LARS Publication 010682

Tempisque Valley (Costa Rica) Case Study

by

Shirley M. Davis

For use with

Flexible Workshop on
Numerical Analysis of
Multispectral Image Data
(LARS Publication 010482)

December 1982

Preface

This case study is designed to accompany LARS Publication 010482, "Flexible Workshop on Numerical Analysis of Multispectral Image Data" by James C. Tilton and Luis A. Bartolucci.

In addition to this publication, you will need the following materials to carry out the activities:

1. A book of computer printouts containing all the output from this analysis.

2. Topographic maps of this portion of Costa Rica at a scale of 1:50,000. Appropriate ones were prepared in 1969, 1972, and 1973 by the Instituto Geografico Nacional, San Jose, Costa Rica.

3. Black-and-white low-altitude aerial photographs of portions of the study area. An appropriate set was taken in December 1977 at a scale of approximately 1:20,000.

It is also convenient to have available one or two assembled gray-scale line-printer maps and an assembled classification map. While these are included in the book of printouts, it is difficult to get a sense of the total area unless they are assembled.

PART I - INTRODUCTION

The case study analysis featured in this workshop is a LARSYS-based analysis of Landsat data collected in 1975 over the Tempisque Valley in the western portion of Costa Rica. The area has a tropical climate, with large areas devoted to tropical agriculture, in particular rice and sugar cane. Livestock production is one of the major industries, with much prime agricultural land used for grazing. In the lowlands, there are dense forests, and mangroves grow along the coastal regions. Saline soils are found in areas that flood frequently. There are no major population areas in the study site. The mouth of the Tempisque River opens into the Gulf of Nicoya; a 12 km long island is located in the bay. A map of the region is provided in Figure I-2.

During the analysis, we will create a land-cover map of an area of approximately 100,000 hectares, or 1000 sq. km. The analysis will seek to distinguish between water (pure and turbid), agricultural areas with varying crop conditions, forest, mangroves, grass and grazing areas, bare soil or bare rock.

One way to obtain information about a digital MSS data set stored in a CCT is to use a LARSYS processor called IDPRINT. IDPRINT prints the identification record from a multispectral image storage tape.

Study the output from IDPRINT (see pages 1-3 of the computer printouts) and note:

1. the run number for this data set. Each data set has a unique run number, the first two digits indicating the year in which the data was collected.
2. the date and time of day when the data were taken. (Don't confuse the date the data were taken with the reformatting date.)
3. the number of lines and the number of data samples (columns) in this data set.
4. the spectral band for each of the four channels of data. In which portions of the spectrum do these bands occur?
5. the translation from wavelength bands 4, 5, 6 and 7 (as

---

* This case study description, authored by Shirley M. Davis, is based on an analysis performed by Luis A. Bartolucci, Carlos R. Valenzuela, Darlys C. McDonald, and the author.

identified by the EROS Data Center) to data channels 1, 2, 3 and 4.

6. the calibration pulse values. Historically, aircraft scanner systems recorded three calibration signals for each channel. In the case of Landsat, only two calibration sources are used; however, the values shown under C0 and C1 can still be used to convert the data into radiometric units. In this case, C0 has the value 0.0 for each band, and C1 has values that indicates the maximum radiance for each channel in mwatts/cm^2 sr. The values are the same for all data from Landsat 1.

In order to find out what the numerical data values are that are associated with specific locations in the data (pixels), we can use another LARSYS processor, TRANSFERDATA. Pages 80 through 85 of the computer listing show the output of TRANSFERDATA for four individual pixels, selected to represent different ground materials.

Notice that each pixel is identified by its line and column address and has four data values associated with it. This set of values is also referred to as a "data vector." The data values indicate the relative amount of energy returned from the Earth's surface to the sensor, in other words, the brightness of that spot on the Earth's surface as measured by the scanner system in each wavelength band. Data values may range from 0 to 127 in each channel. (Normally, Landsat channel 4 responses, as contained in data from EROS Data Center, span values from 0 to 63; in this and subsequent processing, however, channel 4 has been expanded so that it is equivalent to the other three in dynamic range.) A data value of zero is dark, signifying that no energy is measured by that detector for that pixel; 127 is bright, indicating saturation of the scanner detector.

The four data vectors printed by TRANSFERDATA (pages 81 through 84) are representative of the many thousands of data vectors within the scene. Later, when the computer classifies the data, it will assign each pixel to a class on the basis of these numerical values.

The data vector below, from a different data set, is a pixel known to represent forest:

		CHANNELS			
		1	2	3	4
LINE	COL	DATA			
115	326	26.0	15.0	58.0	62.0

Compare these values to the ones shown on pages 81 through 84 of the printouts. Which page shows data values which probably

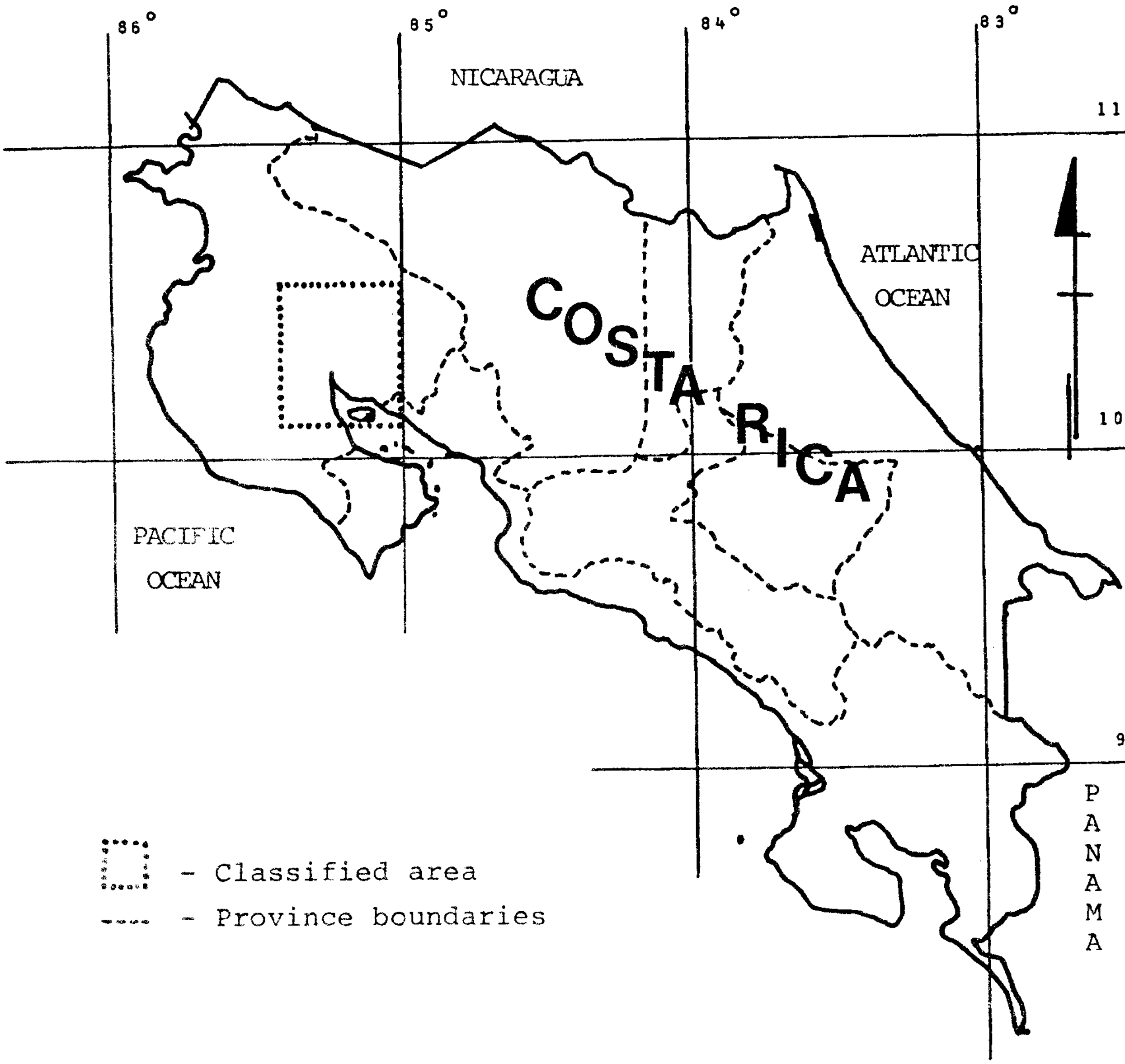


Figure I-2. Map of Costa Rica showing the location of the area classified in this case study.

also represent forest? How did you decide? If you were allowed to use only Channel 3 data, could you have decided as easily? How would it have been if you were limited to only Channels 2 and 3?

From your understanding of spectral characteristics of materials found on the Earth's surface, can you figure out what any of the other pixels represent?

 Pages II-1 to II-10 not in file

PART II - MSS DATA SELECTION, CORRELATION WITH REFERENCE DATA AND TRAINING SAMPLE SELECTION

This portion of the case study demonstrates one possible way in which the first step in the analysis of MSS data might be accomplished. A data set that had been collected by Landsat-1 on March 3, 1975, was available for this region. At this time of year, the analyst decided, the features to be identified could be characterized in the data.

ASSESSING DATA QUALITY

The first thing we want to do with the data set is to examine it in order to assess the data quality, both the amount of cloud cover and any radiometric distortions. Normally, black-and-white and false-color images produced by EROS Data Center are used for this purpose; in this instance they were not available. Using an alternate method, therefore, first, full-frame line-printer images of a visible channel and an infrared channel were produced (at a sampling rate of 1:5) to provide a quick look at the data and to enable us to determine line and column coordinates of the Tempisque Valley area. Using these coordinates, we produced dot-matrix printer images of all four bands for the area of interest. They are shown as Figures II-4 through II-7 in this manual.

These gray scale images were originally produced on an electrostatic printer-plotter that uses dot patterns of varying darkness to represent the gray level to which each pixel is assigned. The printer-plotter creates an image which is picture-like in appearance, revealing the spatial features of the data. We can also print gray scale maps on a line printer, using symbols of varying darkness to represent the gray levels. This creates a larger map in which each pixel can be individually examined. See Figure II-8 for an example of a line printer gray scale map.

