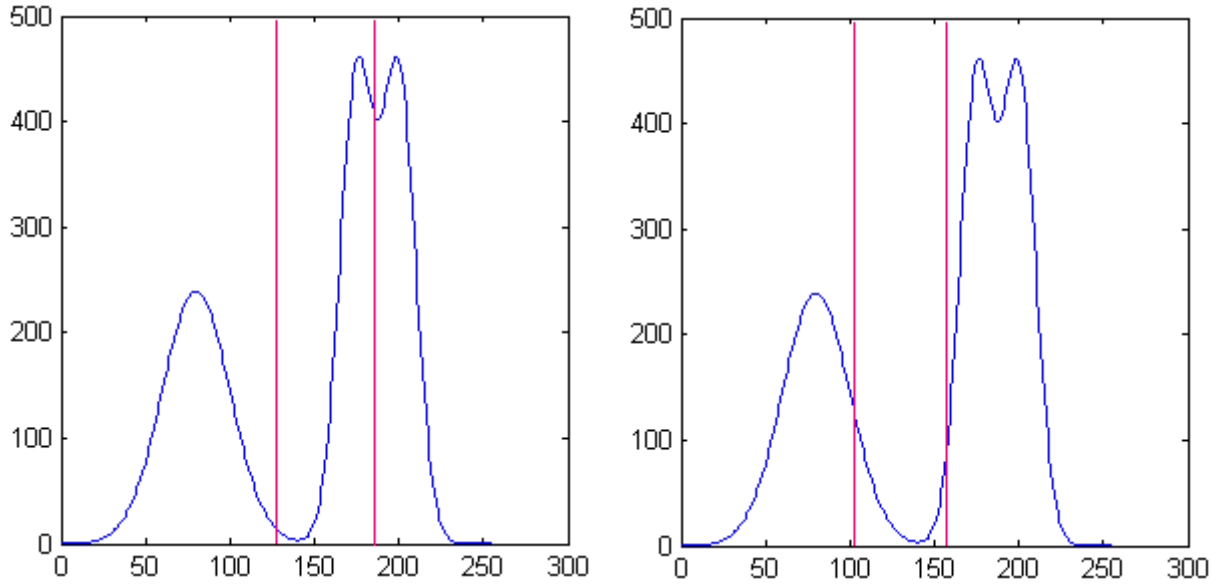
Figure (IV.1)- Histogramme artificiel composé de trois modes (classes).

En fixant le nombre de classes à trois (3), les seuils fournis par l'algorithme génétique ainsi que l'algorithmes itératif sont donnés dans le tableau (IV.2)

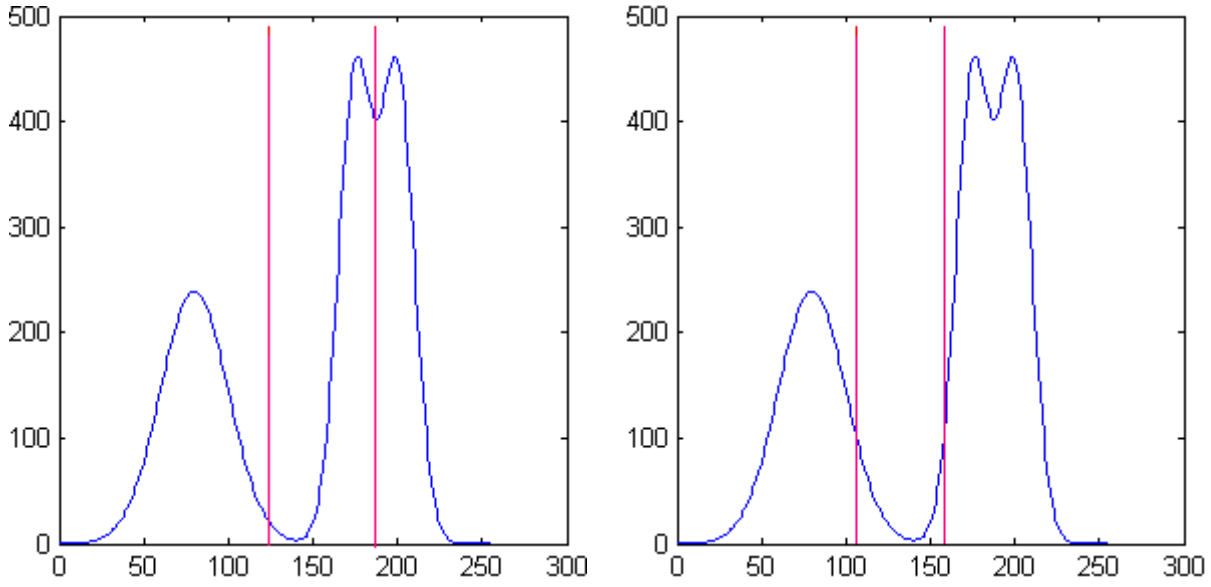
	Algorithme Génétique (AG)	Algorithme Itératif (AI)
Otsu	127-188	126-186
Kapur	106-159	104-159

Tableau (IV.2)- Valeurs des deux seuils obtenus par l'AG et par l'AI

La figure (IV.2) montre la position de ces seuils sur l'histogramme en utilisant respectivement l'algorithme itératif et l'algorithme itératif.



Otsu (AI) Kapur (AI)



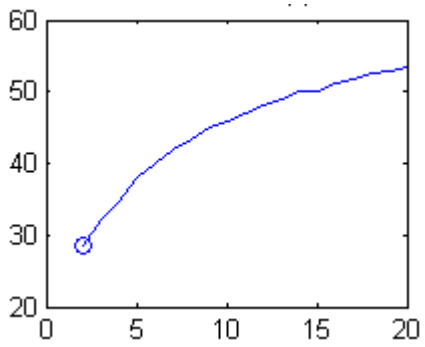
Otsu (AG) Kapur (AG)

La figure (IV.2)- Position des seuils sur l'histogramme de l'exemple 1

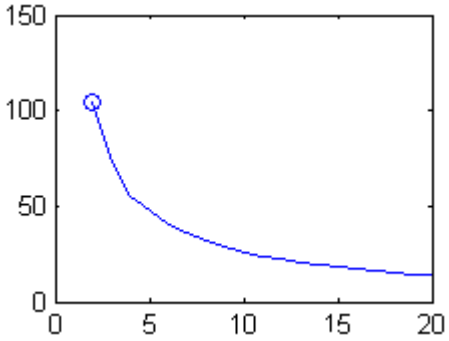
Remarque :

L'algorithmes génétique (AG) donne pratiquement les mêmes résultats que celui de l'algorithme itératif (AI). Cependant l'algorithme itératif consomme peu de temps de calcul par rapport à l'algorithme génétique surtout lorsque le nombre de seuil est élevé.Par conséquent, tous les tests qui suivent seront effectués en utilisant seulement l'algorithme itératif.

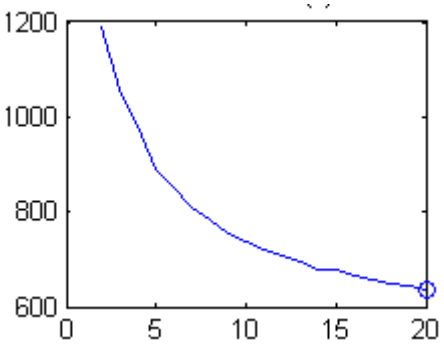
La figure (IV.3) montre la variation des différents indices de validité en fonction du nombre de classes allant de 2 à 20, en utilisant la fonction objective de ''Otsu''. Sur chaque courbe est indiqué le nombre optimal de classes ainsi que la valeur optimale de l'indice correspondant.



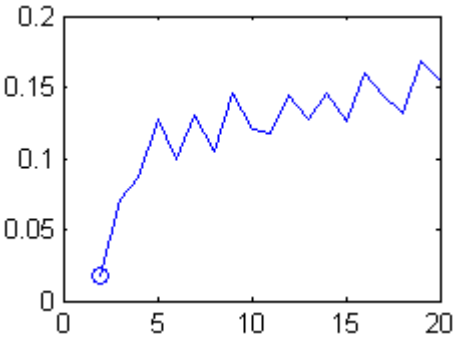
(a)- Indice de FH(k)
FH_{min}=28.6 => k_{opt}=2



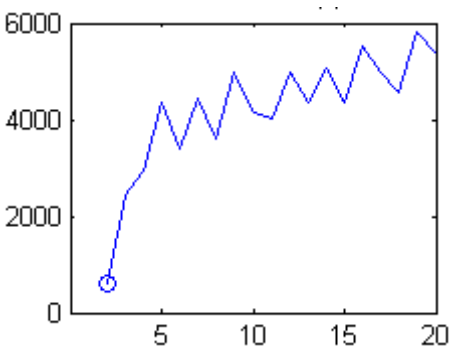
(b)- Indice de PA(k)
PA_{max}=104.7 => k_{opt}=2



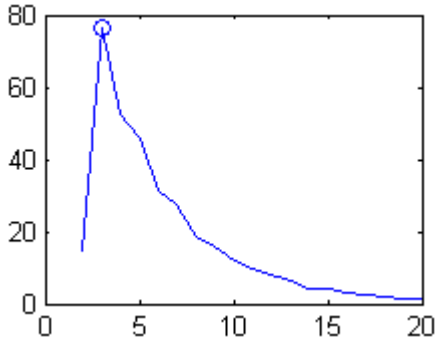
(c)- Indice de DP(k)
DP_{min}=637 => k_{opt}=20



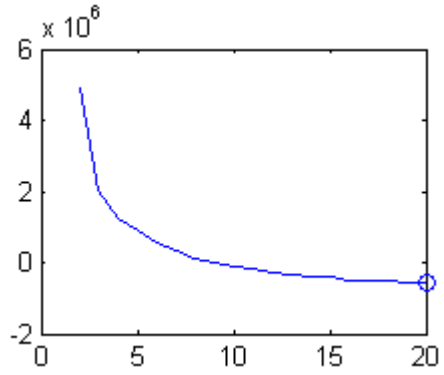
(d)- Indice de Xb(k)
XB_{min}= 0.01801 => k_{opt}=2



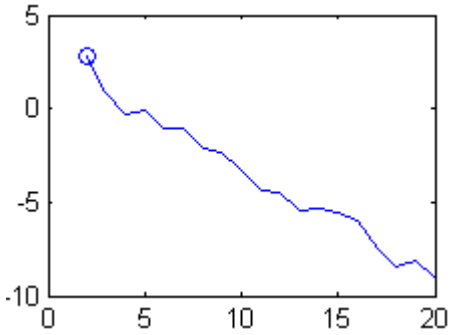
(e)- Indice de Kw(k)
Kw_{min}=612.6 => k_{opt}=2



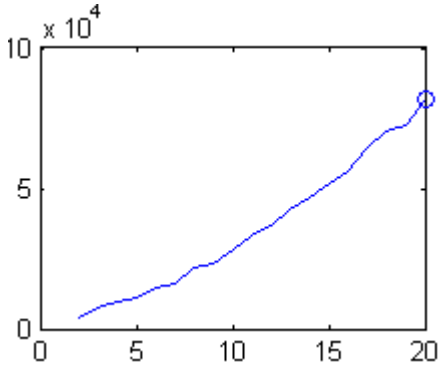
(f)- Indice de I(k)
I_{max}=76.58 => k_{opt}=3



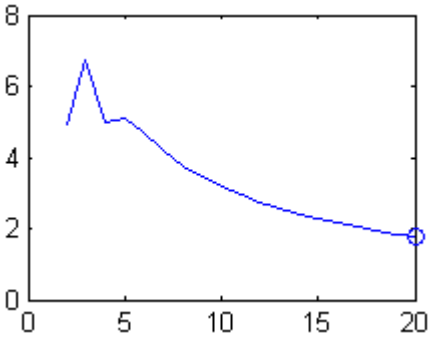
(g)- Indice de FS(k)
FS_{min}= -5.785e+005 => k_{opt}=20



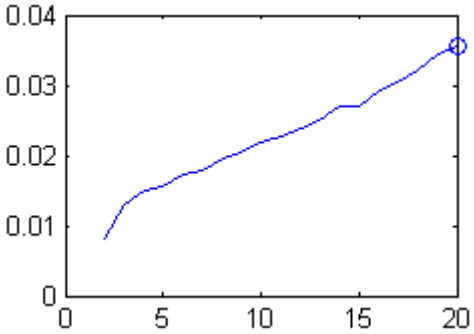
(h)- Indice de PCAES(k)
$PCAES_{max}$=2.783 => k_{opt}=2



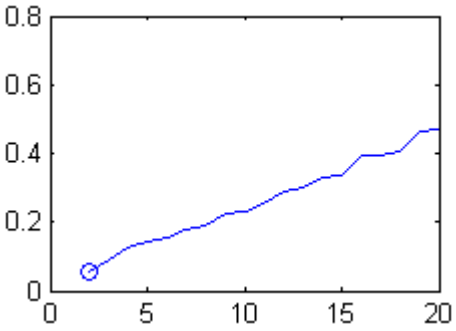
(i)- Indice de CH(k)
CH_{max}= 8.2e+004 => k_{opt}=20



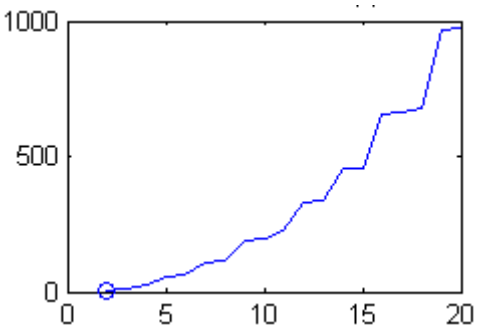
(j)- Indice de DB(k)
DB_{min}=1.748=> k_{opt}=20



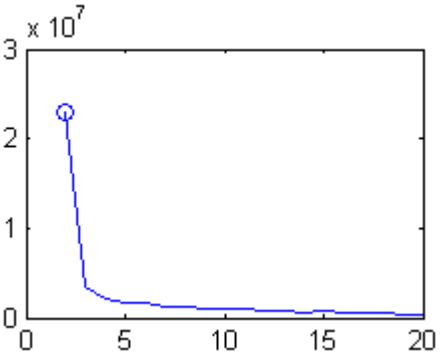
(k)- Indice de D(k)
D_{max}= 0.03571=> k_{opt}=20



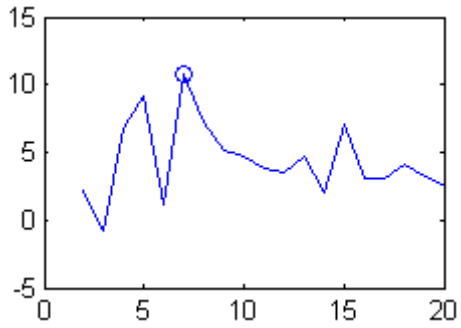
(l)- Indice de R(k)
R_{min}=0.05791 => k_{opt}=2



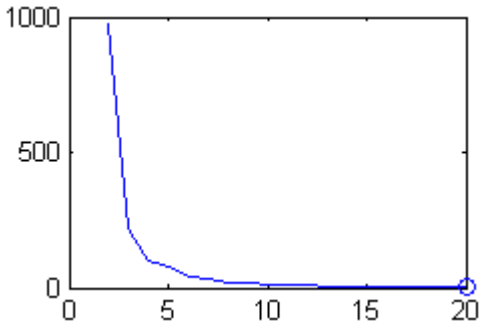
(m)- Indice de B(k)
B_{min}= 1.078=> k_{opt}=2



