統計数理 (1993)
第41巻 第2号 149-167

混合分布モデルを用いた分類法とデータ構造の色彩表示

——LANDSAT画像データの解析——

総合研究大学院大学* 中村永友
統計数理研究所・総合研究大学院大学 小西貞則・大隅昇

(1993年12月 受付)

1. はじめに

地球観測衛星LANDSATから送信される画像データの解析技術の進歩によって，地表上の植生分布，植物の活力度，海洋，河川，湖沼の汚染状態，農作物の作柄状態などが，広範囲にわたって把握できるようになってきた．このような画像データに内在する情報を抽出するとき，“分類”は本質的な操作で，実際に数多くの分類手法が提案されている．

現在広く利用されているLANDSAT 5号のセンサTM (Thematic Mapper) の解像度（分解能または瞬時視野のことで，1画素の大きさ）は30 m×30 m，また，MSS (Multispectral Scanner) は83 m×83 mであり（宇宙開発事業団地球観測センター 編 (1990)），日本のような複雑に入り組んだ土地利用状況を分析するには決して解像度が高いとはいえない．つまり1つの画素中にいくつもの対象物が混入し，その画素を特定の対象物として分類・識別することには限界がある．

LANDSATの画像データの特性は，ディスプレイ上に表示される画像の各画素の位置を表す座標と，いくつかの波長帯の観測値（多重分光の輝度値）から構成される（観測値により作られる空間を“特徴空間”(feature space) と言う)．従来の画像データの分類法には，この観測値のみを用いる手法や，これに座標を加えたものを用いる手法がある．これらの分類法は地表上の細かい部分を識別することが目的で，つまりクラスとして水田，畑，住宅地，道路，市街地などの特定の“対象物”に対する分類精度の向上を，主な研究課題としてきた（ここで“クラス”とは，利用者の目的によって意味付けられる似ているものの集まり，または画像データを何らかの方法で分類して得られる個々の等質な集合のこと)．そして分類結果を画像表示する際には1つのクラスに1つの色を割り当てる方法であった．

このような従来の分類法や配色方法に対して，本稿では混合分布モデルと配色アルゴリズムを用いた分類法を提案する．まず，特徴空間の構造が複数の分布による多変量混合分布からなっていると考える．つまり1つの画素中に複数の対象物が混入している場合，その画素上の観測値は，これらの対象物に固有な分光反射輝度が混在したものとしてとらえ，その様相（状態）を混合分布モデルによりとらえる．ただし特徴空間の次元が高いので観測値に主成分分析を行い，

* 数物科学研究科 統計科学専攻：〒106 東京都港区南麻布 4-6-7.

次元を縮小した主成分スコアに対して混合分布モデルをあてはめる．そして，従来の分類法のような特定の対象物を識別するのではなく，画像全体の特徴をとらえるため，大まかな分類を行う．そのためクラスとしては，ほぼ“水域”（海，河川，湖沼），“植生”，“人工物”（建築物や市街地などの人工的に作られた建造物）のようなものを想定する．次に，推定した混合分布の情報を用いて分類結果を色彩情報として画像化する．これは，従来の分類法のような画素単位でクラスへの所属判定を行い，これに配色をすることによって色彩画像を生成するのではなく，各コンポーネント分布の特徴をある種の平滑化したマクロな色彩イメージ情報として，画像上に視覚化することである．

なお，実際の画像データの観測値には多変量正規分布より裾が重いデータが存在するため，混合分布モデルのコンポーネント分布として多変量正規分布と，これより裾の重い多変量 t 分布の2つを比較検討した．この結果，多変量正規分布にもとづく混合分布モデルのあてはまりが必ずしもよくない場合があり，この解決策としての，多変量 t 分布にもとづく混合分布モデルの有効性を検証することができた．

以上の内容に沿って，第2章では扱う画像データと具体的な分類手順について，第3章では混合分布モデルとそのパラメータ推定法について述べる．第4章は色彩画像表示する際の配色方法について，第5章は3種類の画像データの解析例について述べ，第6,7章では提案した手法の利点や問題点について議論する．

2. LANDSAT 画像データの解析手順

2.1 画像データ

解析に用いる画像データは，LANDSAT 5号の TM で観測された7つのバンド（観測波長帯）からなる多変量の観測値である．バンド1～5,7の分解能は 30 m×30 m，バンド6は 120 m×120 m である．各バンドの観測値は0～255の整数値からなり，バンド1～3は可視光線帯域，バンド4,5,7は近赤外線帯域，バンド6は熱赤外線帯域である．今回の解析では分解能が大きく異なるバンド6を除いた6つのバンドを解析対象とした．扱うデータは各バンドの観測値だけではなく，座標情報，時間的情報の利用も考えられるが，ここでは観測値のみを用いる．

一般に，解析する対象のデータは，その目的に応じたバンドの選択を行ったり，バンド間演算などにより次元を縮小することが行われる．本来は6つのバンドを同時に観察したいが，次元が高いのでここでは主成分分析により次元を縮小してデータの特徴をとらえることにする．そこで分解能の同じバンド（バンド1～5,7）の観測値から得られる分散共分散行列にもとづく主成分分析によって次元縮小を行い，正規化しない第1,2主成分スコアを分類対象のデータとする（主成分スコアの布置により構成される空間を“データ空間”と呼ぶ）．主成分スコアは各バンドの観測値の線形結合であるので，この操作は一種のバンド間演算と考えられる．

2.2 解析手順

各シーンに対して以下の手順で解析を進める．図1にこの手順の概略を示す．

ステップ1. ［次元の縮小］ 1シーンの画像の全画素上の観測値を用いて標本分散共分散行列を求め，主成分分析により次元縮小を行う．正規化しない第1,2主成分スコアを分類対象のデータとする．