The process used to generate the gray scale maps, both for the dot-matrix printer and the line printer, is a combination of level slicing and contrast stretching. First the overall range of the data in each channel is determined; the data to be used are specified by the analyst in the block card and the coordinates are shown in the printouts under the heading "Data Block(s) Histogrammed." For example, if you look at the back of Figure II-4, you see the run number repeated, 75044401, and then the line and column identification of the data used to determine the data range. In this case the data are those in Channel 1 from line 401 through line 771 and in column 420 through column 933. The line interval and sample interval chosen is 2, meaning that only data in every other line and every other column is used for the histogramming, or 25% of the data. A "1" in that position would mean that every data point within that block is used to calculate the histogram. This interval used is determined by the size of the image and the scene complexity. A larger interval would be used for a larger image with equivalent scene complexity, and a smaller interval would be used for a more complex image of the same size. Reducing the sample saves valuable computer time.

Once the actual dynamic range of the data is determined by the processor, the data are then histogrammed and divided into sixteen levels. The printout on the back of Figure II-4 shows the lower and upper limits of each subset or "bin" of data, the level number (from 1 through 16), the number of pixels assigned to each bin, and the percent of the total histogrammed area that those points represent. For example, look at the table under the heading "THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE." The third line in that table indicates that all pixels in the histogrammed area with data values from 16.5 to 17.5 in Channel 1 are assigned to Gray Level 3 and that there are 6549 of them, amounting to 13.7% of the total points histogrammed. All pixels assigned to Gray Level 3 are displayed by the matrix printer with the gray tone shown in the block third from the right. (Note that the half-tone pattern goes from Level 16 on the left to Level 1 on the right.) The remainder of the table specifies the limits for determining which gray tone is used to represent any data value in the scene. These limits are caluculated so as to utilize fully the 16 gray scale levels available from the matrix printer. The same procedure is used for the gray scale map produced by the line printer; the only difference is that instead of using dot patterns to achieve various gray tones, alphanumeric characters are used, with the darkest ones (M, Z, X) assigned to the lowest response levels and the lightest ones (.,-,blank) assigned to the bins with the highest data values.

From the information on the backs of Figures II-4 through II-7, identify which image represents the data values in Channel 1; Channel 2; Channel 3; and Channel 4.

Examine the gray scale images. Why are the images from Channels 1 and 2 similar to each other? Why are the images from Channels 3 and 4 similar to each other but different from those of Channels 1 and 2?

What can you identify on the images?

Are any data quality problems apparent on any of the images? Does there appear to be a problem with cloud coverage?

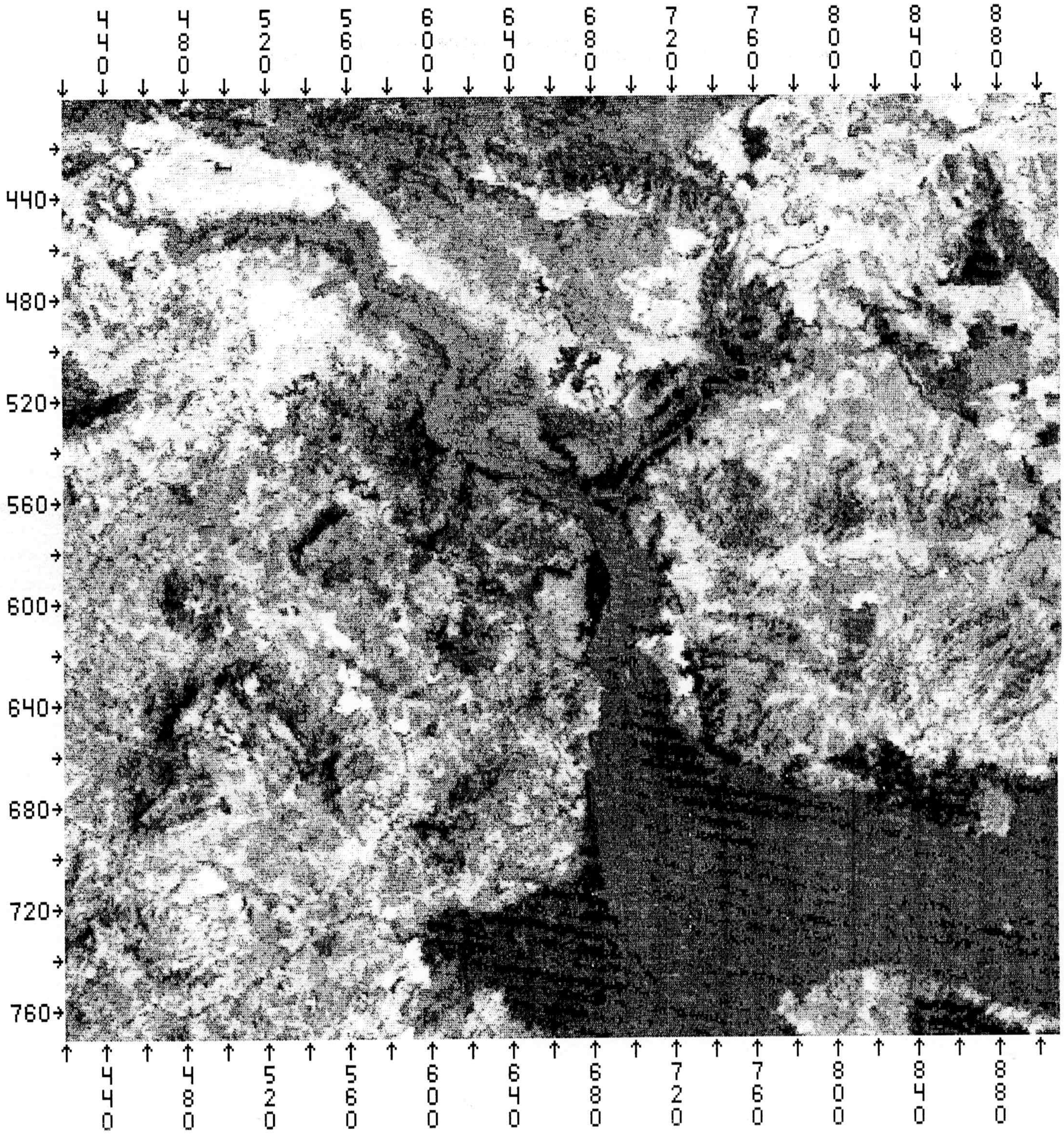


Figure II-4. Gray-scale printout of data in Channel 1 for a portion of the study area.

RUN NUMBER........... 75044401

FLIGHT LINE... 204015163 C.RICA

DATA TAPE/FILE NUMBER.. 3466/ 1

REFORMATTING DATE. SEPT 28,1976

DATE DATA TAKEN... MAR 3,1975

TIME DATA TAKEN..... 1416 HOURS

PLATFORM ALTITUDE..3062000 FEET

GROUND HEADING..... 180 DEGREES

CHANNEL 1 SPECTRAL BAND 0.50 TO 0.60 MICROMETERS CALIBRATION CODE= 1 C0 = .0

HISTOGRAM BLOCK(S)

RUN NUMBER	LINES	COLUMNS	CALIBRATION CODE
75044401	(401, 771, 2)	(420, 933, 2)	1

THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE

LOWER LIMIT	UPPER LIMIT	LEVEL NUMBER	SAMPLE COUNT	PER CENT OF TOTAL SAMPLE
<	16.5	1	3303	6.9
16.5	16.5	2	0	0.0
16.5	17.5	3	6549	13.7
17.5	18.5	4	3963	8.3
18.5	18.5	5	0	0.0
18.5	19.5	6	2799	5.9
19.5	21.5	7	4229	8.8
21.5	22.5	8	4404	9.2
22.5	22.5	9	0	0.0
22.5	23.5	10	6910	14.5
23.5	24.5	11	1260	2.6
24.5	25.5	12	3670	7.7
25.5	25.5	13	0	0.0
25.5	26.5	14	4702	9.8
26.5	29.5	15	3759	7.9
29.5	>	16	2254	4.7

THE TOTAL NUMBER OF SAMPLE POINTS... 47802

THE AVERAGE NUMBER OF SAMPLE POINTS ASSIGNED PER GRAY LEVEL... 2987.625

THE STANDARD DEVIATION OF THE NUMBER OF SAMPLE POINTS PER GRAY LEVEL... 2249.860

HALF-TONE PATTERN 'H5W4GRAY' WILL BE USED FOR THIS PLOT - THE GRAY SCALE LEVELS FOR THIS PATTERN ARE (FROM 16 TO 1) ...