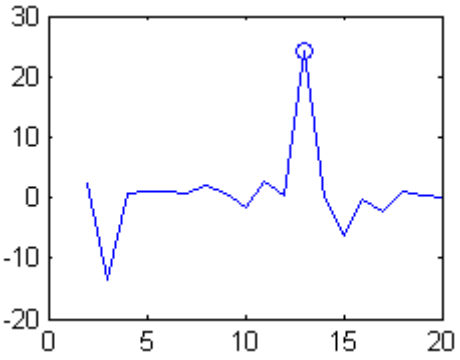
(n)- Indice de Icc(k)
Icc_{max}=2.298e+007=> k_{opt}=2



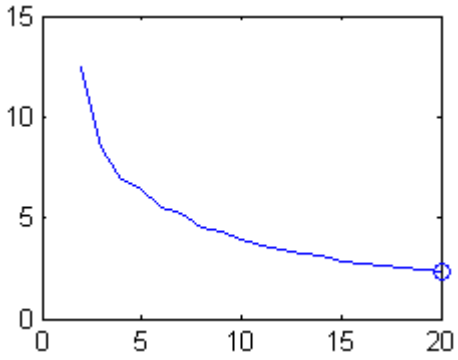
(o)- Indice de V(k)
V_{max}= 10.83=> k_{opt}=7



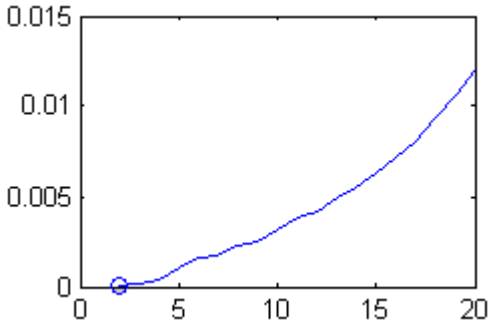
(p)- Indice de VCR(k)
VCR_{min}=1.938 => k_{opt}=20



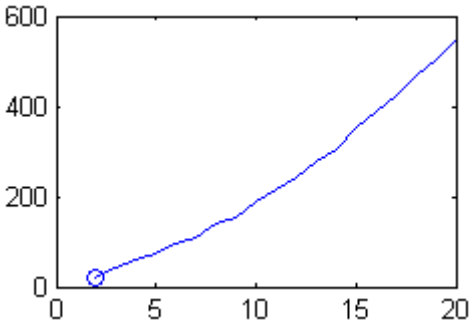
(q)- Indice de KL(k)
KL_{max}= 24.11=> k_{opt}=13



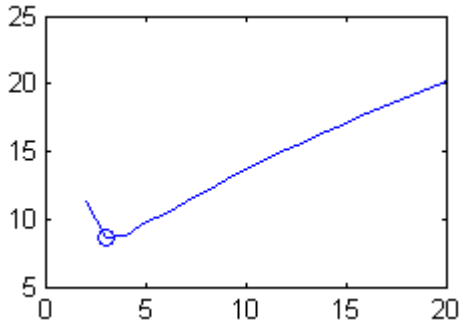
(r)- Indice de RMSSTD(k)
$RMSSTD_{min}$=2.312 => k_{opt}=20



(s)- Indice de CS(k)
CS_{min}=5.675e-005 =>k_{opt}=2



(t)- Indice de J(k)
J_{min}=17.89 => k_{opt}=2



(u)- Indice de F(k)

F_{min}=8.627=>k_{opt}=3

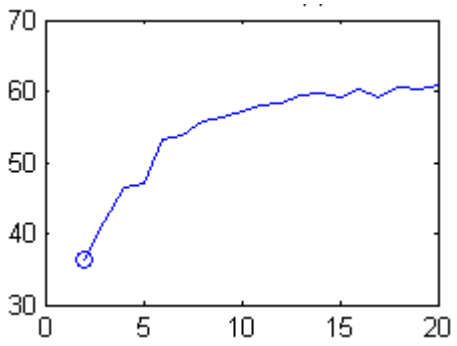
Figure (IV.3)- Indices de validité pour le seuillage d'histogramme de l'exemple (1) avec la fonction objective d' ''Otsu'',

Les indices de PA(k) (figure IV.3.b), de Densité de Partition DP(k) (figure IV.3.c), de Fukuyama Sugeno FS(k) (figure IV.3.g), de Wu et Yan PCAES (k) (figure IV.3.h), de Davies Bouldin DB(k) (figure IV.3.j), de De Franco Icc(k) (figure IV.3.n), de VCR(k) (figure IV.3.p), de RMSSTD(k) (figure IV.3.r) varient d'une manière décroissante, alors que l'indice de Gath et Geva FH(k) (figure IV.3.a), de Calinski Harabasz CH(k) (figure IV.3.i), de Dunn D(k) (figure IV.3.k), de Razae R(k) (figure IV.3.l), de Boudraa B(k) (figure IV.3.m), de Chou et Sun CS(k) (figure IV.3.s), de Deng et al J(k) (figure IV.3.t) varient d'une manière croissante.

Les indices de Xie et Beni Xb(k) (figure IV.3.d), de Kwon Kw(k) (figure IV.3.e) varient d'une manière erratique mais présentent en général une allure croissante.

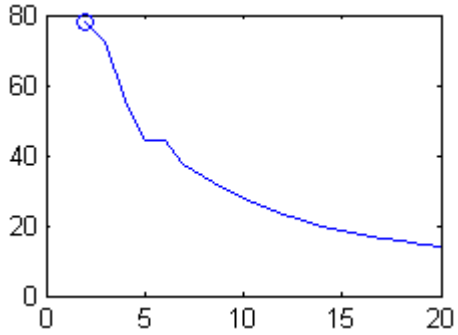
Les indices de Turi V(k) (figure IV.3.o), de Karzanowiski et lai KL(k) (figure IV.3.q), varient également d'une manière erratique. Ces deux indices ont tendance à surestimer le nombre exacte de classes étant donné que le nombre de classes optimal fournit par V(k) est égal à 7 alors que celui obtenu par KL(k) est égal à 13. Tous ces indices cités précédemment ne permettent pas de déterminer le nombre optimal de classes. Cependant, seuls les indices de I(k), de F(k) présentent un seuil optimum. Ces deux indices donne le nombre optimal de classes égale au nombre réel (3) de classes

La figure (IV.4) montre la variation des indices de validité en fonction du nombre de classes variant de 2 à 20, en utilisant la fonction objective de ''Kapur ''



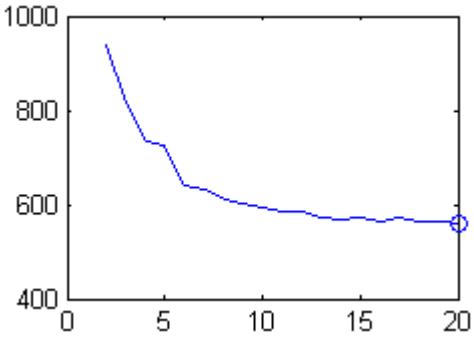
(a)- Indice de FH(k)

FH_{min}=36.23 => k_{opt}=2



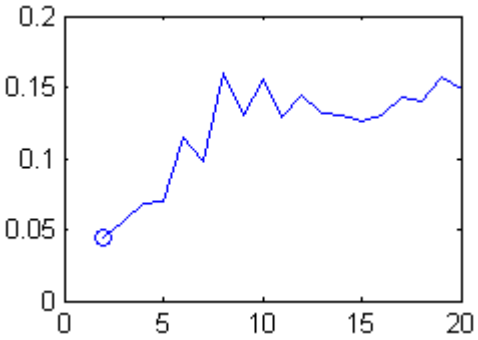
(b)- Indice de PA(k)

PA_{max}=77.92 => k_{opt}=2



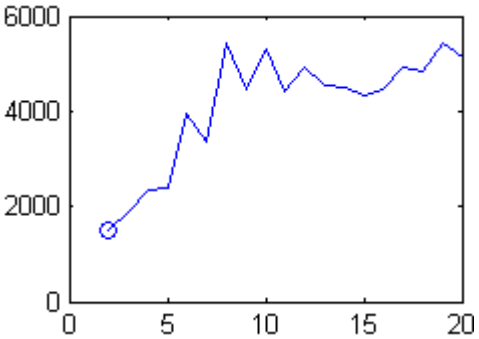
(c)- Indice de DP(k)

DP_{min}=558.6 => k_{opt}=20



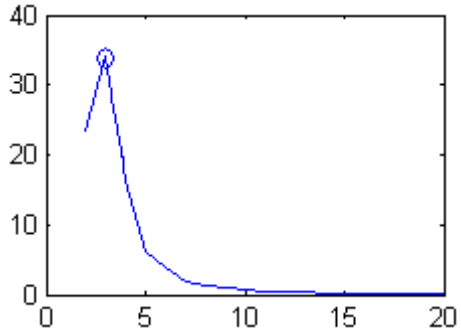
(d)- Indice de Xb(k)

XB_{min}= 0.0444 => k_{opt}=2



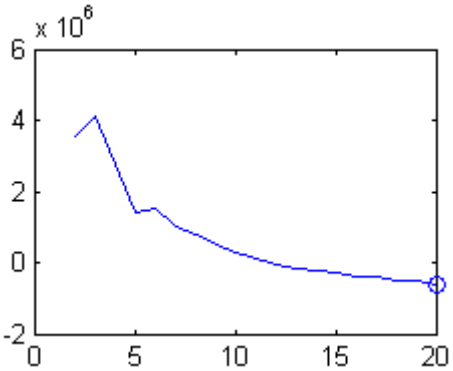
(e)- Indice de Kw(k)

Kw_{min}=1511 => k_{opt}=2



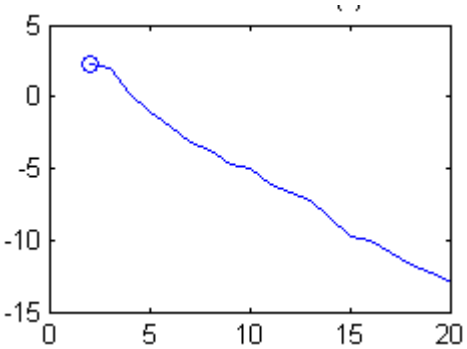
(f)- Indice de I(k)

I_{max}=34.23 => k_{opt}=3



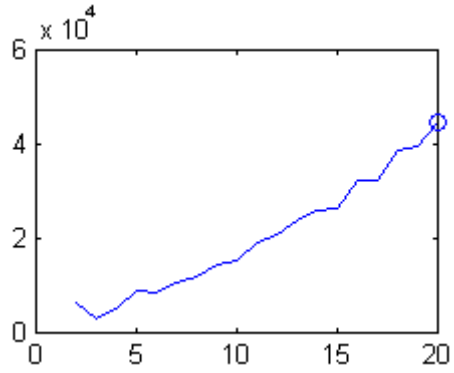
(g)- Indice de FS(k)

FS_{min}= -5.927e+005 => k_{opt}=20



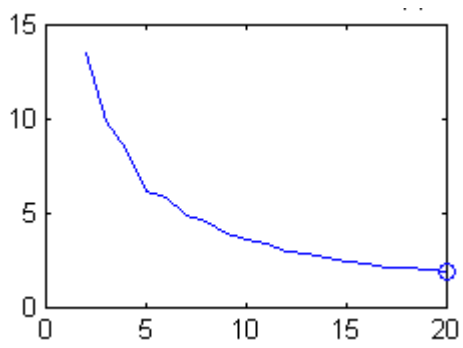
(h)- Indice de PCAES(k)

$PCAES_{max}$=2.311 => k_{opt}=2



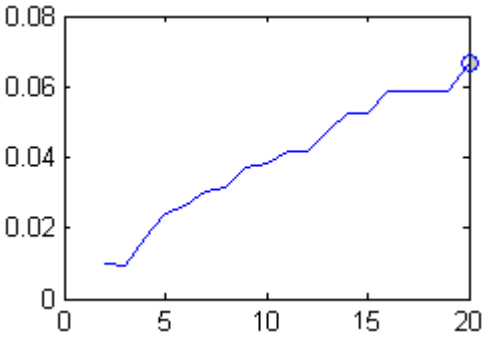
(i)- Indice de CH(k)

CH_{max}= 4.429e+004 => k_{opt}=20



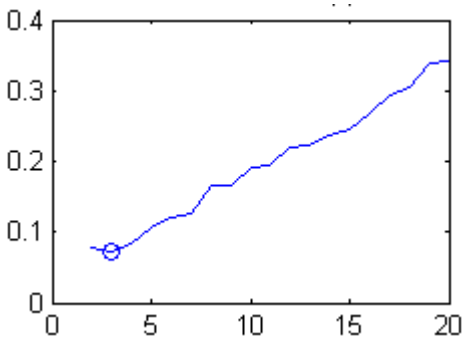
(j)- Indice de DB(k)

DB_{min}=1.83=> k_{opt}=20



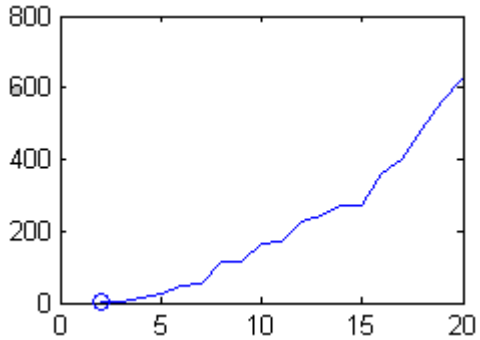
(k)- Indice de D(k)

D_{max}= 0.06667=> k_{opt}=20



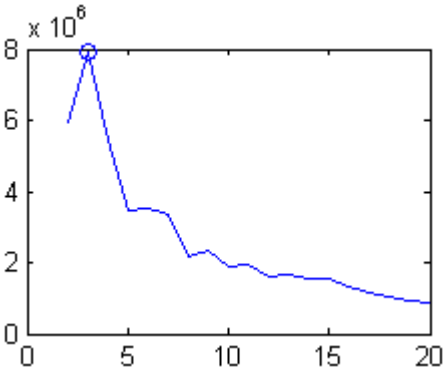
(l)- Indice de R(k)

R_{min}=0.07252 => k_{opt}=3



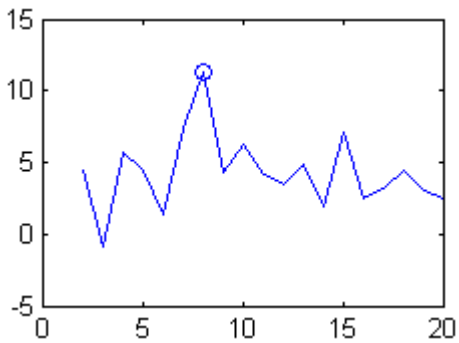
(m)- Indice de B(k)

B_{min}= 1.112=> k_{opt}=2



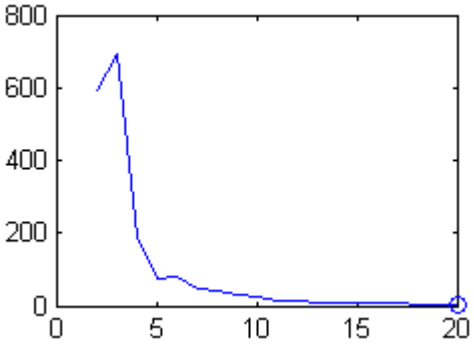
(n)- Indice de Icc(k)

