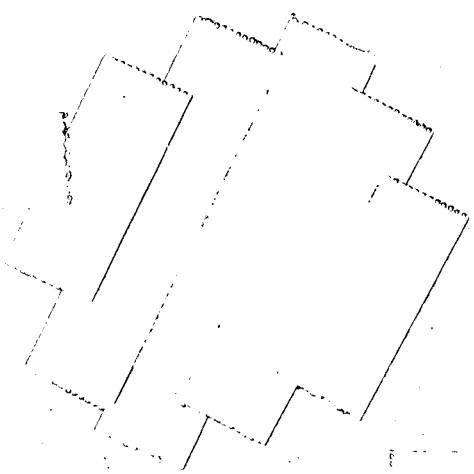
TR-220-1075
APRIL 1972

UNSUPERVISED CLASSIFICATION OF REMOTE MULTISPECTRAL SENSING DATA

Prepared for:

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
GEORGE C. MARSHALL SPACE FLIGHT CENTER
Aero-Astrodynamics Laboratory

UNDER CONTRACT NAS8-27364



NORTHROP SERVICES, INC.
P. O. BOX 1484
HUNTSVILLE, ALABAMA 35807
TELEPHONE (205)837-0580

TR-220-1075

UNSUPERVISED CLASSIFICATION OF REMOTE MULTISPECTRAL SENSING DATA

15 April 1972

by

M. Y. Su

PREPARED FOR:

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
GEORGE C. MARSHALL SPACE FLIGHT CENTER
AERO-ASTRODYNAMICS LABORATORY

Under Contract NAS8-27364

REVIEWED AND APPROVED BY:

A. L. Grady

A. L. Grady, Manager
Advanced Engineering Analysis

NORTHROP SERVICES, INC.
HUNTSVILLE, ALABAMA

FOREWORD

This study was undertaken by Northrop Services, Inc., for NASA/MSFC, Huntsville, Alabama, under Contract No. NAS8-27364. This will constitute the interim report for the period of 12 months ending on April 15, 1972. The program was under the direction of MSFC Aero-Astrodynamics Laboratory, Flight Data Statistics Office, with Mr. R. R. Jayroe, Jr. as the project monitor.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. F. R. Krause, Mr. R. E. Cummings and Mr. R. R. Jayroe, Jr. for their helpful discussions in completing this work. The author wishes also to acknowledge Dr. H. W. Smedes, U. S. Geological Survey, Denver, Colorado, for supplying the remote multispectral data over the Yellowstone National Park, through the Flight Data Statistics Office, and for his comparison of the unsupervised classification maps with the associated ground truth map and to the Purdue University, Laboratory for Remote Sensing Applications, for Supplying the Purdue C-1 Flight Line data. Assistance with data processing from Jerry Durret, NASA Co-op student, was also appreciated.

SUMMARY

This report presents a new automatic processing technique for unsupervised classifications (or clustering) for multispectral remote sensing data. This technique has been implemented into a digital computer program. Applications of the computer program for actual multispectral scanner data from the aircraft survey will also be presented.

Up to the present, main approaches are based on supervised maximum likelihood classification techniques which require reference spectral target signatures from training areas on the ground. One of the most serious drawbacks of the supervised classification techniques is associated with the high variability of the spectral signatures.

The unsupervised classification technique avoids the above drawback by not requiring the reference signatures. Essentially, the technique will group the data sets into a number of classes based on the intrinsic similarity within each class. The physical identification of each class is done by checking a small area belonging to each class after the data processing. In this respect, the application of unsupervised techniques is in the reverse order of the supervised technique. The advantage of processing the data in the former order is that the investigator shall know better where to select the ground truth. Another advantage is for on-board data compression to minimize the rates of data transmission from future spacecrafts to the ground receiving stations. The third advantage is that automatic change analysis of earth resources study can be more logically carried out by the unsupervised technique.

The new unsupervised classification technique for classifying multispectral remote sensing data which can be either from the multispectral scanner or digitized color-separation aerial photographs 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 have been accomplished. Comparisons between the classification maps by the unsupervised technique and the supervised maximum liklihood technique indicated that the classification accuracy is comparable to each other.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Continued)

LIST OF ILLUSTRATIONS (Concluded)

LIST OF TABLES

Section I
INTRODUCTION

Applications of nonsupervised clustering techniques have recently attracted more attention for processing and analyzing multispectral data obtained by remote sensing of the earth's resources and environment (refs. 1-7). In the past, main approaches in dealing with these types of data were based on supervised classification techniques which required reference spectral target signatures from training areas (ref. 8).

One of the most serious drawbacks of the supervised classification techniques is associated with the high variability of the reference spectral signatures. These signatures depend not only on different physical targets of interest, but also on the following factors (some known and some unknown in the process of data gathering):

- Background materials
- Atmospheric and meteorological conditions
- Different physical location and orientation
- Time of day, different reason of data collection
- Sensor scan angle and sun elevation and azimuth
- Different stages of plant growth
- Different land use practices.

With so many variable factors affecting the remote sensing data, it is very difficult, if not impractical, to set up an operational system for establishing the reference spectral signature (or ground truth) library. So far, the users of the supervised classification methods mainly obtain the reference spectral signatures directly from training sets which form parts of the test areas. Even with this practice, it still requires much human judgment and intervention to select proper training areas for obtaining sufficient accuracy of classification.

The nonsupervised classification, or clustering, techniques avoid most of the above difficulties and operational impracticability. The clustering

technique does not require the reference spectral signatures. In essence the technique will group the sample data into a number of classes, all of which are statistically homogeneous. Finally, the physical identification of each class is accomplished by collecting the ground truth from a suitable size area belonging to that particular class. In this sense, the application of clustering techniques to the multispectral data analysis is in the reverse order of the supervised classification technique. The advantage of processing the data in the order according to the clustering techniques is that it will be known better where to select the reference ground truth.

Another advantage of the clustering technique is for data flow compression in the telemetry of data from the spacecraft to the ground data receiving station. It is quite clear now that the data rate collected by the satellite-borne sensors will be so large that present telemetry systems can not handle it. However, Dr. A. Park indicated that if the data can be compressed to 1/50 or greater of the present volume, then the presently available commercial TV receiving station can be used for space data collection. It is quite feasible that the clustering techniques can process onboard the raw data and compress it into the acceptable reduced volume for this purpose. It is also possible in the hydrological applications to augment a relatively few number of ground sensors by the remote sensing data with the clustering techniques.

Under the present contract, a new composite sequential K-means clustering algorithm has been developed with actual applications to two sets of remote sensing data; the Purdue agricultural field (Purdue C-1 Flightline) and the Yellowstone National Park test site. The latter test site is actually in natural wilderness with various terrain types, forest cover and parts of it under cloud shadow. According to Dr. A. Park, NASA Headquarters, Earth Resources Program, this set of remote sensing data is about the most complex data collected under the NASA Earth Resource Program. Thus, it may offer the most critical test to date of the capability of the unsupervised clustering technique. If the technique can obtain an acceptably accurate classification map, then it may be safe to apply to other remote multispectral sensing data for earth resources and environment survey.

The principles behind the composite clustering technique will be presented in Section II. Detailed mathematical algorithms, computer programs and users manual will not be given in this report, but will be included in the final contract report. Applications of the technique to the aforementioned two sets of data together with some supporting processing by other computer programs developed under previous contracts (ref. 9) will be given in Sections III and IV, respectively. A comparison of the unsupervised classification maps with Purdue LARS' results (refs. 8 and 10) will be made. Finally, some future developments and concluding remaks will be made in Section V.

Section II

THE COMPOSITE SEQUENTIAL K-MEANS CLUSTERING TECHNIQUE

The composite clustering technique developed essentially consists of two independent clustering techniques. The first is called the statistical sequential classification technique (SSC) (refs. 11 and 12) and the second is called generalized K-means techniques (GKM) (ref. 13). Therefore, each technique will first be described, and then how they can be merged into one will be described.

2.1 STATISTICAL SEQUENTIAL CLUSTERING

The sensor collects multispectral data from a target which forms an image. An image can be composed of m scan lines of n resolution elements per scan line. Each resolution element yields a K-dimensional observation vector $x(\lambda_i)$, i = 1,2,... K, where λ_i indicates the i^{th} spectral band.

The purpose of the SSC program is to classify the given sequences of multisectional 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 results and statistics.

A flowchart of the main steps of the algorithm is depicted in Figure 2-1. Step 1 - all control parameters and statistical tables are read in. Step 2 - M (M = 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 null-class, which contains all the samples unidentifiable, and then read in a new sample 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 new population. The

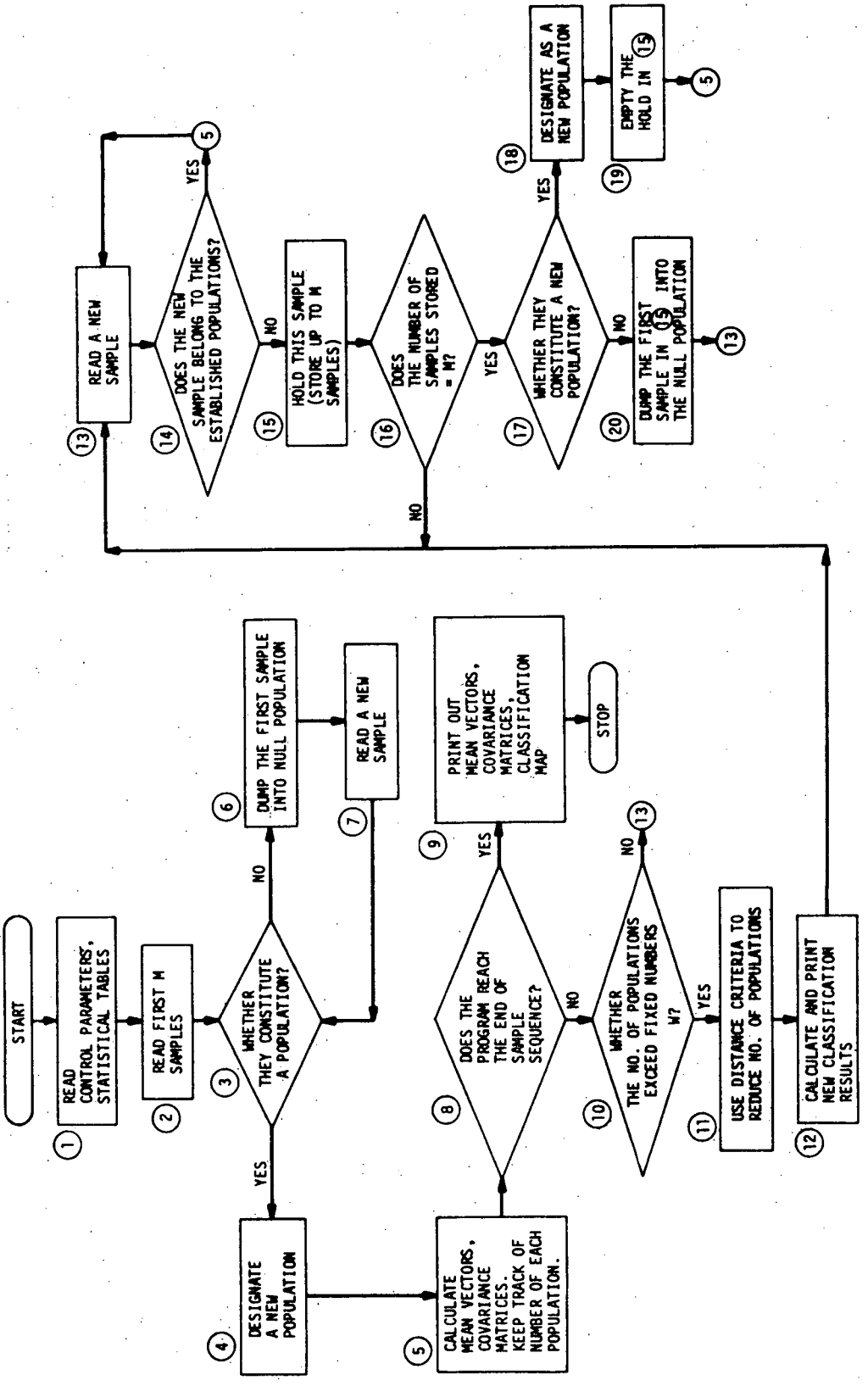


Figure 2-1. FLOWCHART OF STATISTICAL SEQUENTIAL CLUSTERING

above process will be repeated until the first population is established. The statistical parameters of interest for this population are calculated in Step 5.

Next, one proceeds to check whether the end of the entire sample sequence is reached. If it is, the program will print out the final results of the number of samples in that population; the corresponding sample mean vector, covariance matrix, and classification map. The latter map represents a 2-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 exceed the prescribed number. If the answer is yes, the program will proceed to Step 11 in order to reduce the number of established populations back to the prescribed number. This is accomplished by combining two populations that are most similar to each other into a enlarged population encompassing all those samples belonging to the two original populations. Subsequently, the program will also recalculate the corresponding sample mean vector and covariance matrix 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 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 vector and covariance matrix 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 was done for establishment of 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 dumped into the class of unidentifiable class, then proceed to read in a new sample. This process is repeated until all the sample sequences have been processed. The final outputs of the whole algorithm is to print out the number of samples, the mean vector and covariance matrix

for each population, a divergence matrix among all the populations, and finally a 2-dimensional map of spatial locations of samples for each population.

2.2 GENERALIZED K-MEANS CLUSTERING

The generalized K-means algorithm essentially consists of three steps plus an additional step for displaying the 2-dimensional map of clustering results. This algorithm is an improved version of the existing K-means algorithms (refs. 14 through 17).

2.2.1 Step 1 – Estimation of Initial Cluster Centers

Let the sample sequence be denoted by $\{x_i(\lambda_j)$, i = 1, 2, ..., M and j = 1, 2, 3, ..., N$\}$. Here i denotes the sample number and j denotes its components. The first initial cluster center C_1 will be the first sample, i.e.,

$$C_1(\lambda_j) = x_1(\lambda_j) \tag{2-1}$$

The second initial cluster center C_2 will be the sample which has the fartherest distance from C_1, i.e.,

$$C_2(\lambda_i) = x_i(\lambda_j) \text{ with the maximum of}$$

$$\sum_{j=1}^{N} [x_i(\lambda_j) - x_1(\lambda_j)]^2 \text{ over all i.} \tag{2-2}$$

The $(k+1)^{th}$ initial cluster center C_k (for k > 2) will be the sample which has the maximum of the minimum distances among all with respect to the established k initial cluster centers, i.e.,

$$C_{k+1}(\lambda_j) = x_i(\lambda_j) \text{ with}$$

$$\max_i \left\{ \min_k \left[\sum_{j=1}^{N} [x_i(\lambda_j) - x_k(\lambda_j)]^2 \right] \right\} \tag{2-3}$$

The results of the above procedure is to plant evenly the initial cluster centers whose number will be prescribed over that part of the measurement space occupied densely by the given input sample sequence. The step of estimating initial centers is relatively time consuming. The computer time requirement will be proportional to

$$MN\,\frac{(K-1)(K-2)}{2} = \frac{1}{2}\,MNK^2\left(1 - \frac{3}{K} + \frac{2}{K^2}\right) \qquad (2\text{-}4)$$

where K is the total number of cluster centers. Clearly, the computer time required is linearly proportional to the total sample M and number of components per sample N, respectively, but almost to the square of the required number of cluster centers, K. Usually, N is fixed, but in general one would expect, without any prior knowledge, that K would increase with M.

2.2.2 Step 2 – Preliminary Improvement of Cluster Centers

This step of improving accuracy of the initial cluster centers is exactly the same as used in the present K-means algorithm. "Preliminary" is used here because another improvement to the cluster centers will be made after this step as discussed in Step 3. The minimum distance criterion is employed. The entire sample sequence is classified into K groups by calculating the distances of each sample with respect to each cluster center and classifying the sample into that particular center that yields the minimum distance, i.e.,

$$x_i(\lambda_j) \rightarrow C_k(\lambda_j) \text{ if}$$

$$\sum_{j=1}^{N} [x_i(\lambda_j) - C_k(\lambda_j)]^2 \text{ is the minimum over all } k. \qquad (2\text{-}5)$$

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 mean measurement vector (or center-of-gravity). These updated K centers will now be regarded as the 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. In general, one would expect that smaller threshold values result in better (or more accurate) results, but it also requires a larger number of iterations. Some compromise is thus obviously called for and no general rule can be specified.

2.2.3 Step 3 – Final Improvement of Cluster Centers

The reason for requiring some further improvement to the cluster centers as obtained from Step 2 can be illustrated by using Figure 2-2. In this figure there are three natural clusterings in 2-component scattering diagrams. Further, these three clusters are clearly linearly separable, thus it is desirable to separate the samples into three clusters. Using Step 2, 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 samples actually belonging to cluster No. 1 are mis-classified into clusters No. 2 and 3. This resulted from the fact that inter-distance between clusters No. 1 and 2 (similarly for cluster No. 1 and 3) is about equal to the sum of the two intra-distances of the individual clusters of which one is much larger than the other.

This hypothetical example is actually a very common case in the multispectral observations of earth resources and environments. Investigators of spectral signatures have pointed out this difficulty many times.

One way to deal with this difficulty and thus improve the power of the present K-means algorithm will be proposed. To the first approximation, the intra-distance of samples within one cluster will be the sample standard deviation vector that is the square roots of the diagonal elements of the sample covariance matrix. 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 completely correct if each component has the same standard deviation. With this basic understanding, the minimum-distance criterion used in the present K-means algorithm can be replaced by a more general similarity criterion with the standard deviations as weights to better locate the decision hyperplanes.

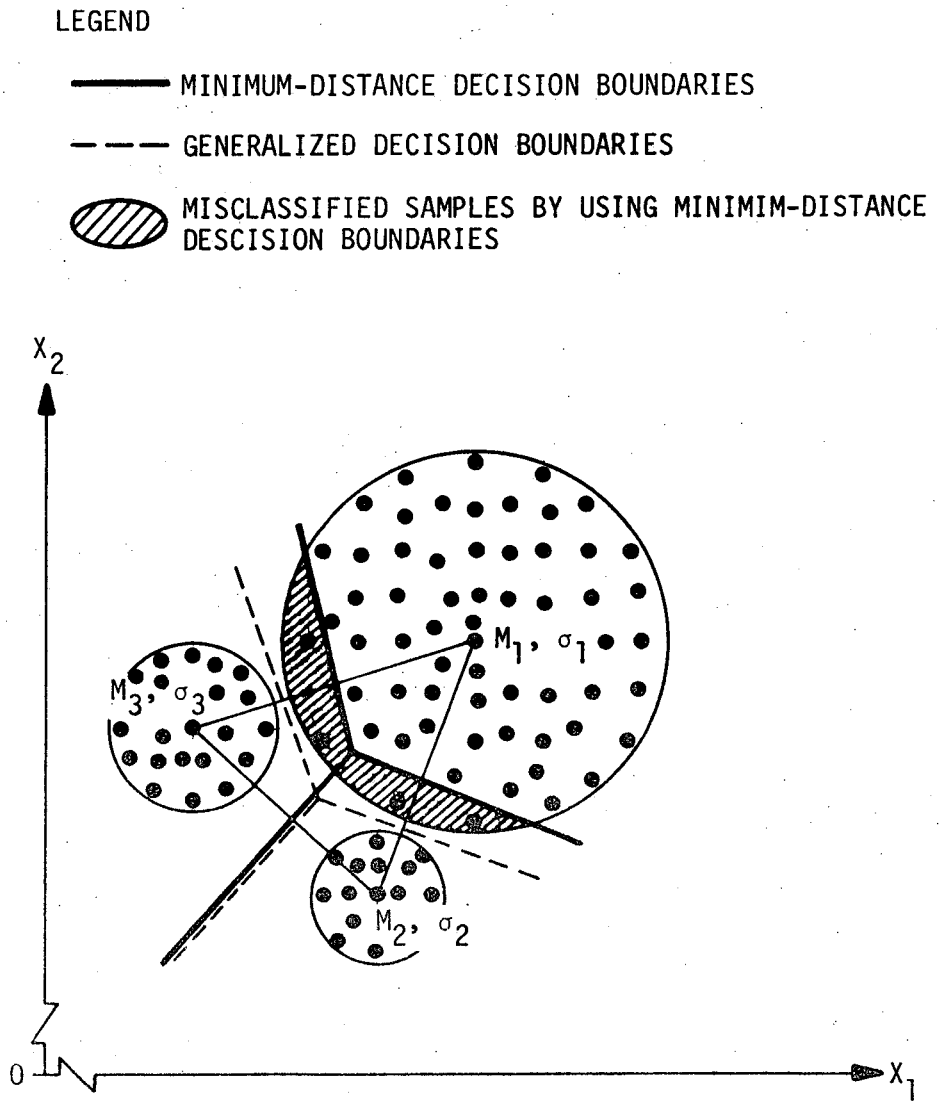


Figure 2-2. COMPARISON OF THE PRESENT AND GENERALIZED K-MEANS ALGORITHMS

The proposed similarity criterion can be expressed as

$$x_i(\lambda_j) \rightarrow C_k(\lambda_j) \text{ if}$$

$$\frac{1}{S_k^2} \sum_{j=1}^{N} [x_i(\lambda_j) - C_k(\lambda_j)]^2 \text{ is minimum over all k.} \qquad (2\text{-}7)$$

Here, S_k is the characterized sample standard deviation for k^{th} cluster center. The rest of the step will be the same as in Step 2.

Three important points germaine to the added step will now be discussed. First, one might ask why not use Step 3 with the more general similarity measure 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 their evaluations are more apt to the influence of misclassified samples than the centers-of-gravity of clusters. Hence, there is no clear indication to expect better performance from Step 3 than Step 2 at the first several iterations. Therefore if Step 2 is employed to its utmost capacity, then the best possible estimate of the sample standard deviations is 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 should be calculated and classify it 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 adding a simple updating routine for accumulation. 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. As has been demonstrated the misclassification may occur by Step 2 only if the intra-distance of samples in any cluster center is greater than half of the

inter-distances between the two clusters. Further, the proposed 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 inter-distance between two clusters is much larger than the sum of the intra-distances for individual clusters. This can be easily shown by noting that the intra-distance for each cluster center should be shorter than only one-half of the inter-distance between the associated pair of cluster centers in order to have a perfect (i.e., completely correct) classification. Consider the most trying but still completely separable clustering by the minimum-distance criterion, namely, the larger of the two intra-distances is equal to one-half of the inter-distance between clusters and the shorter one is much smaller. For such a case, this generalized similarity criterion will set the hyperplane decision boundary at a distance twice the shorter intra-distance from the cluster center. So, a perfect classification will also result.

It is worthwhile to note that the cluster centers established by Steps 1 through 3 can be joined or merged together in order to reduce the total number of cluster centers. However, with regard to saving of computation time, it will be better to start off using fewer clusters than merging the established clusters.

The results of clustering by Step 3 can be displayed in a 2-dimensional map for the multispectral observations such as the multispectral line scanner. In addition, all the statistical parameters and sample probability density functions can also be calculated at the last iteration of Step 3 and printed out together with the 2-dimensional map.

2.3 MERGING OF SEQUENTIAL AND K-MEANS CLUSTERING

Before describing how the statistical sequential clustering technique and the generalized K-means clustering technique can be combined into more powerful clustering techniques, the merits and drawbacks of each technique will be discussed. This review then points out a natural way for combining these two techniques.

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, to the author's knowledge, has never been accomplished in any existing clustering techniques. This feature also permits fairly fast computation. However, because of only one pass of the data sequence, the null class of unidentifiable data samples that resulted from establishing new classes can not be reexamined, which 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 establishing cluster centers. Its main drawbacks is that the procedure for choosing the initial cluster centers is quite arbitrary and requires as many passes of the entire data sequence as the number of cluster centers. Furthermore, because of these rather inaccurate initial cluster centers, many iterations of the entire data sequence will be further required to achieve good clustering accuracy.

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

(1) The given data sequence will be processed by the SSC technique with only a single pass of the entire data sequence. The outputs of the processing will be the mean spectral vectors of clusters.

(2) The mean spectral vectors from (1) will be used as the initial cluster centers to the KGC technique. In order to allow for extra cluster centers from the null class of the SSC in (1), the original KGC procedure for establishing extra initial cluster centers can be used as many times as desired. Next, the initial cluster centers will be iterated about 2 to 3 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 number of clusters.

Section III

UNSUPERVISED CLASSIFICATION OF AGRICULTURAL REMOTE SENSING DATA

In order to test and demonstrate the capability of the non-supervised clustering technique, a set of computer programs has been developed. The program to process and analyze a set of most well-known multi-spectral data which was made available by the Purdue University's Laboratory for Applications of Remote Sensing was employed.

3.1 DATA DESCRIPTION

The data were obtained by the University of Michigan multispectral scanner over an agricultural experiment test site near Lafayette, Indiana, from a flight altitude of 2600 feet on June 28, 1966. This set of data was designated as Purdue Flight Line C-1. In particular, only the results from scans 587 through 797 are presented for the purpose of comparing our nonsupervised classification results with LARS's supervised classification results of the same area (ref. 18).

3.2 PRELIMINARY DATA ANALYSIS

Figure 3-1* shows the aerial photo of the target area (about 1 square mile) with the ground truth designation superimposed. The multispectral scanner recorded simultaneously 12 channels of spectral bands reflecting from the earth's surface between 0.4 and 1.0 μm. These spectral bands are listed in Table 3-1. Again for the purpose of comparison with LARS results only 4 channels were used, i.e., channels 1, 6, 10, and 12. These 4 channels have been determined by LARS to be the optimal 4-channel feature selection (based on the divergence measurement) for the flight line C-1 data (ref. 18).

Figures 3-2 through 3-5 show the probability histograms of each individual channel, respectively. These histograms clearly show that the majority of resolution elements (or target) having the spectral radiance between 140 and 200, with the total radiance range being 0 to 256. Further,

Figure 3-1 through 3-35 are presented following the text at the end of this Section.

Table 3-1. SPECTRAL BANDS OF MICHIGAN MULTISPECTRAL SCANNER

CHANNEL NO.	SPECTRAL BANDWIDTH (microns)	CHARACTERISTIC COLOR	
1	0.40 - 0.44	Violet	
2	0.44 - 0.46	Blue	
3	0.46 - 0.48		
4	0.48 - 0.50	Blue-Green	
5	0.50 - 0.52		
6	0.52 - 0.55	Green	Visible
7	0.55 - 0.58		
8	0.58 - 0.62	Yellow	
9	0.62 - 0.66	Red	
10	0.66 - 0.72	Red	
11	0.72 - 0.80	Reflective near infrared	
12	0.80 - 1.00		

several distinct peaks were observed in each histogram which indicate the mixing of several different populations as expected. However, these peaks are not completely separate. This fact implies that more than one channel out of these four would be required for discrimination between different populations.

Figures 3-6 through 3-9 show the corresponding grey-level plots of the channels used. The road running in the flight direction in the middle of the area is indicated by a blank. Several other rectangular agriculture fields can also be observed from these plots. In particular, one can see the close correspondence of the two wheat fields in Figures 3-1 and 3-8. It should be noted that the complement of the numerical value with respect to 256 is proportional to the spectral radiance received by the scanner. Hence, the larger the numberic used in the grey-level plot, the smaller the spectral radiance.

Figures 3-10 through 3-12 show three scatter plots between channels 1, 6, and 10, The number 1 through 8 used indicates the number of samples in

each spectral cell, while number 9 indicates the number of samples to be 9 or greater. Two things can be observed from these plots. First, the scatter patterns of Figure 3-10 with Channel 1 versus Channel 6 and Figure 3-11 with Channel 1 versus Channel 10 are alike. This indicated that Channel 6 and Channel 10 are possibly linear related. This prediction is further confirmed by the scatter pattern in Figure 3-11 with Channel 6 versus Channel 10. Second, the scatter pattern in each figure does not indicate clear cut clusters, which implies the impossibility of completely correct discrimination basing on any two-channel pairs out of channels 1, 6, and 10. That is, three or more channels of data are needed simultaneously for discrimination between different crops in this set of data.

Figure 3-13 shows an inventory boundary map by the boundary enhancement technique (ref. 7) for the target area. One can see the very clear correspondence of the boundaries of different crop fields between this map and the aerial photo (Figure 3-1). One purpose of generating the boundary map is to establish the spatial registration between the multispectral data and the ground scene based on the aerial photo so that the training set can be selected, if needed, as in the supervised classification by LARS. So much for the preliminary data analysis of this particular target area. In the following, the results from the non-supervised classification techniques will be discussed.

3.3 UNSUPERVISED CLASSIFICATIONS

In order to see more clearly the advantage of the composite clustering technique, the results employing, separately, the SSC technique and GKC technique was presented first.

Figures 3-14 through 3-19 show the classification maps by the statistical sequential clustering technique for the numbers of 18, 17, 16, 15, 14, and 13 classes, respectively. Actually, only Figure 3-14 with 18 classes was processed from the data directly. The other classification maps were obtained consecutively by merging the two most similar classes based on the minimum distance criterion. It is interesting to examine the merging process in this series of classification maps. The 18 classes in Figure 3-14 are designated by

alphanumeric symbols 1, 2, ..., 8, 9, A, B, C, E, F, G, H, and I, respectively. The class (I) appears in the last scan 797 from sample numbers 115 through 221 in Figure 3-14. This class was merged into class (1), as shown in Figure 3-15. Next, the class (H) scattering in the rectangle defined by scans 645 and 699, and sample numbers 1 and 45 in Figure 3-15 was merged into class (4) as shown in Figure 3-16. Next, class (G) occupies the rectangle defined by scans 707 and 797, and sample numbers 1 and 19 in Figure 3-16 were merged into class (C) as shown in Figure 3-17. Next, class (F) occupies the right side of scans 791 and 793 in Figure 3-17 were merged to class (1) as shown in Figure 3-18. Up to this stage, four classes (I, H, G, and F) have been merged into other classes. It was noted that the number of samples for each of these four classes is relatively small compared with the total number of samples in the target area. However, in the next merging, the very large class (2) in Figure 3-18 was merged into another large class (1). Comparing the classification maps of Figures 3-18 and 3-19 with 14 and 13 classes, respectively, against the aerial photo, it clearly shows that classes (1) and (2) should be two separate classes. Thus, one may conclude that 14 classes may be the most natural classification of this set of data. Among these 14 classes, the smallest class containing only 19 samples is designated by symbol E in Figure 3-18 or symbol 2 in Figure 3-19.

Next, Figures 3-20 through 3-27 show the classification maps by the generalized K-means clustering technique alone on the same set of data. Figures 3-20 through 3-22 give the classification maps with 18 classes for three stages of clustering, i.e., no iteration and after one and two iterations, respectively. The rest of the classification maps were generated consecuitvely by merging the two most similar classes based on the minimum distance criterion down to 13 classes.

By comparison of the corresponding classification maps by the statistical sequential technique and by the generalized K-means technique, with regard to the ground truth map (Figure 3-1), it seems that the performances by both techniques are about the same with about 70 to 80 percent correct classification accuracy (or clustering). It should be noted that for the same accuracy

of clustering it took only one pass of the data set by the statistical sequential technique, while it took 20 passes (18 passes for establishing initial cluster centers and 2 passes for updating these cluster centers) by the generalized K-means technique.

The results by the composite sequential K-means clustering technique will be discussed next. Figures 3-28 through 3-35 show the classification maps of the same target area (Figure 3-1) by the composite technique. Figures 3-28, 3-29, and 3-30 give the classification map with 13 classes. The classification maps were generated by using the mean vectors of the four channels 1, 6, 10, and 12 of the 13 most populous classes obtained by the statistical sequential technique (Figure 3-18) as the initial cluster centers into the generalized K-means technique. Figure 3-28 gives the classification without any updating of the cluster centers, while Figures 3-29 and 3-30 give the classification, respectively, after one and two iterations. One can see clearly that even without any updating of the cluster centers, the simple reclassification by the K-means technique has produced great improvement in accuracy. After only two iterations (or updating) of the cluster centers, the classification map corresponds very well with the ground truth map (Figure 3-1). From Figure 3-30, one can see that the wheat fields are classified into three classes (6, 8, and C); corn fields into 3 classes (1, B, and D); oats into two classes (9 and 7); soybeams into 2 classes (A and 4); while hay, alfalfa, red clover, and pasture are collectively into two classes (2 and 7). The fact that each of the four crops - wheat, corn, oats, and soybeans are grouped into more than one class simply implied that there existed variations within each spieces of crop. The important point is that the three classes (6, 8, and C) representing wheat, for example, do not mingle with the other crops. Hence, the clustering results for these four crops should be considered correct. On the other hand, lumping all the other crops - hay, alfalfa, red clover, etc., together into only two classes (2 and 7) is due to their very close resemblance in the spectral signature in the four channels used. The difficulty of distinguishing these crops has also shown up in the supervised classification results by the LARS program which will be discussed further later when a comparison is made with these and LAR's results.

The road running vertically through the center of the target area will now be discussed. In Figure 3-30, the road is designated by class symbols 8, B, A, D, 4, and C. Certainly, this designation is not correct. This misclassification of the road, however, can be easily corrected. This is accomplished by increasing one more class, i.e., from 13 to 14, using the K-means technique prior to updating the initial cluster center. The results of this processing are shown in Figures 3-31 through 3-33. Figure 3-31 is obtained without updating, while Figure 3-32 and 3-33 are obtained, respectively, after one and two iterations. This 14^{th} class (E) unmistakably indicates the road as one can see in the middle part (vertically) in Figure 3-33.

Table 3-2 summarizes the quantitative results from the last classification map (Figure 3-33). It may be noted that there are 2 small classes, i.e., classes (3) and (5), with samples 5 and 28, respectively. Both of them belongs to the wheat field at the left of the map in Figure 3-33. They show much stronger spectral radiances in channels 6 and 10 compared with other classes. The causes for this fact is not clear, because of insufficient ground truth information available. Note that only four passes of the data set were required for the classification map by the composite clustering technique.