Figure II-4 legend

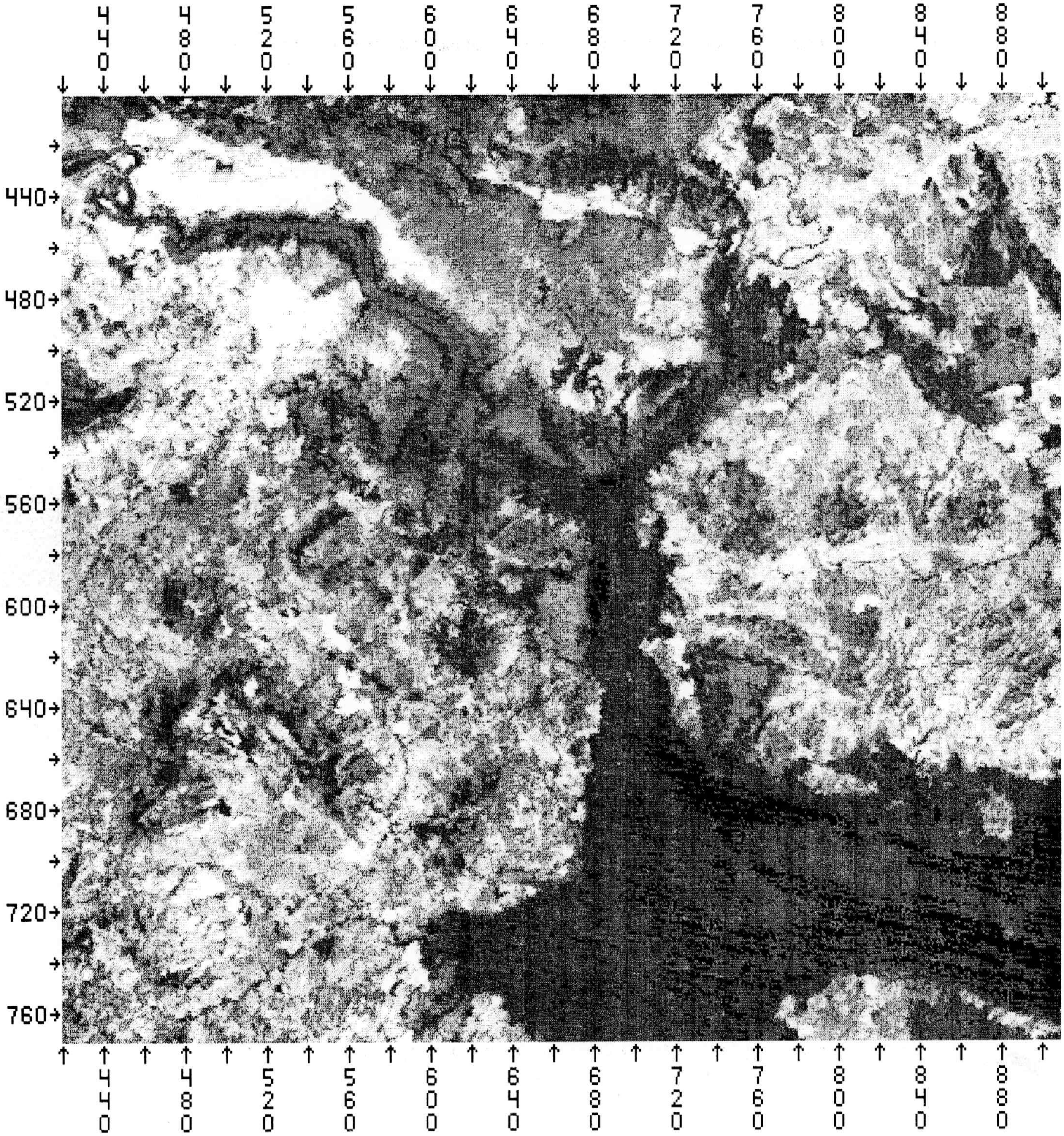


Figure II-5. Gray-scale printout of data in Channel 2 for a portion of the study area.

RUN NUMBER........... 75044401

FLIGHT LINE... 204015163 C.RICA

DATA TAPE/FILE NUMBER.. 3466/ 1

REFORMATTING DATE. SEPT 28,1976

DATE DATA TAKEN... MAR 3,1975

TIME DATA TAKEN..... 1416 HOURS

PLATFORM ALTITUDE..3062000 FEET

GROUND HEADING..... 180 DEGREES

CHANNEL 2 SPECTRAL BAND 0.60 TO 0.70 MICROMETERS CALIBRATION CODE= 1 C0 = .0

HISTOGRAM BLOCK(S)

RUN NUMBER	LINES	COLUMNS	CALIBRATION CODE
75044401	(401, 771, 2)	(420, 933, 2)	1

THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE

LOWER LIMIT	UPPER LIMIT	LEVEL NUMBER	SAMPLE COUNT	PER CENT OF TOTAL SAMPLE
<	12.5	1	1792	3.7
12.5	14.5	2	4637	9.7
14.5	17.5	3	2383	5.0
17.5	20.5	4	3120	6.5
20.5	23.5	5	3191	6.7
23.5	25.5	6	3293	6.9
25.5	27.5	7	3903	8.2
27.5	28.5	8	2303	4.8
28.5	30.5	9	1790	3.7
30.5	32.5	10	3580	7.5
32.5	33.5	11	2634	5.5
33.5	35.5	12	3196	6.7
35.5	37.5	13	2406	5.0
37.5	39.5	14	3739	7.8
39.5	43.5	15	2809	5.9
43.5	>	16	3026	6.3

THE TOTAL NUMBER OF SAMPLE POINTS... 47802

THE AVERAGE NUMBER OF SAMPLE POINTS ASSIGNED PER GRAY LEVEL... 2987.625

THE STANDARD DEVIATION OF THE NUMBER OF SAMPLE POINTS PER GRAY LEVEL... 768.730

HALF-TONE PATTERN 'H5W4GRAY' WILL BE USED FOR THIS PLOT - THE GRAY SCALE LEVELS FOR THIS PATTERN ARE (FROM 16 TO 1) ...

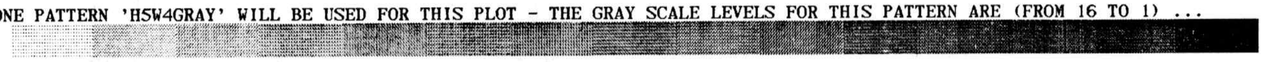


Figure II-5 legend

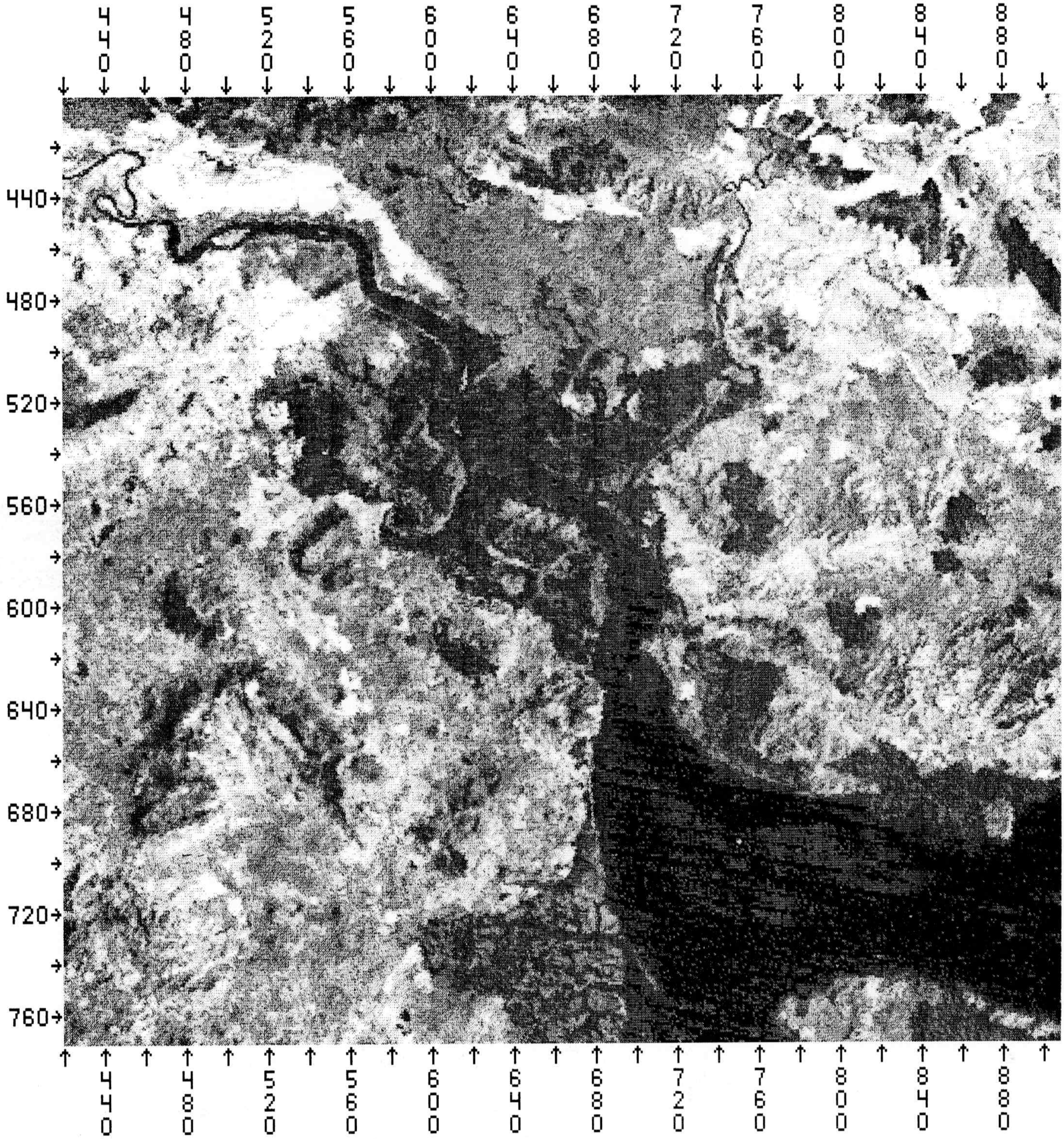


Figure II-6. Gray-scale printout of data in Channel 3 for a portion of the study area.

RUN NUMBER........... 75044401

FLIGHT LINE... 204015163 C.RICA

DATA TAPE/FILE NUMBER.. 3466/ 1

REFORMATTING DATE. SEPT 28,1976

DATE DATA TAKEN... MAR 3,1975

TIME DATA TAKEN..... 1416 HOURS

PLATFORM ALTITUDE..3062000 FEET

GROUND HEADING..... 180 DEGREES

CHANNEL 3 SPECTRAL BAND 0.70 TO 0.80 MICROMETERS CALIBRATION CODE= 1 C0 = .0

HISTOGRAM BLOCK(S)

RUN NUMBER	LINES	COLUMNS	CALIBRATION CODE
75044401	(401, 771, 2)	(420, 933, 2)	1

THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE

LOWER LIMIT	UPPER LIMIT	LEVEL NUMBER	SAMPLE COUNT	PER CENT OF TOTAL SAMPLE
<	6.5	1	3768	7.9
6.5	11.5	2	2249	4.7
11.5	28.5	3	3083	6.4
28.5	33.5	4	3123	6.5
33.5	37.5	5	3192	6.7
37.5	39.5	6	2213	4.6
39.5	41.5	7	4298	9.0
41.5	42.5	8	1498	3.1
42.5	44.5	9	2794	5.8
44.5	45.5	10	3670	7.7
45.5	47.5	11	3404	7.1
47.5	49.5	12	2903	6.1
49.5	51.5	13	2332	4.9
51.5	54.5	14	3252	6.8
54.5	59.5	15	3047	6.4
59.5	>	16	2976	6.2

THE TOTAL NUMBER OF SAMPLE POINTS... 47802

THE AVERAGE NUMBER OF SAMPLE POINTS ASSIGNED PER GRAY LEVEL... 2987.625

THE STANDARD DEVIATION OF THE NUMBER OF SAMPLE POINTS PER GRAY LEVEL... 680.534

HALF-TONE PATTERN 'H5W4GRAY' WILL BE USED FOR THIS PLOT - THE GRAY SCALE LEVELS FOR THIS PATTERN ARE (FROM 16 TO 1) ...

