ISSN 0389-4010
UDC 519. 272

NAL TR-854

航空宇宙技術研究所報告
TECHNICAL REPORT OF NATIONAL AEROSPACE LABORATORY

TR-854

大規模衛星画像クラスタ解析の研究
ヒストグラム・モード法

松本甲太郎・中　正夫・山本浩通

1985 年 3 月

航空宇宙技術研究所
NATIONAL AEROSPACE LABORATORY

大規模衛星画像クラスタ解析の研究*
——ヒストグラム・モード法——

松本甲太郎** 中 正夫** 山本浩通**

A Study of the Clustering Method for a Huge Satellite Image
— Histogram Mode Method —

By
Kohtaro Matsumoto, Masao Naka, Hiromichi Yamamoto

ABSTRACT

The clustering technique is important as the unsupervised classifier in the realization of an automated satellite image analysis system. But current cluster analysis methods are not suitable to deal with huge, multi-dimensional satellite images because of the dependency on initial partitions and the tremendous computational time required.

In this paper, to analize a huge satellite image, we will describe two new algorithms, i.e., Multi-Layered Hashing Scheme and Histogram Mode Method. The first one is for the construction of a multi-dimensional histogram from the huge satellite image data. And the second one is for deriving natural clusters from the histogram using the basic idea above.

The algorithms were implemented on the large scale pipeline computer (FACOM 230/75 APU), and have been successfully used. Some results are shown in the paper.

はじめに

クラスタ解析手法は衛星画像の自動解析の教師無し分類手法として重要であるが，従来のクラスタ解析手法では少数のデータしか取り扱うことが出来なかったので，膨大な衛星画像の極く小さな部分画像の解析が可能であるにすぎなかった。このため，部分画像の選定が問題となり，教師無し分類としてのクラスタ解析のメリットが十分発揮できなかった。本論文はこのような点から，衛星画像のフレーム全体を1度に解析できる高速なクラスタ解析手法について論ずる。

* 昭和60年1月9日受付
** 計算センター

広義のクラスタ解析は〝似た者同士を集める〟統計学的手法であり，解析対象データから単純に似た者同士をまとめる場合と，解析対象データからある特徴をもった少数の代表的クラスタを抽出する場合がある（文献1）。リモートセンシングにおいて衛星画像の自動解析に用いるクラスタ解析は，主に後者の場合である。解析対象（衛星画像）から地上のカテゴリーに応じた特徴をもったクラスタを抽出し，これによって衛星画像を弁別分類し地上の被覆分類図を作成する事が衛星画像クラスタ解析の目的である。

衛星画像のクラスタ解析には，これまで大別して2種類の手法が使われて来た。1つはBall & Hallの非階層的再配置法の流れをくむもので代表的なプ

ログラムにISODATA, ISOCLS, LARSYS（文献12,5,20）等があり,手軽で高速なことから広く用いられている。他の手法は，データのヒストグラムを構成し，対話的に頻度による閾値処理を繰り返し，クラスタ分割を行なうもので，カナダのCCRSで開発された（文献２）。

本論文では，後者と同様にヒストグラムを用いるが，その基本的考え方は，Wishartの階層的モード法（文献13）に近い新しい手法（ヒストグラム・モード法と名付ける）について報告する。その基本的な考えは，衛星画像から多次元ヒストグラムを構成し，そのモード（即ち，確率密度の山）を多次元空間でのヒストグラム頻度のパターンとして捉え，このモードによってクラスタ解析するものである。

ヒストグラム・モード法の基礎となった階層的モード法では，各データの近傍での標本密度から密度の極大点を求めてクラスタを構成していた。ヒストグラム・モード法では，予めヒストグラムを構成し，n次元空間での確率密度の極大点をヒストグラムの形から決定し，その周囲にクラスタを定める。衛星画像をこのようにしてクラスタ解析するためには，

(1) 多次元空間でのモードの把握方法
  (a) 頻度の極大値，谷間決定手法
  (b) モード（クラスタ）の表現方法
(2) 多次元ヒストグラムの効率的構成方法
(3) 多次元ヒストグラムの標本誤差除去

などの問題に対処していかなければならない。又，このような手法的問題とは別に，衛星画像の大規模性から必要となる膨大な計算量をどのように効率的に処理するかという計算機の高度利用も重要な課題となる。

以下，１章においては，衛星画像及びその分類解析の概略と，本論文で提案するヒストグラム・モード法の位置付けを述べる。２章においては前記(2)の問題を解決する，大規模画像データの多次元ヒストグラム構成方法について述べる。３章では前記(1),(3)の問題を解決するヒストグラムモード法について述べる。又，４章では，本プログラムにおいて用いたベクトル型計算の高度利用法について述べる。最後に，５章において本論文において述べる手法により衛星画像を解析した例について述べる。

1. 序　論

本論に入る前に，問題の背景として，リモートセンシングに用いられる衛星画像の分類解析手法の概略について述べる。

1.1 衛星画像の分類解析手法

宇宙から人工衛星で地上を分光観測して得られる情報には

(1) 地上の各地点の分光特性
(2) 地上の表面の幾何的形状

がある。後者の情報は得られた画像を２次元的に観察することによって得られるもので，地形のリニアメント抽出などが主に行なわれているが，本論文では取り扱わない。ここでは，前者の分光特性に基づく分類解析の問題を取り扱う。

分光特性情報は，地上の各地点を何バンドかに分光し各地点の離散的分光スペクトルを観測して得られる。この情報から各地点のカテゴリー（即ち，森，都市，など）をスペクトルパターン（色）によって判別できると期待される。例えば，図１－１に示すように，森は波長の短い方にスペクトルのピークがみられ，乾燥地面では逆に波長の長い方に高原状のピークが見られる，というように分光スペクトルパターンが異なるのでこの違いをキーに両者を弁別する。

このような分光スペクトルの違いに基づいて，衛星画像をいくつかのカテゴリーに分類する手法は，人間の持つ知識をどう利用するかによって，教師付き分類／教師無し分類の２種類に分かれる。

教師付き分類とは，判別のための弁別関数を人間

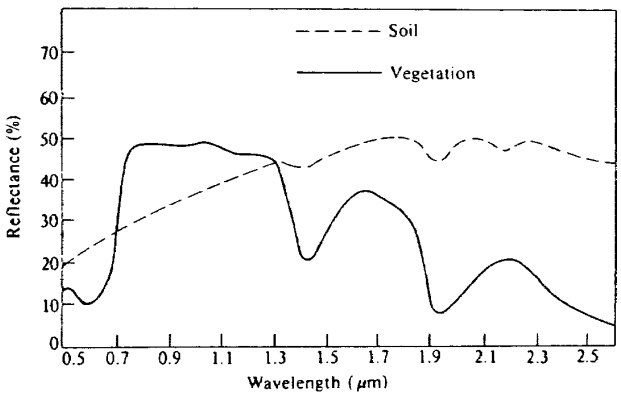


図１－１ 分光スペクトル曲線の例

（実験者，即ち教師）の知識を直接的に利用したトレーニング過程によって定め，これによって衛星画像をカテゴリー分類するものである。この方法は簡単なため広く用いられているが，トレーニング過程の妥当性が極めて重要なファクターであるにもかかわらず，どのようなトレーニング処理が適切かについて未だ手法が確立していない難点がある。

これに対して教師無し分類は，いわゆるクラスタ解析手法を用いて，画像からスペクトルパターンが似通っている事を基準にいくつかのクラスタを抽出し，各クラスタを弁別カテゴリーに対応づけてカテゴリー分類するものである。実験者の知識はクラスタとカテゴリーの対応づけに用いられるだけなので，教師付き分類に不可避であったトレーニング処理の妥当性の問題が軽減され，より客観的な解析が可能である。

クラスタ解析手法には，従来，解析対象のデータの質，解析の目的などによって様々な手法が考案されている。表1－1に従来のクラスタ解析手法のうち代表的なものを衛星画像に適用した場合の問題点について示す。多くの手法では衛星画像の膨大なデータ量をうまく取り扱えず，わずかにBall & Hallの再配置法とCCRSのヒストグラム法が実際に用いられているのみである。

再配置法を用いたプログラムではISODATA，ISOCLS，LARSYS等がある。この手法では，適切な初期分割を与えられれば，反復繰り返しにより最適なクラスタ分割に至る事が知られている。しかし，適切な初期値が与えられる事が極めて重要であり，任意の初期分割の下では準最適な解にしか収束しない事が知られている（文献6）が，適切な初期値設定方法が未だ確立されていない点が問題である。

第2のヒストグラムによる方法は，Shlien 等によって提案されたもので，衛星画像の多次元ヒストグラムを作り，対話型にヒストグラムの閾値処理を行ないクラスタを分離して行くものである。この方法では処理が対話型である点に問題が残る。対話型である事は，"教師無し"である事によって得られる客観性というメリットを捨てるもので，対話過程の妥当性という問題が新たに起きてくる。

1.2　ヒストグラムモード法

本論文で提案するヒストグラム・モード法は従来は衛星画像のクラスタ解析には用いられていなかったWishartの階層的モード法を基礎にしたものである（文献13）。

階層的モード法は，それまでのクラスタ解析手法においてややあいまいなまま用いられて来たクラスタ（データの魂）の定義を，確率密度関数のモードと定め，理論的に明確にした。このため，従来の手法より"自然な"クラスタを生成できる特徴を持っている。

表1－1　クラスタ解析手法を衛星画像に適用する場合の問題点

	計算負荷	メモリー量	問　題　点
階層的手法	N^2	N^2	膨大な計算負荷とメモリー
非階層的手法			
再配置法 (Ball & Hall)	$N \cdot M \cdot A$	M	計算負荷，初期クラスタ
丘登り	$N \cdot M \cdot B$	N	計算負荷，初期クラスタ
強制移動	$N \cdot M \cdot B$	N	計算負荷，初期クラスタ
ヒストグラム法	$N' \cdot M \cdot C$	N'	対話型
階層的モード法	N^2	N^2	膨大な計算負荷とメモリー

N ： サンプル数，　M：クラスタ数

N'： ヒストグラムセル数（10^4程度）

A, B, C：計算負荷係数　$A < C \ll B$

（但し，データの再サンプリング，次元数の減少は行なわないとした。）

階層的モード法は個体間の距離計算を基本的に用いているため，表1－1に示されるように計算負荷メモリーの点で少数データ用のアルゴリズムであった。

しかし，階層的モード法のアルゴリズムを詳細に検討し，

(a) 同手法において重要な役割を果たしている〝データ密度〟は確率論的には確率密度関数の近似的推定である事，

(b) クラスタの分割・融合によって探索決定される確率密度関数のモードが，多次元空間のヒストグラム頻度の極大点（即ち，密度の山）である事（図1－2），

の2点に注目し，前節のヒストグラム法と組み合わせる事によって，大規模な衛星画像にも適用できる新しい手法，ヒストグラム・モード法を以下に提案する。

ヒストグラム・モード法の基本的アプローチは，

(1) 大規模データから多次元ヒストグラムを構成する，

(2) ヒストグラムのモード（即ち，頻度の極大点）によって混合確率密度関数のモードを推定する，

(3) クラスタの境界は，極大点の周囲に順次クラスタを成長させ，クラスタが互いに接触した所（即ち，頻度の谷間）とする，

に基づいている。ヒストグラム・モード法によって大規模データを効率よく解析するためには，

1) 多次元ヒストグラムを効率よく構成する方法（次元の呪いの問題），

2) 多次元ヒストグラムの極大点を核にクラスタを構成する方法（特に計算負荷の小さい手法），

3) ヒストグラムに不可避の標本誤差の処理方法，

等が問題となる。

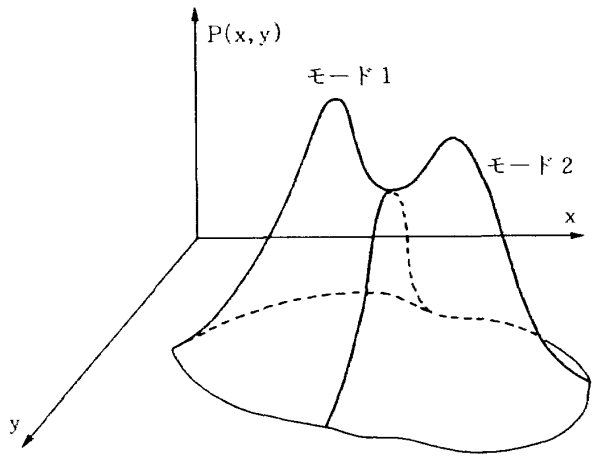


図1－2 確率密度分布関数のモードによるクラスタ分割

以下，2章において，多次元ヒストグラムの作成方法について述べ，2)，3)に関しては3章において述べる。

2. 多次元ヒストグラムの構成法 — 多層ハッシング法 —

本論文で述べるヒストグラム・モード法には多次元ヒストグラムが重要な役割を果たす。本章では多次元ヒストグラム構成に伴う問題について論じ，新たに考案した多層ハッシング法について述べる。

2.1 ヒストグラムの構成手法

n次元ヒストグラムの構成手法には大別して2種類ある（図2－1）。

(1) 完全型 n次元配列を用いる方法

(2) 近似型

(a) Karhunen-Loeve変換を用いる方法

(b) テーブル探索手法を用いる方法

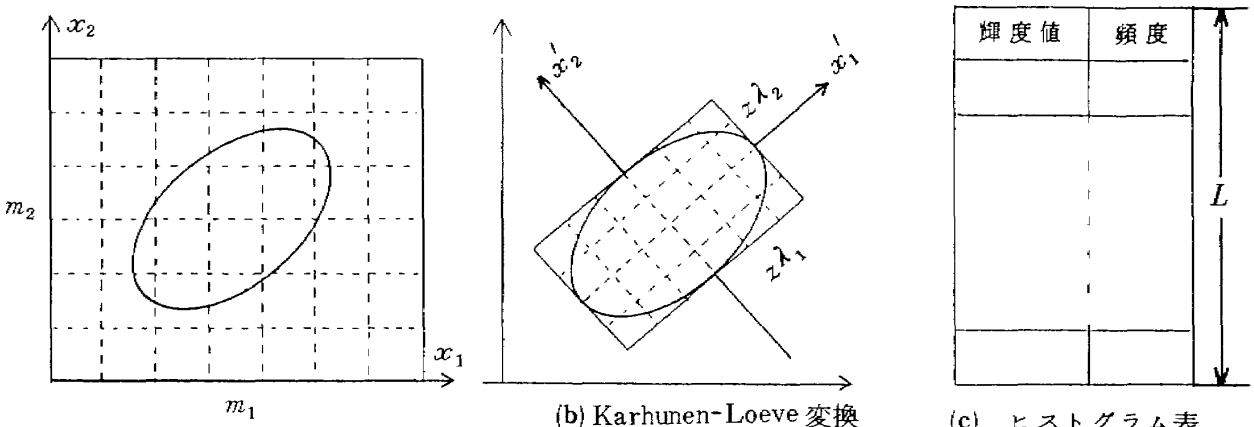
以下，これらのそれぞれについて述べるが，その前提として，ヒストグラムを構成する対象の観測データについて次のように仮定する。

1) 観測データにはN個の母集団C_iから生成された標本xがランダムに存在する。

2) 観測データはn次元ベクトルVで表わされ，各次元毎のデータは区間〔0，a_i〕に存在する。

3) 各母集団は平均μ_i，分散K_i，先験確率P_iのガウス分布をなす。

このような仮定は極めて一般的であり，クラスタ解析する殆どのデータに対して仮定してよい。又，仮定3）は2．1．2節において，近似的ヒストグラムに必要な領域体積を推定するためだけに必要な仮定であり，ヒストグラム構成アルゴリズムとは本質的な関係は無い。従って，2.2節で述べる多層ハッシング法は非ガウシアンなデータに対しても適用できる。



(a)　多次元配列　　(b) Karhunen-Loeve 変換　　(c)　ヒストグラム表

図2－1　多次元ヒストグラムの作成法

2.1.1　完全なヒストグラムの構成方法
—　n次元配列を用いる方法　—

この方法は，観測データの完全なヒストグラムを作ろうとするものである。n次元の各データ区間〔0，a_i〕を適当に量子化し，離散値ベクトル$\boldsymbol{v}$を作る（式(2.1)）。離散値ベクトル$\boldsymbol{v}$の各次元毎の量子化のレベル数m_iから，ヒストグラムを構成するために用いるn次元配列は，式(2.2)の形で定められる，いわゆる"配列"でよい。その方法に必要なメモリー量はV_1（式(2.3)）に比例する。

$$\boldsymbol{v}=(v_1, v_2, \cdots v_n) \qquad (2.1)$$

$$m_1 \times m_2 \times m_3 \times \cdots \times m_n \qquad (2.2)$$

$$V_1 = \prod_{i=1}^{n} m_i \qquad (2.3)$$

この方法は，次元数が1，2のように小さいあいだは極めて有効である。事実，2次元以下のヒストグラム作成には他の方法を考慮する必要は全くない。しかし，次元数が3次元以上になってくると，いわゆる「次元の呪い」のために，ヒストグラム構成に必要なメモリー領域が天文学的になりはじめ，4次元以上ではこの方法は殆どの場合役に立たない。

データは多次元空間の全てに一様に存在するわけではなく，局所的にかたまって存在している（従ってクラスタ解析を行なう意味がある）。このため，この莫大なメモリー領域の殆どは結果的に頻度0で本質的には不必要なメモリー領域である。

2.1.2　近似的ヒストグラムの構成

観測データの"完全な"ヒストグラムを作ろうとすると2.1.1節で述べたように"次元の呪い"のために莫大なメモリー領域が無駄になるが，ヒストグラムの完全性を多少犠牲にした近似的なヒストグラムならば，莫大な無駄領域のほとんどを必要としない手法が開発できる。

観測データになんらかの統計的母集団を想定すれば，観測データは多次元空間のある部分に密に集中しているはずで，他の大半の領域では極めて疎にしか存在せず無視し得るであろう。従って，データが集中する部分でだけ集中的にヒストグラムを作り，他の部分は無視しても観測データの大多数を表わすヒストグラムが構成でき，これによって全空間のヒストグラムが近似できるであろう。

各母集団からの標本がその先験確率に比例してヒストグラムに含まれるものとし，ガウス分布を仮定すると，各母集団の平均の周りに部分領域A_iが次式で定められる。

$$A_i=\{X \mid C_i \ni X,\ (X-\mu_i)^t K_i^{-1}(X-\mu_i)<\chi^2(n,1-P)\} \qquad (2.4)$$

$(X-\mu_i)^t K_i^{-1}(X-\mu_i)$は自由度$n$の$\chi^2$分布をなすので，式(2.4)で定まる領域$A_i$には母集団$C_i$から生成される標本$X$のうち$P$%のものが含まれる。従って，これらの領域$A_i$（$i=1..N$）にだけしかヒストグラムセルを設定しなくても，確率的に全データのP%に相当するデータからなるヒストグラムが構成できる。必要なヒストグラムセルの数は領域同士の重なりを無視すれば，各領域A_iの体積の総

和V_2に比例する。

$$W_i=\frac{(\pi z^2)^{n/2}\cdot|K_i|^{1/2}}{\Gamma(n/2+1)},\ i=1\cdots N. \quad (2.5)$$

$$V_2=\sum_{i=1}^{N}W_i=\frac{(\pi z^2)^{n/2}}{\Gamma(n/2+1)}\cdot\sum_{i=1}^{N}|K_i|^{1/2} \quad (2.6)$$

但し $z^2=\chi^2(n,1-P)$

$$R_{12}=\frac{V_1}{V_2}=\prod_{i=1}^{n}\frac{m_i}{(\pi z^2)^{1/2}}\cdot\frac{\Gamma(n/2+1)}{\sum_{j=1}^{N}|K_j|^{1/2}} \quad (2.7)$$

完全なヒストグラムを作る場合に必要な体積との比は図2－2に示すように，次元数の冪で大きくなる。従って，近似的ヒストグラム構成法は完全なヒストグラムを作る場合よりもメモリー領域の有効利用が図れ，より高次元なデータに対して適用できると言える。

近似的ヒストグラムを構成する方法として，

Karhunen－Loeve変換を用いる方法

テーブル探索手法を用いる方法

が考えられる。本論文で述べる多層ハッシング法は後者のテーブル探策による手法のうちハッシング法を発展させたものである。

(1) Karhunen－Loeve変換を用いる方法

この方法は2.1.1節の多次元配列法を用いて近似的ヒストグラムを構成する方法である。

ヒストグラムを作るときに各母集団の統計量が既知であるということは，一般には仮定できない。従って，領域A_iを個別に予め定めることはできない。しかし全観測データの平均μ_0，共分散K_0は簡単に求められるから，各個別領域A_iの代わりに全体で

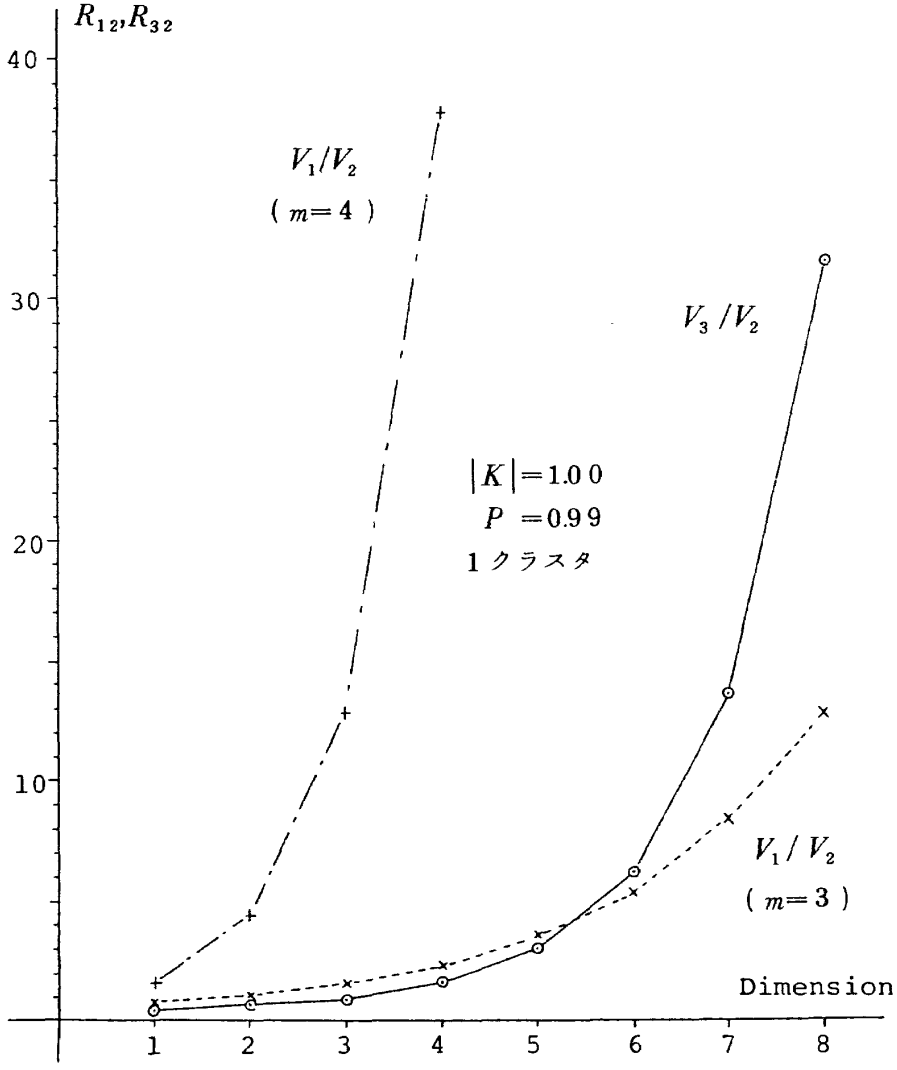


図2－2 多次元ヒストグラムを構成するのに必要なメモリー量の比較

1つの領域A_0をμ_0, K_0によって定められる。このときヒストグラムの近似度Pがある程度大きければA_0によって各個別領域A_iの和領域A_Tを近似してもよかろう（付図1参照）。

$$A_T = \bigcup_{i=1}^{N} A_i \tag{2.8}$$

$$V_3 = \frac{(\pi z^2)^{n/2}}{\Gamma(n/2+1)} \cdot |K_0|^{1/2} \tag{2.9}$$

A_0はn次元楕円体であるが，Karhunen-Loeve変換には，これに外接するn次元直方体A_0'を用いる。A_0'はその体積がV_3'（式(2.10)）で表わされA_0と殆ど同程度であり，かつアドレス計算は極めて簡単である。

$$V_3' = (2z)^n \cdot |K_0|^{1/2} \tag{2.10}$$

$$R_{33} = \frac{V_3'}{V_3} = \frac{2^n \cdot \Gamma(n/2+1)}{(\pi)^{n/2}} \tag{2.11}$$

