Symposium on Machine Processing of Remotely Sensed Data

June 29 - July 1, 1976

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
76CH1103-1 MPRSD

LANDSAT SIGNATURE DEVELOPMENT PROGRAM

Royce N. Hall*, Ken G. McGuire,* Roy A. Bland**

*Federal Electric Corporation, Kennedy Space Center, Florida
**NASA Earth Resources Office, Kennedy Space Center, Florida

I ABSTRACT

The LANDSAT Signature Development Program, LSDP, is designed to produce an unsupervised classification of a scene from a LANDSAT tape. This classification is based on the clustering tendencies of the multispectral scanner data processed from the scene. The program will generate a character map that, by identifying each of the general classes of surface features extracted from the scene data with a specific line printer symbol, indicates the approximate locations and distributions of these general classes within the scene.

Also provided with the character map are a number of tables each of which describes either some aspect of the spectral properties of the resultant classes, some inter-class relationship, the incidence of picture elements assigned to the various classes in the character map classification of the scene, or some significant intermediate stage in the development of the final classes.

II INTRODUCTION

Numerous analysis techniques are available for the interpretation and display of MSS data. Most of these, to a more or less degree, require judgments from the analyst. Those programs that operate in an unsupervised mode necessitate preinstructions for the program that require technical expertise in order to successfully use the software.

The need arose for a "first-look" totally unsupervised classification for LANDSAT MSS scenes designed for a user who is not necessarily trained in computer science techniques. Such a software package was developed at the NASA/Kennedy Space Center for use on a Honeywell 635. Results from this program have been compared with scenes analyzed by the GE IMAGE-100 system and sophisticated clustering programs. The LSDP results, which were obtained more economically, compared favorably with these other methods of analysis.

III PROCESSING FEATURES

Processing relies on the clustering properties of the data and is designed to provide a 1:24,000 scale character map. The maximum map, per computer run, contains 130 pixels across and 130 pixels down the page. A nearest neighbor scheme as reported in LARS Information Note 103073 by Paul E. Anuta has been adopted as the geometric correction method.

The principle assumption made concerning the data is that the coordinate system can be realigned, via a rotation matrix computed from the matrix of eigenvectors, in order to improve the overall effectiveness of a band-by-band classification approach. Once transformed, the covariant terms are assumed not to be significant and therefore treated as zero. This concession was made primarily because of the additional computer core required by not doing so and because it does not seem to preclude the accuracy sought in the classification. The transformed data set is reduced before rotating by not considering pixels which did not occur at least four times in the scene. This again was a trade-off of classification effectiveness versus computer impact.

The spatial organization of the rotated data is not retained, only the unique transformed pixel values and their frequency of occurence. This data set is then reduced to a set of clusters defined by a mean frequency, and a mean and variance in each band. Each cluster is formed by collecting all pixels in the set within a fixed distance about a seed pixel and then accepting only pixels in the set that do not change the variance by more than the chi-square statistic would permit at a selected level, and this is not more than the associated standard deviations from the mean.

The first seed pixel is the most frequent in the data set and the next seed is the most frequent in the set remaining after forming the first cluster. All non-seed pixels are checked for acceptance to each subsequent cluster formed provided their frequency is less than the seed frequency. The fixed distance about the seed is two maximum projections of the original scale intervals on the rotated axis. This distance is used to compute an initial mean and variance for each cluster before letting them adapt with the chi-square and standard deviation test.

Clusters are next subjected to a merge test. Cluster pairs with mean separation within a certain hyperellipsoid region are merged. The merge region is a function of the clusters mean, variances and mean frequencies and the object of the merge is to insure a significant resultant set of clusters. When all clusters are stable, i.e., do not pass the merge test, they are next inspected for overlap at the three standard deviation range. All overlaps are resolved by the maximum likelihood rule, using the mean frequencies as the "a priori" factors. This results in a set of non-overlapping regions in the data space. Pixels which fall in these regions are assigned unique characters, then mapped by reading again the data set. The mean and covariant matrix of the pixels that fall within these regions constitute the signatures associated with the character map.

IV PROCESSING STEPS

A selected area on a LANDSAT tape is determined and the coordinates placed on a single user input card. The steps below specify the processing to be performed on the pixels.

STEP 1

There is a set of N pixels to be processed in the selected area. First, any pixel which does not occur more than three times is deleted from the set. N is reduced accordingly.