3.4 COMPARISON WITH SUPERVISED CLASSIFICATION

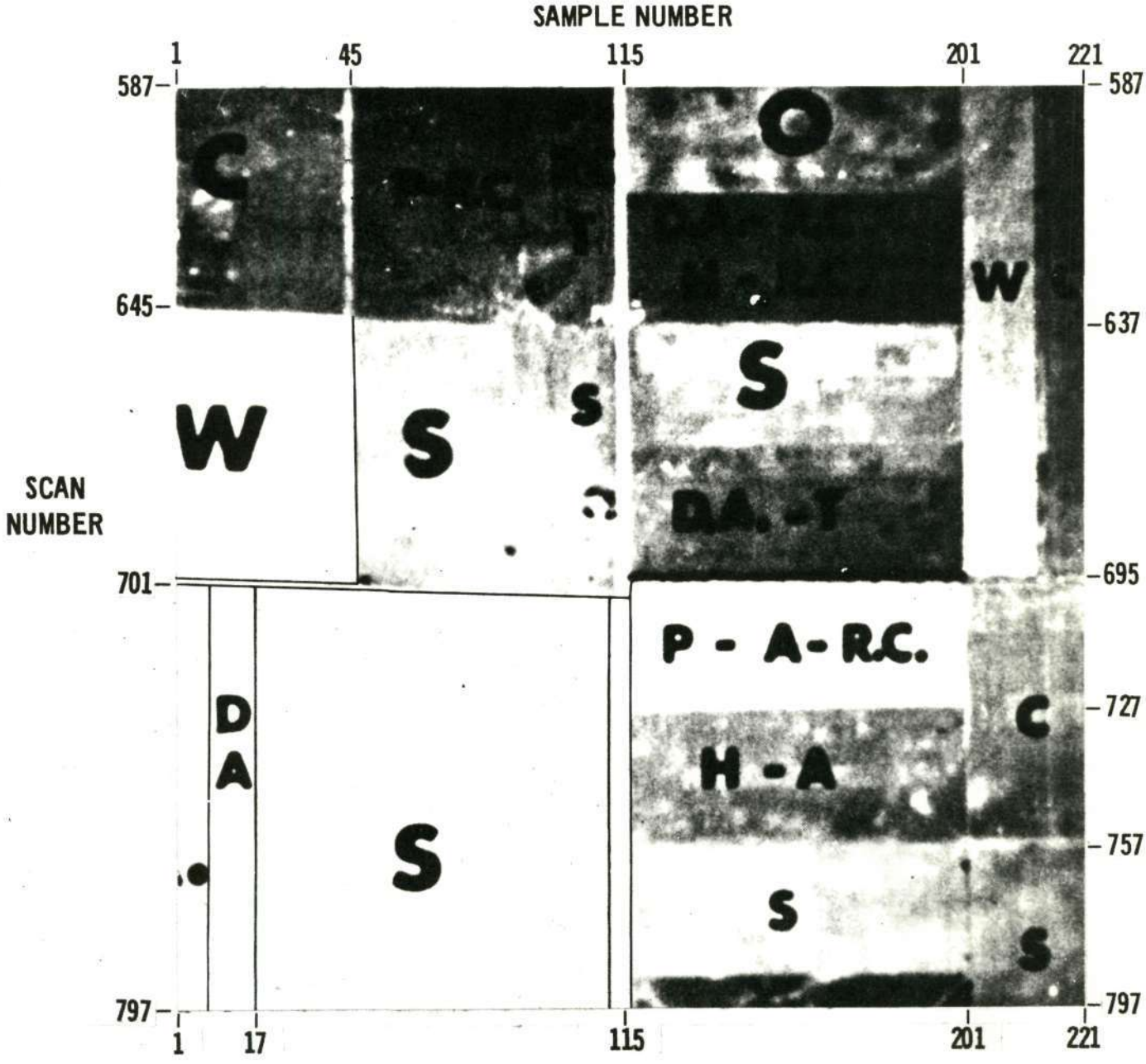
As mentioned earlier, the reason for choosing the particular data set for testing our composite clustering technique is for comparison with the supervised classification results obtained by LARS using the maximum liklihood classification technique. LARS's classification map is reproduced in Figure 3-34 employing the same four channels as were used for the composite clustering discussed above (ref. 18). The training fields used in the classification program are outlined with asterisks (*) and the test fields are outlined with plus (+) signs. The tabulation of classification results of the test fields is also reproduced in Figure 3-35. The test fields chosen in LARS classification covered only about 5989/11660 = 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

Table 3-2. SUMMARY OF MEAN SPECTRAL RADIANCES OF 14 CLASSES BY THE COMPOSITE CLUSTERING TECHNIQUE (FIGURE 3-33)

CLASS NUMBER	CLASS SYMBOL	NO. OF SAMPLES	CHANNEL 1 .4-.44 μm	CHANNEL 6 .52-.55 μm	CHANNEL 10 .66-.72 μm	CHANNEL 12 .8-1.0 μm
1	1	1575	174.2	173.0	183.8	177.9
2	2	3103	179.5	171.3	175.0	150.0
3	3	5	162.8	114.8	88.8	160.2
4	4	1798	166.5	160.5	165.9	172.6
5	5	28	166.8	127.8	106.5	163.7
6	6	523	178.2	166.2	152.3	182.3
7	7	768	181.9	176.3	182.2	163.8
8	8	318	174.2	152.7	138.3	176.2
9	9	255	180.4	172.6	165.0	172.2
10	A	2134	159.3	154.8	159.4	181.1
11	B	852	169.0	167.3	172.4	184.9
12	C	165	168.6	140.9	127.8	170.6
13	D	1736	172.1	165.8	169.4	169.8
14	E	49	140.7	142.5	149.4	178.7

Total No. of Samples = 11,766

on 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, but about 80 percent or less, if the overall performance is based on the entire field. It is also noted from the LARS classification results, as well as the map, that red clover, hay, and alfalfa are fairly similar to each other, with very little discrimination among them. Comparing the classification map by the composite clustering technique (Figure 3-33) with the LAR'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. However, it is important to recall that no training fields or any other ground truth information has been employed in applying the unsupervised technique.



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. AIR PHOTO OF PURDUE FLIGHT LINE C-1 (SCAN 587-797)

This page is reproduced again at the back of this report by a different reproduction method so as to furnish the best possible detail to the user.

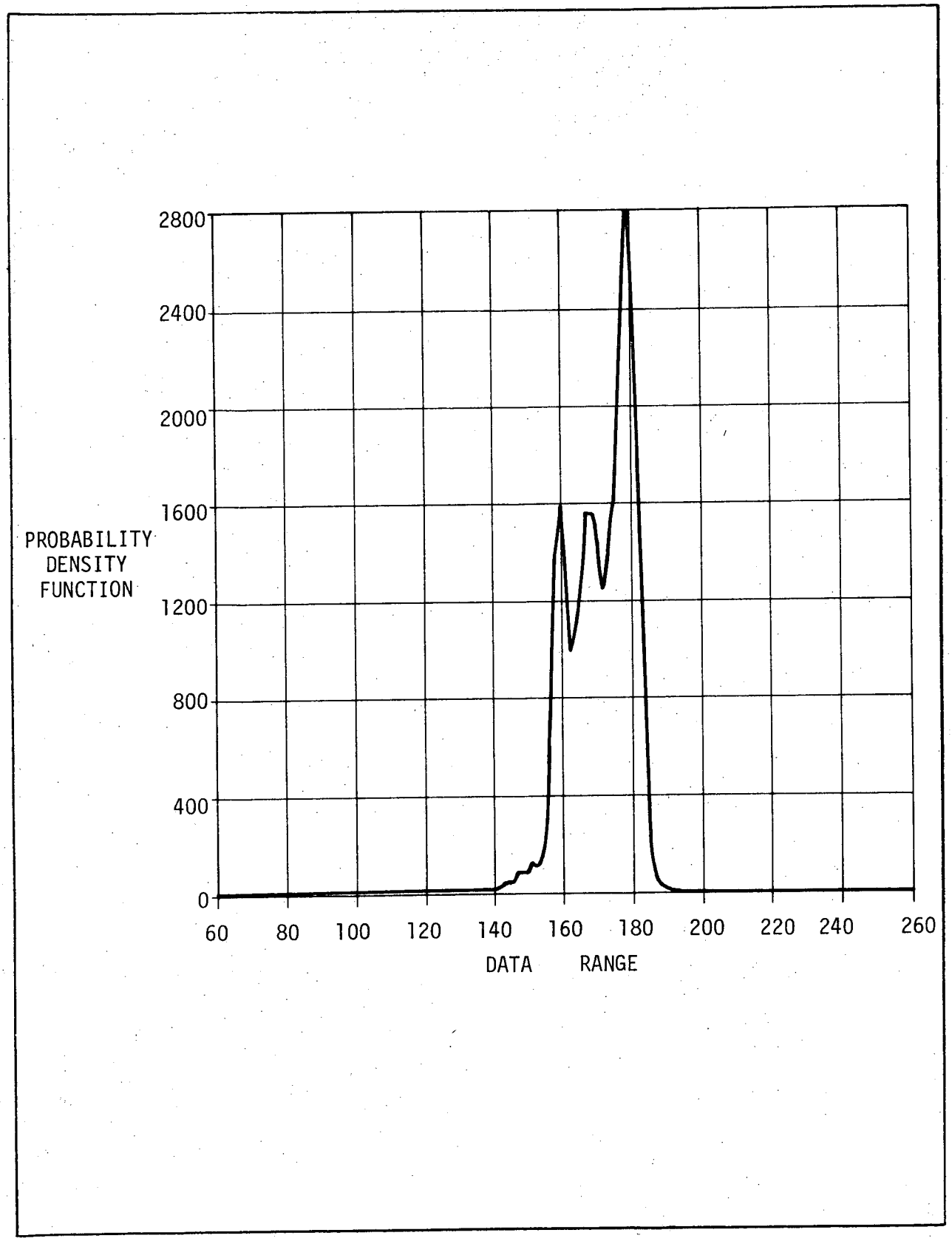


Figure 3-2. PROBABILITY HISTOGRAM OF CHANNEL 1 (0.4-0.44 μm)

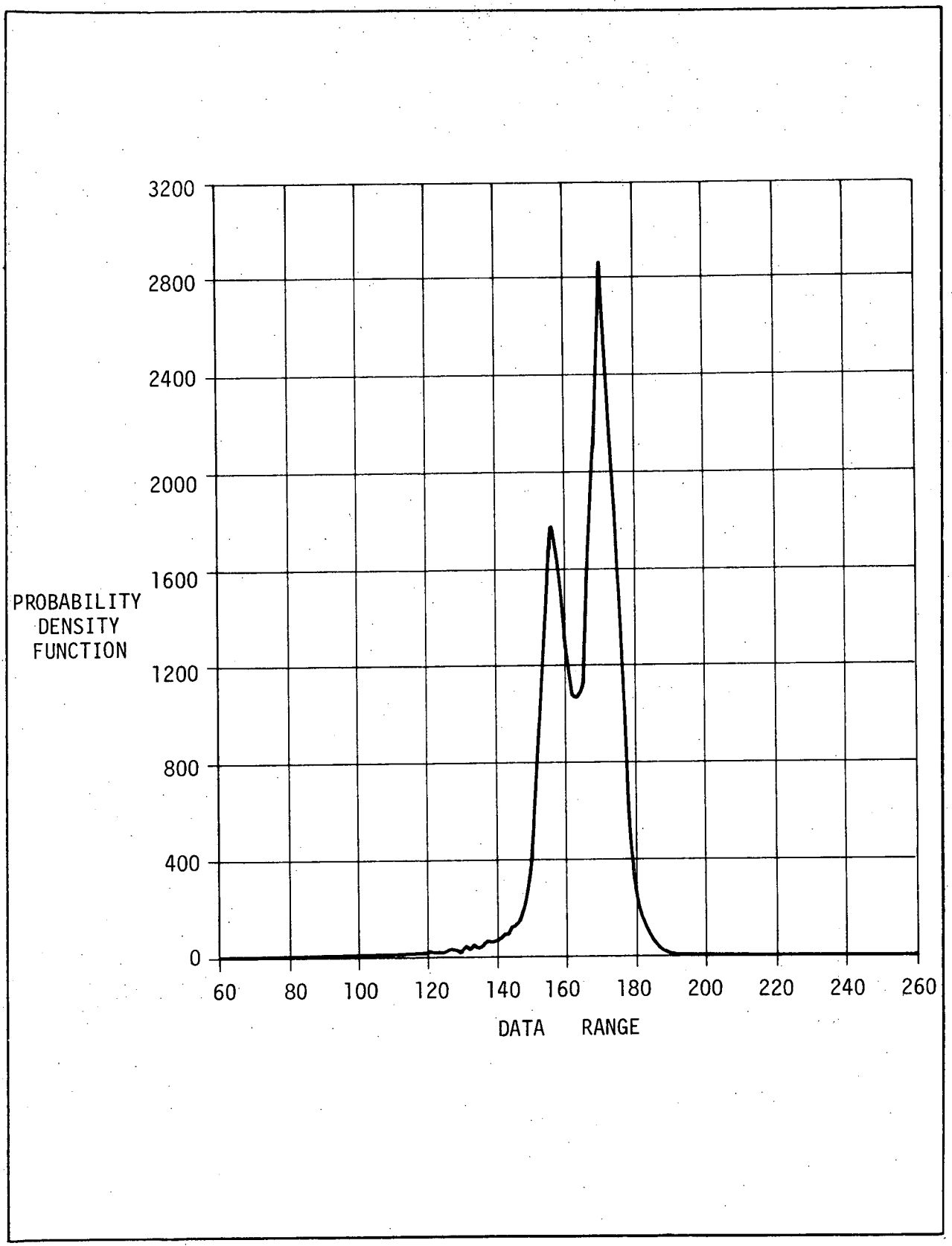


Figure 3-3. PROBABILITY HISTOGRAM OF CHANNEL 6 (0.52-0.55 μm)

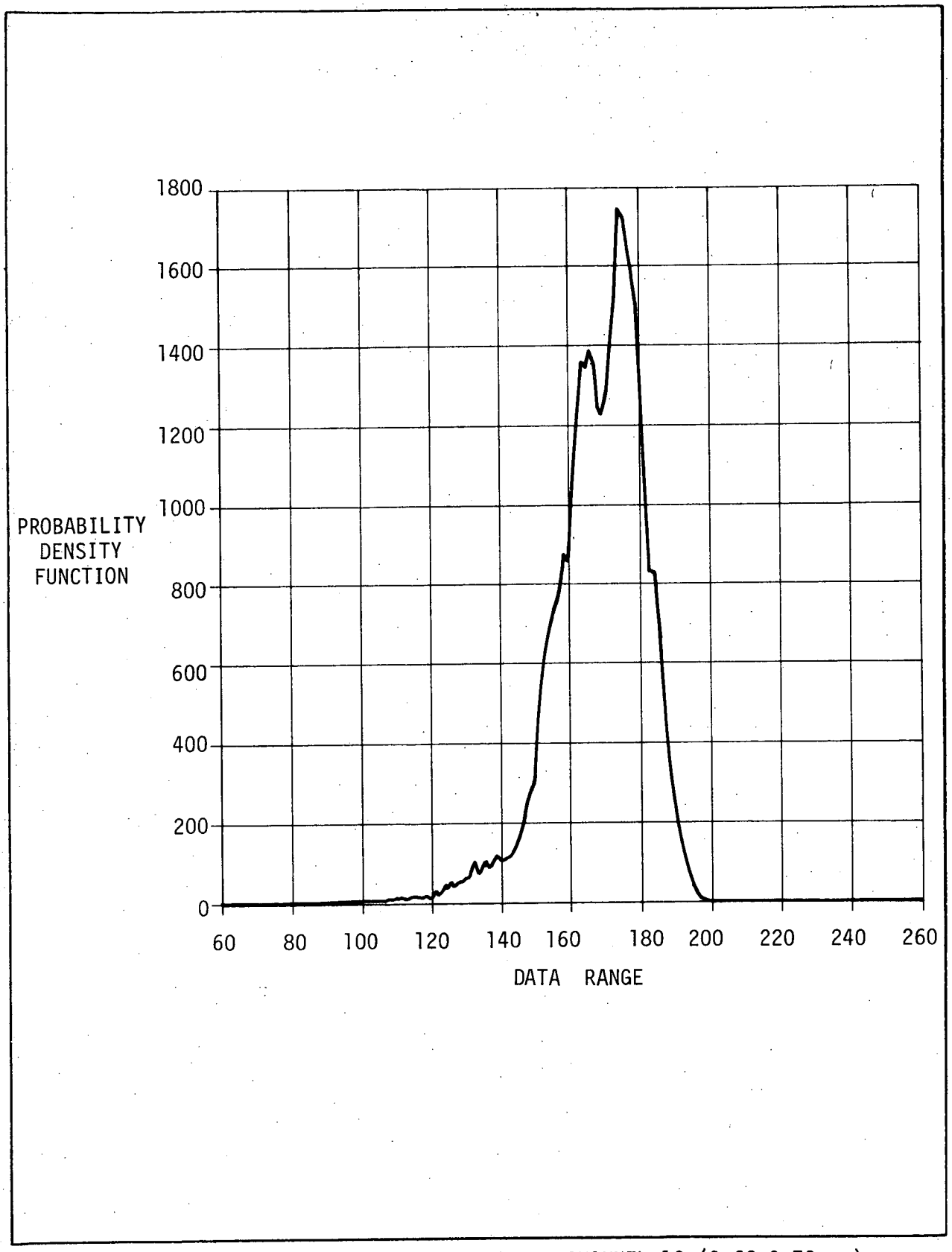


Figure 3-4. PROBABILITY HISTOGRAM OF CHANNEL 10 (0.66-0.72 μm)

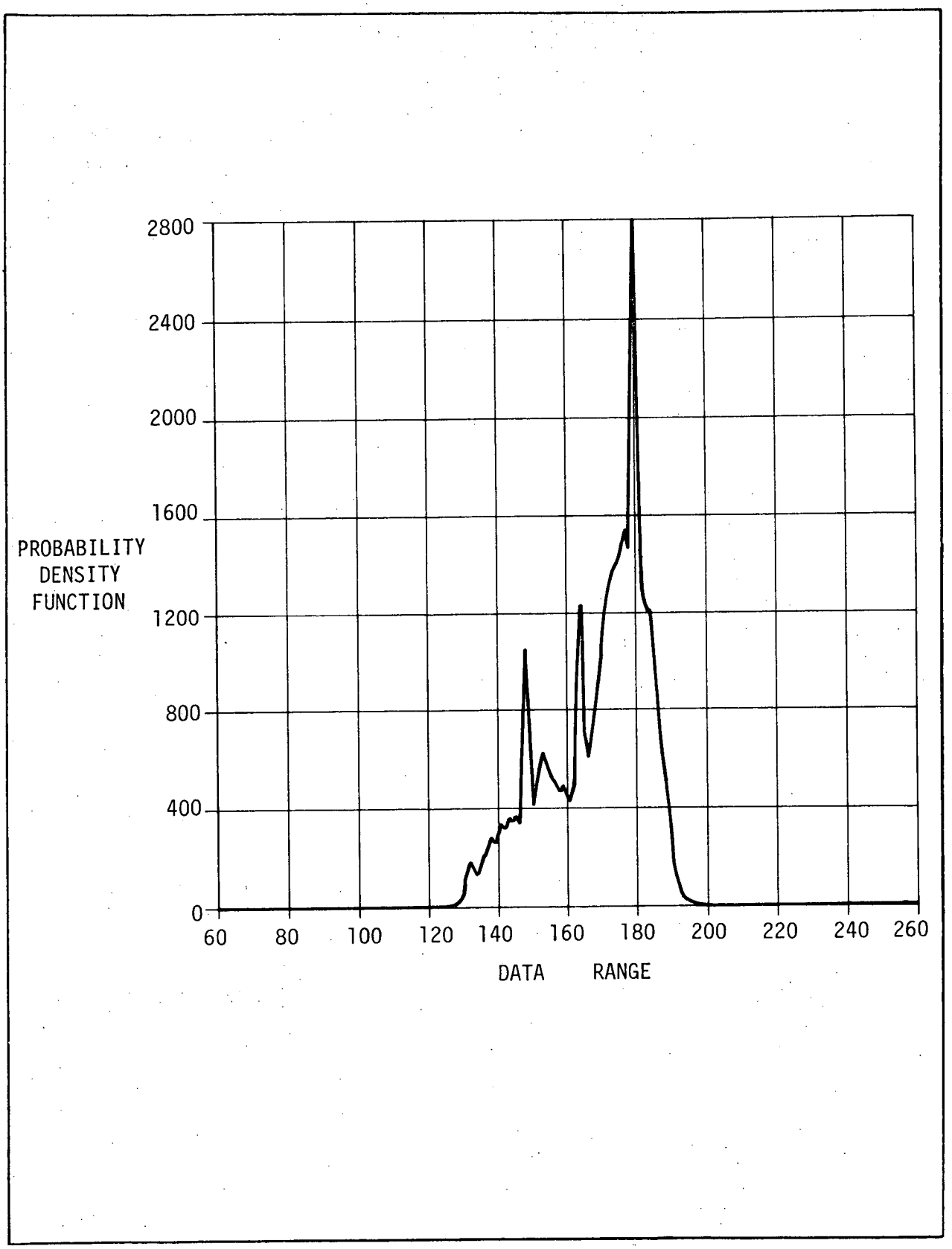


Figure 3-5. PROBABILITY HISTOGRAM OF CHANNEL 12 (0.8-1.0 μm)

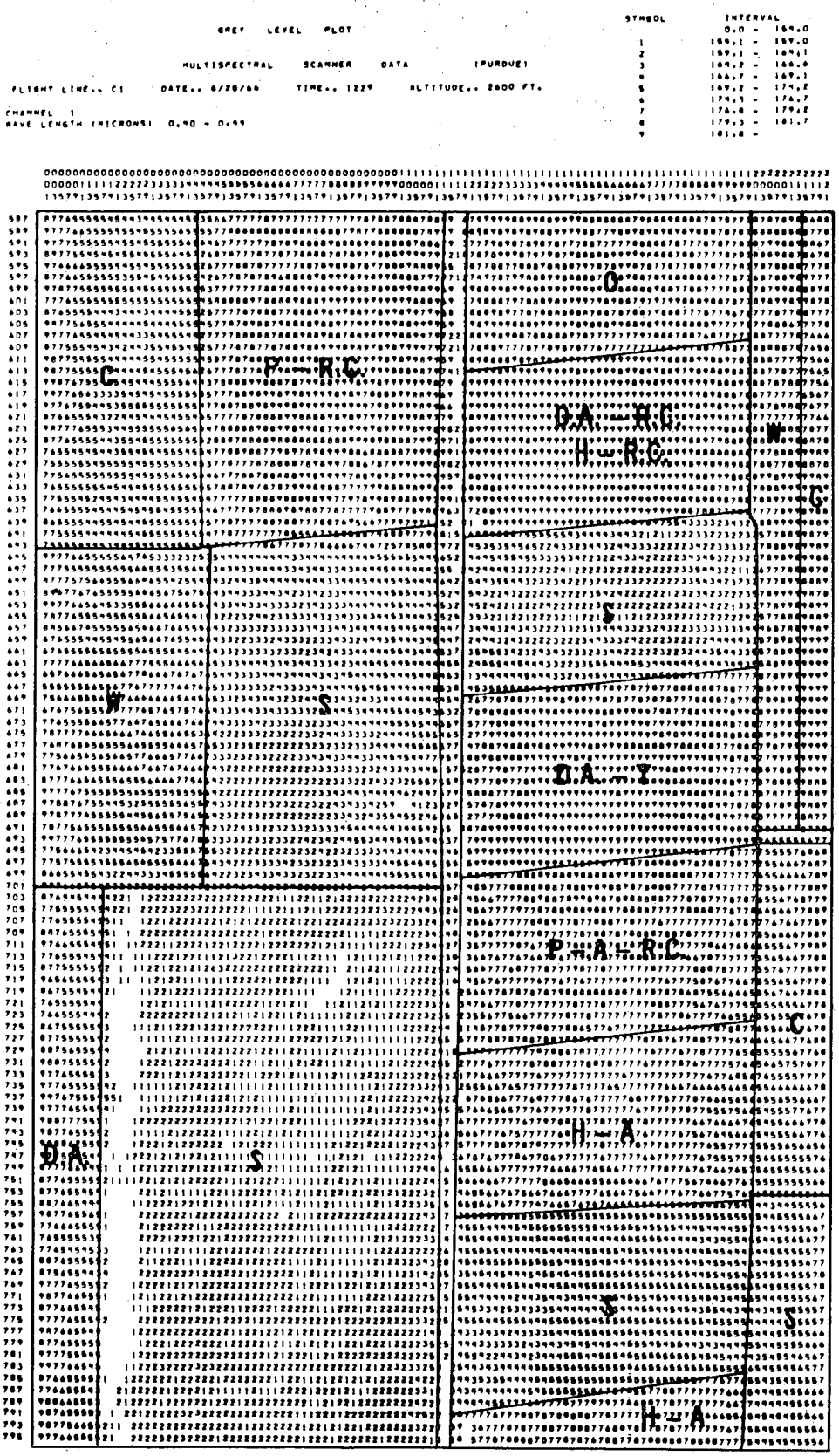


Figure 3-6. GREY-LEVEL PLOT OF CHANNEL 1 (0.4-0.44 μm)

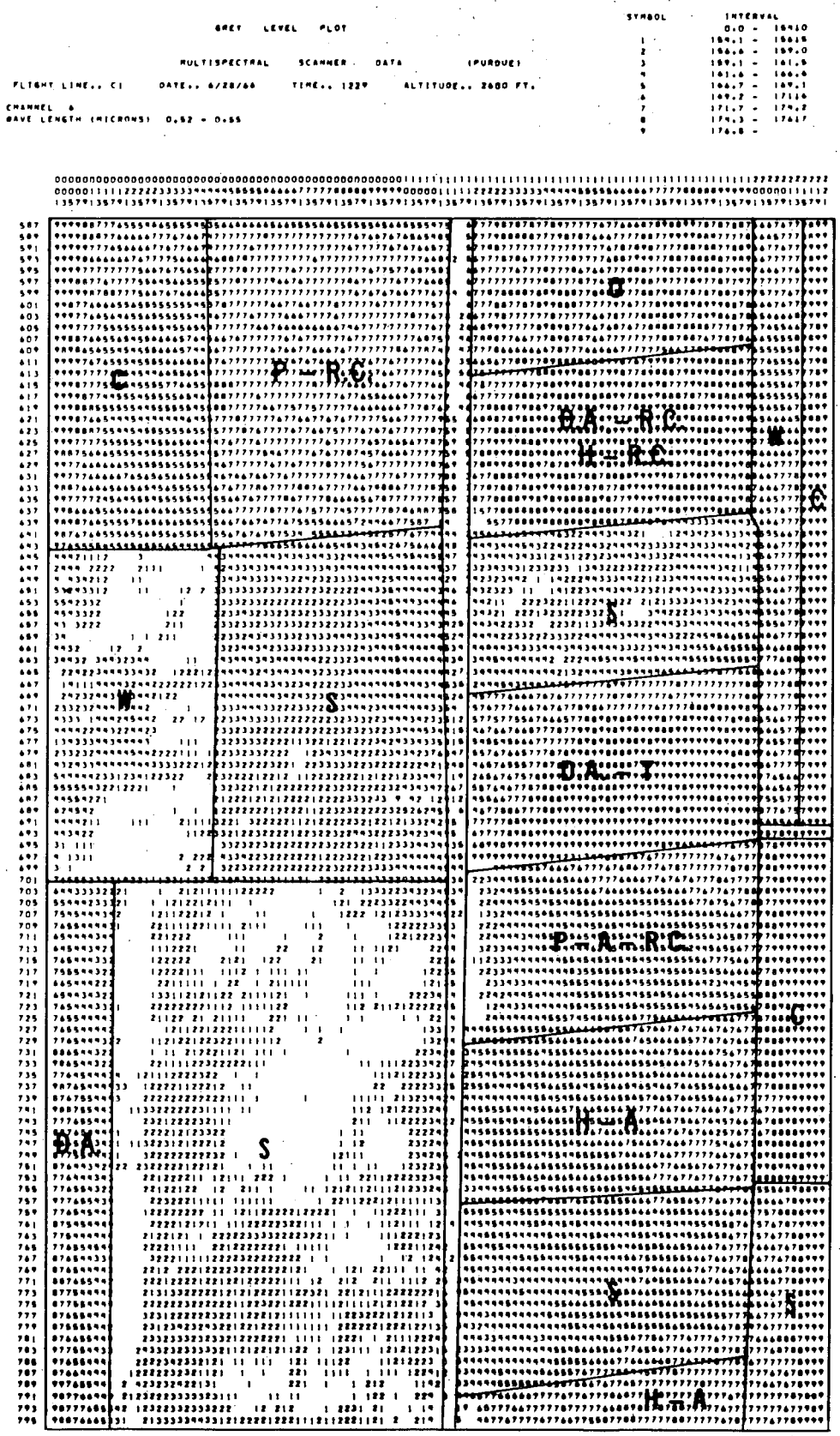


Figure 3-7. GREY-LEVEL PLOT OF CHANNEL 6 (0.52-0.55 μm)

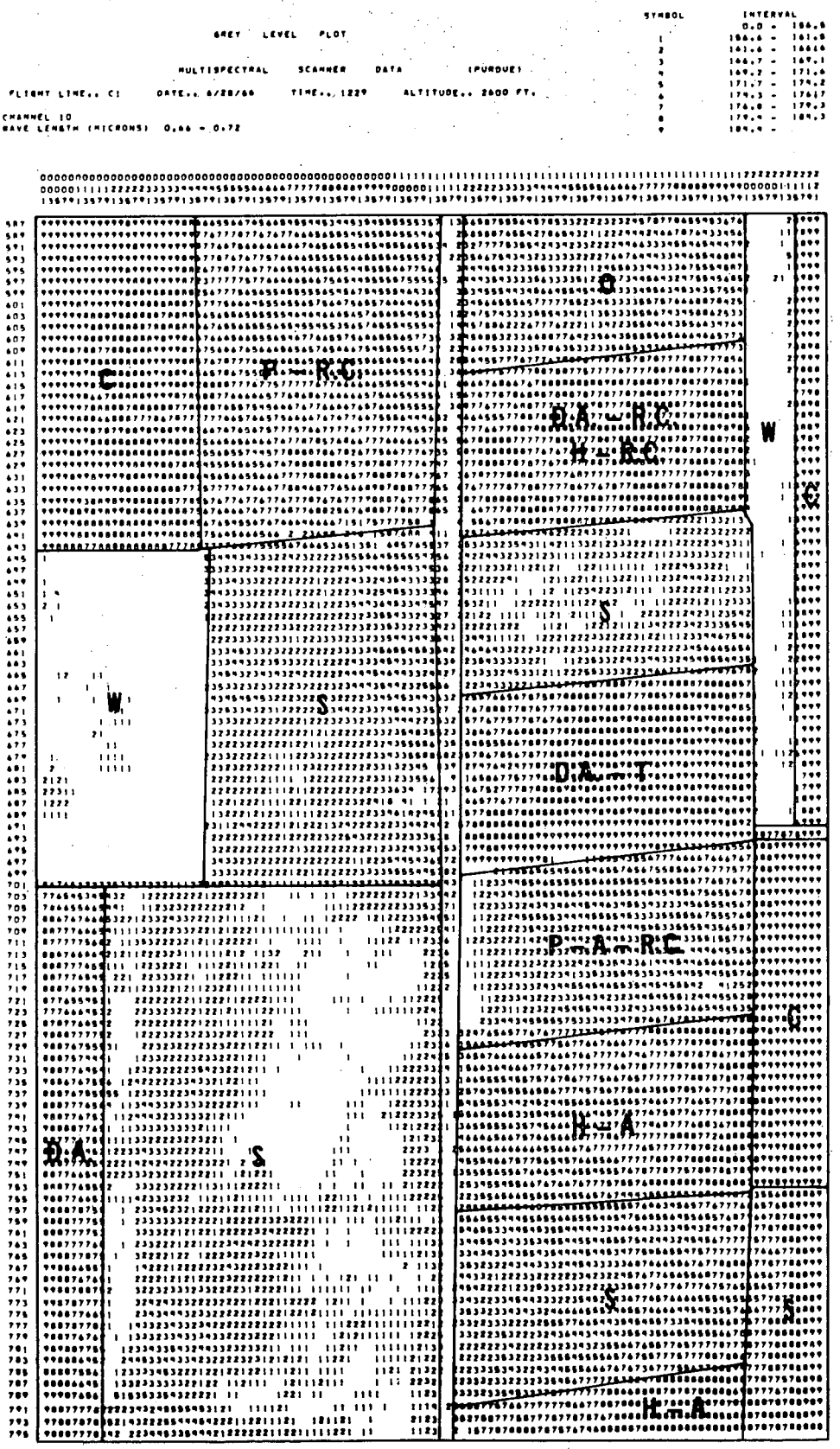


Figure 3-8. GREY-LEVEL PLOT OF CHANNEL 10 (0.66-0.72 μm)

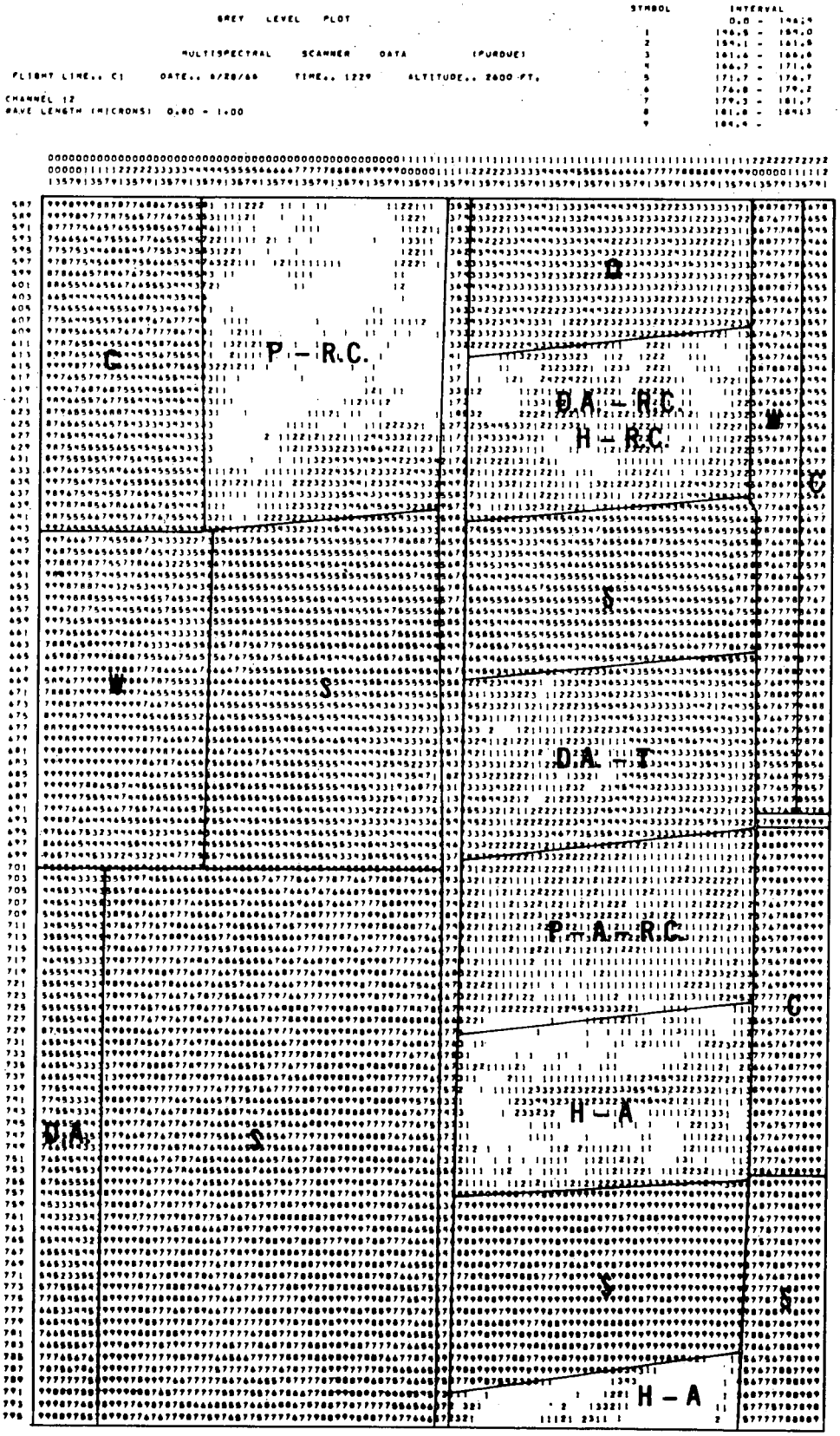


Figure 3-9. GREY-LEVEL PLOT OF CHANNEL 12 (0.8-1.0 μm)