観測値Xを，μ_0，K_0によってKarhunen-Loeve変換しX'を得る（式(2.12)TはKarhunen-Loeve変換行列。n次元直方体領域A_0'は，共分散行列K_0の固有値λ_iによって式(2.13)で規定されるから，各軸毎に区間$[-Z\lambda_i, Z\lambda_i]$を所要の分解能に従って適当に量子化してヒストグラムを作成すればよい（図2－1(b)）。

$$X' = T \cdot (X - \mu_0) \tag{2.12}$$

$$A_0' = \{X' \mid -Z\lambda_i < x_i' < z\mu_i\} \tag{2.13}$$

この方法に必要なメモリー領域の大きさは式(2.10)のV_3'に比例するが，これはV_2に比べると（式(2.14)），母集団の数N及びμ_i, K_iによって変わるが，数倍程度は大きいと言わざるを得ない。

$$R_{32} = \frac{V_3'}{V_2} = \frac{2^n \cdot \Gamma(n/2+1)}{(\pi)^{n/2}} \cdot \frac{|K_0|^{1/2}}{\sum_{i=1}^{N} |K_i|^{1/2}} \tag{2.14}$$

(2)　テーブル探索手法を用いる方法

本節で述べる手法は，2.2節で述べる多層ハッシング法の基礎をなす手法である。

Karhunen-Loeve変換を用いてヒストグラム構成範囲をデータが密に存在する領域に限ってメモリー領域の節約を図る前者の方法は，力まかせの多次元配列法に比べればはるかにメモリーを効率よく使えるが，それでもP%の観測データに相当するヒストグラムを作るのに本当に必要な領域の広さV_2と比べると無駄な領域が残る。

本節では原理的にメモリー領域の広さがV_2に比例するテーブル探索型のヒストグラム構成方法について述べる。この方法はヒストグラム表及びテーブル探索手法からなる。ヒストグラム表（図2－1(c)）は，

（頻度値，観測データ）$=(Hf(i), Hv(i))$

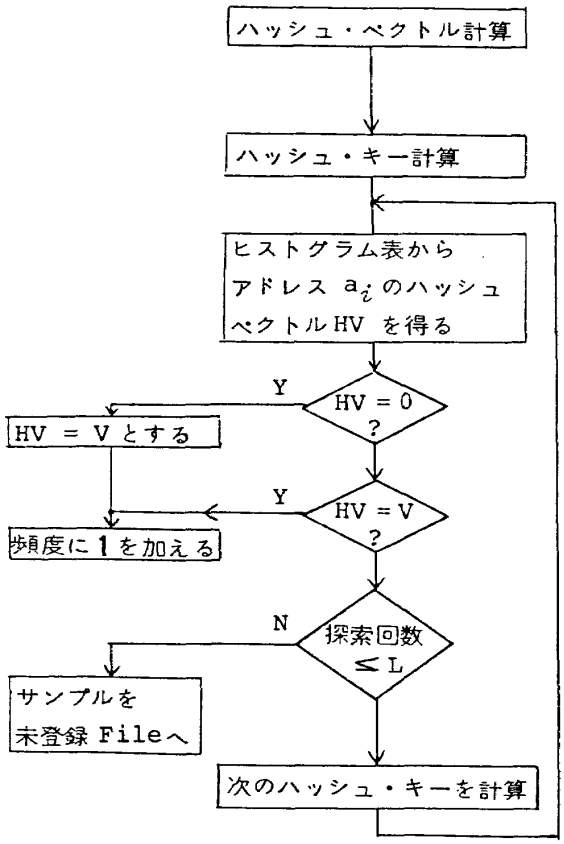
を1セットにしたヒストグラムセルを1つの要素とする長さLの表である。表の長さLは，データvの取り得るあらゆる場合の数V_1よりずっと小さく定める，表の各要素とデータvの間には1：mの写像関係が成り立つ。このため，あるヒストグラムセルの位置kに対応する観測データvは何通りも存在し，ヒストグラムセルの衝突が起きる。この衝突の問題を解決し，観測データvからセルの位置kを決定するために，テーブル探索手法が用いられる。

ハッシング法は，このようなm：1の写像関係を効率よく取り扱えるテーブル探索手法として知られている。ハッシング法では，ヒストグラム表の長さLを素数P_1になるように選ぶ。P_1に近くP_1を超えない他の素数をP_2とする。観測データvは式(2.1)のように量子化されているとする。この時，ハッシング法では，ヒストグラム表探索のためのキーa_i, bを式(2.16)で定め，これを用いてStep 1-Step 3（図2－3(b)）のアルゴリズムで観測データvを格納するヒストグラムセルを探索する。

$$V = (\cdots(v_1 * m_1 + v_2) * m_2 + v_3) * m_3 + \cdots) * m_n + v_n \tag{2.15}$$

$$\begin{aligned} a_1 &= \mathrm{mod}(V, P_1) + 1 \\ b &= \mathrm{mod}(V, P_2) + 1 \\ a_i &= \mathrm{mod}(a_{i-1} + b, P_1) + 1, \quad i \geq 2 \end{aligned} \tag{2.16}$$

図2－3にハッシング法のアルゴリズムを示す



X：サンプル
$= (x_1, \cdots, x_n)$

$V = \sum_1^n x_k \cdot (2^m)^{k-1}$

$a_1 = \mathrm{mod}(V, P_1) + 1$

$b = \mathrm{mod}(V, P_2) + 1$

P_1, P_2：ヒストグラム表の長さを超えない相異なる最大の素数

$a_i = \mathrm{mod}(a_{i-1} + b, P_1) + 1$

(a)

Step1：観測データ v からハッシュ・キー a，b を求める。

Step2：ヒストグラム表の a 番目の要素を調べ，

a) $Hf(a) = 0$ ならば
$Hv(a) = v$，$Hf(a) = 1$ として，v の処理終了。

b) $Hv(a) = v$ ならば
$Hf(a) = Hf(a) + 1$ として，v の処理終了。

c) それ以外ならば，Step 3へ。

Step3：$i = i + 1$ として，
i が N 以下ならば，式(2.16)で a_i を求め Step 2 へ，
i が N 以上ならば，v を未登録データとして終了。

(b)

図2－3 ハッシング法によるヒストグラム作成のアルゴリズム

(これを以下基礎的ハッシング法と呼ぶ)。

Shlien 等はこのハッシング法をランドサット MSS 画像に応用し，ヒストグラム表の長さ 6,000 で 250,000画素の 95% 以上を表わす4次元ヒストグラムが殆どのシーンに対して，構成できる事を示した。

2.2 多層ハッシング法

本節では多次元ヒストグラムを効率よく作成するために，筆者等が新たに考案した多層ハッシング法について述べる。

2.2.1 多層ハッシング法

ハッシング法は，本来はハッシング表に登録される相異なる事象の数が，表の長さ L より小さい場合

のアルゴリズムである。このためハッシング表（ここではヒストグラム表）が満杯になった時の処理アルゴリズムは不十分なものしかない。しかし，衛星画像のような膨大なデータの多次元ヒストグラムを，比較的小規模なヒストグラム表によって作る場合には，ハッシング表が満杯になる事は不可避なのでその場合を効率的に取り扱う手法がハッシング法に付加されなければならない。

従来はこの問題は殆ど論じられておらず，わずかに

「表が一杯になりかけたら，頻度の少ないヒストグラムセルを取り除いて空き領域を作り，残ったヒストグラムセルを再度ハッシングしてヒストグラム表を作り直す。」（再ハッシング処理）

と言う簡単な方法が知られているのみであった。このような方法によってもデータ量が少ない間は余り重大な問題にならなかった。しかし，データ量が増して来た時

(1) ある程度以上のデータを処理した後は，頻度の少ないヒストグラムセルを除く再ハッシング処理が頻繁に必要になる。この結果，計算時間が急激に増大し，ヒストグラムの計算負荷よりも再ハッシング処理の計算負荷の方がはるかに大きくなる。

(2) ヒストグラムセルを取り除くときに，これから処理するデータにおける頻度分布が全く考慮されない。このため，実際には頻度の大きなセルであるにもかかわらず，構成されたヒストグラムに存在しない事が在り得る。

と言う問題がデータ量が増すにつれ重要になって来る。特に，第2の点は衛星画像においては重要である。各クラスタ（母集団）のデータは衛星画像の中で空間的に偏って存在しているが，ハッシング法では結果が処理にかける画像データの順序に依存するので，画像の後半部分に存在する大きなクラスを逃す事につながってくる。

2.2.2　多層ハッシング法のアルゴリズム

ここで述べる多層ハッシング法は，このような問題を多数のヒストグラム表を動的に使用する事によって解決するものである。

多層ハッシング法では図2－4に示すように，最初のステップでは衛星画像の帯状の部分画像のそれぞれに対して別々のヒストグラム表を用いてヒストグラムを作る。部分画像の大きさは一定ではなく，画素当りの計算時間の増加，ヒストグラム表への未登録画像数の増加（図2－3 Step3）等，ハッシングの効率が低下して来た場合に，新しいヒストグラム表に切り替える。このようにして衛星画像全体に対応するヒストグラム表を作った後に，次のステップでこれらのヒストグラム表を融合して1つのヒストグラムとする。ヒストグラム表の融合にもやはりハッシング法を用いる。

基礎的ハッシング法で問題となった再ハッシング処理は多層ハッシング法では行なわず，複数のヒストグラム表の使用とそれらの融合によって計算負荷を軽減する事が出来る。

又，クラスタが空間的に偏在しても，多層ハッシング法では最初のステップで部分画像毎にヒストグラム表を更新するので，クラスタが十分な割合を占めていさえすれば，いずれかのヒストグラム表で大きな頻度を持つヒストグラムセルとなり，最終的なヒストグラムに適切に反映される。

多層ハッシング法のアルゴリズムを以下に示す

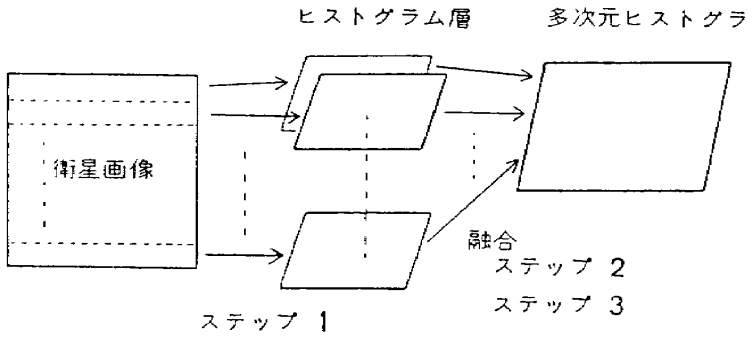


図2－4　多層ハッシング法

(図2－5，6，7)。本手法の基本的考えは，大量の生データ(衛星画像)を取り扱う部分(Step A)は，できるだけ効率の良い高速処理を行ない，ある程度圧縮したデータ(ヒストグラムセル)に対しては手が込んで時間がかかるが，高精度な処理(Step B,C)を行なうものである。このような高速処理と高精度処理の組み合わせにより，処理時間の短縮が図れるばかりでなく，衛星画像に見られるクラスタの空間的な偏りにも効果的に対処できる。

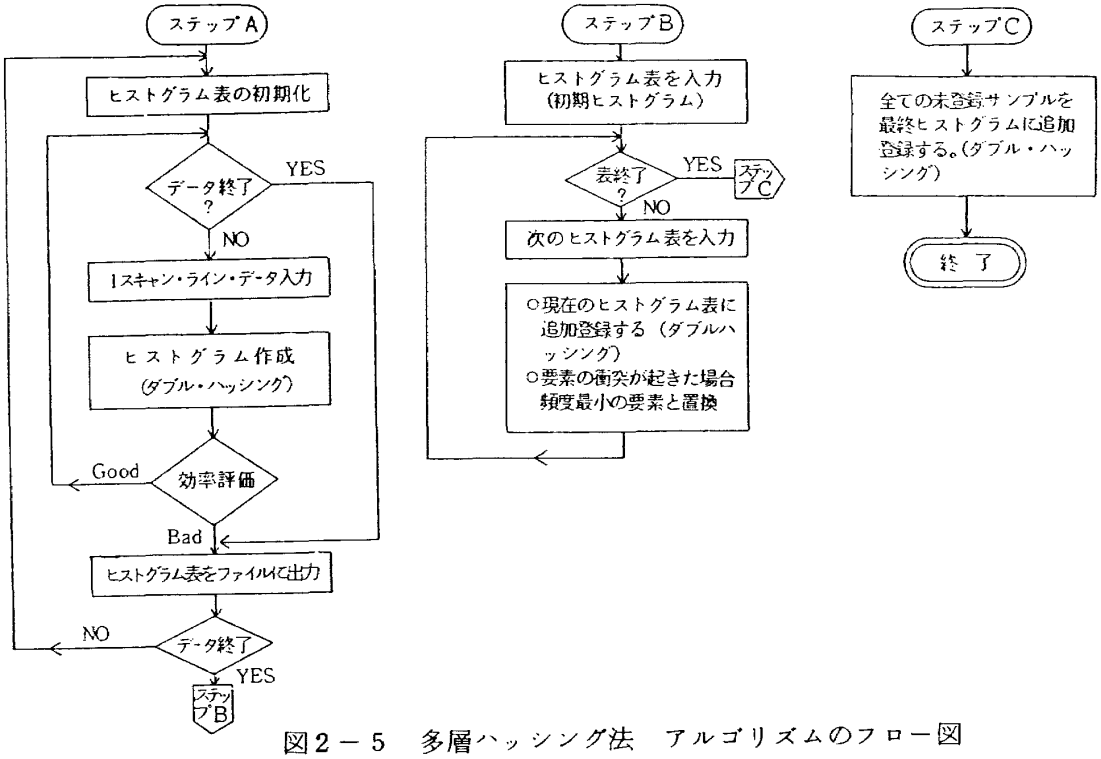


図2－5　多層ハッシング法　アルゴリズムのフロー図

[Step A]

Step A-1: 衛星画像の1スキャンラインのデータを取り，従来のハッシング法で多次元ヒストグラムを作る。ハッシングで未登録となったデータはFile Bにしまう。

Step A-2: ハッシング法の効率を計算し，一定の条件が満たされた時，Step 3に行く。
データが終了した時はStep Bへ行く。
それ以外はStep 1に行く。

Step A-3: 現在のヒストグラム層を終結しヒストグラム表をFile Aにしまう。新しいヒストグラム表を次のヒストグラム層作成のために用意する。

[Step B]

Step B-1: File Aからヒストグラム表を1つ取り出し，現在のヒストグラム表とする。

Step B-2: File Aから次のヒストグラム表を取り出し，現在のヒストグラム表にハッシング法に従って登録して行く。“衝突”の場合は衝突の連鎖上でセルの頻度を比較し頻度最小のセルを現在のヒストグラム表から取り除き，File Bにしまう。

Step B-3: File Aにヒストグラム表がなくなるまでStep B-2を繰り返す。

[Step C]

Step Bで構成したヒストグラム表にFile Bにしまわれている画素及びヒストグラムセルをハッシング法で登録する。

図2－6　多層ハッシング法のアルゴリズム

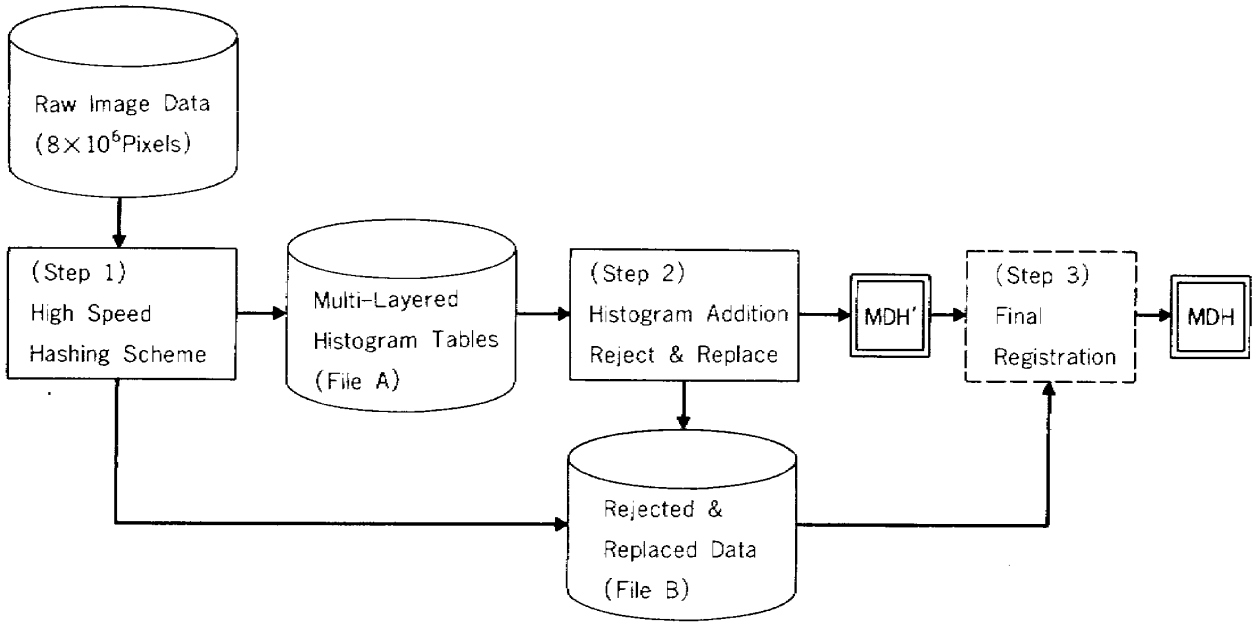


図２－７　多層ハッシング法

3. ヒストグラム・モード法

3.1 ヒストグラムモード法

本論文で述べるヒストグラム・モード法はWishart の階層的モード法の考え方を基礎にしたものである。階層的モード法は計算負荷，メモリーの点で小規模データにしか適用できなかったが，ヒストグラム・モード法では，多次元ヒストグラムを用いる事によって必要な計算負荷，メモリー量が大幅に軽減され，これにより大規模な多次元衛星画像を一度にクラスタ解析できるようになった。

3.1.1 階層的モード法の大規模化の検討

階層的モード法のアルゴリズムを図３－１に示す。階層的モード法では確率的分布状態を把握するために，各点の近傍でのデータ密度を

"k番目に近い点までの距離"

によって定義する（Step-1）。ここでkは適当に定めるパラメータで，通常２～５が用いられる。距離のリストPDを作るためには，データ数をNとして，$N(N-1)$回の距離計算及び比較計算が必要になる。又，Step-2 の並べ換え計算は本質的に不可避であり，このためには$N \cdot \log N$のオーダー の計算が必要である。更に，Step-4 のクラスタを構成していく処理でも$N(N-1)$回の距離計算及び比較計算が原理的に必要である。

このように，階層的モード法の基本的アルゴリズムは計算負荷が標本数Nの自乗に比例するという点で，大規模データのクラスタ解析には向いておらず，主として少数標本の場合に向けて考えられていると言えるであろう。

しかし，階層的モード法のアルゴリズムを更に詳しく検討した結果，多次元ヒストグラムを用いる事によって，大規模データにおいても同様の手法が導出できる事が以下の如く明らかになった。

(a) この手法において重要な役割を果たす"データ密度"とは，確率論的には確率密度関数の値と同義であるから，各点のデータ密度の指標配列PDの代わりに各点近傍でのヒストグラムの頻度値を用いても同じ結果が得られる。

(b) 多次元空間でのヒストグラムの構成方法については第２章で詳しく述べたが，その計算負荷はN^2ではなくNのオーダーである。又，ヒストグラムを構成する際に2.2節で述べた多層ハッシング法を用いれば，10^6のオーダーの標本を10^4 個程度のヒストグラムセルで表わせる。このため，Step-2のソーティングの計算負荷（$N \log N$）も大幅に軽減できる。

(c) 最後に，クラスタ分割生成のStep-4- a～cのアルゴリズムは，基本的には，クラスタの核を

Step-1: データ間の距離を計算し，各点i毎に各点からk番目に近い点の番号αiとその距離βiのリスト$PD(i)=\beta i$，$KP(i)=\alpha i$ を作る。

Step-2: 距離リストPDを距離の小さい順にソーティングする。点リストKPもPDと同じ順に並べかえる。

Step-3: $KP(1)$を第1のクラスタとする。クラスタ1の密な点のリストに$KP(1)$を加える。
$j=2$とする。

Step-4: $KP(i)$の点をaとし，PMIN$=PD(j)$とする。点aとそれまでに作られたクラスタの「密な点リスト」上の各点との距離を計算し，

(a) 全ての距離がPMIN以上なら点aによって新しいクラスタを作る。

(b) 1つのクラスタの密な点からだけPMIN以内の距離にあるならば，そのクラスタの点として「密な点リスト」に付加する。

(c) 2つ以上のクラスタの密な点からPMIN以内の距離にあるならば，それらのクラスタと点aが1つのクラスタに融合される。

Step-5: $j=j+1$としてKPが尽るまでStep-4を繰り返す。

図3－1 階層的モード法のアルゴリズム

"データ密度の高い点"に定め，データ密度がそれよりも低い点を順次その核の周りに付加していく（Step-4-b）事と，新たなクラスタの設定を，新たに付加する点がそれまでのクラスタから遠く離れている場合（Step-4-a）に行なう事からなっている。このアルゴリズムは多次元空間において，"データ密度（＝ヒストグラム頻度）"の極大点を抽出するアルゴリズムであるから，より計算負荷の小さい極大点探索手法に置き換えてもよい。

3.1.2 ヒストグラム・モード法

本手法では，前節で述べた考えに基づき，多次元ヒストグラムの形（極大点）によってクラスタを決定する。

大量のデータからヒストグラムを作った場合，ヒストグラムによってそのデータの母集団の確率密度関数を推定する事が出来る。これは，元のデータが複数の母集団からの標本が混合したものであっても同様で，この場合は混合確率密度関数がヒストグラムから推定される。各母集団の確率密度関数が単峰であると仮定すると，母集団同士の分布関数の重なりの程度が低いならば，これらM個の母集団によって作られる混合確率密度関数にはM個の極大点が現われ。Bayesの弁別境界はこれらM個の極大点の間の確率密度の谷間となる。（図1－2）（厳密には次節に述べるように若干異なる）

従って，ヒストグラムの極大及び谷間の位置によって混合確率密度関数の極大・谷間位置を推定し，これによってクラスタ分割をすれば，ほぼ最適なクラスタ分割が得られる，と言えるであろう。

ヒストグラム・モード法はこのような考えに基づき，衛星画像データの多次元ヒストグラムの極大点と谷間を探索決定する事によって，クラスタ分割を行なうものである。

3.1.3. ヒストグラムの谷間と最適弁別境界

ヒストグラム・モード法ではヒストグラム頻度の谷間によってクラスタを分割する。しかし，ヒストグラムの谷間，即ち確率密度の谷間は，厳密にはBayesの最適弁別境界にならないので，その違いを以下に例をとって示す。

