NASA CONTRACTOR
REPORT

NASA CR-128999

THE COMPOSITE SEQUENTIAL CLUSTERING TECHNIQUE FOR ANALYSIS OF MULTISPECTRAL SCANNER DATA

By M. Y. Su
Northrop Services, Inc.
P. O. Box 1484
Huntsville, Alabama 35807

October 1972

Prepared for

NASA-GEORGE C. MARSHALL SPACE FLIGHT CENTER
Marshall Space Flight Center, Alabama 35812

TECHNICAL REPORT STANDARD TITLE PAGE

1. REPORT NO.
NASA CR-128999

2. GOVERNMENT ACCESSION NO.

3. RECIPIENT'S CATALOG NO.

4. TITLE AND SUBTITLE
THE COMPOSITE SEQUENTIAL CLUSTERING TECHNIQUE FOR ANALYSIS OF MULTISPECTRAL SCANNER DATA

5. REPORT DATE
October 1972

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
M. Y. Su

8. PERFORMING ORGANIZATION REPORT #
TR-250-1141

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Northrop Services, Inc.
P.O. Box 1484
Huntsville, Alabama 35807

10. WORK UNIT NO.

11. CONTRACT OR GRANT NO.
NAS 8-27364

12. SPONSORING AGENCY NAME AND ADDRESS
National Aeronautics and Space Administration
Washington, D. C. 20546

13. TYPE OF REPORT & PERIOD COVERED
CONTRACTOR

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES
Prepared under the technical monitorship of Mr. Robert E. Cummings, Flight Data Statistics Office, Aerospace Environment Division, Aero-Astrodynamics Laboratory, NASA-Marshall Space Flight Center.

16. ABSTRACT

A new clustering technique is presented. It consists of two parts: a) a sequential statistical clustering which is essentially a sequential variance analysis, and b) a generalized K-means clustering. In this composite clustering technique, the output of (a) is a set of initial clusters which are input to (b) for further improvement by an iterative scheme. This unsupervised composite technique was employed for automatic classification of two sets of remote multispectral earth resource observations. The classification accuracy by the unsupervised technique is found to be comparable to that by traditional supervised maximum likelihood classification techniques.

The mathematical algorithms for the composite sequential clustering program and a detailed computer program description with job setup are given.

17. KEY WORDS
Earth Resources
Remote Sensing
Clustering
Classification

18. DISTRIBUTION STATEMENT
Unclassified - unlimited

E. D. Geissler
Director, Aero-Astrodynamics Laboratory

19. SECURITY CLASSIF. (of this report)
Unclassified

20. SECURITY CLASSIF. (of this page)
Unclassified

21. NO. OF PAGES
61

22. PRICE
NTIS

FOREWORD

This study was undertaken by Northrop Services, Inc., Huntsville, Alabama, for NASA-Marshall Space Flight Center, under contract NAS 8-27364. The work was under the direction of the Flight Data Statistics Office, Aerospace Environment Division, Aero-Astrodynamics Laboratory, with Mr. R. E. Cummings as the project monitor. The computer program developed under this contract will initially be used for processing ERTS-1 data for land use studies in the state of Alabama through a joint effort involving the University of Alabama, the Geological Survey of Alabama, and Marshall Space Flight Center. The author would like to thank Mr. Cummings for his helpful suggestions and support in completing this work.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Section I

INTRODUCTION

This report presents a new automatic processing technique for unsupervised classification (or clustering) of multispectral remote sensing data. This technique has been implemented in a digital computer program. Applications of the computer program to actual multispectral scanner data and digitized multispectral photographs from aircraft surveys will be presented.

In the past, the main approaches for classification were based on various supervised techniques which require reference spectral signatures of targets from training areas on the ground (refs. 1-4). One of the serious drawbacks of these supervised techniques is associated with the high variability of the spectral signatures. It is thus very difficult to set up an operational reference signature library. In general, application of these supervised techniques has required obtaining the reference signatures directly from training areas which form parts of, or are near, each particular survey area. Even with this practice, considerable human judgment and intervention with iterations are required to select proper training areas such that the classification can be of acceptable accuracy.

The unsupervised classification technique avoids the above difficulty by avoiding the reference signatures in the data processing phase (ref. 5-7). Essentially, the technique will group the multispectral data into a number of classes based on some intrinsic similarity within each class. The physical identification of each class is then done after the data processing by checking a small area belonging to each class. In this respect, the procedures in the unsupervised technique are in the reverse order of the supervised technique. The advantage of the processing order in the unsupervised technique is that the investigator will know where to select the ground truth based on the resulting classification map. Another application of unsupervised techniques is for onboard information extraction to minimize the rates of data transmission from future spacecraft to the ground receiving stations. The third advantage is that automatic temporal change detection of earth resources can be more logically carried out by the unsupervised technique.

Section II presents this new unsupervised technique for classifying multispectral remote sensing data which can be either from the multispectral scanner or digitized color-separation aerial photographs. It consists of two parts: a) a sequential statistical clustering which is a one-pass sequential variance analysis, and b) a generalized K-means clustering. In this composite clustering technique, the output of (a) is a set of initial clusters which are input to (b) for further improvement by an iterative scheme.

Applications of the technique using an IBM 7094 computer on multispectral data sets over Purdue's Flight Line C-1 and the Yellowstone National Park test site are given in Section III. Comparisons between the classification maps obtained from the unsupervised technique and from the supervised maximum likelihood technique were also made. Section IV presents a users manual for the computer program.

Section II

MATHEMATICAL ALGORITHMS

2.1 INTRODUCTION

The unsupervised classification technique which will be described is called the composite sequential clustering technique (refs. 8-10). It is a composite of two independent unsupervised techniques, each of which can be used separately for processing the given data set. The accuracy of classification by either technique alone was found to be less accurate than those obtained by the supervised maximum likelihood classification method. However, the composite sequential technique (without using any training data) has produced classifications with accuracy comparable with the above mentioned supervised techniques.

The first part of the composite sequential technique is called sequential statistical clustering. It consists of three main steps:

- Establishing new classes,
- Classifying new data samples into established classes,
- Merging excessive classes.

The second part of the composite sequential technique is called generalized K-means clustering (refs. 11 and 12). It consists of four main steps:

- Estimating initial classes,
- Preliminary improvement of classes,
- Final improvement of classes,
- Classification map and statistical parameters.

The generalized K-means clustering is an improvement to the existing K-means algorithm. The first two steps (of the generalized K-means clustering) are identical to the ones described in references 11 and 12, but the third step is an improved criterion for classification.

Detailed descriptions of these two clustering algorithms separately, and then the integration of these two into the composite sequential clustering technique are presented in the following subsections.

2.2 STATISTICAL SEQUENTIAL CLUSTERING (SSC)

The sensor collects multispectral data from a target which forms an image. An image can be composed of m scanlines of n resolution elements per scanline. Each resolution element is represented by a K-dimensional observation vector.

The purpose of the SSC program is to classify the given sequences of multispectral data into a specified number of subclasses; each of which is statistically homogeneous or similar in their spectral characteristics. To accomplish this goal, the program consists of four main steps:

- Establishing new classes,
- Classifying new samples into established classes,
- Merging excessive classes,
- Displaying classification map and statistics.

A flowchart of the algorithm is depicted in Figure 2-1. These steps as shown in the figure are: Step 1 - all control parameters and statistical tables are read in. Step 2 - M($M \geq 6$) samples are read in, which shall be tested to decide whether they come from the same population. If they do, they will be designated as the first population. If they do not, then the first sample will be dumped into a class of unidentified samples which contains all of the samples unidentifiable, and then a new sample is read as shown in Steps 6 and 7. These new M samples will be tested once again in Step 3 to see whether they constitute a homogeneous population. The above process will be repeated until the first homogeneous population is established. The statistical characteristics of interest for this population are calculated in Step 5.

The program then proceeds to check whether the end of the entire sample sequence is reached. If it is, the program will print out the final results; i.e., the number of samples in that population, the corresponding sample mean, variance and classification map. The latter map represents a two-dimensional spatial location of samples from each population. After this printout, the program will terminate itself. If there are still samples left, the program will proceed to check whether the total number of established homogeneous populations exceeds the prescribed number. If the answer is yes, the program

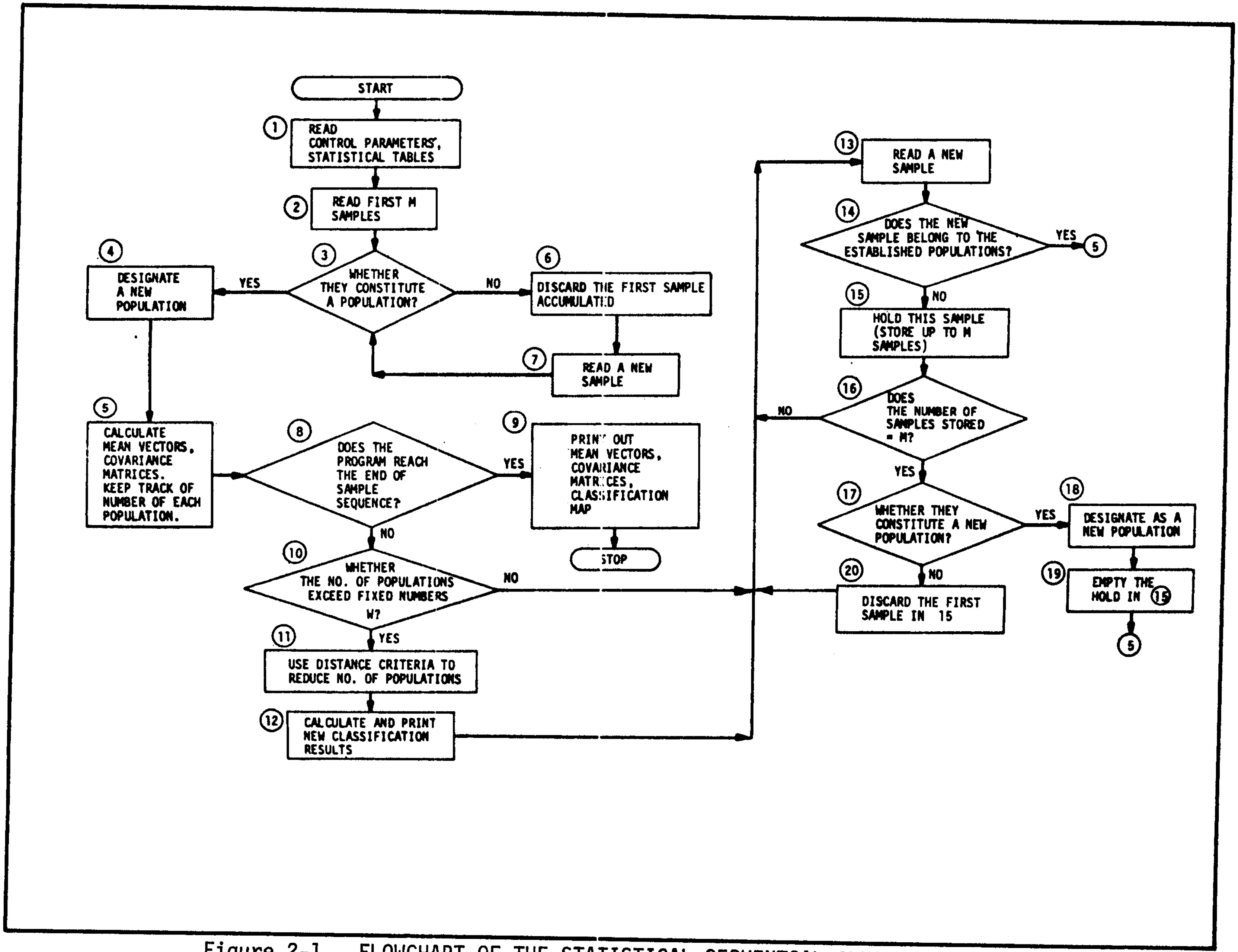


Figure 2-1. FLOWCHART OF THE STATISTICAL SEQUENTIAL CLUSTERING

will proceed to Step 11 to reduce the number of established populations back to the maximum prescribed number. This is accomplished by combining the two populations that are most similar to each other into an enlarged population. Subsequently, the program will also recalculate the corresponding sample mean and variance in Step 12. If the answer is no, then the program proceeds to Step 13 to read in a new sample. The sample is then subjected to another test to see whether it belongs to any of the established population in Step 14. If the answer is yes, the sample is added to that population where it belongs, and the corresponding sample mean and variance are updated. This process is repeated until a new sample is encountered, which does not belong to any of the established populations. This new sample will be held in a temporary hold location until M such samples have been accumulated. These M samples are then tested to see whether they constitute a new population, as is done to establish the first population. If the test is affirmative, then a new population will be set up for them and then continue to Step 5. If the test is negative, the sample which is held first in the temporary hold will be discarded as an unidentifiable sample and then proceed to read in a new sample. This process is repeated until all of the sample sequence has been processed. The final outputs of the algorithm is to print out the number of samples, the mean, the variance of the spectral intensity for each population, and a two-dimensional map of spatial locations of samples for each population, each designated by a different alphanumeric symbol.

The following describe the equations used and the tests performed, if any, in every step of the algorithm. Underlying reasons or justification of the equations or tests will also be given.

- <u>Step 1</u> - Read control parameters and statistical tables,
- <u>Step 2</u> - Read first M samples,
- <u>Step 3</u> - Test for establishing a new population.

This test is to see whether a given small number, M, of samples constitutes a reasonable homogeneous population. Consider each sample as a K-dimensional vector. Let M samples be denoted by $x_{i,k}$, i=1,2, ..., M and k=1,2, ..., K. First, calculate the mean vector $\bar{x}_{1,k}$ and the square of the distance between each sample and the mean vector, i.e.,

$$\bar{x}_{1,k} = \frac{1}{M} \sum_{i=1}^{M} x_{i,k} \quad , \tag{2-1}$$

and

$$\Delta x_i^2 = \sum_{k=1}^{K} (x_{i,k} - \bar{x}_{1,k})^2 \quad . \tag{2-2}$$

Find the maximum along the above M values, i.e.,

$$\Delta x_{max}^2 = \text{Max } [\Delta x_i^2] \quad . \tag{2-3}$$

Now, if

$$\frac{\Delta x_{max}^2}{|\bar{x}_{1,k}|^2} \leq T \tag{2-4}$$

then these M samples will be considered to form a new population, where T is some threshold value to be given and $|\bar{x}_{1,k}|$ is the length of the mean vector. If the above inequality is not satisfied, then the first sample will be discarded. A new sample will replace this discarded sample and repeat the above whole process until the first new population is established.

- Step 4 - Designates a new population when it is established,
- Step 5 - Calculation of statistics for populations.

The purpose of these steps is to maintain and update an inventory of the various statistical parameters for every population established. Let W denote the number of populations already established. The number of samples, mean, mean square, and variance for the i^{th} population will be denoted as m_i, $\bar{x}_{i,k}$, $\bar{R}_{i,k}$, $\bar{V}_{i,k}$, respectively. The recursion formulae are used to compute these parameters.

$$\bar{x}_{i,k}\Big|_{m_i+1} = \frac{m_i}{m_i+1} \bar{x}_{i,k}\Big|_{m_i} + \frac{1}{m_i+1} x_{m_i+1,k} \tag{2-5}$$

$$\bar{R}_{i,k}\Big|_{m_i+1} = \frac{m_i}{m_i+1}\bar{R}_{i,k}\Big|_{m_i} + \frac{1}{m_i+1}x^2_{m_i+1,k} \tag{2-6}$$

and

$$\bar{V}_{i,k}\Big|_{m_i+1} = R_{i,k}\Big|_{m_i+1} - \bar{x}^2_{m_i+1,k} \ . \tag{2-7}$$

- Step 6 - Discard the first of M accumulated samples which do not form a new population,
- Step 7 - Read a new sample,
- Step 8 - Check the end of the sample sequence,
- Step 9 - Print out final results.

The final results to be printed out include the mean and variance for all populations and the classification map.

- Step 10 - Check the number of populations established,
- Step 11 - Reduce excessive populations.

The purpose of these steps is to reduce the number of established populations to a prescribed maximum allowable number, W_{max}. This operation combines the two most similar populations into a single population. The measure of similarity employed is defined to be the Euclidean distance between means of these populations. Calculate the matrix of the pairwise distance square between the means of all the populations established. Next, search for the pair of populations which has the minimum distance and then combine them into one population.

$$D(i,j) = \sum_{k=1}^{K} (\bar{x}_{i,k} - \bar{x}_{j,k})^2 \tag{2-8}$$

$i,j = 1, 2, \ldots, (W_{max} + 1)$.

Next, search for the smallest D(i,j). Let

$$D(i_1,i_2) = \text{minimum}\ [D(i,j)]. \tag{2-9}$$

The i_1th and i_2th population will then be combined into one single population.