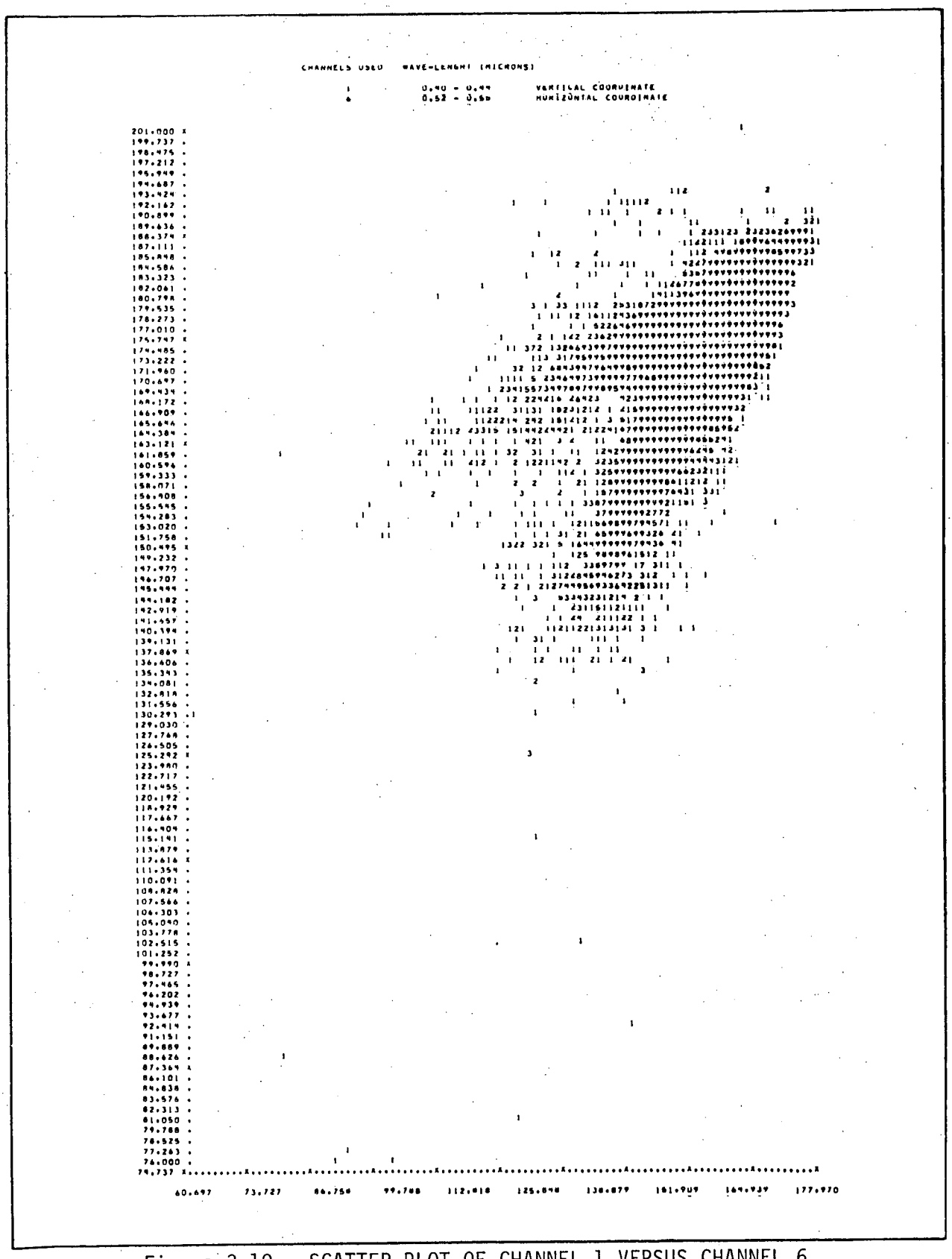


Figure 3-10. SCATTER PLOT OF CHANNEL 1 VERSUS CHANNEL 6

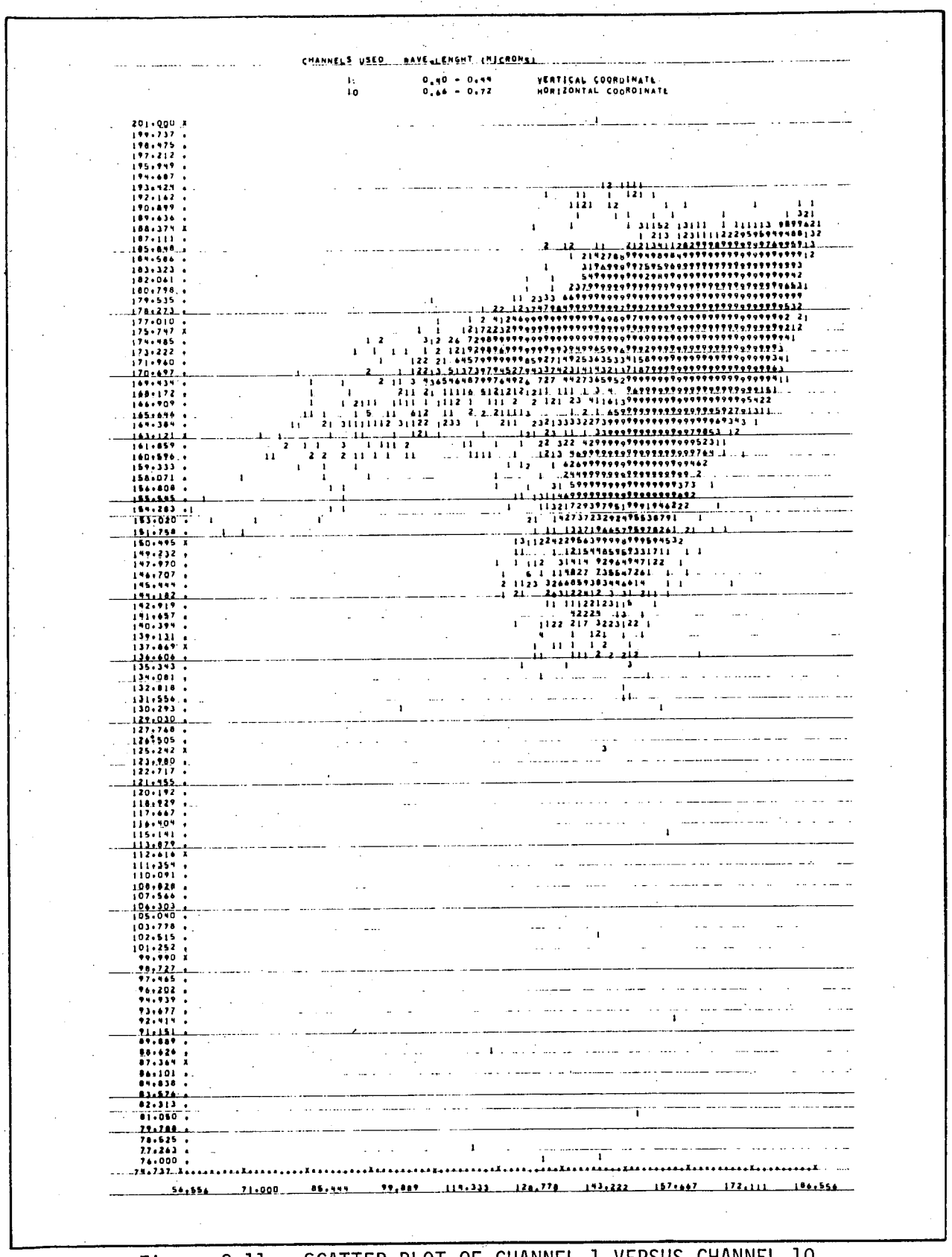


Figure 3-11. SCATTER PLOT OF CHANNEL 1 VERSUS CHANNEL 10

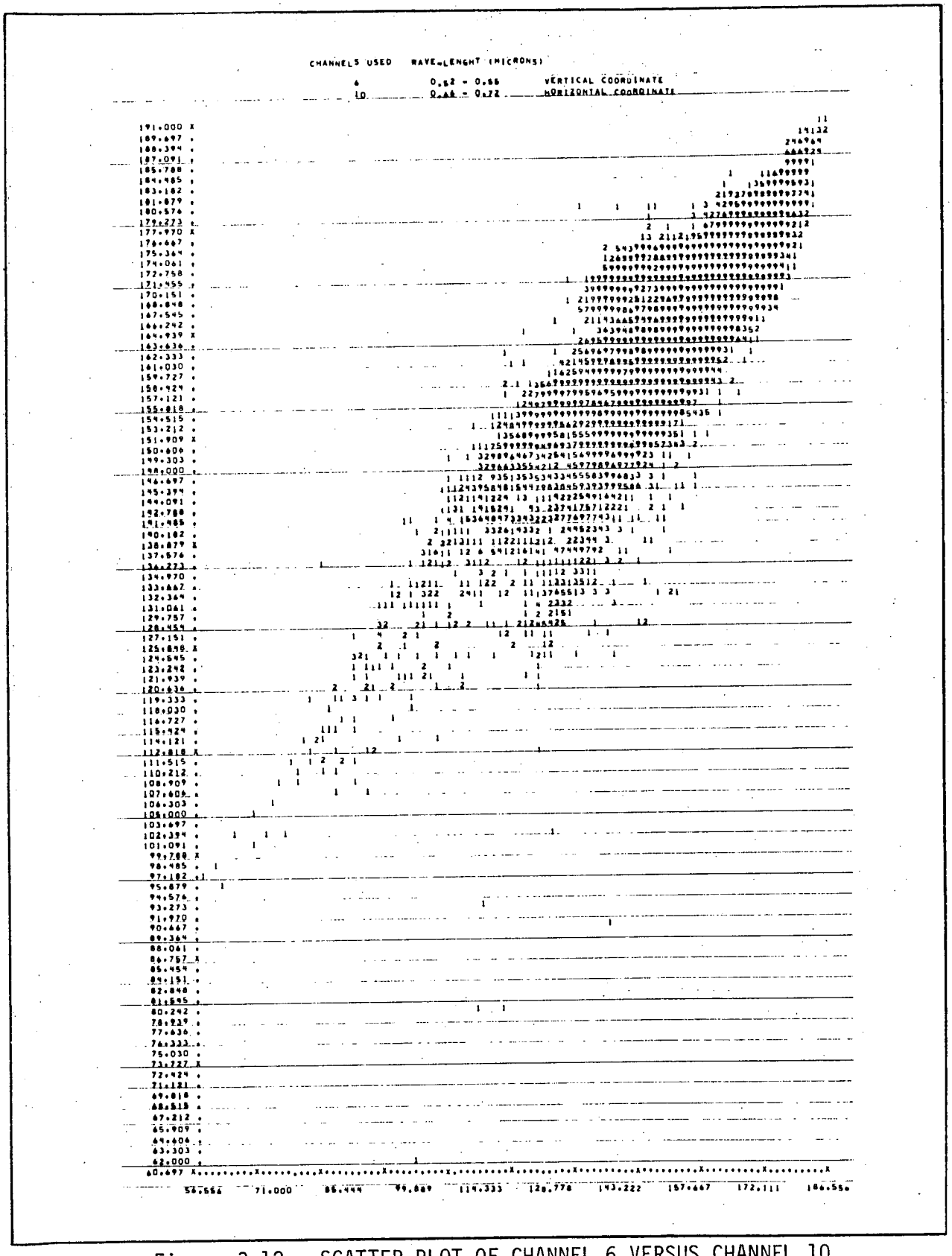


Figure 3-12. SCATTER PLOT OF CHANNEL 6 VERSUS CHANNEL 10

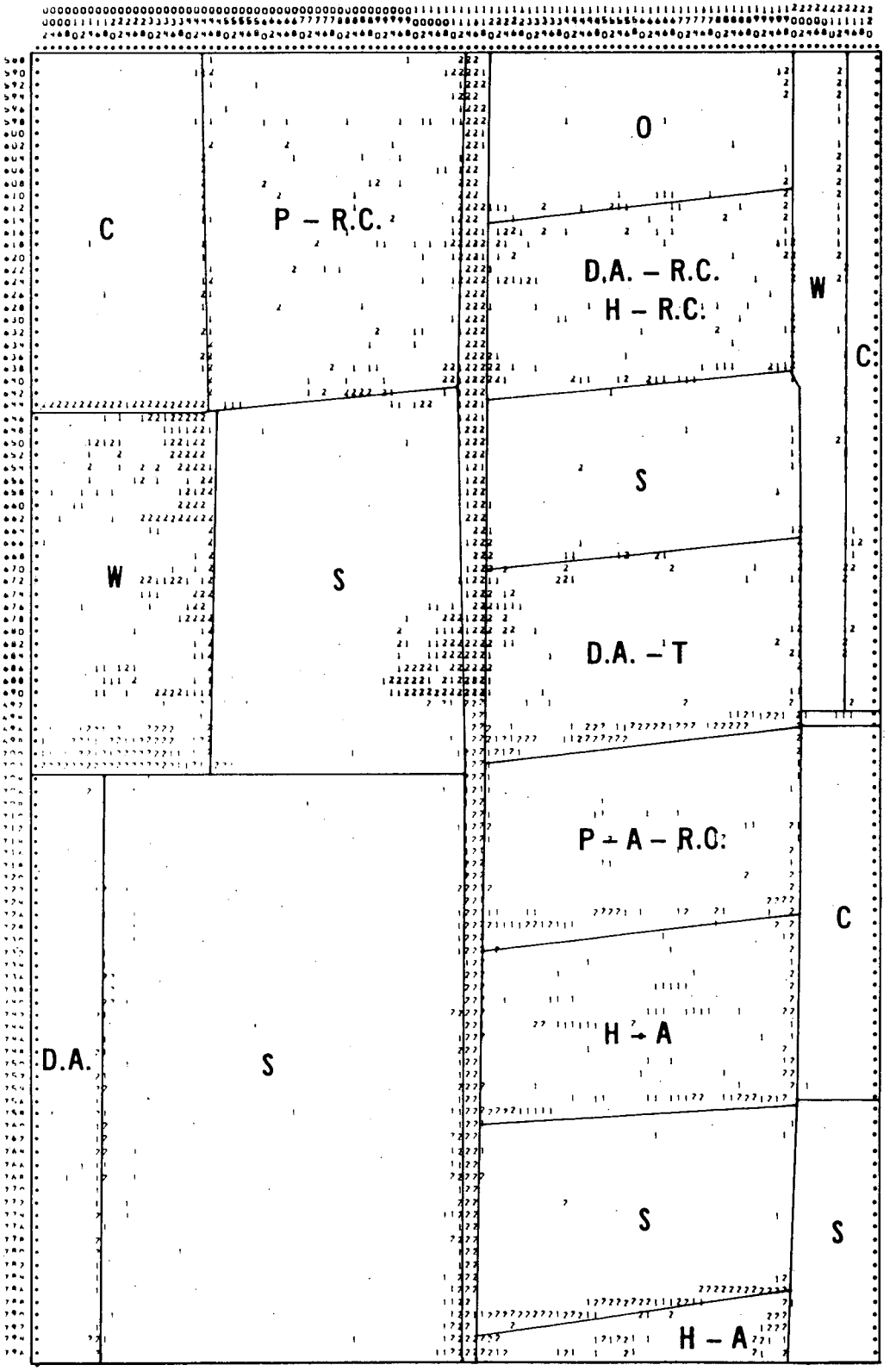


Figure 3-13. INVENTORY BOUNDARIES BY THE BOUNDARY ENHANCEMENT TECHNIQUE FOR PURDUE C-1 FLIGHT LINE

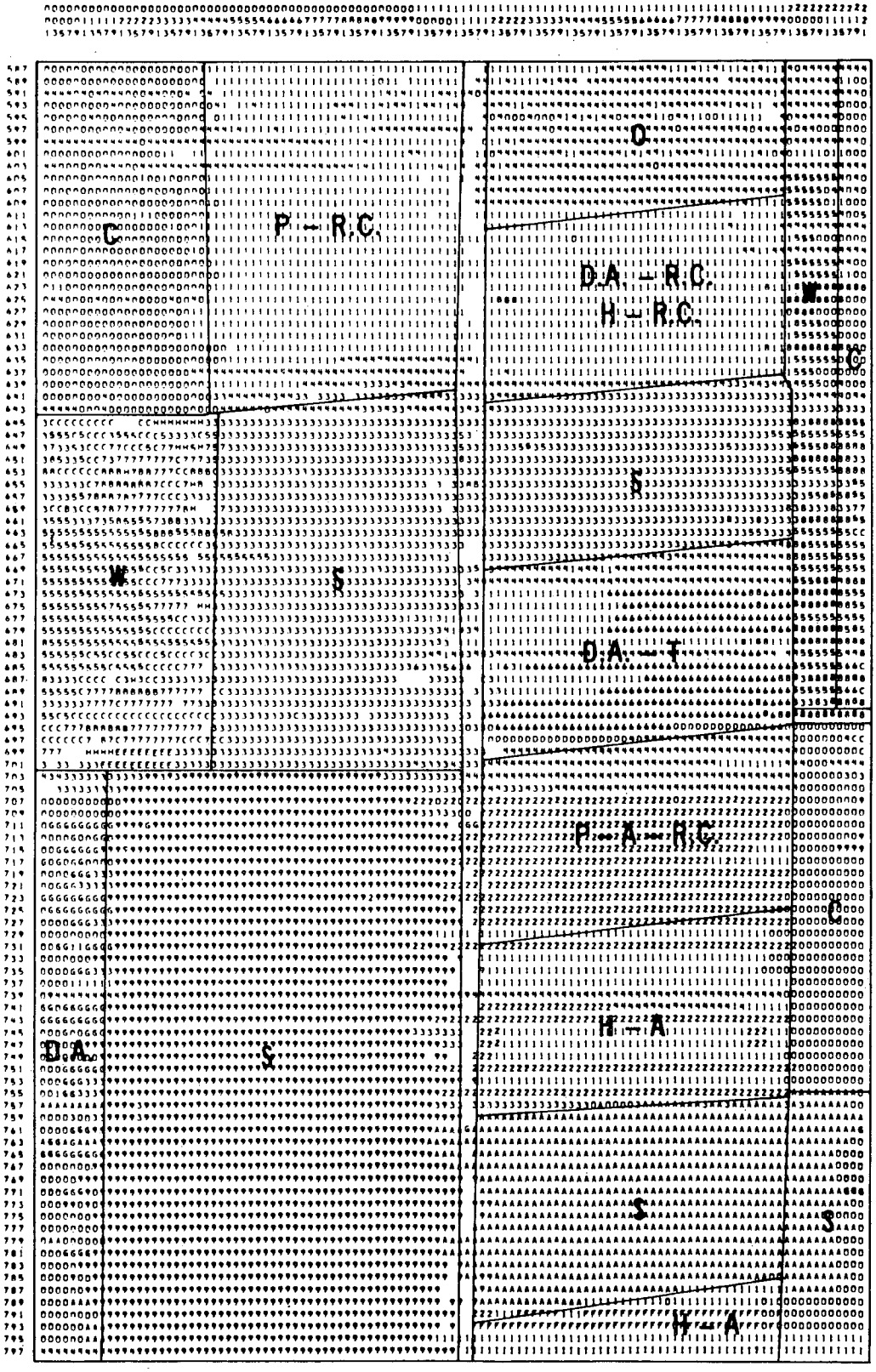


Figure 3-14. CLASSIFICATION MAP BY THE STATISTICAL SEQUENTIAL TECHNIQUE WITH 18 CLASSES

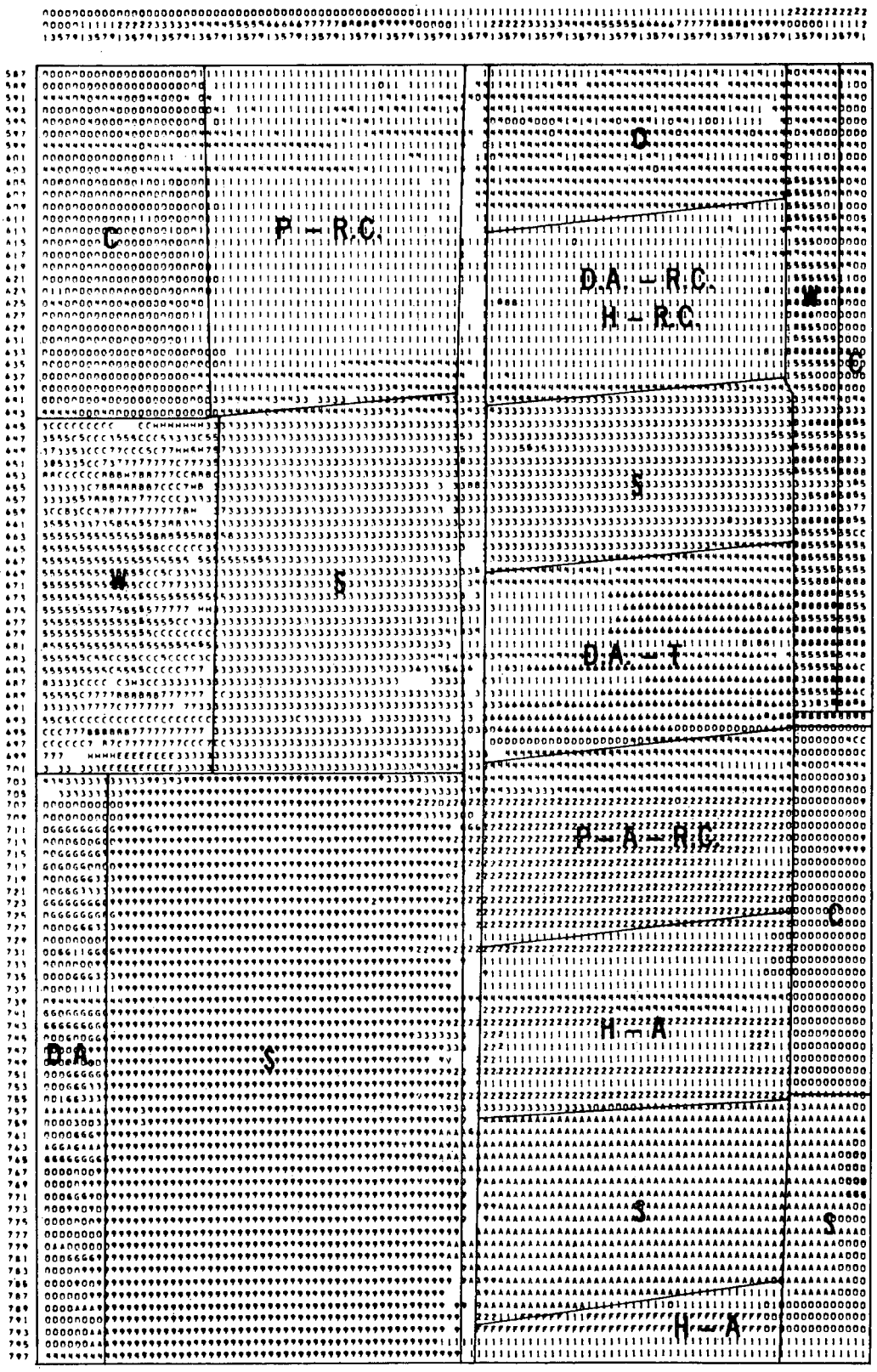


Figure 3-15. CLASSIFICATION MAP BY THE STATISTICAL SEQUENTIAL TECHNIQUE WITH 17 CLASSES

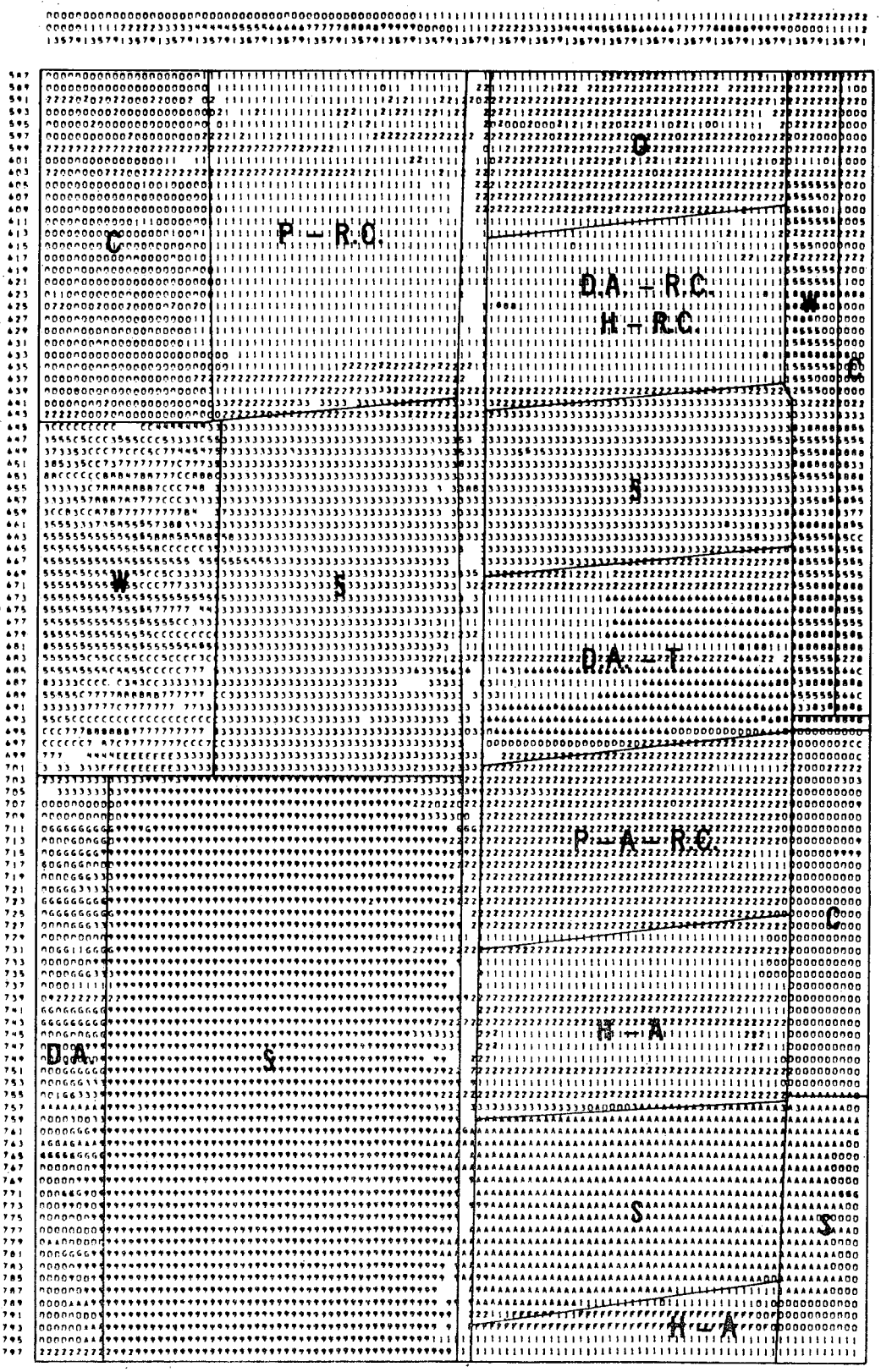


Figure 3-16. CLASSIFICATION MAP BY THE STATISTICAL SEQUENTIAL TECHNIQUE WITH 16 CLASSES

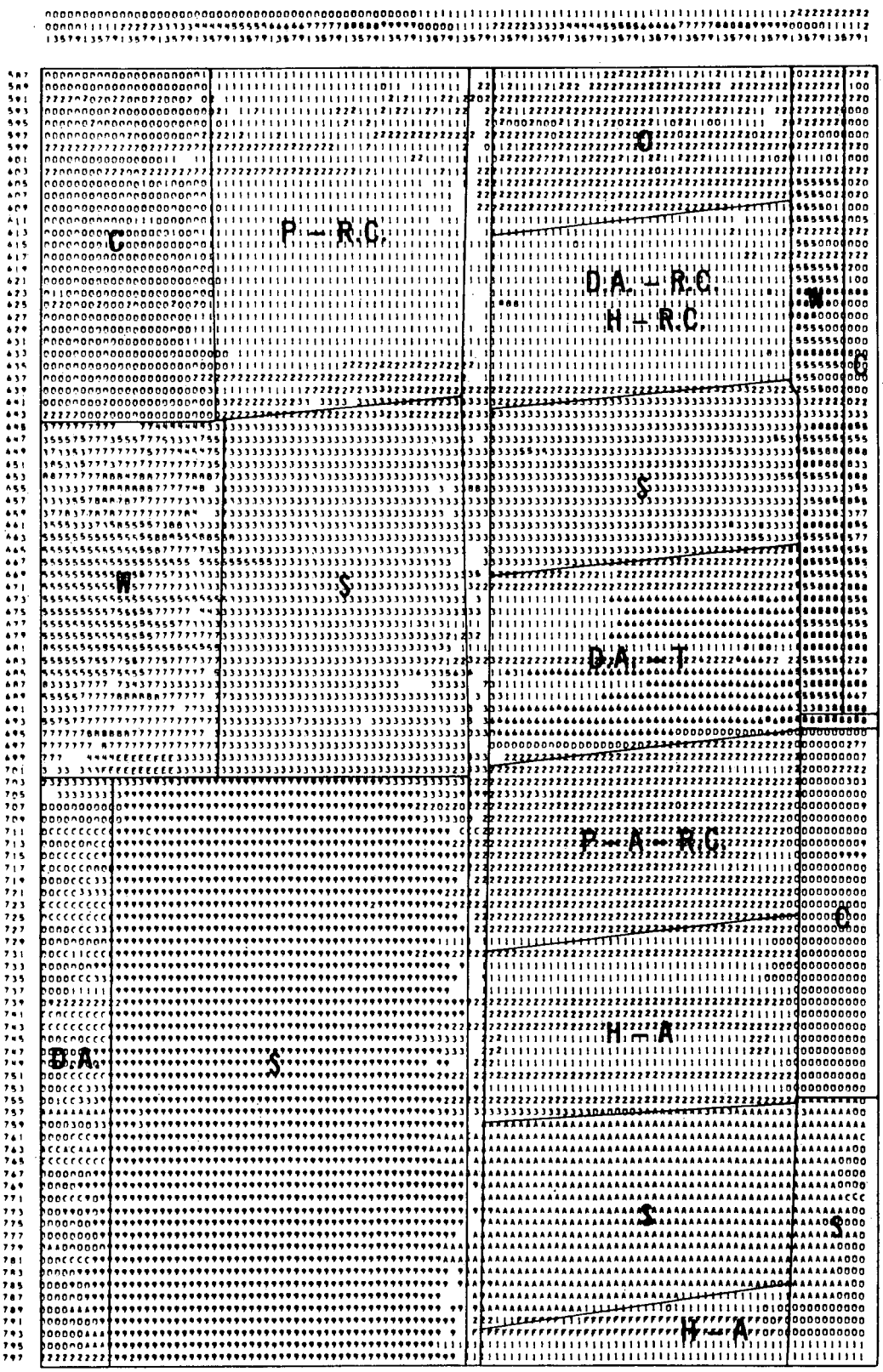


Figure 3-17. CLASSIFICATION MAP BY THE STATISTICAL SEQUENTIAL TECHNIQUE WITH 15 CLASSES

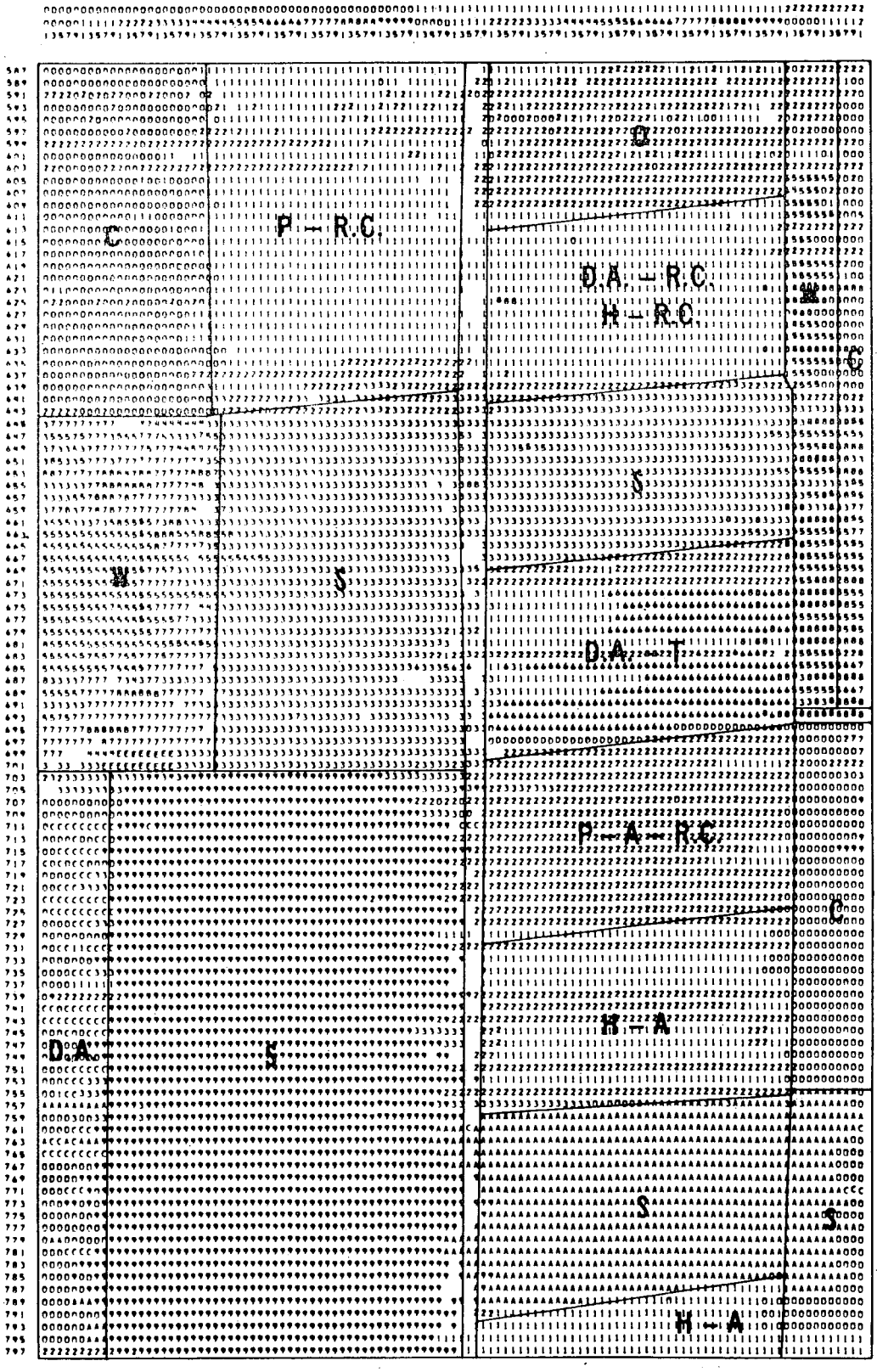


Figure 3-18. CLASSIFICATION MAP BY THE STATISTICAL SEQUENTIAL TECHNIQUE WITH 14 CLASSES

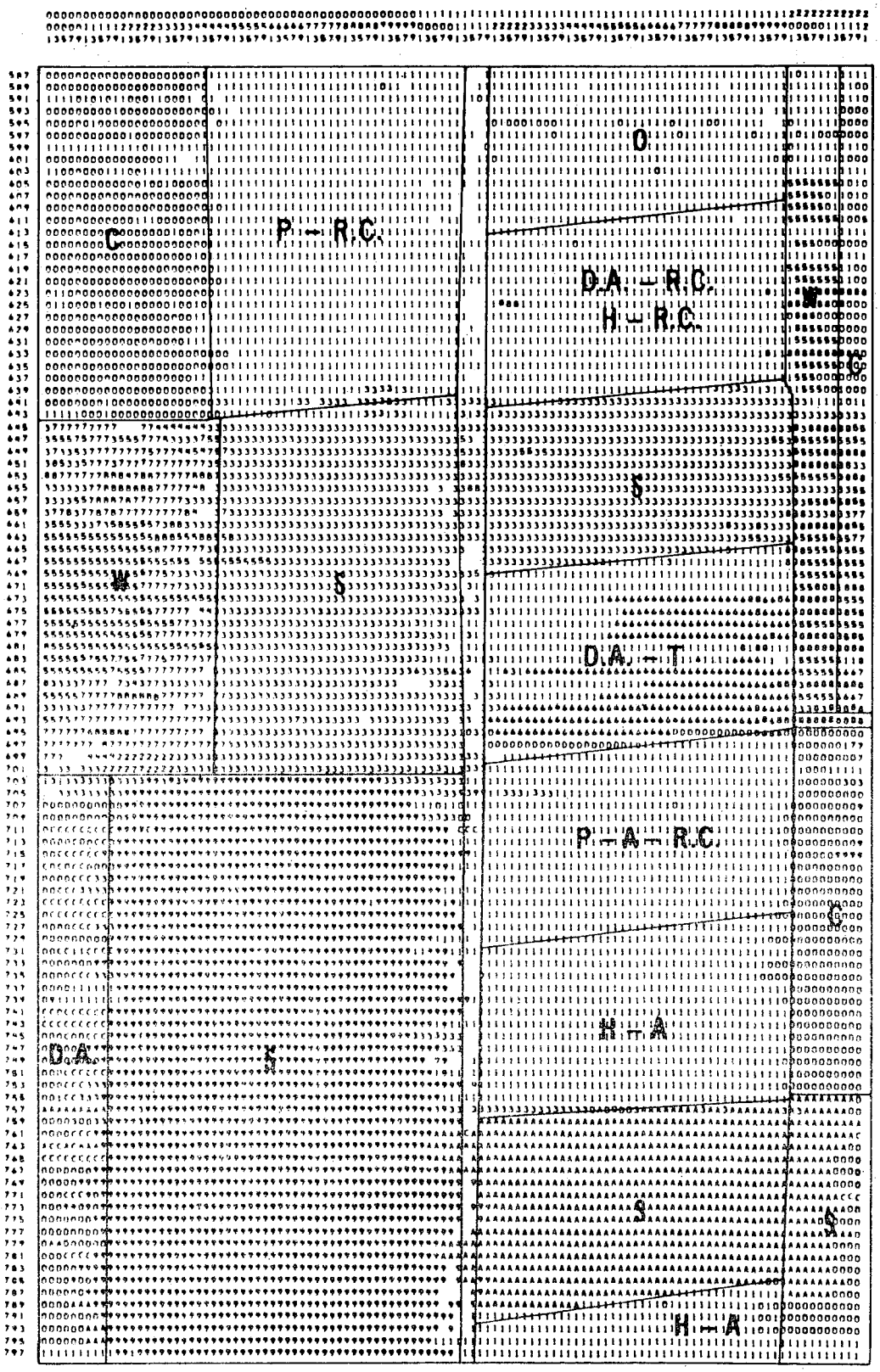


Figure 3-19. CLASSIFICATION MAP BY THE STATISTICAL SEQUENTIAL TECHNIQUE WITH 12 CLASSES

