AN AUTOMATIC CLASSIFICATION METHOD FOR LANDSAT DATA AS RESULTING FROM DIFFERENT EXPERIENCES IN THE ITALIAN ENVIRONMENT

Angelo Zandonella

Telespazio S.p.A.

Presented to the.
"Secondo Congresso Nazionale sul Telerilevamento delle Risorse Terrestri" organized by S.I.T.E.
Varenna, 1979.

AN AUTOMATIC CLASSIFICATION METHOD FOR LANDSAT DATA AS RESULTING FROM DIFFERENT EXPERIENCES IN THE ITALIAN ENVIRONMENT

Angelo Zandonella

Telespazio S.p.A.

ABSTRACT

This work describes a method for Landsat images classification, set up after some years of experience in Italy, particularly in the thematic cartography field.
This method consists of two parts:
- a first part concerning the training sets selection,
- a second part concerning the classification itself.

Training sets are selected by means of a feedback feature selection procedure, employing a method of conditioned hierarchical classification for evaluating the differences between them and their representativeness in the scene.
Classification is performed through K-means algorithm, duly improved and integrated by methods for eliminating non-significant classes, for merging similar classes, for increasing the convergence.

SOMMARIO

Questo lavoro illustra un metodo per la classificazione delle immagini Landsat, messo a punto dopo alcuni anni di esperienza in particolare sulla cartografia tematica in Italia.
Due sono le parti essenziali del metodo:
- la prima riguarda la selezione dei "training sets",
- la seconda, invece, la classificazione propriamente detta.

I "training sets" vengono individuati mediante una procedura di selezione delle caratteristiche, utilizzando un metodo di classificazione gerarchica condizionata per valutare la loro diversità e la loro rappresentatività nella scena.
La classificazione propriamente detta viene effettuata mediante l'algoritmo delle K-medie, opportunamente migliorato nella logica e integrato con metodi per eliminare le classi non significative, per fondere le classi simili, per accelerare la convergenza.

1. INTRODUCTION

Automatic classification methods of Landsat data can be divided into supervised or unsupervised, depending on whether or not spectral signatures extracted by homogeneous sample areas are available for training a classifier.

Considerable literature exists on these methods and on relevant advantages and disadvantages.

In relation to the expected results, the main limit of supervised techniques consists in the extraction of representative spectral signatures in sample areas with small dimensions or with hard geophysical characteristics. On the other hand, the unsupervised techniques do not allow to obtain satisfactory results as supervised techniques. In fact, they do not require initial information about ground-truth and user intervention to control their running is limited. These techniques start with raw data and group pixels into homogeneous sets. The correspondence between these sets and ground-truth data is established in a second step.

Therefore, the best classification approach seems to be the one for which supervised and unsupervised techniques are combined for obtaining the best results.

The purpose of this paper is to describe a Landsat data classification method, based on previous considerations and set up after some years of experience in Italy, particularly in the thematic cartography field.

This method consists of two parts concerning:

- training sets selection,
- classification itself.

Both parts require an user intervention to control their running.

Training sets are selected by means of a learning procedure, based on a feedback system.

Classification requires decisions on: metric selection, elimination of non-significant classes, fusion of similar classes, algorithm convergence.

2. TRAINING SETS SELECTION

THE BASIC CRITERIA

The procedure for training sets selection consists in using minimum-entropy transformation of pair contiguous spectral bands in different steps (refer to Zandonella + Sellman, 1979).

A first step concerns the presentation, on an image processing system video display, of the color composite of the transformed images.

The analysis of these images permits to identify the position and the extension of sample areas (e.g. training sets).

A second step concerns the evaluation of their uniqueness and their representativeness in the scene. This process is effected by a feedback loop (see Figure 1) using:
- feature vectors of spectral band pair, generated via min. entropy method, as input data;
- a conditioned hierarchical classification method for training sets evaluation.