簡単のために，データを1次元，2母集団とする。各々の母集団A，Bの確率密度関数$P_a(x)$，$P_b(x)$が，

(1) 三角分布

(2) ガウス分布

の2通りについて示す。

(1) 三角分布の場合：

確率密度関数$P_a(x)$，$P_b(x)$ は

$$P_a(x)=\begin{cases}0 & \cdots a_1<x\\ \dfrac{P_a}{a_1^2}(a_1-x) & \cdots 0<x\leq a_1\\ \dfrac{P_a}{a_1^2}(a_1+x) & \cdots a_1<x\leq 0\\ 0 & \cdots x\leq -a_1\end{cases}$$

$$P_b(x)=\begin{cases}0 & \cdots a_2+a_3\leq x\\ \dfrac{P_b}{a_2^2}(a_2+a_3-x) & \cdots a_3<x\leq a_2+a_3\\ \dfrac{P_b}{a_2^2}(a_2+a_3+x) & \cdots a_3-a_2<x<a_3\\ 0 & \cdots x\leq a_3-a_2\end{cases}$$

(3.1)

とする（図3－2）。その混合確率密度関数P_0は

$$P_0(x)=P_a(x)+P_b(x) \qquad (3.2)$$

であり，図3－2の点線のような分布になる。

この時，Bayesの最適弁別境界点b_0は，クラスA，Bの損失係数を等しいとして

$$\int_{b_0}^{a_1} P_a(x)\,dx=\int_{a_3-a_2}^{b_0} P_b(x)\,dx \qquad (3.3)$$

から求められ，

$$b_0=\frac{a_1 a_2(1-\alpha)+a_1 a_3\alpha}{a_2+a_1\alpha}$$

但し　$\alpha=(P_b/P_a)^{1/2}$　　(3.4)

$\alpha=1$として

$$b_0=\frac{a_1 a_3}{a_2+a_1} \qquad (3.5)$$

これに対して，ヒストグラムに現われる頻度の谷間点b_1は図3－2では

$$b_1=a_1 \qquad (3.6)$$

この差Δは

$$\Delta=\frac{(b_0-b_1)}{a_3}=\frac{a_1}{a_3(a_1+a_2)}(a_3-(a_1+a_2)) \qquad (3.7)$$

となる。従って分布A，Bの重なりが少なければ，即ち$a_3-(a_1+a_2)$が小さければ，最適弁別境界b_0と頻度の谷間点b_1の差Δは小さく無視し得る。

(2)　正規分布の場合：

2つの母集団の確率密度関数を

$$P_a(x)=\frac{P_a}{(2\pi)^{1/2}\sigma_1}\cdot\exp\left(-\frac{x^2}{2\sigma_1^2}\right)$$

$$P_b(x)=\frac{P_b}{(2\pi)^{1/2}\sigma_2}\cdot\exp\left(-\frac{(x-d)^2}{2\sigma_2^2}\right) \qquad (3.8)$$

$$P_a=P_b \quad , \quad \sigma_1=\sigma_2$$

の時以外は，最適弁別境界b_0と頻度の谷間点b_1は若干異なる。

図3－3に一次元の場合のいくつかの計算例を示す。

3.1.4　クラスタの定義

リモートセンシングにおけるクラスタ解析の目的は，観測データからある特徴をもつ代表的クラスタ

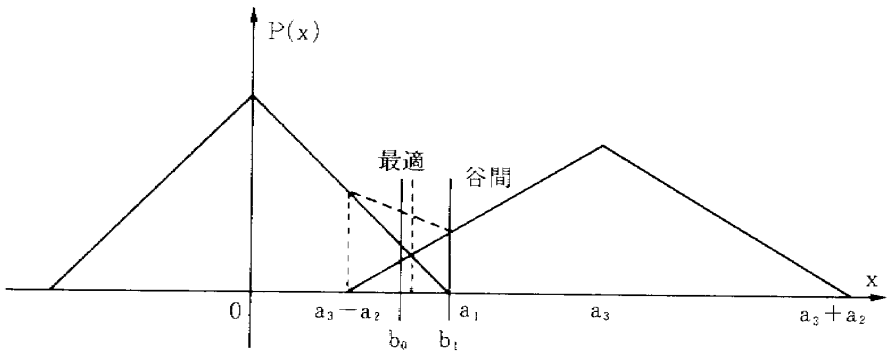


図3－2　一次元三角分布の弁別境界

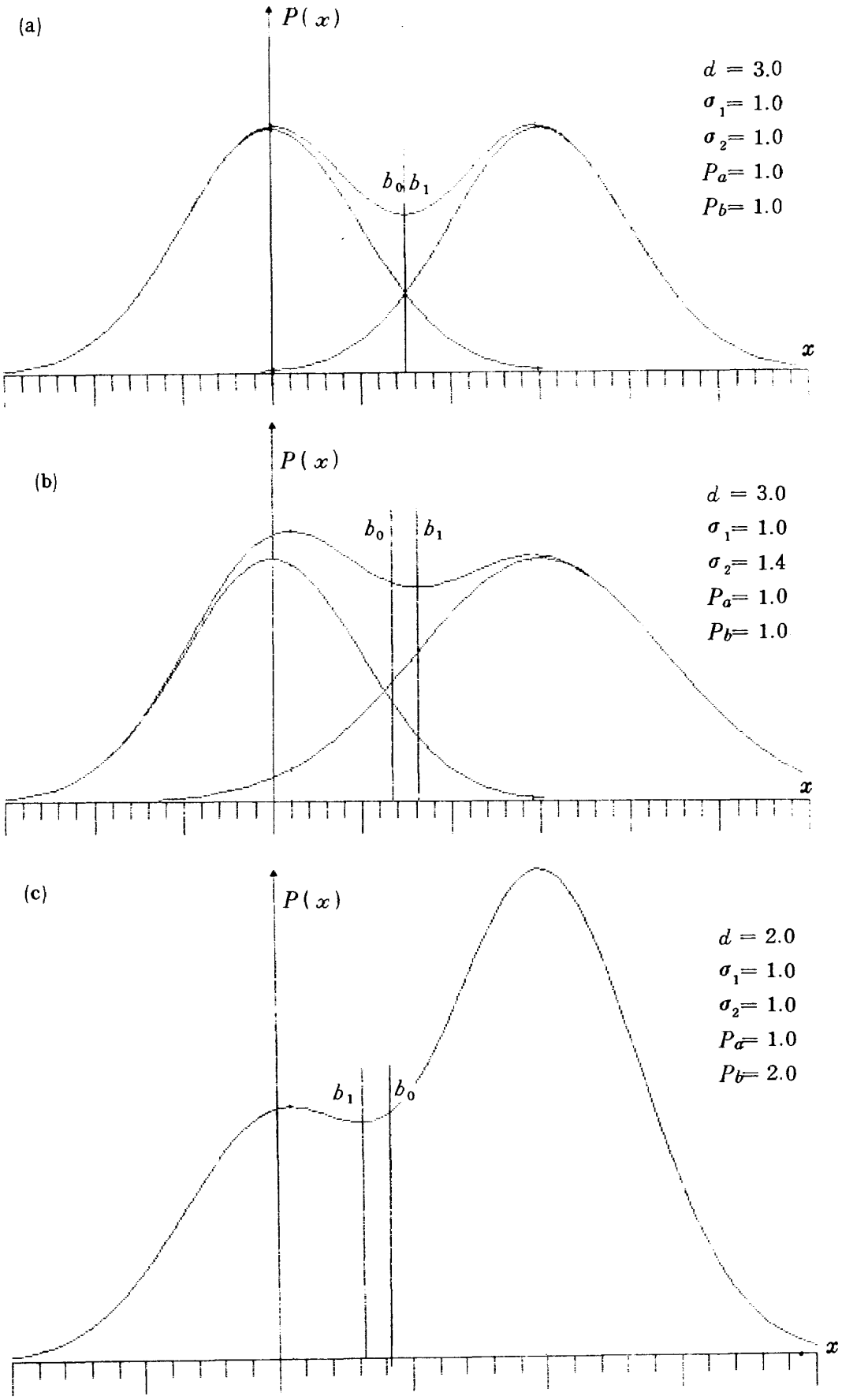


図3－3　一次元正規分布の弁別境界

を抽出し，これによって観測画像の弁別分類を行なう事であった。抽出されたクラスタが目的（リモートセンシングデータ解析）にあっているか否かは，最終的には弁別分類の結果であるクラスタ図を実験者が判定して決める。

しかし，クラスタ解析の処理過程ではこの判定は行なえないので，クラスタ解析処理が正しく行なわれているか，又，いつ処理を終了するかなどの判定にクラスタの妥当性の基準が必要である。更に，妥当性の基準が一般性にとみ，合理的なものであればあるほど，最終的なクラスタ図の妥当性が強くなり，解析処理の独自性が発揮される。

クラスタの妥当性基準としては，従来

(a)　minimum variance　基準

(b)　エントロピー基準

等の，全体的な傾向を示すなんらかの統計量を求め，それによって定量的に評価されていた。しかし，このような定量的評価によって得られるクラスタが統計論的にどのような意味を持つのか若干あいまいであった。

Wishartの階層的モード法は，この点を明確にするもので，クラスタ解析によって確率密度関数のモードを把握する事を目的にアルゴリズムが構成されている。

確率密度関数のモードとは図1－2に見られるような，確率密度の極大点を中心とし，確率密度の谷間を隣接する領域との境界とする領域である。このようなクラスタの定義は，統計論的に明確であるばかりでなく，定性的にも極めて受け入れやすいものである。

本論文で述べる，ヒストグラム・モード法においてはクラスタは，

ヒストグラムの極大点を核とし，

頻度の谷間を隣接するクラスタとの境界とする

多次元空間の領域，

と定義する。これは，確率密度関数のモードの概念を含み，さらに複雑な形状の領域をも含むように拡張された定義である。クラスタは単峰でありさえすれば十分で，正規分布以外にも三角分布や馬蹄形のクラスタもこの定義に含まれ，ヒストグラム・モード法によって抽出され得る。

この事はリモートセンシングにおいては特に重要なメリットである。衛星画像から抽出されるクラスタが，必ずしも正規分布していない可能性が指摘されているので（文献5），単純な形以外のクラスタもそのまま抽出できる事は重要である。

3.1.5　ヒストグラム・モード法の限界

ヒストグラム・モード法では，クラスタ同士をヒストグラム頻度の谷間で分割する。このことは，

仮定：N個の単峰なクラスタが観測データに存在するならばその混合確率密度関数はN峰となる，

ということが大前提として仮定されていることを意味する。

この仮定が成立しない場合には，ヒストグラム・モード法によってはクラスタの分割ができない。このような場合の若干の例を図3－4に示す。

図3－4に見られるように，クラスタ同士が近ずき過ぎ混合確率密度関数の形が高原状になった場合には，ヒストグラム・モード法ではこれを2つのクラスタに分割できない。しかし，この場合でも，第3のクラスタとの識別は行なえる。従って，クラスタ分割をしたのち，必要ならばクラスタ個々の妥当性を正規性検定等の統計的検定によって検証すればよかろう。

3.2　ヒストグラム・モード法のアルゴリズム

3.2.1　概念の定義

ヒストグラム・モード法で用いる概念を定義する。

(1)　クラスタ：“クラスタ候補”のうち“クラスタ確立条件”を満たしたものを“クラスタ”と呼ぶ。クラスタは他のクラスタと融合されない。

(2)　クラスタ確立条件：“クラスタ候補”を取り巻く谷間の深さの最小値がN_v以上である事。N_vは谷間の深さの閾値。

(3)　クラスタ候補：“孤立点”の周囲に順次拡張融合された“領域”の集合。

(4)　孤立点：多次元空間の一点で，その点が処理されるまでに構成された，どの“領域”とも“連結”しない点。孤立点の周りには新しいクラスタ候補の核となる領域が作られ，その領域だけからなるクラスタ候補となる。

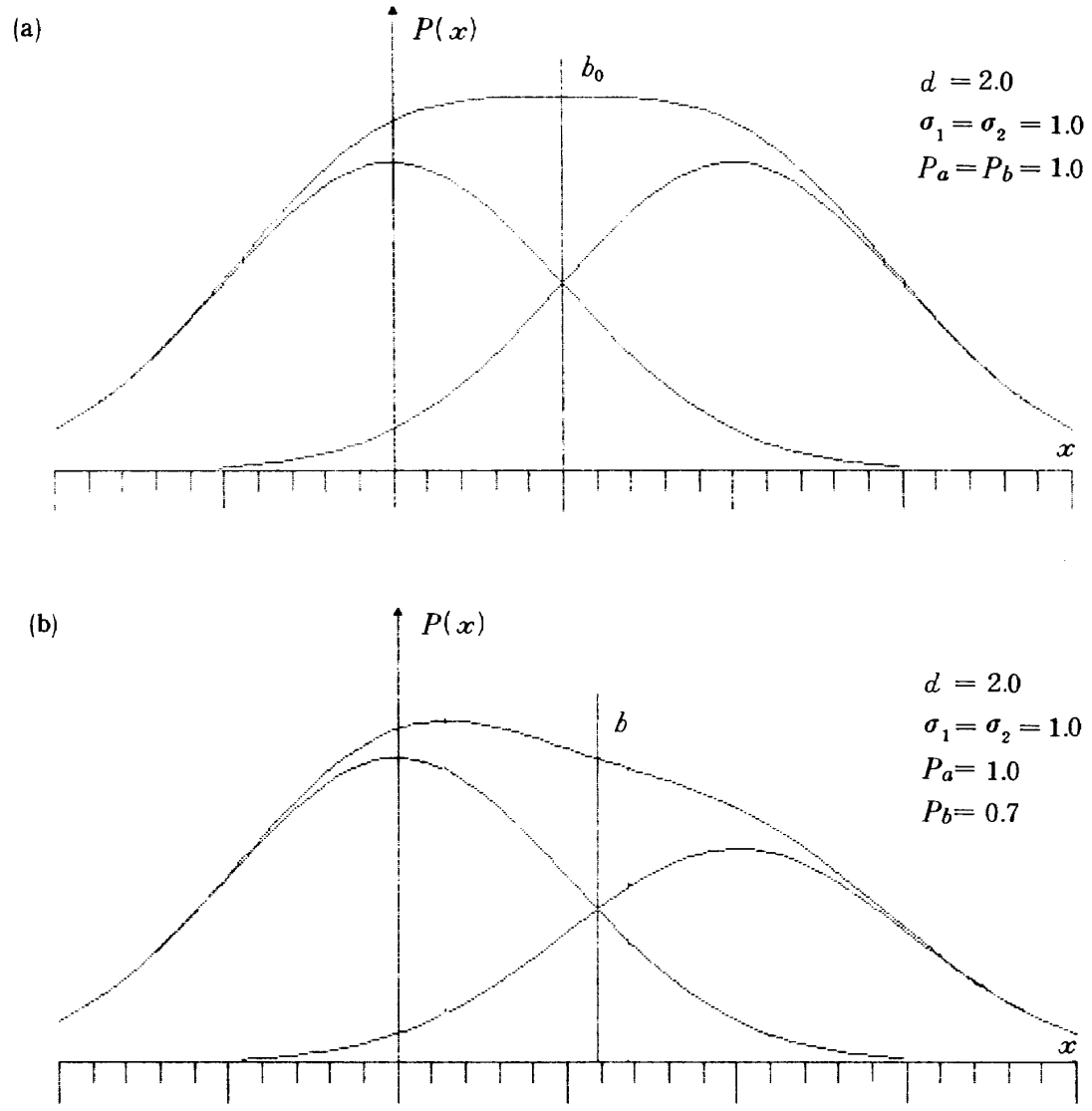


図3－4　ヒストグラム・モード法によって分割できないクラスタ例

(5) 領域：クラスタ/クラスタ候補の要素。領域に含まれるヒストグラムセルの集合に外接する，座標軸に平行な多次元直方体であり，各次元毎の上・下限 R_i^U, R_i^L で表わされる。

$$(\text{領域 I}) = [(R_1^L, R_1^U), (R_2^L, R_2^U), \ldots, (R_n^L, R_n^U)]$$

属性として“レコード”（後述）と実セル数を持つ。

(6) 連結：点Xと領域Iの連結の定義には4連結又は8連結を用いる。ある点が領域に連結されるとその領域は“再定義”される（図3－5）。

$$X = (x_1, x_2, \ldots, x_n)$$

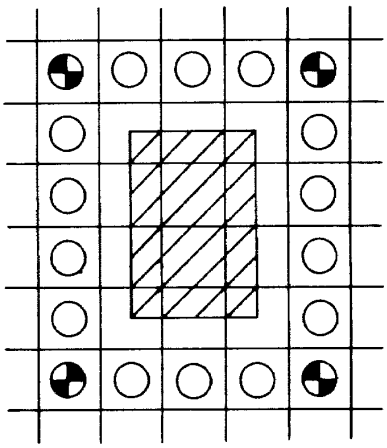


図3－5　領域と点の連結

1) 4連結；

$$\exists i \ , \ R_i^L \leqq x_i \leqq R_i^U$$
$$R_j^L - 1 \leqq x_j \leqq R_j^U + 1 \quad (j \neq i) \tag{3.9}$$

2) 8連結；

$$\forall i \ , \ R_i^L - 1 \leq x_i \leq R_i^U + 1 \tag{3.10}$$

(7) 再定義：領域にヒストグラムセル点が連結した時，その領域は再定義される。再定義には連結の仕方によって3通りある。

1) 点が領域Iに外接する時；

(a) その点を含むように領域Iを拡張しても他のクラスタ候補/クラスタの領域と交差せず，かつ領域Iの“実体積比”が閾値R以上である場合には，領域Iを拡張する。

(b) それ以外の場合は，その点だけを含む新しい領域Jを作り，領域Iと同じレコードを付ける。

2) 点が領域Iの内側にある時；

領域Iの実セル数に1を足す。

(8) 融合：クラスタ，クラスタ候補があるヒストグラムセルを介して連結した場合，連結したものの組み合わせにより以下の3種類の融合が起きる。

1) クラスタ同士の場合；

何もしない(即ち，融合は行なわない)。

2) クラスタとクラスタ候補の場合；

クラスタの核領域の属性を

クラスタ候補の核領域に書き込む。

3) クラスタ候補同士の場合；

処理レベルの新しいクラスタ候補の核領域のポインターを，処理レベルの古いクラスタ候補の核領域を指すように書き換える。

(9) レコード：その領域が属するクラスタ/クラスタ候補の最初の核領域へのポインター及びその核領域が作られた時の処理レベル値(図3－6)。

(10) 処理レベル：

ヒストグラム頻度値，

又は，ヒストグラム頻度値の変化した回数。

(11) 核領域のレコード：クラスタ/クラスタ候補の核領域のレコードのポインターにはO又はクラスタ番号が書かれている。

(12) 実体積比：領域Iの実体積比Rとは，領域の体積Vと実セル数Sの比で　$R = S/V$　で定める。

(13) 領域選択基準：セルが複数の領域と連結した時に所属させる領域を決める基準

1) その点が内包される領域で，

実体積比が最小の領域。

2) 再定義の際に拡張が可能な領域で，

実体積比が最大の領域。

3.2.2 ヒストグラム・モード法のアルゴリズム

ヒストグラム・モード法のアルゴリズムを図3－7，8に示す。又，1，2次元の場合の動作例を図

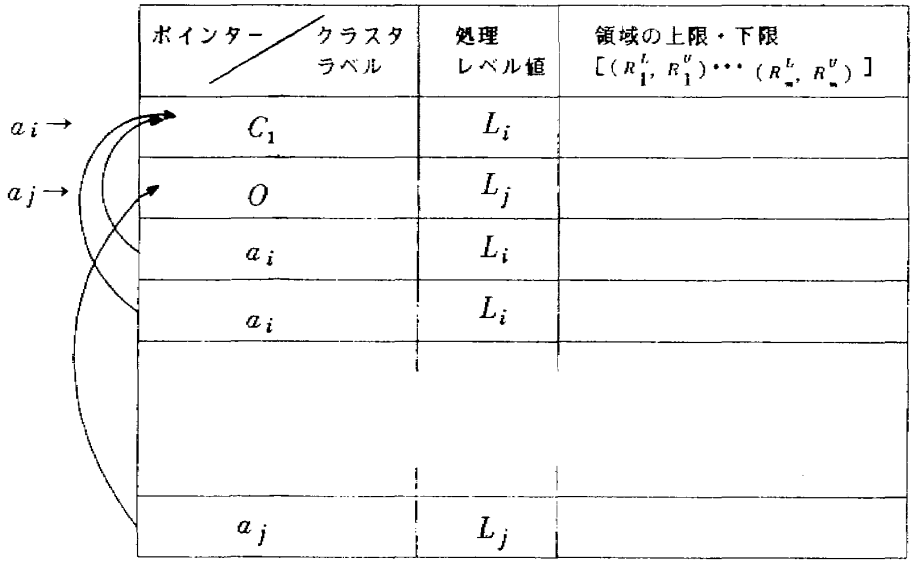


図3－6　領域表の構成

Step-1: 多次元ヒストグラム表を頻度値の降冪の順に並べ換え，スペクトル値の配列ＳＶと頻度値の配列ＨＦを作る。ＨＦ(1)は最大の頻度を示す。

Step-2: 処理レベルをＬ１とする。

$$L1 = HF(1)$$

ＳＶ(1)の点を核に領域を作りクラスタ候補とする。領域のレコードは［０，Ｌ１］となる。

ｉ＝１とする。

Step-3: ｉ＝ｉ＋１として，処理レベルを更新する。

$$Li = Li + 1 - \triangle L$$

Step-4: クラスタ候補のうち処理レベルの値１が

$$1 > Li + N$$

であるものをクラスタとし，その核領域にクラスタ番号をつける。この条件はクラスタ候補を取り巻く谷間の深さがＮ以上あることを意味している。

Step-5: 頻度値 ＨＦ(j)

$$Li < HF(j) < Li - 1$$

のヒストグラムセルに対してStep-6 Step-7を順次適用する。

Step-6: 現在処理中のヒストグラムセル点ＳＶ(j)とこれまでの処理によって作られたクラスタ候補の領域群との空間的連結関係を調べ；

☆孤立点：どの領域とも連結しない時は点ＳＶ(j)を核に新しい領域を作り，クラスタ候補とする。その処理レベル値はＬｉとする。

☆単連結：領域ｊとのみ連結した場合は，点ＳＶ(j)を含むように領域Ｊを再定義する。

☆複連結：２つ以上の領域と連結した場合，領域決定基準に従ってそれらのうち１つの領域を決め，その領域を再定義する。 連結している領域に異なるクラスタ／クラスタ候補がある場合は，これらのクラスタ／クラスタ候補を融合する。