- Step 12 - Update statistical parameters of the combined population.

The purpose of this step is to calculate the mean and variance for the newly combined population from the i_1th and i_2th population. The new population will be called the ith population. Let the number of samples in the original two populations be m_{i_1} and m_{i_2}. Then the mean, mean square, and variance become, respectively,

$$\bar{x}_{i,k} = \frac{1}{m_{i_1} + m_{i_2}} (m_{i_1} \bar{x}_{i_1,k} + m_{i_2} \bar{x}_{i_2,k}) \tag{2-10}$$

$$\bar{R}_{i,k} = \frac{1}{m_{i_1} + m_{i_2}} [m_{i_1} \bar{R}_{i_1,k} + m_{i_2} \bar{R}_{i_2,k}] \tag{2-11}$$

$$\bar{V}_{i,k} = \bar{R}_{i,k} - \bar{x}^2_{i,k} . \tag{2-12}$$

- Step 13 - Read a new sample,
- Step 14 - Classify new sample into established populations.

The purpose of this step is to check whether a new sample belongs to any of the homogeneous populations established so far. This test will consist of two parts. The first part is based on χ^2 (Chi-square) distribution and the second part is based on the normal distribution. They will be designated simply as χ^2-test and N-test hereafter and described in detail below.

2.2.1 χ^2-Test For Classification

The sample variance vector of the ith population with the number of m_i samples accumulated is given by:

$$S^2_{m_i,k} = \frac{1}{m_i-1} \sum_{j=1}^{m_i} (x_{j,k} - \bar{x}_{i,k})^2 \tag{2-13}$$

This variation can be used to construct a random variable, i.e., Chi-square variable with $(m_i - 1)$ degrees of freedom, i.e.,

$$\chi^2(m_i-1) = \frac{(m_i-1)\, S^2_{m_i,k}}{\sigma^2_{i,k}} , \tag{2-14}$$

where $\sigma^2_{i,k}$ is the unknown expected variance vector.

In order to estimate the value of $\sigma^2_{i,k}$, the most probable value of $\chi^2(m_i-1)$, is used, i.e.,

$$\chi^2(m_i-1)\Big|_{\text{most prob.}} = m_i - 3 \qquad \text{for } m_i \geq 4 . \tag{2-15}$$

Thus, we get the most probable variation

$$\hat{S}^2_{i,m_i} = \frac{m_i-3}{m_i-1}\, \hat{\sigma}^2_{i,k} \tag{2-16}$$

where $\hat{\sigma}^2_{i,k}$ is the most probable value of $\sigma^2_{i,k}$. The value of $\hat{\sigma}^2_{i,k}$ can be estimated by the least square fit between the expected most probable variance and the actual sample variance. That is,

$$\frac{\partial}{\partial\hat{\sigma}_{i,k}} \sum_{m_i'=4}^{m_i} \left[\frac{S_{m_i',k} - \hat{S}_{i,m_i'}}{\hat{S}_{i,m_i'}}\right]^2 = 0 \tag{2-17}$$

This yields

$$\hat{\sigma}_{i,k} = \sum_{m_i'=4}^{m_i} \frac{m_i'-1}{m_i'-3}\, S^2_{m_i',k} \Big/ \sum_{m_i'=4}^{m_i} \sqrt{\frac{m_i'-1}{m_i'-3}}\; S_{m_i',k} \tag{2-18}$$

where m_i is the current number of samples in i^{th} population.

The actual variance will oscillate around the most probable error curve. The range of oscillation can be expressed by the following confidence interval with (1-2p) x 100 percentage of significance.

$$\frac{\chi_{1-p}(m_i-1)}{\sqrt{m_i-1}} \leq \frac{S_{m_i,k}}{\hat{\sigma}_{m_i,k}} \leq \frac{\chi_p(m_i-1)}{\sqrt{m_i-1}} \quad , \tag{2-19}$$

where $\chi_{1-p}(m_i-1)$ is the (1-p) x 100 percentage point of χ^2-distribution with (m_i-1) degree of freedom.

Now, the quantity to be tested, i.e., $\frac{S_{m_i,k}}{\hat{\sigma}_{m_i,k}}$ is a vector with K components. Therefore, the inequality (2-19) actually consists of K inequalities. Further, the lower and upper limits of these inequalities are independent of the component, i.e., k. Thus, the inequality is replaced by

$$\frac{\chi_{1-p}(m_i-1)}{\sqrt{m_i-1}} \leq \max_k \left[\frac{S_{m_i,k}}{\hat{\sigma}_{m_i,k}}\right] \leq \frac{\chi_p(m_i-1)}{\sqrt{m_i-1}} \ . \tag{2-20}$$

A new sample will not be considered to belong to the i^{th} population, if the inequality (2-20) does not hold. On the other hand, if the inequality holds, the new sample may belong to the i^{th} population. However, since a new sample under test for classification may satisfy the inequality for more than one established population, a further test (or criterion) is needed to make the final classification. This is accomplished by the following N-test.

2.2.2 N-Test For Final Classification

The N-test for resolving the uncertainty in the χ^2-test is to examine the normalized quantities for the new sample $x_{j,k}$ under test using the current sample means and standard deviations, respectively, for all established populations which satisfy the inequality (2-20). That is,

$$y_{i,k} \equiv \frac{x_{j,k} - \bar{x}_{i,k}}{S_{m_i,k}} \tag{2-21}$$

for all i satisfying the inequality (2-20). Next, search for the minimum overall $y_{i,k}$ and assign the new sample to the population in which $y_{i,k}$ is the minimum.

- Step 15 - Hold samples pending for establishing a new population,
- Step 16 - Check the number of samples held in Step 15 to be equal to M,
- Step 17 - Same as Step 3,
- Step 18 - Same as Step 4,
- Step 19 - Empty the samples in Step 15,
- Step 20 - Same as Step 6.

2.3 GENERALIZED K-MEANS CLUSTERING (GKMC)

The generalized K-means clustering consists of four steps as described below.

2.3.1 Step 1 – Estimation of Initial Cluster Centers

The speed of convergence as well as the clustering accuracy depends on the choice of initial cluster centers (refs. 11 and 12). In general, it is desirable that the initial cluster centers be distributed over the populated region of the same space rather than concentrated in one part of it. One procedure used to obtain such a distribution of cluster centers is as follows. The first sample in the sample sequence to be processed is designated cluster center number 1. The distances of the remaining samples from this one are calculated, and the farthest sample is designated cluster center number 2. The smaller of the two distances from each sample to these two centers is listed, and the sample having the greatest minimum distance is selected as cluster center number 3. The remaining centers are chosen in turn to have maximum separation from the existing centers. These initial cluster centers are well scattered over the multispectral sample space that is an intuitively desirable property.

2.3.2 Step 2 – Preliminary Improvement of Cluster Centers

The W initial cluster centers to be improved through this step are the mean spectral vectors for all the established populations output from Step 1. The minimum distance is employed as the similarity criterion. The entire data sample sequence is classified into these W populations by calculating the distance of each sampel with respect to every cluster center and assigning the sample into that particular population that yields the minimum distance. That is,

$$x_{j,k} \rightarrow W_i$$

if

$$\sum_{k=1}^{K} (x_{j,k} - \bar{x}_{i,k})^2 \text{ is the minimum overall i.}$$

This classification is equivalent to set up a system of hyperplane decision boundaries to separate K clusters. Once this is done, the sample belonging to each cluster center is used to calculate its updated mean vector. These updated K centers will now be regarded as the new initial cluster centers for the next iteration. The procedure will be repeated until the difference (or distance) between two successive iterated values of every cluster center is smaller than some prescribed threshold value, or until some fixed number of iterations is performed. At the present, the latter approach is adopted.

2.3.3 Step 3 – Final Improvement of Cluster Centers

The reason for requiring some further improvement to the cluster centers as obtained from Step 2 is illustrated in Figure 2-2. In this figure, there are three natural clusterings in two-component scattering diagrams. Further, these three clusters are clearly linearly separable. To separate the samples into three clusters, Step 2 is used. The best results obtainable, after a sufficient number of iterations, is shown by the linear minimum-distance decision boundary as indicated by the solid lines. Parts of the samples actually belonging to cluster No. 1 are misclassified into clusters No. 2 or 3. This resulted from the fact that the intradistance of cluster No. 1 is greater

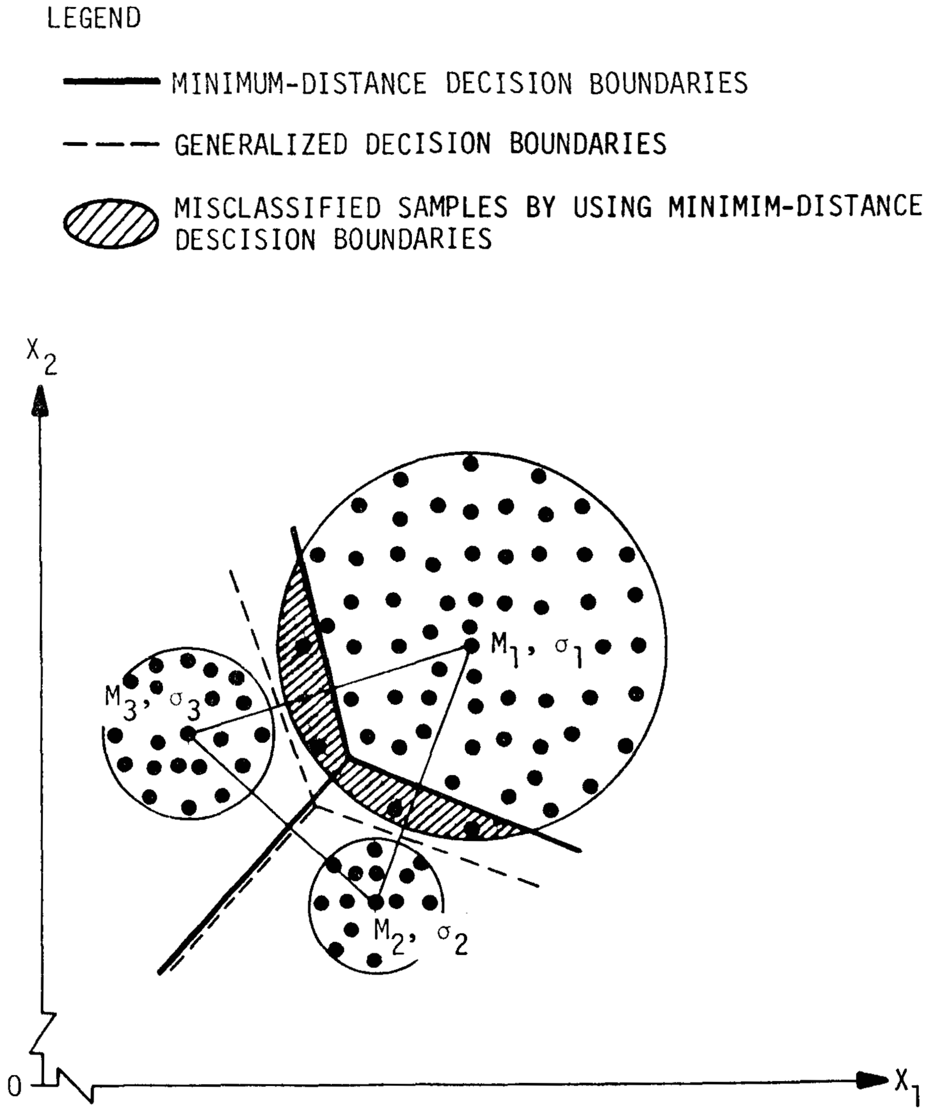

LEGEND
MINIMUM-DISTANCE DECISION BOUNDARIES
GENERALIZED DECISION BOUNDARIES
MISCLASSIFIED SAMPLES BY USING MINIMIM-DISTANCE DESCISION BOUNDARIES
X_2
M_1, σ_1
M_3, σ_3
M_2, σ_2
0
X_1

Figure 2-2. COMPARISON OF THE PRESENT AND GENERALIZED K-MEANS ALGORITHM

than half of the interdistance between clusters No. 1 and 2 (similarly for cluster No. 1 and 3).

One way to deal with this difficulty and thus improve the power of the original K-means algorithm is proposed as follows. The intradistance of samples within one cluster will be proportioned to the sample standard deviation vector. Except for the very elongated cluster, this sample standard deviation vector may be characterized by a single scalar, i.e., the root mean square of the standard deviations of the components. This characterization is correct if each component has the same standard deviation. With this basic assumption, the minimum-distance criterion used in the traditional K-means algorithm is replaced by a more general measure of similarity using the standard deviations as weights to locate the decision hyperplanes.

This improved similarity criterion can be expressed as follows. Compute

$$\frac{1}{S_i^2} \sum_{k=1}^{K} [x_{j,k} - \bar{x}_{i,k}]^2 \qquad \text{for } i = 1,2, \ldots, W \qquad (2\text{-}22)$$

If this is minimum over all k, then

$$x_{j,k} \rightarrow W_k$$

where S_i is the characterized sample standard deviation for the ith cluster center. The rest of the step will be the same as in Step 2.

Three important points germane to the added step will now be discussed. First, one might ask why not use Step 3 with the more general similarity exclusively, i.e., eliminating Step 2 altogether. The answer is that the sample standard deviations for K cluster centers may not be accurate enough at the first few iterations in improving the cluster centers and that they are more sensitive to the influence of misclassified samples than the mean vector of clusters. Hence, there is no strong reason to expect better performance from Step 3 than Step 2 at the first few iterations. On the other hand, if Step 2 is employed to its utmost capacity, then followed by Step 3, a better estimate of the sample standard deviations can be obtained and the true power of the more general similarity will prevail.

The second point is concerned with whether Step 3 with additional evaluation of sample standard deviations will be very time consuming. The answer is no, since in Step 3, as well as Step 2, the square of the distance of each sample with respect to each cluster center needs to be calculated and classified to the cluster center with the shorter distance. The evaluation of sample variance for each cluster center can make use of the above calculation by simply adding an accumulation operation for each sample. Therefore, each iteration of Step 3 will take only slightly more time than that of Step 2.

The last point is that Step 3 will not degrade the results from Step 2. It has been demonstrated that misclassification may occur by Step 2 only if the intradistance of samples in any cluster center is greater than half of the interdistances between the two clusters. Further, the added Step 3 can remedy this difficulty. Step 3 will do just as well as Step 2 for the cases that Step 2 can do perfectly, i.e., the cases in which the interdistance between two clusters is much larger than twice of the intradistances of either individual cluster.

It is worthwhile to note that the cluster centers established by Steps 2 and 3 can be joined or merged together to reduce the total number of cluster centers. However, to save computation time, it will be better to start off using fewer clusters than merging the established clusters.

2.3.4 Step 4 – Classification Map and Statistical Parameters

The results of clustering by Step 3 can be displayed in a two-dimensional map for the multispectral observations such as the multispectral scanner. Each population is designated by a given alphanumeric symbol. In addition, all the statistical parameters and sample probability density functions can also be calculated at the last iteration of Step 4 and printed out together with the classification map.

2.4 COMPOSITE STATISTICAL SEQUENTIAL CLUSTERING

So far, the statistical sequential clustering and the generalized K-means clustering have been described separately. Now, the adyantages and limitation for each technique will be examined. The single most significant advantage of the SSC is that it requires only one pass of the entire data sequence to achieve fairly good clustering of the given data. This truly sequential feature is unique among existing clustering techniques. However, because of only one pass of the data sequence, the class of those unidentifiable samples that are passed over during establishing new classes cannot be reexamined. This is the main drawback of the SSC technique.

The most significant advantage of the GKC technique is that it possesses the capability for repetitive correction and updating of the established cluster centers. Its main drawbacks are that the procedure for choosing the initial cluster centers is either arbitrary or requires as many passes of the entire data sequence as the number of cluster centers (ref. 11). Furthermore, because of these inaccurate initial cluster centers, many iterations of the entire data sequence will be further required to achieve good clustering accuracy.

From the above assessment of these two techniques, it is clear that they can complement each other since the drawbacks of each technique can be eliminated by properly combining the two techniques. The composite sequential clustering technique is then composed of two steps:

- The given data sequence will be processed by the SSC technique with only a single pass of the data sequence. The outputs of the processing will be the mean spectral vectors of clusters.
- The mean spectral vectors from the first step will be used as the initial cluster centers for the GKC technique. To allow for extra cluster centers from the unidentified samples of the SSC, Step 1 of the original K-means clustering can be used for establishing as many extra initial cluster centers as desired. Next, the initial cluster centers will be iterated about two or three times to obtain the final clustering.

In short, the above composite clustering technique can accomplish good unsupervised classification of a given data sequence with about four passes of the entire data set regardless of the preset maximum number of clusters.