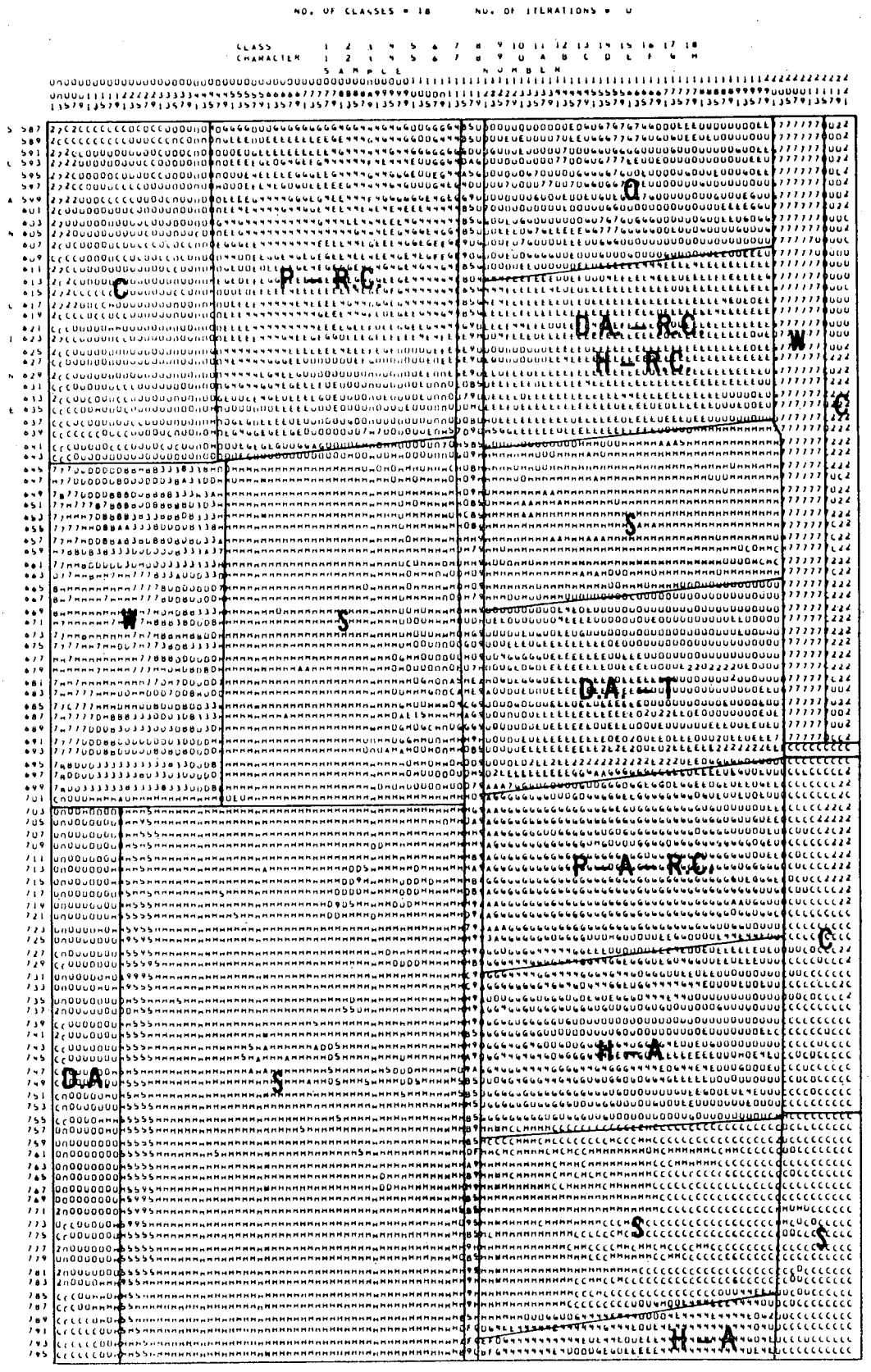


Figure 3-20. CLASSIFICATION MAP BY THE GENERALIZED K-MEANS TECHNIQUE WITH 18 CLASSES AND NO ITERATION

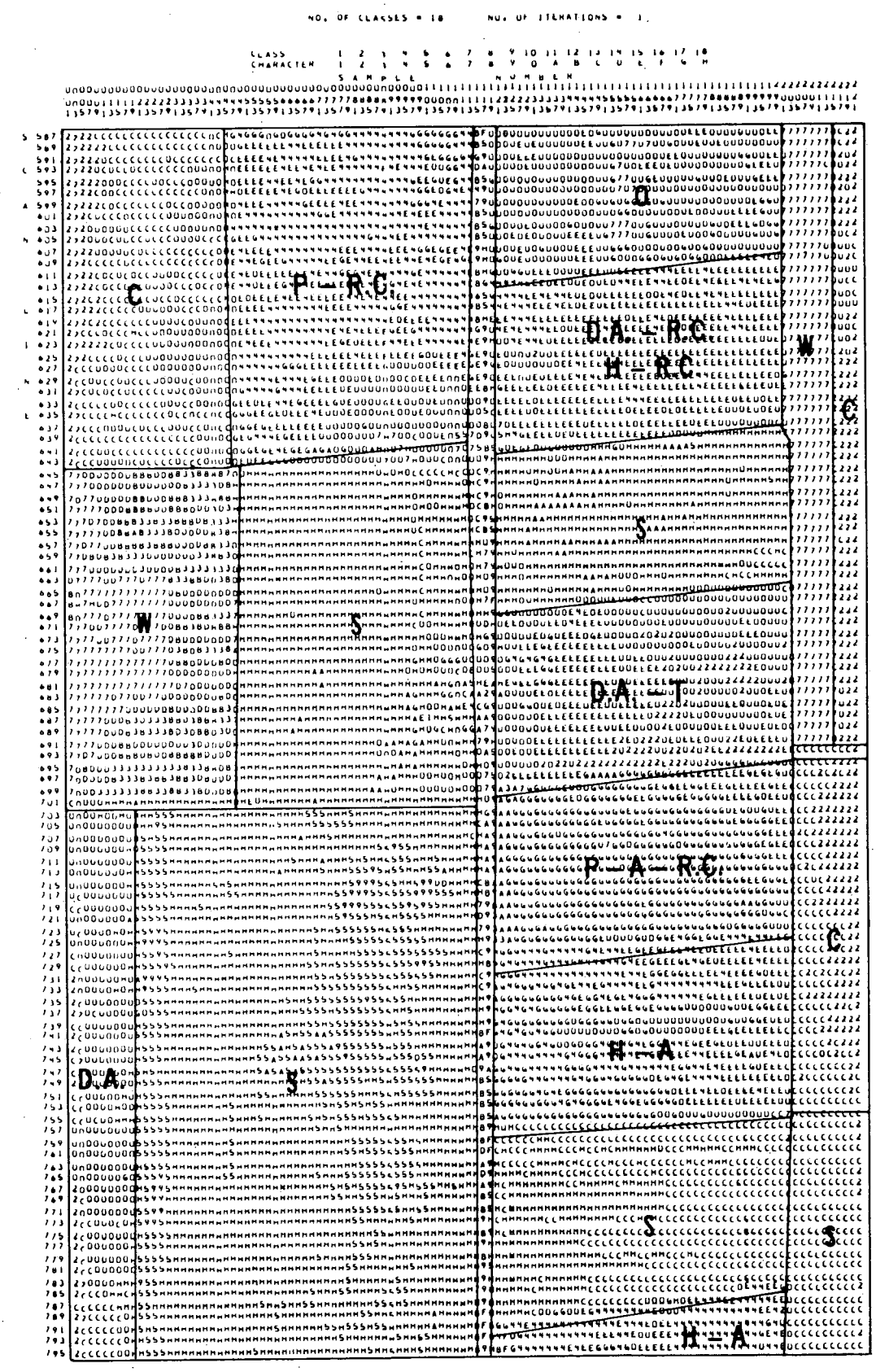


Figure 3-21. CLASSIFICATION MAP BY THE GENERALIZED K-MEANS TECHNIQUE WITH 18 CLASSES AFTER ONE ITERATION

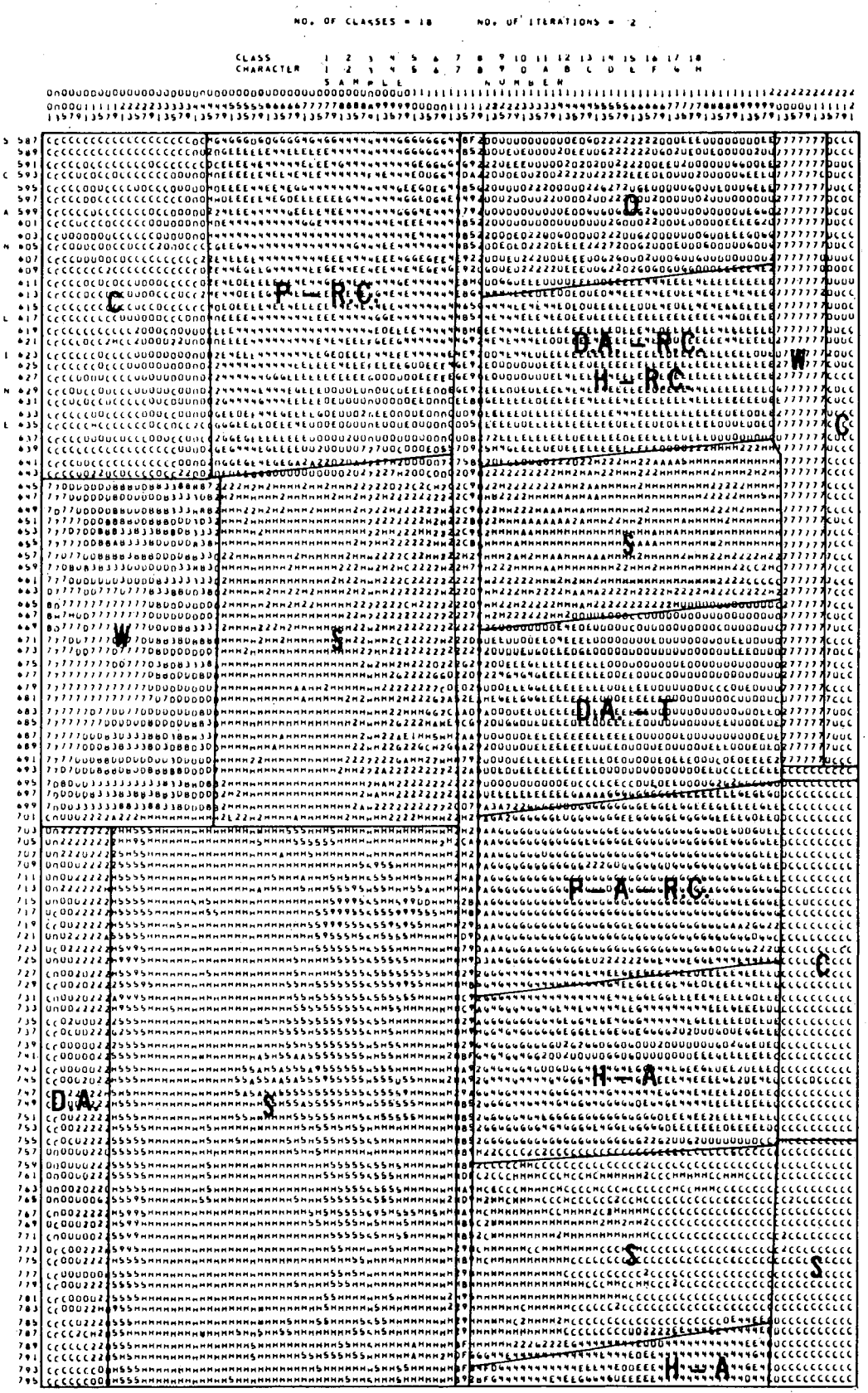


Figure 3-22. CLASSIFICATION MAP BY THE GENERALIZED K-MEANS TECHNIQUE WITH 18 CLASSES AFTER 2 ITERATIONS

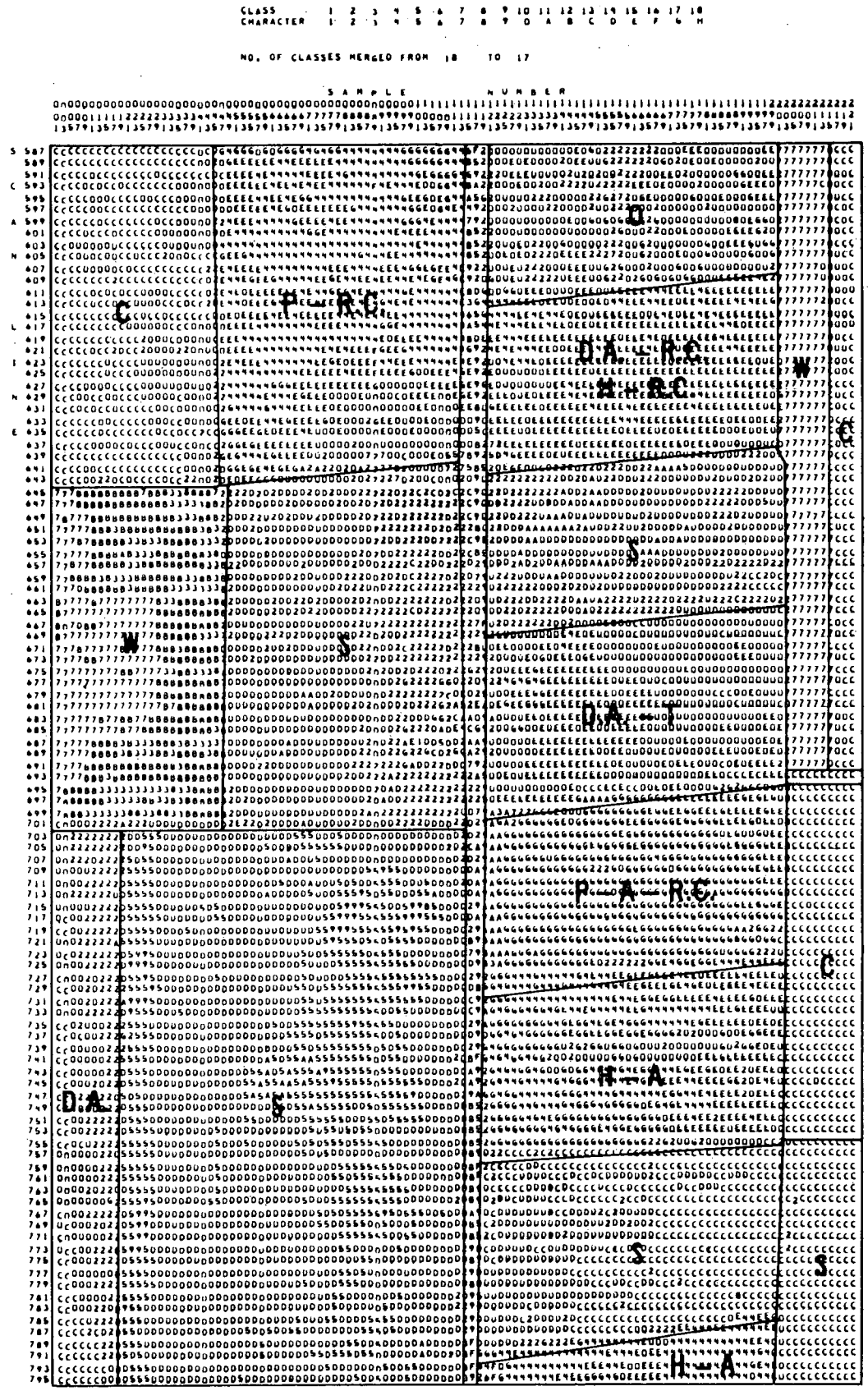


Figure 3-23. CLASSIFICATION MAP BY THE GENERALIZED K-MEANS TECHNIQUE WITH 17 CLASSES

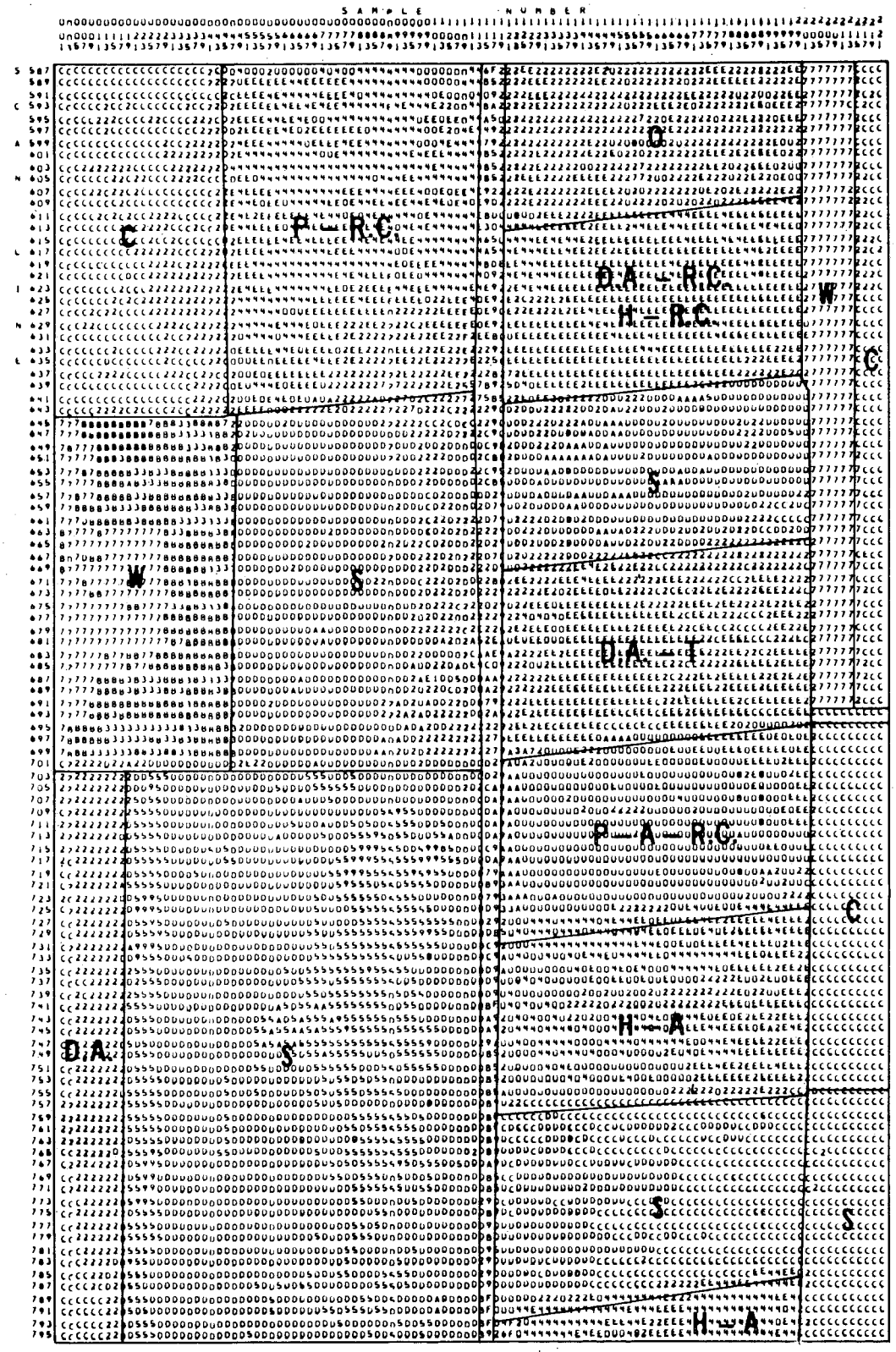


Figure 3-24. CLASSIFICATION MAP BY THE GENERALIZED K-MEANS TECHNIQUE WITH 16 CLASSES

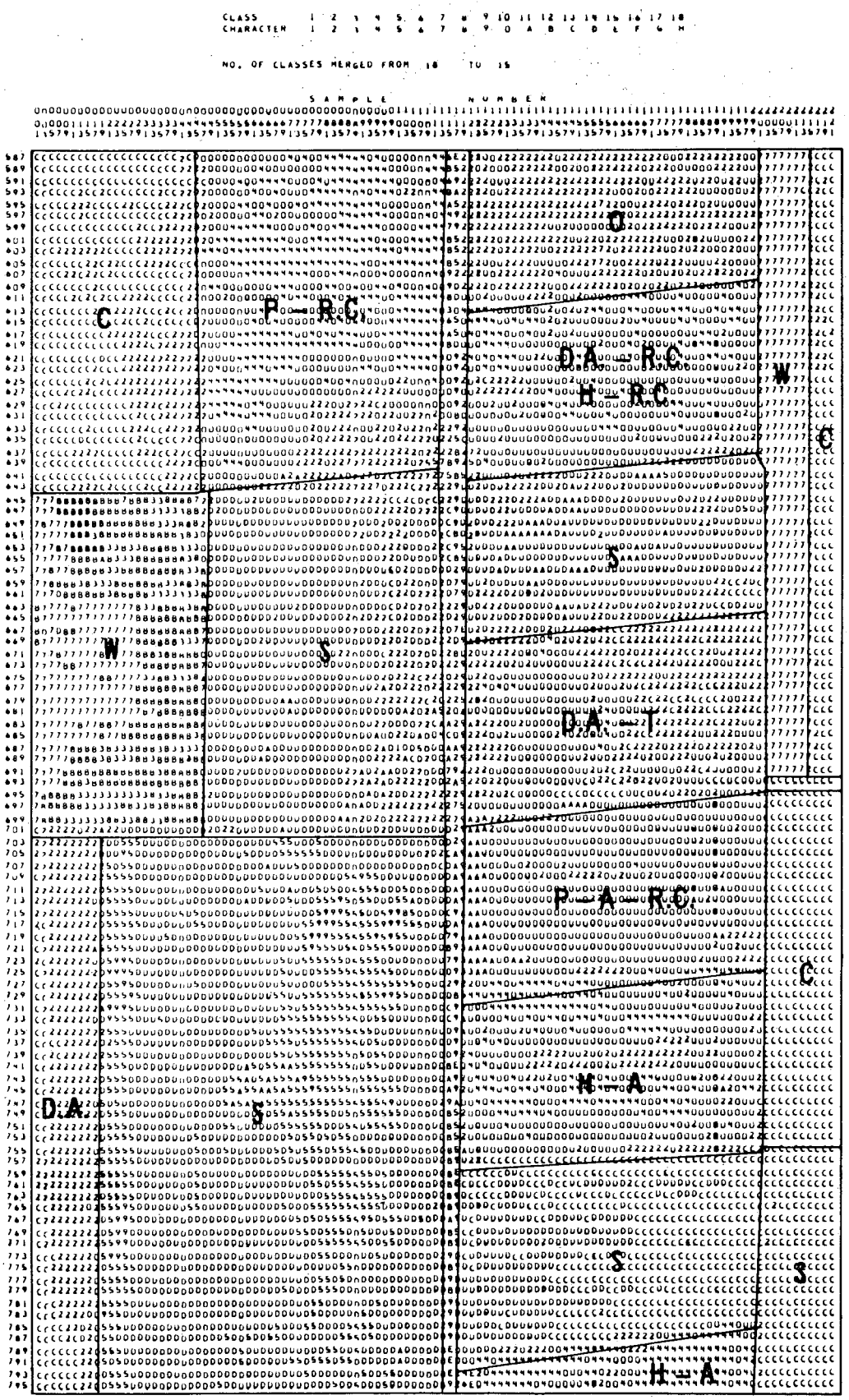


Figure 3-25. CLASSIFICATION MAP BY THE GENERALIZED K-MEANS TECHNIQUE WITH 15 CLASSES

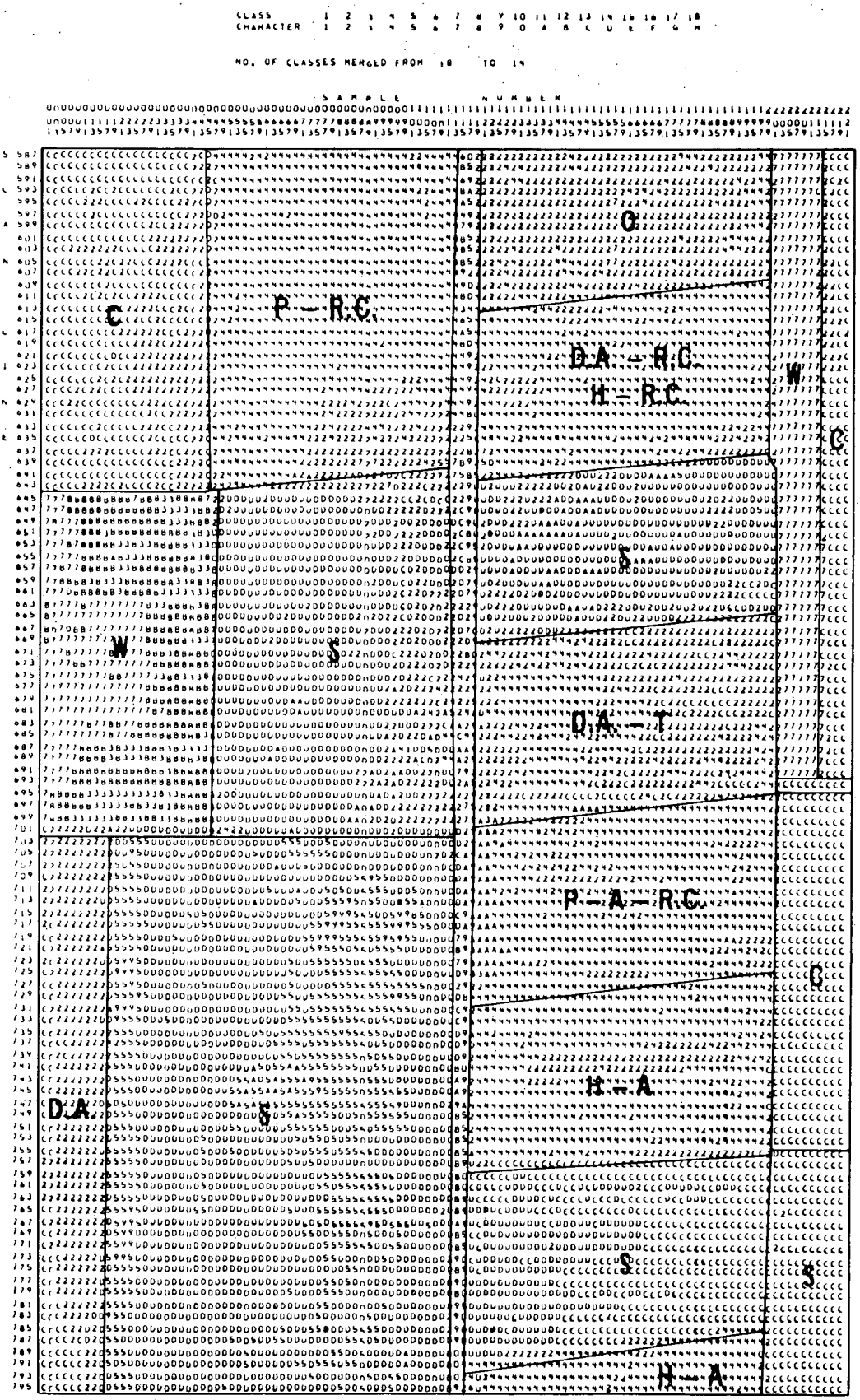


Figure 3-26. CLASSIFICATION MAP BY THE GENERALIZED K-MEANS TECHNIQUE WITH 14 CLASSES

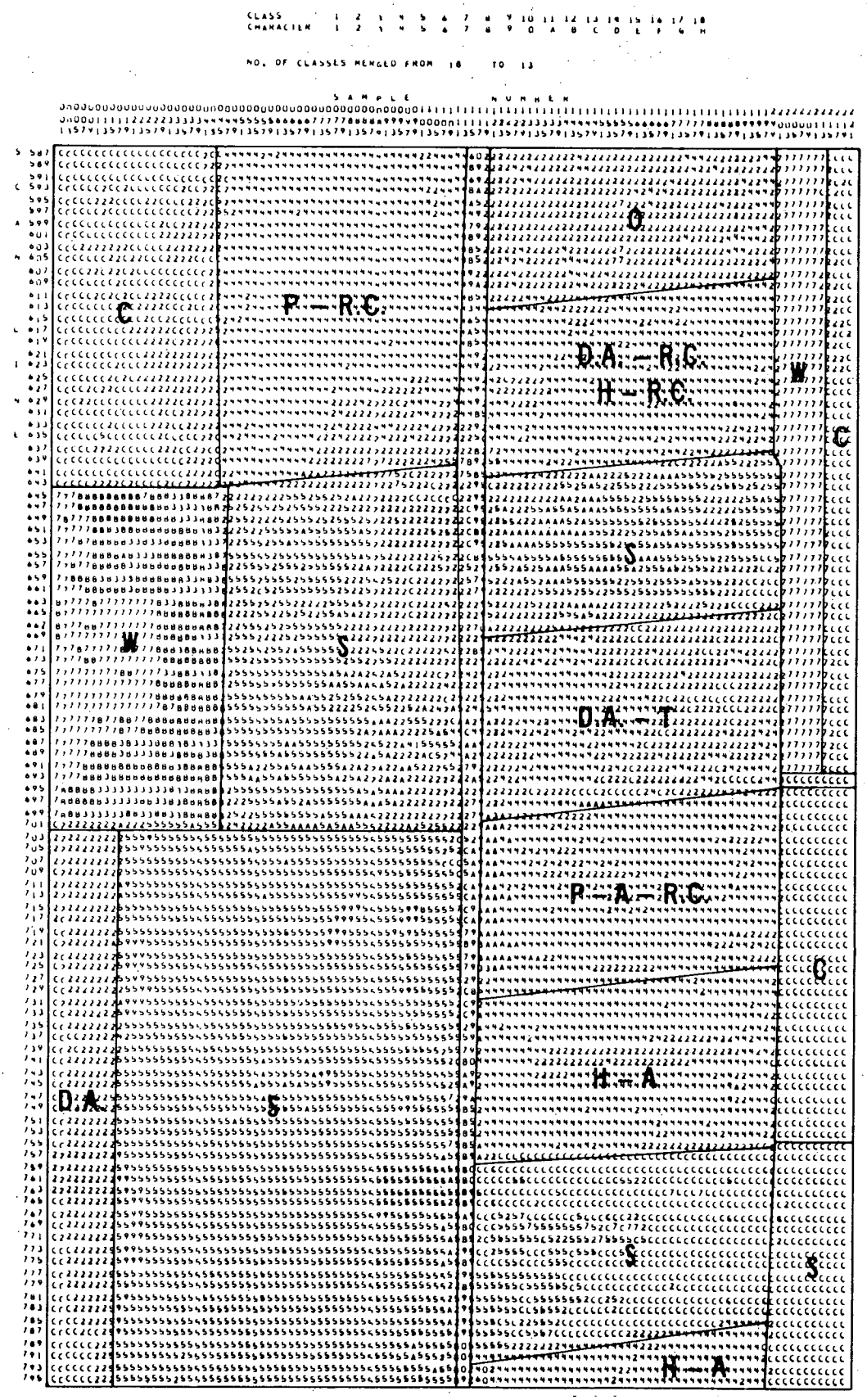


Figure 3-27. CLASSIFICATION MAP BY THE GENERALIZED K-MEANS TECHNIQUE WITH 13 CLASSES

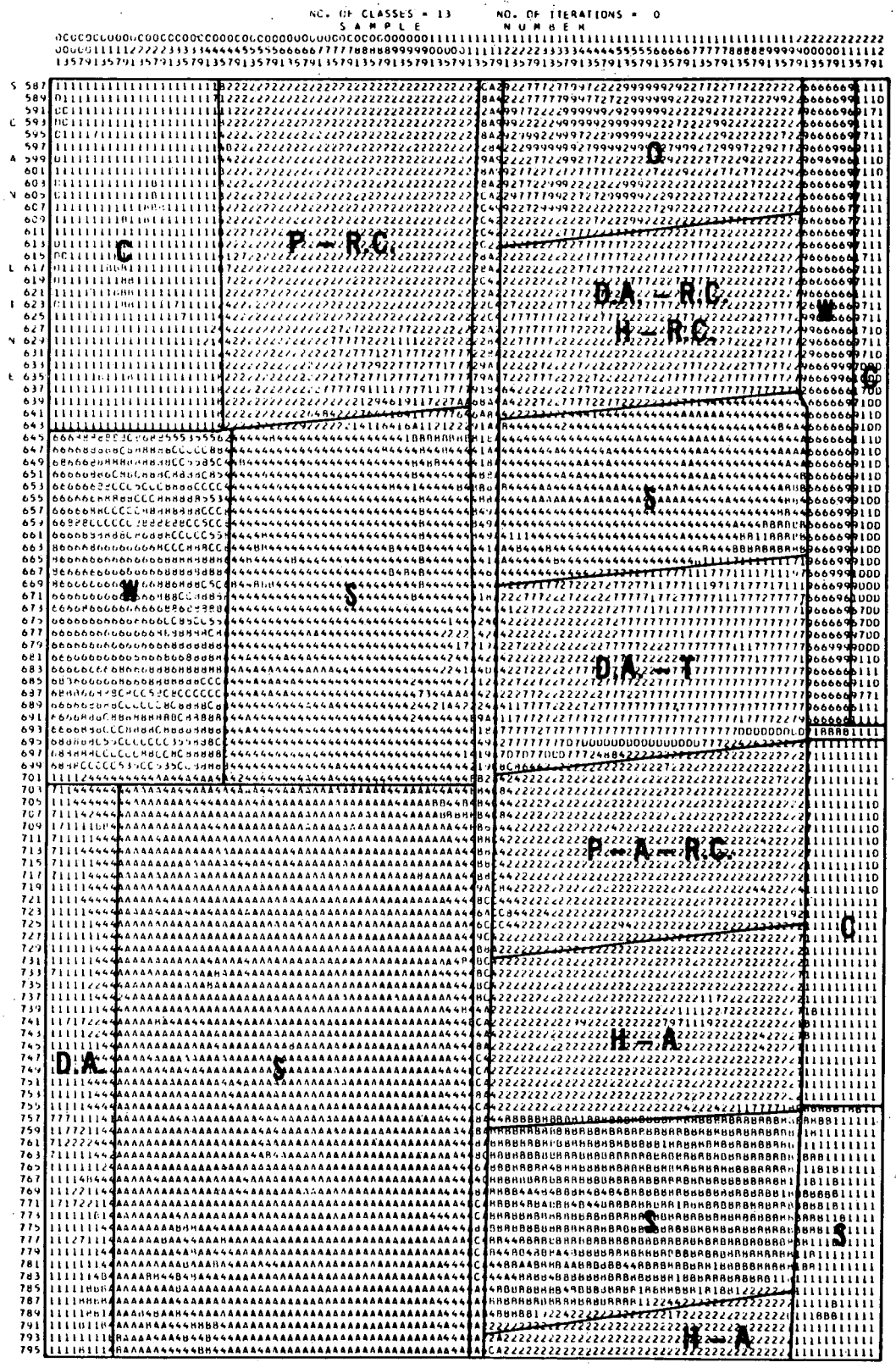


Figure 3-28. CLASSIFICATION MAP BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 13 CLASSES AND WITHOUT ITERATION