Icc_{max}=7.896e+006=> k_{opt}=3



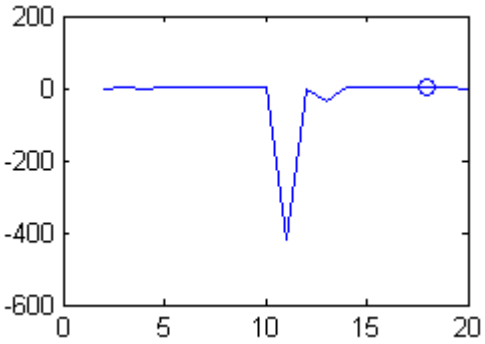
(o)- Indice de V(k)

V_{max}= 11.26=> k_{opt}=8

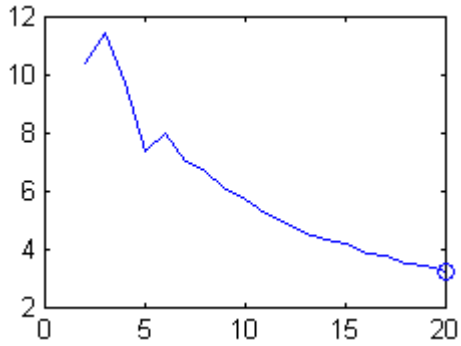


(p)- Indice de VCR(k)

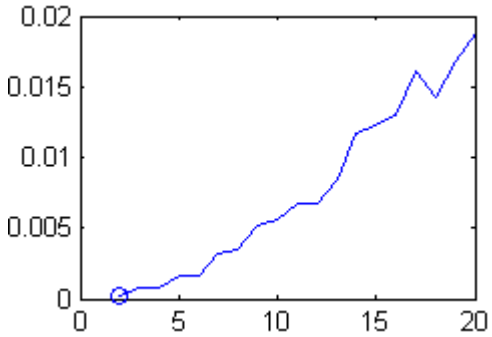
VCR_{min}=3.213=> k_{opt}=20



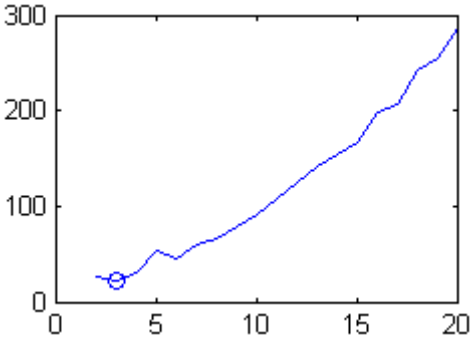
(q)- Indice de KL(k)
KL_{max}= 4.352 => k_{opt}=18



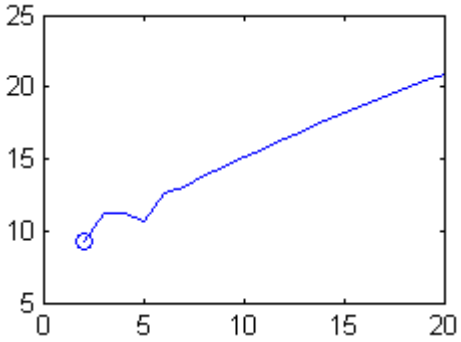
(r)- Indice de RMSSTD(k)
$RMSSTD_{min}$=3.2 =>k_{opt}=20



(s)- Indice de CS(k)
CS_{min}=0.0001552 =>k_{opt}=2



(t)- Indice de J(k)
J_{min}=21.6 =>k_{opt}=3



(u)- Indice de F(k)

F_{min}=9.313 =>k_{opt}=2

Figure (IV.4)- Indices de validité pour le seuillage d'histogramme de l'exemple 1 avec la fonction objective de ''Kapur'', de l'exemple (1)

Les indices de PA(k) (figure IV.4.b), de Densité de Partition DP(k) (figure IV.4.c), de Fukuyama Sugeno FS(k) (figure IV.4.g), de Wu et Yan PCAES (k) (figure IV.4.h), de Davies Bouldin DB(k) (figure IV.4.j), de VCR(k) (figure IV.4.p), de RMSSTD(k) (figure IV.4.r) varient d'une manière décroissante, alors que l'indice de Gath et Geva FH(k) (figure IV.4.a), de Calinski Harabasz CH(k) (figure IV.4.i), de Dunn D(k) (figure IV.4.k), de Boudraa B(k) (figure IV.4.m), de Chou et Sun CS(k) (figure IV.4.s), de Yen et Chang F(k) (figure IV.4.u) varient d'une manière croissante.

Les indices de Xie et Beni Xb(k) (figure IV.4.d), de Kwon Kw(k) (figure IV.4.e) varient d'une manière erratique mais présentent en général une allure croissante.

Les indices de Turi V(k) (figure IV.4.o), de Karzanowiski et lai KL(k) (figure IV.4.q) varient également d'une manière erratique. Ces deux indices ont tendance à surestimer le nombre exacte de classes car le nombre de classes optimal fournit par V(k) est égal à 8 alors que celui obtenu par KL(k) est égal à 18. Tous ces indices cités précédemment ne permettent pas de déterminer le nombre optimal de classes. Cependant, seuls les indices de I(k), de R(k), de J(k), Icc(k) présentent un seuil optimum. Ces trois indices donnent le nombre optimal de classes égales au nombre réel de classes (3).

Pour ce premier exemple, l'indice commun aux deux fonctions objectives (Otsu et Kapur) qui donne le nombre optimal de classes est l'indice de Maulik I(k).

Exemple 2

Dans ce deuxième exemple l'histogramme généré est composé de cinq distributions gaussiennes dont les paramètres statistiques seront présentés dans le tableau (IV.3). La figure (IV.5) montre l'allure de cet histogramme.

m_1	m_2	m_3	m_4	m_5	P_1	P_2	P_3	P_4	P_5	σ_1	σ_2	σ_3	σ_4	σ_5
20	70	115	160	237	12000	12500	13000	12000	15000	20	15	17	11	18

Tableau (IV.3)-Paramètres statistiques réels de l'histogramme

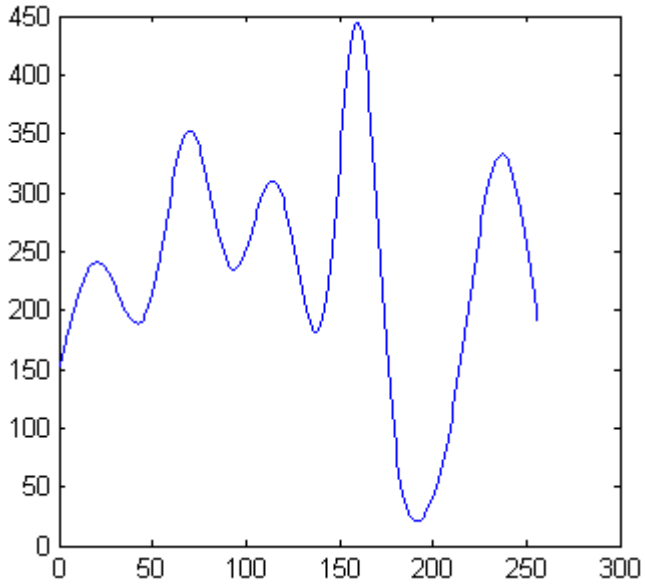


Figure (IV.5)- Histogramme artificiel composé de cinq modes(classes).

En fixant le nombre de classes à cinq (5), les seuils fournis par l'algorithme itératif sont donnés dans le tableau (IV.4)

	Algorithme itératif
Otsu	147-91-135-195
Kapur	49-95-142-205

Tableau (IV.4)- Valeurs des seuils obtenus par l'algorithme itératif

La figure (IV.6) montre la position de ces seuils sur l'histogramme

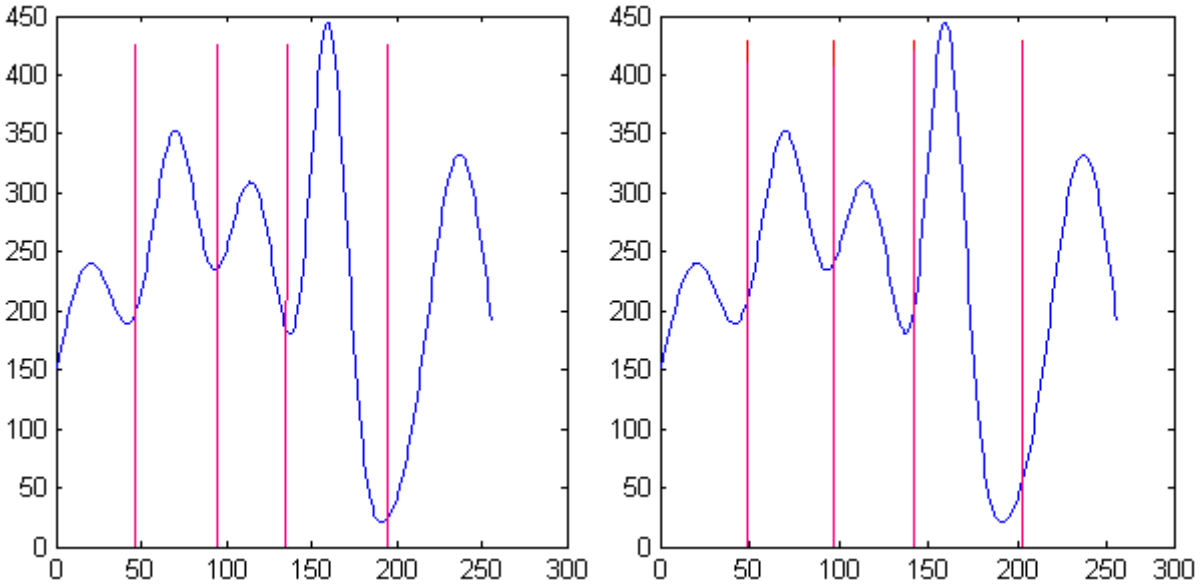
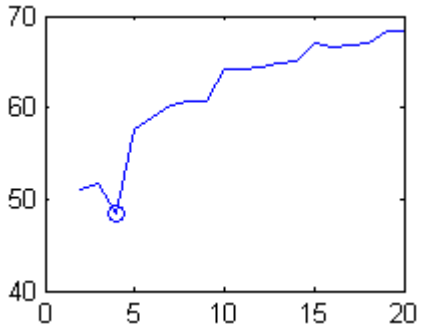
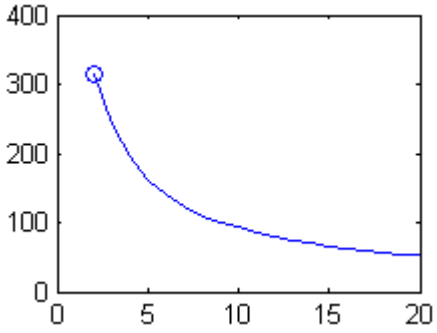


Figure (IV.6)- Position des seuils sur l'histogramme de l'exemple (2)

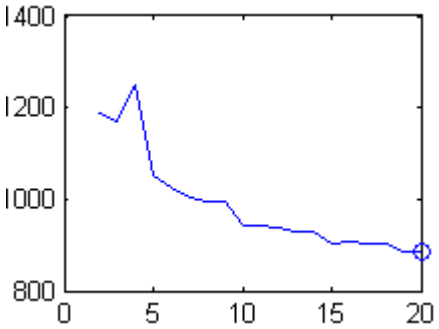
La figure (IV.7) illustre les variations des indices de validité en fonction du nombre de classes obtenues en optimisant la fonction d'Otsu.



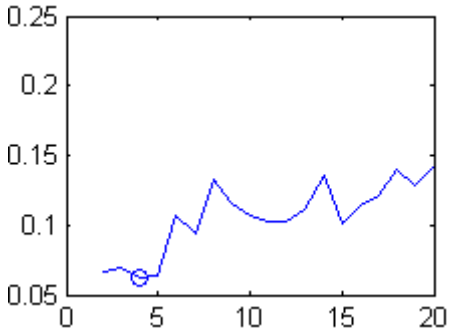
(a)- Indice de FH(k)
FH_{min}=48.48=> k_{opt}=4



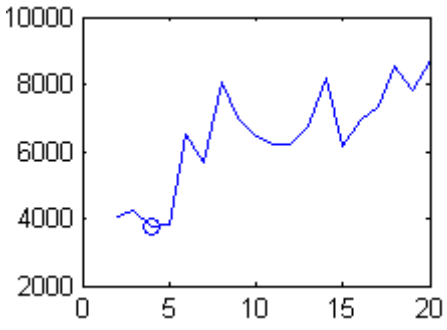
(b)- Indice de PA(k)
PA_{max}=314.7 => k_{opt}=2



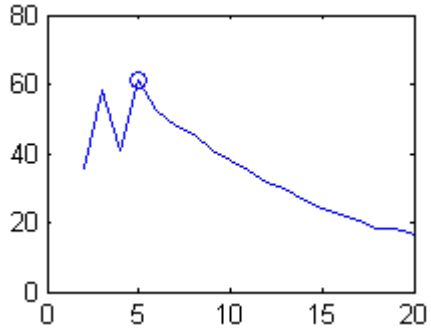
(c)- Indice de DP(k)
FH_{min}=883.7 => k_{opt}=20



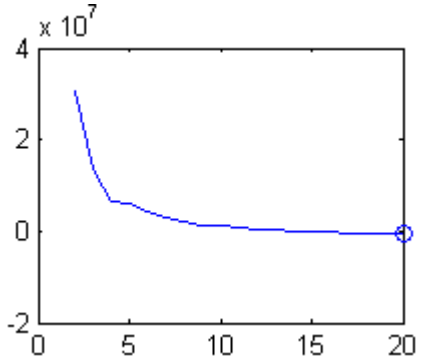
(d)- Indice de Xb(k)
XB_{min}=0.06196 => k_{opt}=4



(e)- Indice de Kw(k)
KW_{min}=3748 => k_{opt}=4

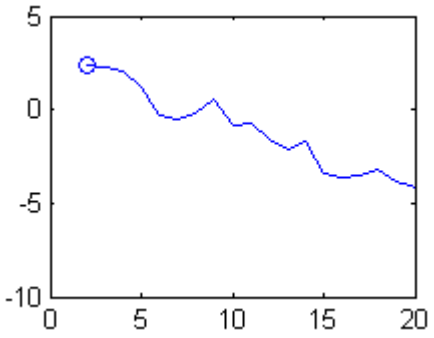


(f)- Indice de I(k)
IK_{max}=61.41 => k_{opt}=5



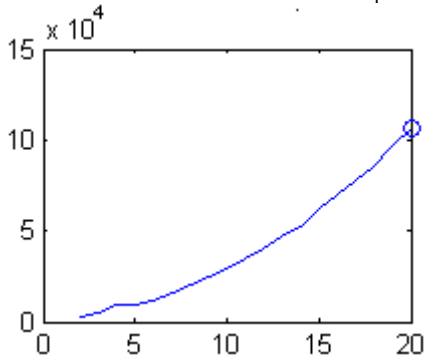
(g)- Indice de FS(k)