図３－７ ヒストグラム・モード法のアルゴリズム

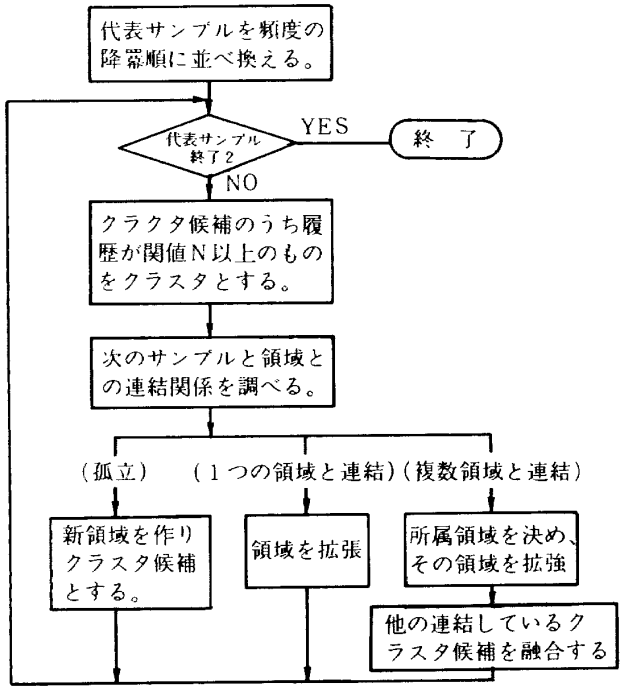


図３－８ ヒストグラム・モード法のフロー図

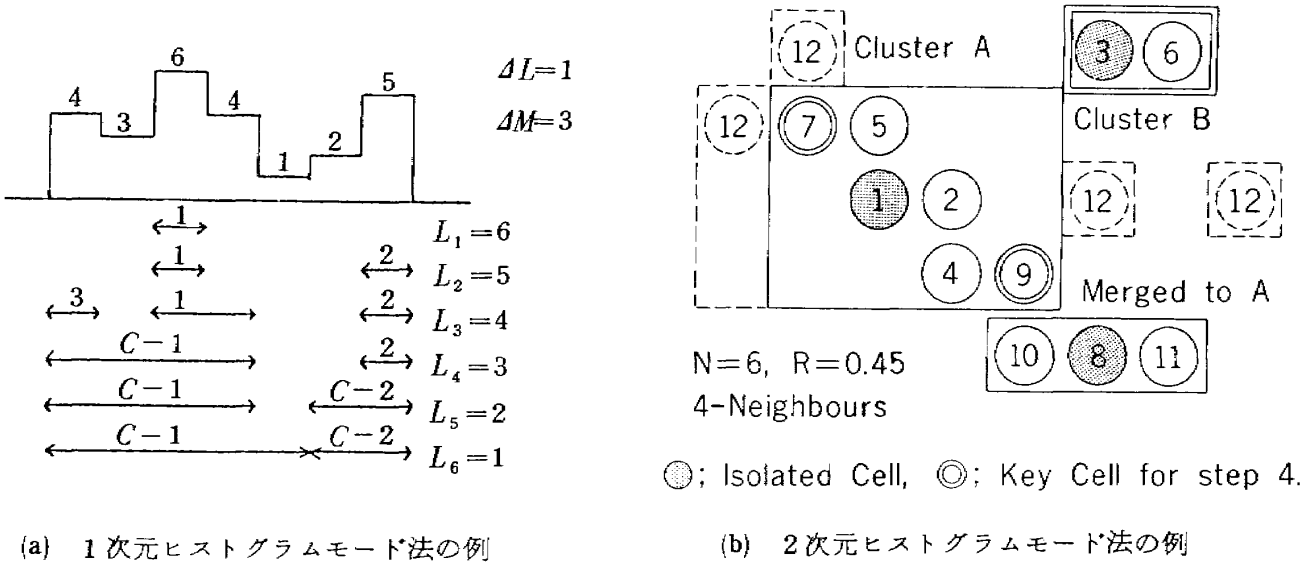
３－９に示す。

多次元ヒストグラムは2.2 節で述べた多層ハッシング法で構成されているとする。各ヒストグラムセルは図２－１(c)のように輝度ベクトルと頻度値が対になっている。

階層的モード法とヒストグラム・モード法を比較すると両手法ともに，多次元の確率密度の極大点を探索し，各々の極大点を核にクラスタを順次成長させて確率密度関数のモードによってクラスタ分割をしている点では基本的には同じである。しかし，処理するデータ量の違いから表３－１に示すような違いがある。

解析にかけるデータ量が少ない場合には，ヒストグラムを作成するよりも，階層的モード法のような形で“データ密度“を定義した方が，確率密度関数を高精度に表現できる。しかし，衛星画像のような大規模データの解析処理では，

(1) ヒストグラムによって高精度に確率密度関数を表わせる，



(a)　1次元ヒストグラムモード法の例　　(b)　2次元ヒストグラムモード法の例

図3－9　ヒストグラム・モード法の解析過程例

表3－1　ヒストグラム・モード法と階層的モード法の違い

	確率密度分布関数	クラスタの表現	データ規模	計算負荷オーダー
階層的モード法	k番目に近い点までの距離	クラスタ内の全ての点のリスト	小規模データ	N^2
ヒストグラム・モード法	ヒストグラムの頻度値	多次元直方体の集合	大規模データ	N'

(2)　個々のデータ間距離計算荷の点で不可能に近い，

という理由でヒストグラムを用いる方が優れている。

クラスタの表わし方に関しては，クラスタ内の全てのデータのリストを作成する階層的モード法の方が優れている。ヒストグラム・モード法でも同様にする事は可能であるがヒストグラムセル数が10^4程度であり，現在のアルゴリズムで10^6程度の距離計算が10^8と約100倍になるので，計算負荷の点で望ましくない。

3.2.3　ヒストグラム・モード法のパラメータ

ヒストグラム・モード法で解析結果に影響する主要なパラメータは，次の2つである：

谷間の深さの閾値　N_v，

実体積比　R　。

(1)　谷間の深さの閾値　N_v

極大点は各領域から孤立した点として検出されるが，孤立点が現われるには次の2つの場合がある：

1)　真の極大点，即ち新しいクラスタの核，

2)　ヒストグラムの標本誤差による雑音。

問題なのは後者の雑音の場合である。確率密度関数が図3－10(a)のように滑らかであっても，そのヒストグラムには必ず同図(b)のような微小極大点が標本誤差のために現われる。これらの微小極大点も当初は孤立点として検出され，数レベル後に周囲と連結してしまい，微小極大点，即ち標本誤差である事がわかる。従って，全ての孤立点を直ちにクラスタの核とする事は適当でない。

この標本誤差による微小極大点と，分布のモードによる真の極大点を区別するために，ヒストグラム・モード法では“極大点の周りの谷間の深さ”を用いる。ある極大点が標本誤差によるものであれば，その周囲の点との間の頻度の谷間はそれほど深くないであろう。しかし，その極大点が真に分布のモードであれば隣接モードとの間には十分深い頻度の谷間があるはずである。

閾値N_vは，この谷間の最小の深さを制御する。

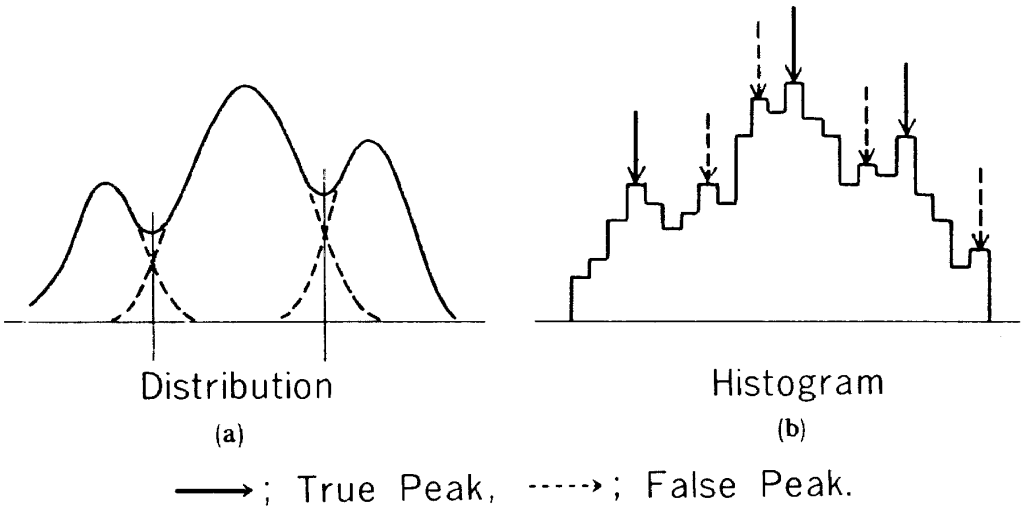


図3－10 確率密度分布関数とヒストグラム

より大きなN_v程ヒストグラムの標本誤差に強い解析が可能となるが，あまりN_vを大きくしすぎると，距離の小さいクラスタ同士が分離できなくなる。

(2) 実体積比 R

ヒストグラム・モード法では計算負荷を軽減するために，クラスタの領域を座標軸に平行な多次元直方体で表現している。このような形での領域の表現は，領域内に含まれた点の近似的表現にしか過ぎないので，領域内の点が座標軸に斜めに分布した場合，実際の点の分布状況と領域の形が掛け離れてくる。

実体積比Rは多次元直方体がどれくらい忠実に実際の状況を表わすかを制御するものである。Rが1に近いほど領域と実際の点の分布状況は一致するが，領域の数が大量に必要になり，計算負荷とメモリー量が問題となる。Rが0の場合は，最も簡略な解析が行なわれ，最小の計算負荷とメモリーで済む。

3.3 クラスタ図の解釈について

衛星画像をクラスタ解析して得られるクラスタ及びクラスタ分類図の持つ意味を考えてみる。

リモートセンシングの分類解析の目的は，1．2節で述べたように衛星画像を利用目的に応じたいくつかの分類カテゴリーに弁別し，各々のカテゴリーが地上でどのように分布しているか，どれ位存在しているかを把握する事にある。従って，クラスタ解析によって作られるクラスタ分類図が地上のカテゴリーの分布状況を忠実に表現している事が必要である。

しかし，クラスタの処理過程には解析結果をどのような目的に用いるか，という利用上の要求は直接的には反映されない。クラスタ解析では実用的目的を一般化したパラメータ（初期クラスタ，クラスタの数，距離の定義，等）によって間接的に解析の目的を表わすことができるだけである。データ利用のための分類カテゴリーとは全く無関係にクラスタ分類図が作られるとも言えるだろう。

クラスタ解析によって抽出した個々のクラスタが分類上のどのカテゴリーに対応するかは，クラスタ解析だけでは決められない。分類カテゴリーに関するグランドトルースデータを用いて，クラスタをカテゴリーに対応付ける解釈処理が，図3－11のように，クラスタ解析の後処理として必要である。解釈処理では，

(a) クラスタとカテゴリーの距離関係

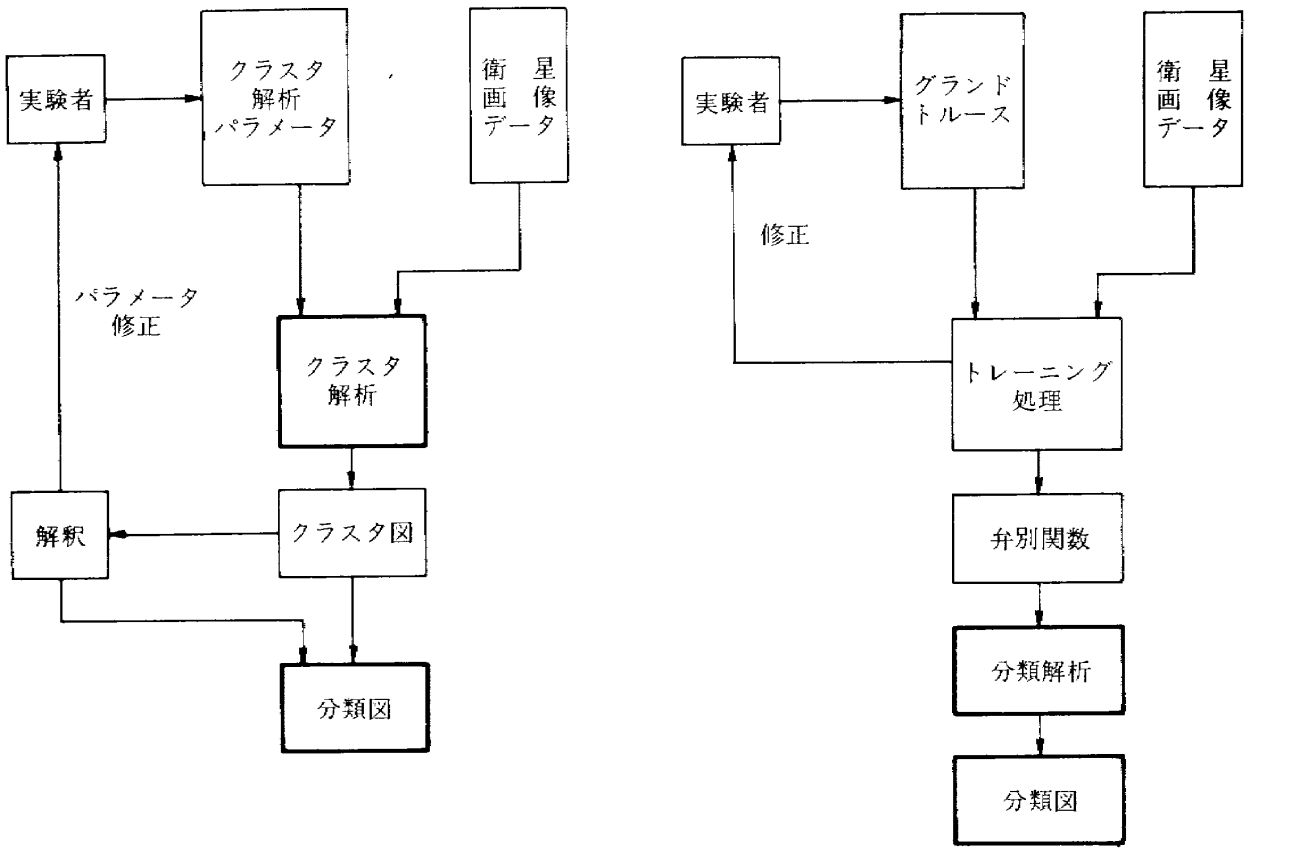
(b) クラスタ分類図とトレーニングエリアの画素毎の一致度，

などによってクラスタとカテゴリーが対応付けられる。

この時，クラスタとカテゴリーが1対1に対応すれば問題ないのであるが，通常は次のような事がしばしば起きる。

(1) 1つのカテゴリーに複数のクラスタが対応する。

カテゴリー設定が大まかであると，その中に複数のクラスタが抽出される事がある。この場合は殆ど問題ない。



（a）クラスタ解析　　（b）教師付分類

図3－11　分類解析における解釈処理過程の違い

(2) どのカテゴリーとも似ていないクラスタが存在し得る。

分類カテゴリーを十分に設定しない場合に，全てのクラスタをカテゴリーに対応付けようとすると，全く無意味な結果を導く事がある。

(3) 分光的に弁別不加能なカテゴリー設定があり得る。

カテゴリーを設定する場合，それがスペクトルとして弁別可能か否かの検討がなされない事がある。この場合，複数のカテゴリーが1つのクラスタを構成する。

第1の状況は問題ない。一般的にはクラスタ解析は分類しようとするカテゴリーの数よりも多いクラスタを生成するように，パラメータを調節するのが普通である。第2，3の状況は逆に，トレーニングデータの妥当性の検証としてしばしば用いられている。

4. 算法の並列化

ヒストグラム・モード法は，当初から多次元大規模データのクラスタ解析を目的に研究してきたものである。このため，初期においてはFACOM230/75-APUを，後期にはCRAY-1の大型ベクトル計算機を使用し，算法は全面的に並列コードで書かれている。以下，その主要な部分について例示し，衛星画像のような大規模データをベクトル処理する際の考え方について述べる。

ヒストグラム・モード法で大規模な計算負荷が発生する部分は，主に次の3ケ所である：

(1) 多次元ヒストグラムを作る際のハッシング計算，

(2) 多次元ヒストグラム表の頻度順の並べ換え計算，

(3) ヒストグラムセルとクラスタ領域の連結計算。

以下，各々の場合について述べる。

(1) 並列ハッシング

ベクトル計算機では，同時に多数の同じ計算を行なう場合に高い計算効率が発揮できる。このとき，

同時に計算するデータの集まりを計算ベクトルと称する。

衛星画像のハッシング計算で，計算ベクトル長を長くするには，計算ベクトルをスキャンライン方向に設定するとよい。

画素データvを式(4.1)で定めると，計算ベクトルは式(4.2)となる。これに対応したハッシュベクトルHVは式(4.3)で定義される。ランドサットMSSでは$d=4$なので1語当たりのbit長に納まるが，TMのように7バンドもある場合，このままでは計算できない。しかし，式(4.5)～(4.17)に示すアルゴリズムを用いれば，単精度整数演算しかできない計算機でも，式(4.4)でbit長が任意のハッシュ・ベクトルに対してハッシュ・キーを並列に計算できる。

ハッシングの繰り返しループで必要な計算は，式(4.13)とハッシュ法のStep-2（2.1.2(2)節）の部分である。これらは付図2に示すようにすれば完全に並列化出来る。但し，付図2において，点線で囲ったベクトル再構成部分は，ベクトル再構成の負荷と処理の完了したデータに対する余分な計算の負荷のバランスによっては適宜取り外す方がよい。

(2) ヒストグラム表のソーティング

表をあるキーの大きさの順に並べ換えるソーティング計算には，多くのアルゴリズムがあるが，そのいずれもが逐次計算機のための手法で，ベクトル計算機のためのアルゴリズムの報告はない。

そこで，ソーティングのアルゴリズムのうち“入れ替え”によるアルゴリズムをベクトル化したものを付図3に示す。ベクトル化アルゴリズムでは入れ替え位置を探索するステップ及びデータの入れ替えが，完全にベクトル化されるので極めて高速なソーティングが可能となる。

(3) 連結計算

ヒストグラム・モード法では頻度の順に各ヒストグラムセルと，それまでに作成されたクラスタ（候補）の領域との連結関係を調べ，これによって融合するか，新しいクラスタ候補を作るかを決める。

この計算では，各領域との連結関係の計算負荷が非常に大きい。領域の数は当初1であるが，処理が進むにつれ増加して数百程度にまでなる。これらの領域と10^4のヒストグラムセルとの連結関係を調べるので，平均して10^6回の距離計算が必要になり，並列化によって計算の効率を図らなければならない。

ベクトル計算の長さという点からは，$v\ell$個のヒストグラムセルと1つの領域の距離計算の方が効率がよいが，ヒストグラム・モード法のアルゴリズム上この並列化は行なえない。計算は，c個の領域と1つのセル点の距離計算を同時に行なうという形で並列化する。

付図4に並列コードを示す。(1)のハッシュ計算と同様に，ベクトル再構成の計算負荷と余分な距離計算の計算負荷とのトレードオフが必要である。

$$v_i = \{v_1^i, v_2^i, \cdots v_n^i\} \quad (i=1, \cdots v\ell) \tag{4.1}$$

$$CV = \{v_1, v_2, \cdots v_{v\ell}\} \tag{4.2}$$

$$HV = \{V_1, V_2, \cdots V_{v\ell}\} \tag{4.3}$$

$$V_i = \sum_{k=1}^{n} v_k^i \cdot \left\{ \prod_{j-k}^{n-1} m_j \right\} \tag{4.4}$$

$$V_i^k = \sum_{\ell=1}^{n} v_{\ell+4(k-1)}^i \cdot M^{(\ell-1)} \tag{4.5}$$

$$\begin{aligned} L &= \begin{cases} 4 \\ \mathrm{mod}(n, 4) \end{cases} \\ k &= 1, \cdots K \\ K &= \left[\frac{n+3}{4} \right] \\ M &= \max_i m_i \end{aligned} \tag{4.6}$$

$$V_i = \sum_{k=1}^{M} V_i^k \cdot \alpha^{(k-1)} \quad (i=1 \cdots v\ell) \tag{4.7}$$

$$\alpha = M^4 \tag{4.8}$$

（ハッシュキーベクトル）

$$HA^j = \{a_1^j, a_2^j, \cdots a_{v\ell}^j\} \tag{4.9}$$

$$HB = \{b_1, b_2, \cdots b_{v\ell}\} \tag{4.10}$$

$$a_i^1 = \mathrm{mod}\left\{ \sum_{k=1}^{M} \gamma_{k,i}^1 \cdot \beta_k^1, P_1 \right\} + 1 \tag{4.11}$$

表５－１　クラスタ解析のデータ

	2次元疑似データ	4次元疑似データ	ランドサット全体画像	ランデサット補正部分画像
大きさ	600×100	600×100	3228×2340	1381× 773
次元数	2	4	4	4
クラスタ数	3	10	?	?
確率密度分布関数	正規分布	正規分布	正規分布?	正規分布 ?
ヒストグラム表	997	9973	19997	9973
データ収集率(%)	100	97.3	91.4	98.4
解析精度(%)	99.2	79.6	?	?
Bayesの限界精度	99.3	91.9	?	?

$$b_i = \mathrm{mod}\left\{ \sum_{k=1}^{M} \gamma_{k,i}^{2} \cdot \beta_k^2, P_2 \right\} + 1 \quad (4.12)$$

$$a_i^{j+1} = \mathrm{mod}\ (\ a_i^{j} + b\ , P_1) + 1 \quad (4.13)$$

$$\beta_1^h = 1\ ,\quad h = 1\ ,2 \quad (4.14)$$

$$\beta_2^h = \mathrm{mod}\ (\ \alpha\ ,\ P_h\ ) \quad (4.15)$$

$$\beta_k^h = \mathrm{mod}\ (\ \beta_{k-1}^h \cdot \beta_2^h, P_h) \quad (4.16)$$

$$\gamma_{k,i}^h = \mathrm{mod}\ (V_i^k, P_h) \quad (4.17)$$

5. クラスタ解析実験・考察

5.1 データの記述

ヒストグラム・モード法によるクラスタ解析実験には表５－１に示す４種のデータを用いた。

5.1.1 疑似データ

アルゴリズムの検証に用いた疑似データは，２次元及び４次元のデータである。いずれも分布関数は正規分布を仮定し，ランダムに与えた平均，共分散を元にBox Mullerの正規乱数からデータを生成した（文献８）。

ランドサットデータをシミュレーションするために各クラスタの分布は適当に重なり合っている（図5−10，5−11）。表５－２に２，４次元の各クラスタ相互の距離関係を示す。

5.1.2 ランドサットデータ

解析に用いたデータは，Scene ID 1145-00542，1972年12月15日の東海地方のデータである。

図５－１(a)にシーン全体のバンド４の写真を示す。北部の山岳地帯には積雪が見られ，南部の伊豆半島（右下）上空には若干雲がかかっているのが見られる。

画像の全体の大きさは3,240画素 × 2,340画素であるが左右に４バンド全てのデータが揃わない部分があり，解析に利用できるのは3,228画素 × 2,340画素の領域である。全体の画素数は7.55

表５－２　疑似データの分離度（Separability）

(a)　２次元疑似データ

	1	2	3
1	—	3.5	3.6
2		—	3.1
3			—

(b)　４次元疑似データ

	1	2	3	4	5	6	7	8	9	10
1	—	2.8	1.6	2.0	2.7	3.1	2.2	3.3	2.4	3.2
2		—	2.7	2.0	1.8	2.8	3.1	2.4	2.8	2.6
3			—	1.6	2.6	3.7	2.8	3.5	2.6	2.3
4				—	2.3	2.7	2.5	2.9	1.1	2.2
5					—	2.7	2.5	3.7	2.7	3.8
6						—	2.5	2.6	2.0	3.9
7							—	4.0	2.3	4.3
8								—	2.7	2.1
9									—	2.6