2.5 PREPROCESSING OF MULTISPECTRAL SCANNER DATA

For higher classification accuracy of the multispectral data, it is sometimes necessary to preprocess the original multispectral data. The types of preprocessing required depend on the types of data themselves, as well as the objectives of classification. In this subsection, several types of preprocessing are described. In addition, several useful options for selecting and rearranging the multispectral data from the original data tape are described. All of the data preprocessings discussed below have been implemented into a subroutine which constitutes part of the composite statistical sequential clustering program. A more detailed description of each option is given below.

2.5.1 Spectral Channel Selection

Let the input (original) multispectral scanner data under consideration be denoted by

$$x_{ij,k}$$

where

i is scan number, (1, 2, ..., N)

j is sample number along a scan (1, 2, ..., M)

k is spectral channel number for each sample (1, ., ..., K).

The option for the spectral channel selection will transform the given data set into a new set with fewer spectral channels as specified by the user.

2.5.2 Spectral Normalization

This option will transform the given data set into a new set, $y_{ij,k}$, normalizing each sample by the sum of amplitudes from all channels belonging to that sample. That is,

$$x_{ij,k} \rightarrow y_{ij,k}$$

where

$$y_{ij,k} = \frac{x_{ij,k}}{\frac{1}{K} \sum_{k=1}^{K} x_{ij,k}} .$$

Hence

$$\sum_{k=1}^{K} y_{ij,k} = 1.$$

2.5.3 Scan Angle Correction

This option will provide a rough scan angle correction to the original data set. This is effected by normalizing each original data sample by the mean spectral intensity along the same scan angle which corresponds to the sample number. That is,

$$x_{ij,k} \to y_{ij,k}$$

where

$$y_{ij,k} = \frac{x_{ij,k}}{\bar{x}_{j,k}}$$

$$\bar{x}_{j,k} = \frac{1}{N} \sum_{i=1}^{N} x_{ij,k} .$$

2.5.4 Equal Spectral Weight Transformation

This option will provide an equal weighting to each spectral channel based on the mean spectral intensity over the entire data set of interests. That is,

$$x_{ij,k} \to y_{ij,k}$$

where

$$y_{ij,k} = \frac{x_{ij,k}}{\bar{\bar{x}}_{k}}$$

$$\bar{\bar{x}}_{k} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} x_{ij,k} .$$

2.5.5 Principal Axis Transformation

This option will first compute the covariance matrix of the entire data set of the multispectral data. The covariance matrix is then used for computing the eigenvalues and corresponding eigenvectors. Each sample is then projected onto a selected number of eigenvectors with larger eigenvalues.

Specifically, let the symmetric covariance matrix be C. Then, the eigenvalue λ_i, i = 1,2, ..., K (arranging in the descending order) will be given by the characteristic equation

$$|C - \lambda_i I| = 0$$

where I is the identity matrix. The following equality holds for these eigenvalues

$$\sum_{i=1}^{K} \lambda_i = \text{Trace of C.}$$

The eigenvector, $\bar{e}_i$, (column vector), associated with λ_i, is given by

$$C \bar{e}_i = \lambda_i \bar{e}_i .$$

The eigenvector associated with the largest eigenvalue is called the principal axis of the entire data set. Choose the first few eigenvectors, say $\bar{e}_1$, $\bar{e}_2$, $\bar{e}_3$, associated with the first three larger eigenvalues to form a new orthogonal feature space. The original data set is then projected onto this feature space. That is,

$$x_{ij,k} \rightarrow y_{ij,k'} \qquad k' = 1, 2, 3$$

where

$$y_{ij,k'} = x_{ij,k} (\bar{e}_1, \bar{e}_2, \bar{e}_3)$$

where $y_{ij,k'}$ and $x_{ij,k}$ are row vectors.

2.5.6 Data Selection and Rearrangement

This option contains three independent suboptions for generating an intermediate data tape for the unsupervised classifications. They are described below.

- Selection of a rectangular portion of data, which is defined by the desired starting and ending scan numbers, and the starting and ending sample numbers.
- Skipping every fixed number of samples along each scan.
- Skipping every fixed number of scans along the flight direction.

Section III

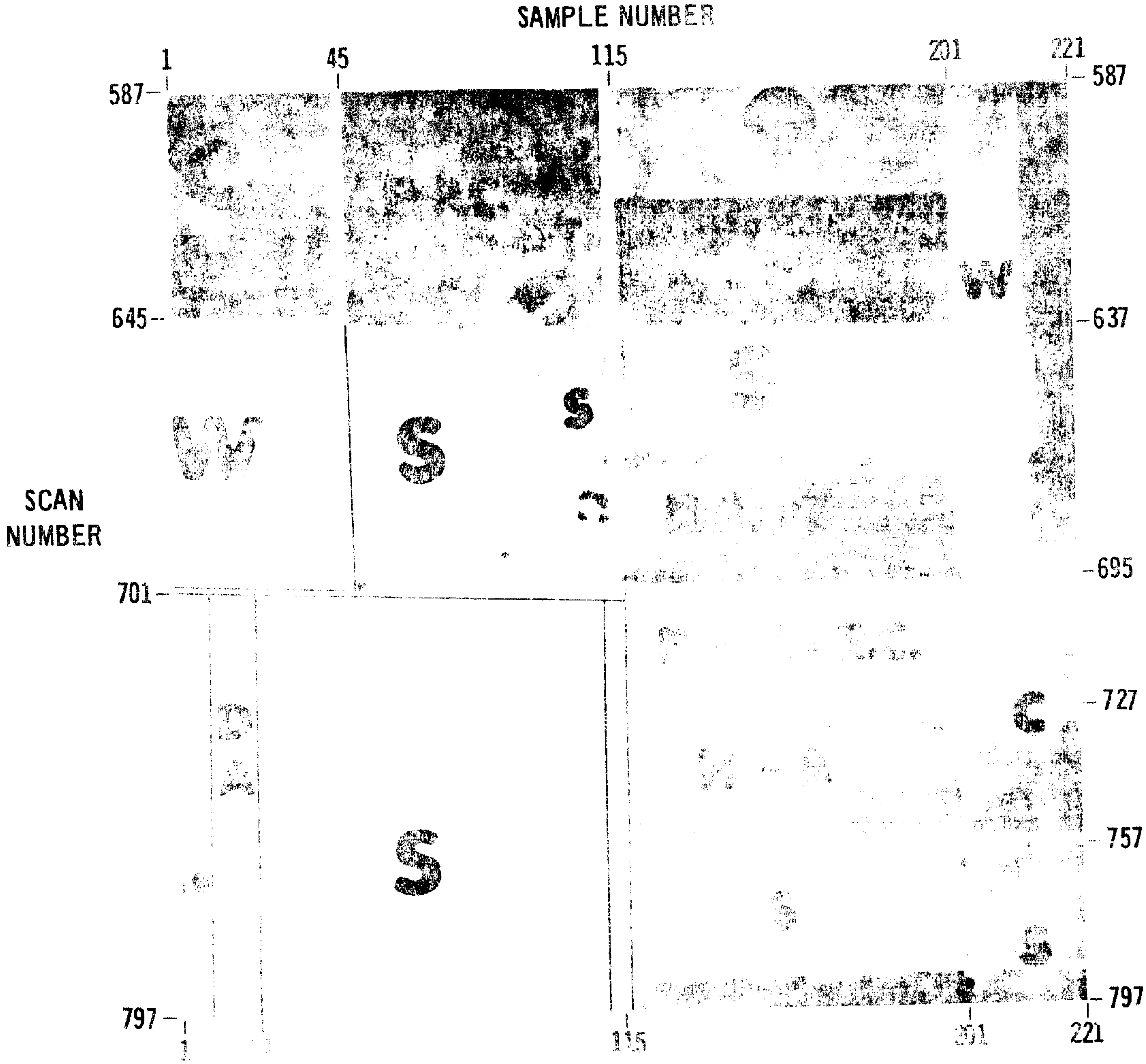
APPLICATIONS

The composite sequential clustering technique has been applied to two sets of multispectral data obtained by the University of Michigan's 12-channel scanner over Purdue's Flight Line C-1 (ref. 2) and the Yellowstone National Park test site (ref. 3). To compare the unsupervised classification results by using the composite clustering technique with results obtained by the Purdue's LARS supervised maximum likelihood technique, the same portions of the test site shall be presented and the same best four-channel selections for corresponding test site will be used as reported by LARS. Thus, for Flight Line C-1 Channels 1 (0.4-0.44 μm), 6 (0.52-0.55 μm), 10 (0.66-0.72 μm), and 12 (0.8-1.0 μm) are used. For the Yellowstone data Channels 2 (0.44-0.46 μm), 9 (0.62-0.66 μm) 10 (0.61-0.72 μm) and 12 (0.8-1.0 μm) are used.

Figure 3-1 shows an aerial photo of the test area which forms part of Flight Line C-1. The area is about one square mile with the ground truth designation superimposed.

In order to show the advantages of the composite clustering technique over the statistical sequential clustering or the generalized K-mean clustering, individually, the classification maps by these three clustering techniques are presented as given in Figures 3-2, 3-3, and 3-4, respectively. In these three maps, the same number of classes, namely 13, has been specified to be the maximum allowable number of classes.

Figure 3-2, which is obtained by employing only the statistical sequential clustering, shows some blank areas in the map, which belong to the class for unidentifiable samples, at the end of the single pass. It is noted that even with only one pass of the data set, the classification has already about 60-70 percent accuracy according to the ground truth map. Figure 3-3 shows the classification map by the generalized K-means clustering only obtained after 15 passes of the entire data set, that is, 13 passes for initial cluster estimates plus two for improvement. The classification accuracy seems to be



LEGEND:

A - Alfalfa	S - Soybeans
C - Corn	T - Timothy
H - Hay	W - Wheat
O - Oats	DA - Diverted Acres
P - Pasture	RC - Red Clover
R - Rye	

Figure 3-1. AERIAL PHOTO OF PURDUE FLIGHTLINE C-1 (SCAN 587-797)

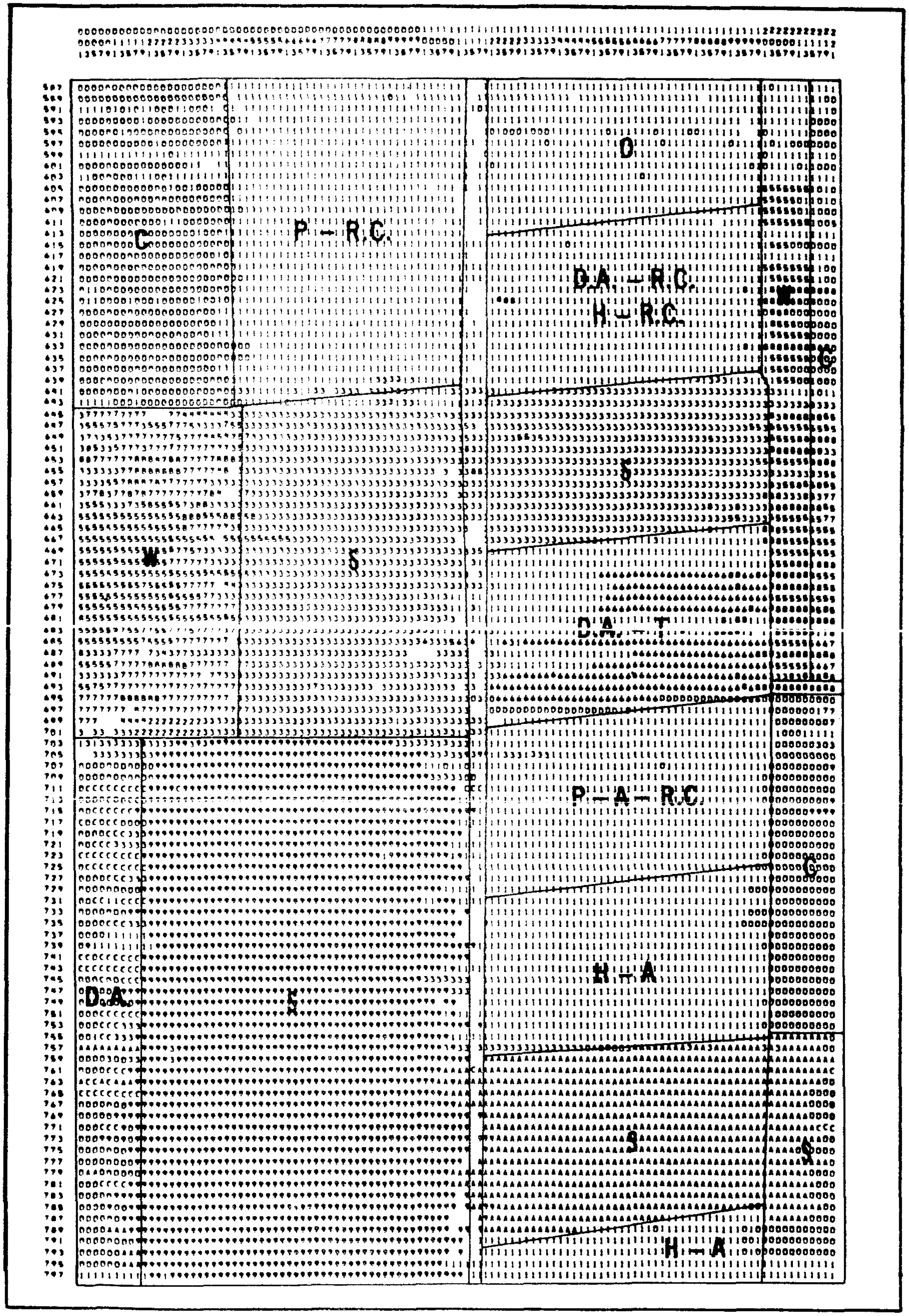


Figure 3-2. UNSUPERVISED CLASSIFICATION MAP OF C-1 FLIGHT LINE BY THE STATISTICAL SEQUENTIAL TECHNIQUE WITH 13 CLASSES AND ONLY ONE PASS OF THE DATA SET

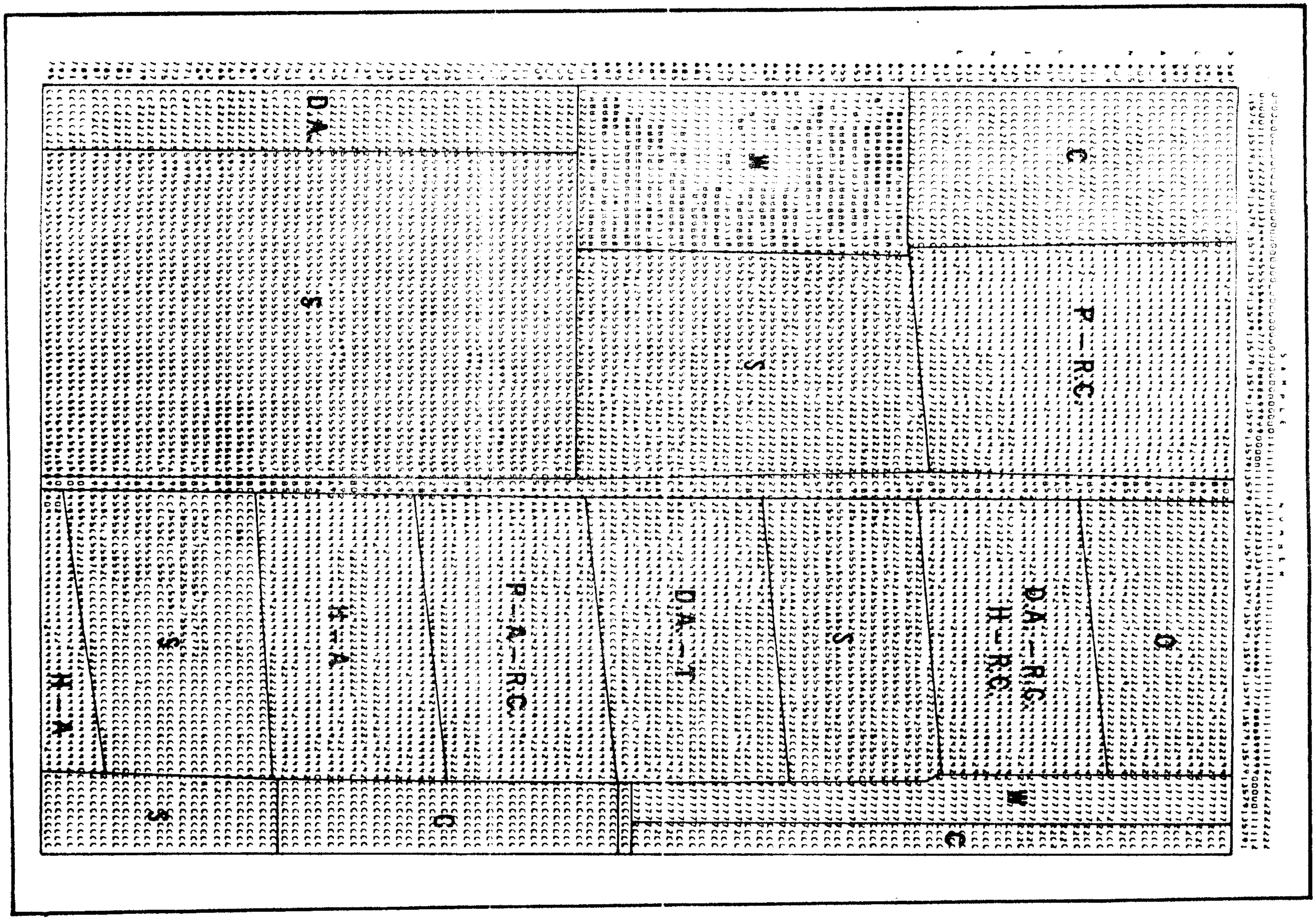


Figure 3-3. UNSUPERVISED CLASSIFICATION MAP OF C-1 FLIGHT LINE BY THE GENERALIZED K-MEANS TECHNIQUE WITH 13 CLASSES AND 15 PASSES OF THE DATA SET