THE CONCEPT OF MINIMUM-ENTROPY TRANSFORMATION APPLIED TO LAND SAT DATA

Consider K pattern classes to be recognized.
Assume that each of the K pattern populations is characterized by a normal probability density function and the covariance matrices, describing the statistics of the classes, are equal.
Let $_{(n,2)}X$, for Landsat satellite data, be a matrix of "n" pixels in 2 pair contiguous spectral bands.
The basic idea of the method consists of determining a linear transformation vector $_{(2,1)}A$ which operates on $_{(n,2)}X$ to yield an image vector $_{(n,1)}Y$ so that the relevant population entropy (used as measure of intraset dispersion of the classes) is minimized. This transformation may be written as:

$$_{(n,1)}Y = {}_{(n,2)}X \cdot {}_{(2,1)}A$$

If one assumes a multivariate normal distribution for each pattern population, this function is characterized by its mean vector and covariance matrix which is, in turn, characterized by its eigenvalues and eigenvectors. These eigenvectors carry the information describing the properties of the patterns under consideration.
In relation to this, it is possible to find a function along which the projection of the dispersion ellipses of the pattern classes are separated the most and inflated the least. This discriminant property can be graphically shown, for pair contiguous bands, as in Figure 2.
For minimum-entropy model, the length of this function is equal to the smallest eigenvalue of the covariance matrix, while the orientation is due to the associate eigenvector.
In fact, the entropy function of $_{(n,1)}Y$ is minimized when we select this type of eigenvector for forming the transformation vector $_{(2,1)}A$, (refer to Tou + Heydorn, 1967, or Tou, 1969).
The vector of six component, so calculated for pair contiguous bands, is used as feature vector for training sets evaluation (refer again to Zandonella + Sellman, 1979).

FORMAL DESCRIPTION OF THE PROCEDURE FOR TRAINING SETS EVALUATION

Let be:

$G'_q \in G_q$: The q-th group of N-q training sets. On this group we would evaluate the dissimilarity degree.

$G''_q \in G_q$: The q-th group of two elements: e.g. the entire scene and the cumulated behaviour of the training sets. On this group we would evaluate the similarity degree.

$\left[S_q(t_i,t_j)\right]$: The correlation coefficient matrix between the elements (t_i,t_j), $\forall i,j$ and for $i \neq j$. Each element is characterized by the above indicated six component feature vector.
The correlation coefficients are used as measures of similarity and/or dissimilarity between (t_i,t_j).

This procedure, derived from a modification to Johnson algorithm (see Johnson, 1967), consists of the following steps:

Step 1
Find two elements of G'_q more correlated.
If t_i and t'_i are these elements, then the following relation is satisfied:

$$S_q(t_i,t'_i) < S_q(t_k,t_l) \quad \text{for } \forall (t_k,t_l) \in G'_q \times G'_q \quad \text{and for } (t_k,t_l) \neq (t_i,t'_i)$$

Step 2
Fusion of t_i and t'_i in one element that we indicate by the symbol $\{t_i,t'_i\}$.

Step 3
Define the ensemble G'_{q+1} which has the same elements of G'_q minus t_i and t'_i. These elements are replaced in G'_{q+1} by the element $\{t_i,t'_i\}$.

Step 4
Compute the max correlations between the N-q-1 elements of G'_{q+1} so defined. If for $\forall t_k,t_l \in G'_{q+1}$, where $t_k \neq t_l$ and $t_k,t_l \neq \{t_i,t'_i\}$, we have:

$$S_{q+1}(t_k,t_l) = S_q(t_k,t_l)$$

the max correlation is the one which satisfies the following relation:

$$S_{q+1}(t_k, \{t_i,t'_i\}) = \text{Max}\left\{S_q(t_k,t_i), S_q(t_k,t'_i)\right\}$$

Step 5
Reiterate again.

The procedure is terminated when the esemble $\{G'_q, G''_q\} \in G_q$ has the following characteristics:
- G'_q exceed the threshold;
- G''_q don't exceed the threshold.

Otherwise we select new training sets.

The results of this procedure may be represented diagrammatically as in Figure 3.
This figure summarizes the relationship between every pair of groups (training sets, entire scene and cumulative behaviour of training sets) in dendrogram form, for a visual evaluation of the training sets.

OUTPUT OF THE TRAINING SETS SELECTION PROCEDURE

So selected training sets are considered as prototypes of relevant classes. They are characterized by measures concerning mainly their position into the spectral bands space (or transformed space via a factor analysis method), their dispersion and their intersection.
Measures of position and dispersion are, respectively:
- means vector of spectral bands,
- covariance matrix among bands.