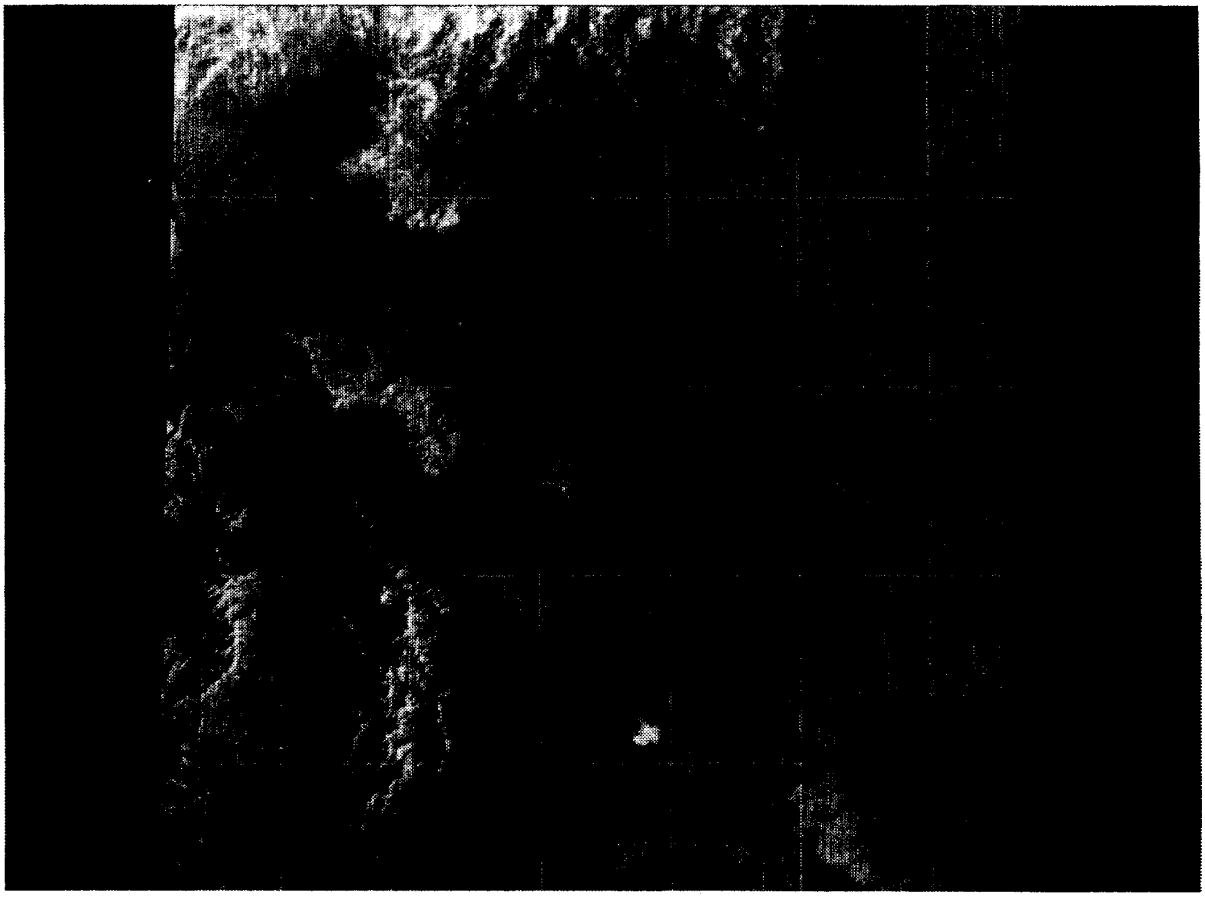


図5－1(a) ランドサットID 1145-00542 東海地方 1972年12月15日 バンド4

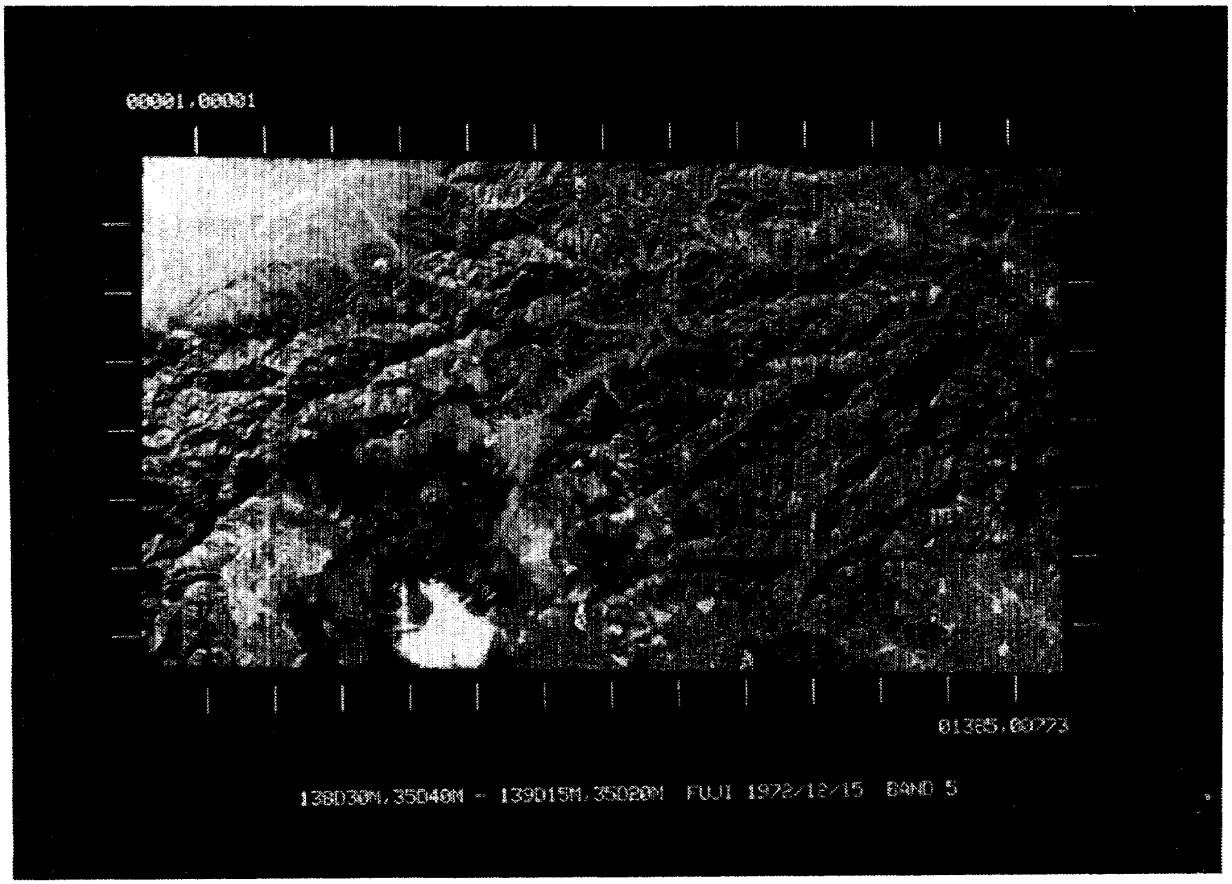


図5－1(b) 地理補正画像 （バンド5） 東経138°30′～139°15′ 北緯35°40′～35°20′

$\times 10^6$ になる。

補正部分画像は東海地域全体の画像から東経138°30′～139°15′，北緯35°40′～35°20′の地域を抜き出し，幾何学的地理補正を施したもので，1,381画素×773画素の大きさである。図5－1(b)にそのバンド5の画像を示す。

5.2　多次元ヒストグラムの構成

本節ではランドサットの画像全体のヒストグラム構成に関して述べる。多次元ヒストグラムは，4章に述べたように並列ハッシング法によって構成している。ランドサット画像の場合，2スキャンラインを単位に並列化をしており，ベクトル長は6,456になる。従って，多層ハッシング法でヒストグラム表を更新する操作は2スキャンライン毎にしか起動しない。ヒストグラム作成に用いた画像の画素数は 7.55×10^6 になる。

5.2.1　評価指標の選定及びその閾値

多層ハッシング法ではアルゴリズムStep A（図2－6）で，ハッシングの効率を常に評価しなければならない。ハッシング法の効率は1画素当たりの計算負荷で定義できる。この計算負荷は，図5－2に示すようにハッシングの探索の平均回数とほぼ完全な比例関係にある。

しかし，Step－Aは膨大な原データを取り扱う部分なのでできるだけ計算負荷を減少させることが必要である。平均探索回数は，原データの全ての探索回数が必要なので，計算負荷が大きく，評価指標として適切ではない。このため，より負荷の小さな評価指標として，

指標(1)　ヒストグラムセルの使用率（Puse）

指標(2)　そのヒストグラム表において未登録となった画素の割合（Prej）

を考える。実際のランドサットデータにおいてこれらの評価指標と平均探索回数との関係を調べたものを図5－3に示す。図5－3に見られるように指標(2)Prejは平均探索回数と線形関係にある。指標(1)Puseは必ずしも平均探索回数と線形関係にはなく，又，計算効率悪化の始まり近くで急激な変化を示している。従って，指標(2)Prejの方が効率判定には優れている。

図5－4からは，ヒストグラム表を更新するためのPrejの閾値としては，1％程度が良いと言える。Prej＝1％はヒストグラム表の95-98%が使用された状況に相当する。これ以上同じヒストグラム表に登録を続けると，効率が急激に悪化し，この後は処理するスキャンラインの大半のデータが未登録処理となることが見られる。

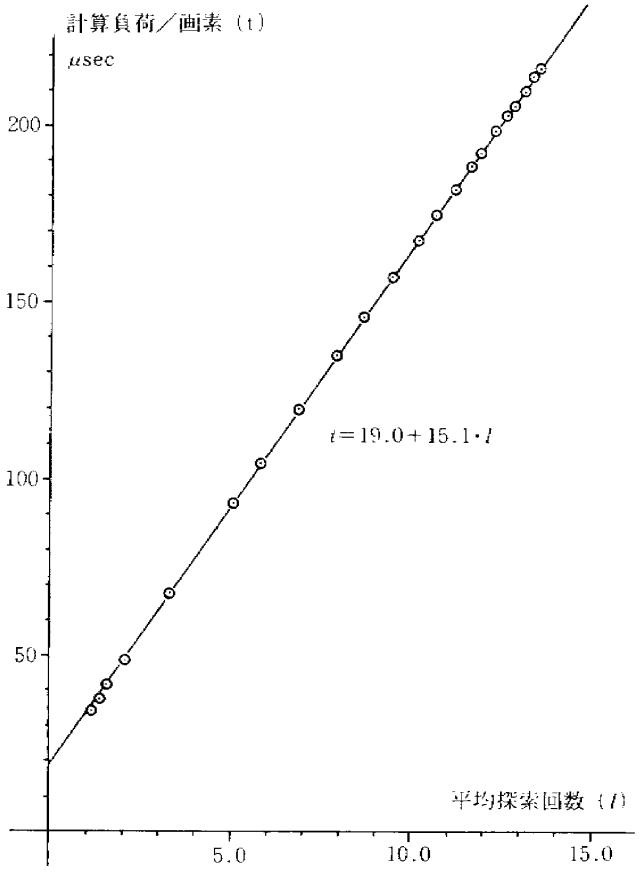


図5－2　ハッシング法の計算負荷

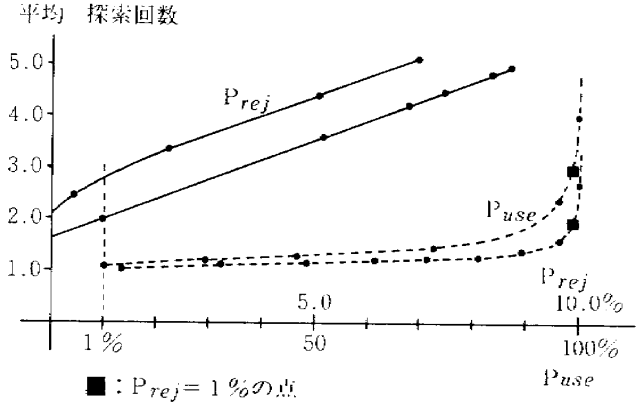
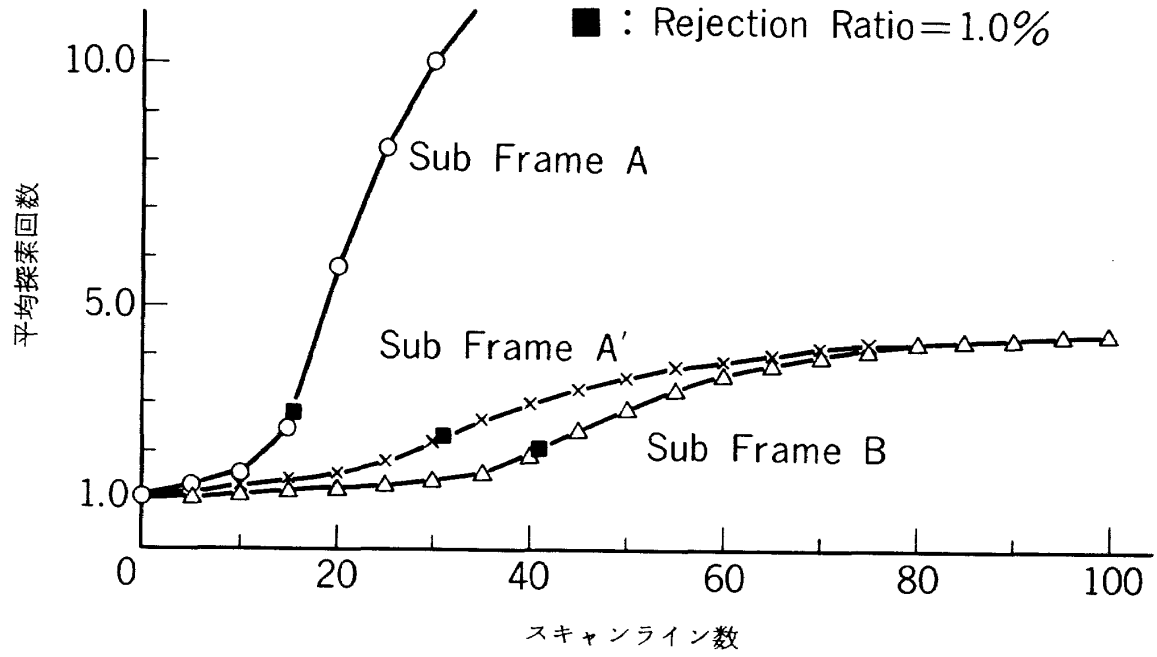
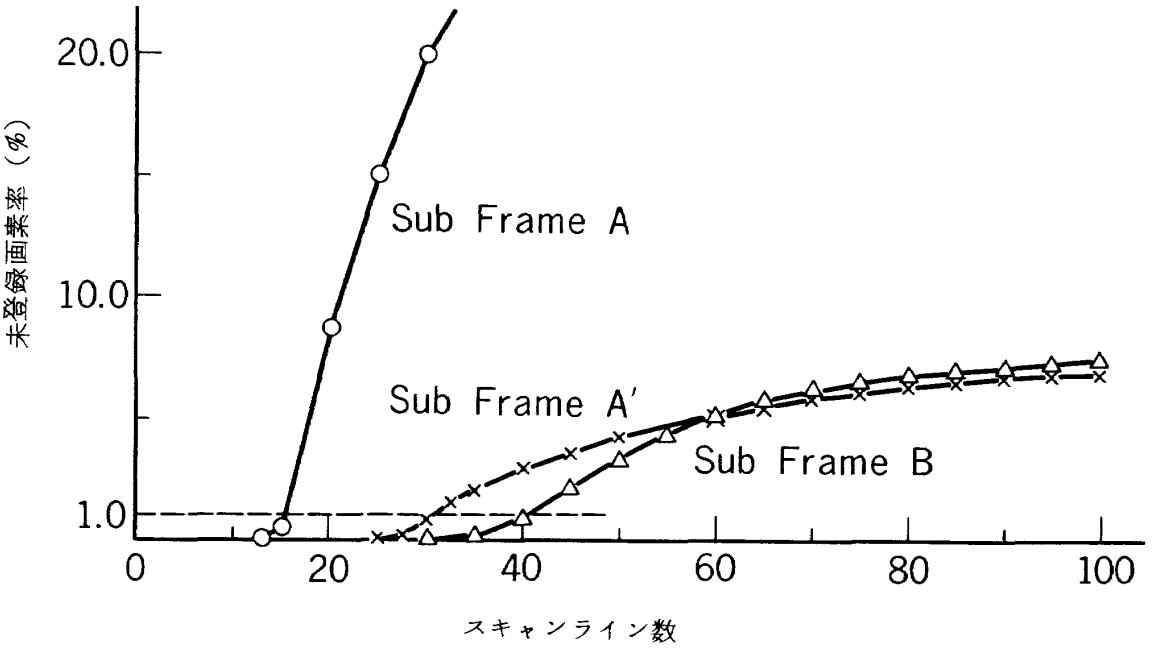


図5－3　ハッシング法の効率評価指標



(a)



(b)

図5－4 ハッシング処理の進行による効率の変化

5.2.2　多次元ヒストグラムの構成

多層ハッシング法をランドサットデータに適用し，ID.1145-00542 の東海地方のシーン全体から多次元ヒストグラムを構成した結果を表５－３に示す。Step-A，B，C各々がほぼ等しい計算負荷である。

Step-A では，４５枚のヒストグラム表に全データの98.8%が登録された。使用したヒストグラムセルの総数は約９２万個であり，各セルの平均頻度は8.1である。各ヒストグラム表毎に処理されたスキャンライン数を図５－５に示す。画像の変化の激しい部分（画面北部に相当する部分）ではヒストグラム表の更新が頻繁に行なわれているが，変化の比較的少ない部分ではより多くのスキャンラインが１枚のヒストグラム表に収容されている。

９２万個のヒストグラムセルのうちには大量に同じ輝度値を持ったセルが存在しているので，これらをStep－Bにおいて融合して１つの多次元ヒストグラムとする。Step－B では，それぞれが９８～９９％のセル使用率のヒストグラム表を融合するので，衝突が起きる確率は極めて大きい。平均探索回数は図５－６に示すように，処理が進むにつれ増加し，最終的には１２.２回にもなる。Step－Aの平均探索回数が2.0回であったのに比べると大幅な計算負荷の増大であるが，Step-B で処理する対象は1/8にデータ圧縮されたヒストグラムセルなのでStep-A とほぼ同等な計算時間で済んでいる。

表５－３　LANDSAT ID.1145-00542のシーンにおける多次元ヒストグラム作成結果

	基礎的ハッシング法		多層ハッシング法	
	収集率(%)	所要時間(秒)	収集率(%)	所要時間(%)
Step-A	65.8	1784.7	98.8	331.7
Step-B	—	—	89.6	439.8
Step-C	—	—	91.1	288.5
合　計	65.8	1784.7	91.1	1060.0

ヒストグラム表サイズ：19997
ヒストグラム表枚数：　45
効率評価開数　　：Prej＝1％

Step－Cにおいては，Step-A,Bでヒストグラム表に未登録となった画素及び，ヒストグラムセルの回収を行なう。理論的には，図2－7 File Bに含まれている画素やヒストグラムセルと同じ観測ベクトルを持つセルが，最終的なヒストグラム表に登録されている可能性があるので，Step－C がないと最終的なヒストグラムが正しい部分ヒストグラムにならない。しかし，表５－３に示されるように，そのようなデータの割合は全データの1.5％と極めて少ない。又，Step-C で処理するデータの大半

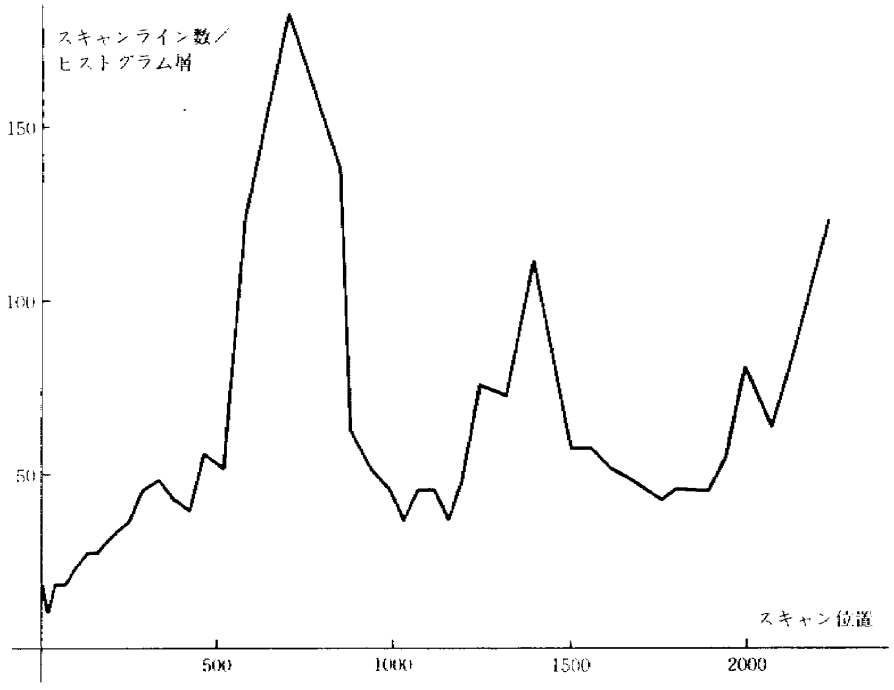


図５－５　各ヒストグラム表によって処理されたスキャンライン数の変化

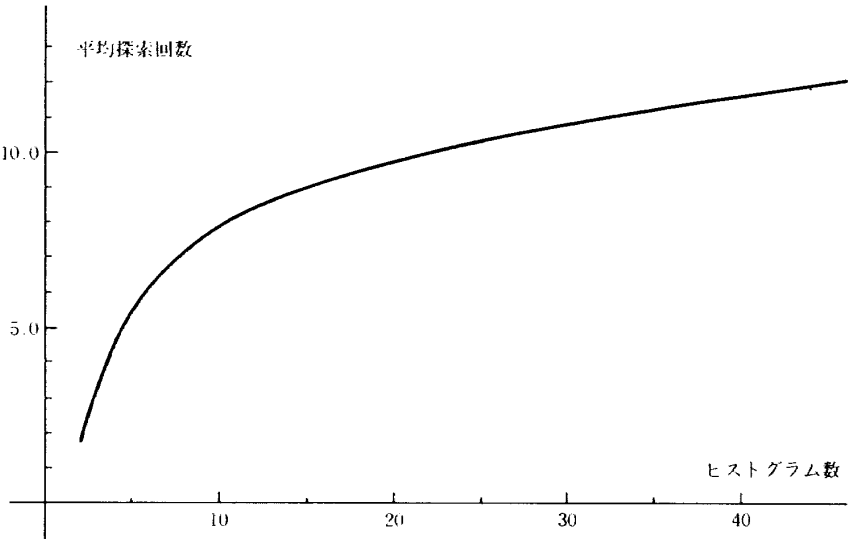


図5－6　ヒストグラムの融合にともなうハッシュ法の平均探索回数の変動

(8.9/10.4)は結局は未登録になる。この場合，ハッシングの探索は制限回数まで行なわれてしまうので，計算負荷が収集されるデータの量に比べて相当大きい。従って，コストパーフォーマンスの点からStep-Cを省き，処理速度を向上させてもあまり問題はない。