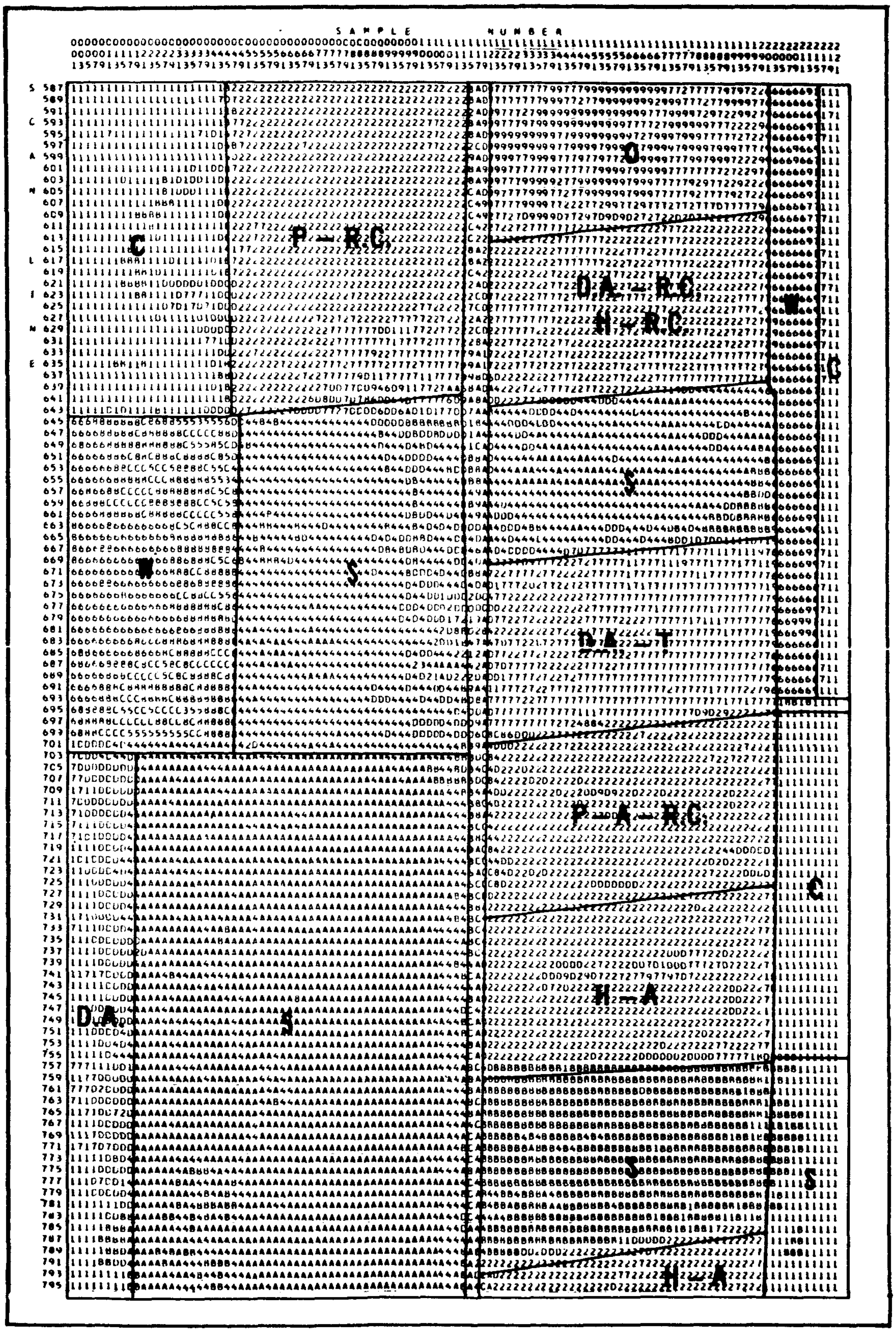


Figure 3-4. UNSUPERVISED CLASSIFICATION MAP OF C-1 FLIGHT LINE BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 13 CLASSES AND 3 PASSES OF THE DATA SET

slightly higher than that by the statistical sequential clustering. This slight increase in accuracy was obtained by an additional 14 passes of the data which is a high price to pay. Figure 3-4 shows the classification map obtained by the composite clustering technique with a total of three passes of the data set. The classification accuracy is about 80 percent which is higher than that obtained by the K-means clustering. The above comparison among the three different clustering techniques clearly shows that the composite sequential clustering technique has both higher accuracy and higher efficiency.

As mentioned earlier, the reason for choosing the particular data set for testing the composite clustering technique is for comparison with the classification results obtained by Purdue LARS using the supervised Bayes classification technique. LARS's classification map is reproduced in Figure 3-5 (ref. 8). The training fields used in the LARS classification program are outlined with asterisks (*) and the test fields are outlined with plus (+) signs. The test fields chosen in LARS classification covered only about 51.5 percent of the entire field. The overall performance of correct classification is 87.5 percent. Actually, the entire field has been classified by the LARS program, as is evidenced by the classification symbols covering the entire field. The so-called test fields in the map are just the "selected" areas for computing the accuracy of correct classification. One can see clearly that the overall performance would be less than the cited 87.5 percent if the overall performance is based on the entire field. It is also noted from the LARS classification results, as well as the unsupervised classification, that red clover, hay, and alfalfa are fairly similar to each other. Comparing the classification map by the composite clustering technique (Figure 3-4) with the LARS's results and with the ground truth aerial photo (Figure 3-1), the overall performance by the composite clustering technique over the entire field is close to 80 percent. That is, the overall performance by the LARS supervised classification technique and by the unsupervised composite technique are comparable.

The Purdue Flight Line C-1 covers a cultivated agricultural area, so it is not a very critical test case for the composite clustering technique. On

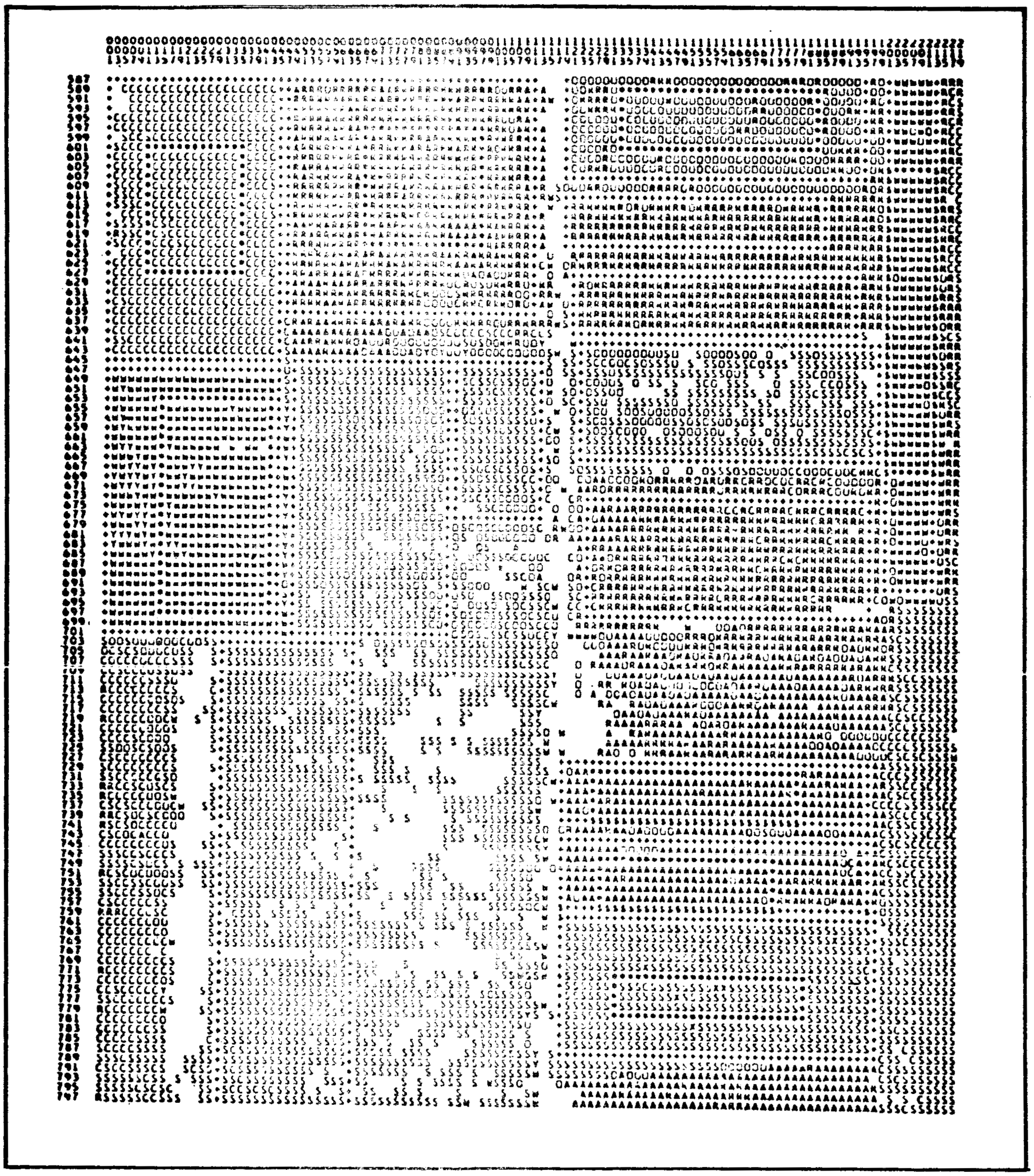


Figure 3-5. SUPERVISED CLASSIFICATION MAP BY THE PURDUE LARS'S MAXIMUM LIKELIHOOD TECHNIQUE (REF. 2, PAGE 31)

the contrary, the Yellowstone National Park test site is a natural uncultivated area with complex terrain structure as is evidenced by the video display of Channel 9 (Figure 3-6). This area is about 2 miles by 2 miles. The resolution element on the ground is about 20 feet by 20 feet. Figure 3-7 shows the ground truth map of the test site, consisting mainly of water, exposed bedrock, forest, kame, till, talus and cloud shadow over the forest. Detailed physical descriptions of these terrain types are given in reference 3. This target area is complex, having no distinctive boundary between different ground covers, and many parts were mixtures of two or three of the above terrain types.

Figure 3-8 presents the unsupervised classification map by the composite clustering techniques with three passes of the data set. Seventeen classes were specified in this classification. Evidently, several terrain types contains more than one subclass. For easier comparison with the ground truth map, the 17 classes established were merged down to 12 classes as shown in Figure 3-9. The criterion of similarity for merging is the Euclidean distance measure in the color space. A comparison of this map with the ground truth survey map was made and the classification accuracy for this unsupervised map was estimated to be about 80 percent (ref. 2). On the other hand, the supervised maximum likelihood technique obtained about 86 percent classification accuracy on the same test site (ref. 3). However, to obtain this higher accuracy, much human intervention and manipulation was needed; a) knowing before hand where to choose the typical training areas for every terrain type of interest, b) classifying the data and calculating the classification accuracy, and c) new training areas were selected based on results from (b). Steps (b) and (c) are repeated until no more improvement can be made. Contrasted to this iterative processing with close human supervision, the unsupervised clustering map was obtained with no human intervention after initially specifying the maximum allowable classifications.

Figure 3-10 shows the double-track boundary map corresponding to the classification map in Figure 3-9. The blank spaces indicate the homogeneous areas.