ステップ2. ［トレーニングデータの作成］ 扱う1シーンの全画素数は数万～数十万画素と

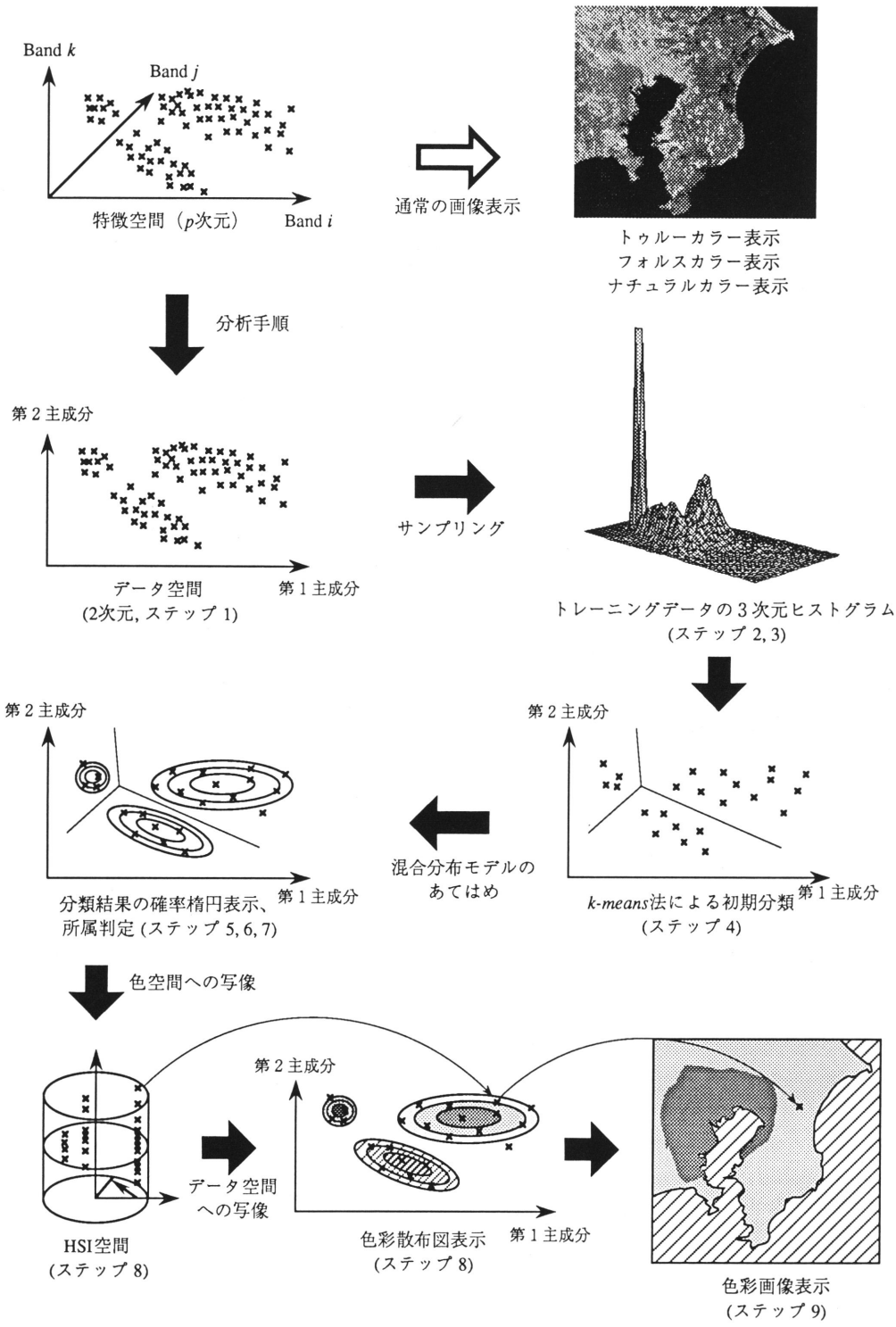

Band k
Band j
特徴空間（p次元）
Band i
通常の画像表示
トゥルーカラー表示
フォルスカラー表示
ナチュラルカラー表示
分析手順
第2主成分
データ空間
(2次元, ステップ 1)
第1主成分
サンプリング
トレーニングデータの3次元ヒストグラム
(ステップ 2, 3)
第2主成分
k-means法による初期分類
(ステップ 4)
第1主成分
混合分布モデルの
あてはめ
第2主成分
分類結果の確率楕円表示、
所属判定 (ステップ 5, 6, 7)
第1主成分
色空間への写像
第2主成分
データ空間
への写像
HSI空間
(ステップ 8)
色彩散布図表示
(ステップ 8)
第1主成分
色彩画像表示
(ステップ 9)

図1.　解析手順の概略.

非常に膨大であるため，シーンごとにそれぞれのデータ空間から5%のデータをランダムサンプリングし，これを“トレーニングデータ”とする．今回解析した画像データはいずれも600×800＝480,000画素で，トレーニングデータの標本数は，N＝24,000画素になる．

ステップ3. [コンポーネント分布の数の指定] トレーニングデータの3次元ヒストグラムを描き（図5(a), 図6, 図7），これを観察して混合分布モデルのコンポーネント分布の数を指定する．

ステップ4. [初期分類] ステップ3で決めたコンポーネント分布（クラス）の数にしたがって，トレーニングデータを分割型分類法（k-means法，MacQueen (1967) など）で初期分類し，各クラスの統計量（平均ベクトル，分散共分散行列，混合比率）を計算する．

ステップ5. [正規混合モデルのあてはめ] ステップ4で求めた統計量をEMアルゴリズム（次章で説明）の初期値として，正規混合モデル（多変量正規分布の混合分布モデル）のパラメータ推定を行う．

ステップ6. [t混合モデルのあてはめ] ステップ5で推定した正規混合モデルのパラメータの値を初期値としてEMアルゴリズムと*re-weighting*法（次章で説明）を用いてt混合モデル（多変量t分布の混合分布モデル）のパラメータ推定を行う．

ステップ7. [事後確率による判別] 2つの混合分布モデルから推定されたパラメータを用いてそれぞれ判別ルール（事後確率による判別）を構成し，1シーンの全データの各コンポーネント分布に対する所属を決定する．この結果は配色ルールで用いる．

ステップ8. [配色ルール] ステップ5, 6のパラメータを用いて，色彩画像表示のための配色ルールを構成する．

ステップ9. [色彩散布図表示] 配色ルールにもとづき1シーンのすべてのデータについて配色を行い，“色彩散布図”（配色されたデータ空間）を描く（図9(a), (c), 図10(a), (c), 図11(a), (c)).

ステップ10. [色彩画像表示] 色彩散布図での配色をもとに，“色彩画像”（配色された画像）の表示を行う（図9(b), (d), 図10(b), (d), 図11(b), (d)).

次章ではステップ5，ステップ6の混合分布モデルのパラメータ推定について，4章ではステップ8の配色ルールについて述べる．

3. 混合分布モデル

観測値に対してr個のコンポーネント分布$g_k(\cdot)$からなる混合分布モデル

$$f(\boldsymbol{x}|\boldsymbol{\Theta})=\sum_{k=1}^{r}\pi_k g_k(\boldsymbol{x}|\boldsymbol{\theta}_k) \tag{3.1}$$

のあてはめを考える．ここで，π_kはコンポーネント分布の混合比率パラメータとし（$\sum_{k=1}^{r}\pi_k=1, 0<\pi_k<1$），$\boldsymbol{\Theta}=\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, \ldots, \boldsymbol{\theta}_r, \pi_1, \ldots, \pi_k, \ldots, \pi_r\}$とおく．

各コンポーネント分布$g_k(\cdot)$の確率密度関数を楕円分布族（狩野 (1992), Kano et al. (1993) など）とするとき，それは一般的に次のように書くことができる：

$$g_k(\boldsymbol{x}|\boldsymbol{\theta}_k)=|\boldsymbol{V}_k|^{-1/2}h_k\{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{V}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)|\nu_k\}. \tag{3.2}$$

ここで，^{T}は転置を表し，$h_k(\cdot)$は$(\boldsymbol{\mu}_k, \boldsymbol{V}_k)$に依存しない非負の関数，$\boldsymbol{\mu}_k$は第$k$コンポーネント分布の位置ベクトル，$\boldsymbol{V}_k$は尺度行列，$\nu_k$は$\boldsymbol{\mu}_k$, $\boldsymbol{V}_k$とは独立なあるパラメータ，$\boldsymbol{\theta}_k=\{\boldsymbol{\mu}_k, \boldsymbol{V}_k, \nu_k\}$である．

混合分布モデル (3.1) のパラメータの推定は，観測値 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ を用いて以下の手順により行う．パラメータ ν_k $(k=1, \ldots, r)$ を事前に与えておき $(\nu_k=\nu_k^{(0)})$，目的関数として擬対数尤度関数 (pseudo log likelihood function)

$$\mathscr{L}(\boldsymbol{\Theta})=\sum_{i=1}^{N}\log\sum_{k=1}^{r}\pi_k|\boldsymbol{V}_k|^{-1/2}h_k\{(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{V}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)|\nu_k^{(0)}\} \tag{3.3}$$