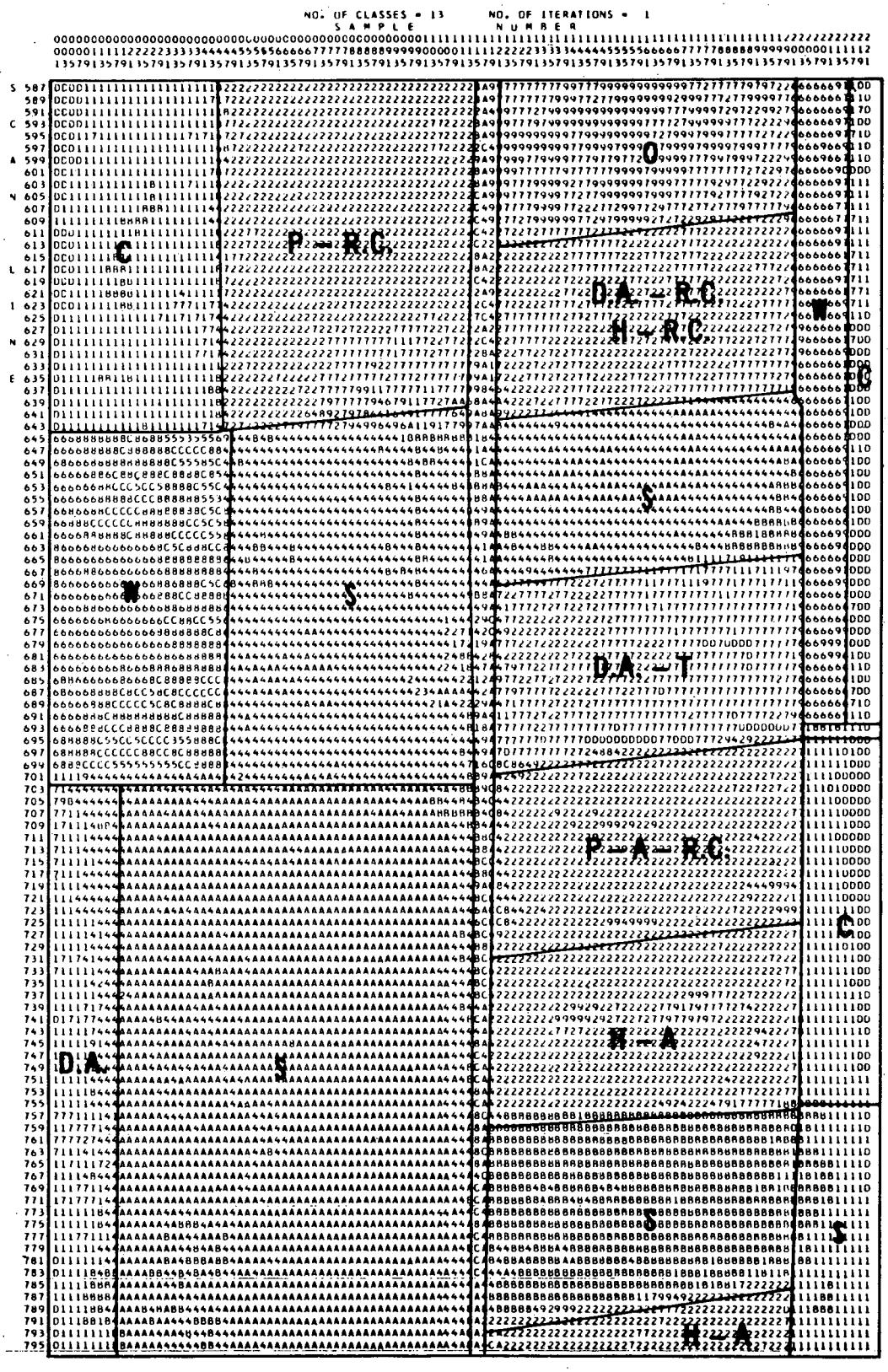


Figure 3-29. CLASSIFICATION MAP BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 13 CLASSES AFTER ONE ITERATION

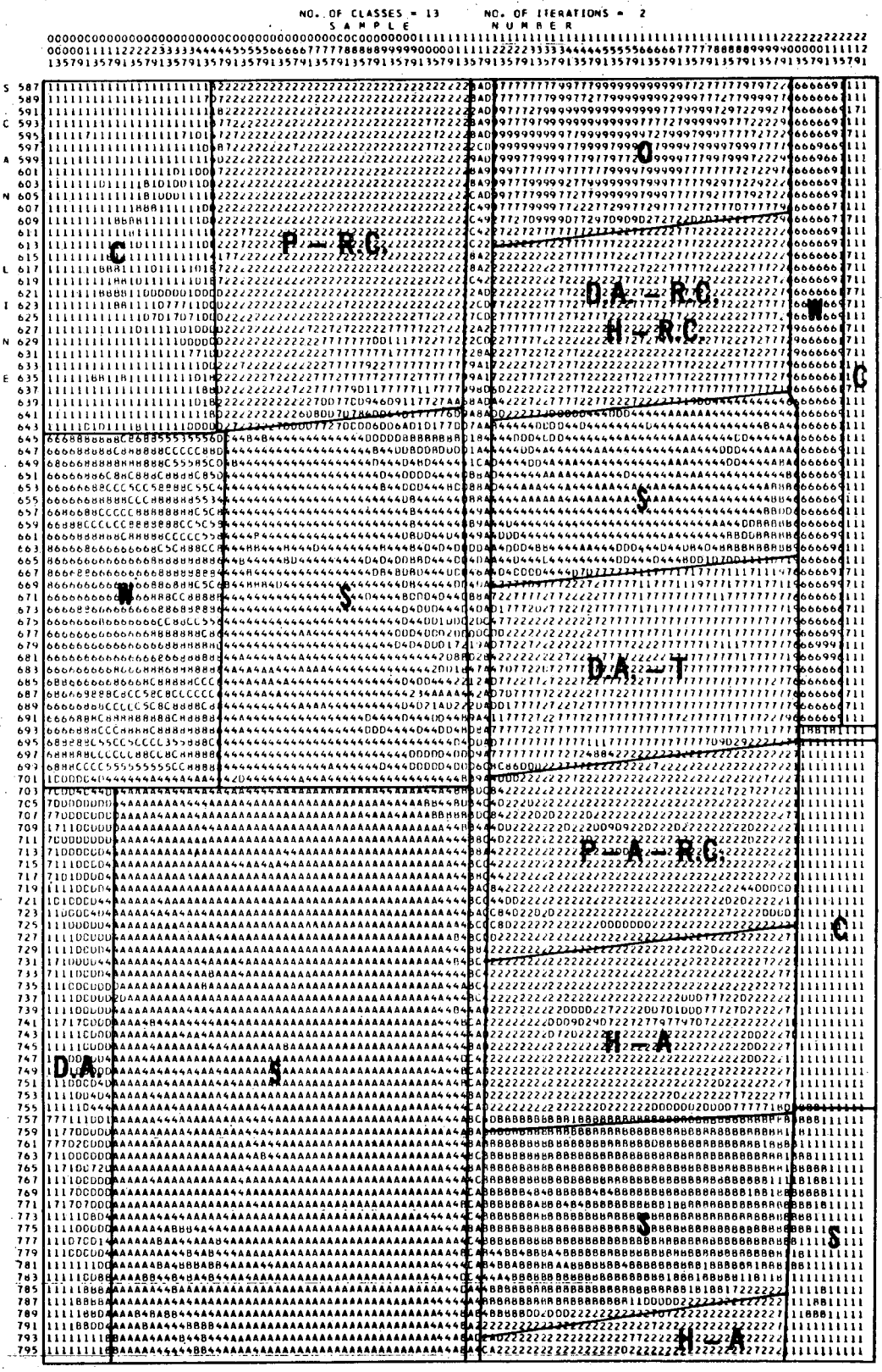


Figure 3-30. CLASSIFICATION MAP BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 13 CLASSES AFTER 2 ITERATIONS

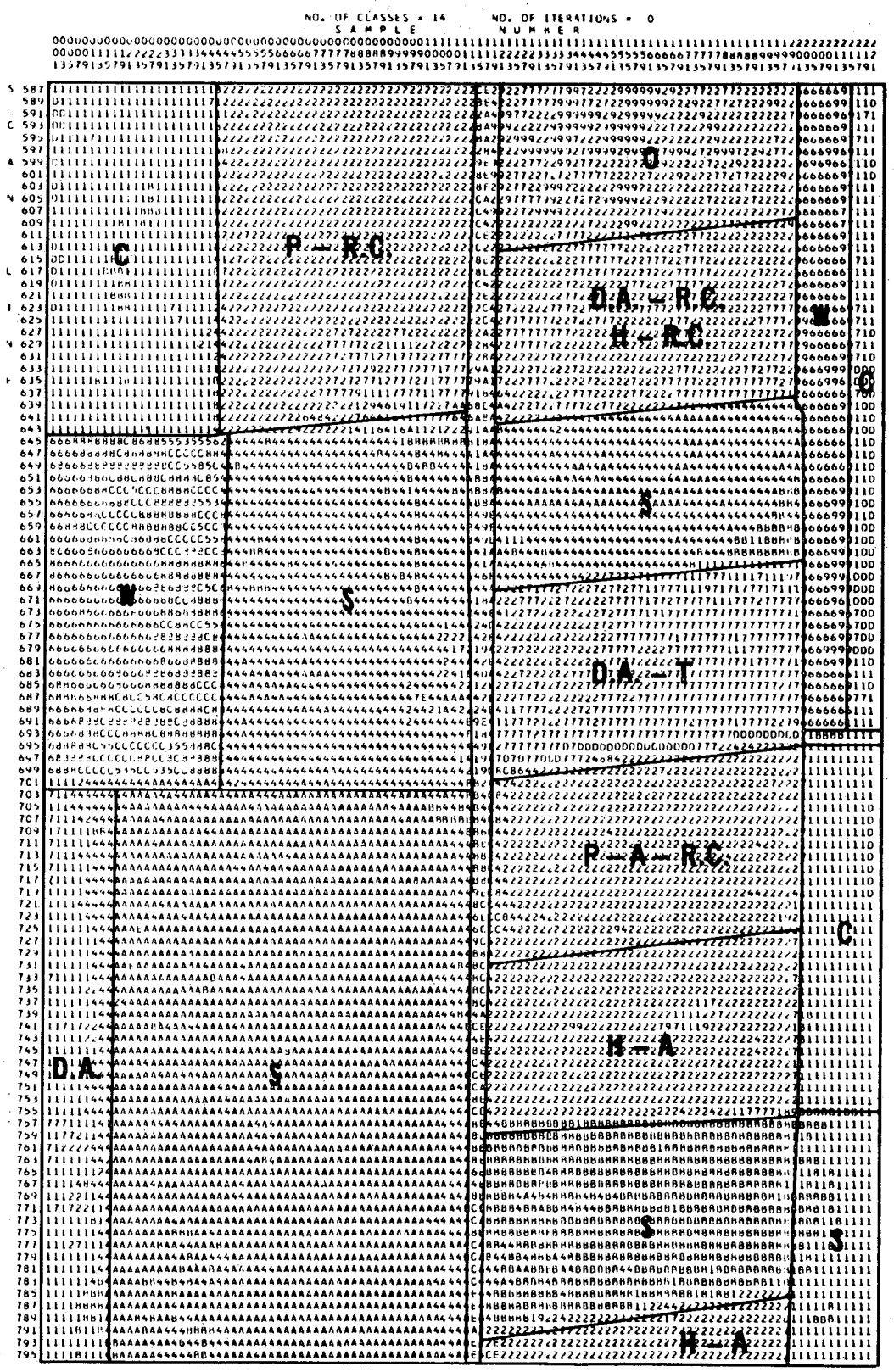


Figure 3-31. CLASSIFICATION MAP BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 14 CLASSES AND WITHOUT ITERATION

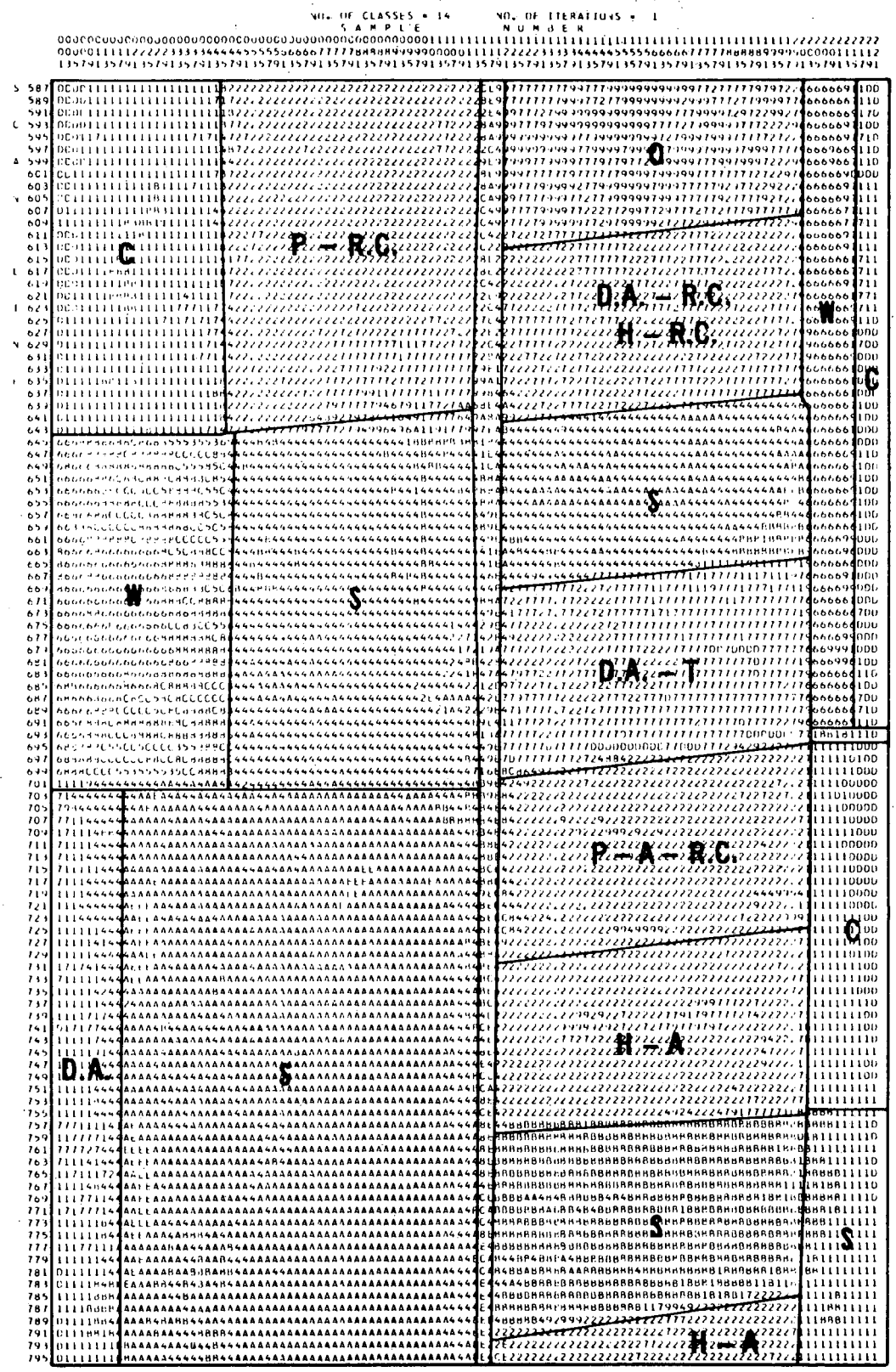


Figure 3-32. CLASSIFICATION MAP BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 14 CLASSES AFTER ONE ITERATION

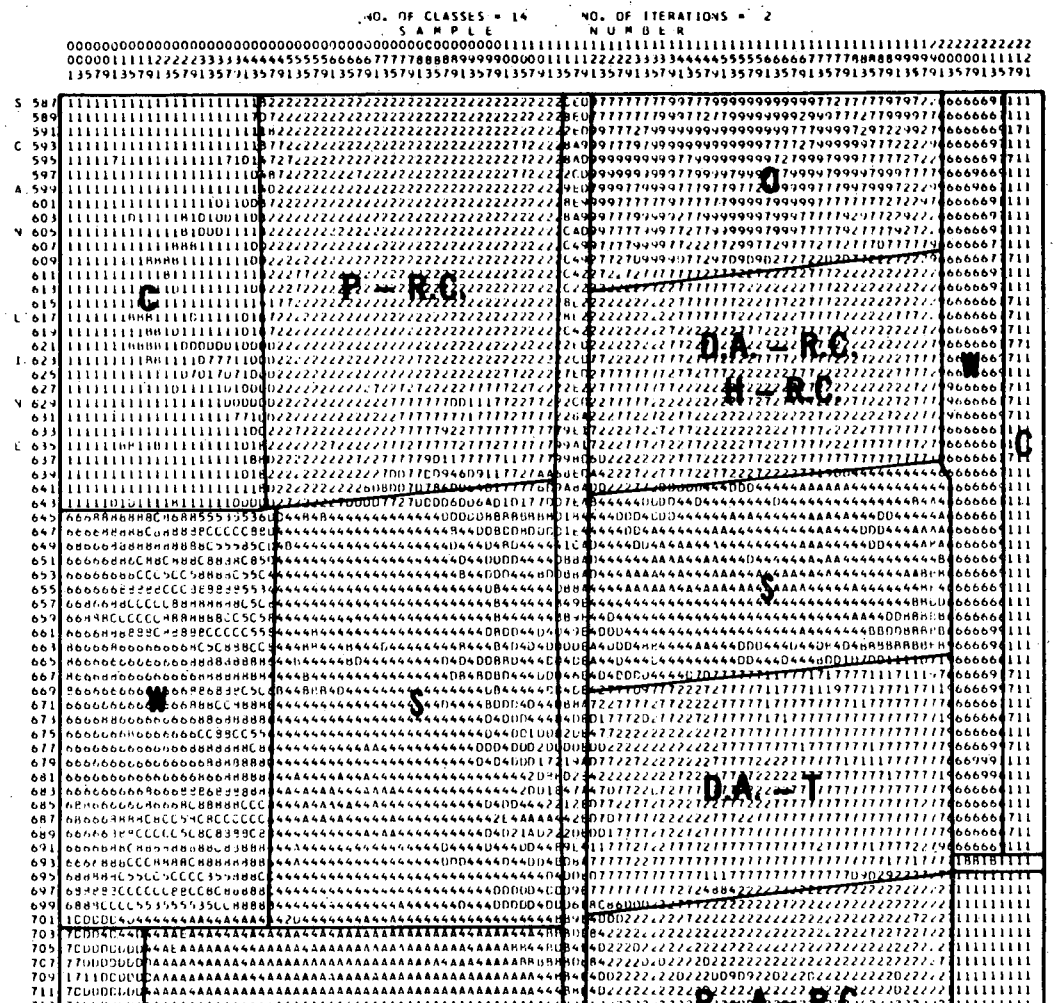


Figure 3-33. CLASSIFICATION MAP BY THE COMPOSITE CLUSTERING TECHNIQUE WITH 14 CLASSES AFTER TWO ITERATIONS

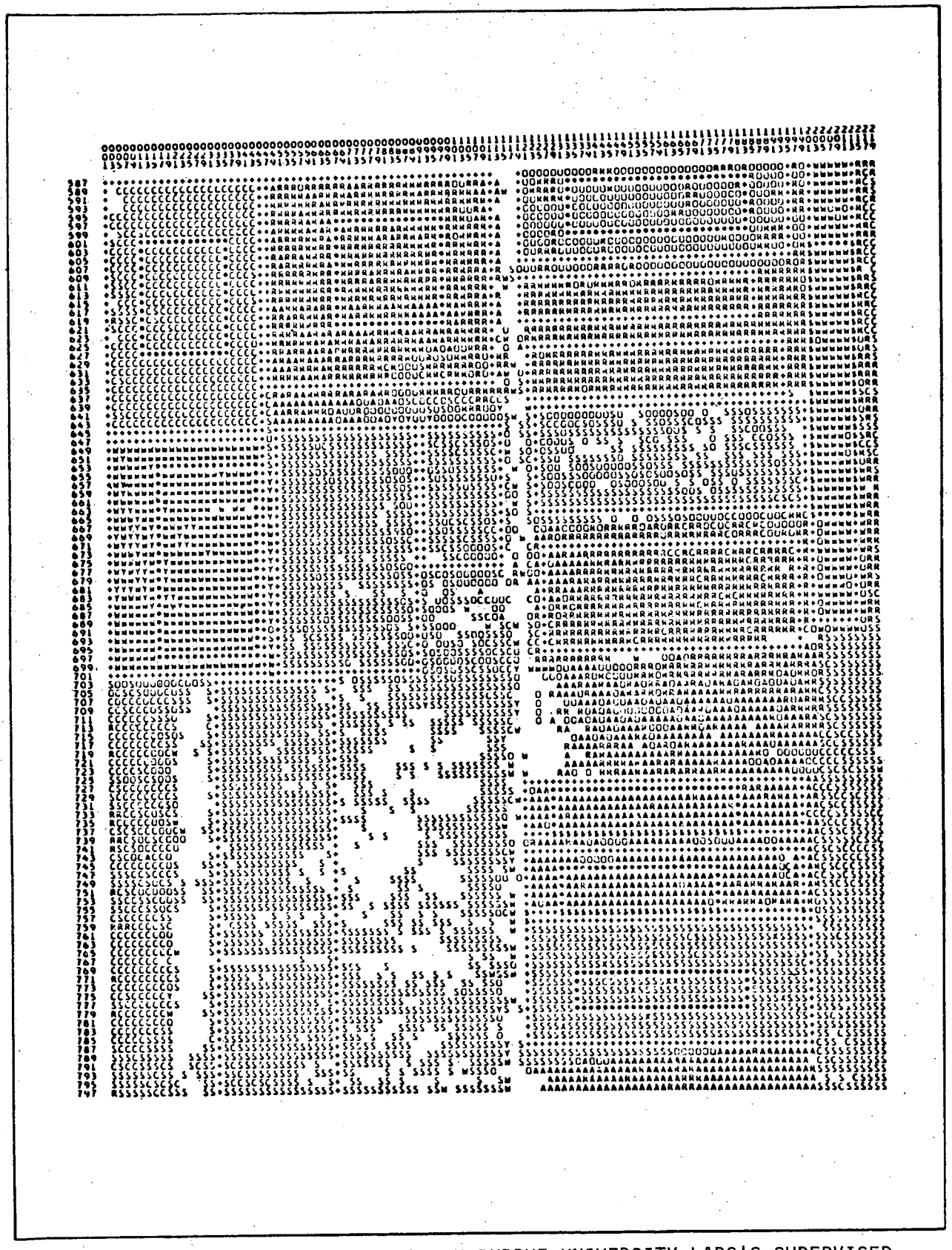


Figure 3-34. CLASSIFICATION MAP BY PURDUE UNIVERSITY LARS'S SUPERVISED BAYES CLASSIFICATION TECHNIQUE (Ref. 10, p. 40)

LABORATORY FOR AGRICULTURAL REMOTE SENSING
PURDUE UNIVERSITY

••• LARSYSAA ILLUSTRATION •••

CLASSIFICATION STUDY .. SERIAL NO. 705807300
CLASSIFICATION DATE .. JULY 5, 1968

RUN NUMBER-----26600061 DATE------ 6/28/66
FLIGHT LINE---- C1 TIME------1229
TAPE NUMBER---- 102 ALTITUDE-- 2600 FEET

CLASSES CONSIDERED

SYMBOL	CLASS	THRESHOLDS
S	SOYBN I	14.900
C	CORN I	14.900
O	OATS	14.900
W	WHEAT I	14.900
R	RD CL I	14.900
A	ALFALFA	14.900
Y	RYE	14.900
X	BR SOIL	14.900
W	WHEAT II	14.900

FEATURES CONSIDERED

CHANNEL NO.	SPECTRAL BAND
1	0.40 0.44
6	0.52 0.55
10	0.66 0.72
12	0.80 1.00

CLASSIFICATION SUMMARY BY TEST FIELDS

	CLASS	NO OF SAMPS	PCT. CORCT	NO OF SAMPLES CLASSIFIED INTO SOYB	CORN	OATS	WHEA	RED	ALFA	RYE	SOIL	THRS
7-27	SOYB	407	63.4	258	12	84	0	0	0	0	0	53
12-7	SOYB	513	88.9	456	4	31	0	0	0	0	0	22
12-2	SOYB	150	79.3	119	11	18	0	0	0	0	0	2
12-3	SOYB	752	89.2	671	8	0	0	0	0	0	0	73
7-23	SOYB	546	97.3	531	4	0	1	0	0	0	10	0
12-9	CORN	588	94.0	25	553	1	0	1	0	0	0	8
7-1	OATS	370	84.9	0	0	314	0	56	0	0	0	0
7-2	WHEA	260	93.5	0	0	17	243	0	0	0	0	0
12-10	WHEA	546	90.1	0	0	0	492	0	0	48	0	6
12-8	RED	713	80.2	2	3	27	0	572	109	0	0	0
7-29	RED	128	96.9	0	0	4	0	124	0	0	0	0
7-28	RED	175	98.9	0	0	2	0	173	0	0	0	0
	RED	385	86.8	0	17	5	0	334	24	0	0	5
7-24	ALFA	190	93.2	0	0	2	0	11	177	0	0	0
7-24	ALFA	266	83.8	0	4	19	0	18	223	0	0	2
	TOTAL	5989		2062	616	524	736	1289	533	48	10	171

OVERALL PERFORMANCE = 87.5

Figure 3-35. TABULATION OF CLASSIFICATION RESULTS OF TEST FIELDS (Ref. 10, p. 41)

Section IV

UNSUPERVISED CLASSIFICATIONS OF NATURAL TERRAIN TYPES

To give a more critical test and establish the capability of the unsupervised clustering technique, the most complex remote sensing data ever collected by the University of Michigan Multispectral scanner under NASA's sponsorship was chosen - the aircraft survey data over the Yellowstone National Park test site. These data have been kindly made available by Dr. W. H. Smedes, U. S. Geological Survey, Denver, Colorado.

4.1 DATA DESCRIPTION

These particular data were collected by the multispectral 12-channel scanner onboard an aircraft at the altitude of about 6,000 feet (ref. 7). The scanner resolution is 3 milliradians. Each scan line contains 220 ground resolution cells about 20 feet square. The multispectral scanner recorded simultaneously 12 channels of spectral bands reflecting from the earth's surface between 0.4 and 1.0 μm. These spectral bands are listed in Table 3-1. For the purpose of comparison with Purdue LARS's supervised classification results, only four channels were used, i.e., channels 2, 9, 10 and 12. These 4 channels have been determined by LARS' feature selection program to be the optional channels (based on the divergence criterion) for this particular set of data (ref. 10).

4.2 PRELIMINARY DATA ANALYSIS

Figure 4-1* shows a gray-scale video display of reflectance for channel 9 (0.62-0.66 μm) over the area. Also shown in the figure is the ground truth survey. Containing water, bedrock, forest, kame, till, talus and cloud shadow over forest. Detailed physical descriptions of these terrain types are given in reference 10. It is clear that the terrain feature is very complex, and that many parts of the test site do not have clear-cut boundaries between

* *Figures 4-1 through 4-26 and Tables 4-1 and 4-2 are presented following the text at the end of this section.*

different terrain types. This is quite different from the Purdue C-1 Flight Line in which the boundaries between different crop types are very clear (see Section III).

Figures 4-2 through 4-5 show the univariate probability histograms for these four channels for the data set from scan 200 through 500. Very few distinct modes show in each histogram, which indicates that the spectral signatures from different terrain types overlap each other, and that more than one channel would be needed for discrimination among different terrain types.

Figures 4-6 and 4-7 show the corresponding digital gray-scale plots of the test area in channels 2 and 10, respectively. Comparing these gray-scale plots with the gray-level video display, one can see clear correspondences for several main areas with large contrast. It should be noted that the complement of the numerical value with respect to 256 is proportional to the spectral radiance collected by the scanner. Hence, the larger the numerical number as indicated by the interval, the smaller the spectral radiance. Figures 4-8 through 4-13 show the scatter plots between channels 2, 9, 10 and 12. The numeric 1 indicates the number of samples in each spectral cell to be between 1 and 9; numeric 2 is between 10 and 19 and so forth. From these scatter plots, one can note that channels 2, 9 and 10 are linearly correlated, while channel 12 is not correlated with the other three channels. This implies that the terrain types possess quite different reflectance characteristics in the visible and reflected IR ranges. It is also noted that no distinct cluster is visible in these scatter plots, which, in turn, indicates the overlapping of spectral signatures of different terrain types as observed from the invariate probability histograms.

Figure 4-14 shows a boundary map of the test area obtained by using the boundary enhancement principle (ref. 8). In this map, the symbol (·) indicates that the enhanced spectral difference among adjacent resolution elements lies between the mean enhanced value over the entire target area plus one standard deviation and the mean plus two standard deviations. The symbol (+)

indicates that the enhanced difference lies between the mean plus two standard deviations and the mean plus three standard deviations. The symbol (x) indicates that the enhanced difference is greater than the mean plus three standard deviations. Finally, the area with the enhanced difference smaller than the mean plus one standard deviation is left blank. In other words, the blank area implies a relatively homogenous region, while the area indicated by the symbol (x) has the largest spectral contract between adjacent resolution elements. These boundaries are found to be in good correspondence with the gray-level video display in Figure 4-1.

4.3 UNSUPERVISED CLASSIFICATION OF TERRAIN TYPES

Figures 4-15 through 4-24 show the intermediate and final unsupervised classification maps of the test area by the composite statistical and K-mean technique. The purpose of presenting the intermediate results is to show how the composite technique performs at its various stages so that some types of automatic decision logic may be formulated and built into the present computer program to achieve a more autonomous unsupervised classification scheme.

For processing the set of data, a maximum of 18 classes was initially specified for the statistical sequential clustering. The output from this processing after only one pass of the entire data set is a set of mean spectral signatures for 18 initial classes. (Note: If the K-mean clustering had been used, 17 passes of the entire data set would have been required to estimate the 18 initial cluster centers. Further, these initial cluster centers would not be as accurate as those obtained by the statistical sequential technique). The choice of a maximum of 18 classes for the data was based on a rough examination of the video display of the test area (Figure 4-1), and 18 classes were believed to be sufficient. Actually, the number is twice as large as the main terrain types indicated by the ground truth survey map (Figure 4-25) supplied by Dr. W. H. Smedes, U. S. Geological Survey. The study is presently underway on how to decide on a suitable number of initial classes for any given data set. This study will be presented in the final contract report.