Measure of intersection between training sets pair is the Swain-Fu distance (refer to Swain + Fu, 1970).
The interest for this measure justifies a short analytical description.
Assume, for a generic class prototype, that:
μ : is the means vector,
C : is the covariance matrix,
N : is the number of spectral bands used.
The Swain-Fu distance between two classes prototypes i and j is defined as:

$$SF = \frac{\| \mu_i - \mu_j \|}{D_i + D_j}$$

where, for a generic class prototype:

$$D = \left\{ \frac{\| \mu_i - \mu_j \|^2 \cdot (N+2)}{tr \left[(C^{-1}) \cdot (\mu_i - \mu_j) \cdot (\mu_i - \mu_j)^T \right]} \right\}^{1/2}$$

This measure is important, from a numerical point of view, for evaluating similar training sets that in the scene are different. This may be due, for instance, to the fact that they are selected in areas with different geophysical characteristics.
In fact, the relationships between Swain-Fu distance and intersection/separation between two classes prototypes are:

SF = 0.	intersecting for 100%
SF = 0.5	intersecting for 50%
SF = 1.	incidents
SF > 1.	separated.

3. CLASSIFICATION

FORMAL DESCRIPTION OF THE PROCEDURE

The algorithm used is based on the minimization of the sum of square error of all pixels, in a class domain, respect to the relevant center. This procedure, known as K-means algorithm, has been duly improved to the aim of reducing the computer time and minimizing the classification error.

Assume that:

x_I : is a generic pixel to classify.

S_1^I : is the 1-th class of x at the iterative step I.

μ_1^I : is the means vector of the S_1^I class.

The essential steps of the procedure are the following:

Step 1

Input of the means vectors and the covariance matrices for all class prototypes.

Step 2

Compute, for each class, the max diameter of the dispersion ellipsoid, to use as correction factor of the distance functions.

Step 3

Assign x to the 1-th class using the relation:

$$\bar{D}(x - \mu_1^I) < \bar{D}(x - \mu_m^I) \quad \forall 1 < m$$

where $\bar{D}$ is the Euclidean or Mahalanobis distance function duly corrected.

Step 4

If required, fusion of similar classes.
Otherwise go to Step 5.

Step 5

If required, elimination of non-significant classes and/or classes with pixels scattered at random in the scene.
Otherwise go to Step 6.

Step 6

Compute, for each class, the means vector and the covariance matrix, to use at the iterative step I+1, such that the sum of square errors from all pixels in S_1^I to the new relevant center is minimized.

Namely:

$$[\text{Min}] \sum_{x \in S_1^I} (x - \mu_1^{I+1})^2 \quad \forall 1$$

Sample means vector and sample covariance matrix of S_1^I minimize this function (refer to Tou + Gonzales, 1974). Therefore they are used as new position and dispersion measures.

Step 7

If there are no significant differences between the classes centers at the iterative steps I and I+1, the algorithm has converged and the procedure is terminated.
Otherwise go to Step 2.

DISTANCE FUNCTIONS

Assume that:

x : is a generic pixel to classify,

μ_1 : is the means vector of the 1-th class,

M : is the number of spectral bands to use in the classification,

W_1 : is an M-dimension quadratic form of 1-th class, defined as positive,

$\emptyset_1$: is the max diameter of the 1-th class dispersion ellipsoid.

The distance function between x and μ_1 is defined as:

$$\bar{D}(x - \mu_1) = D(x - \mu_1) \cdot f(\emptyset_1)$$
$$= \left[(x - \mu_1)^T \cdot W_1 \cdot (x - \mu_1) \right] \cdot f(\emptyset_1) \quad \forall 1$$

where $f(\emptyset_1)$ is a correction factor.
f-function is limited, monotonic and decreasing with the ratio:

$$\emptyset_1 / D(x - \mu_1) \quad \forall 1$$

In relation to this, $\emptyset_1$ has been evaluated, for each class, as two times the sum of the spectral bands unbiased sample variances:

$$\emptyset_1 = 2 \cdot \sum_{i=1}^{M} (\sigma_i^1)^2 \quad \forall 1$$

It is possible to demonstrate (refer again to Tou + Gonzales, 1974) that $\emptyset_1$ is the intraset distance for the 1-th class of pixels.
Different are the criteria of choosing the elements (weight) of W_1 (refer also to Zandonella, 1979).

a. If there are no correlations among bands, then W_1 is chosen as an identity matrix.