5.2.3　多層ハッシング法の効果

本節では多層ハッシング法の効果を基礎的ハッシング法と比較して述べる。

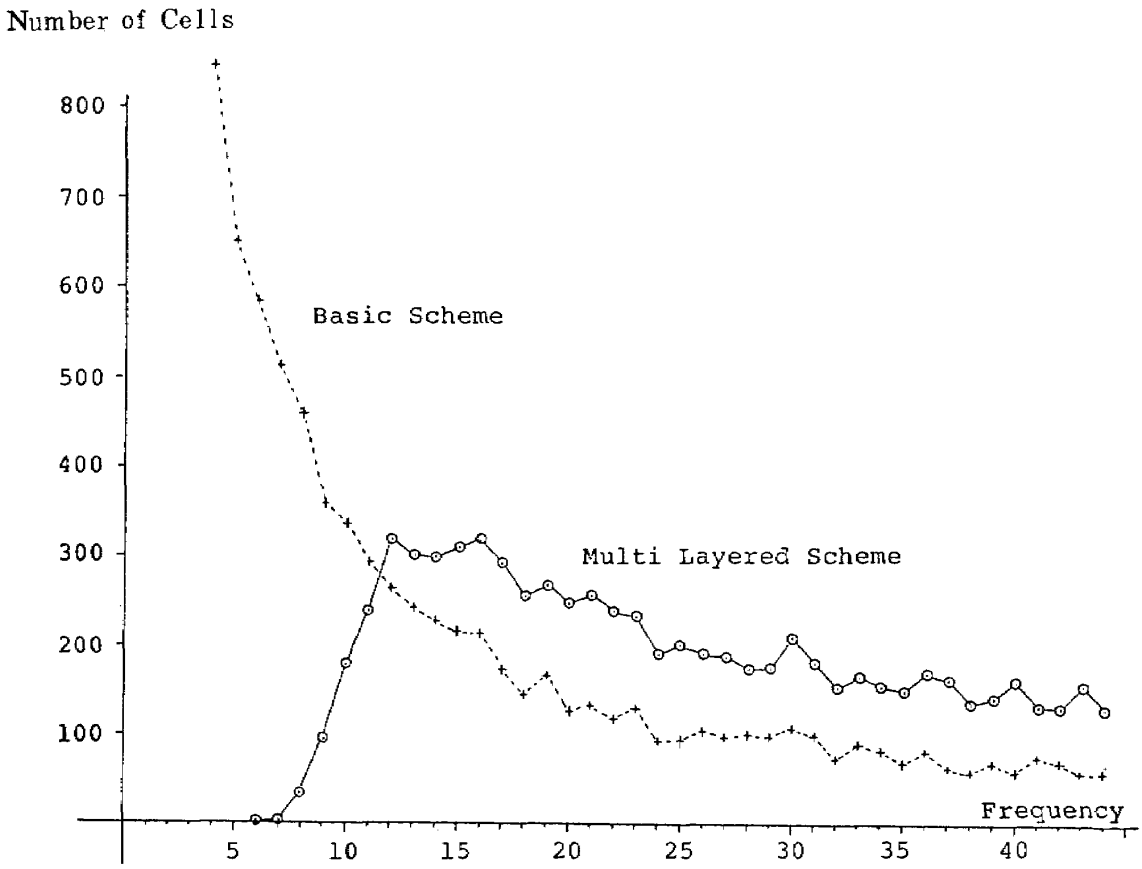
表5－3には，基礎的なハッシング法で同じランドサット画像から多次元ヒストグラムを構成した結果を示す。但し，基礎的ハッシング法ではヒストグラムセルが100%使用された際の“再ハッシッグ”処理は膨大な計算時間がかかる事が判明したので行なっていない。従って，ヒストグラムセル使用率が100%になった後は新しいセルの登録は全く行なわれない。

表5－3に見られるように多層ハッシング法は基礎的ハッシング法に比べて，全く同じ大きさのヒストグラム表を用いて，処理時間が1,785秒から1,060秒に43%短縮され，かつデータの収集率は65.8%から91.1%へと25%以上向上した。

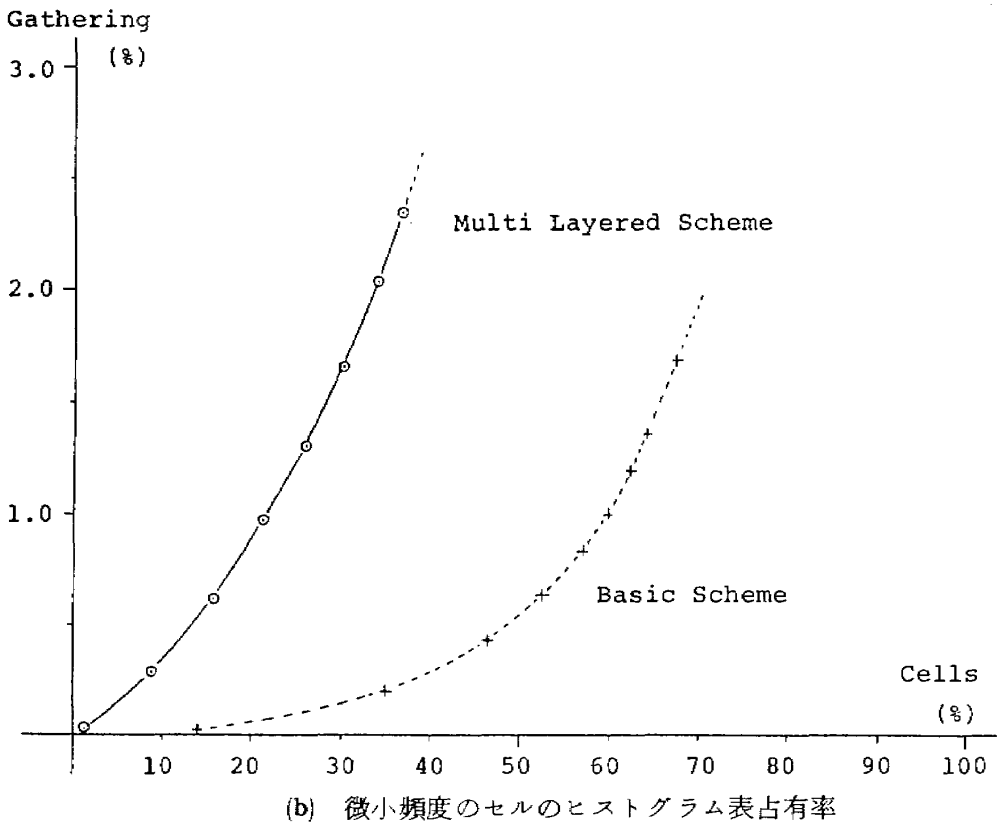
図5－7には両手法によって構成された多次元ヒストグラムのセルのうち，最も頻度数が少ない部分の違いを示した。基礎的ハッシング法では頻度5以下のセルが総セルの35%を占めているが，多層ハッシング法ではこの部分は少頻度セルとして切り捨てられており，この効率の悪いセルの代わりに頻度5以上のセルを余分に集め，ヒストグラムの裾野部分を忠実に表現している。更に，両手法によるヒストグラムの違いは，このような裾野部分のみならず，表5－4に示すように，高い頻度の所にも現われており，基礎的ハッシング法では頻度順位7番，9番という高頻度のセルが見逃されている。

図5－8には構成された多次元ヒストグラムに登録されていない画素の割合の，スキャンライン方向の変化を示す。両手法とも，画像の後半を除いて，ほぼ同じような傾向を示し定常的に20-28%の差がある。しかし，スキャン1,800-2,300にかけて両者の差は著しく拡大している。これは，図5－1に見られるように，この近辺から海が画像に現われて来るためである。多層ハッシング法では各部分画像毎にヒストグラム表を作っているので，海がここで現われた事がStep-Aのヒストグラム表に十分反映されるが，基礎的ハッシング法では既にヒストグラム表が100%使用されていて，海の輝度ベクトルを登録できず，全て未登録としてしまっている。更に，海は輝度ベクトルの変化が少ないので，多層ハッシング法では基礎的ハッシング法とは逆にこの部分のデータ収集率が他より若干良くなっている。

両手法の大きな差に，使用したランドサット画像に依存している所もある。図5－9(a)に見られるように，データの最初の100スキャン程は非常に輝



(a)　頻度の小さい部分での同頻度セル数



(b)　微小頻度のセルのヒストグラム表占有率

図５－７　多次元ヒストグラムの裾野部分における多層ハッシング法と基礎的ハッシング法の違い

度の変化が激しくヒストグラムが作り難い画像であった。そのため，初期段階で出来上がったヒストグラム表の低効率がそれ以後に波及し，全体としての効率を低下させたとも言える。

図5－9に，部分画像のヒストグラムを基礎的ハッシング法と多層ハッシング法で作った場合に，未登録となった画素の割合の変化，計算時間，及びヒストグラム作成の効率を示す。部分画像は図5－1(a)の画像から150ラインづつ帯状に切り出して作った。未登録画素率と所要計算時間の積でヒストグラム構成の効率を評価すると，同図(c)に見られるように殆どの場合，多層ハッシング法の方が高効率でヒストグラムを作っていると言える。

以上の事から多層ハッシング法について以下の事が言える。

(1) 画像に変化の激しい部分が存在する場合，多層ハッシング法の効果が顕著になる。

(2) 多層ハッシング法では，クラスタが画面のどこにあっても十分な割合で存在すれば，必ず多次元ヒストグラムに反映される。

(3) 基礎的ハッシング法ではヒストグラムの裾野の部分が十分表現されていないのみならず，頻度の大きなセルにおいても見落としがある。多層ハッシング法にはこの欠点はない。

(4) 画像の変化が少ない所では両手法とも大きな差はないが，多層ハッシング法の方が計算時間の長い分だけ結果も良くなる。

表5－4 LANDSAT ID.1145-00542のシーンにおける最頻輝度ベクトル例

順位	頻度値	輝度ベクトル	
1	17267	(12, 6, 3, 1)	
2	16497	(21, 17, 18, 8)	
3	14769	(21, 17, 19, 9)	
4	12721	(12, 4, 3, 0)	
5	12717	(12, 6, 4, 1)	
6	12125	(21, 17, 18, 9)	
7	12087	(17, 8, 4, 0)	**
8	12071	(14, 6, 4, 1)	
9	12053	(17, 8, 4, 1)	**
10	10983	(21, 17, 20, 10)	

**：基礎的ハッシング法では検出されなかった輝度ベクトル

5.3 解析実験結果及び考察

ヒストグラム・モード法によるクラスタ解析実験を5.1に述べた4種類のデータに対して行なった。

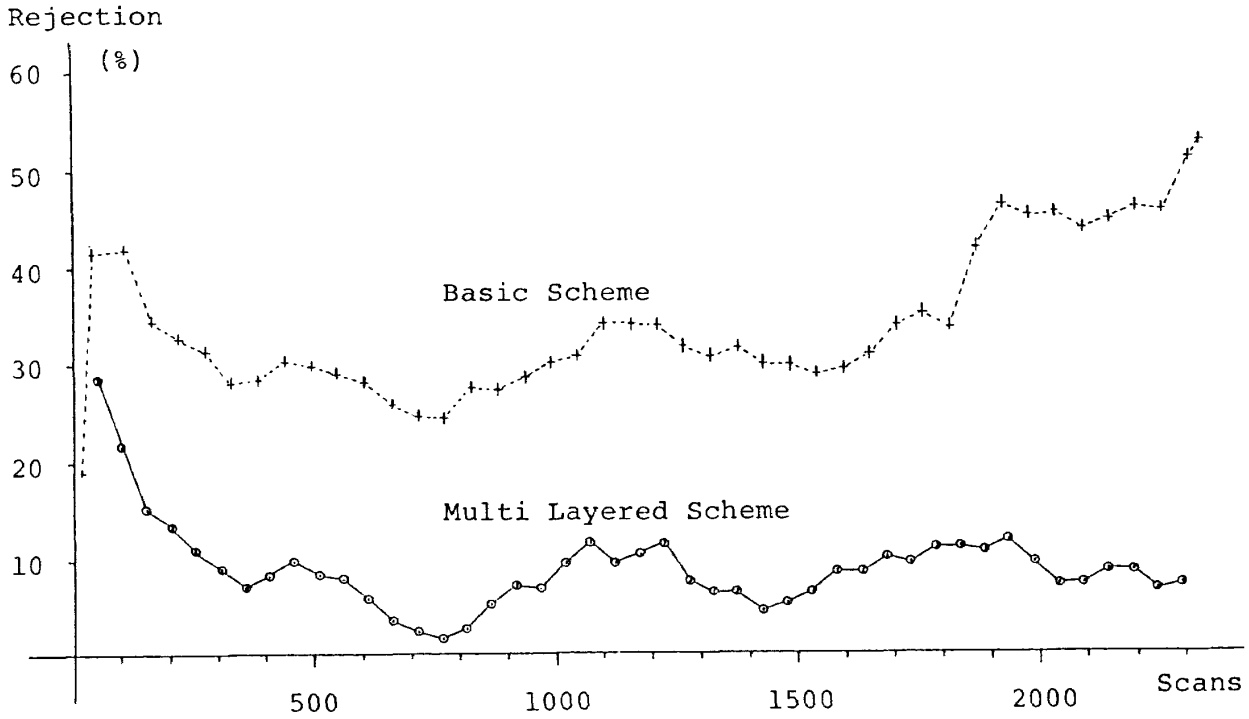
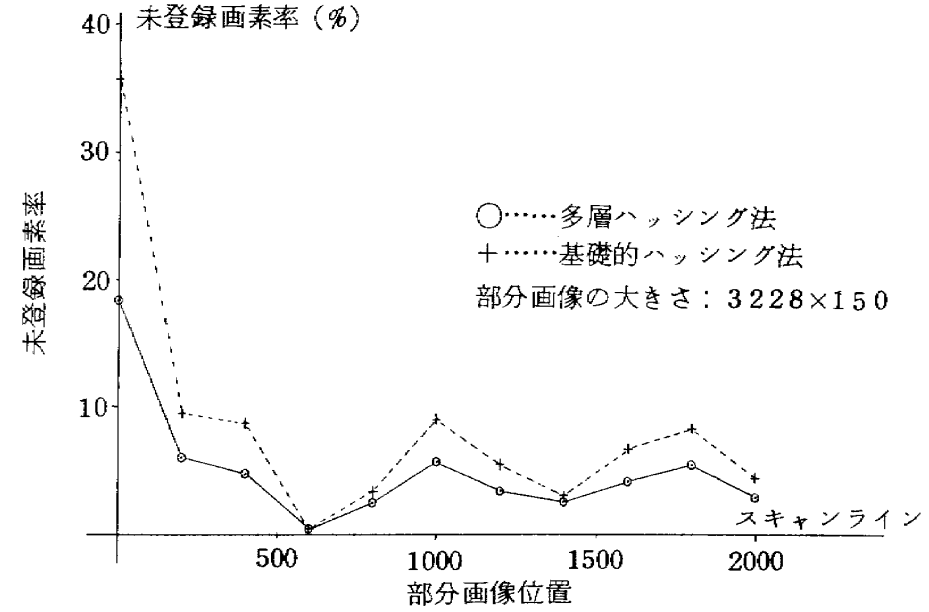
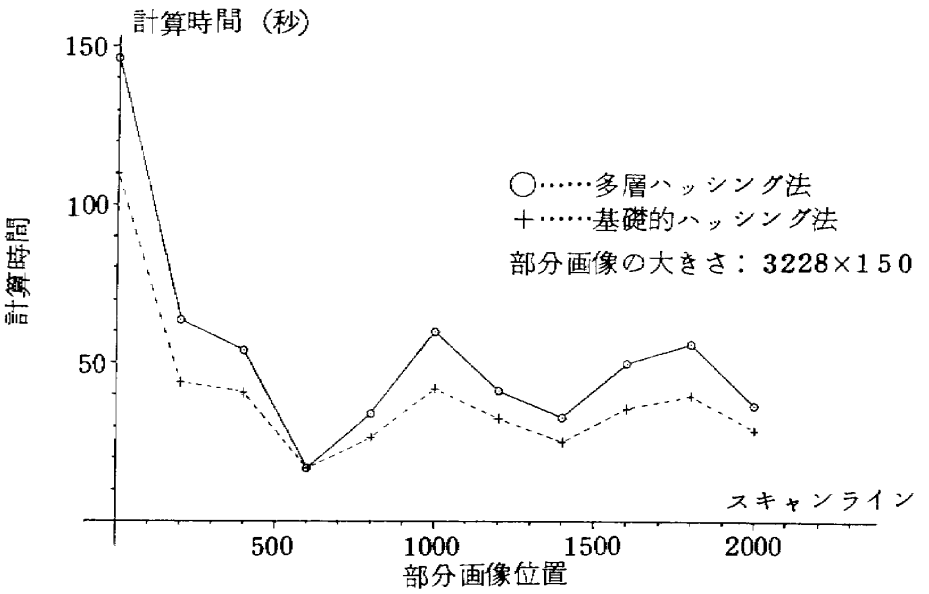


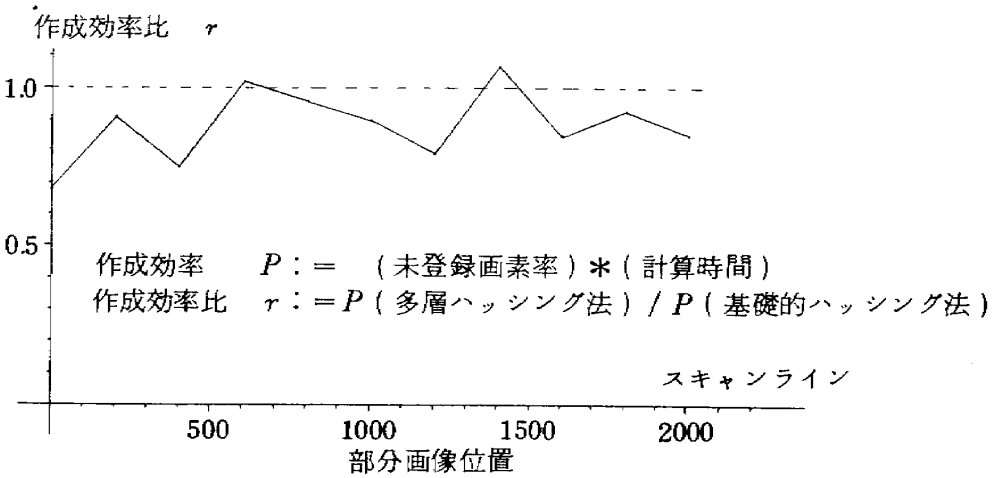
図5－8 ヒストグラムに未登録な画素の各画像位置における違い



(a)　部分画像の未登録画素率



(b)　部分画像のヒストグラム計算時間

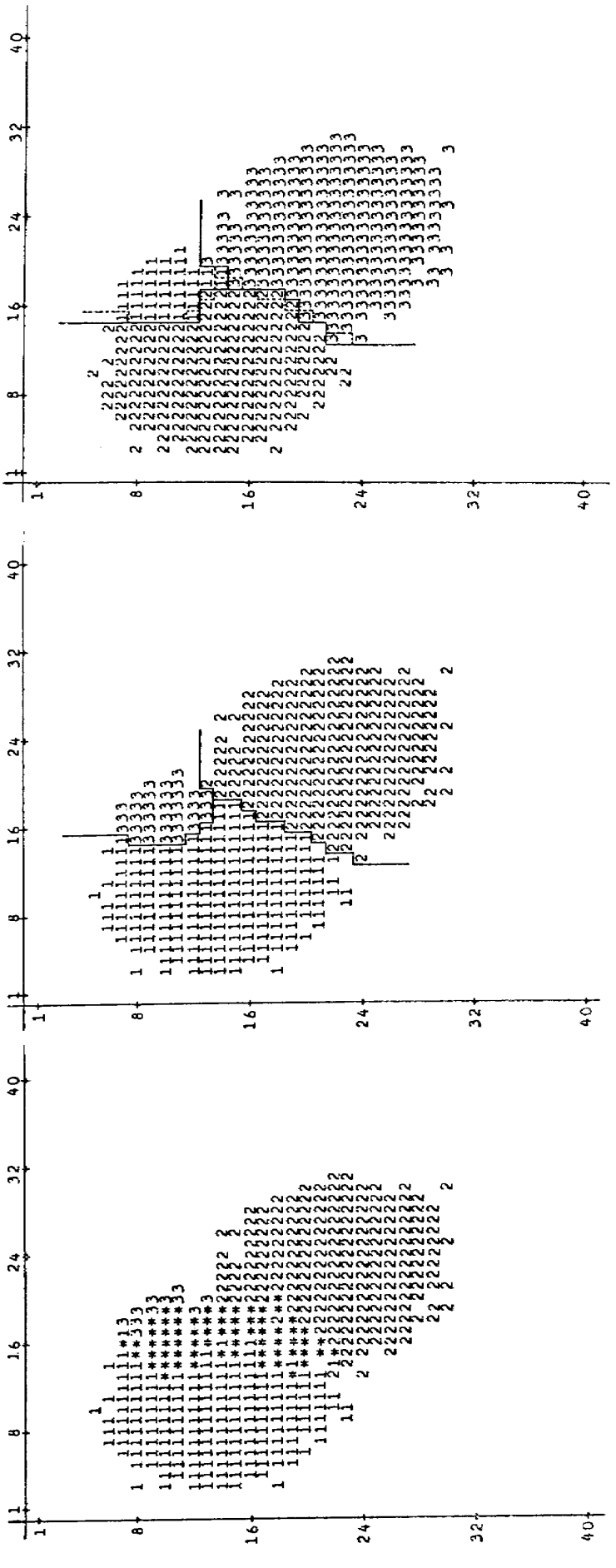


(c)　部分画像のヒストグラム作成効率比

図5－9　多層ハッシング法の効果

表5－5　LANDSAT ID1145-00542.
トレーニング・クラスの誤差行列

	トレーニングサンプル数	1 市街地	2 日陰	3 耕地1	4 耕地2	5 耕地3	6 果樹園	7 湖1	8 湖2	9 雪1	10 雪2	11 雪3	12 雪4	13 雪5	14 河川敷	15 河川	16 草地1	17 草地2	18 草地3	19 草地4	20 磨岩地	21 ササ地	22 新植地	23 裸地	24 カラマツ	25 ウラジロモミ	26 シラベ	27 草地2′	28 広葉樹1	29 広葉樹2	30 針葉樹	31 アカマツ	32 半日陰森林1	33 半日陰森林2	34 半日陰森林3	35 草地1′
1 市街地	226	86.3		1.9											3.2	3.3																			2.4
2 日陰	455		87.9						7.7																									5.0		
3 耕地1	262	3.5		65.6	2.2	1.5													5.7									0.4							2.4
4 耕地2	45			1.9	77.8	0.8											2.0	6.2									1.7								
5 耕地3	133			1.5		80.5	1.6										2.0	2.1										6.2							
6 果樹園	63					1.5	92.1															4.0						0.9							
7 湖1	72							88.9						5.0		6.7																			
8 湖2	1486		8.1						86.3					5.0																					
9 雪1	153									100	7.8																								
10 雪2	77										81.8																								
11 雪3	30											100																							
12 雪4	72												94.4																						
13 雪5	20							9.7	5.4					90.0							5.3														
14 河川敷	31	0.4													96.8																				
15 河川	30							1.4	0.1				2.8			90.0					2.0														
16 草地1	50			7.6	11.1	2.3											88.0	37.2	2.9									0.4							
17 草地2	145			1.1	8.9	5.3											6.0	44.8									1.7	1.8							
18 草地3	35			12.6														6.2	88.6								3.3	2.2							
19 草地4	40																			95.0															
20 磨岩地	150																				90.7			5.6											
21 ササ地	25																					56.0	12.0												
22 新植地	25						1.6															28.0	88.0												
23 裸地	18	8.4	0.2	1.1																	1.3			83.3											
24 カラマツ	205																								76.6							3.6	2.5	3.3	9.5
25 ウラジロモミ	114																									78.1	7.4			4.5	16.5				
26 シラベ	108																									1.8	66.7				1.7	1.8	10.0		
27 草地2′	60																	0.7		5.0								90.0							
28 広葉樹1	227					6.0	1.6											2.1											67.0	45.0					
29 広葉樹2	40					0.8																4.0							16.3	42.5	2.3		12.7		
30 針葉樹	44																									1.8				88.6					
31 アカマツ	115																									14.9	5.6				71.3		5.0	10.0	
32 半日陰森林1	55																								3.4	1.8	3.7		2.6	12.5		67.3	5.0		
33 半日陰森林2	40								0.1																0.5		10.2				2.6	5.5	50.0	16.7	
34 半日陰森林3	30																								1.5		1.9				5.2	1.8	22.5	70.0	
35 草地1′	42			5.7																					15.1							7.3			85.7
36 未分類		1.3	3.7	0.8		1.5	3.2		0.5		10.4		2.8				2.0	0.7	2.9		0.7	8.0		11.1	2.9	1.8	4.6	3.3	2.2		4.5	2.6				