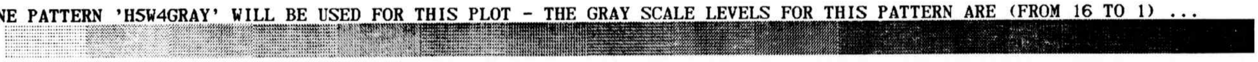


Figure II-6 legend

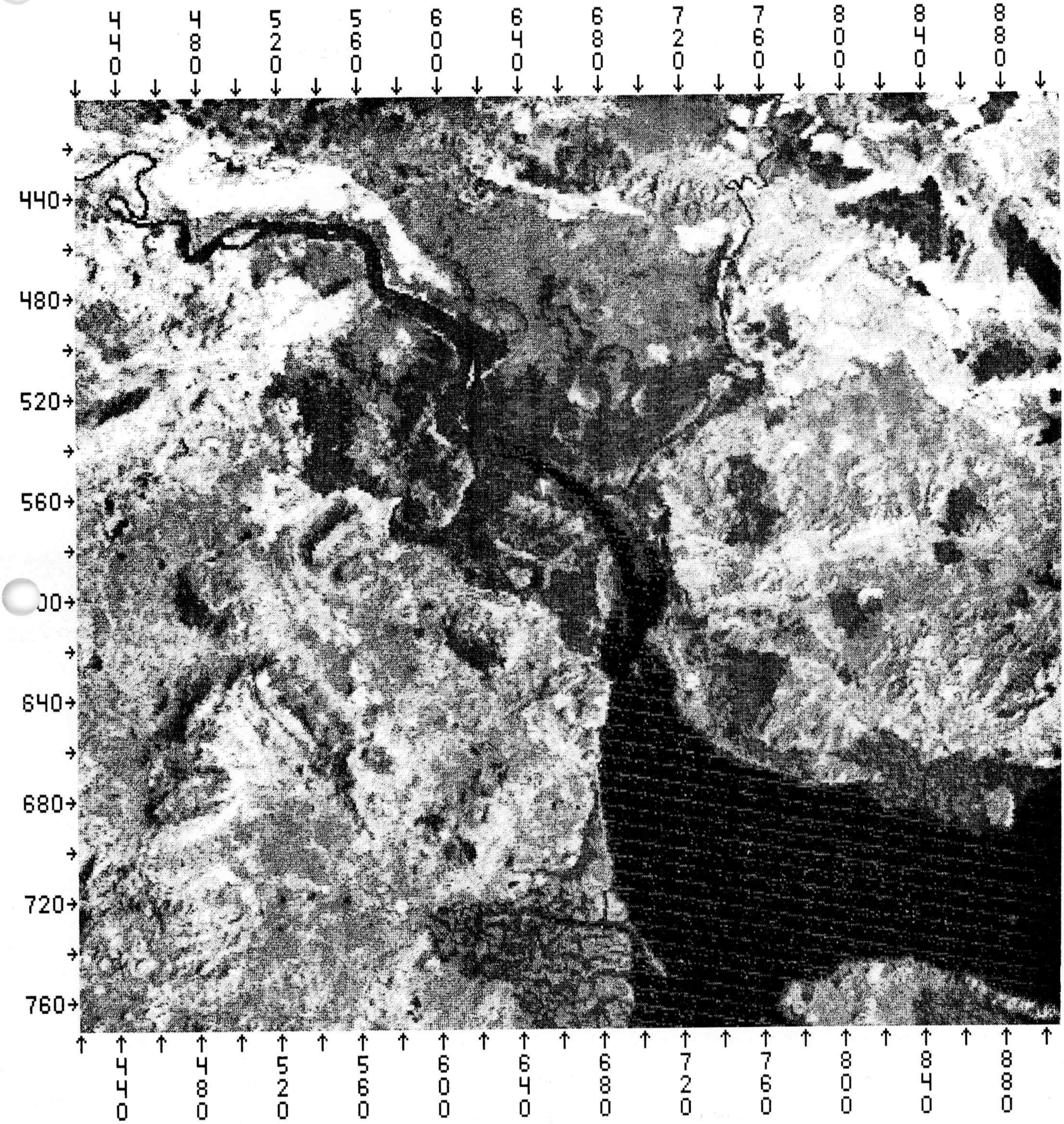


Figure II-7. Gray-scale printout of data in Channel 4 for a portion of the study area.

RUN NUMBER........... 75044401

FLIGHT LINE... 204015163 C.RICA

DATA TAPE/FILE NUMBER.. 3466/ 1

REFORMATTING DATE. SEPT 28,1976

DATE DATA TAKEN... MAR 3,1975

TIME DATA TAKEN..... 1416 HOURS

PLATFORM ALTITUDE..3062000 FEET

GROUND HEADING..... 180 DEGREES

CHANNEL 4 SPECTRAL BAND 0.80 TO 1.10 MICROMETERS CALIBRATION CODE= 1 C0 = .0

HISTOGRAM BLOCK(S)

RUN NUMBER	LINES	COLUMNS	CALIBRATION CODE
75044401	(401, 771, 2)	(420, 933, 2)	1

THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE

LOWER LIMIT	UPPER LIMIT	LEVEL NUMBER	SAMPLE COUNT	PER CENT OF TOTAL SAMPLE
<	0.5	1	4786	10.0
0.5	1.5	2	1277	2.7
1.5	12.5	3	3288	6.9
12.5	15.5	4	2680	5.6
15.5	17.5	5	2384	5.0
17.5	19.5	6	3460	7.2
19.5	20.5	7	2798	5.9
20.5	21.5	8	3670	7.7
21.5	22.5	9	3561	7.4
22.5	23.5	10	3788	7.9
23.5	23.5	11	0	0.0
23.5	24.5	12	3552	7.4
24.5	25.5	13	3229	6.8
25.5	27.5	14	3833	8.0
27.5	29.5	15	2217	4.6
29.5	>	16	3279	6.9

THE TOTAL NUMBER OF SAMPLE POINTS... 47802

THE AVERAGE NUMBER OF SAMPLE POINTS ASSIGNED PER GRAY LEVEL... 2987.625

THE STANDARD DEVIATION OF THE NUMBER OF SAMPLE POINTS PER GRAY LEVEL... 1128.414

HALF-TONE PATTERN 'HSW4GRAY' WILL BE USED FOR THIS PLOT - THE GRAY SCALE LEVELS FOR THIS PATTERN ARE (FROM 16 TO 1) ...

Figure II-7 legend

Figure II-8. Example of an alphanumeric gray scale map of a portion of Channel 4 data; produced by a line printer.

LOCATING A STUDY AREA AND CORRELATING IT WITH REFERENCE DATA

This data set has already been geometrically corrected. You can verify this by noting that the "Ground Heading" (listed by IDPRINT) is 180 degrees. An uncorrected Landsat data set would have a ground heading of about 190 degrees. The data have also been rescaled so that a line printer image of the data would have a 1:25,000 scale, which can be used conveniently with the 1:50,000 topographic maps available for Costa Rica.

The matrix-printer images we looked at earlier were made from this geometrically corrected Landsat data. Looking at the reverse side of these images, we can see that the analyst chose to display only lines 401 through 771 and column 420 through 933, but at an interval of 1. Only this portion was displayed because of the limitations of page size. The actual study area to be used goes from line 401 through 911 and from columns 420 through 933; i.e., 260,000 pixels. This is the area for which maps and aerial photographs are available, about 3.5% of a Landsat frame.

Your instructor will provide you with the following reference data:

a) topographic maps at a scale of 1:50,000 prepared in 1969, 1972, and 1973 by the Instituto Geografico Nacional, San Jose, Costa Rica.
b) black-and-white low-altitude aerial photographs of portions of the study area taken in December 1977 at a scale of approximately 1:20,000.

> Associate the reference data with the matrix printer images. Using the images from Channels 1 and 4, find and outline (with a pen) two or three examples of each cover type of interest -- agriculture, forest, water, bare soil, and mangroves. Some features are more apparent on one image than the other. (Note: do not mark on the Channel 2 and 3 images at this time. They will be needed later.)

SELECTING THE TRAINING SAMPLE

As you will recall, the basic manual describes three approaches to selecting the training sample used to train the classifier: the supervised approach, the unsupervised approach, and a hybrid approach that draws on both. We will use the hybrid approach in this sample case study.

Remember the two rules for selecting training samples: that each training area selected should include more than one cover type, and that every cover type should appear in at least one training area. For the four-channel data we are using, you should select three to six training areas of sizes ranging from 50 lines by 50 columns to 100 lines by 100 columns.

> Select three training areas which are representative of the scene. Use the maps and photographs along with the gray scale images of the scene and the guidelines described earlier. Make sure that every cover type of interest (agriculture, water, bare soil, and forest) is included in at least one of the training areas and that the training areas are distributed throughout the scene.
>
> Outline the areas you select on the Channel 2 gray scale image with a felt tip pen, and note the line and column coordinates of each area in terms of first line, last line, first column and last column. Be able to justify your selection of training areas.

PART III - STATISTICAL DEFINITION OF (SPECTRAL) TRAINING CLASSES

This portion of the case study demonstrates one possible way to define statistically the spectral training classes for the Tempisque Valley study area. We will continue to use the hybrid approach in this analysis.