$D(x - \mu_l)$ becomes the Euclidean distance.

b. If there are correlations among bands, W_l is chosen as the inverse of covariance matrix of the l-th class, in order to decorrelate them.
$D(x - \mu_l)$ becomes the Mahalanobis distance.

c. If the spectral bands are not comparable each other, because for instance they refer to different acquisition times, it is necessary to prescale them. In this case W_l is chosen as a diagonal matrix, whose elements are inversely proportional to the variance of these bands.
$D(x - \mu_l)$ becomes the weighted Euclidean distance.

FUSION OF SIMILAR CLASSES

Similar classes are merged using two different methods. That is:
- Swain - Fu distance or
- a statistical test.

In the first case, fusion is effected in relation to the intersection percent of similar classes, checked during the training sets selection procedure.
In the second case, fusion is effected without user intervention.
The tested hypothesis is the one for which there are no significant differences between means vectors of two generic classes K and l.
Hotelling T^2 test is used for this evaluation, where:

$$T^2 = \frac{N_K \cdot N_l}{N_K + N_l} \cdot (\mu_K - \mu_l)^T \cdot S^{-1} \cdot (\mu_K - \mu_l)$$

and for:
N : the number of pixels of a generic class,
μ : the means vector of a generic class,
S^{-1} : the pooled covariance matrix of the classes K and l.
Degrees of freedom are:

$$2 \text{ and } N_K + N_l - 2 - 1$$

while the significance level is 0.05.

ELIMINATION OF NON-SIGNIFICANT CLASSES AND/OR CLASSES WITH PIXELS SCATTERED AT RANDOM IN THE SCENE

In these cases it is possible to assume the hypothesis that the observed distributions of classified pixels are those expected from random sampling of a Poisson distribution.
It is known that the variance of a Poisson distribution is, in theory, equal to the mean, for which the mean represents the uni

que parameter necessary for defining this distribution. This fact permits to evaluate the characteristics of one distribution comparing mean and variance, therefore a mean significantly different from a variance sets up non-randomness in the dispersions of the observed population.
In relation to this, it is possible to analyze the two cases as follows.
1-st case. Classes with a non-significant number of pixels. The reduction in the number of points of one class implies a reduction of the difference between mean and variance, for which the points distribution approximates the Poisson distribution.
2-nd case. Classes with pixels scattered at random in the scene.
In this case the mean is not significantly different from the variance, for which the pixels distribution tend to a Poisson distribution.
For both cases the tested hypothesis is the one for which there are no significant differences between mean and variance.
The test used is:

$$\chi_1^2 = \sum_i \frac{\left[m_{il} - p(m_{il}) \right]^2}{p(m_{il})} \qquad \forall 1$$

where:

m_{il} : is the mean of pixels, belonging to the l-th class, for the spectral band i

$p(m_{il})$: is the estimated mean using the Poisson distribution.

Degrees of freedom are:

Number of spectral bands-2

and 0.05 is the level of significance.

CONVERGENCE TEST

The criteria used for the algorithm convergence is based on the statistical analysis of the classes centers changing in two contiguous iterative steps.
The tested hypothesis is the one for which there are no significant differences between the classes centers distributions at iterative steps I and I+1.

χ^2 test is used, to verify this hypothesis, where:

$$\chi^2 = \sum_{i,l} \frac{(m_{il}^I - m_{il}^{I+1})^2}{m_{il}^{I+1}}$$

and for:

m_{il}^{I} : the mean of the pixels, belonging to the l-th class, for the spectral band i at the iterative step I.

Degrees of freedom are:

Number of classes x Number of spectral bands-1

while 0.05 is the level of significance.

4. CONCLUSIONS

The described method for classifying Landsat data is based on the criteria by which supervised and unsupervised techniques are combined for obtaining the best results.
To this aim, the method requires a large user intervention for analyzing images data and, consequently, for deciding the type of mathematic operators to use in the procedure.
Particular techniques are utilized, both for training sets selection/evaluation and for metrics correction, for elimination of non-significant classes, for fusion of similar classes.
In relation to the above, this paper intends to give not only a methodological contribution to the classification problem, but also to invite the development of new techniques for obtaining a correct equilibrium between classification accuracy and computer time reduction.