FS_{min}= -5.526e+005 => k_{opt}=20



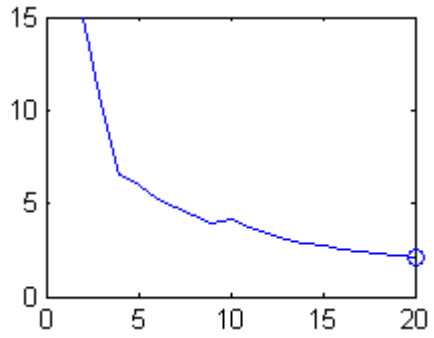
(h)- Indice de PCAES(k)

$PCAES_{max}$=2.382 => k_{opt}=2



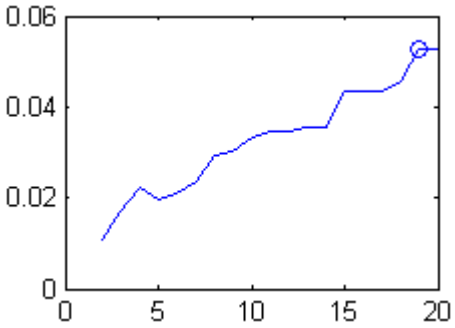
(i)- Indice de CH(k)

CH_{max}=1.061e+005 => k_{opt}=20



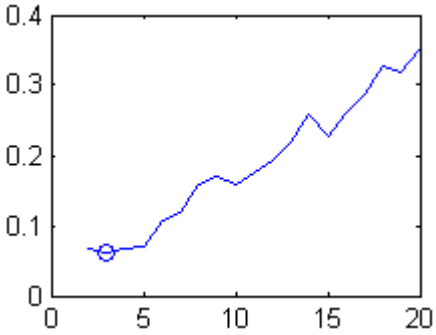
(j)- Indice de DB(k)

DB_{min}=2.089 => k_{opt}=20



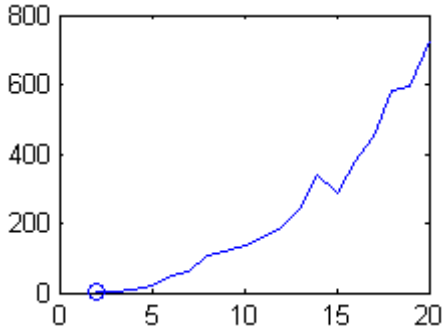
(k)- Indice de Dun(k)

Dun_{max}=0.05263 => k_{opt}=19



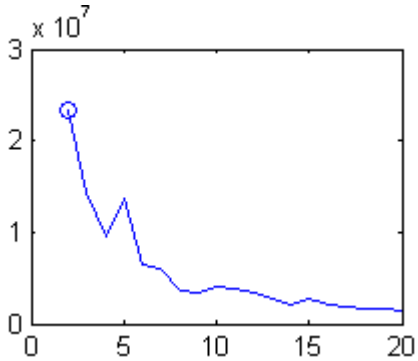
(l)-Indice de R(k)

R_{min}=0.06045 => k_{opt}=3



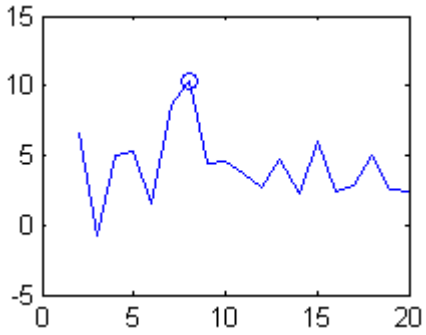
(m)- Indice de B(k)

$Boudraa_{min}$=1.125 => k_{opt}=2

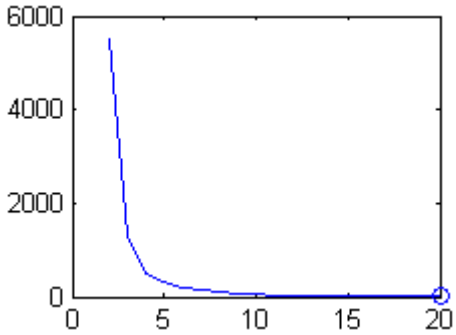


(n)- Indice de Icc(k)

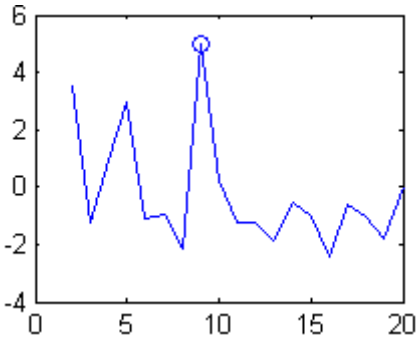
ICC_{max}=2.332e007 => k_{opt}=2



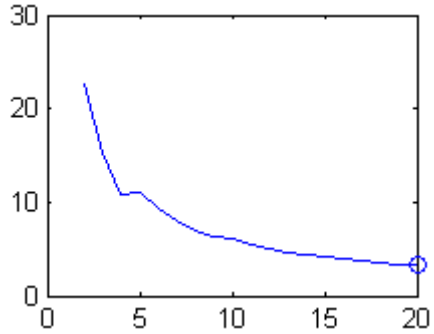
(o)- Indice de V(k)
$Turi_{max}$=10.33 => k_{opt}=8



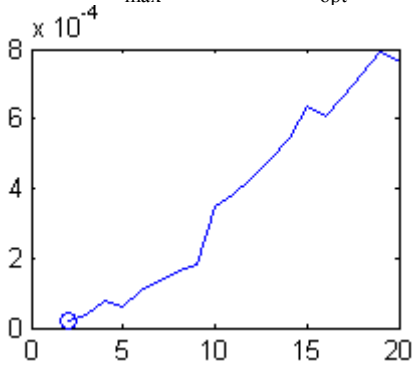
(p)- Indice de VCR(k)
VCR_{min}=6.407 => k_{opt}=20



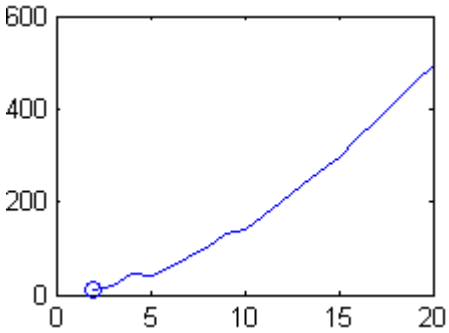
(q)- Indice de KL(k)
Dun_{max}=4.963 => k_{opt}=9



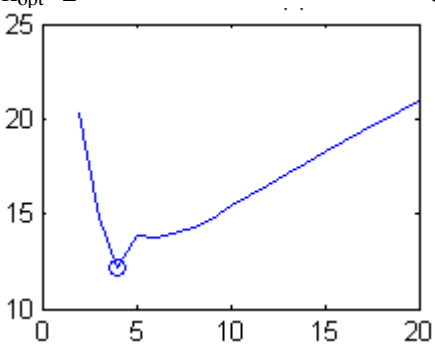
(r)- Indice de RMSSTD(k)
R_{min}=3.233 => k_{opt}=20



(s)- Indice de CS(k)
CS_{min}=1.852e-005 => k_{opt}=2



(t)- Indice de J(k)
J_{min}=9.197 => k_{opt}=2

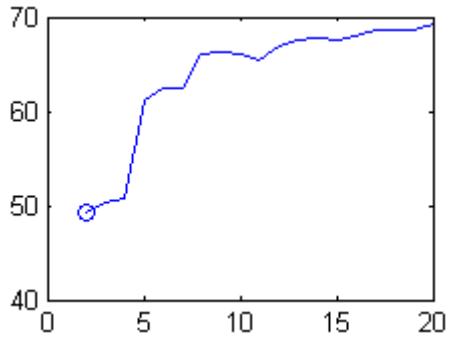


(u)- Indice de F(k)
F_{min}=12.17 => k_{opt}=4

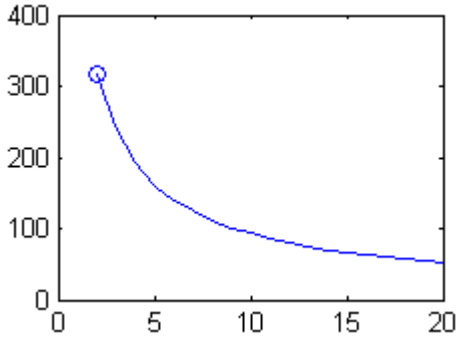
Figure (IV.7) Indices de validité pour le seuillage d'histogramme utilisant ''Otsu '' comme fonction objective.

Pour cet exemple, seul l'indice de Maulik a permis d'obtenir le nombre de classes réel puisque sa valeur maximale est atteinte pour k=5. Tous les autres indices varient d'une manière monotone ou erratique.

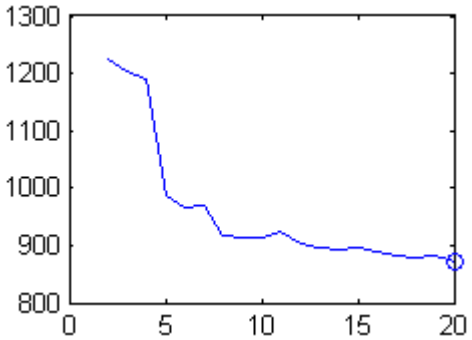
La figure (IV.8) montre l'allure des indices de validité lorsque la fonction de ''Kapur'' est utilisée comme fonction objective.



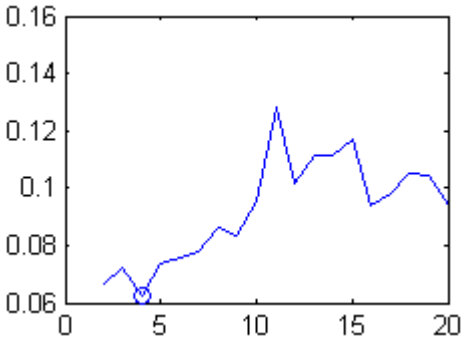
(a)- Indice de FH(k)
FH_{min}=49.3=> k_{opt}=2



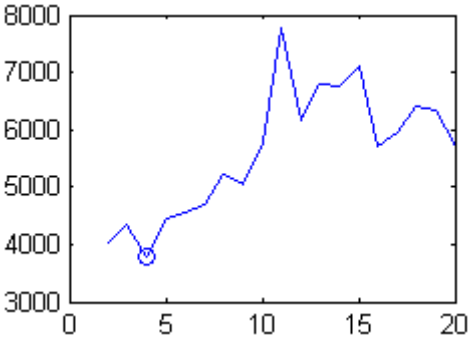
(b)- Indice de PA(k)
PA_{max}=315.7 => k_{opt}=2



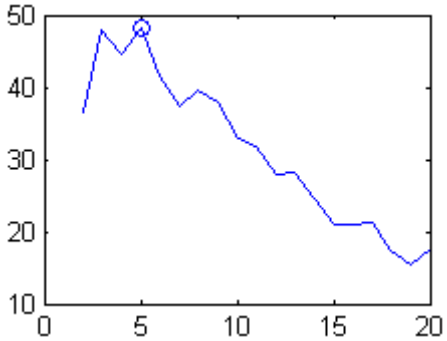
(c)-Indice de DP (k)
DP_{min}=873=> k_{opt}=20



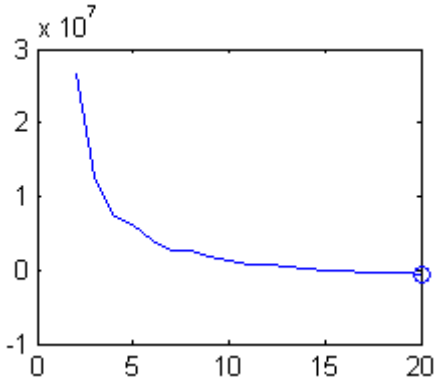
(d)-Indice de XB(k)
XB_{min}=0.06267 => k_{opt}=4



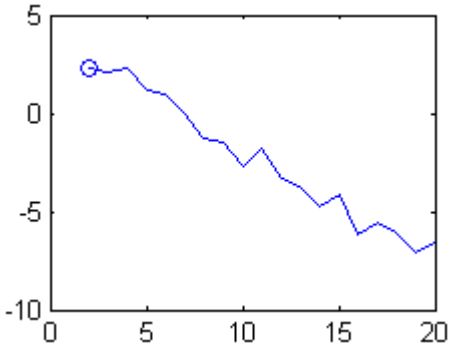
(e)- Indice de Kw(k)
KW_{min}=3791 => k_{opt}=4



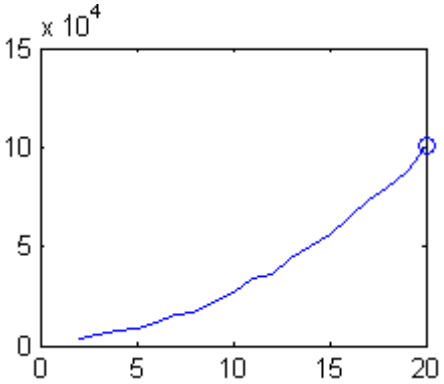
(f)- Indice de I(k)
I_{max}=48.25 => k_{opt}=5



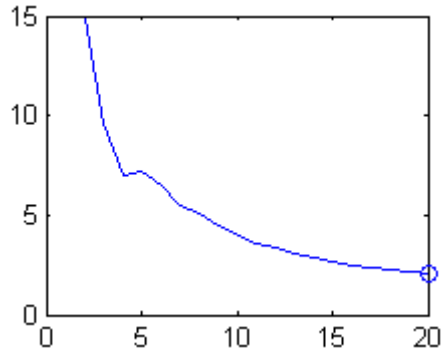
(g)- Indice de FS(k)
FS_{min}= -5.094e+005=> k_{opt}=20



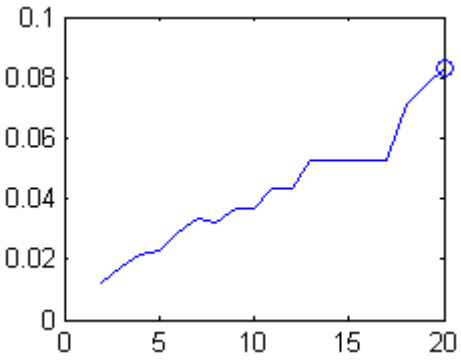
(h)- Indice de PCAES(k)
$PCAES_{max}$=2.298=> k_{opt}=2



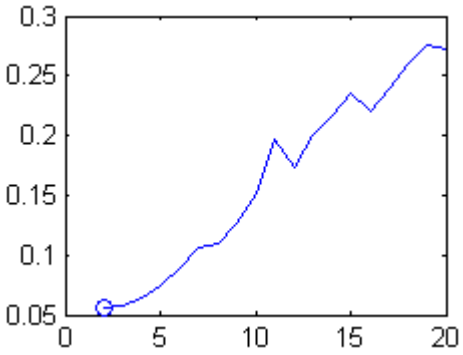
(i)- Indice de CH(k)
CH_{max}=1.005e+005 => k_{opt}=20



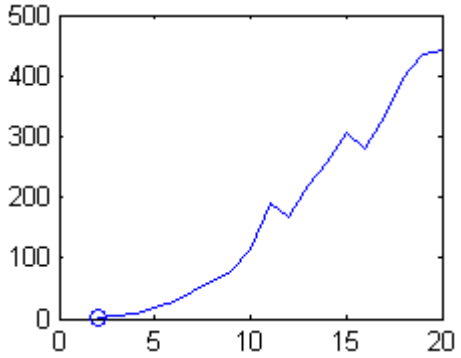
(j)- Indice de DB(k)
DB_{min}=2.024 => k_{opt}=20