を構成する．これを $\boldsymbol{\mu}_k$, $\boldsymbol{V}_k$ でそれぞれ偏微分して 0 とおくことにより，各コンポーネント分布の位置ベクトルと尺度行列の推定値は，次式で与えられる：

$$\hat{\boldsymbol{\mu}}_k=\frac{1}{\sum_{i=1}^{N}\hat{P}r(G_k|\boldsymbol{x}_i)w(\hat{s}_{ki}^2|\nu_k^{(0)})}\sum_{i=1}^{N}\hat{P}r(G_k|\boldsymbol{x}_i)w(\hat{s}_{ki}^2|\nu_k^{(0)})\boldsymbol{x}_i, \tag{3.4}$$

$$\hat{\boldsymbol{V}}_k=\frac{1}{N\hat{\pi}_k}\sum_{i=1}^{N}\hat{P}r(G_k|\boldsymbol{x}_i)w(\hat{s}_{ki}^2|\nu_k^{(0)})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k)^T. \tag{3.5}$$

ここで，

$$\hat{s}_{ki}^2=(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k)^T\hat{\boldsymbol{V}}_k^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_k), \tag{3.6}$$

$$\hat{P}r(G_k|\boldsymbol{x}_i)=\frac{\hat{\pi}_k|\hat{\boldsymbol{V}}_k|^{-1/2}h_k(\hat{s}_{ki}^2|\nu_k^{(0)})}{f(\boldsymbol{x}_i|\hat{\boldsymbol{\Theta}})}, \tag{3.7}$$

$$\hat{\pi}_k=\frac{1}{N}\sum_{i=1}^{N}\hat{P}r(G_k|\boldsymbol{x}_i), \tag{3.8}$$

$$w(\hat{s}_{ki}^2|\nu_k^{(0)})=-2\frac{\partial}{\partial\hat{s}_{ki}^2}\log h_k(\hat{s}_{ki}^2|\nu_k^{(0)}). \tag{3.9}$$

G_k は推定したコンポーネント分布を表すものとし，$\hat{s}_{ki}^2$ は分布 G_k と標本 $\boldsymbol{x}_i$ との擬マハラノビス距離（尺度行列 $\boldsymbol{V}$ を考慮した距離で，$\boldsymbol{V}$ が分散共分散行列の場合はマハラノビス距離になる），$\hat{P}r(G_k|\boldsymbol{x}_i)$ は $\boldsymbol{x}_i$ がコンポーネント分布 G_k に所属する確率（事後確率），$w(\hat{s}_{ki}^2|\cdot)$ は第 k コンポーネント分布に対する $\boldsymbol{x}_i$ のウェイトである．

これらの式をもとに，ここではEMアルゴリズム (Dempster et al. (1977), Redner and Walker (1984)) を用いて未知パラメータの推定を行う．すなわち，事後確率 $Pr(\cdot|\cdot)$ とウェイト $w(\cdot|\cdot)$ を欠測データとして扱い，E-step (expectation step) でこれらの期待値推定を行う．次に，M-step (maximization step) として各パラメータの推定を行う．詳しい手順は後で述べる．

とくに，各コンポーネント分布 G_k が多変量 t 分布（p 変量）の場合，その確率密度関数は，

$$\frac{\Gamma((p+\nu_k)/2)}{(\pi\nu_k)^{p/2}\Gamma(\nu_k/2)|\boldsymbol{V}_k|^{1/2}}\left\{1+\frac{(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{V}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}{\nu_k}\right\}^{-(p+\nu_k)/2} \tag{3.10}$$

と書けるので (Lange et al. (1989)など)，コンポーネント分布 G_k に対する各 $\boldsymbol{x}_i$ のウェイトは次のように計算される：

$$w(s_{ki}^2|\nu_k)=\frac{\nu_k+p}{\nu_k+s_{ki}^2}. \tag{3.11}$$

このように t 分布などのモデルで各点 $\boldsymbol{x}_i$ にウェイトをつけてパラメータを推定する方法を *re-weighting* 法 (Dempster et al. (1980)) と言う．ν_k を形状パラメータと呼び，整数値とは限らず正の実数値をとる．ν_k の推定値は陽な形で書くことができないため，π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{V}_k$ と同時には推定せずに，準ニュートン法（北川 (1993), 付録Aのプログラムを用いた）で推定を行う．

詳しくは次の手順でパラメータ推定を行う．

パラメータ推定アルゴリズム

ステップ1. [初期値設定] 各コンポーネント分布のパラメータ，事後確率，ウェイトの初期値を設定する．まず事後確率の初期値は k-means 法により r 群に分類を行い，

$$Pr^{(0)}(G_k|\boldsymbol{x}_i)=\begin{cases}1 & \text{if } \boldsymbol{x}_i\in G_k\\ 0 & \text{otherwise}\end{cases}$$

($k=1,\ldots,r$; $i=1,\ldots,N$) とする．事後確率の初期値をもとにして，$\pi_k^{(0)}$, $\boldsymbol{\mu}_k^{(0)}$, $\boldsymbol{V}_k^{(0)}$ は，それぞれ (3.8), (3.4), (3.5) 式を用いて計算を行う．ウェイト $w^{(0)}(\cdot|\cdot)$ は (3.11) 式を用いて推定し，形状パラメータ ν_k の初期値 $\nu_k^{(0)}$ は 4 とする．また，反復回数のカウンターは $t\leftarrow 1$ とする．

ステップ2. [E-step] 第 t ステップ ($t\geq 1$) の E-step として，事後確率 $Pr^{(t)}(\cdot|\cdot)$ とウェイト $w^{(t)}(\cdot|\cdot)$ の推定を行う．すなわち，事後確率は (3.7) 式で，ウェイトは (3.11) 式により推定する．同時に (3.8) 式により $\pi_k^{(t)}$ の推定を行う．

ステップ3. [M-step(1)] 第 t ステップの M-step は，尤度の最大化により各コンポーネントのパラメータ推定を行う．すなわち $\boldsymbol{\mu}_k^{(t)}$ と $\boldsymbol{V}_k^{(t)}$ を (3.4), (3.5) 式により推定する．

ステップ4. [M-step(2)] t 混合モデルの場合，$\pi_k^{(t)}$, $\boldsymbol{\mu}_k^{(t)}$, $\boldsymbol{V}_k^{(t)}$ を固定して ν_k に関する擬対数尤度関数の最大化により，$\nu_k^{(t)}$ の推定を行う．

ステップ5. [収束判定] 次の収束条件を満足すれば反復計算を終了し，そうでなければ $t\leftarrow t+1$ としてステップ2へ戻る．

$$|\mathscr{L}^{(t)}(\hat{\boldsymbol{\Theta}}^{(t)})-\mathscr{L}^{(t-1)}(\hat{\boldsymbol{\Theta}}^{(t-1)})|<\varepsilon \quad \text{or} \quad \|\hat{\boldsymbol{\Theta}}^{(t)}-\hat{\boldsymbol{\Theta}}^{(t-1)}\|<\delta.$$

ここで $\mathscr{L}^{(t)}(\cdot)$ は第 t ステップで推定された対数尤度の値，$\hat{\boldsymbol{\Theta}}^{(t)}$ は第 t ステップで推定されたパラメータである．ε と δ は十分に小さい正数である (10^{-5}〜10^{-7} 程度)．

多変量 t 分布における形状パラメータ ν の推定方法に関しては Lange et al. (1989) の研究があるが，この推定方法を混合分布モデルに陽に拡張できなかったため，このような方法によりパラメータの推定を行った．なお，正規分布の場合ウェイト $w(\cdot|\cdot)$ は 1 になり，正規混合分布モデルのパラメータ推定法と一致する (Everitt and Hand (1981), McLachlan (1992), McLachlan and Basford (1988) など)．