REFERENCES

Johnson, S.C.
"Hierarchical Clustering Schemes"
Psychometrica, vol. 32, N. 3, 1967.

Swain, P.H., and Fu, K.S.
"Nonparametric and Linguistic Approaches to Pattern Recognition"
LARS Information Note 051970, Purdue University, June, 1970.

Tou, J.T.
"Feature Selection for Pattern Recognition Systems"
Methodologies of Pattern Recognition (edited by S. Watanabe), Academic Press 1969.

Tou, J.T. and Gonzales, R.C.
"Pattern Recognition Principles"
Addison-Wesley Publishing Company, 1974.

Tou, J.T. and Heydorn, R.P.
"Some Approaches to Optimum Feature Extraction"

Computer and Information Sciences-II (edited by J.T. Tou), Academic Press, 1967.

Zandonella, A.
"Objects classification: some methods of interest in astronomical image analysis"
International Workshop on Image Processing in Astronomy, Miramare, June, 1979.

Zandonella, A., and Sellman, B.
"Feature Selection Via Entropy Minimization: An Example Using Landsat Satellite data".
Secondo Convegno sulle Metodologie di Trattamento dell'Informazione organized by S.A.It, Trieste, 1979.

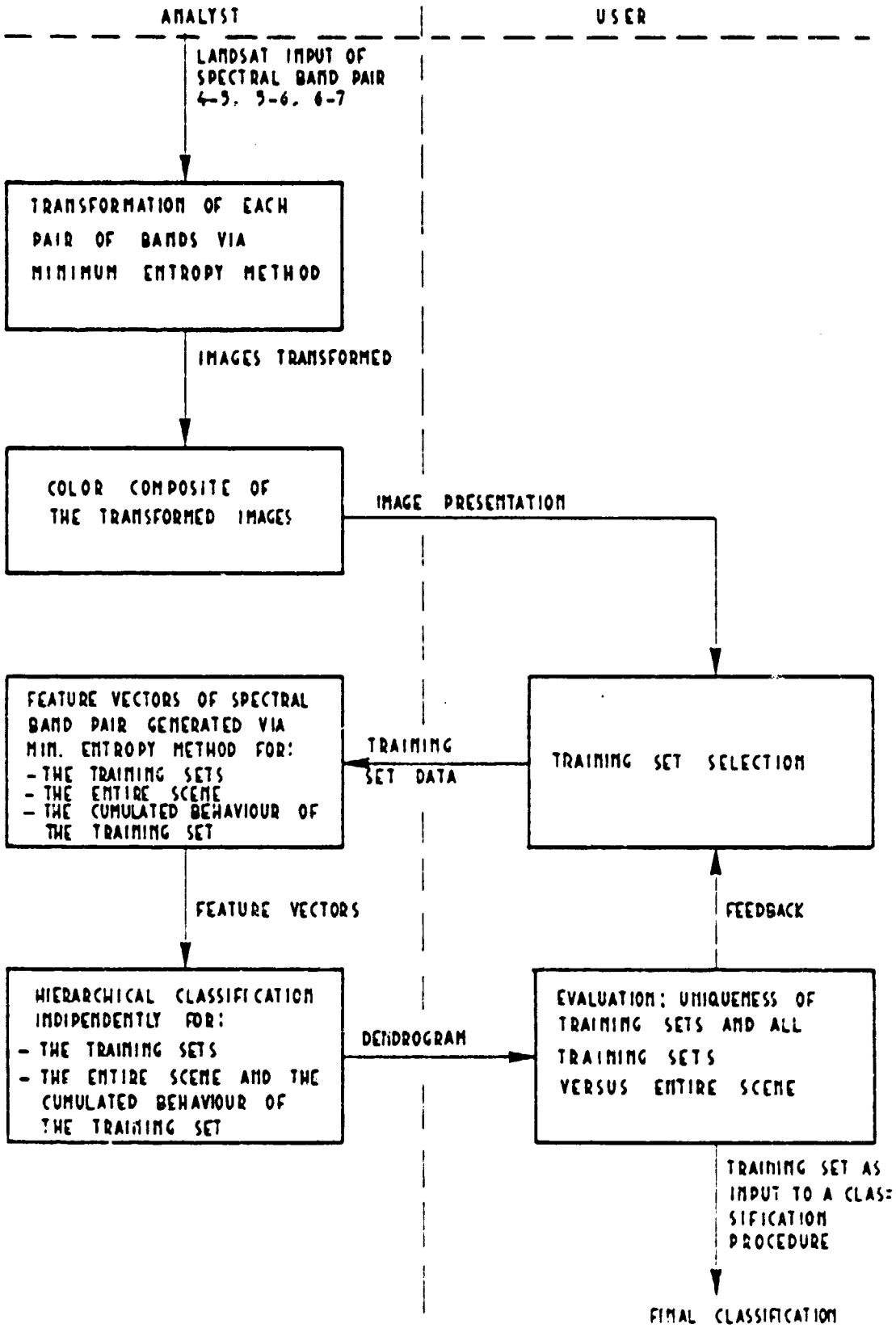

FIGURE 1 - Scheme of the procedure for training sets selection.