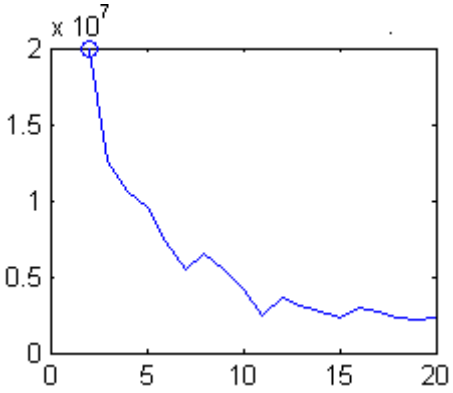
(k)- Indice de D(k)
Dun_{max}=0.08333 => k_{opt}=20



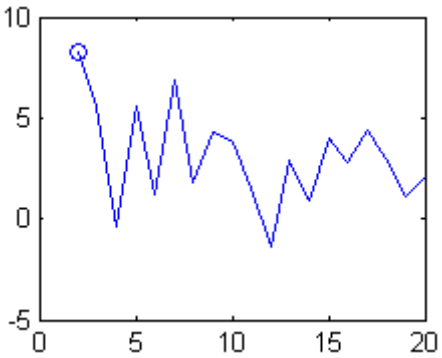
(l)- Indice de R(k)
R_{min}=0.05634 => k_{opt}=2



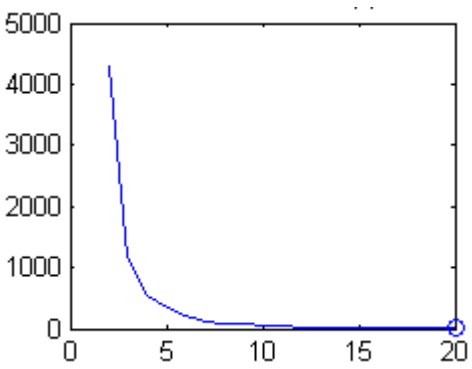
(m)- Indice de B(k)
$Boudraa_{min}$=1.117 => k_{opt}=2



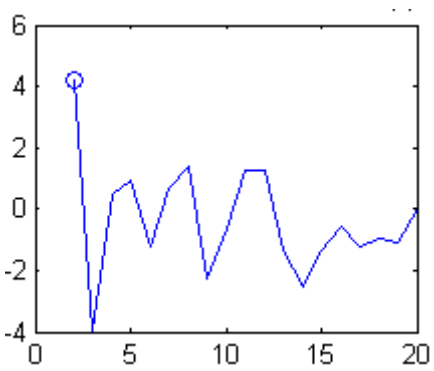
(n)- Indice de Icc(k)
ICC_{max}=1.993e+007 => k_{opt}=2



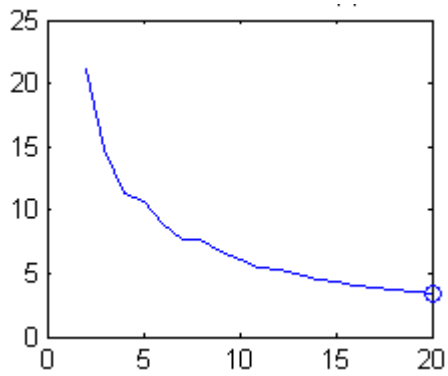
(o)- Indice de V(k)
$Turi_{max}$=8.334 => k_{opt}=2



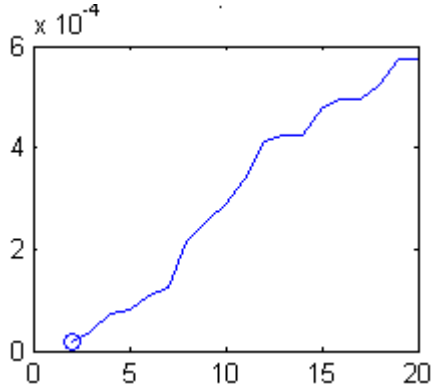
(p)- Indice de VCR(k)
VCR_{min}=6.951=> k_{opt}=20



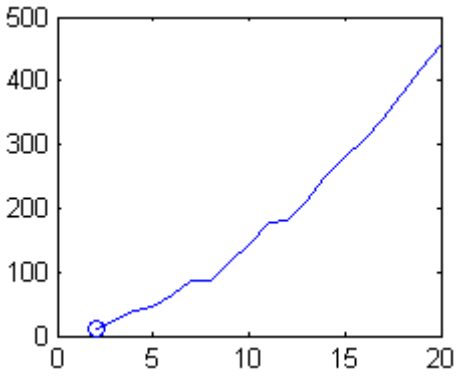
(q)- Indice de KL(k)
Dun_{max}=4.234 => k_{opt}=2



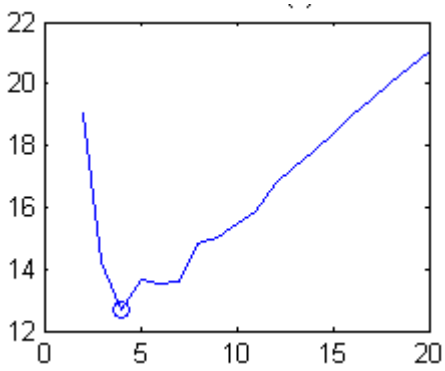
(r)- Indice de RMSSTD(k)
R_{min}=3.364 => k_{opt}=20



(s)- Indice de CS(k)
CS_{min}=1.84e-005 => k_{opt}=2



(t)- Indice de J(k)
J_{min}=10.64 => k_{opt}=2



(u)- Indice de F(k)
F_{min}=12.69 => k_{opt}=4

Figure (IV.8) Indices de validité pour le seuillage d'histogramme utilisant ''Kapur'' comme fonction objective

Comme pour la fonction d' Otsu, seul l'indice de Maulik fournit le nombre exact de classes, par conséquent, pour ce deuxième exemple, seul cet indice est plus performant.

IV.3- Images réelles

Les deux exemples précédents nous ont permis de juger objectivement l'efficacité des indices de validité. A présent, nous allons voir comment varient ces indices en présence d'histogramme issus d'images réels.

Image 1

Cette image ''Airplane'', de taille 256*256, codée sur 8 bits, est représentée sur la figure (IV.9). Son histogramme est illustré sur la figure (IV.10).

Il est difficile d'observer le nombre exact de modes sur cet histogramme. Cependant, on peut distinguer principalement trois modes.

Figure (IV.9) : Image ''Airplane''

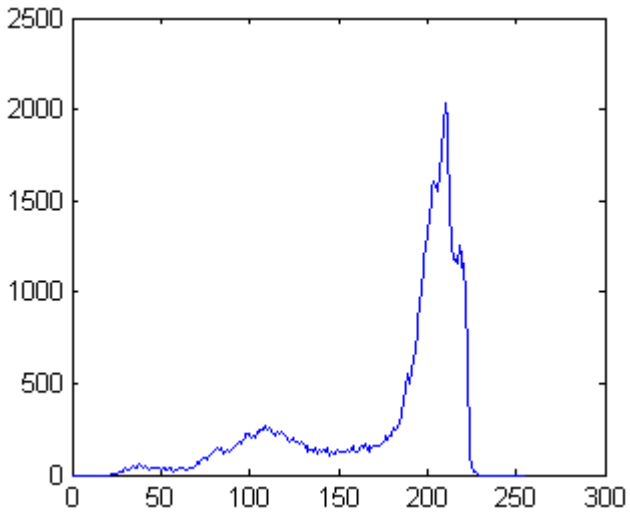


Figure (IV.10) : histogramme de l'image ''Airplane''

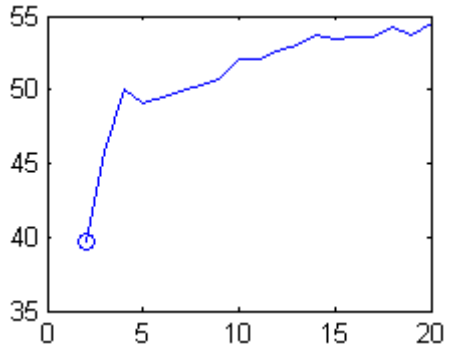
Le tableau (IV.5) donne es valeurs des seuils obtenus par l'algorithme itératif et par la recherche exhaustive lorsque le nombre de seuils varient de 2 à 4

		Nombre de seuils		
		2	3	4
Otsu	Algorithme itératif	115-173	92-141-187	86-129-171-202
	Recherche exhaustive	118-176	95-147-192	89-133-175-204
Kapur	Algorithme itératif	91-160	74-126-178	63-104-145-186
	Recherche exhaustive	77-175	73-129-184	70-107-145-185

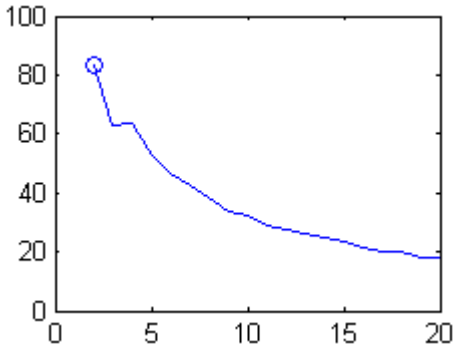
Tableau (IV.5)- Valeurs des seuils obtenues par l'algorithme itératif et par la recherche exhaustive.

D'après le tableau (IV.5), on constate que les seuils optimaux obtenus on utilisant la méthode itérative et la fonction objective d' Otsu et Kapur sont très proches des seuils trouvés avec la recherche exhaustive

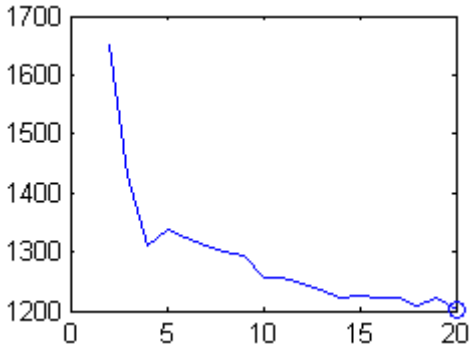
La figure (IV.11) montre la variation des indices de validité en fonction du nombre de classes allant de 2 à 20, en utilisant la fonction objective de ''Otsu''.



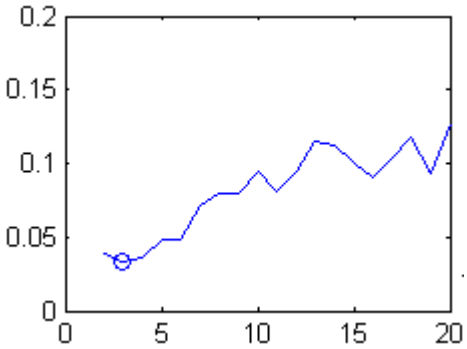
(a)- Indice de FH(k)
FH_{min}=39.68 => k_{opt}=2



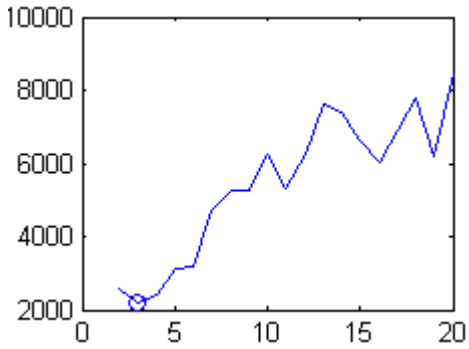
(b)- Indice de PA(k)
PA_{max}=83.33 => k_{opt}=2



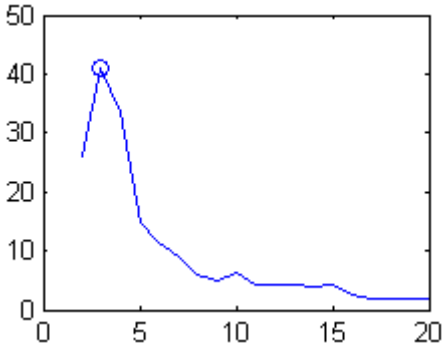
(c)- Indice de DP(k)
DP_{min}=1202 => k_{opt}=20



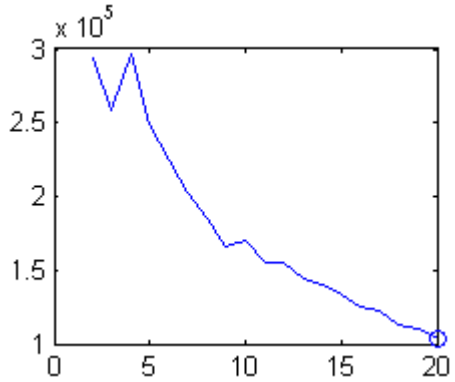
(d)- Indice de Xb(k)
XB_{min}= 0.03352 => k_{opt}=3



(e)- Indice de Kw(k)
Kw_{min}= 2200 => k_{opt}=3

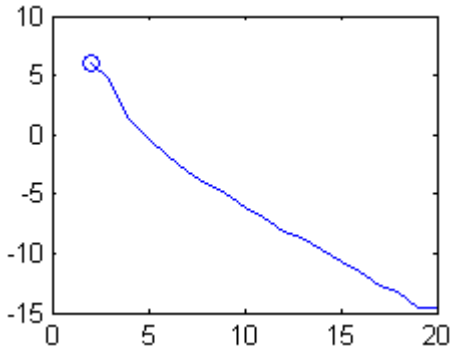


(f)- Indice de I(k)
I_{max}= 41.13 => k_{opt}=3



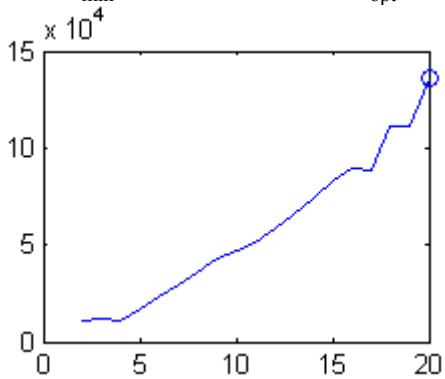
(g)- Indice de FS(k)

FS_{min}= 1.037e+005 => k_{opt}=20



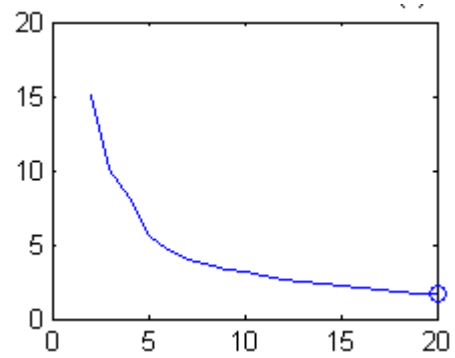
(h)- Indice de PCAES(k)

$PCAES_{max}$=6.074 => k_{opt}=2



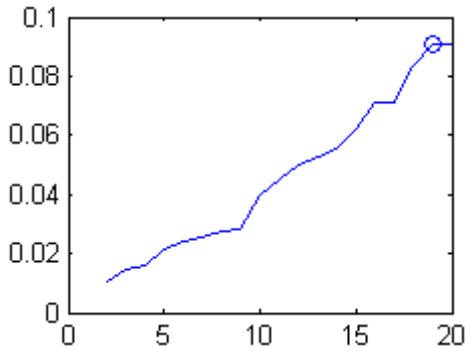
(i)- Indice de CH(k)