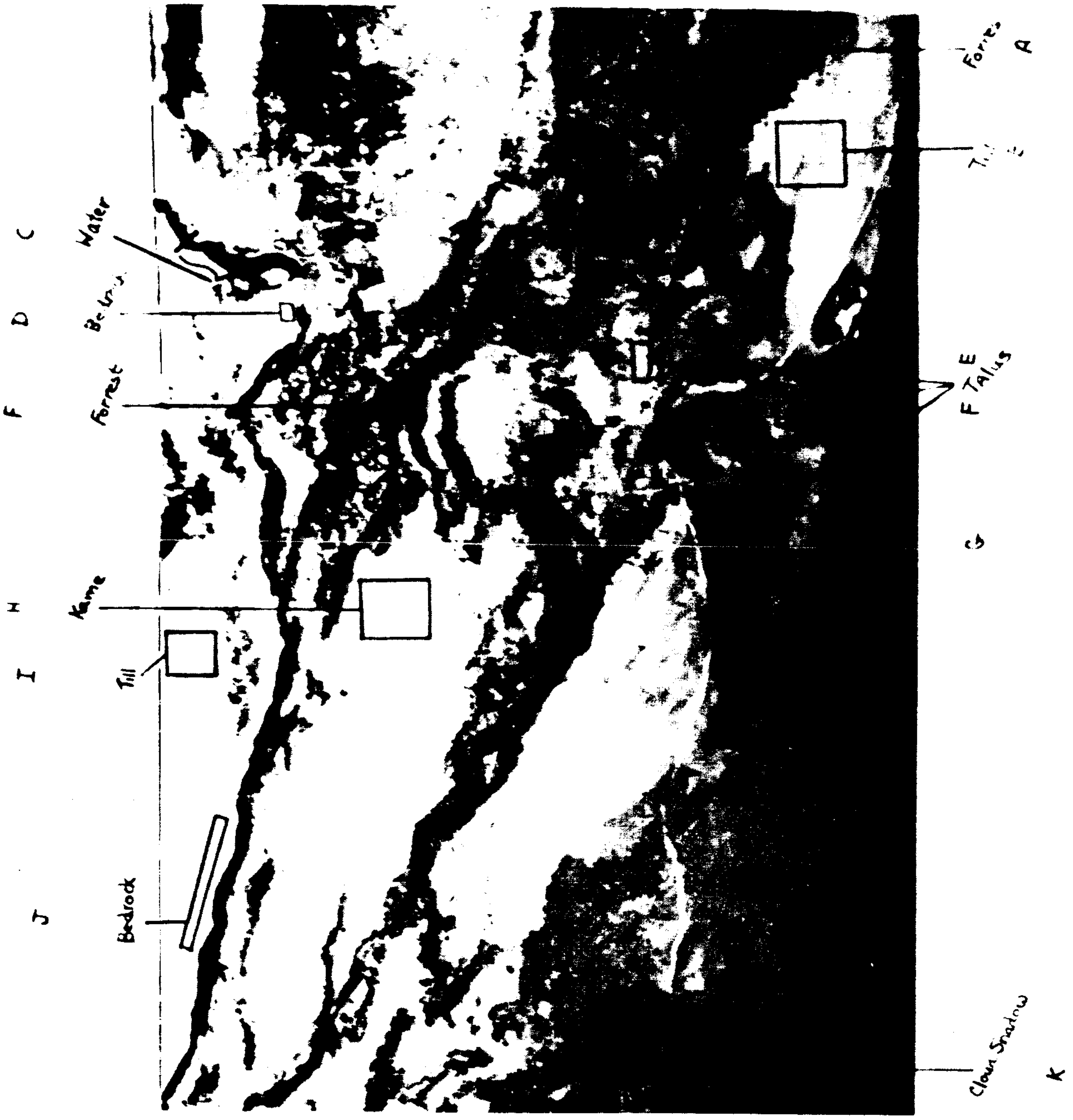


Figure 3-6. GRAY-SCALE VIDEO DISPLAY OF REFLECTANCE FOR CHANNEL 19, YELLOWSTONE NATIONAL PARK TEST SITE

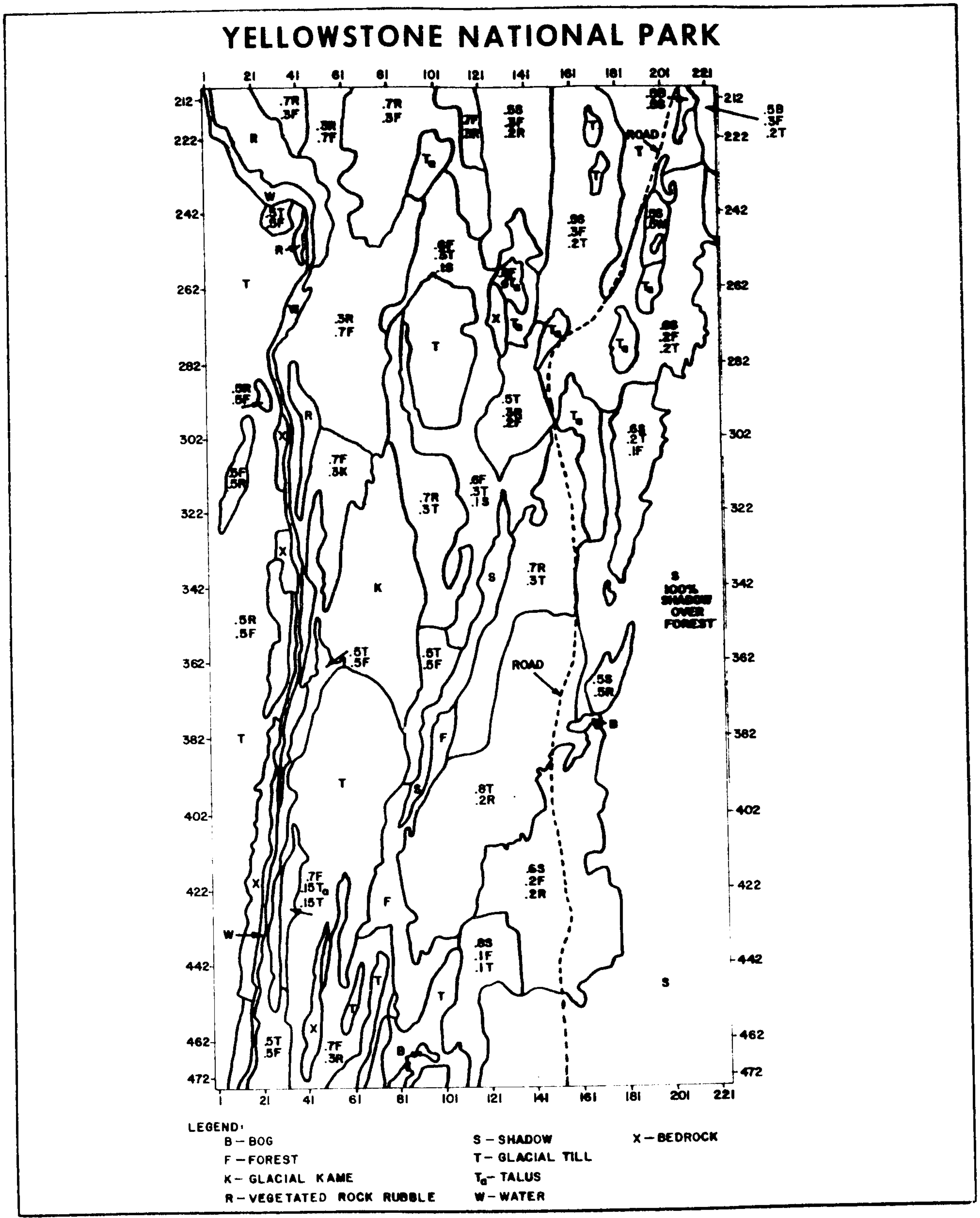


Figure 3-7. GROUND TRUTH MAP OF THE YELLOWSTONE NATIONAL PARK TEST SITE

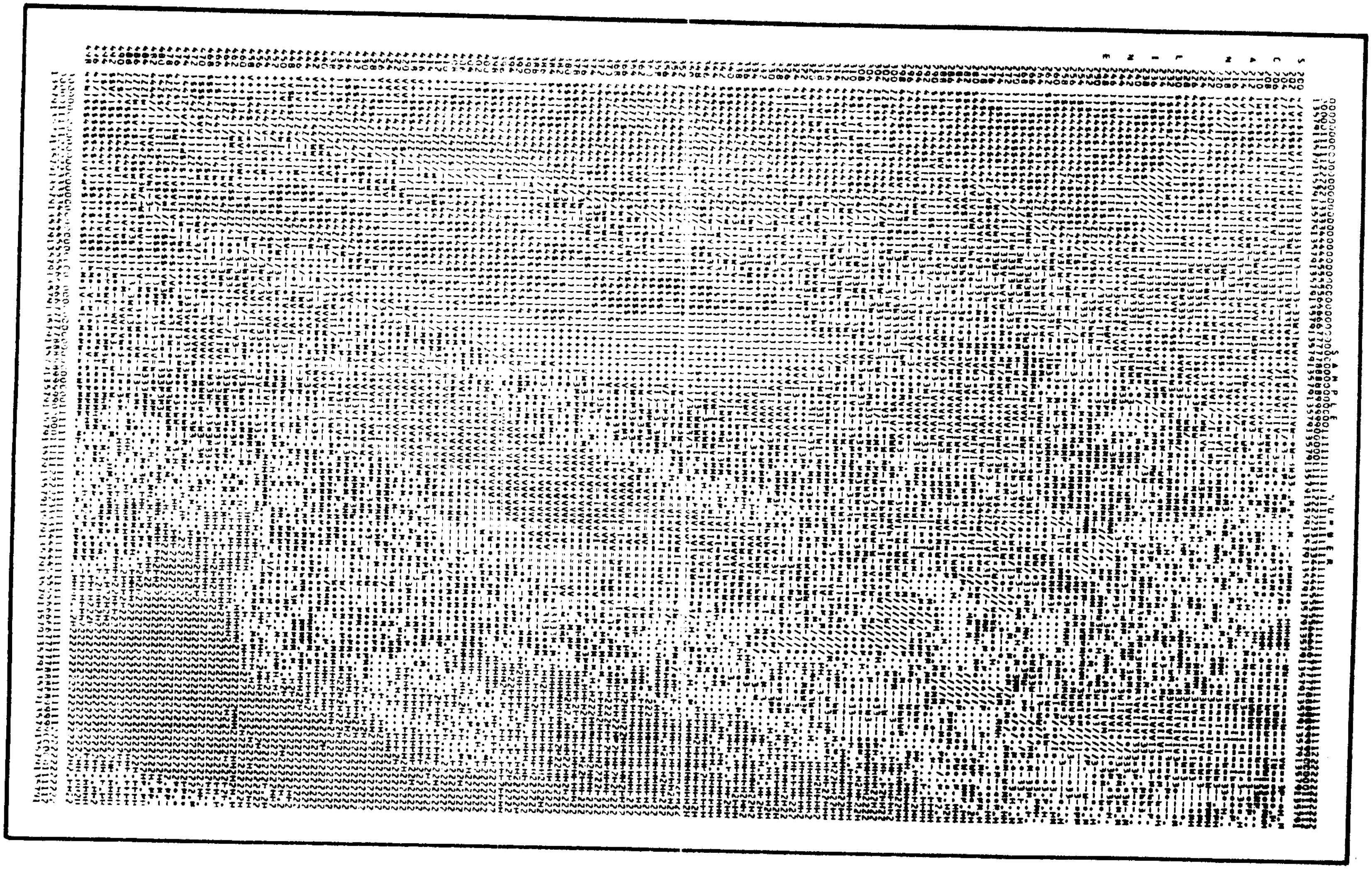


Figure 3-8. UNSUPERVISED CLASSIFICATION MAP OF YELLOWSTONE NATIONAL PARK BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 17 CLASSES AFTER 3 PASSES OF THE DATA SET. SYMBOL:) KAME; I, W, = VEGETATED RUBBLE; ., H, 2 SHADOW OVER FOREST; V, +, $, 4 TILL; -, 3, *, M FOREST; Z BEDROCK; / WATER OR TALUS

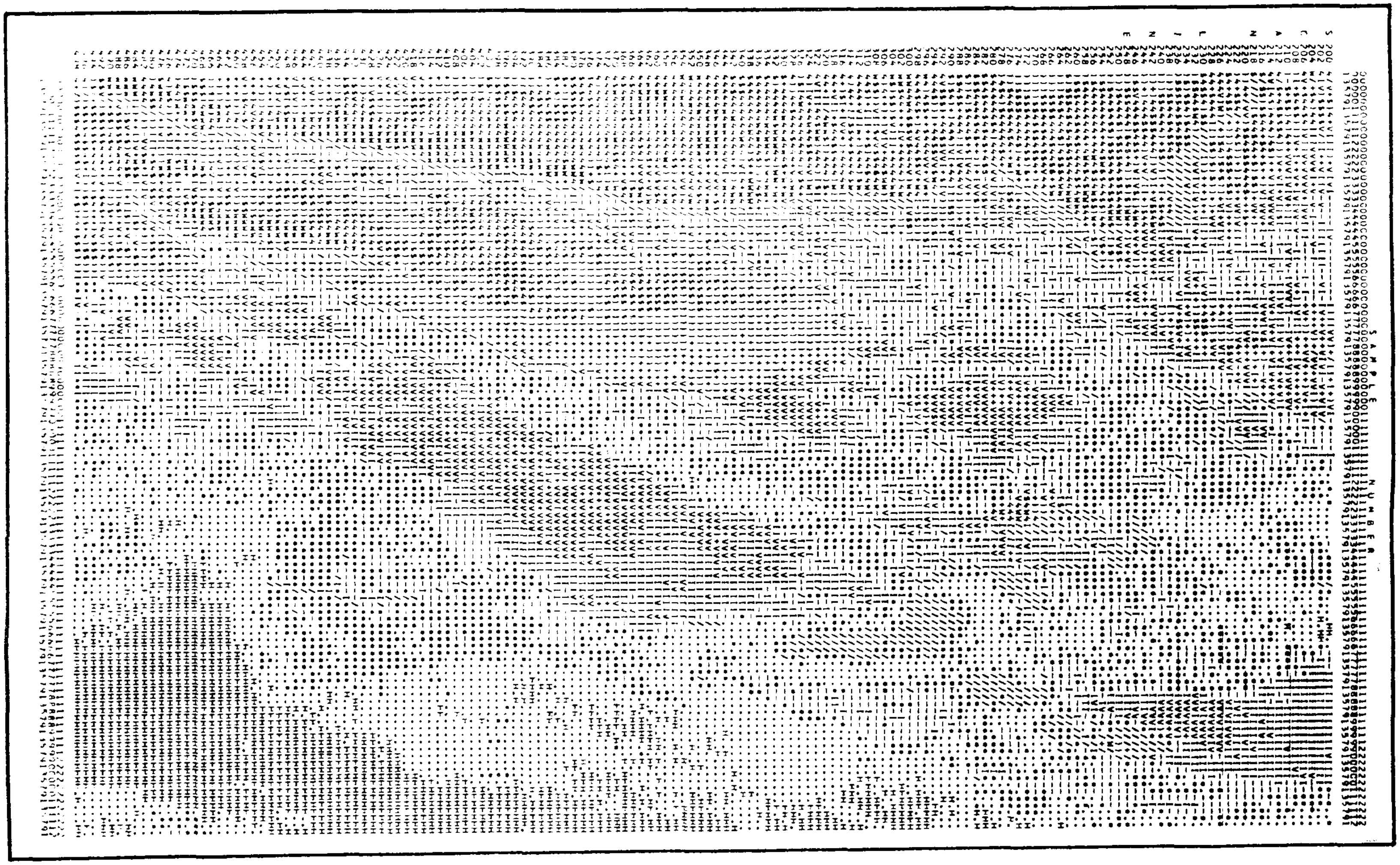


Figure 3-9. UNSUPERVISED CLASSIFICATION MAP OF YELLOWSTONE NATIONAL PARK BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 12 CLASSES. SYMBOL:) KAME; I VEGETATED RUBBLE; H,. SHADOW OVER FOREST; V, +, $, 4 GLACIAL TILL; -, * FOREST; W BEDROCK; AND / WATER OR TALUS

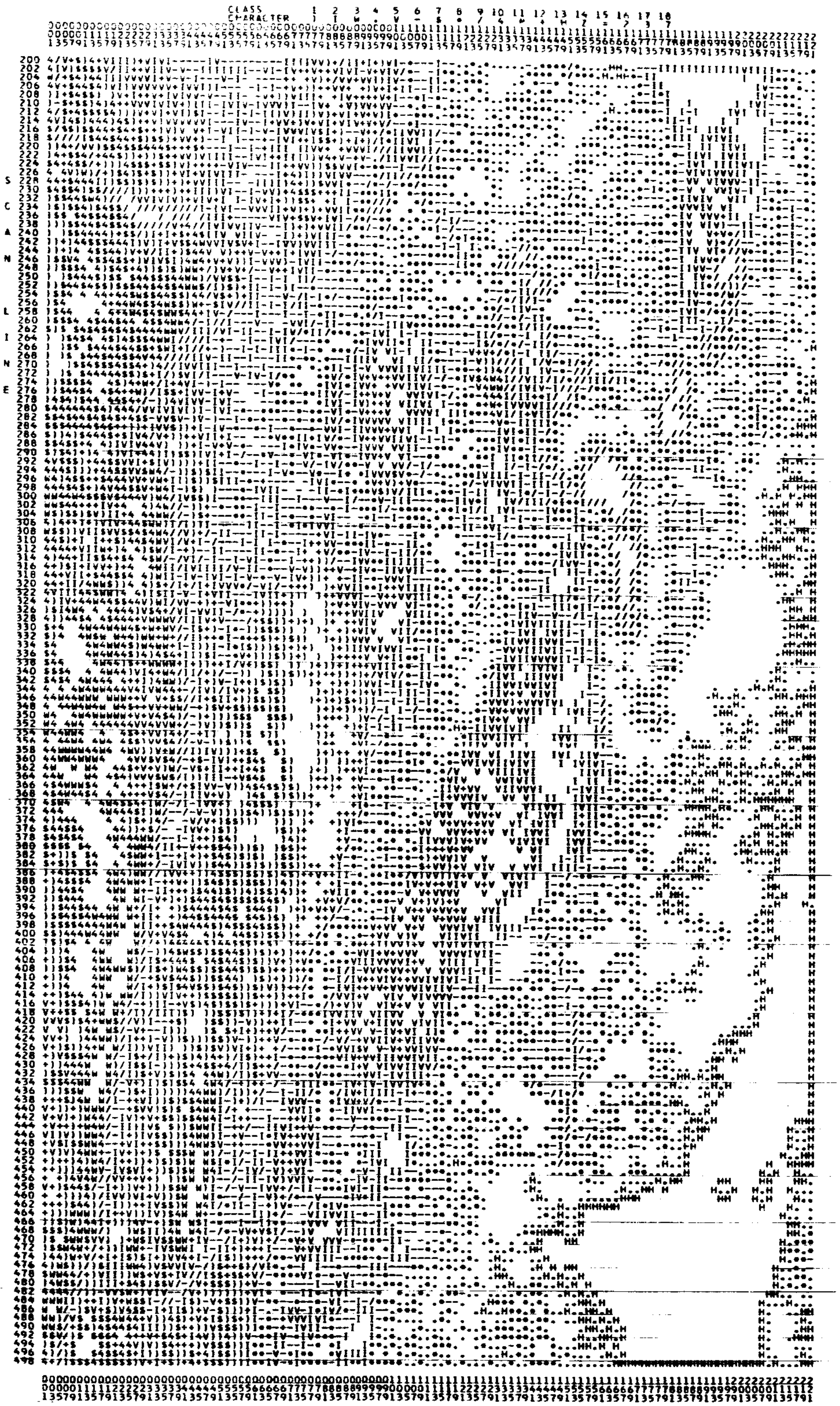


Figure 3-10. BOUNDARY MAP CORRESPONDING TO CLASSIFICATION MAP IN FIGURE 3-9

Section IV

COMPUTER PROGRAM DESCRIPTION

4.1 GENERAL PROGRAM STRUCTURE

The composite statistical sequential and K-means program contains two programs that are independent of each other. The statistical sequential segment is executed first and is never called again. It generates a predetermined number of mean spectral vectors for every population. These are passed along to the K-means segment which further improves these mean vectors, classifies the data sample and prints out a map of these populations along with statistics for each population. The general structure is depicted in Figure 4-1, which contains twelve links. Each of these are explained below.

LINKO (MAIN)

This link does all of the calling of subroutines in the statistical sequential segment and after this segment is finished, it calls a subprogram KMEANS. This subprogram controls the K-means clustering segment. The data preprocessing subroutine is also controlled by this link.