The above mean spectral signatures of the 18 initial classes were input to the K-mean clustering program for further improvement. Figures 4-15 and 4-16 show the classification maps after one and two iterations, respectively, of these initial classes. A comparison of these two maps shows that the majority of class 7 is grouped into class M in the second iteration. Otherwise, no noticeable change has occurred in the second iteration. From the ground truth map, one finds that classes M and 7 both belong to the same class - forest. The very few changes between the first and second iteration classification results indicates that the cluster centers have very rapidly converged to their true locations in the color space. In turn, this may imply that the initial cluster centers obtained by the statistical sequential clustering using only one pass of the data are quite good indeed. Hence, by only three passes of the data sets, i.e., one for the statistical sequential clustering and two for the K-mean clustering, good classification of the data set has been accomplished. By the K-mean clustering, more than 20 passes of the data set would have been required and the clustering results would not be as accurate as those obtained by the composite technique.

The ground truth map (Figure 4-25) does not give a resolution element-by-resolution element terrain type specification. Instead, it shows only the average percentage descriptions of terrain types. For example, one area at the upper left-hand corner shows 80 percent rubble and 30 percent forest (i.e., .7 R, .3 F). Thus, it is not possible to make an exact assessment of the classification accuracy. Furthermore, the ground truth map gives only nine terrain types. For the easier comparisons, the 17 classes resulting from the K-mean program were further reduced one class at a time to nine classes. The criterion used for merging classes is the simple Enclidean distance in the color space. In other words, first the pairwise distances of all the 17 classes are calculated, and then the two classes which have the shortest distance among all possible pairs are combined. This process is repeated on the resulting 16 classes and so on. The classification maps for each of these merging processes are shown in Figures 4-17 through 4-24. The actual merging processes are summarized in Table 4-1. Two meeting arrows denotes the merging

of two classes at that particular stage of merging with the new symbol for the merged class given above the arrows. The single arrow denotes the change of class symbol, for example, from class 2 to class = at the 3rd stage of merging. The latter change of class symbol is due to the computer program coding and is of no significance.

The physical identities of each of the nine classes are determined for this case by comparing the unsupervised classification map (Figure 4-24) with the ground truth survey map (Figure 4-25). The result is shown also in Table 4-1. In actual operation, the physical identities will be established by checking a small percentage of each class on site. As mentioned earlier, it is not possible for this set of data to make an exact assessment of classification accuracy. The overall accuracy is about 80 percent. This comparison was made by Dr. W. H. Smedes who has the detailed knowledge on this test site (ref. 4). The main misclassification came from mingling two terrain types - water and talus even prior to merging classes. The mean spectral signatures of water and talus are given below, as obtained from small areas in the test site,

	Ch-2	Ch-9	Ch-10	Ch-12
Water	85.3	84.2	81.7	67.2
Talus	77.3	75.5	82.1	50.1

which are quite similar to each other for comparing with the mean spectral signatures of the other 16 classes before merging of classes (Table 4-2).

4.4 COMPARISON WITH SUPERVISED CLASSIFICATION

For a better appraisal of the performance of the composite clustering technique, the unsupervised classification map (Figure 4-24) is compared with the supervised classification map obtained by Purdue University's LARS using the maximum likelihood method over the same test area (refs. 4 and 10). The supervised classification map is shown in Figure 4-26. The accuracy of the classification is found to be about 86 percent as also reported by Dr. Smedes. This accuracy is higher than the 80 percent by the composite clustering technique. However, to obtain this higher accuracy, much human intervention and

manipulation were needed by (a) knowing where to pick the typical training areas for every type of terrain of interest, (b) classifying the entire set of data and calculating the classification accuracy, and (c) new training areas were selected when the accuracy was found to not be good enough. In contrast to this iterative processing with close human supervision, the unsupervised classification map (Figure 4-14 or Figure 4-24) were obtained with very little human intervention, only specifying the maximum number of initial classes to begin with the processing. The computation time required for both the LARS supervised and unsupervised composite classification methods are about the same.

Figure 4-1. GRAY-SCALE VIDEO DISPLAY OF REFLECTANCE FOR CHANNEL 9

This page is reproduced again at the back of this report by a different reproduction method so as to furnish the best possible detail to the user.