CH_{max}= 1.357e+005 => k_{opt}=20



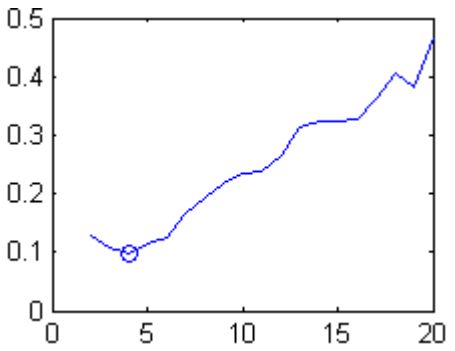
(j)- Indice de DB(k)

DB_{min}=1.643=> k_{opt}=20



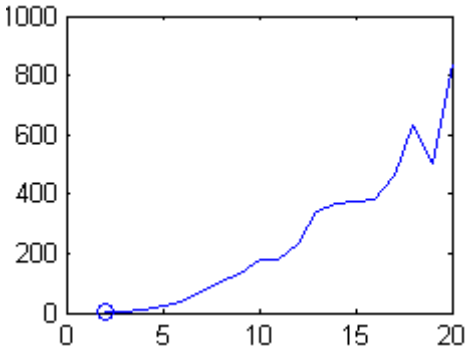
(k)- Indice de D(k)

D_{max}= 0.09091=> k_{opt}=19



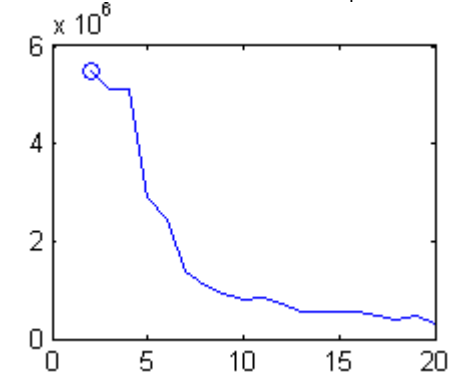
(l)- Indice de R(k)

R_{min}=0.09927 => k_{opt}= 4



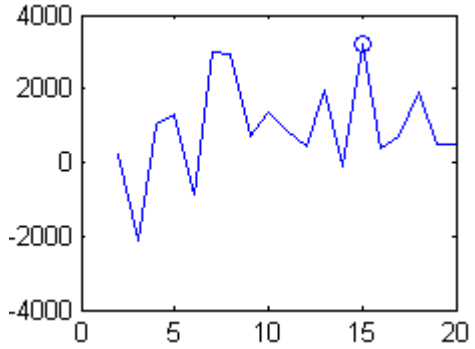
(m)- Indice de B(k)

B_{min}= 1.196=> k_{opt}=2



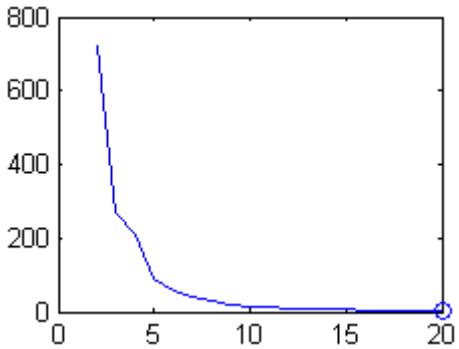
(n)- Indice de Icc(k)

Icc_{max}=5.495e+006=> k_{opt}=2



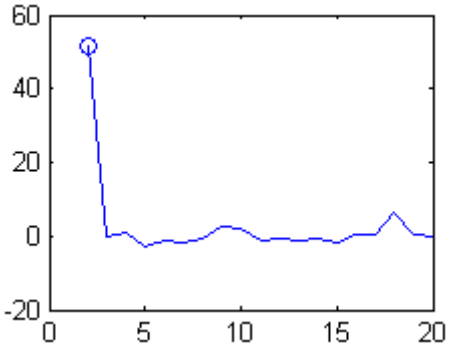
(o)- Indice de V(k)

V_{max}= 3224=> k_{opt}=15



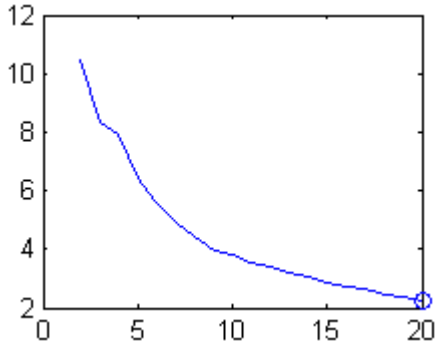
(p)- Indice de VCR(k)

VCR_{min}=2.413=> k_{opt}=20



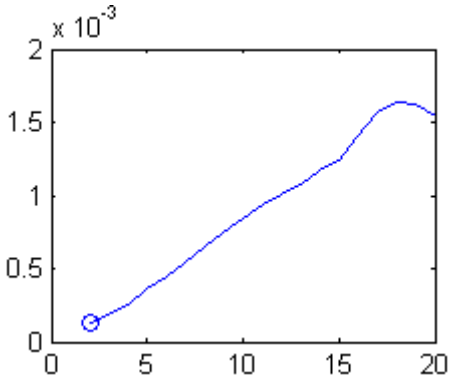
(q)- Indice de KL(k)

KL_{max}= 51.72=> k_{opt}=2



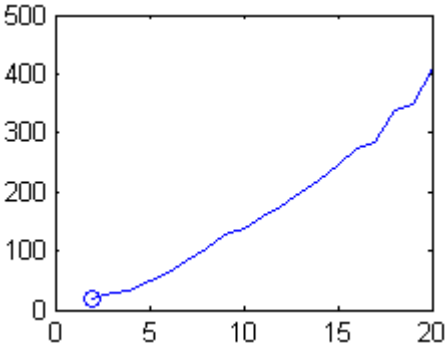
(r)- Indice de RMSSTD(k)

$RMSSTD_{min}$=2.227 =>k_{opt}=20



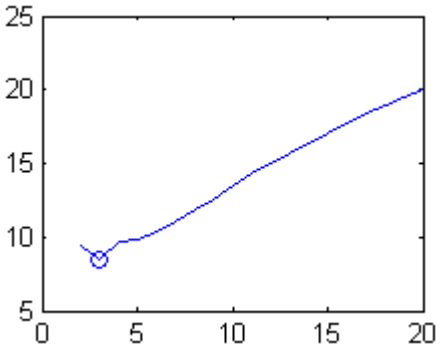
(s)- Indice de CS(k)

CS_{min}=0.0001259 =>k_{opt}=2



(t)- Indice de J(k)

J_{min}=17.46=>k_{opt}=2



(u)- Indice de F(k)

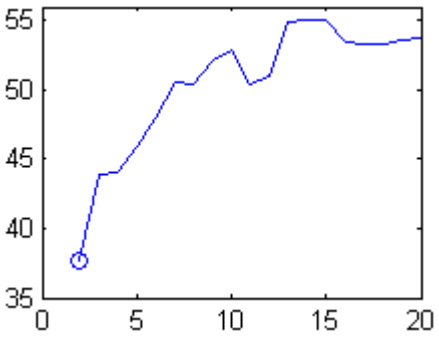
F_{min}=8.502 =>k_{opt}=3

Figure (IV.11)- Indices de validité pour le seuillage d'histogramme de l'image ''Airplane'' on utilisant la fonction objective d' 'Otsu''.

Vu la forme de l'histogramme de l'image ''Airplane'', qui présente approximativement trois modes, seuls les indices de Maulik I(k), Xb(k), Kw(k) et F(k), donnent un nombre de classes égal à trois (3).

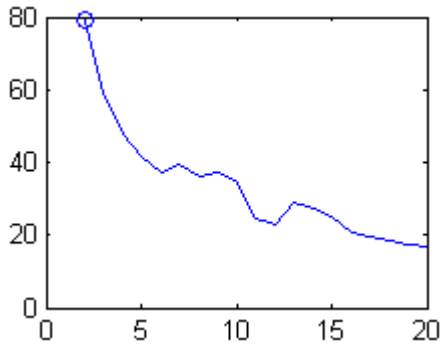
La plus part des indices varient d'une manière monotone ce qui ne permet pas d'obtenir le nombre exact de classes. Par contre l'indice R (k) donne un nombre de classes proches de trois (3) tandis que les indices D(k) et V(k) surestiment le nombre plausible de classes étant donné que le nombre de classes optimale fournit par D(k) est égale à 19 alors que celui obtenu par V(k) est égale à 15.

La figure (IV.12) montre la variation des indices de validité en fonction du nombre de classes allant de 2 à 20, en utilisant la fonction objective de ''Kapur''.



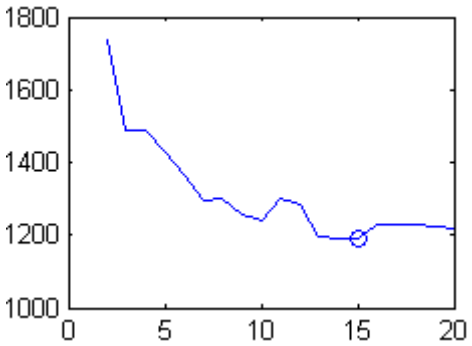
(a)- Indice de FH(k)

FH_{min}=37.74 => k_{opt}=2

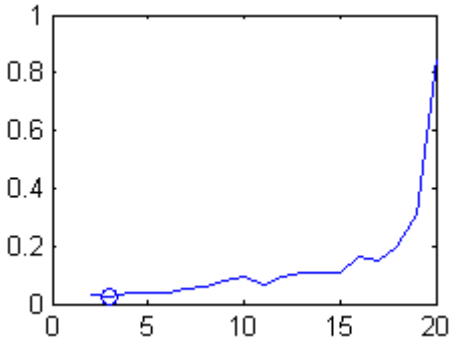


(b)- Indice de PA(k)

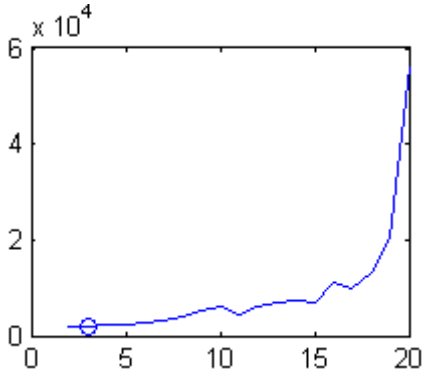
PA_{max}=79.12 => k_{opt}=2



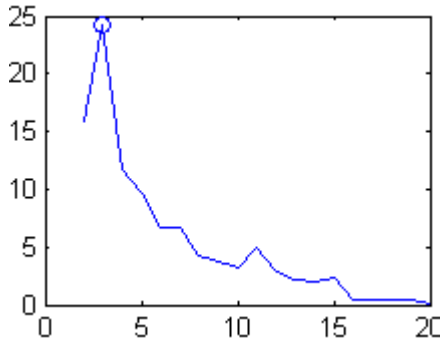
(c)- Indice de DP(k)
DP_{min}=1192 => k_{opt}=15



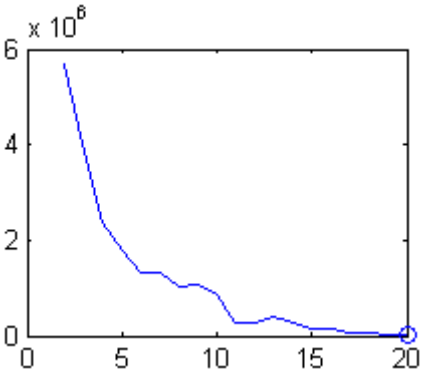
(d)- Indice de Xb(k)
XB_{min}= 0.02592 => k_{opt}=3



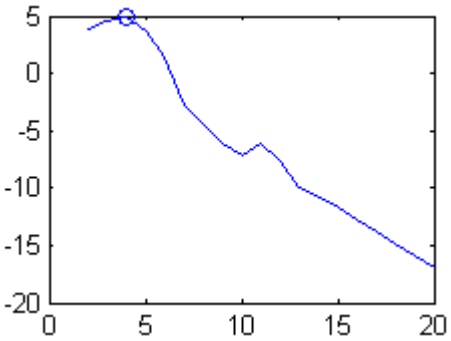
(e)- Indice de Kw(k)
Kw_{min}=1702 => k_{opt}=3



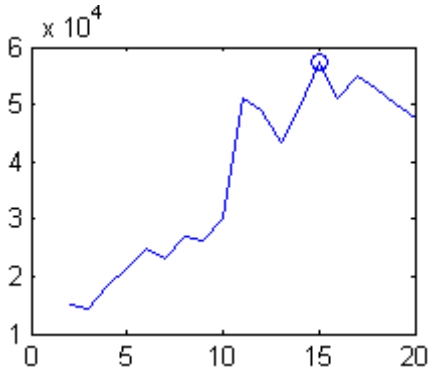
(f)- Indice de I(k)
I_{max}=24.18 => k_{opt}=3



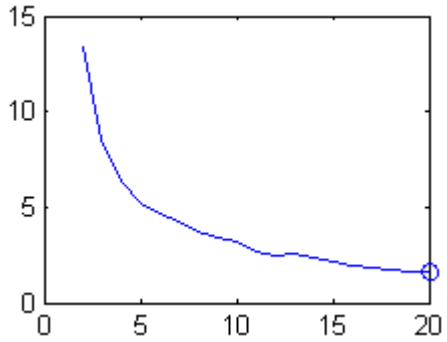
(g)- Indice de FS(k)
FS_{min}= 9112 => k_{opt}=20



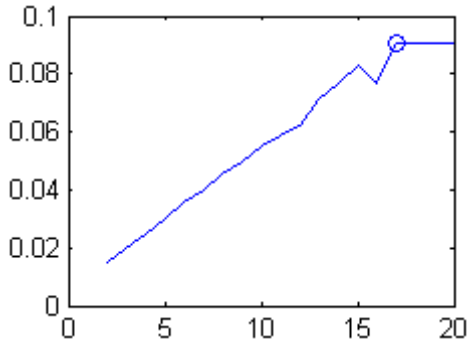
(h)- Indice de PCAES(k)
$PCAES_{max}$= 4.918 => k_{opt}=4



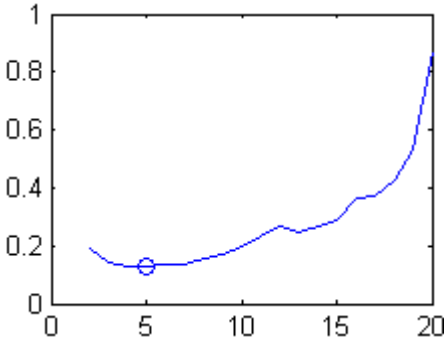
(i)- Indice de CH(k)
CH_{max}= 5.741e+004 => k_{opt}=15



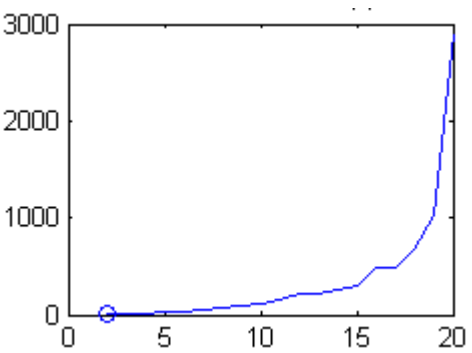
(j)- Indice de DB(k)
DB_{min}=1.59 => k_{opt}=20