Each pixel is a 4-dimensional vector,

$$X^T = (x_1, x_2, x_3, x_4)$$

To begin, each pixel in the set is to be transformed as follows,

$$Y = \Phi^T X$$

where Φ is the normalized column eigenvector matrix.

The normalized eigenvector matrix is computed as follows,

(a) Compute the mean pixel vector $(\bar{x})$ of the set.

$$\bar{X}^T = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4)$$

$$\text{where, } \bar{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_{in} \quad : i \varepsilon \{1,2,3,4\}$$

(b) Compute the covariance matrix (V) of the set.

$$V = [v_{ij}]$$

where, i is the row index
j is the column index

$$vij = \frac{1}{N}\sum_{n=1}^{N} (x_{in} - \bar{x}_i)(x_{jn} - \bar{x}_j); \quad i,j \varepsilon \{1,2,3,4\}$$

(c) Solve the expression for the eigenvalues (λ).

$$|V - \lambda I| = 0$$

$$\text{where } \lambda^T = (\lambda_1 \ \lambda_2 \ \lambda_3 \ \lambda_4)$$

(d) Now solve for the four eigenvectors (ϕj) from the expression below,

$$(V - \lambda_j I)\phi_j = 0 \ ; \ j \ \varepsilon \ \{1,2,3,4\}$$

(e) Normalize the eigenvectors by dividing by their lengths.

$$\text{length } (\phi j) \equiv ||\phi_j|| = \left[(\phi_{1,j})^2 + (\phi_{2,j})^2 + (\phi_{3,j})^2 + (\phi_{4,j})^2\right]^{1/2}$$

$$\text{normalized } (\phi_j) \equiv \hat{\phi}_j^T = \left(\frac{\phi_{1,j}}{||\phi_j||}, \frac{\phi_{2,j}}{||\phi_j||}, \frac{\phi_{3,j}}{||\phi_j||}, \frac{\phi_{4,j}}{||\phi_j||}\right)$$

(f) The normalized eigenvector matrix is then,

$$\Phi = [\hat{\phi}_1 \ \hat{\phi}_2 \ \hat{\phi}_3 \ \hat{\phi}_4]$$

STEP 2

Many of the pixels in the area will have the same value, thereby resulting in fewer unique pixel values (M) than total pixels (N). The first step is to order by frequency (F) the unique pixel values. This will result in the following set.

$$\{Y_m\} \ ; \ m = 1,M$$

where,

$$F(Y_m) \geq F(Y_{m+1})$$

The next step is to reduce the order set to a smaller set. This smaller set will be the clusters (W). The number of clusters (L) will be less than the number of unique pixels (M).

$$\{W_\ell\} \; ; \; \ell = 1,L$$

Before describing the procedure for determining the pixels that belong to each cluster, it is necessary to specify the description of a cluster. A cluster will be described here by the mean and standard deviation in each dimension (band) of those pixels belonging to the cluster. Also, a factor (q) will be associated with each cluster to enable the computation of a priori probability for later pair-wise comparison. Therefore, a cluster is defined by;

$$W(\bar{Y},\bar{S},q)$$

where, (a) $\bar{Y}^T = (\bar{y}_1,\bar{y}_2,\bar{y}_3,\bar{y}_4)$

$$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{N} y_{ij} \; ; \; i \in \{1,2,3,4\}$$

n = number of pixels belonging to the cluster

(b) $\bar{S}^T = (s_1,s_2,s_3,s_4)$

$$s_i = \left[\frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2\right]^{1/2}$$

$$i \in \{1,2,3,4\}$$

(c) $q = F(\bar{Y})$

STEP 3

Compute the initial set of clusters

$$W_\ell(Y,S,q) \; ; \ell = 1,L$$

This process begins by taking the first unique pixel from the ordered set $\{Y_m\}$. This pixel will serve as a seed pixel,

$Y_s^T = (ys_1,ys_2,ys_3,ys_4)$ for determining the first cluster.

The next steps are done on a band-by-band basis, all tests must be passed in each band to be true.