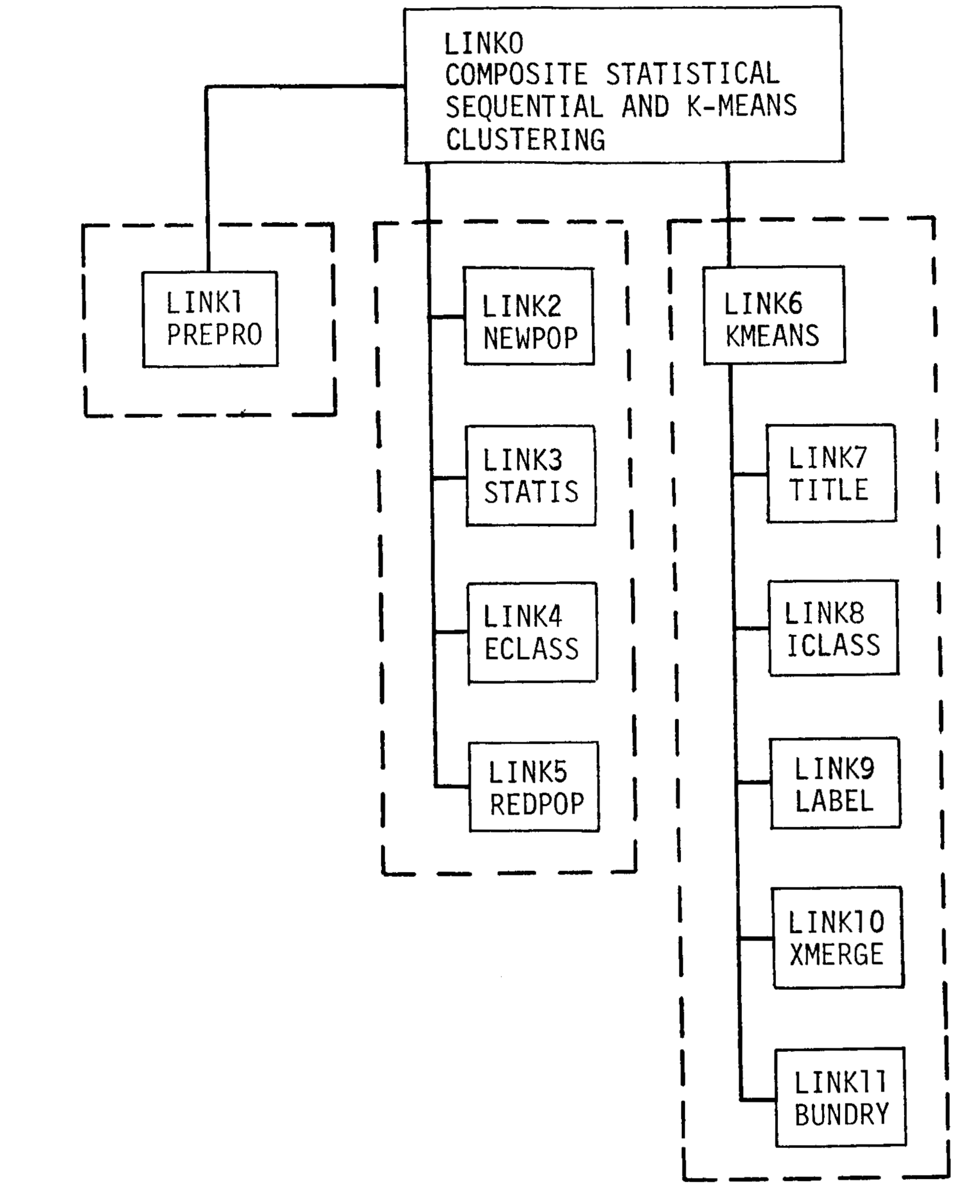
LINK1 (PREPRO)

This link performs the following options for data selection, for preprocessing prior to composite sequential clustering, or both:

- Selection of portion of the data tape to be processed,
- Skipping every other N scans in the data,
- Skipping every other N samples along a scan,
- Selection of spectral channels,
- Spectral normalization,
- Spectral equal weight transformation,
- Scan angle correction,
- Principal axis transformation.

LINK2 (NEWPOP)

This subroutine establishes the first population and every new population after the first one. M (NVARBS) samples are taken to see if they constitute a population. The mean of these M samples is computed, then the distance of



NOTE:

THESE THREE LARGE BLOCKS SURROUNDED BY THE DOTTED LINES ARE OVER-LAID.

Figure 4-1. GENERAL PROGRAM STRUCTURE

each sample from this mean is calculated. The maximum of these distances is found and the ratio of the maximum distance with the mean is computed. If this ratio is less than or equal to a preset threshold (THRESH), then the M samples are assigned to be a new population. If the ratio is greater than THRESH, the M samples do not qualify as a population and the sample which is the farthest distance from the mean is discarded. If LINK2 is establishing the first population, a new sample replaces the sample that was discarded and LINK2 again attempts to establish the first population. If LINK2 is trying to establish a population after the first population has been established, the program is transferred to LINK4 which tries to classify the next sample to existing populations. LINK4 adds additional samples to existing populations as long as they meet the criteria of similarity. When LINK4 runs across a sample it cannot add to an existing population, it adds this sample to the samples that were retained in the temporary position. Once LINK4 has collected M samples in the temporary position, control is returned to LINK2 which tries to establish a new population with these M samples.

LINK3 (STATIS)

This subroutine calculates statistics for newly established populations and updates the statistics of established populations as additional samples are added to them. These statistics are the populations' means, variance, number of samples, and some other parameters which are used in the CHI test.

LINK4 (ECLASS)

This subroutine is to classify new samples to established populations. When it is unable to do this, it stores the sample in the array (XDATA) until XDATA array contains M samples. At this time, LINK2 is called to establish a new population with these samples. Two tests, Chi-square test and N-test as described in Section II, are used for classification purposes. When a sample is added to an existing population, control is returned to LINK3 which updates the statistics of the population.

LINK5 (REDPOP)

This subroutine merges the two most similar populations once the maximum number (MAXPOP) of populations has been exceeded. This is accomplished by calculating the Euclidean distances between the means of all populations. The smallest of these Euclidean distances is found, and the two populations which have this distance between them are merged into one population. Thus the number of populations is reduced by one. Merging consists of adding the number of samples of the two merging populations together and assigning these samples and updating the mean vectors and correlation.

It is permissible to merge more than once. This is accomplished by setting NMRG greater than one. The subroutine then merges NMRG times after MAXPOP populations are established. The number of populations is reduced to (MAXPOP-NMRG). After this control is returned to the main program and MAXPOP populations must be accumulated before LINK5 is called again.

LINK6 (KMEANS)

This subprogram performs the functions of a main calling program for the K-means clustering. This subprogram is called by LINKO once the statistical sequential segment is finished.

LINK7 (ICLASS)

This subroutine is to perform four related functions: a) improve the cluster centers (mean vectors) input from the LINKO, b) classify data samples into finally improved populations based on the shortest distance criterion, c) generate a classification map, and d) calculate the associated statistics.

LINK8 (XMERGE)

This subroutine merges the two most similar populations after predetermined passes (NITER) have been made through ICLASS. Two populations are merged every time XMERGE is called and XMERGE is called NTOMRG times. The decision of which two populations to merge is based on the minimum Euclidean distance.

LINK9 (TITLE)

This subroutine is called by KMEANS before it calls ICLASS. Its function is to print heading information on the classification map.

LINK10 (LABEL)

This subroutine is called by ICLASS before and after ICLASS prints the classification map. LABEL prints out the sample number which goes with each column of the population map. The two calls to LABEL by ICLASS cause this labeling information to be printed at the top and bottom of the population map.

LINK11 (BUNDRY)

This subroutine is used to generate a boundary map based on the final classification map obtained by the composite sequential clustering of the input multispectral scanner data. Each homogeneous area is bounded by the symbol which represents its classification, and the area within the boundary is blank.

A more detailed flowchart of the composite statistical sequential clustering program is shown in Figure 4-2.

4.2 DEFINITION OF NAMELIST VARIABLES

4.2.1 NAMELIST Variables in LINKO

CHIHI(201)

A table of values that are upper bounds for the χ^2-test.*

CHILO(201)

A table of values that are lower bounds for the χ^2-test.*

No need to be specified by the user. The present program uses 99 percent confidence limits.

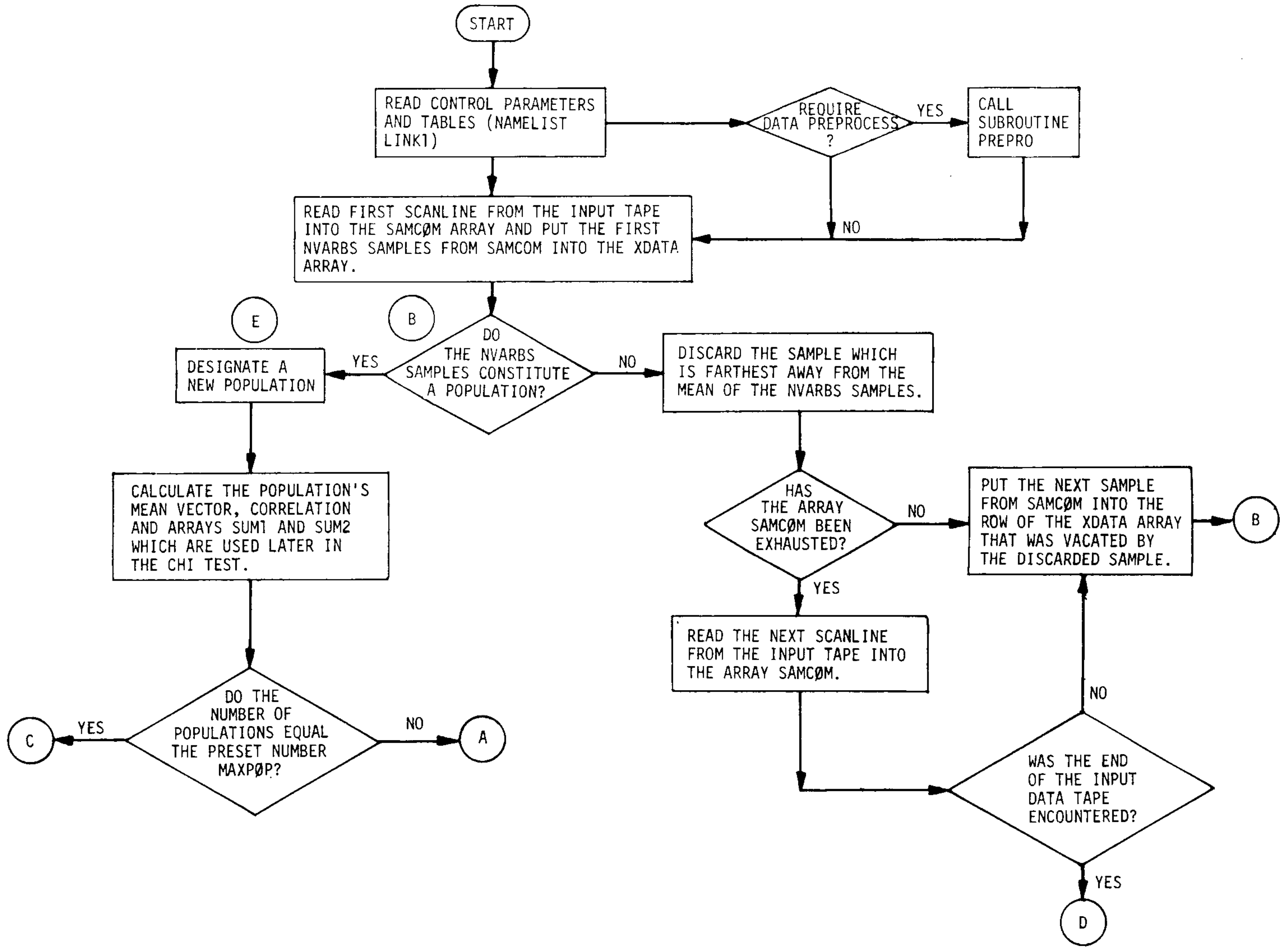

START
READ CONTROL PARAMETERS AND TABLES (NAMELIST LINK1)
REQUIRE DATA PREPROCESS ?
YES
CALL SUBROUTINE PREPRO
NO
READ FIRST SCANLINE FROM THE INPUT TAPE INTO THE SAMCØM ARRAY AND PUT THE FIRST NVARBS SAMPLES FROM SAMCOM INTO THE XDATA ARRAY.
E
B
DO THE NVARBS SAMPLES CONSTITUTE A POPULATION?
YES
NO
DESIGNATE A NEW POPULATION
DISCARD THE SAMPLE WHICH IS FARTHEST AWAY FROM THE MEAN OF THE NVARBS SAMPLES.
CALCULATE THE POPULATION'S MEAN VECTOR, CORRELATION AND ARRAYS SUM1 AND SUM2 WHICH ARE USED LATER IN THE CHI TEST.
HAS THE ARRAY SAMCØM BEEN EXHAUSTED?
NO
PUT THE NEXT SAMPLE FROM SAMCØM INTO THE ROW OF THE XDATA ARRAY THAT WAS VACATED BY THE DISCARDED SAMPLE.
B
YES
READ THE NEXT SCANLINE FROM THE INPUT TAPE INTO THE ARRAY SAMCØM.
NO
DO THE NUMBER OF POPULATIONS EQUAL THE PRESET NUMBER MAXPØP?
YES
C
NO
A
WAS THE END OF THE INPUT DATA TAPE ENCOUNTERED?
YES
D

Figure 4-2. FLOWCHART OF COMPOSITE STATISTICAL SEQUENTIAL CLUSTERING PROGRAM

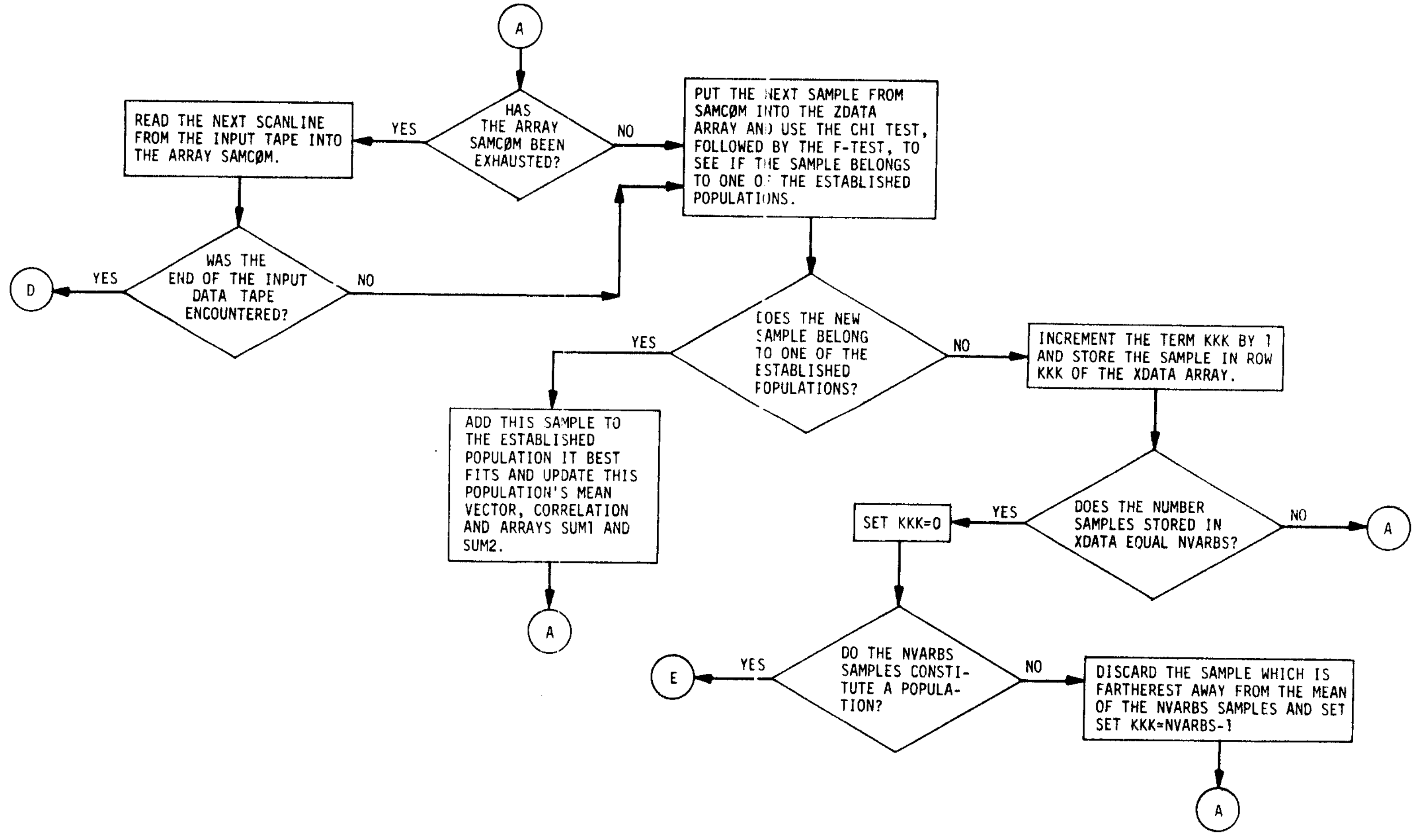

A
HAS THE ARRAY SAMCØM BEEN EXHAUSTED?
YES
NO
READ THE NEXT SCANLINE FROM THE INPUT TAPE INTO THE ARRAY SAMCØM.
PUT THE NEXT SAMPLE FROM SAMCØM INTO THE ZDATA ARRAY AND USE THE CHI TEST, FOLLOWED BY THE F-TEST, TO SEE IF THE SAMPLE BELONGS TO ONE OF THE ESTABLISHED POPULATIONS.
WAS THE END OF THE INPUT DATA TAPE ENCOUNTERED?
YES
D
NO
DOES THE NEW SAMPLE BELONG TO ONE OF THE ESTABLISHED POPULATIONS?
YES
NO
INCREMENT THE TERM KKK BY 1 AND STORE THE SAMPLE IN ROW KKK OF THE XDATA ARRAY.
ADD THIS SAMPLE TO THE ESTABLISHED POPULATION IT BEST FITS AND UPDATE THIS POPULATION'S MEAN VECTOR, CORRELATION AND ARRAYS SUM1 AND SUM2.
A
DOES THE NUMBER SAMPLES STORED IN XDATA EQUAL NVARBS?
YES
SET KKK=0
NO
A
DO THE NVARBS SAMPLES CONSTITUTE A POPULATION?
YES
E
NO
DISCARD THE SAMPLE WHICH IS FARTHEREST AWAY FROM THE MEAN OF THE NVARBS SAMPLES AND SET SET KKK=NVARBS-1
A