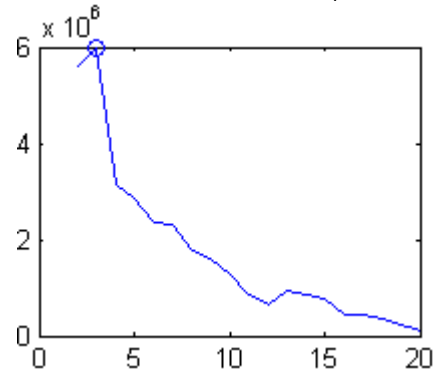
(k)- Indice de D(k)
D_{max}= 0.09091=> k_{opt}=17



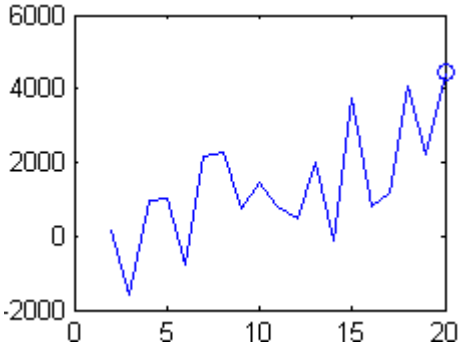
(l)- Indice de R(k)
R_{min}=0.1274 => k_{opt}=5



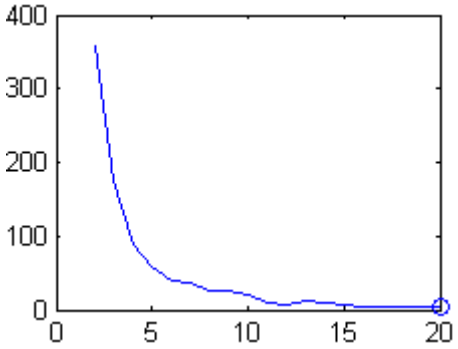
(m)- Indice de B(k)
B_{min}= 1.175=> k_{opt}=2



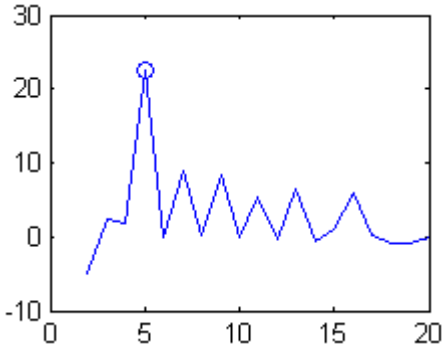
(n)- Indice de Icc(k)
Icc_{max}=5.965e+006 => k_{opt}=3



(o)- Indice de V(k)
V_{max}= 4442 => k_{opt}=20

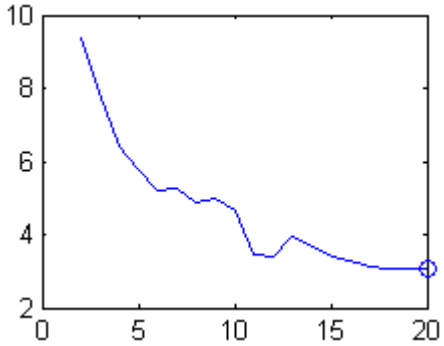


(p)- Indice de VCR(k)
VCR_{min}=3.21 => k_{opt}=20



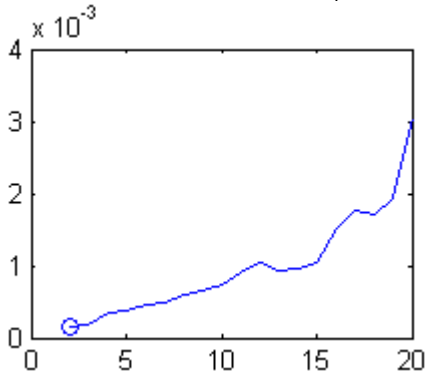
(q)- Indice de KL(k)

KL_{max}=22.42 => k_{opt}=5



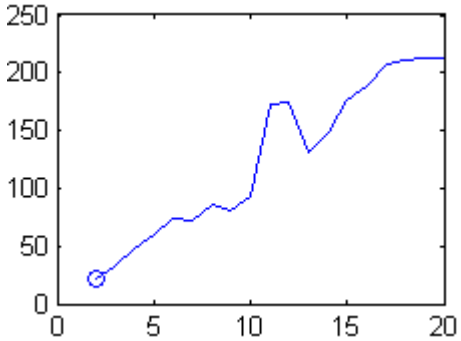
(r)- Indice de RMSSTD(k)

$RMSSTD_{min}$=3.087 => k_{opt}=20



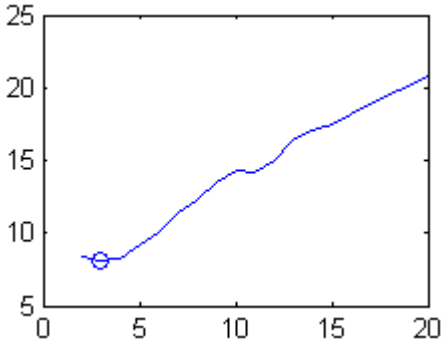
(s)- Indice de CS(k)

CS_{min}=0.0001608 => k_{opt}=2



(t)- Indice de J(k)

J_{min}=22.08 => k_{opt}=2



(u)- Indice de F(k)

F_{min}= 8.115 => k_{opt}=3

Figure (IV.12)- Indices de validité pour le seuillage d'histogramme de l'image ''Airplane'' on utilisant la fonction objective de ''Kapur''

Seuls les indices de Maulik I(k), Xb(k), Kw(k), Icc(k*)* et F(k), donnent un nombre de classes voulu qui est égal à trois (3).

La plus part des indices varient d'une manière monotone ce qui ne permet pas d'obtenir le nombre de classes plausible. Par contre les indices PCAES(k), R(k), KL(k) donnent un nombre de classes proche de trois (3), tandis que les indices DP(k), CH(k), D(k), V(k) surestiment le nombre de classes.

Par conséquent, pour l'image ''Airplane'', les indices communs aux deux fonctions objectives (Otsu et Kapur) qui donnent le nombre adéquat de classes, sont l'indice de Maulik I(k), Xie et Beni Xb(k), Kwon Kw(k) et Yen et Chang F(k).

Image 2

Cette image ''House'' de taille 256*256, codée sur 8 bits, est représentée sur la figure (IV.14). Son histogramme est illustré sur la figure (IV.15).

Il est également difficile d'observer le nombre exact de modes sur cet histogramme. Cependant, on peut distinguer approximativement quatre (4) modes.

Figure (IV.13)- Image ''House''

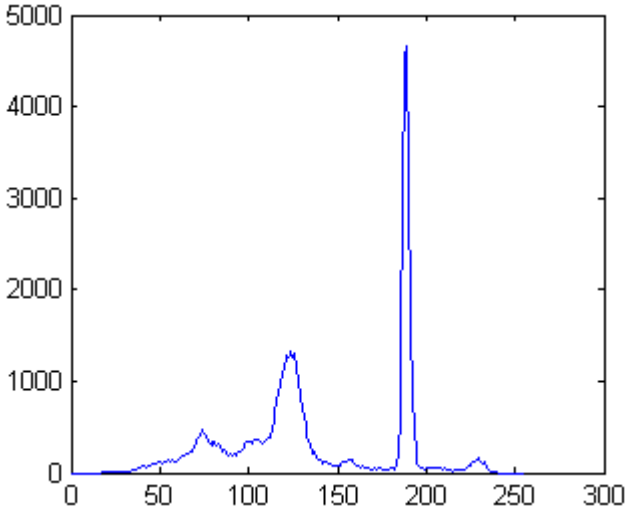


Figure (IV.14) : Histogramme de l'image ''House''

Le tableau (IV.6) donnent les valeurs des seuils obtenus par l'algorithme itératif

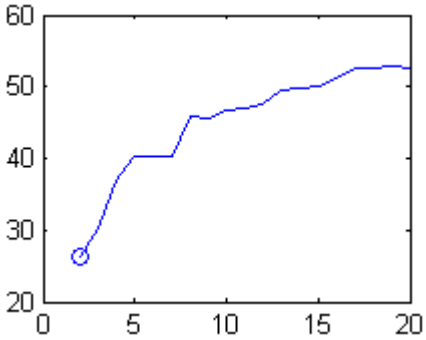
		Nombre de seuils		
		2	3	4
Otsu	Algorithme itératif	98-155	84-113-158	81-111-156-205
	Recherche exhaustive	97-156	83-113-159	83-113-157-206
Kapur	Algorithme itératif	91-165	73-129-185	61-105-149-193
	Recherche exhaustive	103-195	65-113-195	61-89-116-195

Tableau (IV.6)- Valeurs des seuils obtenues par l'algorithme itératif et par la recherche exhaustive.

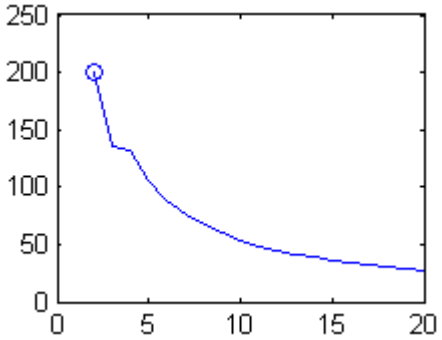
D'après le tableau (IV.6), on constate que les seuils optimaux obtenus on utilisant la méthode itérative et la fonction objective d' Otsu sont très proches des seuils trouvés avec la recherche exhaustive ; ce qui n'est pas le cas pour la fonction objective de Kapur.

Pour cette deuxième image, l'algorithme itératif n'a pas été très efficace pour la fonction objective de Kapur. Par conséquent, tous les tests qui suivent seront effectués en utilisant seulement la fonction objective de Otsu.

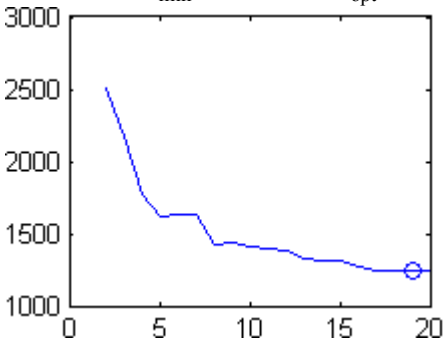
La figure (IV.15) montre la variation des indices de validité en fonction du nombre de classes allant de 2 à 20, en utilisant la fonction objective de ''Otsu''.



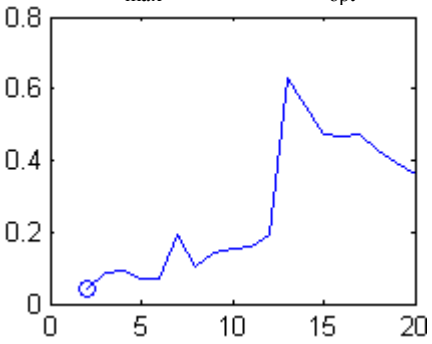
(a)- Indice de FH(k)
FH_{min}=26.17 => k_{opt}=2



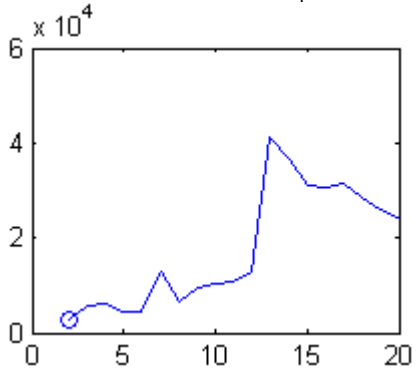
(b)- Indice de PA(k)
PA_{max}=200.2 => k_{opt}=2



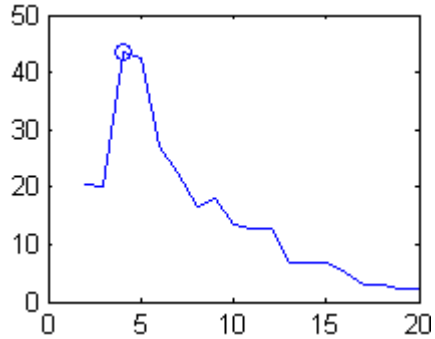
(c)- Indice de DP(k)
DP_{min}=1240 => k_{opt}=19



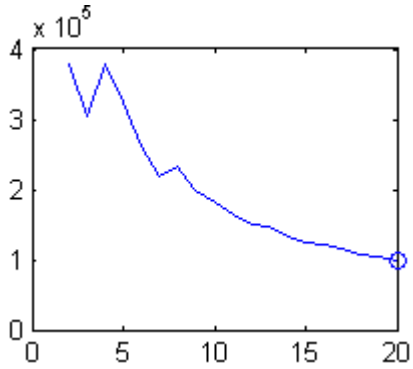
(d)- Indice de Xb(k)
XB_{min}= 0.03974 => k_{opt}=2



(e)- Indice de Kw(k)
Kw_{min}=2605 => k_{opt}=2

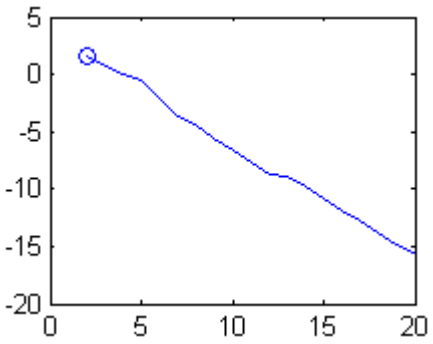


(f)- Indice de I(k)
I_{max}=43.62 => k_{opt}=4



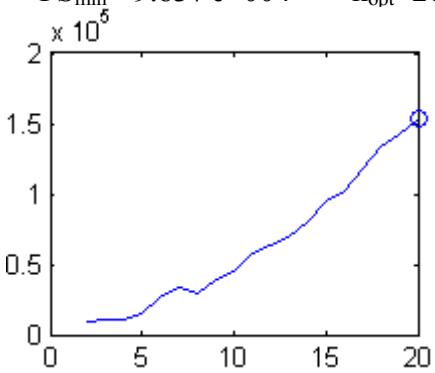
(g)- Indice de FS(k)

FS_{min}= 9.837 e+004 => k_{opt}=20



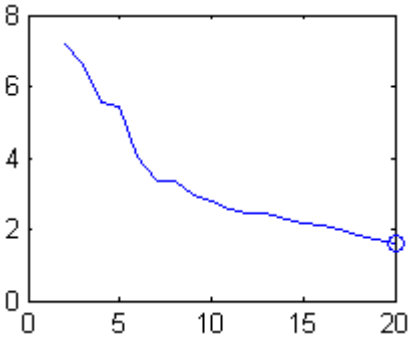
(h)- Indice de PCAES(k)

$PCAES_{max}$= 1.663 => k_{opt}=2



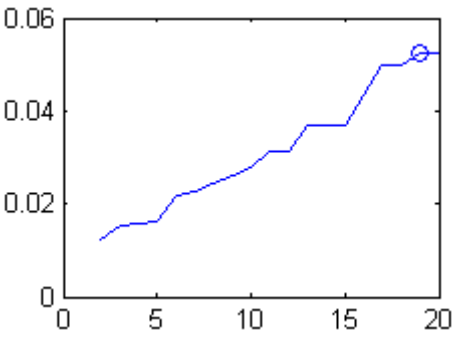
(i)- Indice de CH(k)