4. 分類結果の画像表示および配色方法

混合分布モデルによって推定されたデータ空間の特徴を効果的に反映する配色方法を提案する．

4.1 HSI 空間

まず，カラーモデルである HSI 空間について述べる．カラーディスプレイなどで画素上の色を特定する場合には，RGB (Red, Green, Blue) の数値を指定する方法が一般的であるが，これらの数値の組み合せから合成色の色調をコントロールするのは容易ではない．そこで感覚的に理解しやすい，マンセルの表色系で使われる 3 つの属性，色相 h(hue), 彩度 s(saturation), 明度 i(intensity) を用いる．通常用いる HSI 空間は六角錐を 2 つ張り合わせた双六角錐カラーモデルであるが，ここでは図 2 に示す円柱座標系で表現したカラーモデルを用いる．これを "規格化された HSI 空間" と言う (高木・下田 監修 (1991), 2.1.2.3 節)．各属性の定義域は $h\in[0,1)$, また $s, i\in[0,1]$ とする．この空間の中で色の合成を行い，HSI 空間から RGB 空間への

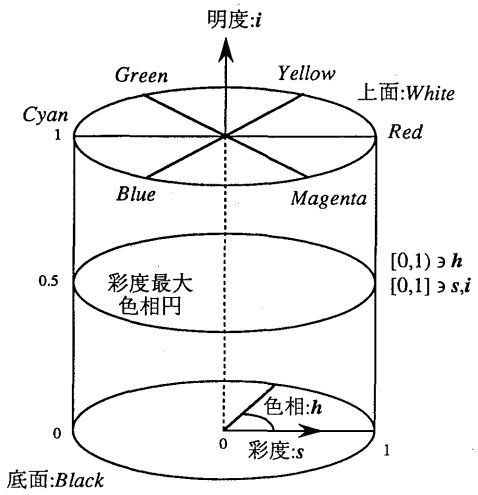


図2.　規格化された HSI 空間.

変換を通してディスプレイに色彩表示する.

HSI 空間内の色彩の構造は次のとおりである. この空間内のある点 $\boldsymbol{P}$ が与えられたとき, 色相はあらかじめ定めておいた基準位置との角度で表される. ここで角度は区間 $[0, 1)$ で規格化されていて, 例えば Red=0 を基準位置にすると, Yellow=1/6, Green=1/3, Cyan=1/2, Blue=2/3, Magenta=5/6 という値をとる. 彩度は点 $\boldsymbol{P}$ から HSI 空間の円柱の中心軸までの最短距離で表され, 軸に近いほど灰色に近くなる. 明度は点 $\boldsymbol{P}$ から円柱の底面への距離になり, 中心軸に沿って底面から上面に向かって黒色から灰色, そして白色に変化する. $s=1$, $i=1/2$ の位置にある円は彩度最大色相円である. なお, この HSI 空間から RGB 空間への変換方法は Foley and Van Dam ((1982), Chapter 17), 高木・下田 監修 ((1991), 2.1.2.3 節) によるアルゴリズムを用いた.

4.2　配色アルゴリズム

推定した混合分布の各統計量などを用いて, データ空間内の任意の点 $\boldsymbol{p}^{DS}$ を HSI 空間内の点 $\boldsymbol{p}^{HSI}$ に 1 対 1 写像する方法を示す. まずコンポーネント分布を従来の分類法におけるクラスと考え, 各コンポーネント分布に対して一定の色相と彩度を対応させる. それらの値は各コンポーネント分布の位置ベクトルと色相の基準位置に対応させるコンポーネント分布の位置ベクトルとの関係により定める. 明度は各コンポーネント分布の位置ベクトルと点 $\boldsymbol{p}^{DS}$ の距離により定める. 点 $\boldsymbol{p}^{DS}$ はすべてのコンポーネント分布の定義域にあるため, このまま各コンポーネント分布ごとに点 $\boldsymbol{p}^{DS}$ を HSI 空間上に対応させると, HSI 空間内でコンポーネント分布と同じ数の点 $\boldsymbol{p}^{HSI}$ が存在するので, 点 $\boldsymbol{p}^{DS}$ と点 $\boldsymbol{p}^{HSI}$ の 1 対 1 の対応を次の写像アルゴリズムにより行う.

HSI 空間への写像アルゴリズム

ステップ 1. [彩度の求め方]　コンポーネント分布 G_k の彩度 S_k は, 混合比率を重みとする位置ベクトルの重心

$$\boldsymbol{g}=\sum_{k=1}^{r}\hat{\pi}_k\hat{\boldsymbol{\mu}}_k$$

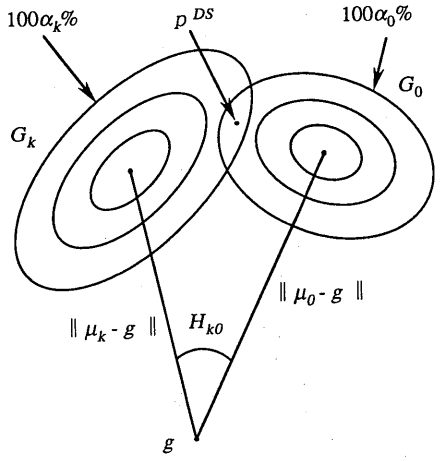


図3. データ空間での分布の位置関係.

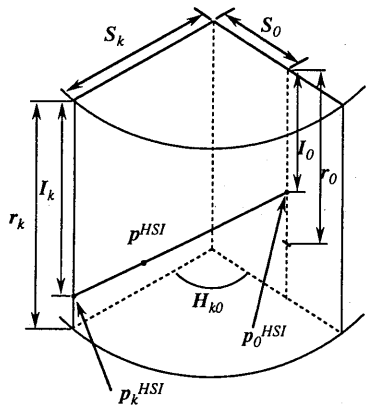


図4. HSI空間.

と G_k の位置ベクトル $\hat{\boldsymbol{\mu}}_k$ との距離により決める（図3）．すなわち，

$$S_k=\frac{\min_{i\in\{1,\ldots,r\}}\|\hat{\boldsymbol{\mu}}_i-\boldsymbol{g}\|}{\|\hat{\boldsymbol{\mu}}_k-\boldsymbol{g}\|}$$

とする．これは重心 $\boldsymbol{g}$ に一番近いコンポーネント分布の彩度が最大になり，これから離れるほど灰色に近くなることを示す．

ステップ2. [色相の求め方] コンポーネント分布 G_k の色相 H_{k0} は，$\hat{\boldsymbol{\mu}}_k$ と色相の基準位置に対応させるコンポーネント分布 G_0 の位置ベクトル $\hat{\boldsymbol{\mu}}_0$ が重心となす角度により決定する（図3）．すなわち

$$H_{k0}=\cos^{-1}\frac{(\hat{\boldsymbol{\mu}}_k-\boldsymbol{g},\ \hat{\boldsymbol{\mu}}_0-\boldsymbol{g})}{\|\hat{\boldsymbol{\mu}}_k-\boldsymbol{g}\|\cdot\|\hat{\boldsymbol{\mu}}_0-\boldsymbol{g}\|}$$

とする． G_0 の決め方は後で述べる．

ステップ3. [明度の変化幅の求め方] 明度は各コンポーネント分布の位置ベクトルとデータの擬マハラノビス距離により決めるが，その変化幅は

$$r_k=\frac{\log|\hat{\boldsymbol{V}}_k|}{\max_{i\in\{1,\ldots,r\}}\log|\hat{\boldsymbol{V}}_i|},\qquad r_k\in[0,1]$$