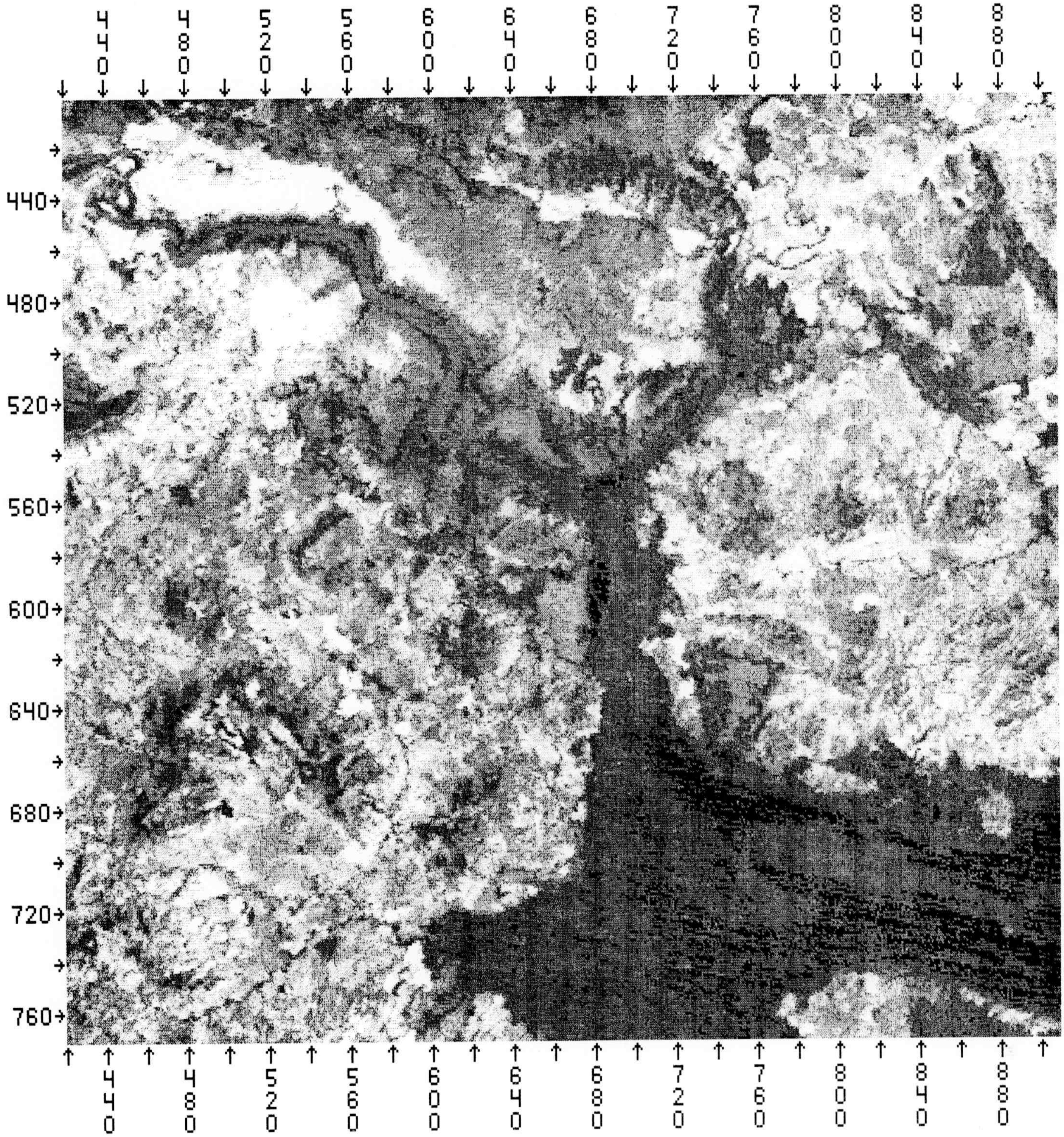
CLUSTERING THE TRAINING SAMPLE

No two analysts would possibly choose the same data samples to use for developing the training statistics. There will be as many different training areas chosen as there are analysts, yet all these areas may be equally valid. In this case study, we have intentionally chosen some less obvious areas as a demonstration of several approaches. The training areas chosen are listed on page 6 of your printout book and are outlined in the line-printer display of the data beginning on page 8. Figure III-17 shows the relative locations of the training sample on one image. (You will recall that the entire study area could not be displayed on this dot-matrix printer output.)

In order to cluster the training data, the CLUSTER processor was run three times; this was necessary because of the 40,000-value limit with this processor as it is implemented in LARSYS. In practice, therefore, when four channels of data are clustered, a maximum of 10,000 pixels may be processed during a single run.

First CLUSTER processing -- the first area clustered is the 100 by 100 rectangle in the upper right of the study area. On page 86 of the computer printouts, the phrase "OPTIONS MAXCLAS(18), CONV(98.5)" indicates first, that we requested that the data submitted be separated into 18 clusters, based on the responses in all four channels, and, second, that the processor continue the iterative cluster assignments until 98.5% of the data samples did not change class membership on repeated iterations. MAXCLAS was set at 18 because, when using reference data, we recognized the presence of at least eight different information classes in the scene: forest, bare soil, standing sugar cane, harvested fields of sugar cane, burned off fields of sugar cane, pasture of at least two types, and other vegetation. It is common practice to ask for two times the number of clusters as there are information classes, plus two or three additional clusters to allow for transition zones between the information classes and for several subclasses in the "other vegetation" class.

The mean values of each cluster class in the first training data and the class variances of each are shown on page 88. Also listed is the number of pixels ("POINTS") assigned to each cluster. In general, the first clusters have higher mean values than the latter ones -- the general trend in the sequencing is from bright


440
480
520
560
600
640
680
720
760
800
840
880
440
480
520
560
600
640
680
720
760

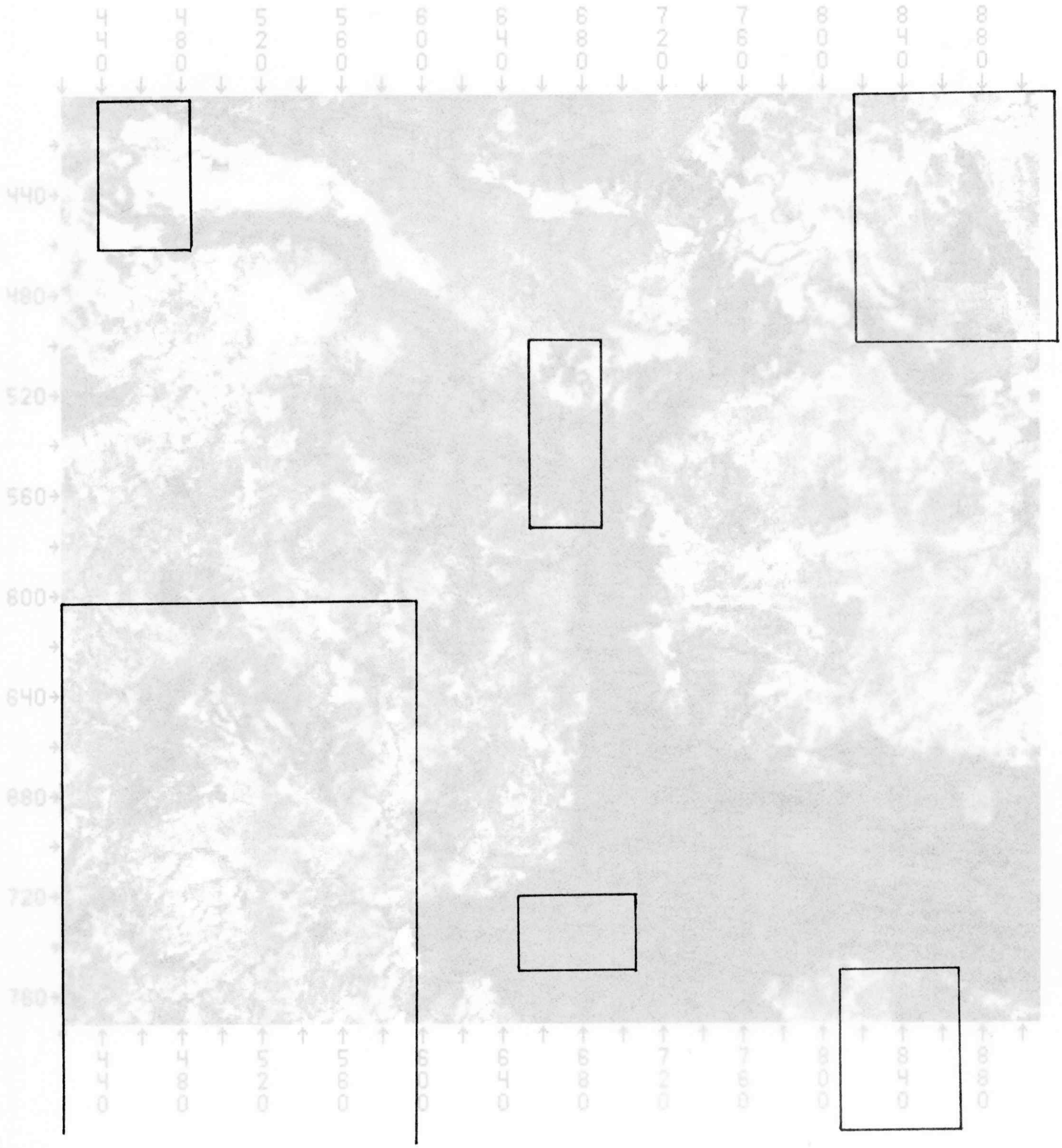


Figure III-17. Training sample outlined as five separate blocks.

(higher mean values) to dark (lower mean values). The variances indicate the spread or dispersion of the data in each channel and serve as a guide to the spectral purity of the class.

The "cluster map" obtained from clustering the data in Area 1 is shown on pages 89 and 90. When you compare the cluster map with the topographic maps and the photography, you can readily determine that Cluster 13 (designated by the symbol G on the map) represents forest. There is a large triangular area of G's in the center, at the top of page 90, that correlates easily to the forest on the aerial photography. The less obvious associations of cluster classes with information classes will be made in the next section.