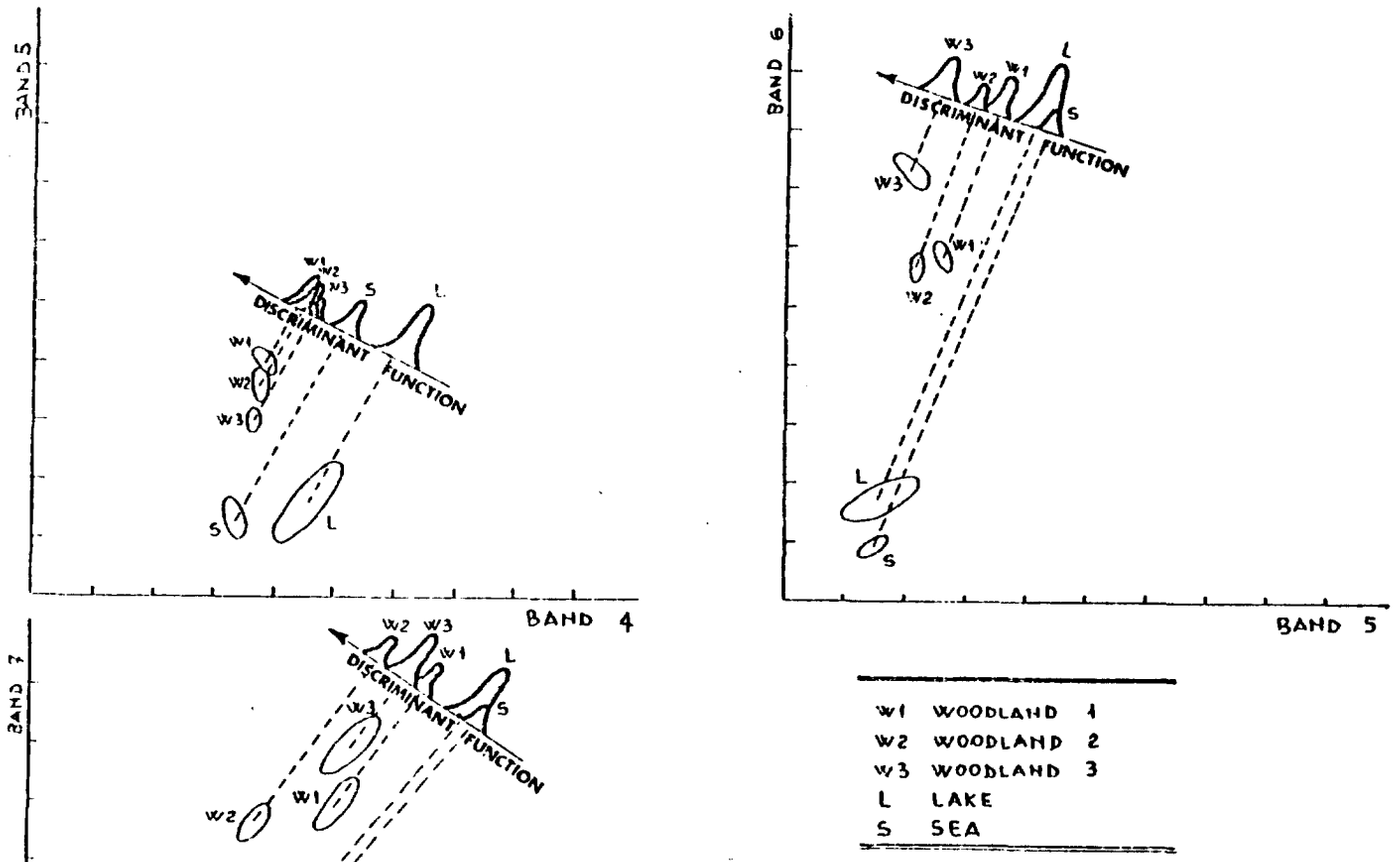

FIGURE 2 - Plot of one sigma dispersion ellipses of different classes in pair contiguous spectral bands.
Classes can be distinguished by projecting ontc the discriminant function line.
Original Landsat-2 data of 13 Sept. 1978 refer to forested shoreline of Migliarino, Massaciuccoli lake, Tyrrhenian sea.