とする（図4）．ここで $|\hat{\boldsymbol{V}}_k|$ はコンポーネント分布 G_k の尺度行列の行列式の値である．

ステップ4. [明度の求め方] コンポーネント分布と点 $\boldsymbol{p}^{DS}$ の距離は，コンポーネント分布の確率楕円を考え，そのパーセント点により与える．これはコンポーネント分布が正規分布の場合はマハラノビス距離に対応する．いま点 $\boldsymbol{p}^{DS}$ がコンポーネント分布 G_k の確率楕円の $100\,\alpha_k$% であるとき，G_k における点 $\boldsymbol{p}^{DS}$ の明度を，

$$I_k=1-\alpha_k r_k+\varepsilon_k$$

により与える（図4）．ここで $\varepsilon_k(\geq 0)$ は点 $\boldsymbol{p}^{DS}$ がどのコンポーネント分布に所属するかを強調するためのものである．つまり，解析手順のステップ7のコンポーネント分布に対する所属判定の結果により，G_k に所属していれば $\varepsilon_k=\text{constant}$ $(\ll 1)$，そうでなければ0とする．実際には constant$=0.005$ とした．ここで α_k は連続量であるが，ディスプレイの表示可能な色の数に

制限があるため，実際は離散化したものを用いた．そして，$\boldsymbol{p}^{DS}$ のパーセント点が 100 α_k% と 100α'_k% の間 ($\alpha_k > \alpha'_k$) にあるとき，$\boldsymbol{p}^{DS}$ は 100 α_k% 点とする（図 3)．実際に離散化したパーセント点としては，0, 5, 10, 25, 50, 75, 90, 95, 97.5, 99% の 9 階調を用いた．

ステップ 5. ［色の合成］ 最後に，データ空間内のある点 $\boldsymbol{p}^{DS}$ と，HSI 空間の円柱座標内での点 $\boldsymbol{p}^{HSI}$ は次のように対応付けする．まず，ステップ 1, 2 によりコンポーネント分布 G_k の色相 H_{k0} と彩度 S_k の値が決まり，次に，任意の点 $\boldsymbol{p}^{DS}$ が与えられると，ステップ 3, 4 により分布 G_k に対する明度 I_k が決まる．その座標は $\boldsymbol{p}_k^{HSI}=(H_{k0}, S_k, I_k)$ である（図 4)．このとき円柱座標内での点 $\boldsymbol{p}^{HSI}$ の位置は，

$$\boldsymbol{p}^{HSI}=\sum_{j,\alpha_j<\beta}\frac{\alpha_j}{A}\boldsymbol{p}_j^{HSI}, \qquad \text{ここで,} \quad A=\sum_{\alpha_j<\beta}\alpha_j$$

で求める．この 2 つの和の記号は，各コンポーネント分布に対して点 $\boldsymbol{p}^{DS}$ のパーセント点が 100β% 点より小さい場合のみについて和をとることを示す．つまり，100β% 点より大きい場合は，そのコンポーネント分布の影響がほとんどないと見なすことである．実際には $\beta=0.99$ とした．

さて，HSI 空間内で基準位置に対応させるコンポーネント分布 G_0 は，次のように定める．r 個あるコンポーネント分布から任意に 1 つを選んで色彩画像表示する．次に，たとえば植生に相当する領域を緑がかった色に配色する場合，植生に相当するコンポーネント分布を G_0 として HSI 空間で緑付近の色相を基準位置として指定する．

ここで示したアルゴリズムによりデータ空間内のすべてのデータは，色彩として変換され，そのデータに対応するもとの画像の画素上に置かれる．データ空間でコンポーネント分布が重なっている部分は，各コンポーネント分布からの距離を考慮した混色が行われる．つまり，画素上の色の情報から，データ空間内のそのデータのコンポーネント分布に対する所属の度合が，色調の変化として視覚的に観察できる．結果として，直感的に画像内の空間的分布の特徴を把握することが可能になるところに，この色彩画像表示の利点がある．

5. 解　析　例

解析対象地域（シーン）は 3 種類あり，三浦半島，横浜市，千葉市～習志野市（それぞれ，600×800 画素，観測日：1986 年 8 月 6 日）である．これらのデータは幾何補正などの基本的な補正処理は施されている．また，画像表示したときに真上が北になるような回転の処理が行われている．

2 章で示した解析手順にしたがって解析した結果を，図表の観察を中心に述べる．

5.1 データ空間の特徴

図 5(a) は三浦半島，図 6 は横浜市，図 7 は千葉市～習志野市のトレーニングデータの 3 次元ヒストグラムである．これらの 3 次元ヒストグラムの特徴として，大きなピークが図 5(a) では 3 つあり，図 6, 7 では 2 つみられる．ただし，図 6, 7 は C の位置に第 3 の小さなピークが確認できる．以上のことから，各シーンに対してコンポーネント分布の数を 3 として分類を行った．

5.2 混合分布モデルのあてはめ

表 1, 2 は 3 シーンのトレーニングデータに正規混合モデルと t 混合モデルをあてはめた結果得られた数値である．表 1 は，推定された対数尤度の値，AIC (Akaike (1973)) の値，そして

表1. モデル推定に関する諸数値.

シーン	モデル	対数尤度 $\hat{\mathscr{L}}$	AIC*1	EMアルゴリズムの反復回数	所要計算時間(分)
三浦半島	正規混合モデル	−157237.5	314509.1	145	14
	t 混合モデル	−156890.3	313820.7	—*2	210
横浜市	正規混合モデル	−187931.0	375895.9	193	18
	t 混合モデル	−187144.1	374328.3	—*2	226
千葉市〜	正規混合モデル	−183540.3	367114.5	484	34
習志野市	t 混合モデル	−183065.5	366171.0	—*2	435

*1 AIC＝−2×(最大対数尤度)+2×(パラメータ数)
*2 EMアルゴリズムと準ニュートン法を併用しているため1つの数値で示せない.

表2. 混合比率と形状パラメータの推定値と主成分寄与率.

シーン		正規混合モデル $\hat{\pi}_k$	t 混合モデル $\hat{\pi}_k$	形状パラメータ $\hat{\nu}_k$	主成分累積寄与率 (pc1, pc2)
三浦半島	A	0.536	0.536	58.44	0.975
	B	0.269	0.237	4.50	
	C	0.196	0.227	∞	(0.914, 0.061)
横浜市	A	0.136	0.137	6.79	0.961
	B	0.823	0.639	3.50	
	C	0.041	0.224	∞	(0.876, 0.085)
千葉市	A	0.179	0.179	28.28	0.957
〜	B	0.596	0.474	3.60	
習志野市	C	0.225	0.347	30.90	(0.888, 0.069)

記号A, B, Cは図5, 6, 7, 8のコンポーネント分布に対応する.

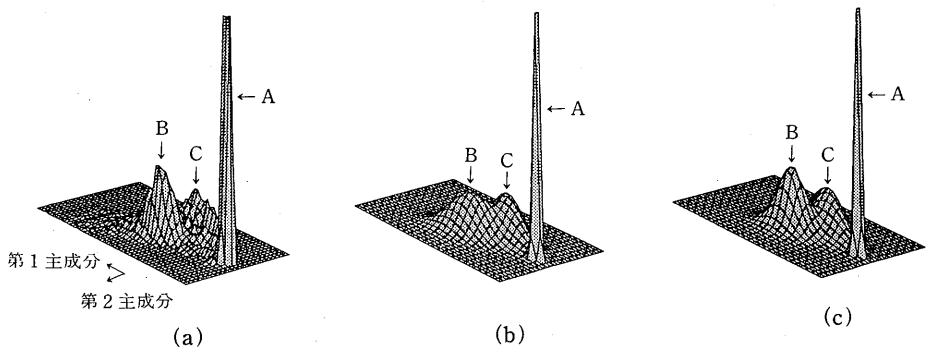


図5. 三浦半島. (a) トレーニングデータの3次元ヒストグラム, (b) 正規混合モデル, (c) t 混合モデル.