(a) Beginning with the seed pixels, $Ys^T = (ys_1,ys_2,ys_3,ys_4)$, get all of the unique pixels $Y^T = (y_1,y_2,y_3,y_4)$ in the set such that

$$|ys_i - y_i| \le 2t_i, \text{ and } F(Y) \le F(Ys)$$

where t_i = maximum absolute value element of $\phi\, e_i$, and where e_i is the unit column vector for the band $i \in \{1,2,3,4\}$

(b) Compute the mean vector,

$\bar{Y}^T = (\bar{y}_1,\bar{y}_2,\bar{y}_3,\bar{y}_4)$ and the standard deviation vector, $\bar{S}^T = (s_1,s_2,s_3,s_4)$, of the m pixels found within the $2t_i$ radii of the seed in all four bands; where

$$\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m}(y_{ij}) \; ; \; i \in \{1,2,3,4\}$$

$$s_i = \left[\frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - y_i)^2\right]^{1/2} ;$$

$$i \in \{1,2,3,4\}$$

(c) Next, test the balance of the unique pixels in order of frequency for acceptance into the cluster. If a pixel does not change the cluster variance by more than the chi-squared variable would permit at the input confidence level and is within the associated distance from the mean of the set formed by adding the pixel to the cluster, it is added to the cluster. When a pixel is added, the cluster mean and standard deviation vectors are recomputed before testing the next pixel.

The tests for acceptance of a pixel Y that is outside the $2t_i$ radii about the seed pixel on each of the $i \in \{1,2,3,4\}$ bands are:

$$(1) \quad |\bar{y}_i{}' - y_i| \le c \cdot s_i{}'$$

$$(2) \quad \frac{r \cdot s_i{}^2}{x_u{}^2} \le (s_i{}')^2 \le \frac{r \cdot s_i{}^2}{x_L{}^2} \; ;$$

$$i \in \{1,2,3,4\}$$

where: $Y^T \equiv (y_1,y_2,y_3,y_4)$, the unique pixel value being tested.

$\bar{S}^T \equiv (s_1,s_2,s_3,s_4)$, the standard deviation vector of the cluster.

$\bar{Y}'^T \equiv (\bar{y}_1{}',\bar{y}_2{}',\bar{y}_3{}',\bar{y}_4{}')$, the mean vector of the set formed by adding Y, the unique pixel value being tested, to the cluster.

$\bar{S}'^T \equiv (s_1{}',s_2{}',s_3{}',s_4{}')$, the standard deviation vector of the set formed by adding Y, the unique pixel value being tested, to the cluster.

c ≡ the percentile of the standard normal distribution at the input confidence level.

r≡ degrees of freedom of the cluster.

$x_u^2, x_L^2 \equiv$ the upper and lower percentiles respectively of the standard chi-squared distribution, at the input confidence level.

(d) If no pixel values are found within the $2t_i$ radii about the seed, a lower bound of one one-thousandth is imposed upon the standard deviations of the cluster in all four bands so that a singularity in the multivariate normal distribution that represents the cluster may be avoided. Pixels accepted into the cluster are to remain in the set, but cannot be used as a seed pixel for the formation of any subsequent cluster.

(e) The factor q is now computed.

$$q = F(\bar{Y}) \approx F(Ys)$$

The frequency of the seed pixel is to be used as the best estimate of $F(\bar{Y})$. This then completes the forming of the initial cluster.

STEP 4

Reviewing Step 3, it can be seen that a set of pixels $\{W_\ell\}$ were taken from the ordered set of unique pixels $\{Y_m\}$.

(a) Now to compute the remaining (L - 1) clusters, the process in Step 3 is repeated until the set contains no pixel that can be used as a seed to form a new cluster. A pixel Y is not eligible to be used as such a seed if Y has already been accepted into one or more previously formed clusters or if $F(Y) \leq 4$. When the set $\{Y_m\}$ is devoid of eligible seed pixels, all L clusters have been formed.

The original set $\{Y_m\}$ has now been replaced with an initial set of clusters $\{W_\ell\}$.

STEP 5

The preceding steps have divided the data set into a set of clusters which will now be assembled into the smaller set $\{W_k\}$ k = 1,K. This will represent a final set of clusters and will be those from which the map is generated.

This step is a "merging" step in that the clusters in the $\{W_\ell\}$ set will be checked in a pair-wise manner to determine if they are to be merged.