Pages 91 through 126 show histograms of the data values from each channel for each of the 18 cluster classes, that is, the channel-by-channel distribution of all the data values from those pixels that were grouped into each cluster by the clustering algorithm. Since the classification algorithm we will use later assumes that each training class is represented by a Gaussian (normal) density function, we must check each histogram to see how closely each cluster class can be approximated by a Gaussian density function. Note the non-Gaussian character of the Channel 2 histogram for Cluster Class 1 (page 91). This is a potential problem. If you look at the relative size of the class, shown on page 88, you will note that this class, with its 201 points, is the smallest of all eighteen classes. In fact, it is not uncommon for such a small class to have a bi-modal histogram. While this deviation is not great enough to require reclustering, we will, however, take this and the class size into account later when we make decisions about which candidate training classes will make up our final set of training classes. Note also that the large variance indicated on page 88 for Class 1, Channel 2 corresponds in size to the comparatively wide spread of the data shown by the histogram. Smaller variances are shown for classes with tighter histograms.

<u>Clustering the second training area</u> -- In order to represent the wide variability in the study area in a second 100 by 100 set of data samples (10,000 pixels) we decided to choose four smaller areas which together total less than 10,000 pixels. The clustering processor clusters the data without respect to its location on the ground, and so this approach is completely valid. In this case, MAXCLAS was set at 16, in recognition of approximately seven different information classes; the convergence remained the same, 98.5%. Output begins on page 129.

The printouts give the same kinds of information they did for the first training area: class mean values and variances, numbers of points assigned to each cluster class, and class histograms. Because the data occurred in the scene in four separate rectangles, four separate cluster maps were created, printed on pages 131 through 136.

Clustering the third training area -- The large area in the lower left of the scene contains few landmarks, yet it is generally quite variable. In order to represent the classes in this large area and still stay below the 10,000 pixel limit, we chose an area that was 301 lines by 180 columns, 54,180 pixels, and instructed the clustering processor to consider only the points in every third line and every other column, a little less than 17% of the pixels in the area, or 9090 points. This time MAXCLAS was set at 15.

The cluster map resulting from this run of CLUSTER is very hard to correlate with the reference data because only one of every six points is shown, but in an area lacking landmarks, as this one does, it serves to represent adequately the spectral variety in the area.

Examine the output from the CLUSTER processor for Training Areas 1, 2, and 3, pages 86 through 205.

1. Look at the mean values for each cluster on pages 88, 130, and 172. Using the table on the next page, try to provide a general identification for as many cluster classes as you can (vegetation, bare soil or water) by comparing the relative values in each band to known spectral reflectance characteristics of basic earth surface features. Although the response values shown there have not been calibrated to facilitate band-to-band comparisons, the general trends can still be observed.

2. Examine the variances associated with each cluster and note on the table any unusually high values, i.e., above 30.

3. Briefly check the cluster maps for any obvious groups of symbols that correspond with features appearing on the reference data. Outline these features on the cluster maps and modify your class identifications, if you wish.

4. Examine the cluster histograms. Note any obviously non-Gaussian characteristics. Select a cluster class with low variances and compare its histograms to those from a cluster class with high variances. Note non-Gaussian classes on the table.

5. Note any clusters with fewer than 40 points.

ASSOCIATING CANDIDATE TRAINING CLASSES WITH INFORMATION CLASSES

Now that we have generated a set of candidate training classes by clustering all the training areas, we need to identify which information class is represented by each candidate training class. To make this identification we will use the line-printer "cluster maps," the aerial photography and the 1:50,000 topographic maps.

Remember that the correspondence between information classes and candidate training classes is not necessarily one-to-one. Most often, more than one candidate training class is associated with one information class; occasionally (we hope, rarely) more than one information class is associated with a single candidate training class. When this occurs, dual class names must be given, e.g., "bare soil and emerging crop" or "bare soil and highway."

> Give each cluster class in Training Area 1 an information class name. Use the computer-generated map on pages 89 and 90, along with reference data, and the class mean values to identify, as accurately as possible, each cluster class in the training area. Record your identification on Table III-1, replacing the general identifications you had previously made.
>
> This Landsat data has been geometrically corrected, and that aids in correlating it with the maps and photographs. Remember that the aerial photography was obtained in December 1977; the fact that it is two years newer than the Landsat data is of less importance than the fact that it was obtained some three months earlier in the year. In the photography, harvesting of sugar cane is not so far advanced as it is in the March Landsat data.
>
> If you have time, go on to identify the classes in Training Areas 2 and 3.

AUGMENTING THE CANDIDATE TRAINING CLASSES

Now that we have associated our candidate training classes with information classes, we should consider whether our training sample is representative of all the information classes in the scene. When we look at the data, we can observe that one class in the scene does not occur in any of the training areas, "sea water." We must now use the supervised approach to define a "sea water" training class. The processor that can return the statistics of a single spectral class is the STATISTICS processor.

Table III-1. Student's identification of cluster classes.

Cluster	Symbol	Area 1	Symbol	Area 2	Symbol	Area 3
1	b̸		b̸		b̸	
2	-		-		-	
3	+		=		=	
4	/		)		1	
5	J		7		C	
6	C		I		I	
7	I		S		Z	
8	Z		3		*	
9	3		&		T	
10	&		X		F	
11	T		E		G	
12	F		K		B	
13	G	Forest	N		R	
14	H		R		Q	
15	$		Q		W	
16	O		M			
17	Q					
18	W					

> Using the gray scale images (either Figure II-4 through II-7 or the alphanumeric printout on pages 8 through 78) and the topographic maps, specify the line and column coordinates for a good sample of sea water. When selecting an area, bear in mind the minimum number of pixels needed to estimate the statistical properties of the training sample. (See page III-7.)

As in the selection of training samples earlier, it is very unlikely that two analysts would choose the same area, yet their choices can be equally valid. We chose a 49-pixel area from line 1378 through 1384, column 334 through 340, an area that lies outside the study area. The full-frame printout we had obtained prior to selecting the study area enabled us to select this area of deep sea water. Pixels within the study area may be equally satisfactory, but we chose this area to demonstrate that it is also possible to use pixels from outside the area to be classified. The processor simply extracts the class statistics from whatever data is provided, and those statistics can be used in the same way the statistics from CLUSTER are used.

> Examine the output from the STATISTICS processor, pages 206 through 212. Note the following:
>
> number of pixels
> the class mean in each channel
> the standard deviation (values here are the square root of the variances, which you saw as output from CLUSTER)
> correlation matrix (showing the correlation of the data values between all pairs of channels)
> histograms
> coincident spectral plot

VISUAL REPRESENTATION OF CANDIDATE TRAINING CLASSES

Previously you associated information classes with the cluster classes from the three training areas. So that you can check your decisions and also have names associated with any classes you may not have identified, Table III-2 lists class names provided by scientists familiar with the area. They used a zoom tranfer scope to visually correlate cluster maps with reference data.

Table III-2. Analyst's assignment of information class names to spectral classes. (Greatful acknowledgement is given to Costa Rican scientists who made these identifications: Edwin Cyrus Cyrus, Francisco Loria Brenes, Arnoldo Sota Mora, and Carlos Elizondo Solis.)

Cluster	Symbol	Area 1	Symbol	Area 2	Symbol	Area 3
1	Ƀ	Bare Soil	Ƀ	Bare Soil	Ƀ	Bare Soil Soil 2
2	-	Sugar Cane	-	Dry Grassland	-	Dry Grassland/ Bare Soil
3	+	Sugar Cane	=	Dry Grassland	=	Dry Grassland Soil 3
4	/	Grassland	)	Dry Grassland	1	Dry Grassland
5	J	Grassland	7	Sparse Forest	C	Dry Grassland
6	C	Bare Soil/ Dry Grassland	I	Dry Grassland	I	Charral
7	I	Grassland	S	Charral	Z	Deciduous Forest
8	Z	Dry Grassland	3	Dry Deciduous Forest	*	Brushes
9	3	Dry Grassland/ Eroded Soil	&	Dry Forest	T	Charral
10	&	Dry Grassland/ Charral	X	Deciduous Forest	F	Deciduous Forest
11	T	Dry Grassland/ Charral	E	Mangrove	G	Charral/Dry Deciduous Forest
12	F	Forest 1	K	Mangrove	B	Deciduous Forest
13	G	Forest 2	N	Charral	R	Charral/Dry Deciduous Forest
14	H	Charral/ Deciduous Forest	R	Marsh	Q	Deciduous Forest
15	$	Charral	Q	Water's Edge	W	Mangrove
16	O	Charral/ Deciduous Forest	M	Water		
17	Q	Burned Sugar Cane				
18	W	Marsh/Burned Charral				

We now have a total of 50 candidate training classes, 18 from Training Area 1, 16 from Training Area 2, 15 from Training Area 3, and the "sea water" class. Since it is quite likely that some of the classes may be similar to others, both in spectral response and in identity, we would like to reduce the number of training classes to a set that is representative of the actual scene. We can reduce the number by deleting some candidate training classes and by combining (pooling) others.

To decide which candidate classes may be deleted or pooled, it is useful to be able to visualize the spectral relationships among all the classes at the same time. The MERGESTATISTICS processor has as one of its output products a two-dimensional plot known as a coincident bi-spectral plot. On one axis is plotted the average of the class means in the two infrared channels, and on the other is plotted the average in the two visible channels. The output is a plot providing a visual comparison of the means of all candidate training classes.

Pages 214-215 of your computer printout contain the bi-spectral plot for the 50 candidate training classes. The following two pages list the classes, the symbol used to represent each, the average mean value for each class in the visible and in the infrared, the class number from the original clustering, and the number of points in each class. The first 18 classes may be recognized as coming from Training Area 1, the next 16 from Training Area 2, and the next 15 from Training Area 3. The last class is the one created using a supervised approach.