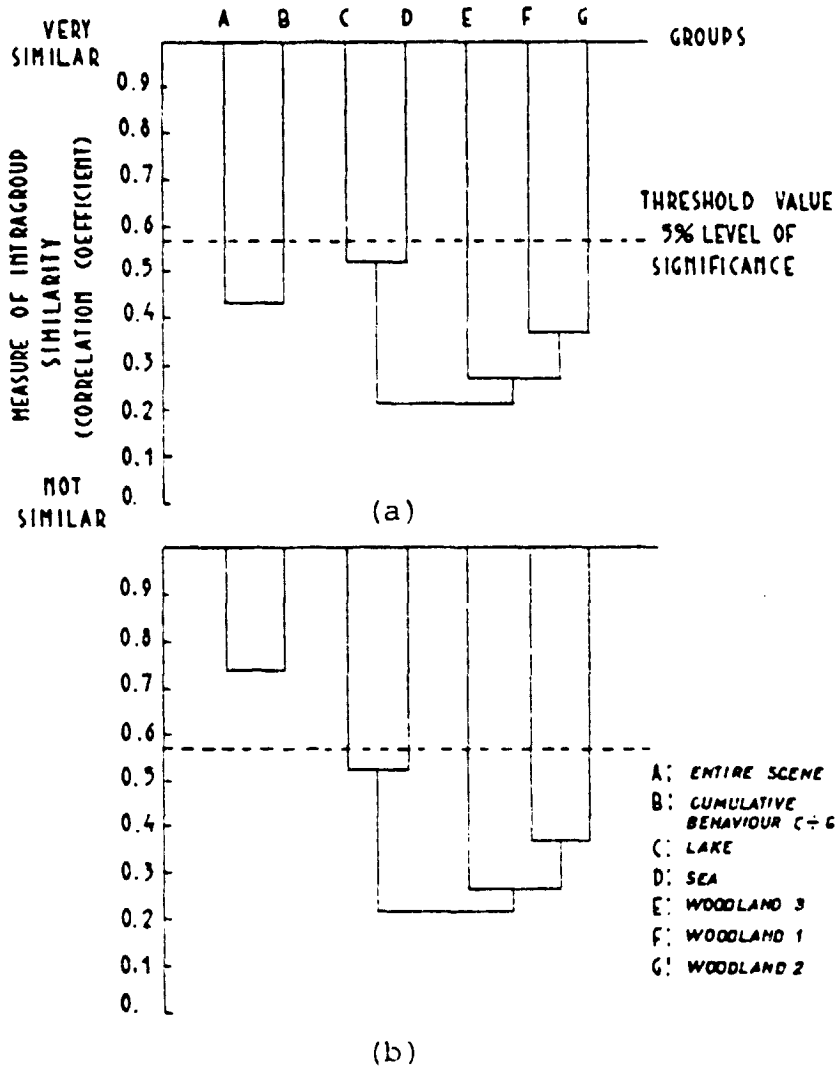

FIGURE 3 - Example of dendrograms obtained after the hierarchical classification.
(a) All training sets are different but not representative of the entire scene.
(b) All training sets are different and representative of the entire scene.