(a) All pair-wise combinations are tested to determine those which can be merged. The test is as follows,

if $\sum_{i=1}^{4} \frac{\Delta_i{}^2}{d_i} \leq 1$ then the pair can be merged,

where: $\Delta_i^2 = (y1_i - y2_i)^2$

and $d_i = \left(\frac{(A1 + A2)}{A1}\right) \cdot s1_i^2 + \left(\frac{(A1 + A2)}{A2}\right) \cdot s2_i^2$; where

$$A1 = q1 \cdot (s1_1 \cdot s1_2 \cdot s1_3 \cdot s1_4)$$

$$A2 = q2 \;.\; (s2_1 \cdot s2_2 \cdot s2_3 \cdot s2_4)$$

the 1 and 2 suffixes differentiate cluster 1 and cluster 2 statistics, $(q1 \leq q2)$

(b) After all pairs are tested, the nearest mergable pair is merged. The nearest pair is the two clusters with the minimum value of:

$$\sum_{i=1}^{4} \frac{\Delta_i^2}{d_i}$$

The merge will consist of pooling the statistics from the two clusters.

$$\bar{y}new_i = \frac{(n1 \cdot \bar{y}1_i) + (n2 \cdot \bar{y}2_i)}{n1 + n2} \quad \text{and}$$

$$snew_i = \frac{((n1 - 1) \cdot s1_i^2) + ((n2 - 1) \cdot s2_i^2)}{n1 + n2 - 2}$$

$$nnew = n1 + n2$$

$$qnew = q1$$

The new cluster now replaces cluster 1's position in the set and cluster 2 is deleted from the set.

The $\{W_k\}$ set is now to become the $\{W_\ell\}$ set again and step (a) repeated until no merges are possible. When no merges are possible, clusters with less than 30 pixels or consisting of only one unique pixel value (i.e., that of the seed pixel) are deleted. This final set (W_k) is now the set of clusters from which the map will be generated.

STEP 6

At this point in the processing, a method of classifying the pixels from the area to the clusters formed will be specified.

(a) This will first be done by specifying an interval of ±3s in each band, for each cluster, about the means. This can be thought of as a set of 4-dimensional rectangular hypervolumes in the data space.

(b) Cluster boundaries which overlap within these limits are to be reset via the maximum likelihood rule in the band with the least overlap.

A pixel Y whose value in all four bands fall in a region where clusters 1 and 2 overlap is assigned to one and only one of the clusters on the basis of the maximum likelihood decision function:

$$(h1 \cdot p1\ (y_i)) - (h2 \cdot p2\ (y_i)) \lessgtr 0 \rightarrow$$

$$Y \in \begin{cases} W_1 \\ W_2 \end{cases}$$

where:

$$p1 = \frac{1}{2\pi s1_i} \cdot e^{-(y_i - \overline{y1}_i)^2/2s1_i^2}$$

$$p2 = \frac{1}{2\pi s2_i} \cdot e^{-(y_i - \overline{y2}_i)^2/2s2_i^2}$$

$$h1 = (2\pi) \cdot s1_i \cdot q1$$

$$h2 = (2\pi) \cdot s2_i \cdot q2$$

and i is the band where the overlap between clusters 1 and 2 is less than on any other band. This new boundary is to be in the overlap region only.

STEP 7

There now exists a non-overlapping region in the data space associated with each cluster in the set $\{W_k\}$.

Now the mapping will consist of assigning a different print character to each cluster. If a pixel does not fall in the region of a cluster, it is to be printed with a blank character. This will result in a character map of the area selected for unsupervised classification.

STEP 8

The signatures (or classes) will be determined by the set of pixels which fall in the above regions. Each signature will be specified as follows,

(a) $C_k\ (Y,\ V_Y, \eta\ , \gamma\ )\ ;\ k = 1,K$

where Y is the mean vector of the transformed pixels

V_Y is the covariant matrix of the transformed pixels

η is the number of pixels

γ is the print character associated with the class

(b) $C_k\ (X,\ V_x,\ \eta, \gamma)\ ;\ k = 1,K$

where X is the mean vector in the original space

Vx is the covariant matrix in the original space

η, γ are as above

This concludes the procedure for unsupervised classification and mapping of the area.