Because of the large number of candidate training classes, some symbols appear twice on the bi-spectral plot. Using the list of average mean values on pages 216-217, add the subscript 2 to the symbols on the bi-spectral plot for classes 31 through 50. For example, class 31, the second symbol A, has a mean in the visible bands of 24.1 (x-axis) and 30.1 in the infrared bands (y-axis). Locate the A that represents this class and label it A_2. Do the same for classes 32 through 50.

To increase further the usefulness of this plot, identify each class by writing the class name, as shown on Table III-2, next to the class symbols on the table below the plot. Verify that soils tend to fall in the upper right, dense vegetation to the left of the diagonal line, and water in the lower left. You will be using this plot extensively in a future step.

CALCULATING STATISTICAL DISTANCES BETWEEN THE CANDIDATE TRAINING CLASSES

The bi-spectral plot studied in the last section shows which classes are spectrally similar to others. However, since the bi-spectral plot is based only on class means, it doesn't tell us anything about which classes overlap in measurement space or how much they overlap. To make good decisions about which candidate training classes to pool or delete, we must also consider the spread of each class in measurement space and the extent of any overlapping between the classes. We can do this by calculating the transformed divergence value (D_T) between each pair of classes using the SEPARABILITY processor.

Pages 220 through 359 are the printouts from the SEPARABILITY processor for all 50 classes when using all four channels and then the best three classes, the best two, and the best single class. Page 221 shows the symbols used to represent the classes. Beginning on page 224, the transformed divergence is shown for every pair of classes using data in all four channels. For example, on page 224, the transformed divergence between classes A and B is 1986, a relatively high value indicating strong separability between the classes when using all four channels of data. The largest transformed divergence value that is assigned is 2000, corresponding to maximum separability. As you look across the transformed divergence values on pages 224-240, you will notice that many of the values are at 2000 or close to that.

As an aid to locating problems in separating classes, we asked that all class pairs with a transformed divergence value of 1700 or less be listed, and that list appears on pages 241-244. This condensed list will be useful in constructing a separability diagram to serve as the basis for further decisions. Note that again, because there are so many classes, some of the class symbols have been used twice. Their positions in the divergence table indicate whether they are A_1 or A_2, etc.

Before progressing to the next step, it will be helpful to look at another aspect of the SEPARABILITY output. Near the top of page 224, you will note that the minimum transformed divergence between any two classes is 243 and the average is 1896 when all four channels are used. Page 247 shows the transformed divergence values using all combinations of three channels. With only three channels, the highest average value is 1886 and the highest minimum value is 192, indicating that if the classification were to be done using only three channels of data, channels 2, 3 and 4 would be the best to use. The three-class separability study ends with a list of class pairs having a transformed divergence value of less than 1700, and this list shows nine more such pairs than the four-class separability study. As you look on to page 281, you see further deterioration in both the minimum and the average transformed divergence values when two channels of data are used and still further deterioration when a single channel is used, as shown on page 326. It is interesting to note that even when only channel 2 data are used, many class pairs still have a transformed divergence value of 2000, but the list of classes that are not separable (pages 352-356) has grown greatly, and the ability to separate the classes that are difficult to separate is what determines the strength of an analysis.

> Using the list of class pairs with transformed divergence values less than 1700 for four channels of data (pages 241-244), add separability information to your bi-spectral plot (pages 214-215). Do this by drawing a solid line between every pair of classes with a transformed divergence of less than 1500. When the transformed divergence value is greater than 1500, do not draw a line.

One additional way of providing a graphic representation of the 50 candidate training samples is shown in Figures III-18 through III-21. Viewing the class means in this way allows you to confirm the general cover type of each class, that is whether the curve is that of vegetation, water, soil, or other surface materials, and provides yet another way to make decisions about spectral similarity or dissimilarity. It is important that the means here have been calibrated, that is channel-to-channel inconsistencies in sensor response have been mathematically removed. The mean values that appear on the CLUSTER output have not been calibrated and so would not produce accurate spectral graphs.

> Select four or five spectrally dissimilar classes from Figure III-18 and check back to see how they were identified on Table III-2. Those with vegetation labels should follow the characteristic form of the response for vegetation: higher in channel 1 than in channel 2, much higher in channel 3 and still higher in channel 4. The same correlation with known spectral characteristics should be exhibited for all other classes.

REFINING THE SPECTRAL TRAINING CLASSES

As aids in selecting our final training classes, we now have a bi-spectral plot with information class labels and separability information, the histograms for each class (from CLUSTER), and the calibrated spectral means for each class.

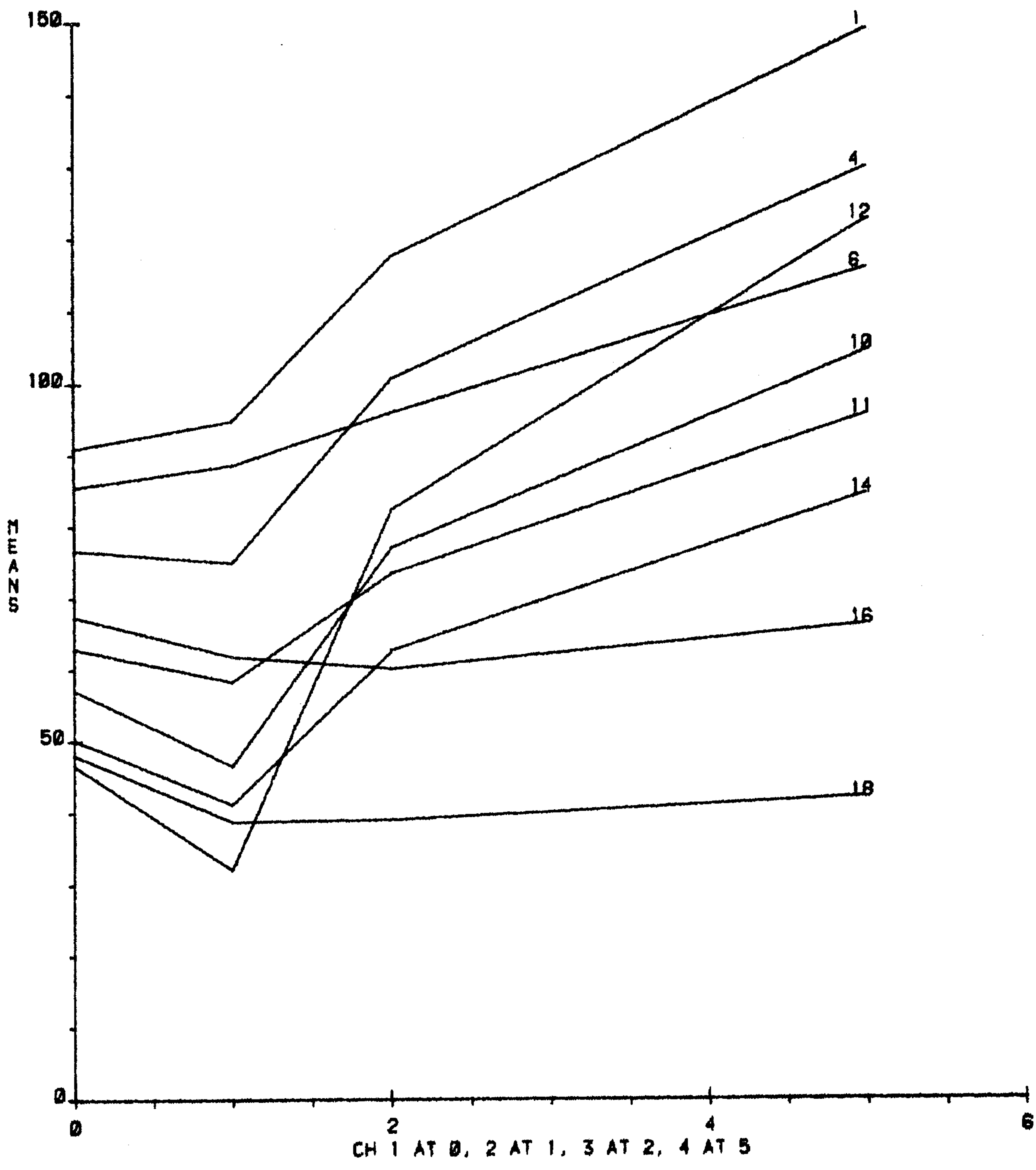


Figure III-18a. Calibrated means for a portion of the candidate training classes from the first clustering, classes 1 through 18.