CH_{max}= 1.535e+005 => k_{opt}=20



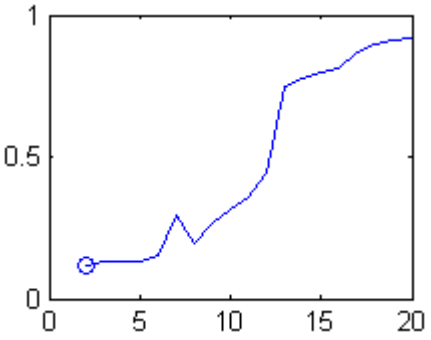
(j)- Indice de DB(k)

DB_{min}=1.631 => k_{opt}=20



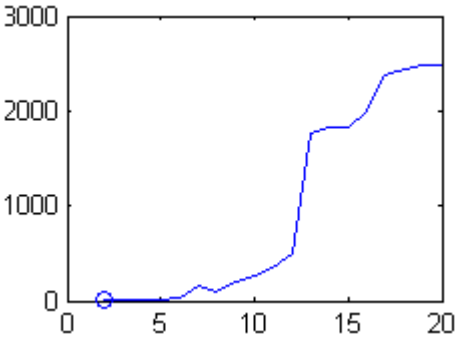
(k)- Indice de D(k)

D_{max}= 0.05263=> k_{opt}=19



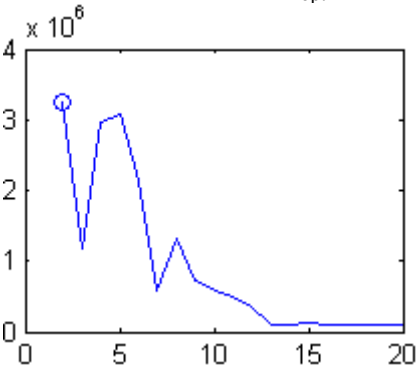
(l)- Indice de R(k)

R_{min}=0.1166 => k_{opt}=2



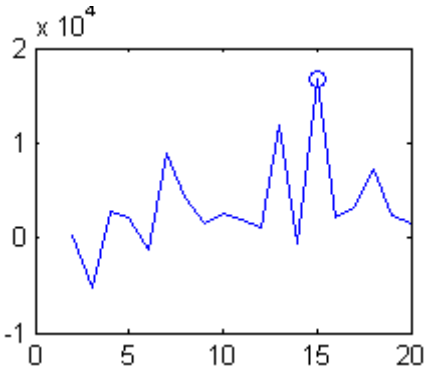
(m)- Indice de B(k)

B_{min}= 1.082=> k_{opt}=2

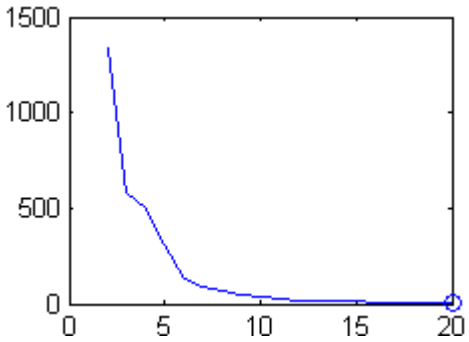


(n)- Indice de Icc(k)

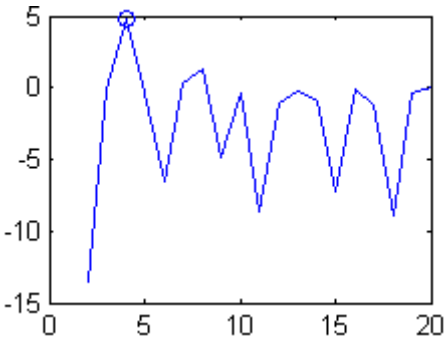
Icc_{max}=3.241e+006 => k_{opt}=2



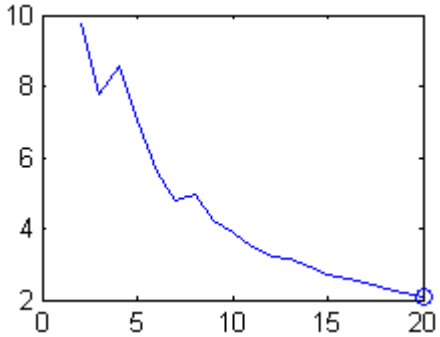
(o)- Indice de V(k)
V_{max}= 1.683e+004 => k_{opt}=15



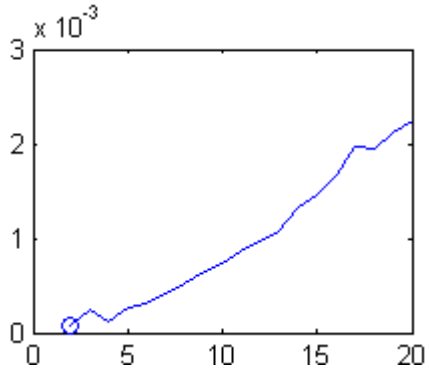
(p)- Indice de VCR(k)
VCR_{min}=4.507 => k_{opt}=20



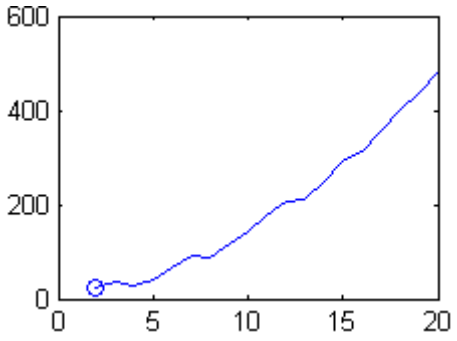
(q)- Indice de KL(k)
KL_{max}=4.751 => k_{opt}=4



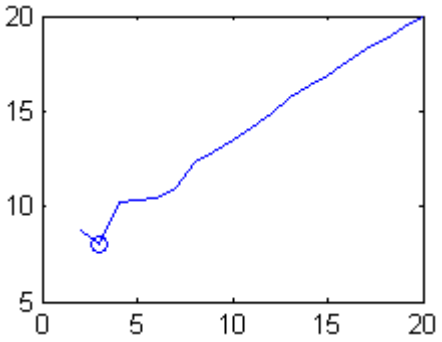
(r)- Indice de RMSSTD(k)
$RMSSTD_{min}$=2.102 => k_{opt}=20



(s)- Indice de CS(k)
CS_{min}=8.1 e-005 => k_{opt}=2



(t)- Indice de J(k)
J_{min}=21.45 => k_{opt}=2



(u)- Indice de F(k)

F_{min}= 7.999 => k_{opt}=3

Figure (IV.15)- Indices de validité pour le seuillage d'histogramme de l'image ''House '' en utilisant la fonction objective d'Otsu.

Vu la forme de l'histogramme de l'image ''House'', qui présente approximativement quatre modes, seuls les indices de Maulik I(k), Kl(k), donnent un nombre de classes égal à quatre (4). La plus part des indices varient d'une manière monotone ce qui ne permet pas d'obtenir le nombre plausible de classes. Par contre l'indice F(k) donne un nombre de classes proche de quatre (4), tandis que les indices DP(k), D(k) et V(k) surestiment le nombre de classes .

IV.4- Conclusion

Ce chapitre a été consacré aux tests, dont l'objectif est de montrer l'efficacité des indices de validité de la segmentation d'images par seuillage d'histogramme.

Il est important de préciser la complication de cette pratique qui est le choix du bon indice qui conduit au meilleur partitionnement.

D'une manière générale, parmi tous les indices évalués, celui de Maulik I(k) est sans doute celui qui a aboutit au nombre exacte de classes.

Bien que l'étape de validation puisse manifestement paraître cruciale, la difficulté provient du fait qu'il n'existe aucun critère universel qui puisse décider si un indice donné soit adapter à une image . Et c'est souvent sur la base des constatations empiriques que l'on se fait une idée sur la répartition réelle des pixels traités.

Conclusion générale

Le travail présenté dans ce mémoire concerne le domaine de traitement d'images et plus précisément celui de la segmentation. Parmi les nombreuses techniques de la segmentation, nous nous sommes intéressé particulièrement aux techniques de seuillage, qui utilisent des informations issues de l'histogramme de l'image, pour rechercher les seuils optimaux. Pour cela nous avons fait appel à l'algorithme itératif qui consiste à optimiser une fonction objective quelconque.

Par la suite, nous avons décrit quelque notions sur les indices de validité et nous les avons transformé et adapté afin de pouvoir les utilisés pour le cas de la segmentation d'image par seuillage d'histogramme.

Les tests réalisés ont montré l'efficacité de l'algorithme itératif à trouver les seuils optimaux en un temps de calcul minimal, ils ont aussi montré que la majorité des indices n'aboutissent pas forcément à de bons résultats. Néanmoins, un seul d'entre eux a attiré notre attention quant a sa justesse et à la fiabilité de ces résultats et aussi a sa compatibilité avec les deux fonctions objective ''Otsu'' et ''Kapur''; Il s'agit de l'indice proposé par Maulik I(k).

Pour finir, ce travail nous a permis d'approfondir nous connaissances théoriques et pratiques concernant le domaine de la segmentation d'image par seuillage et plus particulièrement sur les indices de validité et aussi d'acquérir de nouvelle connaissance sur le langage MATLAB.

Nous espérons que ce modeste travail ainsi réalisé sera utile pour les futures applications.

Bibliographie

[BEZ75] J.C Bezdek and J.C.Dunn, ''Optimal Fuzzy partitions A heuristic for estimating the parameters in a mixture of normal distributions'', IEEE Transactions on computers, vol 24, n. 8, pp. 825-838, 1975

[BOU01] M. Boudraa, Y.Batistakis and M. Vazirgiannis, ''Clustering algorithms and validity measure'', IEE, pp. 3-22, 2001.

[CAH74] R. B. Calinski and J.Harabasz, ''A dendrite method for cluster analysis,comm.. in Statistics'', vol.3, pp.1-27,1974

[CL83] C.H. Lee,'' Minimum cross entropy thresholding''. Pattern recognition. Vol 26N°4, pp.617-625,1983

[CS03] Chien-Hsing Chou, Mu-Chun Sun, Eugene Lai: ''A new clustervalidity measure for clusters with different densities.", 2003.

[CYC95] F.J. Chang, J.C. Yen and Chang,'' A New Criterion for Automatic Multilevel Thresholding'', IEEE trans. image process Vol 4.pp.370-378, 1995

[DA99] Deng Y., et al., "Color image segmentation," Proc.IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '99, Fort Collins, CO, vol.2, pp.446-51, June 1999.

[DAV79] D.L Davies ans D.W. Bbouldin,''A cluster separation measure'', IEEE Trans Pattern Anal. Machine Intell, vol. 1, no. 4, pp. 224-227, 1979.

[DUN74] J.C. Dunn, ''well separated clusters and optimal Fuzzy partitions'', J. Cybern, vol. 4, pp. 95-104, 1974

[FRA02] C. R. De Franco, L. S. Vidal and A.J.O. Crus, ''A validity measure for hard and fuzzy clustering derived from fisher's linear discriminant'', IEEE pp, 1493-1498, 2002.

[GAT89] I. Gath and A. B. Geva, ''Unsupervised Optimal Fuzzy Clustering'' IEEE Trans. Pattern Anal. Machine Intell ., vol. 11,no. 7, pp. 773-781, 1989.

[HAL01] M. Halkidi, Y. Batistakis and M. Vazirgiannis;'' On Clustering Validation Techniques'', Journal of Intelligent Information Systems, Vol. 17, No. 2-3, pp. 107-145, 2001

[HP77] S.L. Horowitz and T. Pavlidis. ''Picture segmentation by a directed split and merge procedure''. In *CMetImAly77*, pages 101.11, 1977.

[KI 86] J. Kittlerand J.Illingworth,'' Minimum error thresholding; pattern recognition'', vol.19, pp.41-47,1986

[K L85] W. J. Krzanowski and Y. T. Lai,'' A criterion for de-terming the number of groups in a data set using sum-of-squares clustering''. Biometrics 44, 23–34, 1985.

[KSW85] E.J.N.Kapur, P. K. Sahoo and A. K.C. Wang,'' A new method for gray-level picture thresholding using the entropy of the histogram'', comput Vision graphics imagepricess,vol29, pp. 273-285,1985

[KWO98] S.H. Kwon, ''Cluster validity index for fuzzy clustering'', Electron, vol. 34, pp. 2176-2177, 1998.

[MAC67] J. MacQueen, ''Some *Methods for Classification and Analysis of Multivariate Observation*'' P, roc. 5th Brekeley Symp, pp. 281-297, 1965.

[MAU04] M. K. Pakhira, S.Bandyopadhyay and U. Maulik,'' Validity index for crisp and Fuzzy clusters'', Pattern Recognitin, Vol. 37, pp. 487-501, 2004.

[MYM01] M. Halkidi, Y. Batistakis and M. Vazirgiannis,'' On Clustering Validation Techniques'', Journal of Intelligent Information Systems, Vol. 17, No. 2-3, pp. 107-145, 2001

[NS79] N. Otsu, ''A threshold selection method for Grey Level histogram'', IEEE Trans. on system, Man and Cybernetics, vol SMC-9,N°.1,1979

[RAZ89] B.L.M.R Razae and J. Reiber, ''A new cluster index for the fuzzy c-means, Pattern recognition'', vol. 19, pp. 237-256, 1989.

[She96] J. Shen,'' On multi-edge detection'', CVGIP, Graphics Models and Image Processing, 58(2):101--114, March 1996.

[TU05] Proceedings Of World Academy Of Science, Engineering And Technology Volume 9 November 2005 Issn 1307-6884

[Wro87] B. Wrobel and O. Monga,''Segmentation d'images naturelles : coopération entre un détecteur contour et un détecteur région'', In Actes du Onzième colloque GRETSI, Nice, France, Juin 1987.

[WY05] K. -L. Wu, M. -S. Yang, ''A Cluster Validity Index for Fuzzy Clustering'', Pattern Recognition Letters 2005, 26: 1275–1291.

[XIE91] X.L.XIE and G.Beni, ''A validity Measure for Fuzzy Clustering'', IEEE Trans. Pattern Anal. Machine Intel, vol. PAMI-13, no. 8, pp. 841-847, 1991.

[YC97] P.Y.Yin,L.H,Chen,''A fast iterative scheme for Multilevel thresholding methods'', Signal processing, vol 60,pp.305-313-1997

[YC95] J.C. Yen, F.J. Chang, S. Chang, ''A new criterion for automatic multilevel Thresholding'', IEEE Trans. Image Process. IP-4 (1995) 370–378

S.ZEGGANE, A.YOUYOU <<Les indices de validité dans la classification automatique>> Mémoire d'ingénieur, département automatique, Tizi Ouzou, 2006.

K.AMARA, N.YAHI <<Multi seuillage des images par les Algorithmes Génétiques>> Mémoire d'ingénieur, département informatique, Tizi Ouzou, 2005.

S.BOUMRAR, F.TAKILT <<UNE méthode de segmentation d'images à base d'un Algorithmes Génétiques>> Mémoire d'ingénieur, département informatique, Tizi Ouzou, 2005.