(a)　2次元疑似データ分布図

(b)　Bayesの最適弁別境界

(c)　ヒストグラム・モード法によるクラスタ境界（点線はBayes境界）

図5−10　2次元疑似データのクラスタ解析

5.3.1 解析結果の解釈

クラスタ解析結果は3.3節で述べたように何らかのグランドトルースデータと突き合わせ，解釈をしなければ意味ある結果にならない。

以下に述べる解析実験では，この解釈は全てトレーニングクラスとの距離（Separability式5．1）によって決定した。

$$d=\frac{B_{12}^{t}(\mu_1-\mu_2)}{(B_{12}^{t}K_1B_{12})^{1/2}+(B_{12}^{t}K_2B_{12})^{1/2}} \tag{5.1}$$

但し $B_{12}=\left(\frac{K_1+K_2}{2}\right)^{-1}\cdot(\mu_1-\mu_2)$

クラスタにラベルを付けるため，各クラスタ分類されたデータから平均，共分散を求め，トレーニングクラスとの距離を計算し，最短距離のクラスのラベルをそのクラスタのラベルとした。

疑似データの解析結果の解釈には，合成した疑似データから計算したトレーニングクラスの平均・共分散を用いた。これは真の平均・共分散の一致推定量となっている。

ランドサットデータに対しては，真のクラス・平均・共分散全てが未知であるから，教師付き分類と同様に実験者によりトレーニングエリアを選定し，トレーニングクラスを作成した。表5－5に選定したトレーニングクラスの一覧及びその誤差行列（Confusion Matrix）を示す。

5.3.2 疑似データの解析結果

(1) 2次元疑似データ

図5－10(a)には，実験に用いた2次元データの分布図を示す。図中，数字1，2，3はそれぞれのクラスタ分布を示し，＊印の所は分布が重なっている部分を示す。同図(b)にはこのデータの確率密度関数が全て既知であるとした場合，Bayesの最適弁別境界を示す。分布に重なりがあるため，いかなる弁別境界を設定してもこれ以上の高精度は達成できない。

このデータをヒストグラム・モード法で解析した。谷間閾値$N=2$，実体積比$R=0.3$で解析した。15の領域が使用され，3個のクラスタが抽出された。

表5－6 クラスタ vs 真のクラスのSeparability距離
（2次元疑似データ）

クラスタ	真のクラス 1	2	3
1	0.009	2.251	＞ 3.0
2	2.327	0.011	2.914
3	＞ 3.0	2.937	0.004

表5－7 クラスタ vs 真のクラスの誤差行列
（2次元疑似データ）(%)

クラスタ	真のクラス 1	2	3
1	99.3	0.2	1.7
2	0.3	99.8	0.0
3	0.4	0.0	98.3

弁別精度：99.15%

解析の結果図5－10(c)に示す境界が設定された。図中，点線はBayesの最適弁別境界である。その違いは12点程度であり，かつ境界のズレは±1に止っている。表5－6には抽出されたクラスタと真のクラスの一致度をクラスタ間距離で示す。又，表5－7には解析結果の誤差行列を示す。

2次元データの解析に関してはヒストグラム・モード法は十分な解析性能を持っていると言えるであろう。

(2) 4次元疑似データ

ランドサットMSSデータを模擬した4次元データの解析を試みた。図5－11(a)には実験に用いた4次元データの各クラスタの1σ境界面の2次元平面への投影図を示す。

全データ60,000の97.4%が9,973個のヒストグラムセルに集約された。この4次元ヒストグラムを谷間閾値$N=2$，実体積比$R=0.3$で解析した所，139の領域を使用して，15のクラスタが抽出された。15個のクラスタのうち5個は要素数1の微小クラスタなので無視した。

得られたクラスタの1σ境界投影図を図5－11(b)に示す。全体の構成はよく似ており，元のクラスの形をよく推定している事が認められる。この結果を

表5－8　クラスタ vs 真のクラスの Separability 距離
（4次元疑似データ）

クラスタ	真のクラス 1	2	3	4	5	6	7	8	9	10
1	0.16		1.44	1.76			2.20		2.45	
2		0.17		1.96	1.57					
3	1.71		0.10	1.40					2.36	2.06
4		2.37	2.03	0.51					0.67	2.21
5	2.29	1.96	2.35	2.40	0.17		2.29			
6				2.45		0.08	2.30		1.86	
7	2.20						0.06		2.25	
8		2.18						0.05		2.12
9	2.36	2.14	2.01	0.71	2.46	2.16	2.42		0.49	
10			2.43	2.31				2.06		0.06

但し，空白は3.0以上を示す。

表5－9　クラスタ vs 真のクラスの誤差行列
（4次元疑似データ）　（%）

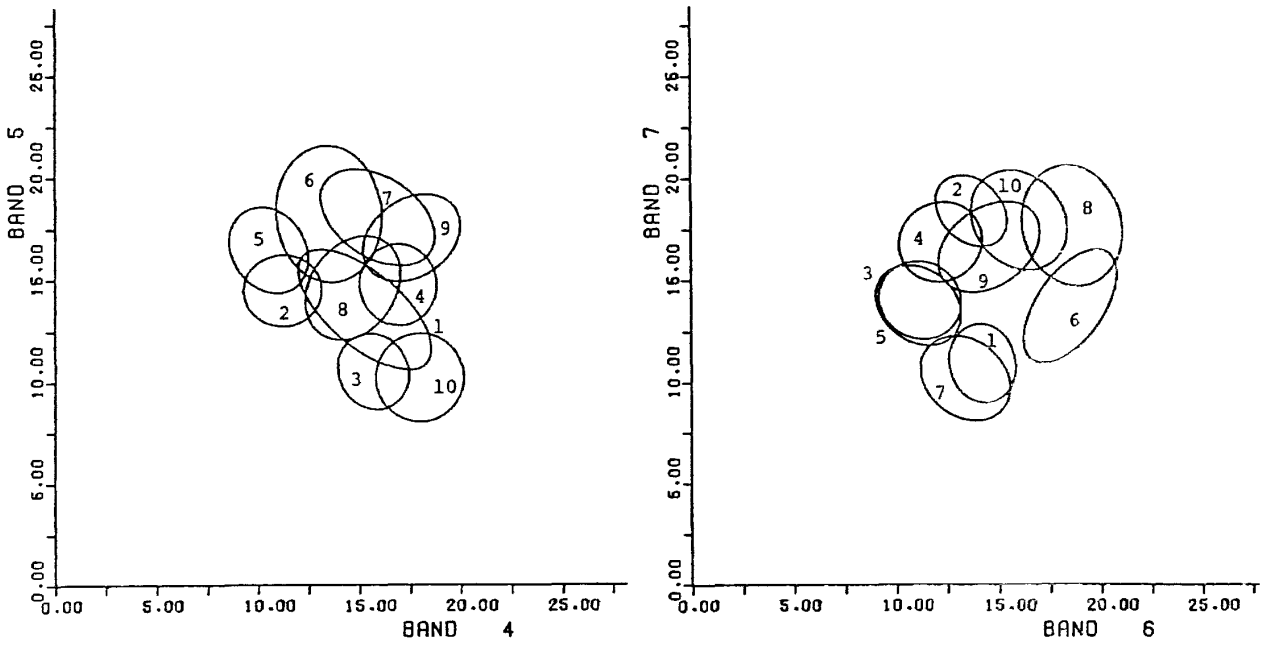
クラスタ	真のクラス 1	2	3	4	5	6	7	8	9	10
1	81.0	0.0	2.8	0.0	3.4	0.1	2.6	0.0	0.7	0.0
2	0.1	85.0	0.4	0.7	2.1	0.2	0.0	4.4	2.1	1.0
3	11.5	0.1	80.1	2.3	2.1	0.0	0.1	0.0	1.1	0.3
4	4.5	3.6	11.9	49.2	0.3	0.8	0.3	0.1	33.7	0.6
5	0.2	10.5	0.3	0.0	89.8	0.4	0.3	0.0	0.2	0.0
6	0.0	0.3	0.0	0.0	0.5	90.0	0.3	0.6	1.9	0.0
7	2.0	0.0	0.1	0.0	1.9	2.2	94.6	0.0	1.0	0.0
8	0.1	0.2	0.1	0.2	0.0	0.8	0.0	91.7	0.4	3.9
9	0.5	0.2	0.4	45.7	0.0	5.4	1.8	0.9	58.4	0.1
10	0.1	0.1	3.9	1.8	0.0	0.0	0.0	2.4	0.6	94.1

弁別精度：79.59%

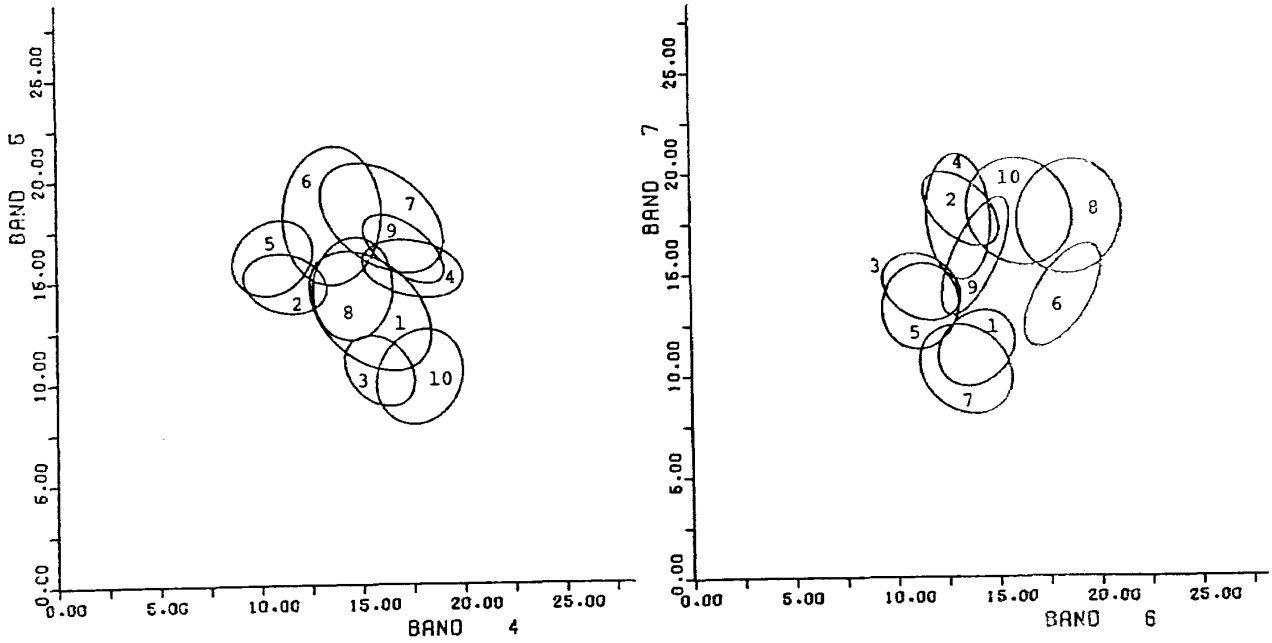
表5－10　最尤法解析結果 vs 真のクラスの誤差行列
（4次元疑似データ）　（%）

最尤法解析結果	真のクラス 1	2	3	4	5	6	7	8	9	10
1	92.5	0.0	3.6	1.3	0.1	0.0	1.0	0.0	0.4	0.0
2	0.0	95.4	0.0	1.9	3.7	0.2	0.0	0.6	0.0	0.4
3	4.2	0.0	90.0	5.2	0.3	0.0	0.1	0.0	0.0	0.8
4	1.3	1.4	5.1	77.1	0.7	0.1	0.2	0.1	13.7	1.0
5	0.3	2.2	0.2	0.8	94.3	0.2	0.6	0.0	0.1	0.0
6	0.0	0.1	0.0	0.0	0.1	97.2	0.2	0.3	1.6	0.0
7	1.3	0.0	0.0	0.3	0.6	0.6	96.8	0.0	0.7	0.0
8	0.0	0.6	0.0	0.0	0.0	0.4	0.0	97.2	0.3	1.6
9	0.3	0.1	0.0	12.3	0.1	1.4	1.0	0.2	82.9	0.1
10	0.0	0.2	1.1	1.3	0.0	0.0	0.0	1.6	0.3	96.0

弁別精度：91.91%



(a) 4次元疑似データ各クラスタの境界投影図



(b) ヒストグラム・モード法によって得られた各クラスタの境界投影図

図5−11 4次元疑似データのクラスタ解析

クラスタ間距離で評価した結果を表5－8に示す。又，解析結果の誤差行列及びBayesの最適弁別での誤差行列を表5－9，5－10に示す。

他のクラスとよく分離している5,8,10のクラスタに関してはほとんど限界に近い解析精度が得られたが，他のクラスと分離の悪いクラスタ4,9などでは解析精度は悪い。しかし，全体としてはBayesの上限91.9%に対して79.6%の分類精度が得られており，クラスタを十分正しく推定できたと言えるであろう。

ヒストグラム・モード法によってクラスタ4，9が分離しにくい原因を考えてみる。クラスタ4と9の分布の重なりの程度が問題であるが，分離度はSeparabilityで1.1（表5－2(b)）であり，これ

は両クラスタが大分重なっている事を意味する。簡単のために，1次元の場合に分離度1.1の2つの分布の重なり具合をプロットしたものを図5−12(a)に示す。同図(b)には分離度2.1（即ちクラス10の最小分離度）の場合を示す。ヒストグラム・モード法の基本的手法が，混合確率密度関数の極大点・谷間の探索であることを考えると，同図(a)のような状況はヒストグラム・モード法の限界に近く，クラスタ推定精度が低下した事は当然と言える。又，同図(b)のように混合確率密度関数が非常に明瞭に分離している場合に（クラスタ5，8，19），ヒストグラム・モード法でほぼBayesの限界に近い推定精度が得られた事は，本手法の正しさを証明していると言えるであろう。

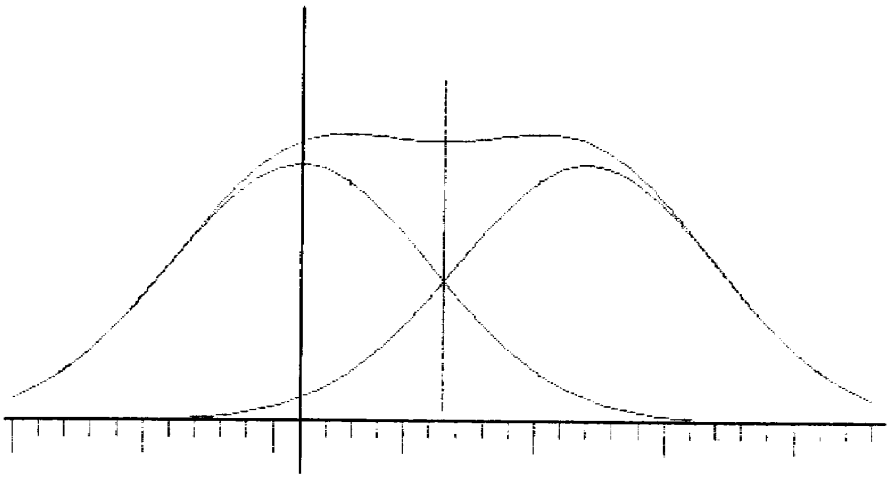
5.3.3　ランドサットデータの解析実験

クラスタ解析は東海地域の全体及びその補正部分画像に対して行なった。

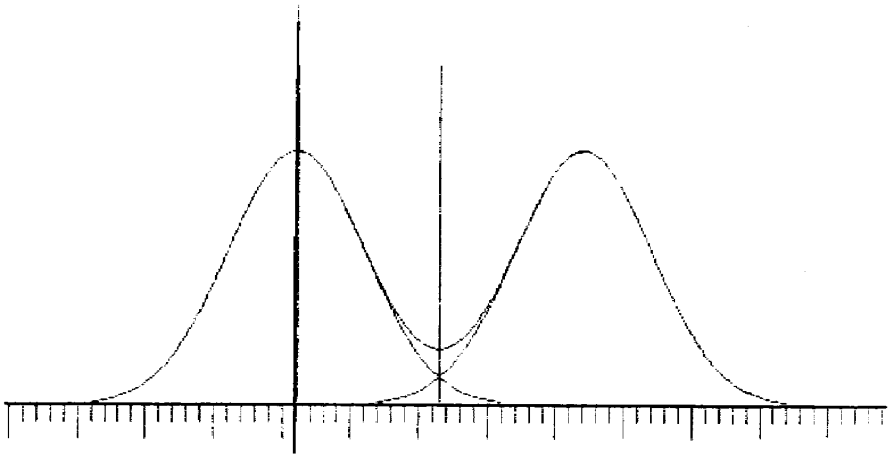
ヒストグラムはシーン全体に対しては19,997個のヒストグラムセルを用い，補正部分画像に対しては9,973個のセルを用いて構成した。ヒストグラム構成の結果を表5−11に示す。

構成したヒストグラムにヒストグラム・モード法を適用し，前述したトレーニングクラスによってクラスラベルをつけた部分画像の分類結果を図5-13に示す。又，同じトレーニングデータによって最尤法分類した結果を図5-14に参考として示す。両図を目視によって比較すると，次のような違いがみられる。

(1)　ヒストグラム・モード法の解析結果の方が，山



(a)　Separability＝1.1



(b)　Separability＝2.1

図5−12　混合確率密度関数（1次元，等分散，等先験確率）

表5−11 LANDSAT ID.1145-00542のシーンにおける多次元ヒストグラム作成結果

	全体画像 収集率(%)	全体画像 所要時間(sec)	補正部分画像 収集率(%)	補正部分画像 所要時間(sec)
多層ハッシング法 Step−A	98.8	331.7	99.1	32.3
多層ハッシング法 Step−B	89.6	439.8	98.2	5.1
多層ハッシング法 Step−C	91.1	288.5	98.4	3.3
多層ハッシング法 合計	91.1	1060.0	98.4	40.7
ヒストグラム表の長さ	19997		9973	
層数	45		2	
処理時間/画素 (μsec)	140		56	

岳地帯に谷間が入り込んでいる様子がよく分かる。

(2) 観測画像全体にわたって，ヒストグラム・モードによる画素の空間的まとめあげが見られ，最尤法分類のような個々の画素単独の分類による解析精度の劣化が抑制されている。

図5−13の結果は谷間の閾値N_vが1の時，即ち最も細かく分割した時の結果である。N_vを大きくすると，互いに近いクラスタ同士から段々に融合していき解析結果も変動して来る。谷間の閾値N_v 1，5，10，50と変えて，ヒストグラム・モード法を実行し，画素数の多い主要なクラスタについて，N_vが大きくなるにつれクラスタが統合されて行く状況のデンドログラムを表5−12(a)，(b)に示す。

$N_v=50$という大きな閾値を設定すると補正部分画像，全体画像ともに，クラスタ同士の連結（チェーニング）が進み，クラスにわたるクラスタの統合が起きて来る。連結されるクラスタのラベルを調べると，$N_v=1,5,10$では異なったクラスのラベルを持っていたクラスタ同士が，$N_v=50$で1つのクラスタに融合する事が，

〔草，カラマツ，耕地，日陰林〕

〔日陰，日陰雪，湖，溶岩地，（日陰林）〕

などのグループ内で見られる。これは，これらのグループに対して設定したトレーニングクラス間の分離度が低く，それに対応したクラスタ間に$N_v=50$という大きな谷間がない事を意味している。

$N_v=5$程度では多くのクラスタは$N_v=1$の時と殆ど同じ構成で同じクラスラベルを持つ。クラスの混合が見られたクラスタが1つあるが，この場合も$N_v=1$でのクラス決定での最短距離が1.54もあり，元々のクラス決定に無理があったものである。

$N_v=10$の場合も$N_v=5$の場合によく似ているが，クラスタの融合が更に進み，クラスにわたる混合が明らかなクラスタが1つあった。

以上の事から，ランドサットデータにおける谷間の閾値N_vについて；

(1) N_vを大きくすると，互いに近いクラスタ同士が連結されるチェーン効果が起きる。

(2) 設定したトレーニングクラスに対しては$N_v=50$は大き過ぎる。特に〔草，カラマツ，耕地，日蔭林〕及び〔日陰，日蔭雪，湖，溶岩地，（日陰林）〕などの間は確率密度の谷間がそれほど大きくない。