EM アルゴリズムの反復回数と所要計算時間である（計算に使用した機種は富士通 S-4/10 model 30）．表 2 は，2 つのモデルの各コンポーネント分布の混合比率と，t 混合モデルの形状パラメータの推定値，第 2 主成分までの累積寄与率と第 1, 2 主成分の寄与率である．

図 5(b), (c) は三浦半島のトレーニングデータに，それぞれ正規混合モデルと t 混合モデルをあてはめ，推定された混合分布の密度関数の立体グラフである．図 5(a) の記号 A, B, C のピークが図 5(b) の A, B, C のコンポーネント分布に対応し，さらに図 8(a) の A, B, C の確率楕円にも対応する．また，同様に図 5(a) の A, B, C は図 5(c) の A, B, C のコンポーネント分布と，図 8(b) の A, B, C の確率楕円に対応する．このトレーニングデータは，図 5(a) を見て分かるようにはっきり大きく 3 つに分かれているため，正規混合モデルによっても t 混合モデルによっても分類結果に大きな差はない．しかし，図 5 の 3 つの図を比較すると図 5(b) の正規モデ

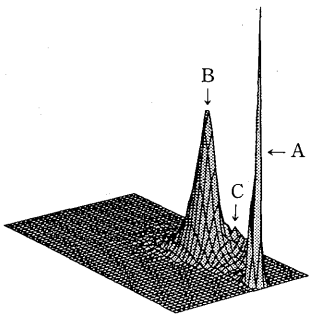


図 6．横浜市のトレーニングデータの 3 次元ヒストグラム．

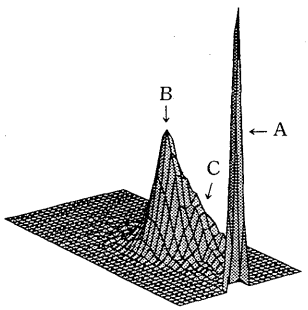


図 7．千葉市～習志野市のトレーニングデータの 3 次元ヒストグラム．

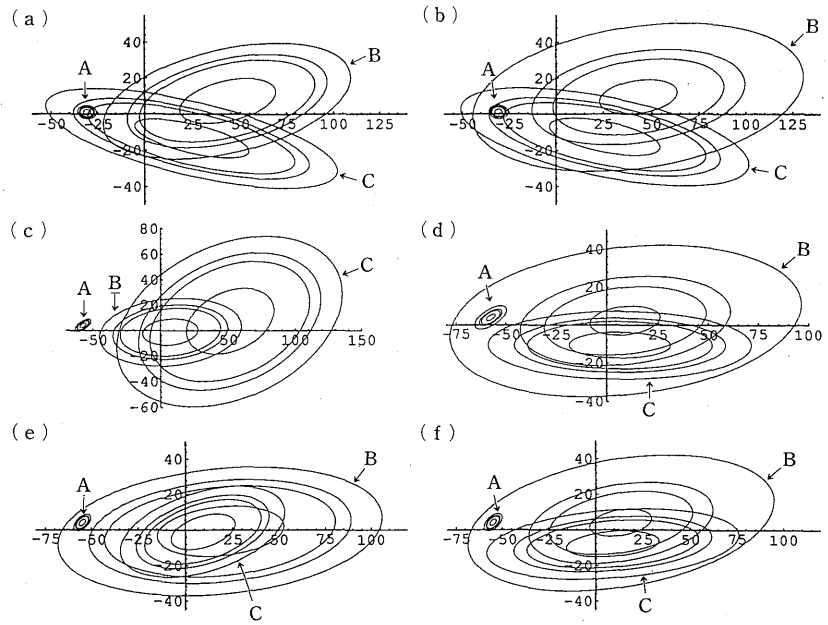
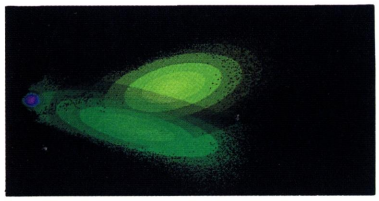
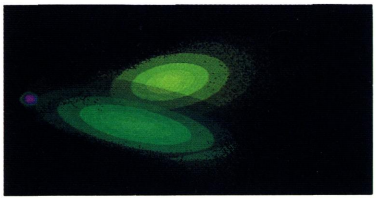


図 8．混合モデルの確率楕円．(a) 三浦半島正規混合モデル，(b) 三浦半島 t 混合モデル，(c) 横浜市正規混合モデル，(d) 横浜市 t 混合モデル，(e) 千葉市～習志野市正規混合モデル，(f) 千葉市～習志野市 t 混合モデル．

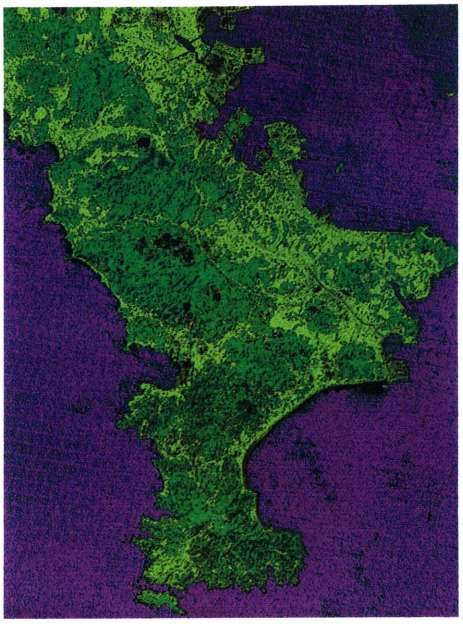


(a) 色彩散布図（正規混合モデル）.

(b) 色彩画像（正規混合モデル）.

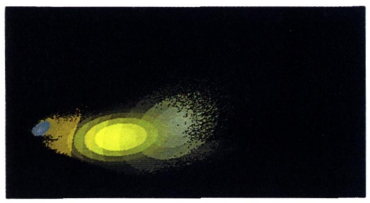


(c) 色彩散布図（t 混合モデル）.

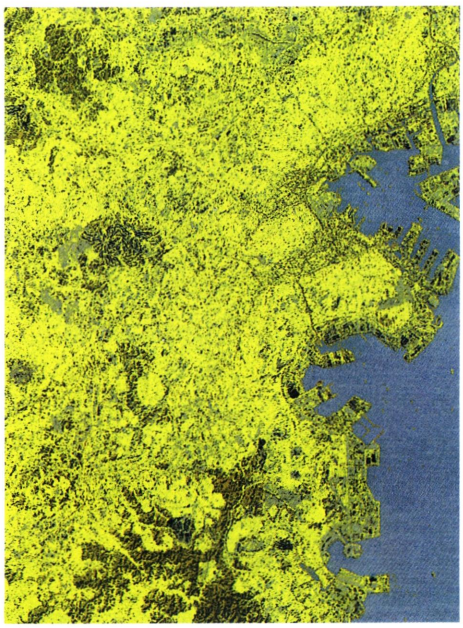


(d) 色彩画像（t 混合モデル）.

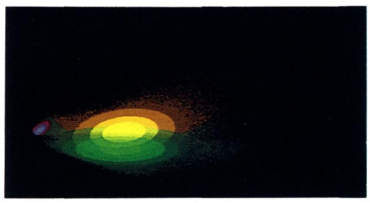
図9. 三 浦 半 島.