Figure 4-2. FLOWCHART OF THE COMPOSITE STATISTICAL SEQUENTIAL CLUSTERING PROGRAM (Continued)

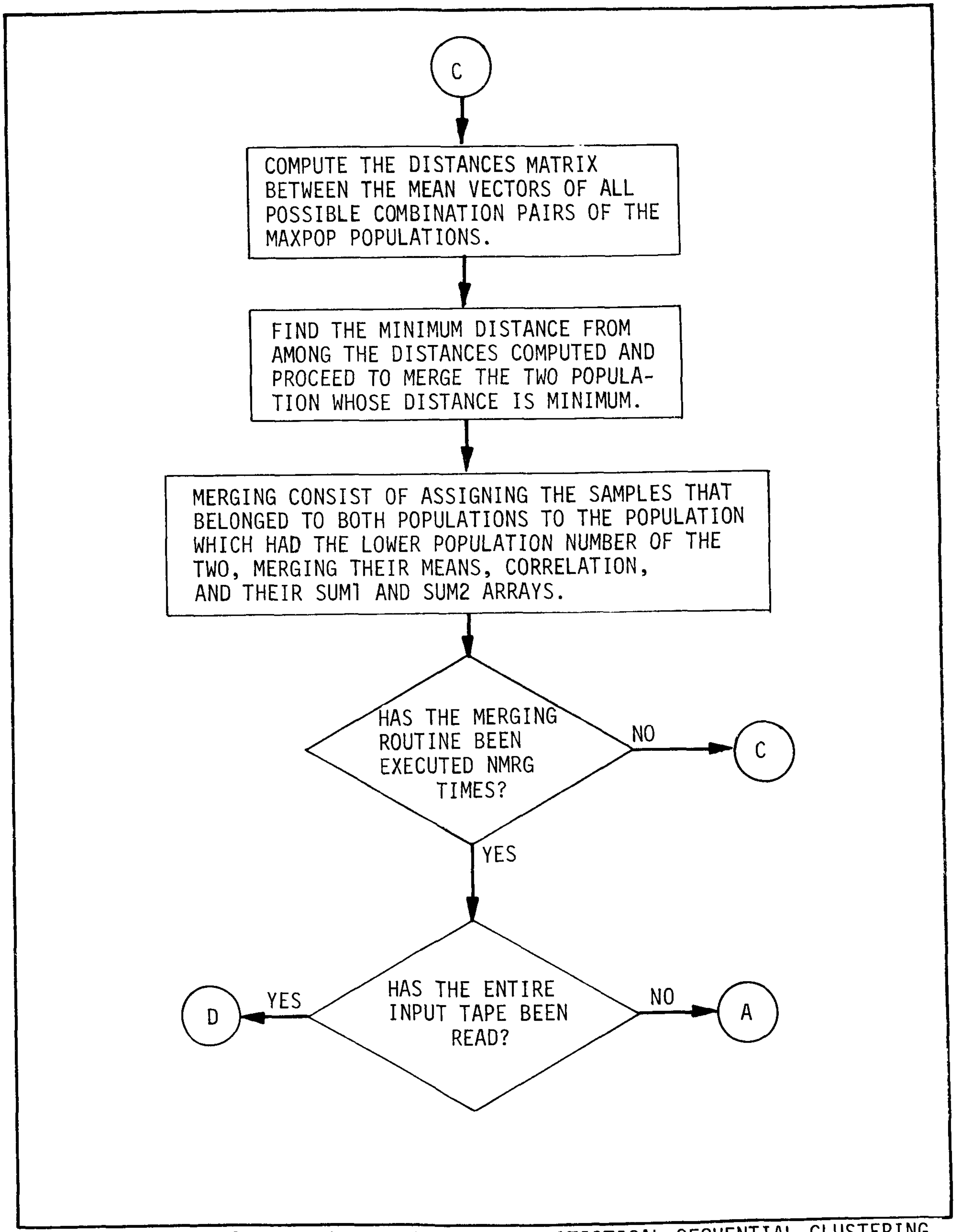

C
COMPUTE THE DISTANCES MATRIX BETWEEN THE MEAN VECTORS OF ALL POSSIBLE COMBINATION PAIRS OF THE MAXPOP POPULATIONS.
FIND THE MINIMUM DISTANCE FROM AMONG THE DISTANCES COMPUTED AND PROCEED TO MERGE THE TWO POPULATION WHOSE DISTANCE IS MINIMUM.
MERGING CONSIST OF ASSIGNING THE SAMPLES THAT BELONGED TO BOTH POPULATIONS TO THE POPULATION WHICH HAD THE LOWER POPULATION NUMBER OF THE TWO, MERGING THEIR MEANS, CORRELATION, AND THEIR SUM1 AND SUM2 ARRAYS.
HAS THE MERGING ROUTINE BEEN EXECUTED NMRG TIMES?
NO
C
YES
HAS THE ENTIRE INPUT TAPE BEEN READ?
YES
D
NO
A

Figure 4-2. FLOWCHART OF THE COMPOSITE STATISTICAL SEQUENTIAL CLUSTERING PROGRAM (Continued)

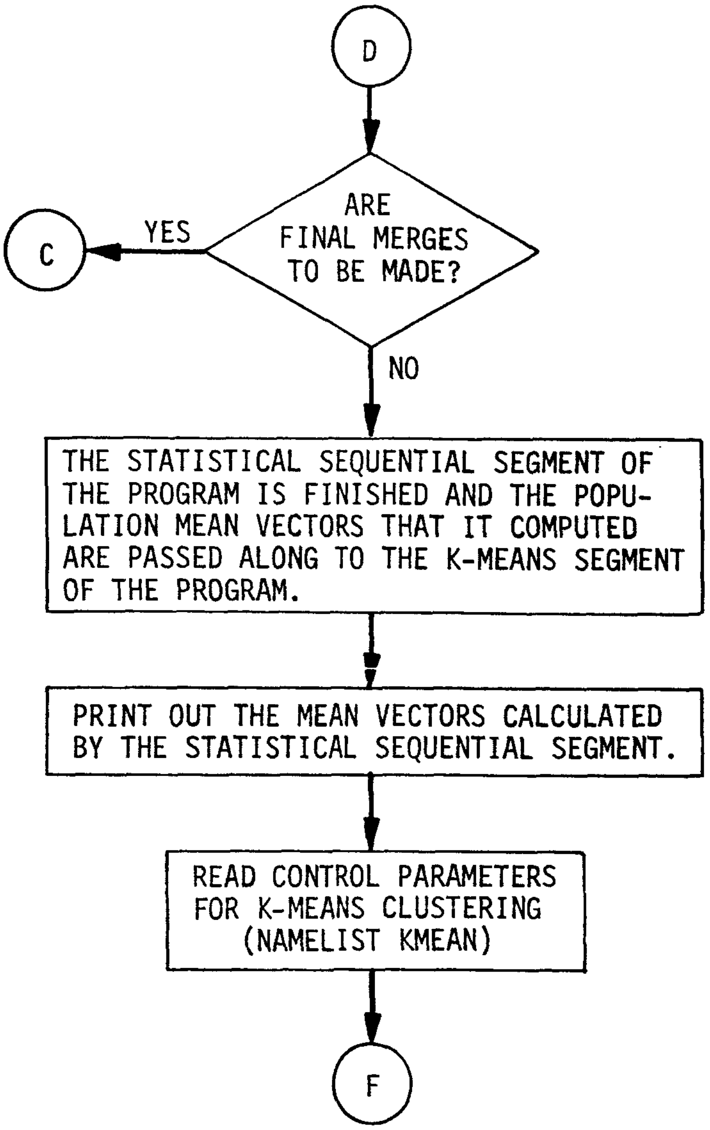

D
ARE FINAL MERGES TO BE MADE?
YES
C
NO
THE STATISTICAL SEQUENTIAL SEGMENT OF THE PROGRAM IS FINISHED AND THE POPULATION MEAN VECTORS THAT IT COMPUTED ARE PASSED ALONG TO THE K-MEANS SEGMENT OF THE PROGRAM.
PRINT OUT THE MEAN VECTORS CALCULATED BY THE STATISTICAL SEQUENTIAL SEGMENT.
READ CONTROL PARAMETERS FOR K-MEANS CLUSTERING (NAMELIST KMEAN)
F

Figure 4-2. FLOWCHART OF THE COMPOSITE STATISTICAL SEQUENTIAL CLUSTERING PROGRAM (Continued)

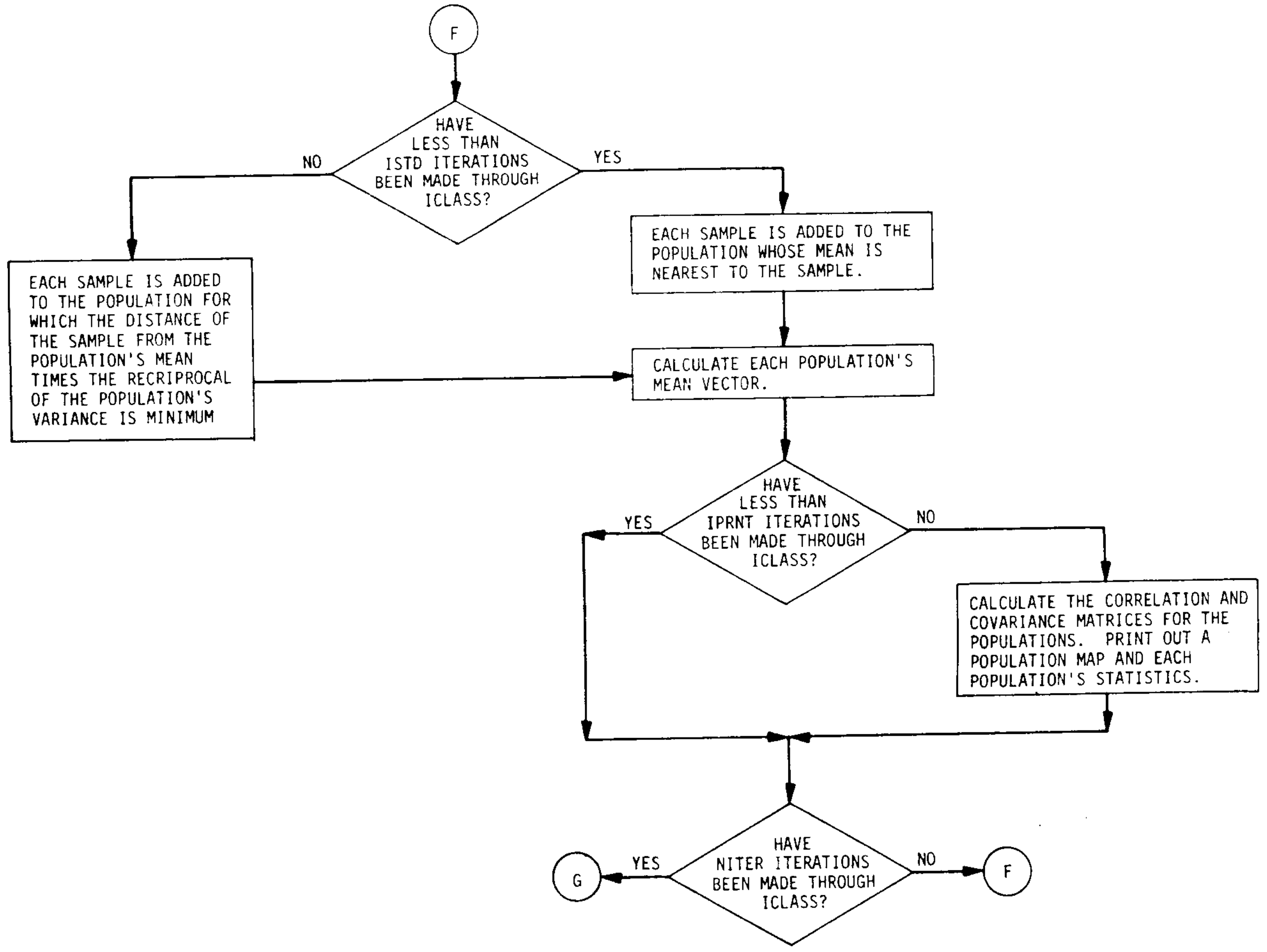

F
HAVE LESS THAN ISTD ITERATIONS BEEN MADE THROUGH ICLASS?
NO
YES
EACH SAMPLE IS ADDED TO THE POPULATION FOR WHICH THE DISTANCE OF THE SAMPLE FROM THE POPULATION'S MEAN TIMES THE RECRIPROCAL OF THE POPULATION'S VARIANCE IS MINIMUM
EACH SAMPLE IS ADDED TO THE POPULATION WHOSE MEAN IS NEAREST TO THE SAMPLE.
CALCULATE EACH POPULATION'S MEAN VECTOR.
HAVE LESS THAN IPRNT ITERATIONS BEEN MADE THROUGH ICLASS?
YES
NO
CALCULATE THE CORRELATION AND COVARIANCE MATRICES FOR THE POPULATIONS. PRINT OUT A POPULATION MAP AND EACH POPULATION'S STATISTICS.
HAVE NITER ITERATIONS BEEN MADE THROUGH ICLASS?
YES
G
NO
F

Figure 4-2. FLOWCHART OF THE COMPOSITE STATISTICAL SEQUENTIAL CLUSTERING PROGRAM (Continued)