(3) $N_v=1\sim5$の間は余り違わない。

などがみとめられる。

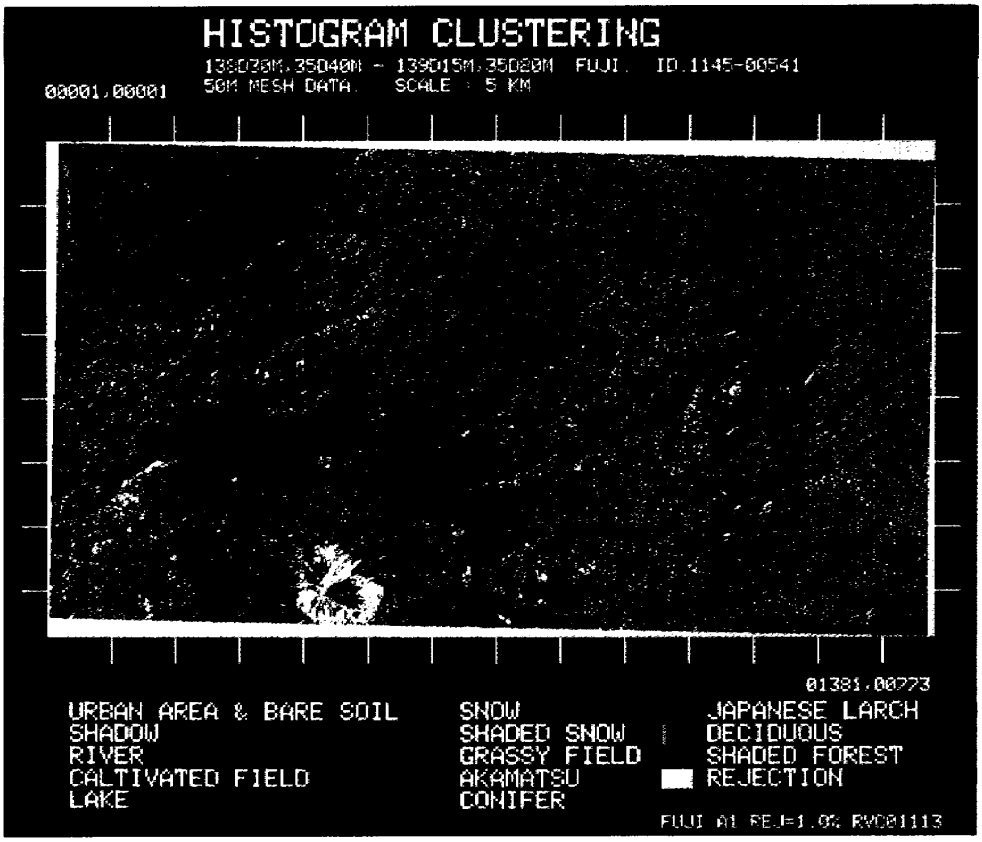


図5−13　クラスタリング解析による植生分類図

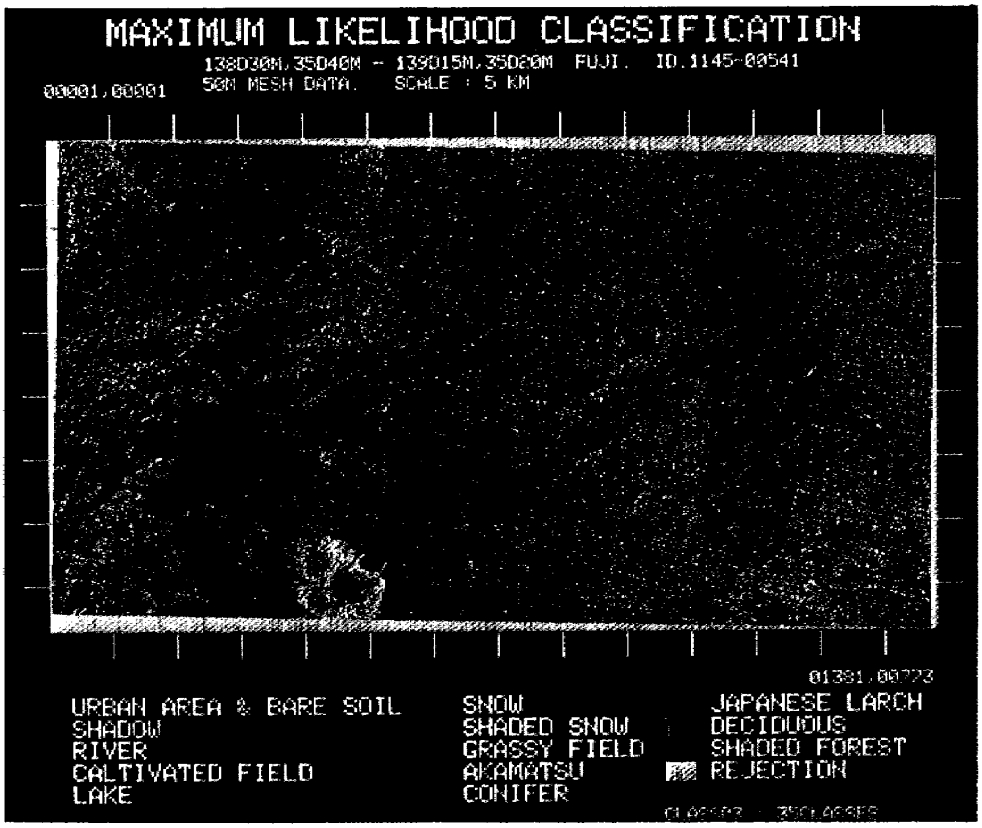


図5−14　最尤法解析による植生分類図

表5−12(a)　谷間の閾値Nによるクラスタ構成の変化（全体画像からのクラスタ）

$N=1$	−>	$N=5$	−>	$N=50$
[草 −1.47, 451539]				
[草 −0.86, 155409]	−>	[草 −1.18, 602048]		
		[草 −0.85, 284605]	−>	[草 −0.78, 889352]
[日 陰−0.73, 29818]				
[日 陰−1.32, 11148]	−>	[日 陰−0.79, 40703]	−>	[日 陰−0.79, 42999]
[草 −0.84, 123377]				
[日陰林−1.54, 38912]	−>	[草 −0.84, 160829]		
		[草 −1.07, 167696]		
		[草 −0.72, 111962]		
		[草 −0.94, 89343]		
		[草 −0.55, 57635]		
		[草 −1.60, 75163]		
		[草 −0.87, 85777]		
		[耕 地−1.19, 94353]		
		[耕 地−1.00, 28954]		
		[カラマツ−0.58, 211173]		
		[カラマツ−1.41, 59724]		
		[カラマツ−1.48, 10946]		
		[日陰林−1.31, 23044]	−>	[草 −0.34, 1176599]
		[日陰雪−1.04, 86030]		
		[日陰雪−1.45, 145102]		
		[日陰雪−0.59, 82020]		
		[日 陰−0.96, 129552]		
		[湖 −0.67, 103195]		
		[溶岩地−0.76, 129528]	−>	[日陰雪−0.69, 675427]
		[広葉樹−1.06, 110857]		
		[広葉樹−1.31, 263584]		
		[草 −1.92, 88118]	−>	[広葉樹−0.69, 467523]
		[針葉樹−1.51, 86907]		
		[日陰林−2.04, 97738]	−>	[針葉樹−1.37, 185786]

[クラス名−D, P]：D；クラスタ距離（Separability）
P；画素数

表5－12(b) 谷間の閾値Nによるクラスタ構成の変化（補正部分画像からのクラスタ）

$N=1$	−> $N=5$	−> $N=10$	$N=50$
[草 −1.27, 5943]			
[草 −1.13, 3977]			
[草 −1.30, 2797]	−> [草 −0.61, 12388]	−> [草 −0.62, 12471]	
		[草 −1.39, 6362]	−> [草 −0.61, 21338]
[広葉樹−0.59, 7528]			
[広葉樹−1.55, 2457]	−> [広葉樹−0.58, 14016]		
	[耕 地−0.75, 9709]	−> [広葉樹−0.48, 23952]	
		[広葉樹−1.64, 24316]	−> [広葉樹−0.61, 52772]
	[針葉樹−1.74, 8377]		
	[広葉樹−1.74, 12933]	−> [針葉樹−1.95, 21310]	
		[草 −1.25, 25557]	−> [草 −1.53, 46867]
[ウラジロモミ−2.50, 6860]	−> [耕 地−2.50, 6998]	−> [広葉樹−2.35, 8919]	−> [広葉樹−1.54, 16026]
		[日陰林−1.56, 28168]	
		[日陰林−0.64, 24856]	
		[日陰林−1.12, 8248]	
		[日 陰−0.19, 17355]	
		[日 陰−1.07, 15365]	
		[湖 −0.97, 15368]	
		[湖 −0.52, 29219]	
		[日陰雪−0.89, 57423]	−> [日陰林−0.92, 236546]
		[日陰林−0.56, 19837]	
		[アカマツ−0.43, 8737]	
		[ウラジロモミ−0.45, 13014]	−> [アカマツ−0.11, 35946]
		[草 −0.84, 70008]	
		[草 −0.79, 26080]	
		[草 −0.93, 11374]	
		[草 −0.86, 8938]	
		[カラマツ−0.81, 19801]	−> [草 −0.51, 129672]
		[カラマツ−1.23, 25557]	
		[カラマツ−0.42, 19751]	
		[アカマツ−0.93, 6534]	−> [カラマツ−0.45, 27999]
		[カラマツ−1.64, 7940]	
		[カラマツ−1.66, 4119]	
		[アカマツ−0.59, 7339]	−> [カラマツ−1.60, 12059]
		[草 −1.91, 14331]	
		[日陰林−1.85, 13600]	−> [日陰林−1.88, 28189]

[クラス名−D, P]：D；クラスタ距離(Separability)
P；画素数

6. 結　　論

従来，衛星画像はその大規模性のために，クラスタ解析にあたって一定のサンプリング処理が不可欠であったり，用い得る手法も限られるなど不満足にしか行なえなかった。これに対して，筆者らは衛星画像の大規模性を逆に有効利用し，いかなるサンプリング処理も必要ないヒストグラム・モード法を提案した。

ヒストグラム・モード法の基本的手法は，大規模データの多次元ヒストグラムの極大点探索，及び極大点から周囲へのクラスタ領域成長による谷間の決定からなる。ヒストグラム・モード法は確率密度関数のモードを把握する階層的モード法を以下の点で発展させたものである。

(1) 大規模データのモード解析を可能にした。

(2) 不明確な概念“データ密度”に代わり，確率密度の推定値であるヒストグラムの頻度値を用いる。

(3) N個のデータに対しN^2のオーダーを要していた計算負荷をNのオーダーに軽減した。

ヒストグラム・モード法の解析能力を示すために，2次元，4次元の疑似データ及びランドサットMSSデータの解析実験を行なった。疑似データの解析では4次元データにおいても十分な解析精度を持つことが示された。又，莫大なランドサットデータの分類解析に十分適用可能であることが示された。

更に，ヒストグラム・モード法によるクラスタ解析の基礎となる多次元ヒストグラムの効率的構成方法に関して，従来用いられていたハッシング法を改良した多層ハッシング法を提案した。多層ハッシング法は複数のヒストグラム表を動的に使用し，全体としての処理効率，ヒストグラムの忠実度を向上させるものである。衛星画像のように，各クラスタの空間分布に偏りがあるデータのヒストグラム構成には，極めて有効である事が示された。

謝　　辞

本研究の一部は科学技術庁特別研究促進調整費によって行なったものである。本研究を行なうにあたり計算センターの本間幸造技官及び㈱センチュリー・リサーチ・センターの田中正明，黒岩好夫氏に協力して頂いた。ここに記して感謝の意を表わします。

参 考 文 献

1) 奥野忠一他，「多変量解析法」日科技連．

2) M. Goldberg & S. Shlien, “A Clustering Scheme for Multispectral Images”, IEEE, SMC-8, No.2, pp. 86—92, 1978.

3) S. Shlien, “Practical Aspects Related to Automated Classification of LANDSAT Imagery Using Lookup Table” CCRS TR-75-2, 1975.

4) K.S. Fu et al., “Information Processing of Remotely Sensed Agricultural Data” Proc. of IEEE vol. 37, No.4, pp. 639—653, 1969.

5) E.P. Kan et al., “The JSC Clustering Program ISOCLS and its Application” Proc. Machine Processing of Remotely Sensed Data, Purdue Univ., 1973.

6) R.O. Duda & P.E. Hart, “Pattern Classification and Scene Analysis”, John Wiley & Sons, 1973.

7) J.C. Simon, “Some Current Topics in Clustering in Relation with Pattern Recognition” Proc. of 4th IJCPR, pp. 19—29, 1978.

8) K. Matsumoto M. Naka & H. Yamamoto, “A New Clustering Scheme for LANDSAT Images Using Local Maximum of a Multi-Dimensional Histogram” Proc. of 7th Machine Processing of Remotely Sensed Data, Purdue Univ., 1981.

9) 津田孝夫，「モンテカルロ法とシミューレーション」，培風館．

10) 地球観測画像情報技術データブック，宇宙開発事業団地球観測センター，1981.

11) P.H. Swain & S.M. Davis edited, “Remote Sensing The Quantitative Approach”, McGraw Hill, 1978.

12) G.H. Ball & D.J. Hall, “ISODATA, A Novel Method of Data Analysis and Pattern Classification”, SRI Tech. Report, April, 1965.

13) D. Wishart, “Numerical Classification Meth-

od for Deriving Natural Classes", Nature, vol.221, Jan. 1969, pp. 97—98.

14) D. Wishart, "An Algorithm for Hierarchical Classifications", Biometrics, vol.25, 1969, pp. 165—170.

15) 和達清夫他，「リモートセンシング」，朝倉書店，1976.

16) 松本甲太郎，中正夫，山本浩通，「衛星画像用大規模データのクラスタリング手法」，第24回航空宇宙技術講演会，1980.

17) 松本甲太郎，中正夫，山本浩通，「衛星画像クラスタリング・プログラムCASIM(I),(II)」，第6回リモートセンシング・シンポジウム，1980.

18) 松本甲太郎，中正夫，山本浩通，「大規模画像の多次元ヒストグラム作成手法ー多層ハッシング法ー」，第7回リモートセンシング・シンポジウム，1981.

19) 松本甲太郎，中正夫，山本浩通，「大規模多次元データのヒストグラム・クラスタ解析手法」，第24回情報処理学会全国大会，1982.

20)「LARSYS 画像処理プログラム 概説書」日本IBM，1976.

付録(1) 2.1.2節の式(2・8)$\mathbf{A}_i$と$\mathbf{A}_T$に関して

2つの2次元正規分布があるとする。

分布f_1：$\mu_1=(a,0)$，　$K_1=I$

分布f_2：$\mu_2=(-a,0)$，$K_2=I$

$$P(\omega_1)=P(\omega_2)$$

このとき，分布f_1，f_2が混合した混合分布f_0は，

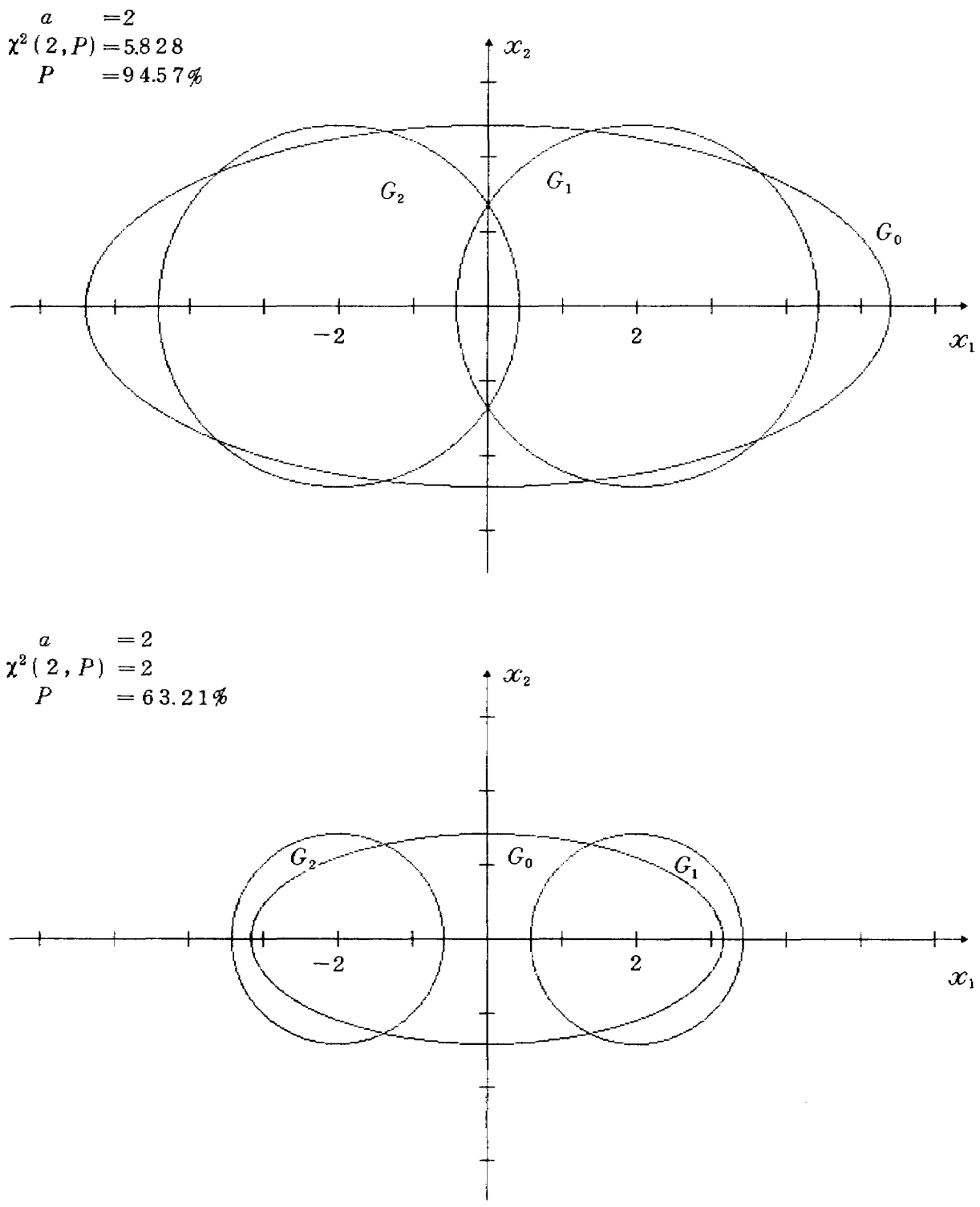
分布f_0：$\mu_0=(\mu_1+\mu_2)/2=(0,0)$

$$K_0=\frac{1}{2}(K_1+K_2)+\frac{1}{4}(\mu_1-\mu_2)\cdot(\mu_1-\mu_2)$$
$$=\begin{bmatrix}1+a^2 & 0\\ 0 & 1\end{bmatrix}$$

分布f_0の中心部分P%のヒストグラムを構成するとする。P%の部分の境界は等確率密度曲線になるように定める。このとき，各分布のP%の領域G_iは

$$G_i=\{X\mid(X-\mu_i)^t K_i^{-1}(X-\mu_i)\leq\chi^2(2,P)\}$$

で定まる。G_0とG_1，G_2の包含関係はP及び両分布の距離aによって変わるが，付図1にいくつかの例を示すように，Pが大きくなるとG_0はG_1，G_2を近似的に包含すると言えるであろう。



付図 1　混合分布と元の分布における中心部分P%領域の違い

```
C              INITIALIZE BETA COEFFICIENTS

      CALL HBETA(IPR1,IPR2,LVEC)                    ←———— (4.14)~(4.16)

C     ********** GRAND LOOP **********************************

C              GET INPUT

 5000 CONTINUE
      CALL SMPGT1(KF,DATA,IFR,NDT,NSMP,NDIM,LNG,INDX,MF)
      IF(LNG.EQ.0) GO TO 2000

      INDX($IDX) = INDX($IDX)+1
      INDX($IDX) = ICHN(INDX($IDX))

C     ******** GENERATE HASH VECTOR ***************************

   90 CONTINUE                                      ←———— (4.4)~(4.8)

       DO 200 IV=1,LVEC
      JVS = (IV-1)*NPACK+1
      JVE = JVS+NPACK-1
      IF(IV.EQ.LVEC) JVE = MZW

      NVEC(*,*,IV) = NVEC(*,*,JVS)
      IF(JVS.EQ.JVE) GO TO 200

        DO 100 JV=JVS+1,JVE
        NVEC(*,*,IV) = NVEC(*,*,IV)*NSHFT + NVEC(*,*,JV)
  100   CONTINUE

  200  CONTINUE

C     ******* CALCULATE HASH KEY "A" & "B" *********************

      LA = LVEC + 1
      CALL HSVEC(NVEC,BETA,IPR1,1    ,LA,MX,MY,LVEC) ←———— (4.11)
      LB = LVEC + 2
      CALL HSVEC(NVEC,BETA,IPR2,2    ,LB,MX,MY,LVEC) ←———— (4.12)
      NVEC(*,*,LA)    = NVEC(*,*,LA) +1
      NVEC(*,*,LB)    = NVEC(*,*,LB) +1

C     ******* HASHING SCHEME **********************************

  400 CONTINUE
      LC  = LB+1
      NDP = 1
      NDT = MXW*MYW

       DO 1000 ILP=1,NLPH

        DO 500 IDP=1,NDT
        KEY  = NVEC(IDP,1,LA)
        IF(NTBL(KEY,NHT) .EQ. 0) GO TO 420

C          ** CHECK CONFLICT **

        DO 410 IV=1,LVEC
        IF( NTBL(KEY,IV) .NE. NVEC(IDP,1,IV) ) GO TO 440
  410   CONTINUE
```

付図2　並列ハッシング・アルゴリズム

```
C           ** VECTOR MATCHING **                          <-------- Step 2(b)

      NTBL(KEY,NHT) = NTBL(KEY,NHT) + 1
      GO TO 500

C           ** BLANK ENTRY **                              <-------- Step 2(a)

  420 CONTINUE
       DO 430 IV=1,LVEC
  430  NTBL(KEY,IV) = NVEC(IDP,1,IV)
      NTBL(KEY,NHT) = 1
      NCRTB = NCRTB+1
      GO TO 500

C           ** CONFLICT **                                 <-------- Step 2(c)

  440 CONTINUE
      NVEC(NDP,1,LC) = IDP
      NDP = NDP+1

C*******NEXTKEYCALCULATION ****************************

  500  CONTINUE
      NDP = NDP-1
      IF(NDP.EQ.0) GO TO 5000
 .....................................................
 .     DO 510 I=1,LB                                 .
 .510  NVEC($JDX,1,I)                                .
 .         = IGAT( NVEC($JDX,1,LC) , NVEC($IDX,1,I) ) .
 .....................................................
      NVEC($JDX,1,LA)                                <-------- (4.13)
           = MOD( NVEC($JDX,1,LA)+NVEC($JDX,1,LB) , IPR1 )+1
      NDT = NDP
      NDP = 1

 1000 CONTINUE

C                GROUND LOOP END

      GO TO 5000

      SUBROUTINE HSVEC(NVEC,BETA,IPR,IBETA,LAA,MX,MY,LVEC)

      INTEGER *4 NVEC(MX,MY,1),BETA(LVEC,2)

C           ** CALCULATE HASH KEY "A" OR "B" **
C              FOR MULTI-WORD HASHING VECTOR

      LA   = LAA                                     <-------- (4.17)
      LB   = LA+1

       DO 340 I=1,LVEC
  340  NVEC(*,*,LA+I) = MOD( NVEC(*,*,I) , IPR )
      IF(LVEC.EQ.1) GO TO 370

       DO 350 I=2,LVEC
  350  NVEC(*,*,LA+I) = MOD( NVEC(*,*,LA+I)*BETA(I,IBETA),IPR )

       DO 360 I=2,LVEC
  360  NVEC(*,*,LB)    = NVEC(*,*,LB) + NVEC(*,*,LA+I)
       NVEC(*,*,LB)    = MOD( NVEC(*,*,LB) , IPR )

  370 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE TBSRT1(NTBL,IPR1,NENT,NPRAC)

      DIMENSION NTBL(IPR1,NENT)
      LOGICAL   LW(NPRAC)
      INDEX     $IDX/IS,IE/,$JDX/JS,JE/
      INTEGER   $IDX,$JDX


      ALLOCATE LW
      NEND= NPRAC+1
      IS  = NPRAC
      IE  = NPRAC
      JS  = MLP = IS-1
      LW(*) = .FALSE.

C.....          ** MAIN LOOP **

      DO 5000 I=1,MLP

C               Find a Position to be inserted

      LW($IDX)  = NTBL(JS,NENT) .GT. NTBL($IDX,NENT)

      IPNT      = IFBON(LW(*))
      IF(IPNT .EQ.IS) GO TO 200
      IF(IPNT .EQ. 0) IPNT=NEND

C               ** DO SORT  BY INSERTING **

      JE   = IPNT-2
      IPNT = IPNT-1

C               Block Move to insert a entry

      DO 100 J=1,NENT
      ISAVE        = NTBL(JS,J)
      NTBL($JDX,J) = NTBL($JDX+1,J)
      NTBL(IPNT,J) = ISAVE
  100 CONTINUE


  200 CONTINUE
      IS  = IS-1
      JS  = IS-1

 5000 CONTINUE
      FREE LW
      RETURN
      END
```

付図3　並列ソーティング・アルゴリズム

```
C.....      ** CHECK CONNECTEDNESS TO SOME AREA **

  400 CONTINUE

C               Check LOWER Bound

      LW(ISDX)  = RSLAND(ISDX,NLOWS).LT.DATA(1,NDPNT)+2
      K  = 2
      DO 410 I=NLOWS+1,NLOWE
      LW(ISDX)  = LW(ISDX)
     .              .AND. RSLAND(ISDX,I).LT.(DATA(K,NDPNT)+2)
  410 K = K+1

C               Check UPPER Bound

      K  = 1
      DO 420 I=NUPPS,NUPPE
      LW(ISDX)  = LW(ISDX)
     .              .AND. RSLAND(ISDX,I).GT.(DATA(K,NDPNT)-2)
  420 K = K+1

C               Get Number of Connected Areas

      LS        = IONC(LW(ISDX))
      IF(LS.EQ.0)  GO TO 500
      LSTC(LDX) = IDXL(LW(ISDX))
      GO TO 600

  500 CONTINUE

CASE ***        Isolated Cell

C.............................................
C               Make a New Area              .
C.............................................
  600 CONTINUE
      IF(LS.GT.1) GO TO 700

CASE ***        Connected to only 1 Area

C.............................................
C               Merge the Cell to the Area   .
C.............................................

  700 CONTINUE

CASE ***        Connected to Some Areas

C.............................................
C               Decide the most suitable Area & .
C               Merge the Cell to the Area   .
C               Merge the Areas if possible  .
C.............................................
```

付図4　領域とセルの連結計算の並列化

航空宇宙技術研究所報告854号

昭和60年3月発行

発行所　航空宇宙技術研究所
東京都調布市深大寺東町7丁目44番地1
電話武蔵野三鷹(0422)47-5911(大代表)〒182

印刷所　株式会社　実業公報社
東京都千代田区九段南4－2－12

Printed in Japan