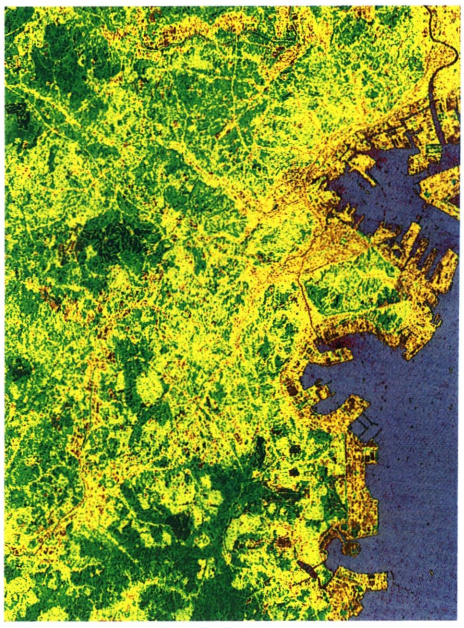
(a) 色彩散布図（正規混合モデル）.



(b) 色彩画像（正規混合モデル）.

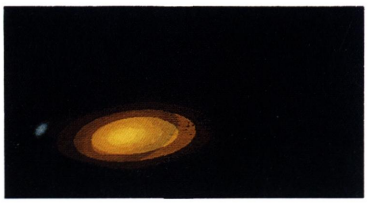


(c) 色彩散布図（t 混合モデル）.



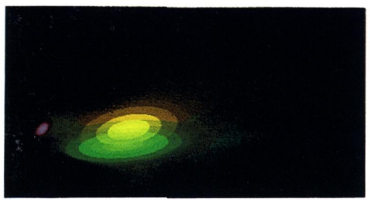
(d) 色彩画像（t 混合モデル）.

図 10.　横　　浜　　市.

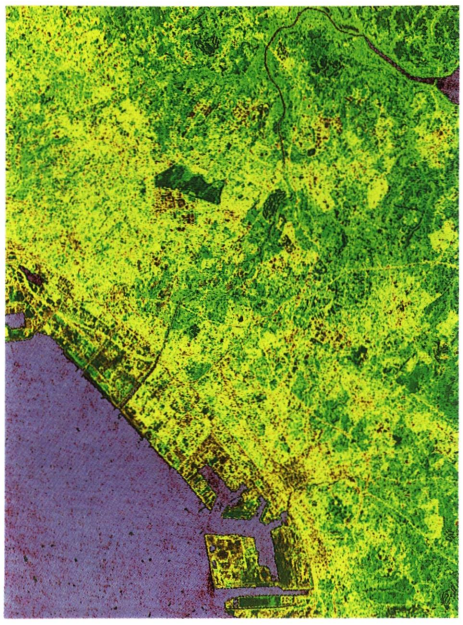


(a) 色彩散布図（正規混合モデル）.

(b) 色彩画像（正規混合モデル）.



(c) 色彩散布図（t 混合モデル）.



(d) 色彩画像（t 混合モデル）.

図 11. 千葉市～習志野市.

ルのBの部分が低く推定されている．それに対して，図5(c) の t 混合モデルはBの高さを良くとらえている．

図6は，横浜市のトレーニングデータの3次元ヒストグラムである．これに正規混合モデルと t 混合モデルをあてはめた結果の各コンポーネント分布の確率楕円はそれぞれ図8(c), (d) である．この図8(d) のBとCのコンポーネント分布の位置は，図8(c) とは違う場所に推定されている．また，図8(c) の正規混合モデルでの推定結果は，図6のAとBの大きな2つのピークを主にとらえている（この2つのコンポーネント分布の混合比率はあわせて0.959である）．これに対して図8(d) の t 混合モデルは，図6のA, B, Cが示すピークとほぼ同じ位置にコンポーネント分布が推定されている．

図7は千葉市～習志野市のトレーニングデータの3次元ヒストグラムである．このデータに正規混合モデルと t 混合モデルをあてはめた結果の各コンポーネント分布の確率楕円はそれぞれ図8(e), (f) である．この図8(e) のAのコンポーネント分布は図8(f) のAと同じ位置にあり，Bのコンポーネント分布もほぼ同じ位置にあるが，Cの位置が異なる．また，図8(e) の正規混合モデルのBとCのコンポーネント分布は位置がほぼ同じであるが，図8(f) の t 混合モデルではCのコンポーネント分布がBの下の位置に推定されている．図7のCの部分に小さなピークがあるが，t 混合モデルはこれをとらえているようである．

5.3 色彩散布図と色彩画像表示

3つのシーンの色彩画像はいずれも水域に相当するコンポーネント分布を G_0 として，青に近い色相で表示してある．

図9(a), (c) は三浦半島の色彩散布図で，それぞれ正規混合モデルと t 混合モデルをあてはめた結果である．これらに対応する確率楕円が図8(a), (b) である．これらのモデルの色彩画像表示はそれぞれ図9(b), (d) である．この色彩画像表示から，図5と図8(a), (b) のA, B, Cは，それぞれ水域，植生，人工物に対応すると考えられる．

同様に図10(a), (c) は横浜市の色彩散布図で，これらに対応するのが図8(c), (d), その色彩画像表示は図10(b), (d) である．この色彩画像表示から，図6と図8(d) の t 混合モデルのA, B, Cは，それぞれ水域，植生，人工物に対応すると考えられる．図8(c) 正規混合モデルの場合はAは水域に，BとCをあわせて陸域として対応がつく．

図11(a), (c) は千葉市～習志野市の色彩散布図で，これらに対応するのが図8(e), (f), その色彩画像表示は図11(b), (d) である．この色彩画像表示から，図7と図8(e) の t 混合モデルのA, B, Cは，それぞれ水域，植生，人工物に対応すると考えられる．図8(e) の正規混合モデルの場合は，横浜と同様にAは水域に，BとCをあわせて陸域に対応がつく．

6. 考　　察

3シーンに共通して，t 混合モデルをあてはめた場合，人工物に相当するコンポーネント分布の形状パラメータ ν_k が3.5～4.5の値で推定され，多変量 t 分布を用いた効果が見られた．ν_k の値が小さいことは正規分布に比べて十分裾を引いていることを示す．ν_k が大きな値になると正規分布に近づくので（具体的には20～30程度以上），t 分布によるモデリングは正規分布を包含するものとして扱うことができる．この意味で ν_k は正規性からの乖離の程度を表すパラメータである．例えば，三浦半島，横浜市のデータの人工物に相当するコンポーネント分布の形状パラメータの推定値はかなり大きな値（100以上）で，正規分布とほとんど変わらない分布としてとらえることができた．

t 混合モデルは水域，植生，人工物と各コンポーネント分布にある程度意味付けのできる（対応関係のつく）推定ができた．これに対して正規混合モデルでは三浦半島の場合を除いて，コンポーネント分布の意味付けは明確にできない．表1のAICの値を比較しても，いずれも正規混合モデルより t 混合モデルの方の値が小さいので，モデル選択規準の観点からも t 混合モデルの方が良いと言える．

また，水に相当する分布の混合比率は，正規混合モデルと t 混合モデル双方の推定値がほぼ同じ値である．水の分布が他の2つに比べて分散が非常に小さく，他の分布から離れているため，このような結果になったと考えられる．

以上のことから，多変量 t 分布にもとづく混合分布モデルを用いることの利点は次のようになる．解析に用いる画像データが正規分布より裾が重い場合がしばしばあり，t 分布はその特徴をとらえるのに有効な分布と考えられる．このようなデータに対して複数の正規分布をコンポーネント分布とする混合分布モデルをあてはめる方法もあるが，各コンポーネント分布の意味付けが困難になると同時に推定するパラメータの数が多くなるなどの問題が生じる．これに対して，t 分布を用いると1つのコンポーネント分布でとらえることができ，推定するパラメータ数の節約ができる．