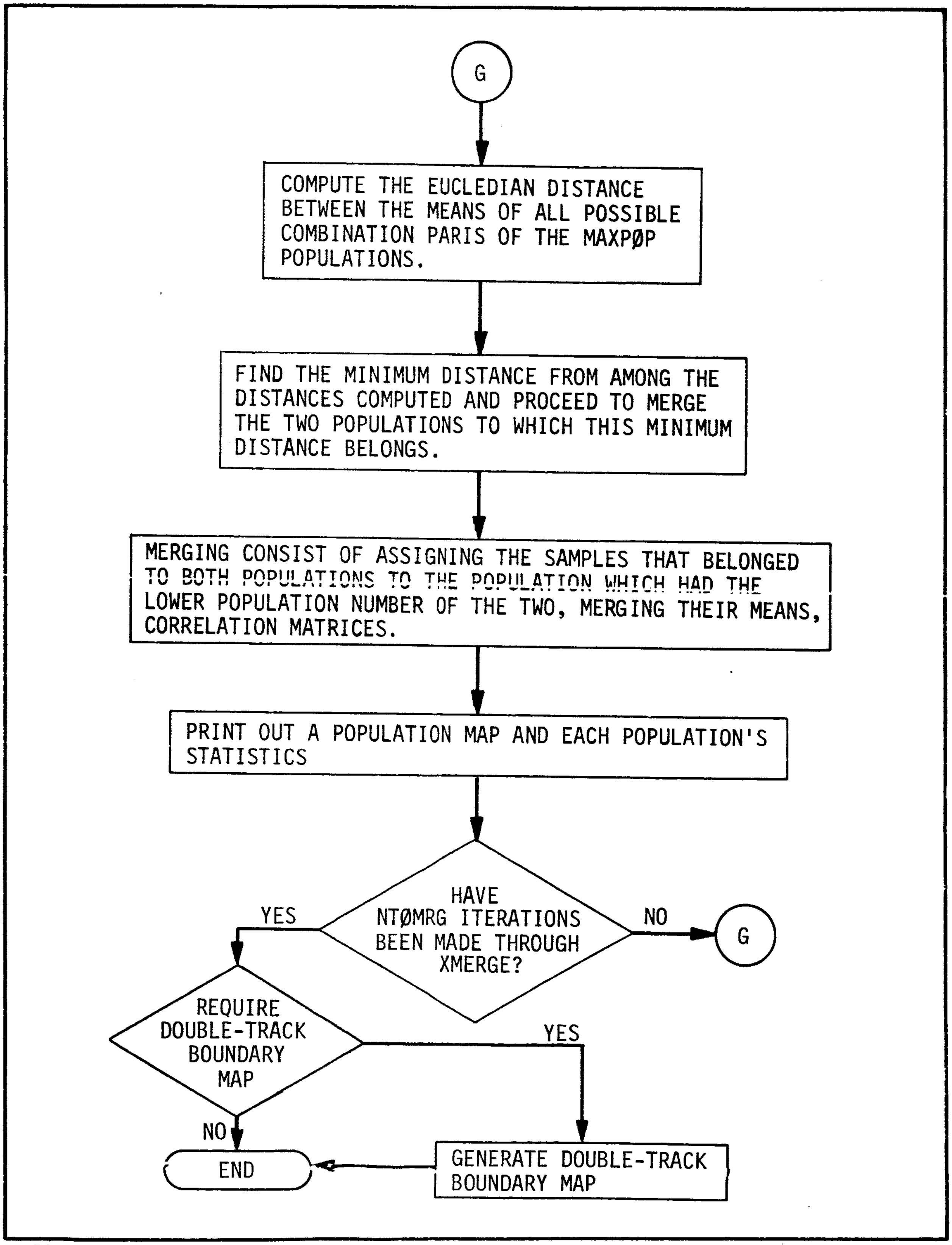

G
COMPUTE THE EUCLEDIAN DISTANCE BETWEEN THE MEANS OF ALL POSSIBLE COMBINATION PARIS OF THE MAXPØP POPULATIONS.
FIND THE MINIMUM DISTANCE FROM AMONG THE DISTANCES COMPUTED AND PROCEED TO MERGE THE TWO POPULATIONS TO WHICH THIS MINIMUM DISTANCE BELONGS.
MERGING CONSIST OF ASSIGNING THE SAMPLES THAT BELONGED TO BOTH POPULATIONS TO THE POPULATION WHICH HAD THE LOWER POPULATION NUMBER OF THE TWO, MERGING THEIR MEANS, CORRELATION MATRICES.
PRINT OUT A POPULATION MAP AND EACH POPULATION'S STATISTICS
HAVE NTØMRG ITERATIONS BEEN MADE THROUGH XMERGE?
YES
NO
G
REQUIRE DOUBLE-TRACK BOUNDARY MAP
YES
NO
END
GENERATE DOUBLE-TRACK BOUNDARY MAP

Figure 4-2. FLOWCHART OF THE COMPOSITE STATISTICAL SEQUENTIAL CLUSTERING PROGRAM (Concluded)

MAXPOP

The maximum desired number of populations plus one, ≤ 30.

NSAMP

Number of samples in a scanline, must be ≤ 880.

NCHAN

Number of channels used (≤ 12).

NSCAN1

The first scan number to be processed.

NSCANE

The last scan number to be processed.

NSCN1

The first scan number on the input data tape (i.e., ITAPE).

NSAMPS

The first sample number along the scan to be processed.

NSAMPE

The last sample number along the scan to be processed.

CHAN(I)

Set the I^{th} element = 1.0 if that channel is used; otherwise, set to 0 ($I \leq 12$).

THRESH

Threshold value used in LINK2 for establishing a new population ($0.01 \leq \text{THRESH} \leq 0.1$).

NDOUBL

Set to 1 for requesting the double-track boundary map. Set to 0 for bypassing.

4.2.2 NAMELIST Variables in LINK1 (PREP)

NCHSL

Set to 1 for performing the channel selection. Set to 0 for bypassing.

NDEBUG

Set to 1 for printing out the input and output data tape for the purpose of program debugging. Set to 0 for suppressing the printouts.

KSKIPM

Every KSKIPM number of samples to be skipped along a scan. When KSKIPM = 0, all samples are processed.

KSKIPC

Every KSKIPC number of scans to be skipped along the flight direction. When KSKIPC = 0, all scans are processed.

NORM

Set to 1 for performing spectral normalization. Set to 0 for bypassing.

NSAC

Set to 1 for performing scan angle correction. Set to 0 for bypassing.

NESWT

Set to 1 for performing equal spectral weight transformation. Set to 0 for bypassing.

NPRIAX

For performing the principal axis transformation set NPRIAX = No. of eigen vectors desired. Set to 0 for bypassing.

NMEAN

Set to 1 for computing the mean vector and covariance matrix over the portion of data to be processed. Set to 0 for bypassing. When NPRIAX = 1, NMEAN is automatically set to 1.

MODE

Set to 1 for FORTRAN input tape and set to 2 for non-FORTRAN input tape.

ITYPE

Set to 1 for floating point input tape and 0 for fixed point input tape.

NBTLG

Number of bits in the input data words.

IHEAD

Set to 1 if the input tape has a header record, otherwise set to 0.

IDWORD

Number of words in the header record ($\leq$ 15).

4.2.3 **NAMELIST Variables in LINK6 (KMEN)**

IALPHA(30)

This array contains alphanumeric characters which are used for the classification map. Each population is assigned a distinct character.

ISTD

Number of iterations that use mean vectors only for K-means iterations.

NITER

Total number of iterations required.

NTOMRG

Number of further population merges required after completion of classification by the GKMC over the entire data set.

WAVE(24)

Array which contains 12 pairs of lower and upper wavelengths in sequence of the channels.

XIDENT(12)

Array which stores alphanumeric information which is to be printed out on the classification map (Example: Purdue/C-1/Flight/Line).

4.3 INPUT DATA TAPE FORMAT

The digital input tape containing the multispectral data should be in the following format to be compatible to the computer program.

One data file consists of two parts: identification record and data record. The identification record contains 15 words or less. It is permissible without this ID record on the input tape. The multispectral data should be stored in the data record scanline-by-scanline. The components (or channels) in each sample on a scanline should be stored in ascending order. The data record can be either in FORTRAN or non-FORTRAN format. It can be either fixed point or floating point.

4.4 COMPUTER PROGRAM DECK SETUP

A typical deck setup for the IBM 7094 is given. The user is referred to Marshall Space Flight Center Manual IBM 7094/7040 Direct Couple Operating System for more detailed description.

The first card (Job Card)

Columns	
1-4	$JOB
16-19	NASA
21-34	Users Name and bin number
35	,
36-41	six digit job number
42	,
43-44	00
45	,
46-47	11 for production, 12 for testing
48	,
49	14MCE

The next card is for the input data tape. The five digit number after the T is the input tape number.

Column 1	Column 8	Column 16
$SETUP	UT6	T12345, Disc,,,1

The next card is $ASSIGN card. A $ASSIGN card is required for each $SETUP card.

Column 1	Column 16
$ASSIGN	SYSUT6

The next two cards are $EXECUTE and $IBJOB.

Column 1	Column 16
$EXECUTIVE	IBJOB
$IBJOB	FIOCS, MAP

The next seven cards are $FILE cards.

Column 1	Column 16
$FILE	'UNIT01', NONE
$FILE	'UNIT07', NONE
$FILE	'UNIT08', NONE
$FILE	'UNIT10', NONE
$FILE	'UNIT12', NONE
$FILE	'UNIT14', NONE
$FILE	'UNIT15', NONE

The next card is a $IBFTC card. A $IBFTC card is required in front of the main program and each subroutine.

Column 1	Column 8
$IBFTC	MAIN

Main program deck (LINKO)

The next card is a $ORIGIN card. It is employed because the program is overlaid. The main program is loaded into core along with the subroutines after the first $ORIGIN card. After the main finishes with these subprograms, the subprograms after the second $ORIGIN card are loaded into the core space that was used by the first set of subroutines. The main program stays in core at all times.

Column 1	Column 16
$ORIGIN	A

Column 1	Column 8
$IBFTC	ONE Subroutine PREPRO
$IBFTC	ONEA Subroutine GET
$IBFTC	ONEB Subroutine EIGEN

Column 1	Column 16
$ORIGIN	A

Column 1	Column 8
$IBFTC	TWO Subroutine NEWPOP
$IBFTC	THREE Subroutine STATIS
$IBFTC	FOUR Subroutine ECLASS
$IBFTC	FIVE Subroutine REDPOP

Column 1	Column 16
$ORIGIN	A

Column 1	Column 8
$IBFTC	SIX Subroutine KMEANS
IBFTC	SEVEN Subroutine TITLE
$IBFTC	EIGHT Subroutine ICLASS
$IBFTC	EIGHTA Subroutine CMAP
$IBFTC	NINE Subroutine LABEL

Column 1	Column 8
$IBFTC	TEN Subroutine XMERGE
$IBFTC	ELEVEN Subroutine BUNDRY

The next cards are data cards that supply the information for the following NAMELIST names: LINKO, PREP, and KMEN. The values that are assigned to these data cards will depend upon what the user wants the program to do. A $DATA card appears before this group of cards.

Column 1

```
$DATA
$LINKO
MAXPOP=15
THRESHOLD=0.05
NSAMP=222
NCHAN=12
CHAN=1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 0.0, 1.0
NSCAN1=10
NSCANE=30
NSAMPS=1
NSAMPE=222
NSCAN1=1
CHILO=Table of 201 constant values \
                                    }    need not be specified by the users.
CHIHI=Table of 201 constant values /
$END
```

Column 1

```
$PREP
KSKIPM=1
KSKIPC=1
NORM=0
NSAC=1
NESWT=0
```

Column 1

```
NPRIAX=0
NMEAN=0
NCHSL=0
MODE=1
ITYPE=1
NBTLG=36
IHEAD=0
IDWORD=15
NDEBUG=1
$END
```

Column 1

```
$KMEN
NITER=2
WAVE=0.4, 0.42, 0.43, 0.44, 0.46, 0.50, etc.
XIDENT=6H  ,    ,6HPURDUE, 6H     , 6HFLIGHT, 6H
         6H  ,6H        6H    , 6H     ,6H     ,6H


ISTD=1
IALPHA=6H1, 6H2, 6H3, etc.
       6HA, 6HB, 6HC, etc.
NTOMRG=1
$END
```

The next card is an end of file card which has a punch for 7 and 8 in column 1, followed by three $EOF cards.

Column 1

```
$EOF
```

This completes the deck setup.

```
Column 1     Column 16
$ORIGIN      A
```

```
Column 1     Column 8     Column 16
$IBFTC       Six
             Subroutine KMEANS
$IBFTC       Seven
             Subroutine TITLE
$IBFTC       Eight
             Subroutine ICLASS
$IBFTC       Nine
             Subroutine LABEL
$IBFTC       Ten
             Subroutine XMERGE
$IBFTC       Eleven
             Subroutine BUNDRY
```

The next cards are data cards that supply the information for the following NAMELIST names: LINKO, PREP, MAINA, and KMEN. The values that are assigned to these data cards will depend upon what the user wants the program to do. A $DATA card appears before this group of cards.

```
Column 1
$DATA
$LINKO
MAXPOP=15
NVARBS=6
THRESH=0.05
ICASE=1
NMRG=1
ITAPE=11
JTAPE=3
KTAPE=4
MTAPE=13
LTAPE=
CHILO=Table of 201 constant values }
CHIHI=Table of 201 constant values }  need not be specified by the users.
```

Column 1

```
NSAMP=222
NCHAN=12
NFNMRG=1
CHAN=1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 0.0, 1.0
NSCAN1=10
NSCANE=30
NSAMPS=1
NSAMPE=222
NSCAN1=1
$END
```

Column 1

```
$PREP
KSKIPM=1
KSKIPC=1
NORM=0
NSAC=1
NESWT=0
NPRIAX=0
NMEAN=0
NCHSL=0
NPORT=1
NDEBUG=1
$END
```

Column 1

```
$KMEN
IOUT=6
ITER=0
MAPOPT=0
NITER=2
XIDENT=Purdue Flight Line C-1
WAVE=0.4, 0.42, 0.43, 0.44, 0.46, 0.50, etc.
```

```
Column 1
IAISKP=1
IAJSKP=1
IALPHA=1, 2, 3, ..., 8, 9, A, B, C, D, E
IDOXMG=0
IHEAD=0
IMPROV=1
IMRD=1
IPRNT=1
ISKIP=1
ISTD=1
MBYPAS=0
NWORD=15
NTOMRG=1
$END

          Col.  17        27        37        45     47     49
One Card        Purdue    Flight    Line      C      -      1
```

The next card is an end of file card which has a punch for 7 and 8 in column 1, followed by three $EOF cards.

```
Column 1
$EOF
```

This completes the deck setup.

Section V

CONCLUSIONS

A new composite statistical sequential K-means technique has been developed for unsupervised classification of multispectral data from earth resources observation. This technique has been applied to two sets of well-known multispectral data, namely, Purdue's Flight Line C-1 and the Yellowstone National Park test site. The classification accuracy for both cases is about 80 percent, while the supervised maximum likelihood technique obtained about 80 percent and 86 percent, respectively, on the corresponding test site. In view of very little human intervention required in the application of the unsupervised technique, these good results are encouraging and are considered sufficient to warrant concentrated research for general application to a wide variety of multispectral data. From the operational viewpoint, the unsupervised technique is more feasible than the supervised technique for many applications including onboard near real time data processing.

We shall apply the unsupervised classification program for processing multispectral scanner data from ERTS-1 for land use survey of the State of Alabama. A study for improving both the mathematical algorithms and computer implementation is also under way.

Section VI

REFERENCES

1. Fu, K. S., Landgrebe, D. A., and Phillips, T. L., "Information Processing of Remotely Sensed Agricultural Data", Proc. IEEE, Vol. 57, No. 4, April 1969.

2. "Remote Multispectral Sensing in Agricultural", Purdue University Laboratory for Remote Sensing Application, Research Bulletin 873, December 1970.

3. Smedes, H. W., Pierce, K. L., Tanguary, M. G., and Hoffer, R. M., "Digital Computer Terrain Mapping from Multispectral Data, and Evaluation of Proposed Earth Resources Technology Satellite (ERTS) Data Channels, Yellowstone National Park: Preliminary Report", AIAA Paper No. 70-309, March 1970.

4. Nalepka, R. F., "Investigation of Multispectral Discrimination Techniques", Willow Run Laboratories, University of Michigan, January 1970.

5. Haralick, R. M. and Kelley, G. L., "Pattern Recognition with Measurement Space and Spatial Clustering for Multiple Images", Proc. IEEE, Vol. 57, No. 4, April 1969.

6. Nagy, G., Shelton, G., and Talaba, J., "Procedural Questions in Signature Analysis", Proc. 7th International Sym. on Remote Sensing of Environment, May 17-21, 1971.

7. Turner, B. J., "Cluster Analysis of Multispectral Scanner Remote Sensor Data", Proc. Remote Sensing of Earth Resources, Vol. 1 edited by F. Shahroki, March 1972.

8. Su, M. Y., "Unsupervised Classification of Remote Multispectral Sensing Data", Northrop Technical Report TR-220-1075, April 1972.

9. Su, M. Y., Jayroe, R. R., and Cummings, R. E., "Unsupervised Classification of Earth Resources Data", Proc. of "Earth Resources Observation and Information Analysis System Conference", University of Tennessee Space Institute, March 13-14, 1972.

10. Su, M. Y. and Krause, F. R., "Automatic Processing of Multispectral Observations", AIAA Paper No. 71-234, AIAA Integrated Information System Conference, February 17-19, 1971.

11. Gasey, R. C. and Nagy, G., "An Autonomous Reading Machine", IEEE Trans. Computers, Vol. C-17, No. 5, May 1968.

12. MacQueen, J., "Some Methods for Classification and Analysis of Multispectral Observations", Proc. Fifth Berkeley Symposium Math. Statistics and Probability, Vol. 1, pp. 281-297, 1967.

13. Smedes, H. W., Linnerud, H. J., Woalaver, L. B., Su, M. Y., and Jayroe, R. R., "Mapping of Terrain by Computer Clustering Techniques Using Multispectral Scanner Data and Using Color Aerial Film", 4th NASA Earth Resource Program Review, February 1972.