V OUTPUT DATA

1. Card image listing of the input options card.
2. Interpreted list of card input parameters.
3. Collect table threshold history.
4. Table of samples of tape input data from the selected tape scene.
5. Number of picture elements (pixels) contained in the selected scene.
6. Number of picture elements (pixels) sampled from the selected scene.
7. Number of unique pixel values contained in the stabilized collect table.
8. Number of unique pixel values, with multiplicity of four or more in the scene, processed. (These pixel values will henceforth be referred to as the data set.)
9. Mean vector and covariance matrix of the data set.
10. Eigenvalues of the covariance matrix of the data set.
11. Rotation matrix for the data set. The rows of this matrix are the normalized eigenvectors of the covariance matrix of the data set.
12. The eigenvalues of the covariance matrix, the normalized eigenvalues, their comulative contributions, and the corresponding column eigenvectors listed in order of magnitude.
13. Table of samples of the transformed pixel values; i.e., of the pixel coordinates relative to the rotated axes.
14. Table of the means, standard deviations, and height at the means of and the number of pixels contained in the initial clusters before any merges.
15. Table of the number of pixels used to compute their statistics and the low and high limits of the clusters after all merges and small cluster eliminations but before the overlap among the clusters is resolved.
16. Table of the height at the mean and low and high limits of the clusters after the overlay among the clusters

has been resolved.

17. List of the number of clusters formed, the number of clusters merged, the number of small clusters eliminated, and the number of clusters kept.
18. The character map.
19. Table of the number of pixels assigned to each of the final classes used to generate the character map, the corresponding class symbols, and the percentages of the scene area covered by each class.
20. Tables of the number of pixles assigned to each class and each class's low and high limits relative to both the original and the rotated axes.
21. The mean vectors and covariance matrixes of each class relative to both the original and the rotated axes.
22. Matrix of Euclidean distances between each pair of class means.

VI EXPERIMENTAL RESULTS

Figure 1 shows a February 4, 1975 photo mosaic of the KSC 3 mile long Space Shuttle runway. Figure 2 is a color coded LSDP character printout of this general area from a February 14, 1975, LANDSAT tape. Colors for the clusters which depict various types of vegetation agree very well with ground truth information. This Shuttle runway itself and adjacent roads are too inhomogeneous to cluster in contrast to the homogeneous natural conditions. Figure 3 represents an Image 100 thematic printout of the area surrounding Lake Washington, Florida, from a March 18, 1974, LANDSAT tape. Figure 4 using the same LANDSAT tape is the Lake Washington area as depicted by the LSDP. Both methods agree generally well with ground truth in defining areas of cypress, wet grasses, willow, willow transition, open water and dry grasses. The Figure 4 LSDP printout clearly defines a power line right of way just north of the lake which again demonstrates the usefulness of the program in showing man's intrusion into natural conditions.