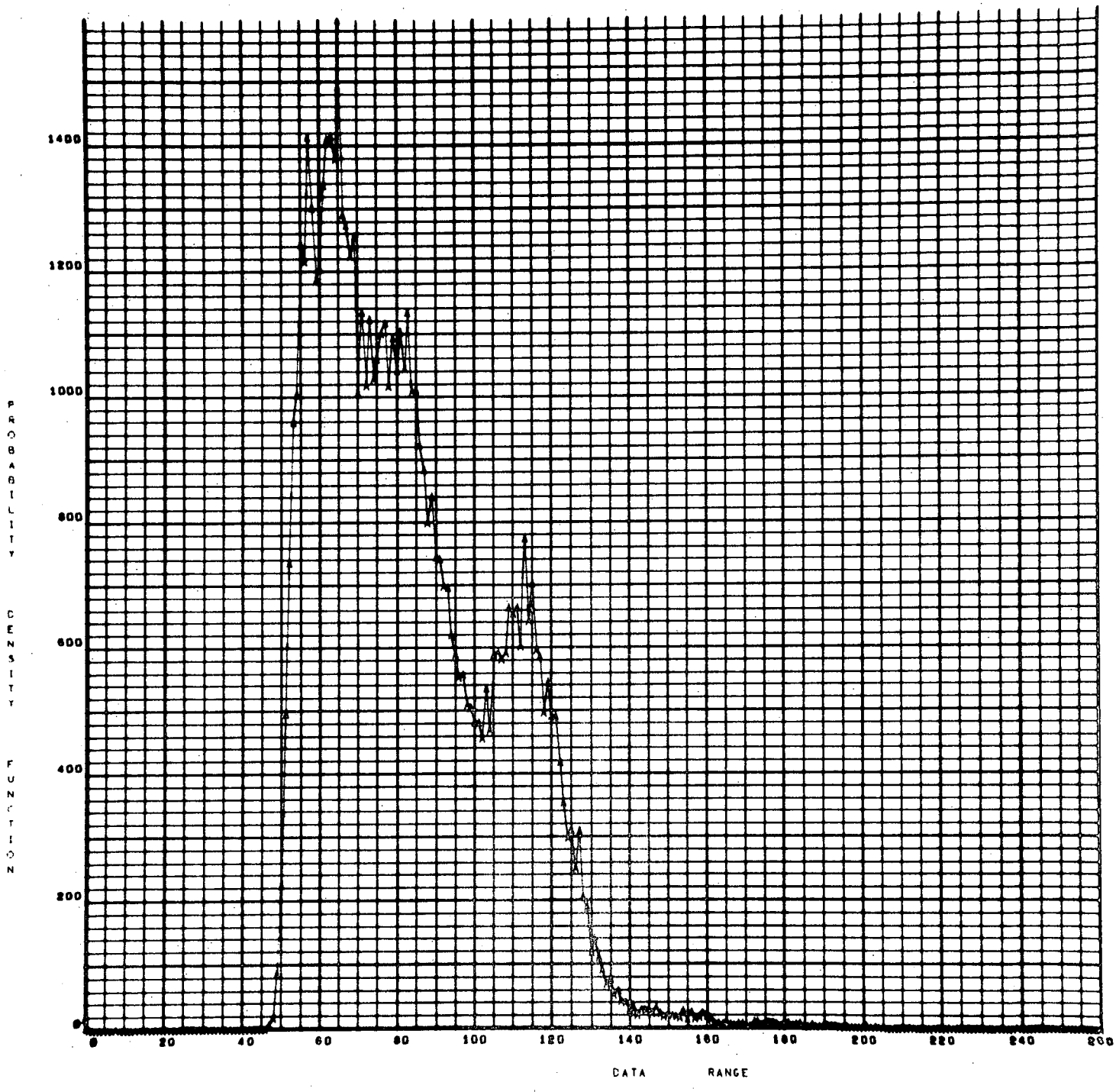


Figure 4-2. PROBABILITY HISTOGRAM OF CHANNEL 2

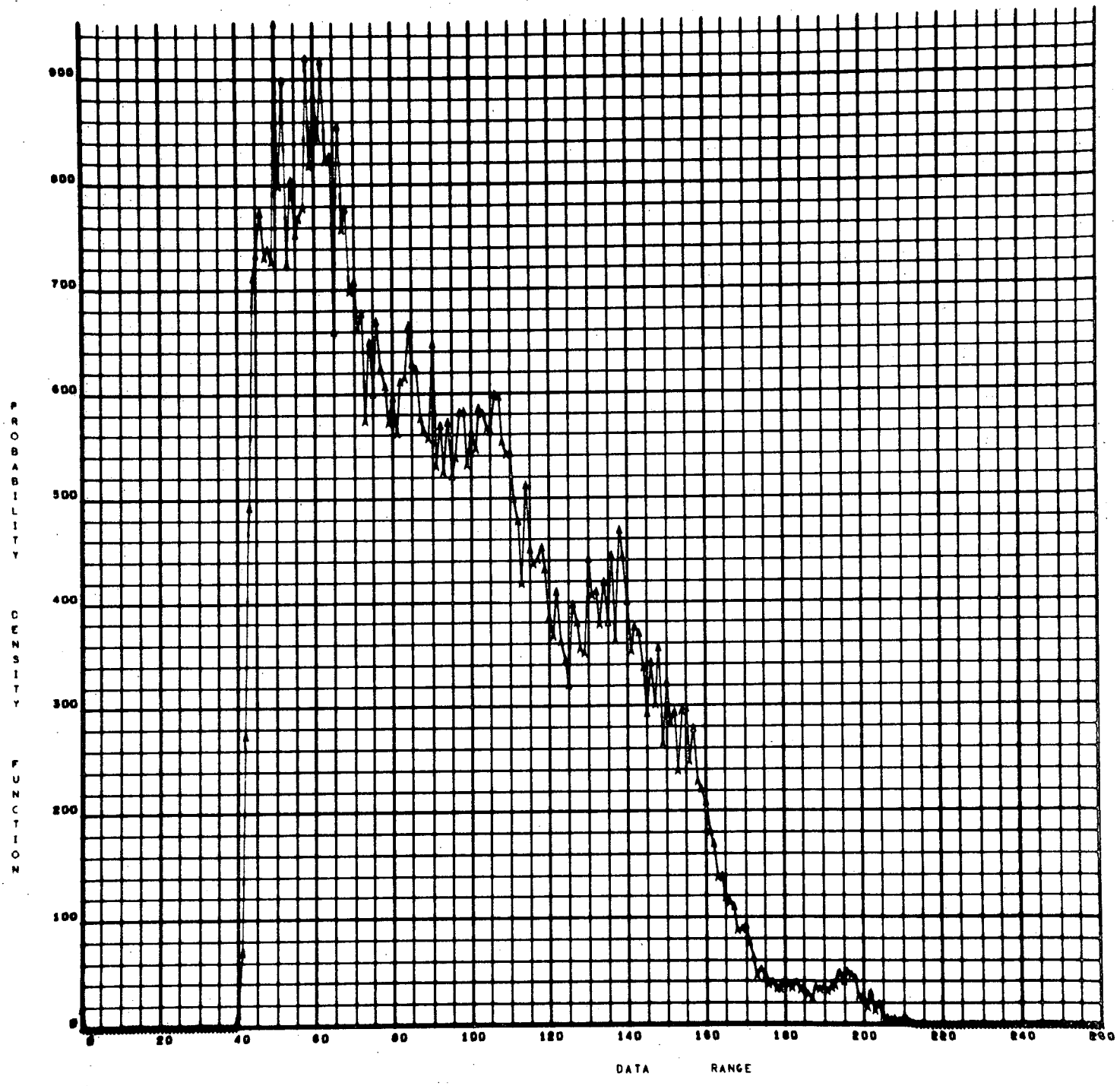


Figure 4-3. PROBABILITY HISTOGRAM OF CHANNEL 9

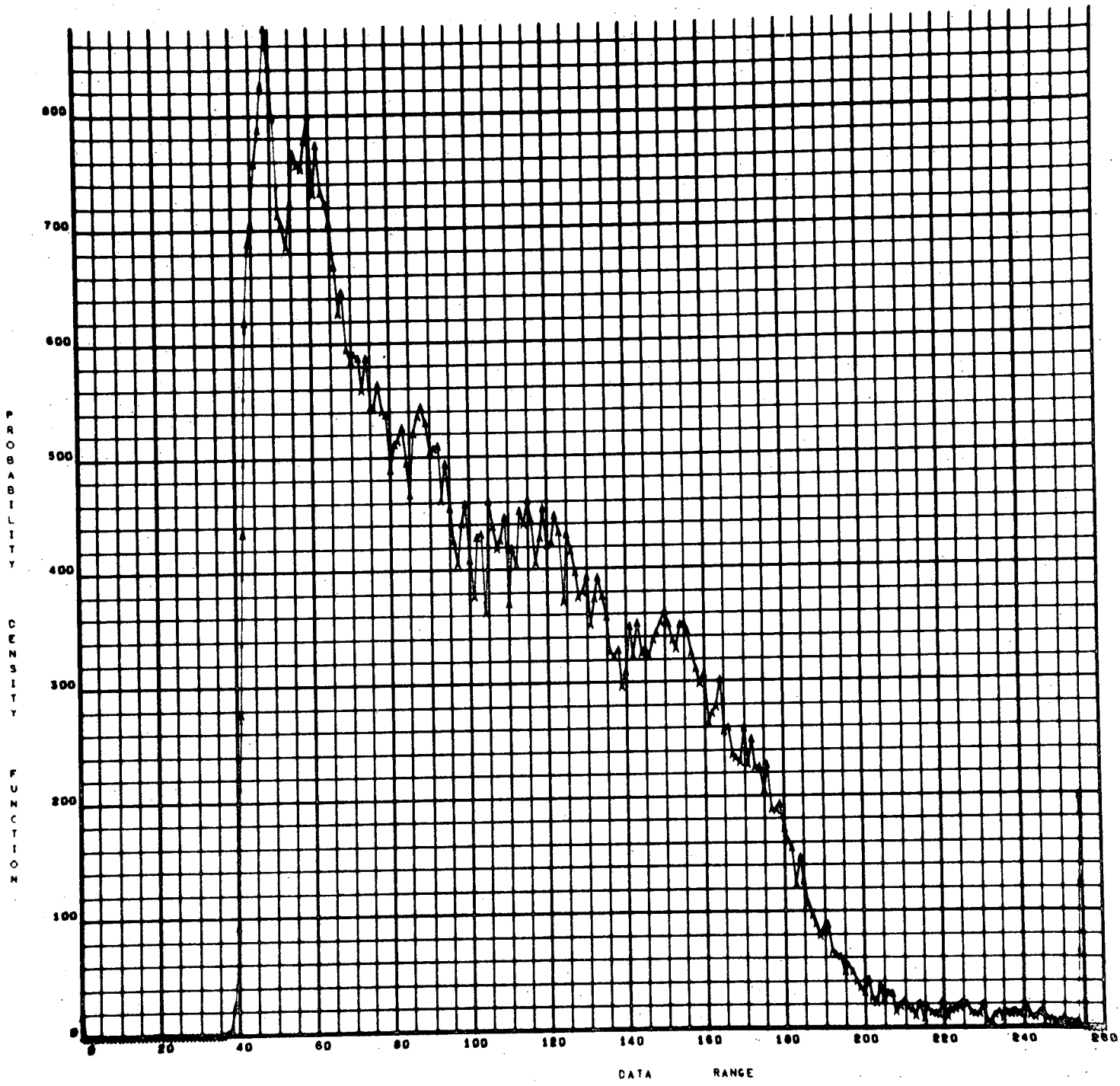


Figure 4-4. PROBABILITY HISTOGRAM OF CHANNEL 10

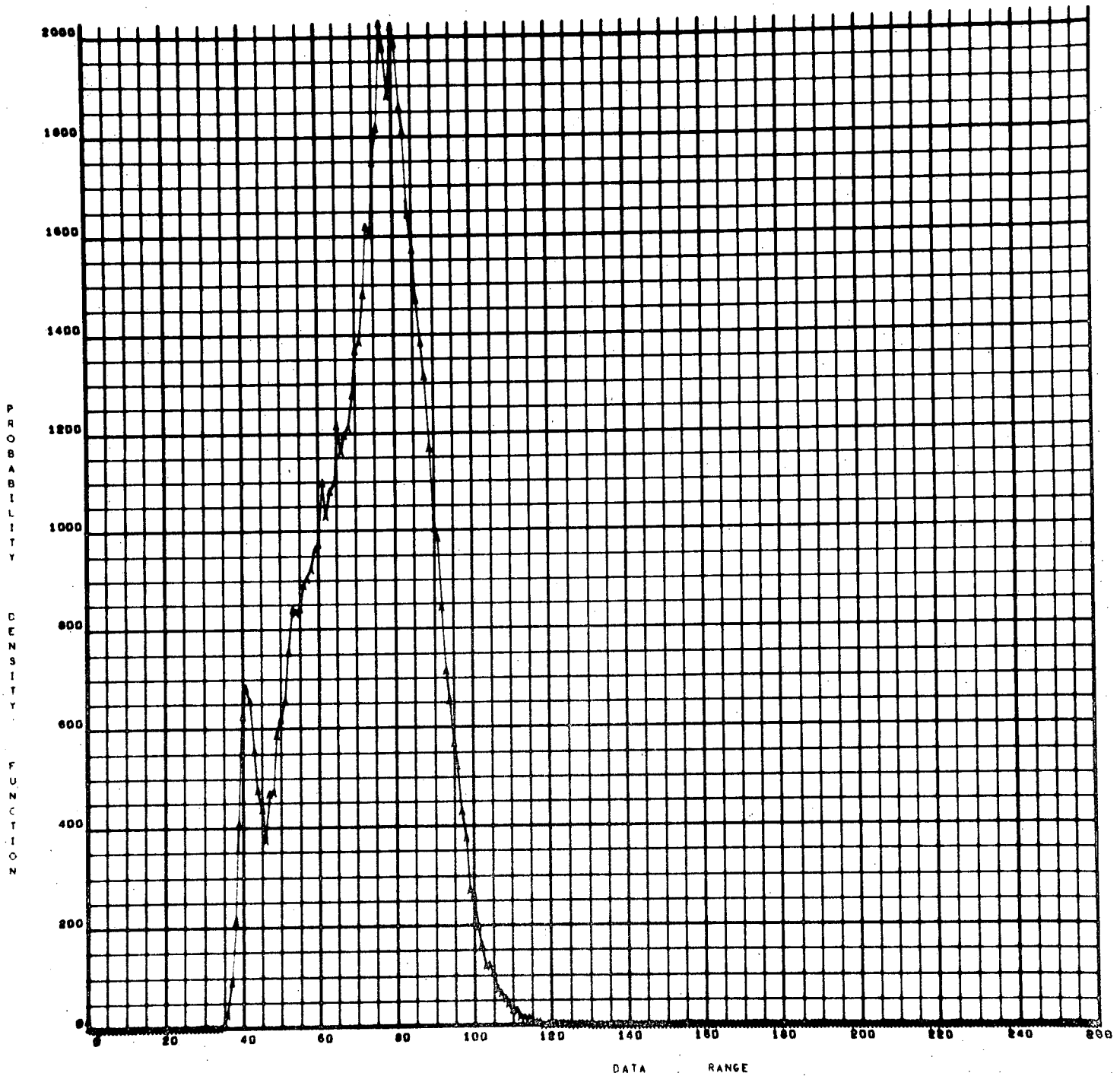


Figure 4-5. PROBABILITY HISTOGRAM OF CHANNEL 12

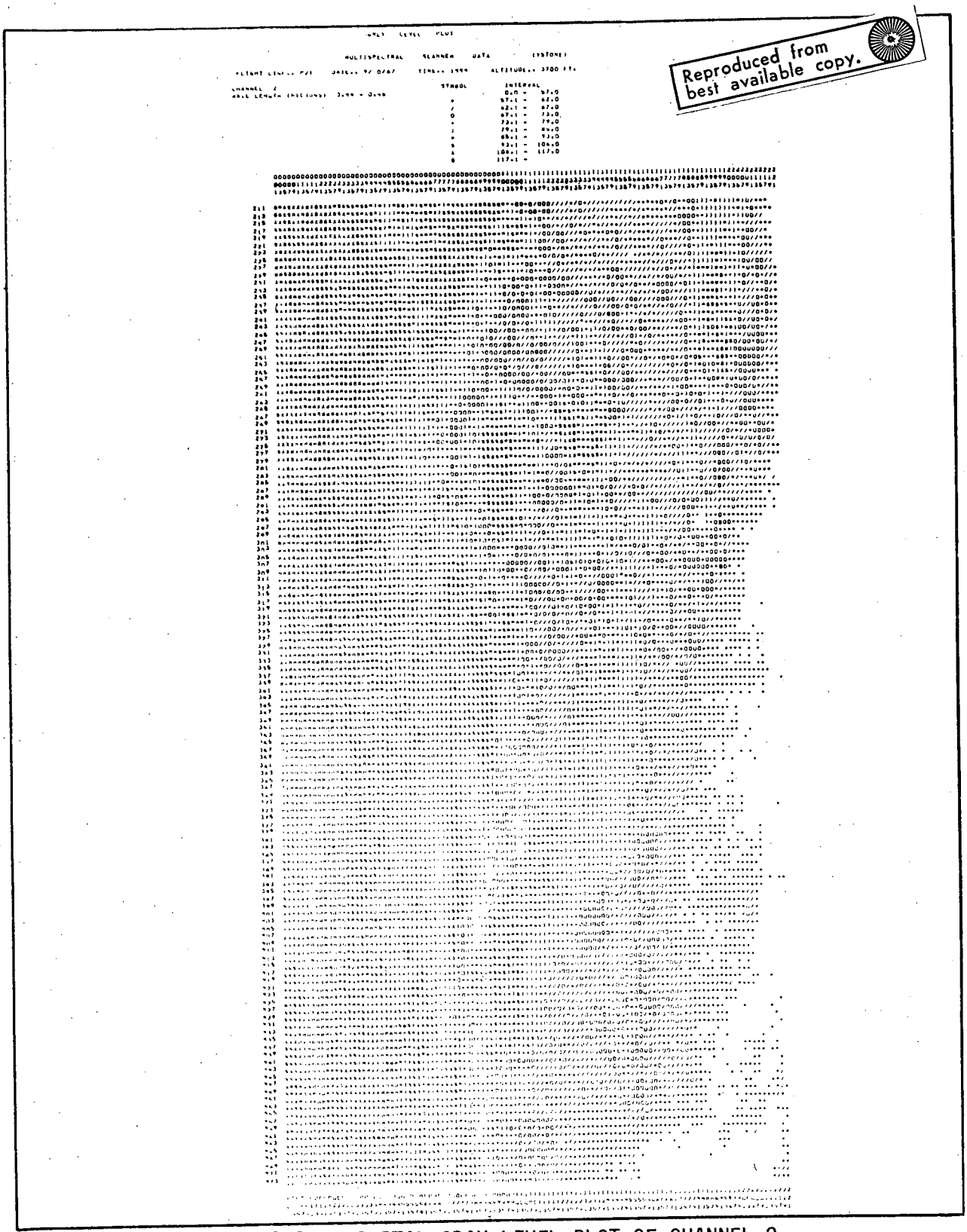


Figure 4-6. DIGITAL GRAY-LEVEL PLOT OF CHANNEL 2

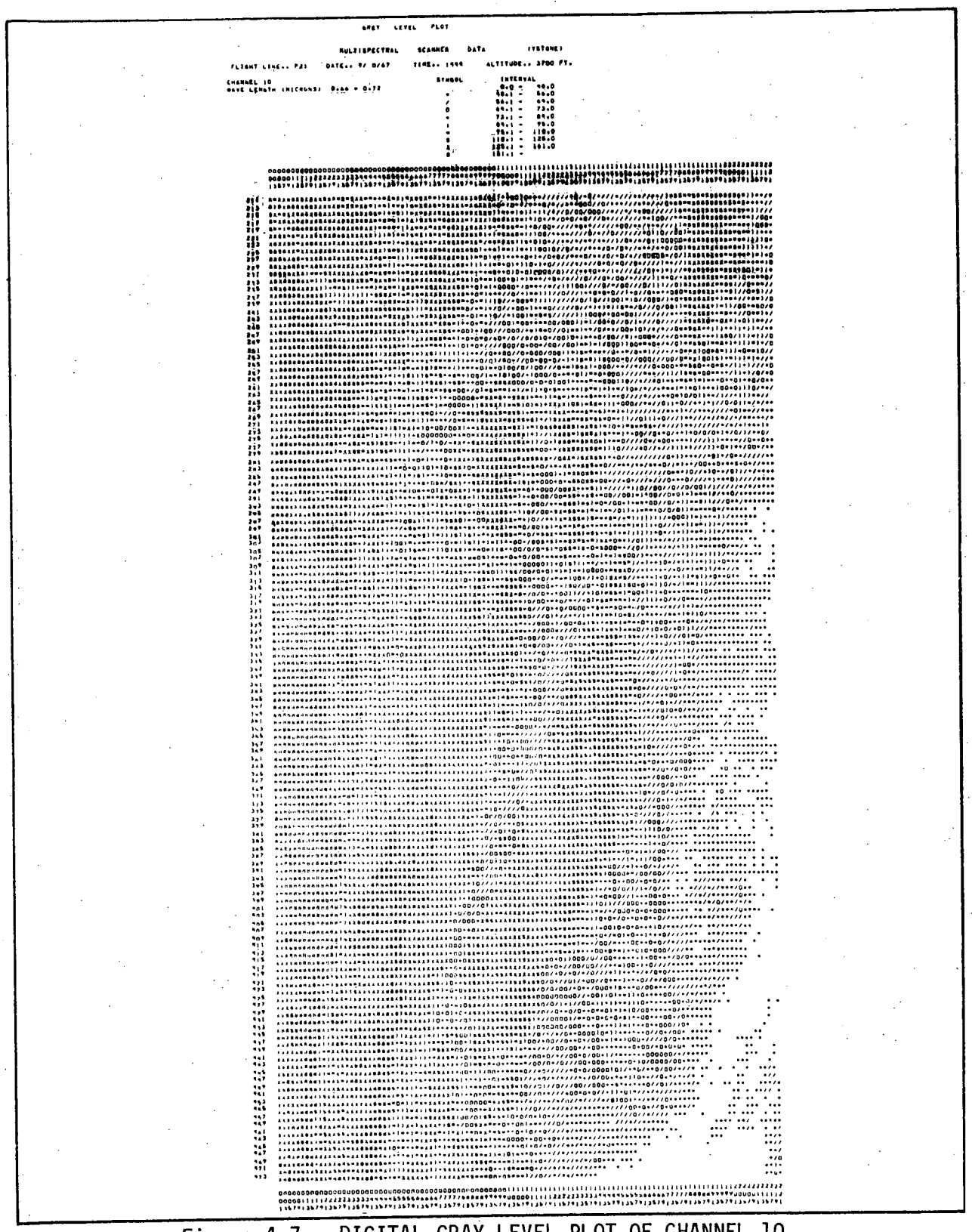


Figure 4-7. DIGITAL GRAY-LEVEL PLOT OF CHANNEL 10

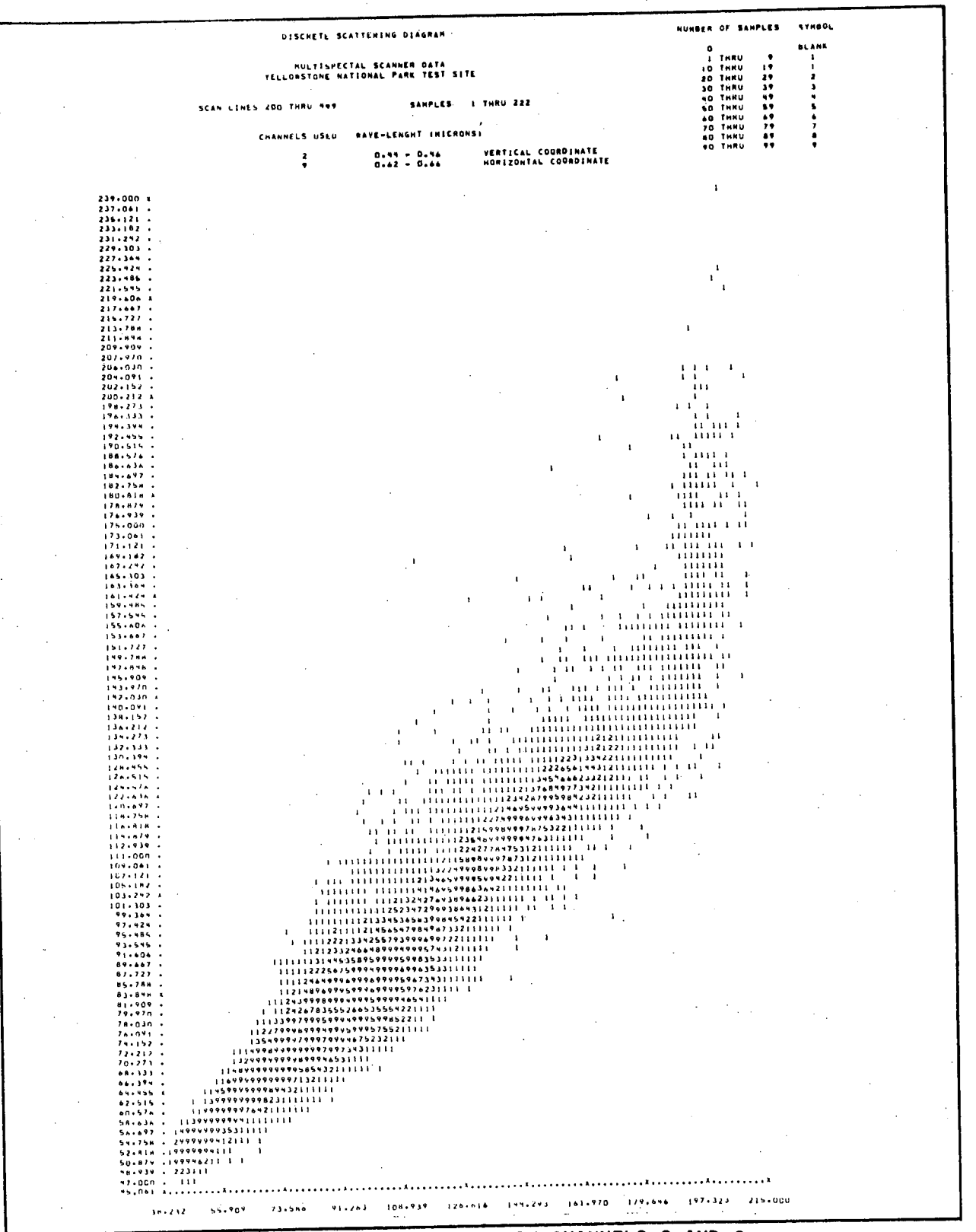


Figure 4-8. SCATTER PLOT OF CHANNELS 2 AND 9

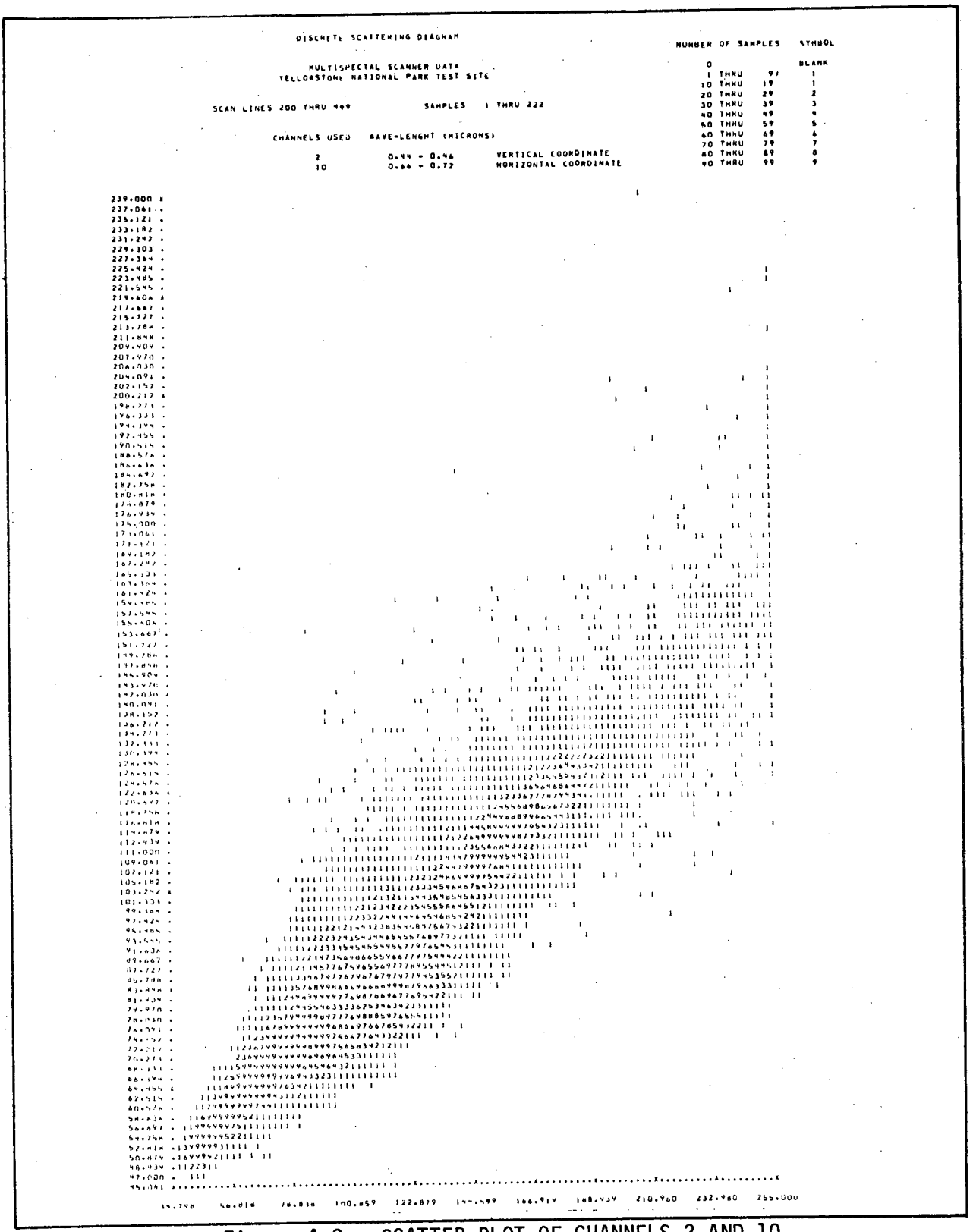


Figure 4-9. SCATTER PLOT OF CHANNELS 2 AND 10

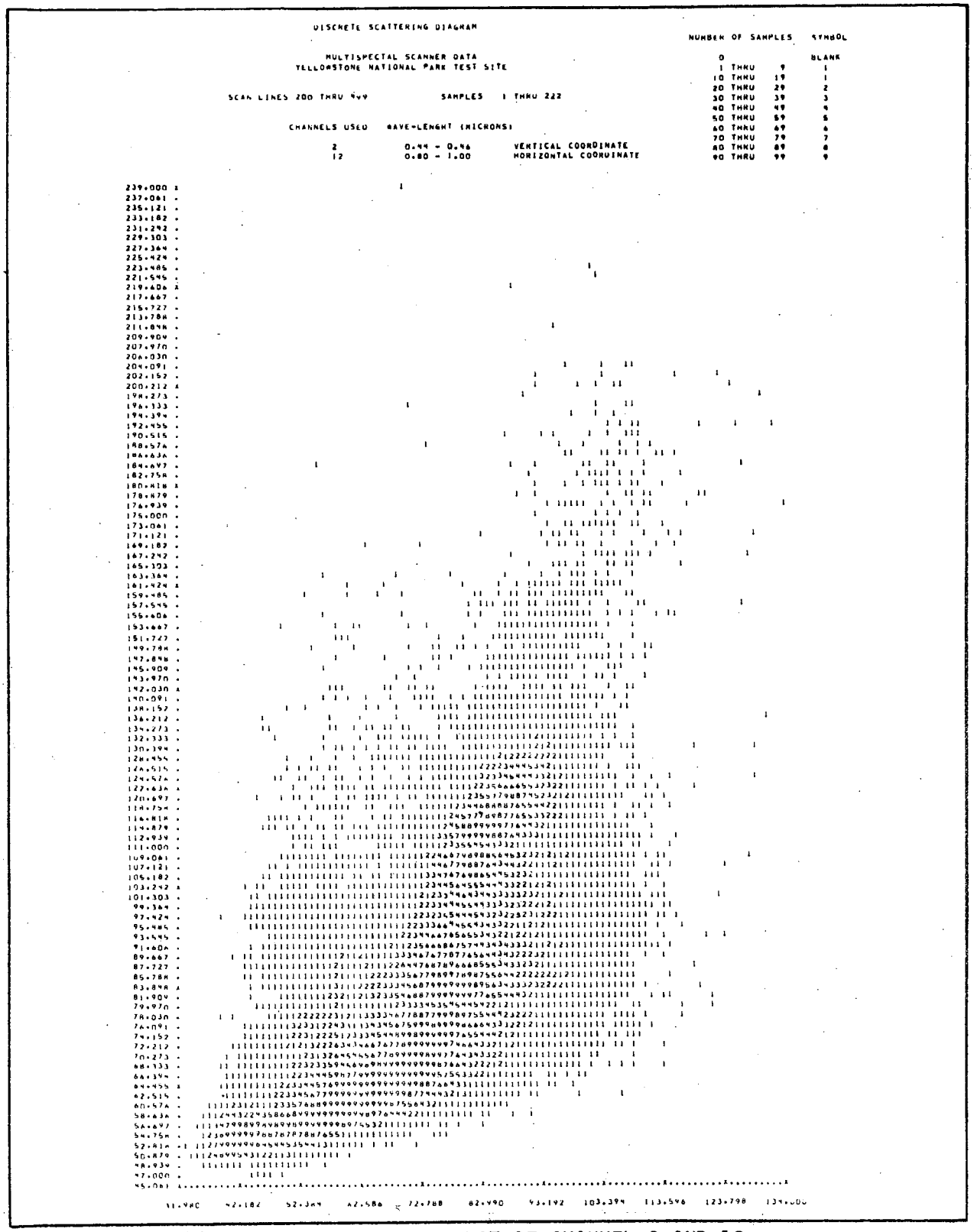


Figure 4-10. SCATTER PLOT OF CHANNEL 2 AND 12

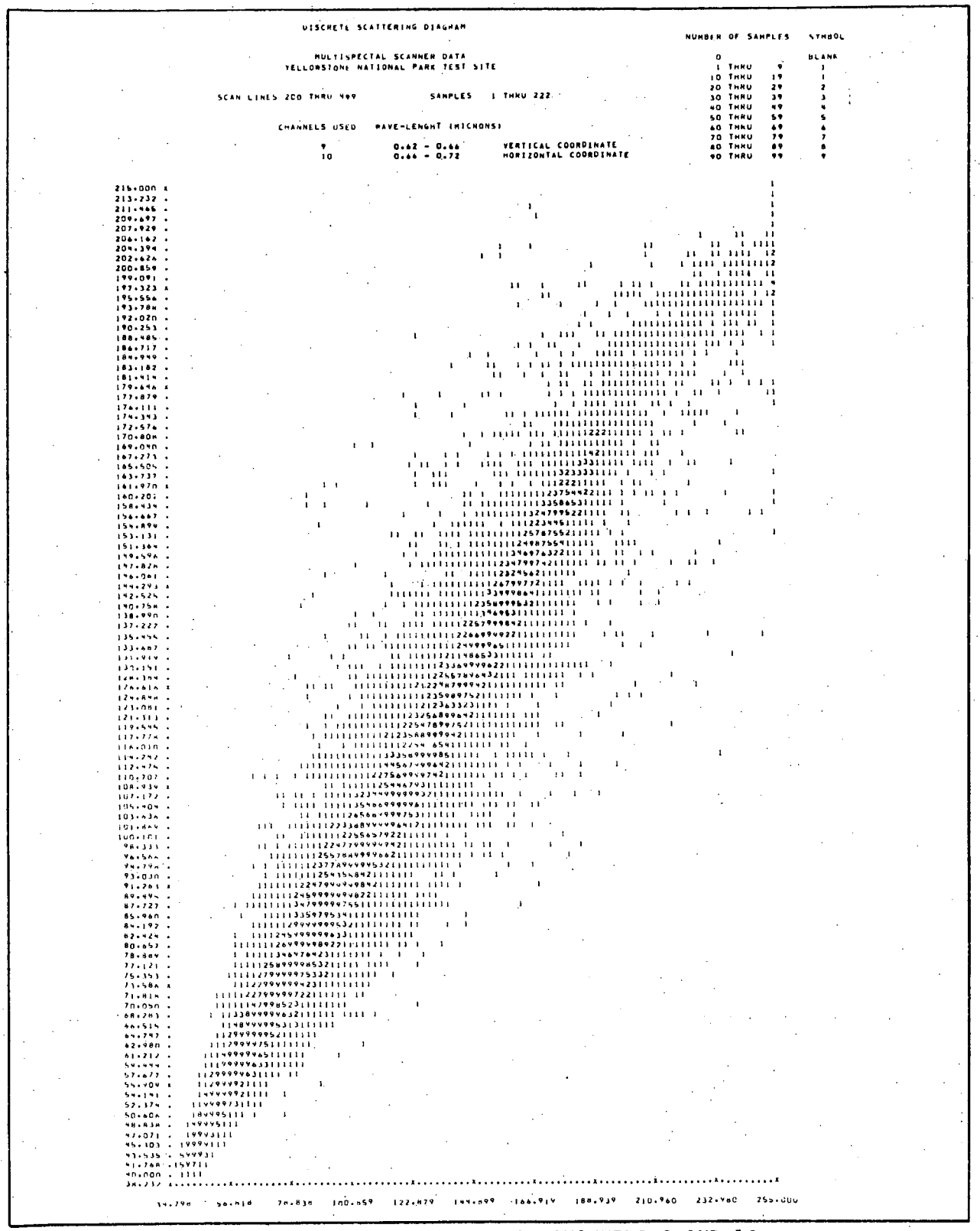


Figure 4-11. SCATTER PLOT OF CHANNELS 9 AND 10

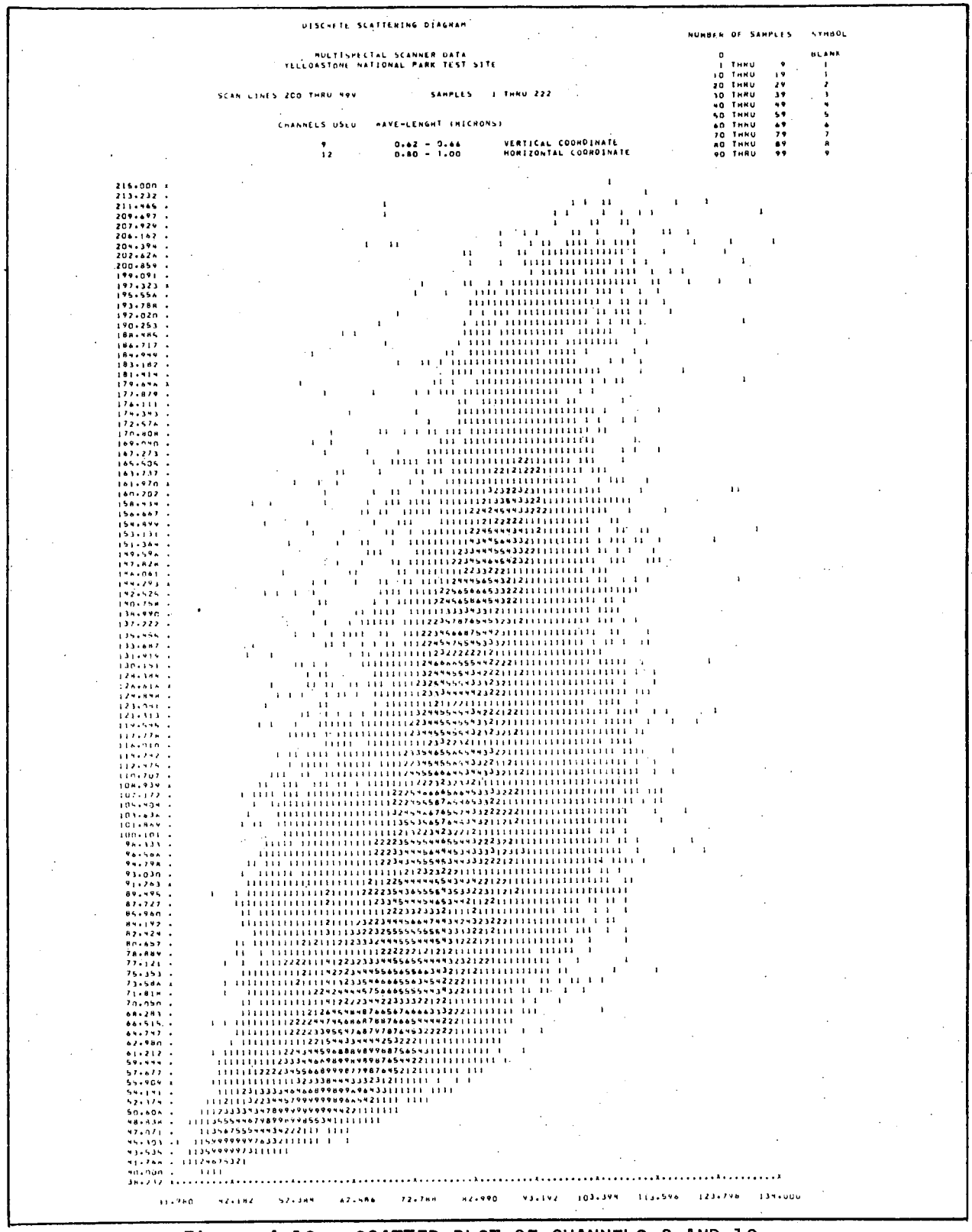


Figure 4-12. SCATTER PLOT OF CHANNELS 9 AND 12

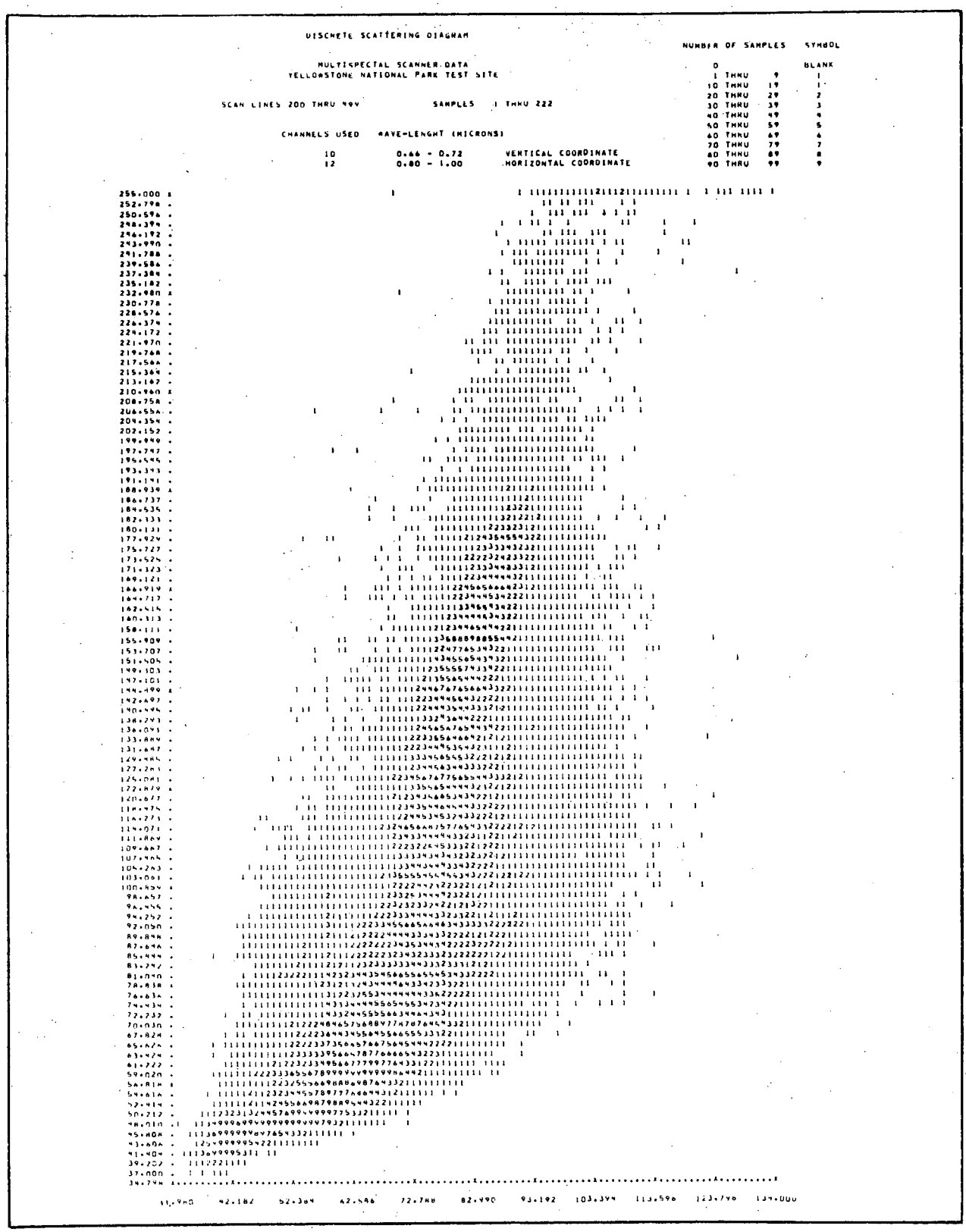


Figure 4-13. SCATTER PLOT OF CHANNELS 10 AND 12

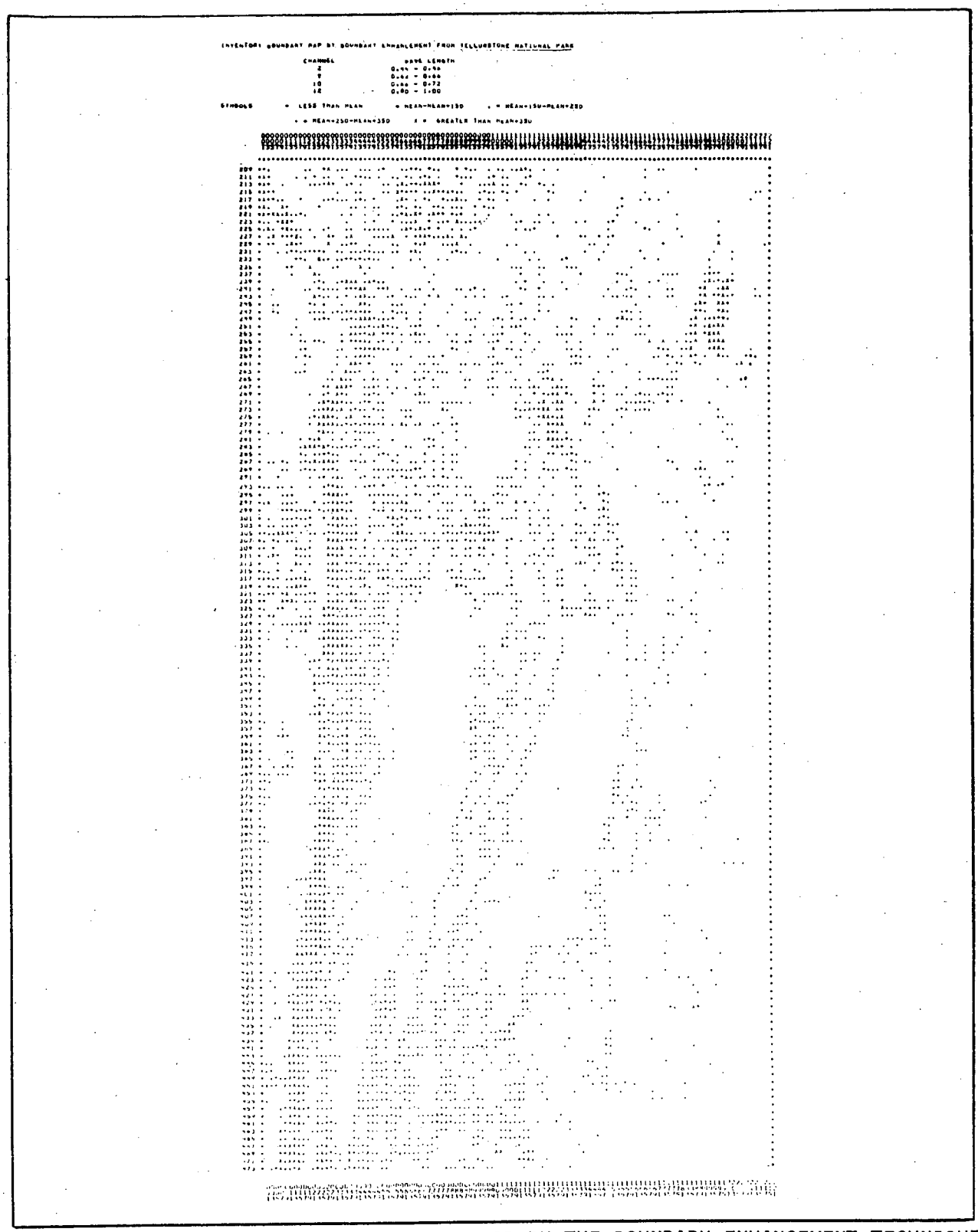


Figure 4-14. THE INVENTORY BOUNDARY MAP BY THE BOUNDARY ENHANCEMENT TECHNIQUE

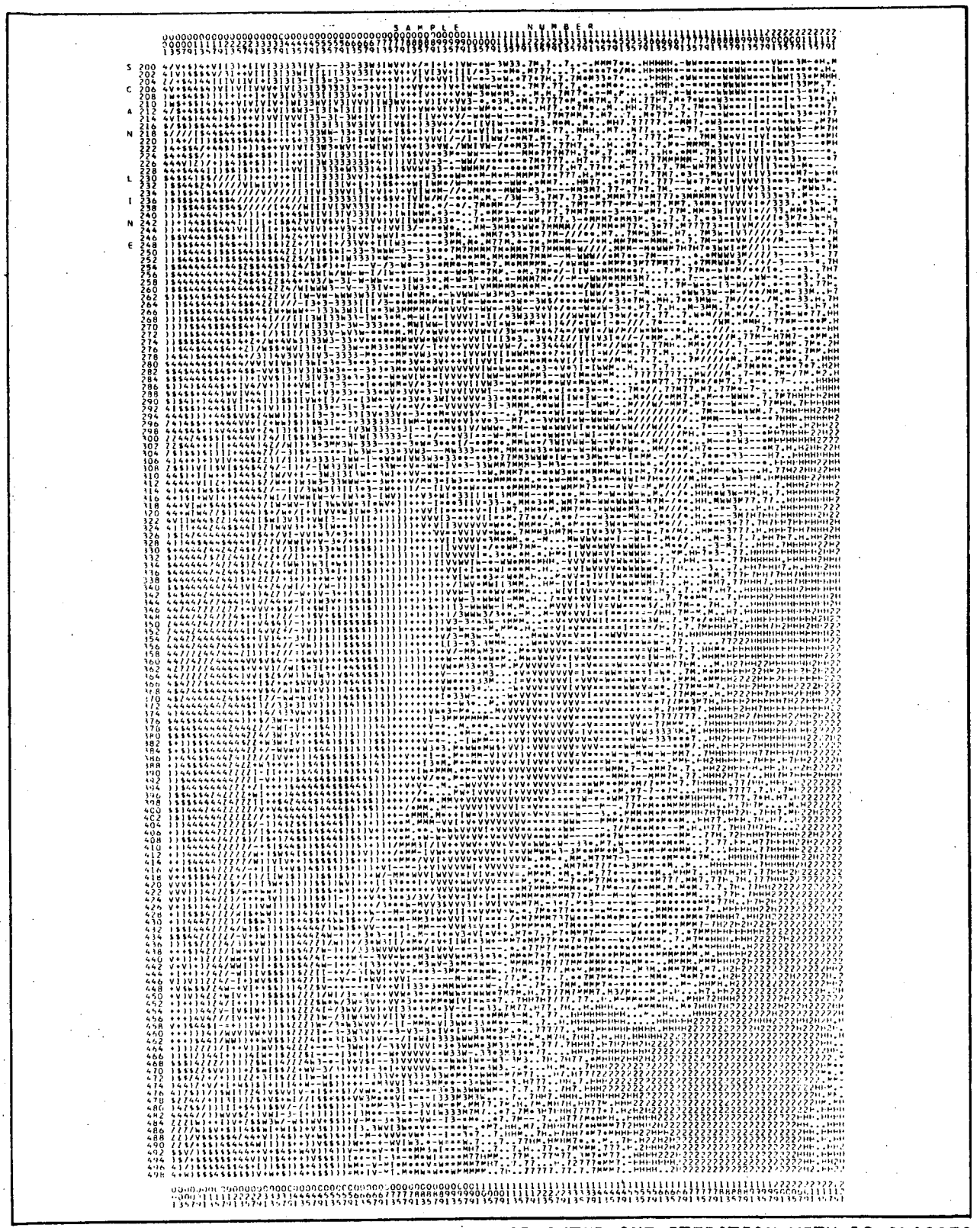


Figure 4-15. UNSUPERVISED CLASSIFICATION MAP AFTER ONE ITERATION WITH 18 CLASSES

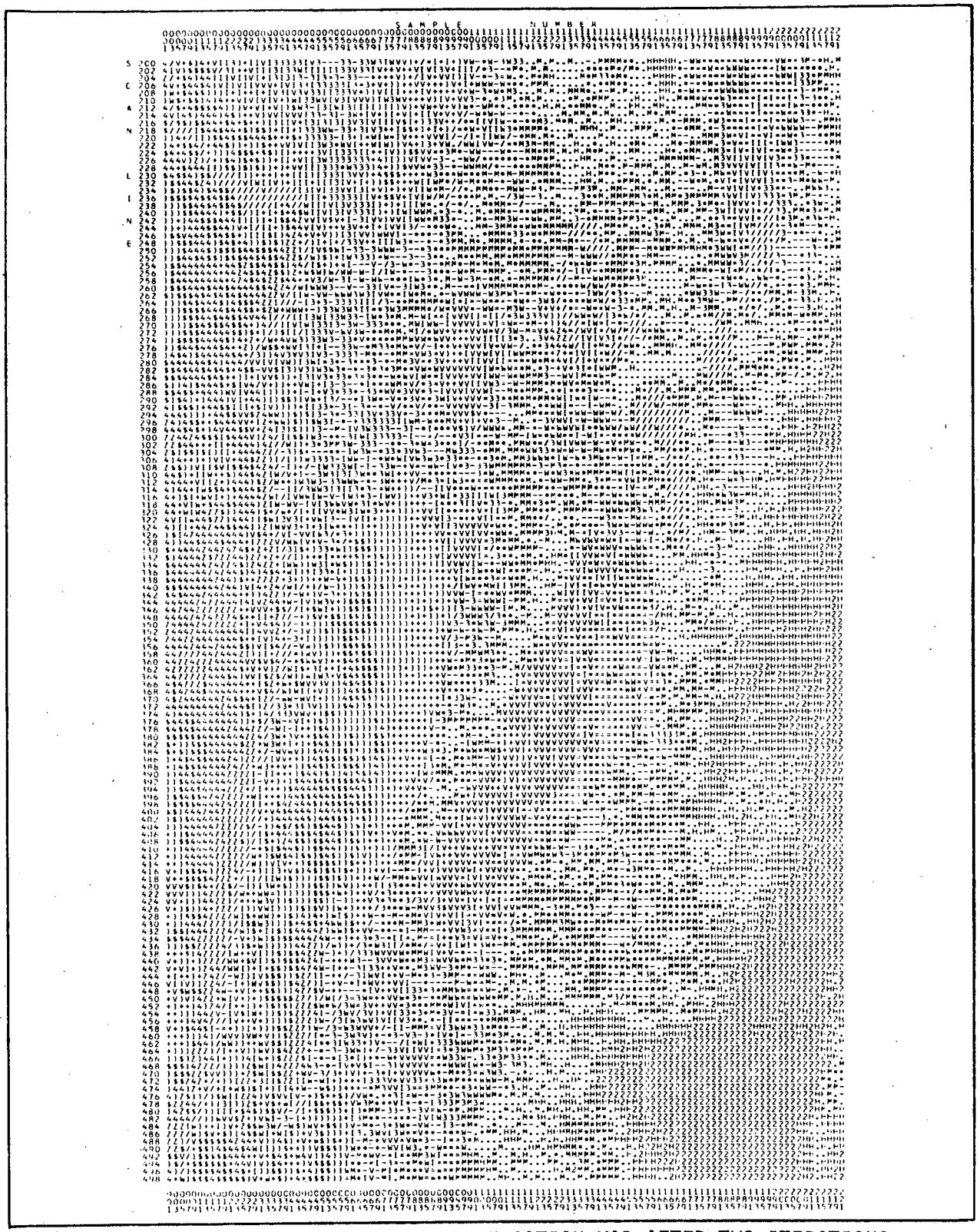


Figure 4-16. UNSUPERVISED CLASSIFICATION MAP AFTER TWO ITERATIONS WITH 17 CLASSES

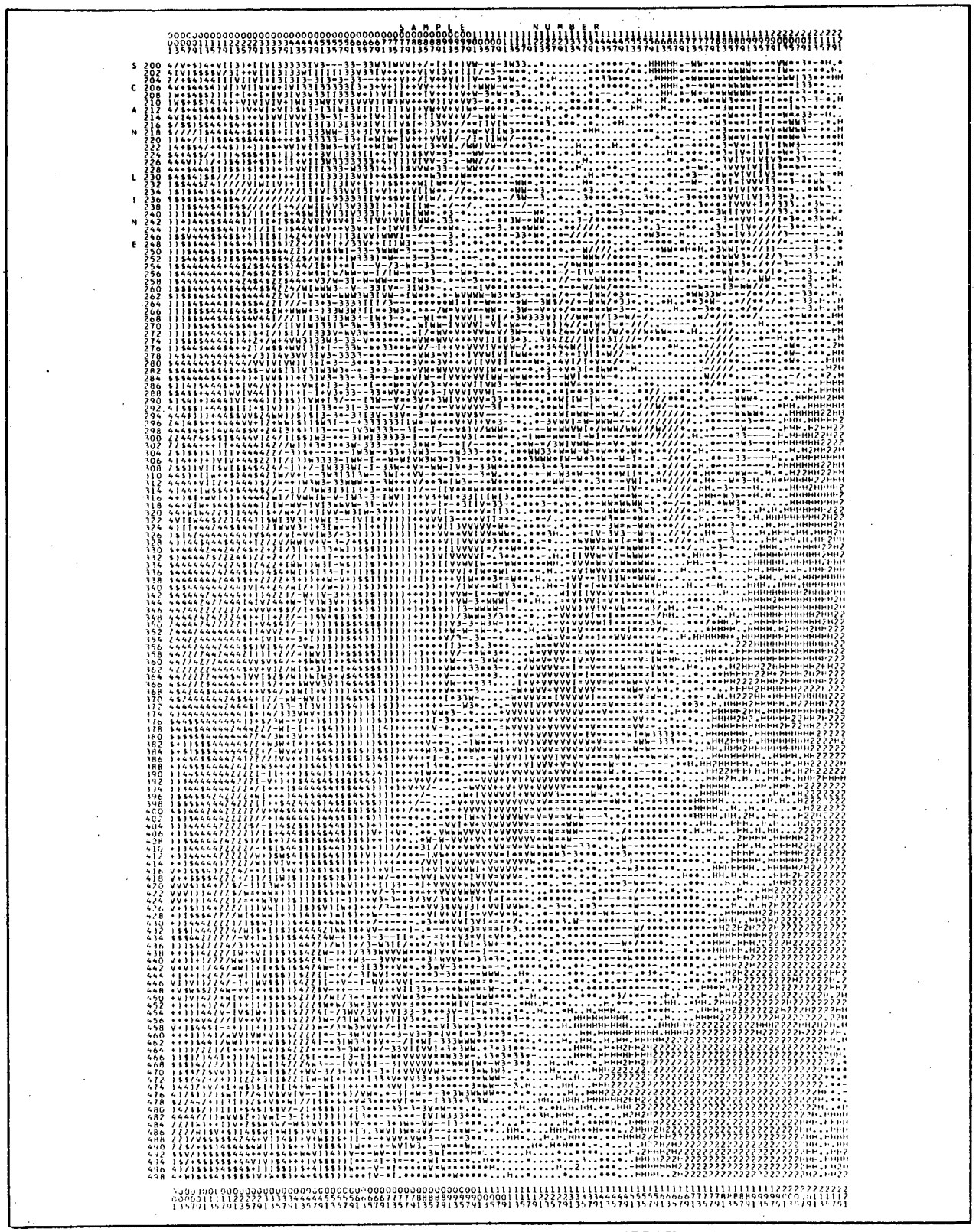


Figure 4-17. UNSUPERVISED CLASSIFICATION MAP AFTER FIRST MERGING WITH 16 CLASSES

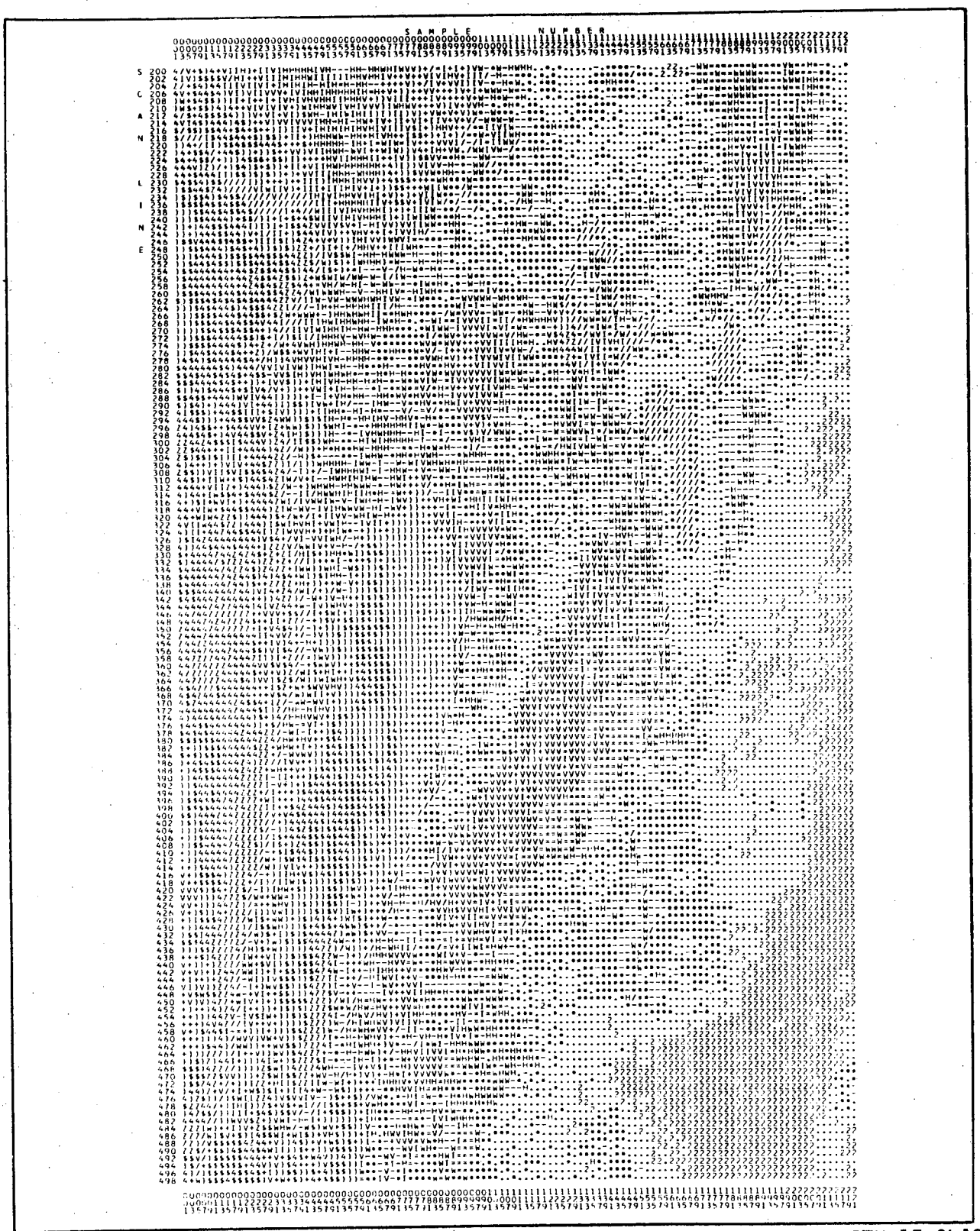


Figure 4-18. UNSUPERVISED CLASSIFICATION MAP AFTER THE SECOND MERGING WITH 15 CLASSES

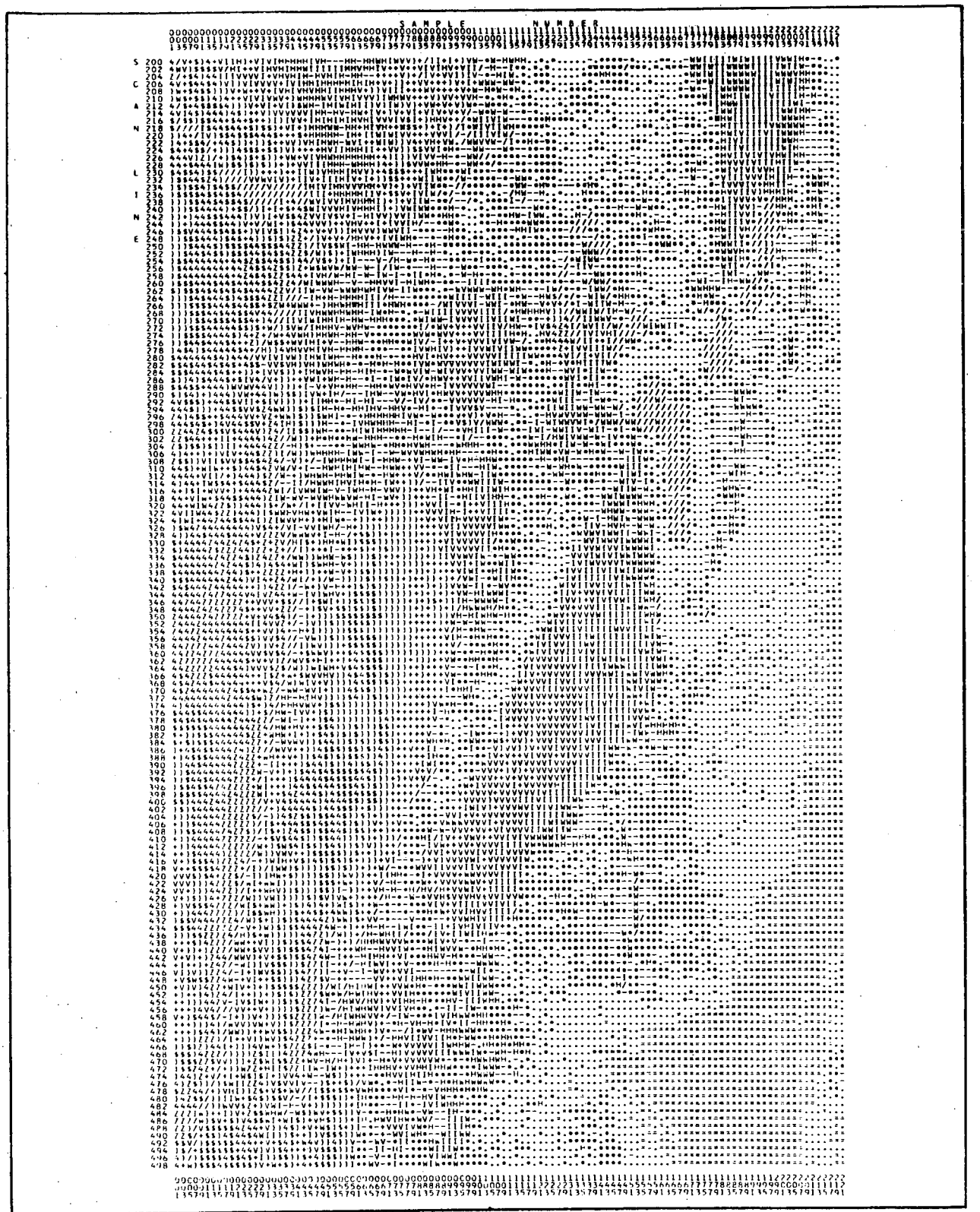


Figure 4-19. UNSUPERVISED CLASSIFICATION MAP AFTER THE THIRD MERGING WITH 14 CLASSES

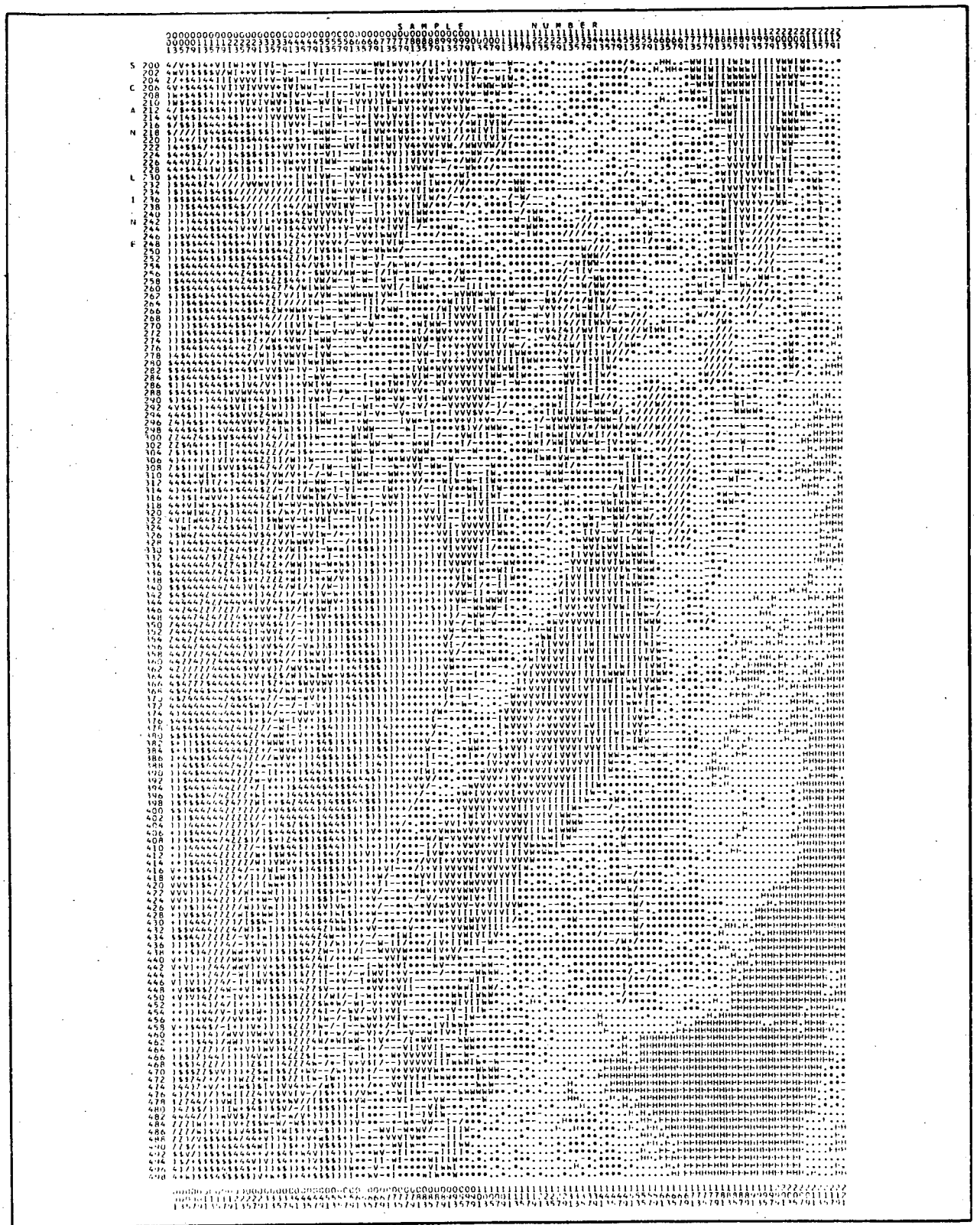


Figure 4-20. UNSUPERVISED CLASSIFICATION MAP AFTER THE FOURTH MERGING WITH 13 CLASSES

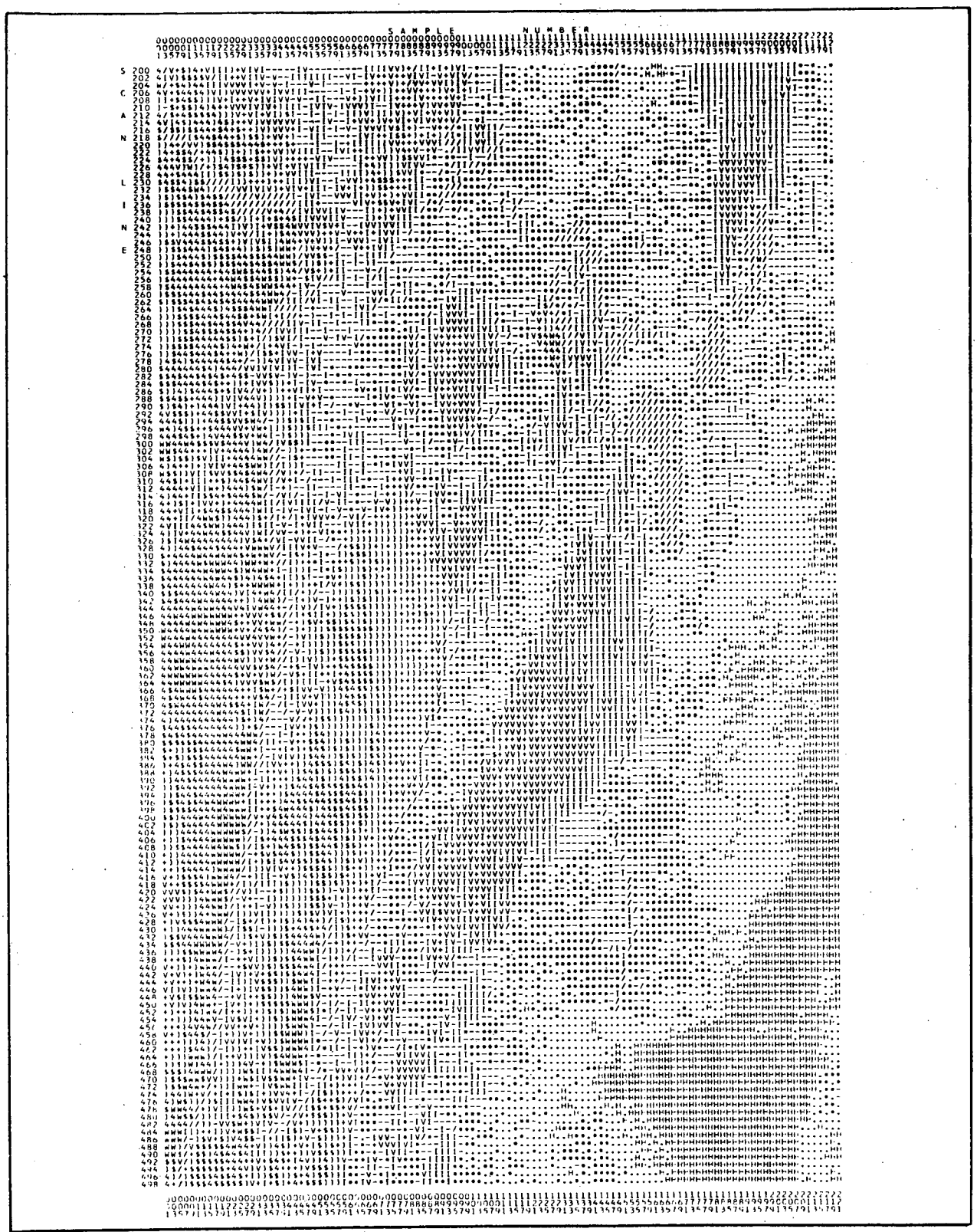


Figure 4-21. UNSUPERVISED CLASSIFICATION MAP AFTER THE FIFTH MERGING WITH 12 CLASSES

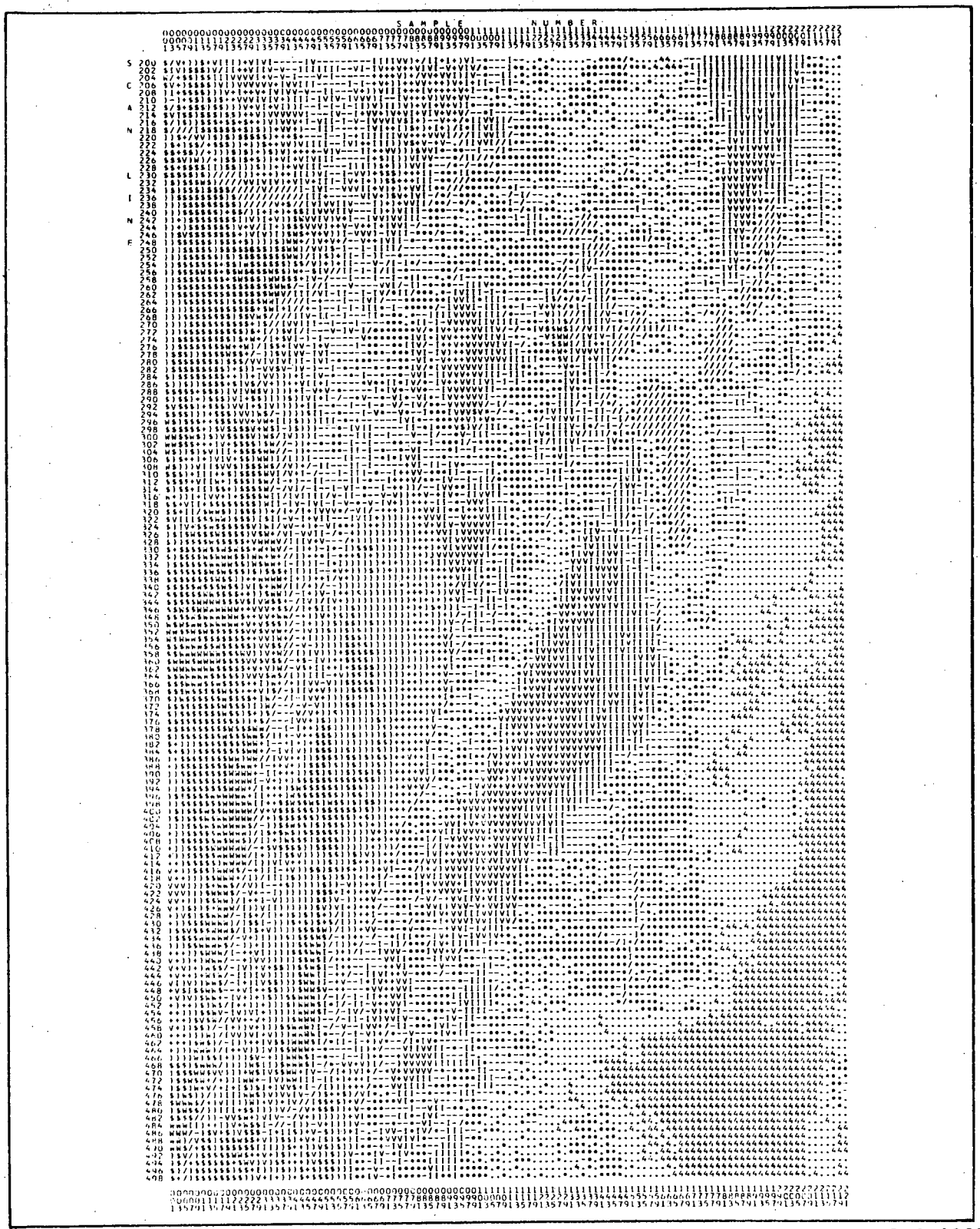


Figure 4-22. UNSUPERVISED CLASSIFICATION MAP AFTER THE SIXTH MERGING WITH 11 CLASSES

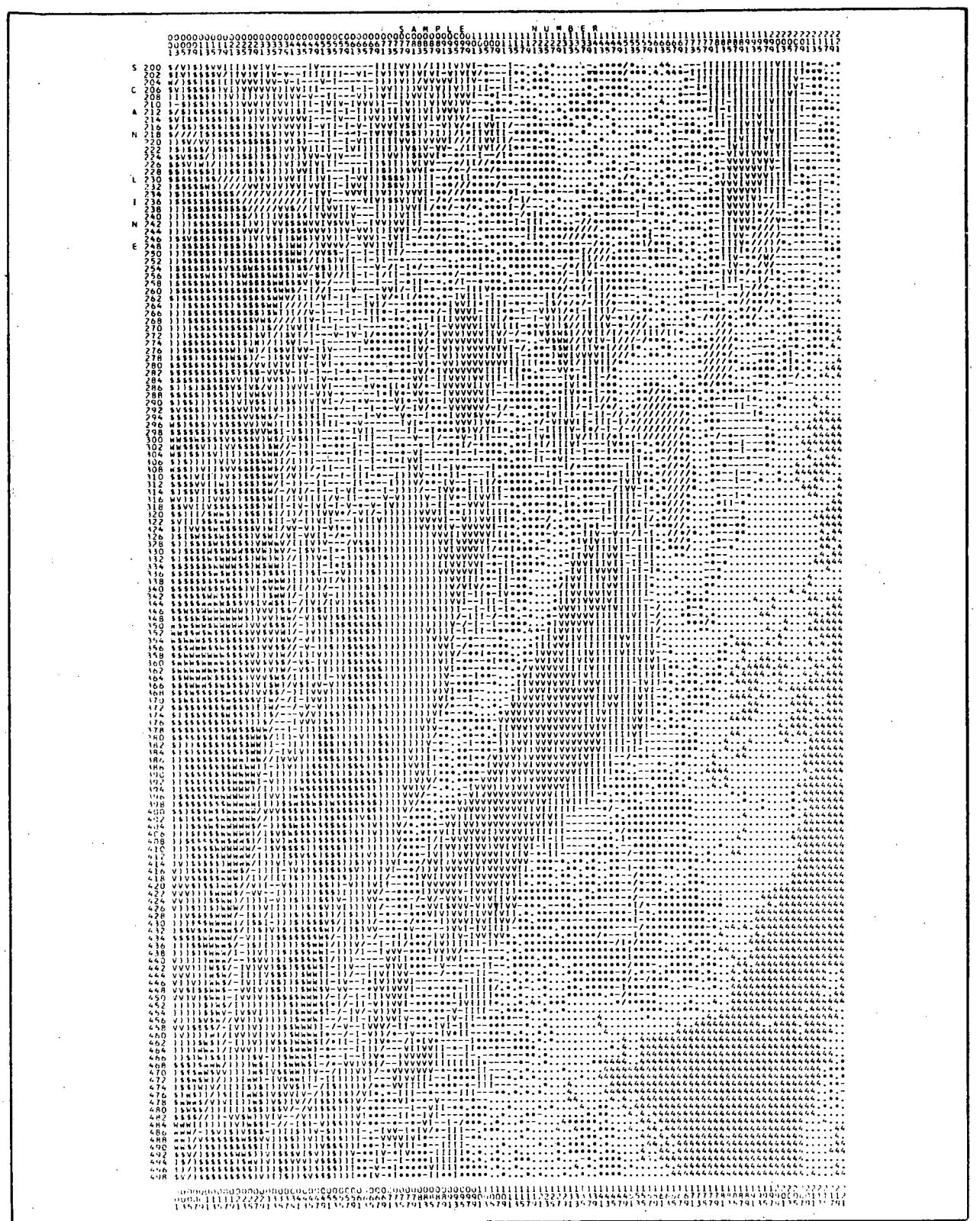


Figure 4-23. UNSUPERVISED CLASSIFICATION MAP AFTER THE SEVENTH MERGING WITH 10 CLASSES

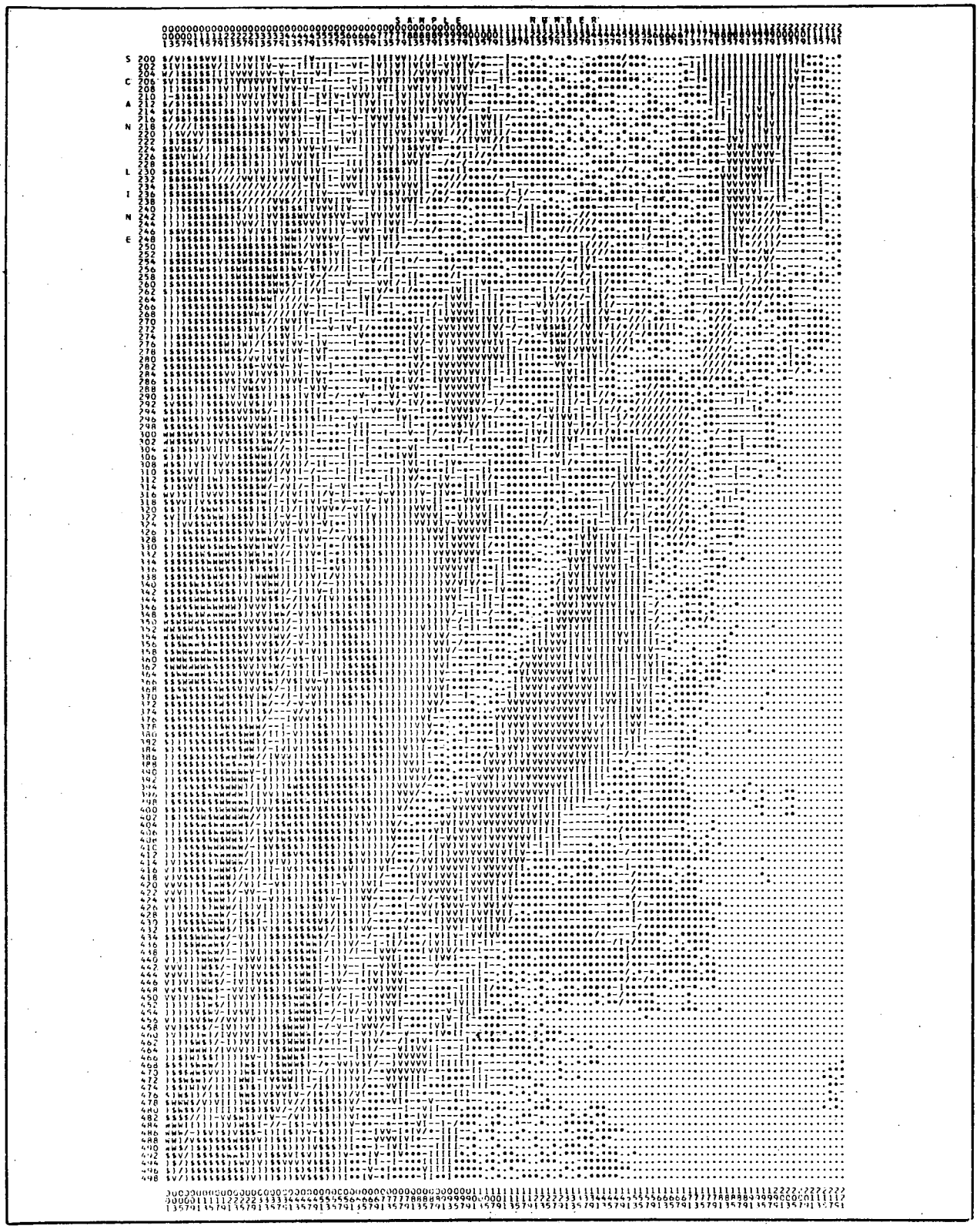


Figure 4-24. UNSUPERVISED CLASSIFICATION MAP AFTER THE EIGHTH MERGING WITH 9 CLASSES

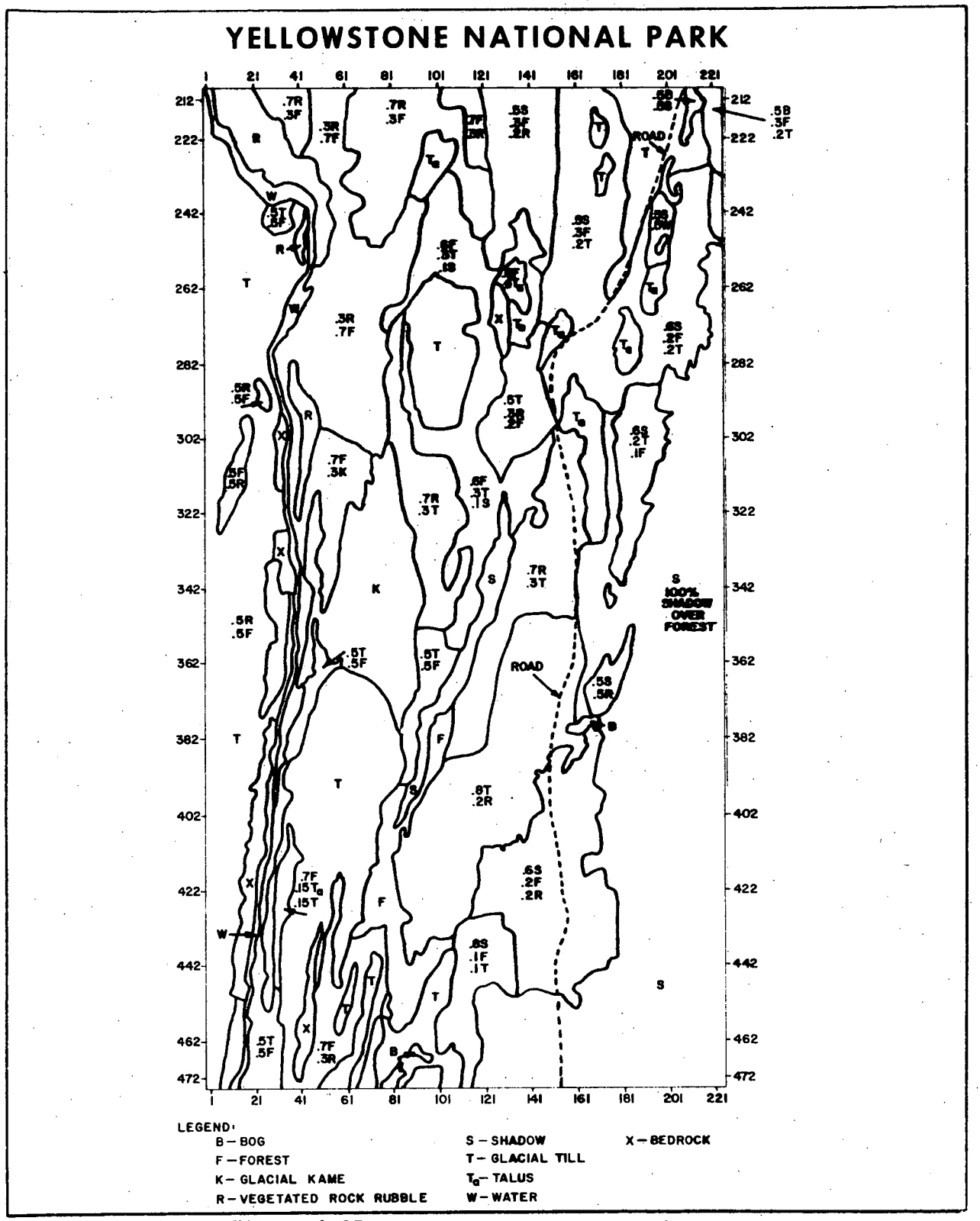


Figure 4-25. GROUND TRUTH SURVEY MAP

Table 4-1. SUMMARY OF UNSUPERVISED CLASSIFICATION AND MERGING OF THE ESTABLISHED CLASSES FOR YELLOWSTONE NATIONAL PARK TEST SITE (SCAN 200-500)

STATISTICAL SEQUENTIAL CLUSTERING		K-MEANS CLUSTERING		MERGING OF CLASSES								PHYSICAL IDENTIFICATION BY GROUND TRUTH MAP
		1ST	2ND	1ST	2ND	3RD	4TH	5TH	6TH	7TH	8TH	
	CLASS NUMBER	NUMBER OF CLASSES										
		18	17	16	15	14	13	12	11	10	9	
OBTAINED MEAN VECTORS OF 18 INITIAL CLASSES	1	)									)	GLACIAL KAME
	2	I				I		I			I	VEGETATED RUBBLE
	3	W										
	4	·		·							·	SHADOW OVER FOREST
	5	V							V		V	GLACIAL TILL
	6	-					-				-	FOREST
	7	$							$		$	GLACIAL TILL
	8	*		*							*	FOREST
	9	/									/	WATER OR TALUS
	10	4										
	11	M	M									
	12	+										
	13	H			H							
	14	Z						W			W	EXPOSED BEDROCK
	15	=	=									
	16	2				=	H		4			
	17	3			H							
	18	7										

Table 4-2. MEAN SPECTRAL VECTORS FOR 18 CLASSES — YELLOWSTONE NATIONAL PARK

CLASS NUMBER	CLASS SYMBOL	NUMBER OF SAMPLES	MEAN SPECTRAL VECTOR			
			CH-2	CH-9	CH-10	CH-12
1	)	1120	110.32	135.89	150.51	83.4
2	I	823	91.45	107.25	115.83	81.89
3	W	836	84.19	96.25	103.65	76.18
4	.	1625	62.49	58.83	58.65	58.80
5	V	923	93.29	115.51	128.92	82.11
6	-	1043	76.32	83.36	88.19	73.18
7	$	905	116.27	145.62	163.33	86.24
8	*	1326	71.94	75.05	77.68	69.70
9	/	433	86.99	86.66	89.57	56.18
10	4	975	121.65	155.46	175.72	88.62
11	M	1518	67.15	67.50	68.46	67.66
12	+	866	103.71	125.70	138.82	82.40
13	H	1272	57.26	51.61	50.94	53.01
14	Z	319	136.69	175.95	204.94	93.73
15	=	779	80.16	101.62	116.34	77.46
16	2	1059	53.86	45.50	44.74	42.20
17	3	828	78.89	89.27	93.61	85.93
TOTAL = 16,650 Samples						

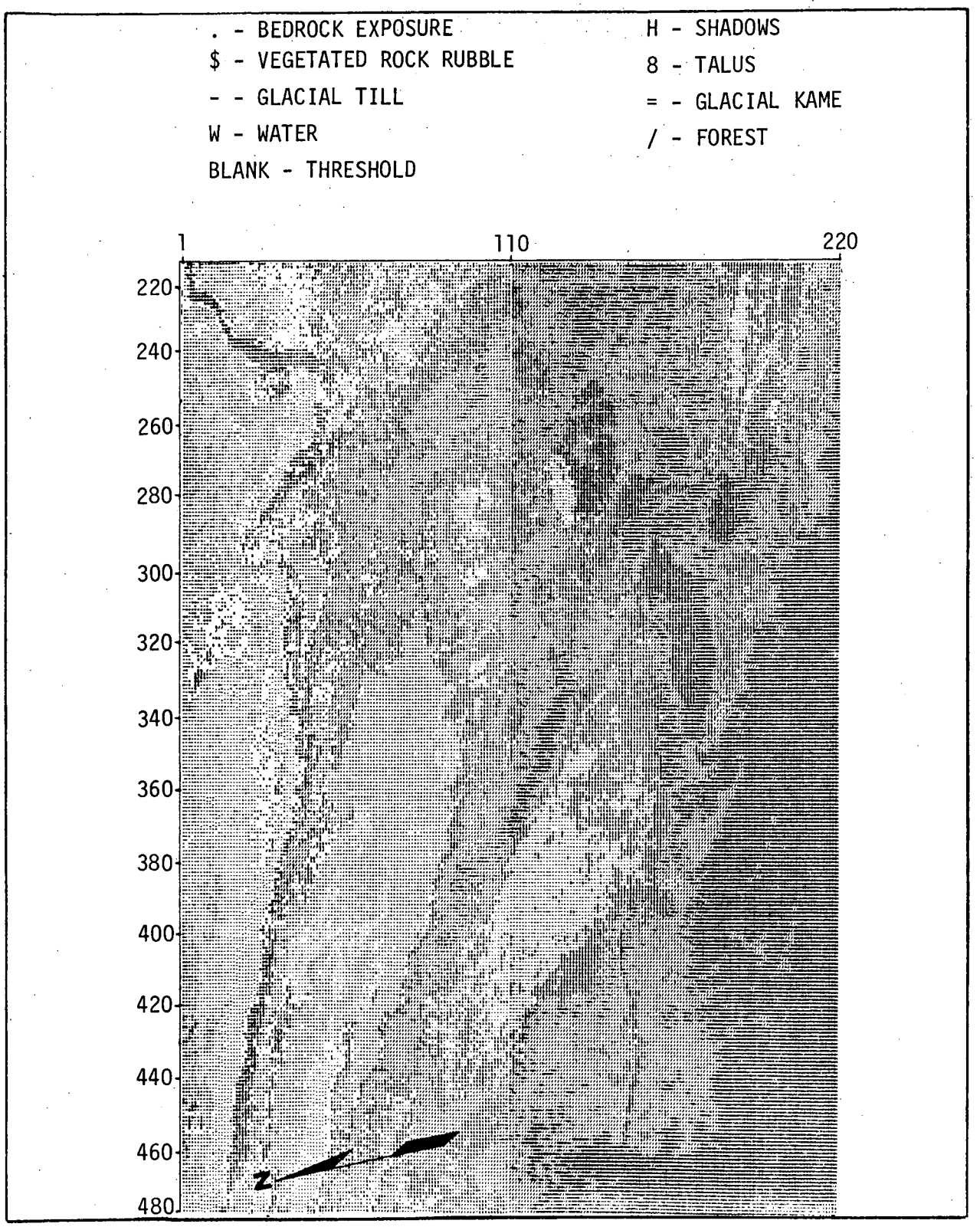