7. 問 題 点

主な問題点は計算時間に関してである．EMアルゴリズム自身の特性として収束が遅いことがあげられるが，初期値の与え方によっては，収束するまでの反復計算の回数を減らせる可能性があるため，初期値の選定方法は十分検討の余地がある．例えば，グランド・トゥルースの情報を用いてトレーニングデータを作ることや，データ空間の散布図からコンポーネント分布の核となる部分を指定し，そこから初期パラメータを求める方法（McLachlan (1988)）など考えられる．今回の収束条件は $\varepsilon=\delta=10^{-7}$ としたが，EMアルゴリズムの反復計算のかなり早いうちにパラメータの収束先の値付近に到達しているので，ε と δ の値を 10^{-3} 程度にすると計算時間はかなり短縮できると考えられる．また，t 混合モデルでは，多変量 t 分布の形状パラメータ ν_k の推定に非線型最適化法を用いるため，正規混合モデルに比べて余計に時間を要する．表1の計算時間を比較すると，t 混合モデルは正規混合分布モデルに比べ約12～15倍要している．形状パラメータの推定を伴うためこのような結果となったが，実用的なアルゴリズムや数値計算上の工夫や改良が今後の課題である．

次に，トレーニングデータの標本数についても考えなければならない．標本数を増やすと収束特性，計算時間の問題が生じ，少なくすると母集団（1シーン内の全データ）の特性をどれくらい表しているか，どの程度信頼できる判別ルールを構成できるか等のジレンマが生じる．

8. おわりに

本稿はLANDSAT画像データに混合分布モデルをあてはめ，混合分布の情報を色彩情報として表現する方法を提案した．この分類法の特徴をまとめると次のようになる．

(1) 主成分分析で次元縮小を行うことで，対象物の分布の様相を容易に観察することができる．また，水域，植生，人工物というクラスで大まかな分類を行うことにより，画像全体の特徴をとらえることができる．分離の良いクラスはこれを取り除くなどの処理を行うことにより，二次分析への手掛りになる．

(2) 本稿では，1つの画素中に複数の対象物が混入している場合，その画素上の観測値は，これらの対象物に固有な分光反射輝度が混在したものとしてとらえる．これは特徴空間におけるその画素上の観測値が，対象物に対応すると推測される複数のコンポーネント分布が重なりあった部分に相当すると考えることができる．このような場合，従来の分類法では画素上の観測値は特定のクラスに所属するものとして，何らかの基準であるクラスに割り当て，分類していたが，本稿は複数の対象物が混入している画素の特徴を，推定したコンポーネント分布の重なりあう様相として色彩画像により視覚化する点に特徴がある．

(3) ここで提案する配色方法は，推定した混合分布のパラメータの値や各コンポーネント分布に対する観測値の確率密度に応じて配色を行う．つまり，特徴空間（データ空間）の構造が画像上で視覚化され，分類結果の画像の解釈が容易になる．

(4) 扱う画像データの画素数が比較的大きく（数十万程度），グランド・トゥルース（地図や現地調査などにもとづく情報）による詳細な情報が入手困難な場合などに有効な手法と考えられる．

LANDSATの画像データの画像表示方法としては，トゥルーカラー（true color）表示，フォルスカラー（false color）表示，ナチュラルカラー（natural color）表示などの方法がある（詳しくは高木・下田 監修 (1991), 2.1.2節を参照）．基本的には各バンドの輝度値をカラーディスプレイのRGBに割り当て，色の合成やバンドの部分的な強調処理をしてバンドの特徴を見る方法である．ここで提案した方法はこれらの方法や従来の分類法による画像表示方法とは根本的に異なることを強調しておきたい．

今回は第2主成分までの2次元のデータに対して混合分布モデルをあてはめ，配色する方法をとったが，3次元以上のデータに対してもこの方法は適用可能である．第3主成分以下を使った解析については別の機会に議論したい．

謝　辞

色空間への写像アルゴリズムについて有益なコメントをいただいた統計数理研究所馬場康維助教授と準ニュートン法のプログラムのソースコードを提供してくださった統計数理研究所北川源四郎教授に深く感謝致します．担当編集委員と査読者の方々に有益なコメントをいただきました．ここに記して厚く御礼申し上げます．

参 考 文 献

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Proceedings of 2nd International Symposium on Information Theory* (eds. B. Petrov and F. Csaki), 267-281, Akademiai Kiado, Budapest.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc. Ser. B,* **39**, 1-38.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed, *Multivariate Analysis V* (ed. P.R. Krishnaiah), 35-57, North-Holland, Amsterdam.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*, Chapman and Hall, London.

Foley, J.D. and Van Dam, A. (1982). *Fundamentals of Interactive Computer Graphics*, Addison Wesley, Massachusetts.

狩野　裕 (1992). 楕円分布と統計的推測，科研費シンポジウム「種々の統計母数モデルとその尤度関数による統計的推測」報告書，26-43.

Kano, Y., Berkane, M. and Bentler, P.M. (1993). Statistical inference based on pseudo-maximum likelihood estimators in elliptical populations, *J. Amer. Statist. Assoc.*, **88**, 135-143.

北川源四郎 (1993). 『FORTRAN 77 時系列解析プログラミング』, 岩波書店, 東京.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989). Robust statistical modeling using the t distributions, *J. Amer. Statist. Assoc.*, **84**, 881-896.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symp. on Math. Statist. Prob.*, Vol. 1 (ed. J. Neyman), 281-297, University of California Press, Berkeley.

McLachlan, G.J. (1988). On the choice of starting values for the EM algorithm in fitting mixture models, *Statistician*, **37**, 417-425.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models : Inference and Applications to Clustering*, Marcel Dekker, New York.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev.*, **26**, 195-239.

高木幹雄, 下田陽久 監修 (1991). 『画像解析ハンドブック』, 東京大学出版, 東京.

宇宙開発事業団地球観測センター 編 (1990). 『地球観測データ利用ハンドブック ―― ランドサット編・改定版 ――』, (財) リモート・センシング技術センター, 東京.

Proceedings of the Institute of Statistical Mathematics Vol. 41, No. 2, 149-167 (1993) 167

Classification of Remotely Sensed Images via Finite Mixture Distribution Models

Nagatomo Nakamura

(Department of Statistical Science, The Graduate University for Advanced Studies)

Sadanori Konishi and Noboru Ohsumi

(The Institute of Statistical Mathematics and
The Graduate University for Advanced Studies)

We discuss how to classify the LANDSAT image data via finite mixture models. The LANDSAT instrument measures the intensity of light in multispectral bands reflected from the surface of the earth. Analyzing the LANDSAT data yields a plenty of geographical knowledge, such as distribution of plants, activity of plants, environmental pollution, crop prospects and investigation of natural resources.

We examine the classification procedure based on a multivariate t mixture model, which is shown to be of much practical use in comparison with a normal mixture model in the analysis of the LANDSAT image data. Mixture models are applied to the scores obtained from the "feature space" by the principle component analysis. This enables us to dispense with a precise segmentation of the objects to grasp the feature of the entire image data. EM algorithm with re-weighting method is proposed as an estimation methodology for multivariate t mixture model. We also develop a coloring procedure which visualizes the characteristics of the obtained classification. Examples show how these estimating and coloring procedures work for actual data sets.

Key words: Finite mixture distribution, multivariate t distribution, EM algorithm, remote sensing, LANDSAT image data, HSI color space.