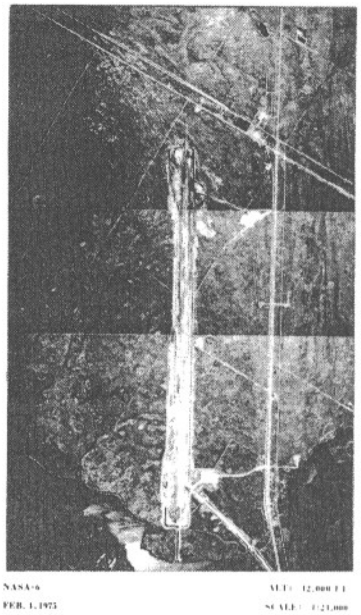

FIGURE 1 - KSC SHUTTLE RUNWAY

*** SPECTRAL DATA CLASSIFICATION PROGRAM ***

FIGURE 2 - KSC SHUTTLE RUNWAY

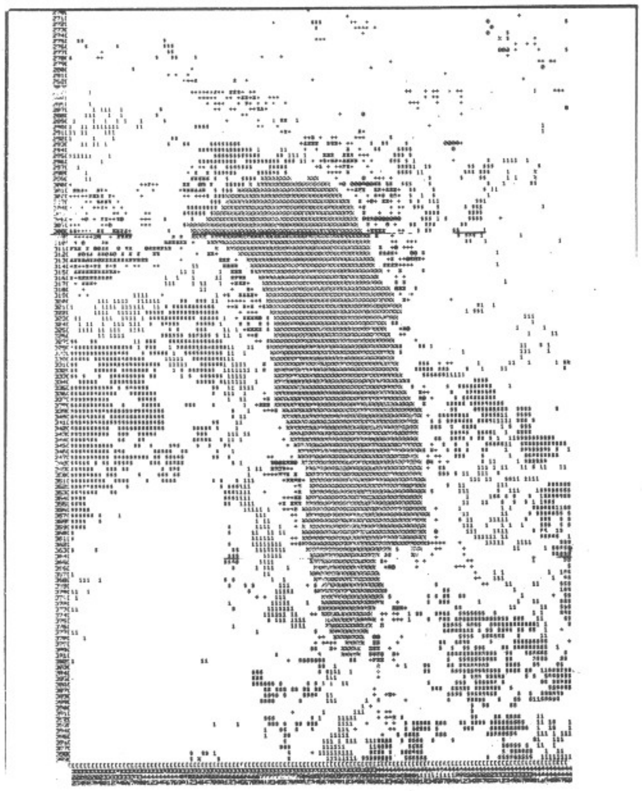

FIGURE 3 - LAKE WASHINGTON, FLORIDA

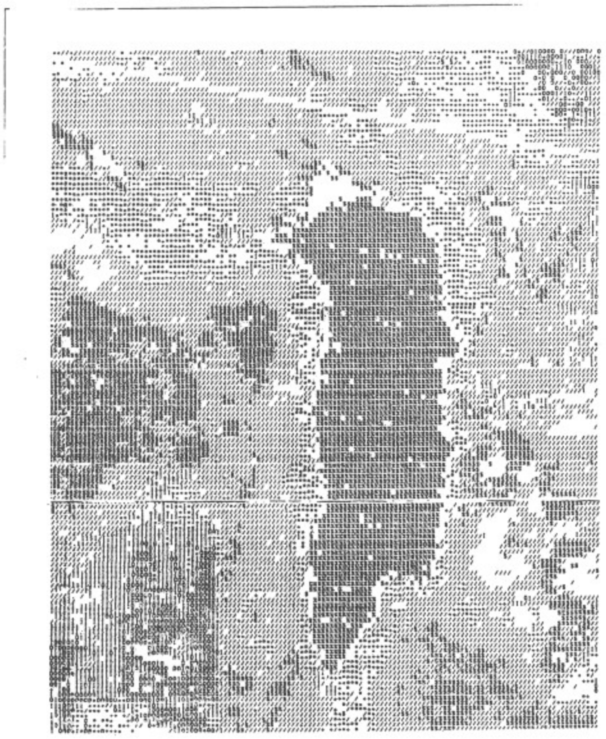

FIGURE 4 - LAKE WASHINGTON, FLORIDA

VII CONCLUSIONS

The LSDP has proven to be a useful and relatively inexpensive tool to classify and analyze the signatures of LANDSAT scenes. LSDP is presently designed to conveniently provide maps and signatures of only a small area not to exceed 130 x 130 pixels. The real value of the program is to develop significant signatures of small areas and the associated map is used to determine the significance of the signatures.

State land use and water quality monitoring officers who have access to modest computing facilities could find this program beneficial to their planning activities. Inquiries may be made to the authors at the KSC Applications Projects Branch, telephone 305-867-7705.

VIII REFERENCES

1. NASA, "Format and Content Specifications for Computer Compatible Tapes," Data Processing Facility, Goddard Space Flight Center, Greenbelt, Maryland, May 1, 1972.
2. General Electric Company, "IMAGE 100 Users Manual," Ground Systems Department, Space Division, Daytona Beach, FL, June 30, 1974.
3. General Electric Company, "Earth Resources Technology Satellite Reference Manual," Space Division, 1972.
4. NASA, "User's Guide and Software Documentation for the Algorithm Simulation Test and Evaluation Program (ASTEP), Revision I," Mathematical Physics Branch, Mission Planning and Analysis Division, Lyndon B. Johnson Space Center, Dec. 4, 1974.
5. Landgrebe, D.A., "Machine Processing of Remotely Acquired Data," LARS Information Note 031573, Purdue University, West Lafayette, Indiana, December 1974.
6. Swain, P.H., "Pattern Recognition: A Basis for Remote Sensing Data Analysis," LARS Information Note 111572 Purdue University, West Layette, Indiana, 1972.
7. Sammon, J.W., "Interactive Pattern Analysis and Classification," IEEE Transaction of Computers, Vol. C-19, No. 7, July 1970.
8. Hunter, H.E., "Application of ADAPT to Integrated Trend Analysis for Checkout of Space Vehicles," ADAPT Service Corporation, reading, Massachusetts, April 1974.
9. Fukunaga, K., "Introduction to Statistical Pattern Recognition," Academic Press, New York, NY, 1972.
10. McGillen, C.D., "Machine Processing of Remotely Sensed Data," Symposium Proceedings, Purdue University, West Lafayette, Indiana, June 3-5, 1975.
11. Veziroglu, N.T., "Remote Sensing," Academia Press, New York, NY, 1975.