Figure 4-26. LARS CLASSIFICATION: YELLOWSTONE NATIONAL PARK (CH-2, 9, 10 AND 12)

Section V

CONCLUSIONS

In this study, a new composite statistical sequential K-means clustering technique has been developed. It was applied for automatic unsupervised classification of remote multispectral sensing data over the Yellowstone National Park test site and Purdue C-1 Flight line. It was found that the classification technique is about 80 percent correct on both data sets, compared to 86 and 85 percent classification accuracy, respectively, obtained by the Purdue LARS supervised maximum likelihood classification method. In view of the very little human intervention required for the application of the unsupervised classification, the slightly lower accuracy seems still rather good. With these two demonstrations on actual data, it seems fair to assert that the new composite technique may be useful for processing various earth resources survey data. From the operational viewpoint, it is also believed that the unsupervised technique is more feasible than the supervised techniques.

There is still some automatic decision logic needed to be developed in the present unsupervised technique such as (a) to decide the number of classes merging optimally suited for any given data set, and (b) to examine the homogeneity of every class established. These two decision logics are closely related and are important for establishing a completely autonomous nonsupervised classification system. The investigating of such decision logics and implementing them into the computer programs is presently underway. Effort is also underway to integrate the statistical sequential clustering and generalized K-means clustering computer programs into a single, more efficient program for operation type data processing. The above developments will be reported in the future.

Section VI

REFERENCES

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

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

3. Smedes, H. W., Linnerud, H. J., Hawks, S. G., and Woolaver, L. B., "Digital Computer Mapping of Terrain by Clustering Techniques Using Color Film as a Three-Band Sensor", Proc. 7 International Symposium on Remote Sensing of Environment, May 17-21, 1971.

4. Smedes, H. W., Su, M. Y., Jayroe, R. R. et al., "Mapping of Terrain by Computer Clustering Techniques Using Multispectral Scanner Data and Using Color Films", Proceedings of NASA 4th Earth Resources Program Review, January 17-21, 1972.

5. Schell, J. A., "A Comparison of two Approaches for Category Identification and Classification Analysis From an Agricultural Scene", paper presented at the Conference on Earth Resources Observation and Information Analysis System, the University of Tennessee Space Institute, Tullahoma, Tennessee, March 13-14, 1972.

6. Turner, B. J., "Cluster Analysis of MSS Remote Sensor Data", paper presented at the same conference as ref. 5.

7. Su, M. Y., Jayroe, R. R., and Cummings, R. E., "Unsupervised Classification of Earth Recources Data", paper presented at the same conferences as reference 5.

8. 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.

9. Su, M. Y., Pooley, J., and Hand, C., "Statistical Algorithms and Computer Programs for Multispectral Observations", NASA CR-103182, December 1970.

10. 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.

11. Su, M. Y., "Algorithms for Sequential Classification of Multispectral Observations into Homogeneous Populations", Northrop-Huntsville Memorandum M-794-808, October 1970.

12. 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.

13. Su, M. Y., "A Generalized K-means Algorithm for Clustering Multispectral Observations", Northrop-Huntsville Memorandum M-794-964, June 1971.

14. Gasey, R. C. and Nagy, G., "Advances in Pattern Recognition", Scientific American, Vol. 224, No. 4, April 1971, pp. 56-71.

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

16. 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.

17. Ball, G. H., "Data Analysis in the Social Sciences: What about the details?", Proc. Fall Joint Computer Conf. pp. 533-559, December 1965.

18. Purdue University, "Remote Multispectral Sensing in Agriculture", LARS Vol. 1, No. 4, October 1968.