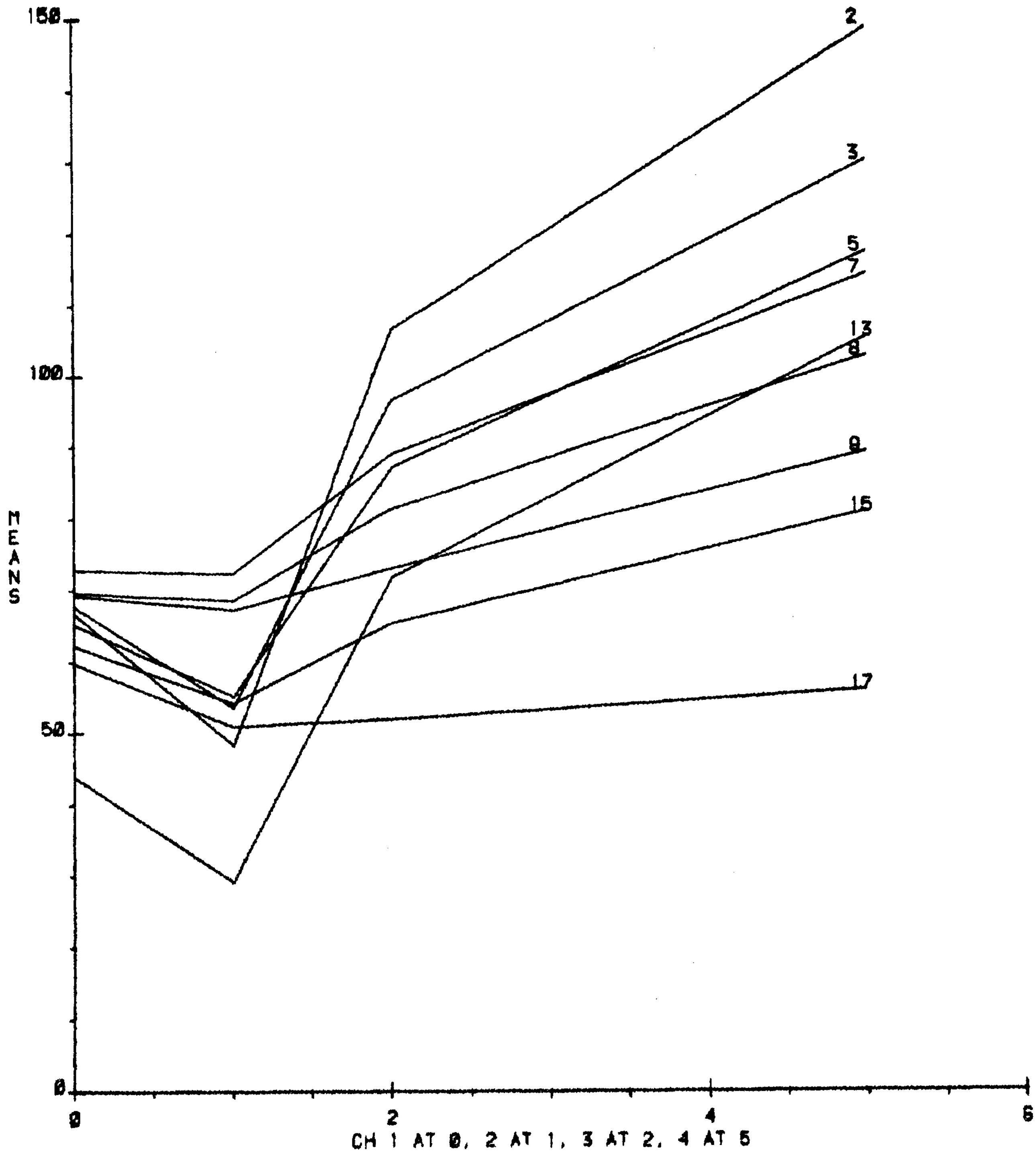


Figure III-18b. Calibrated means for the remaining candidate training classes from the first clustering, classes 1 through 18.

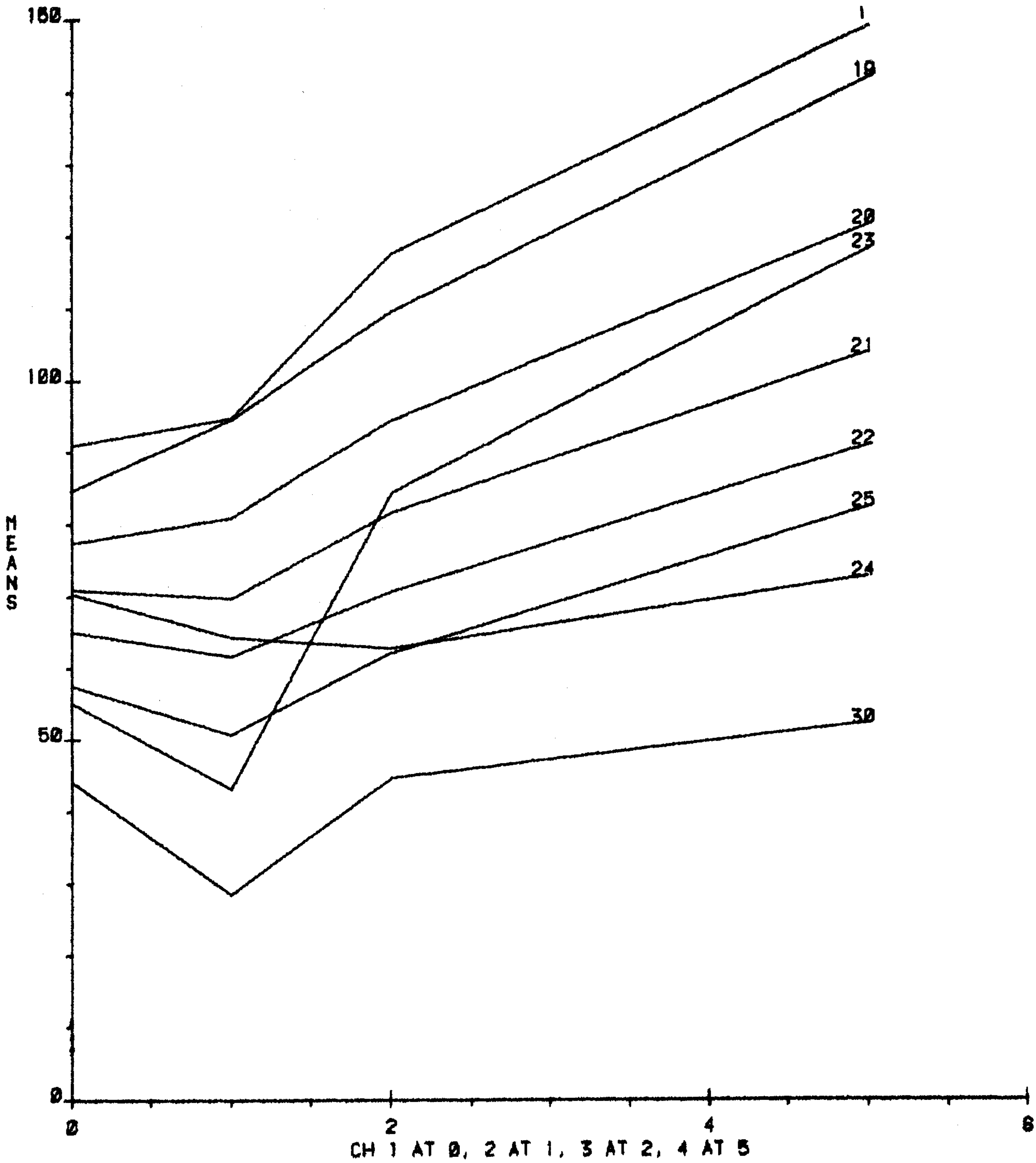


Figure III-19a. Calibrated means for a portion of the candidate training classes from the second clustering, classes 19 through 34. (Class 1 is retained for scaling purposes.)

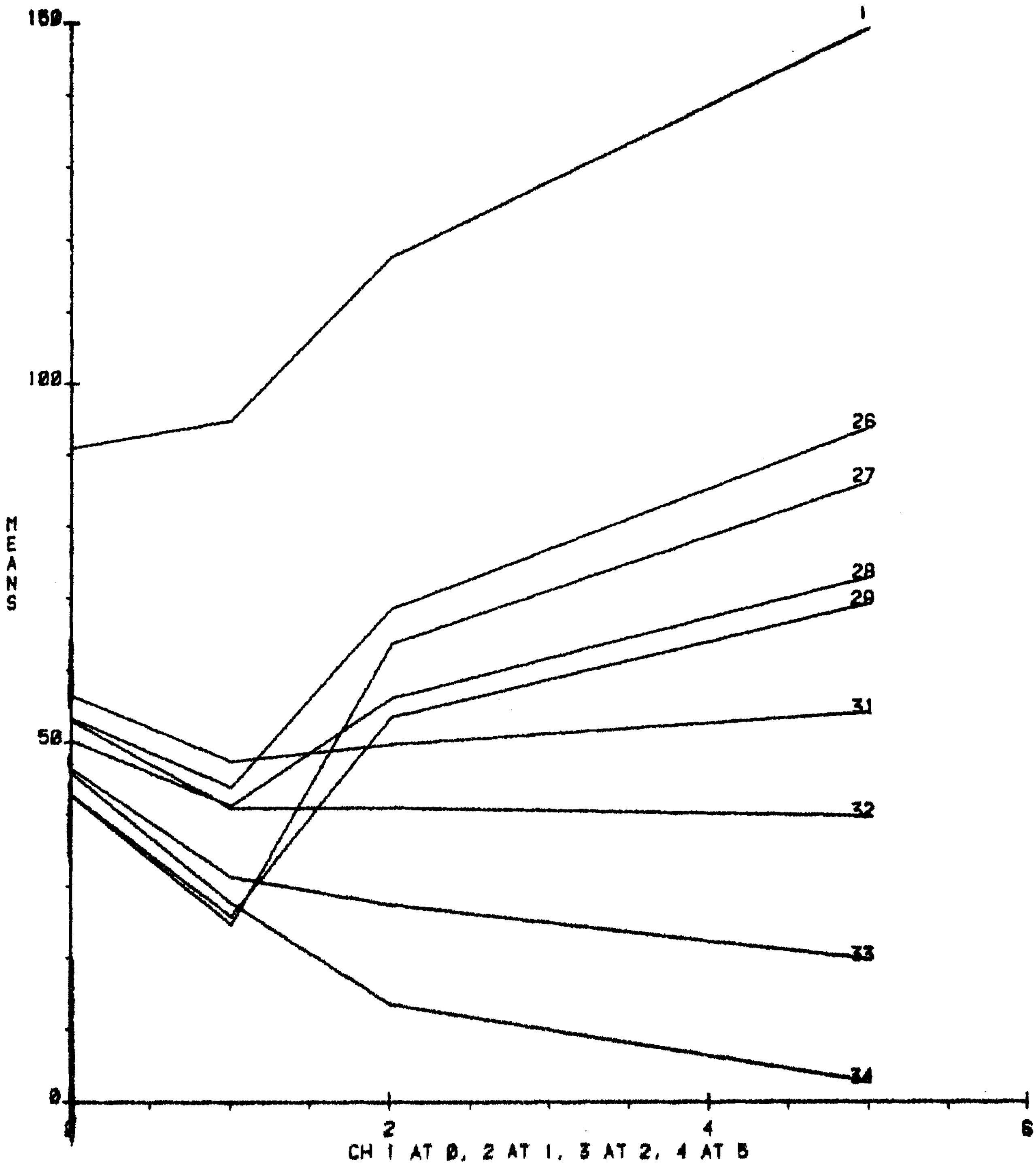


Figure III-19b. Calibrated means for the remaining candidate training classes from the second clustering, classes 19 through 34. (Class 1 is retained for scaling purposes.)

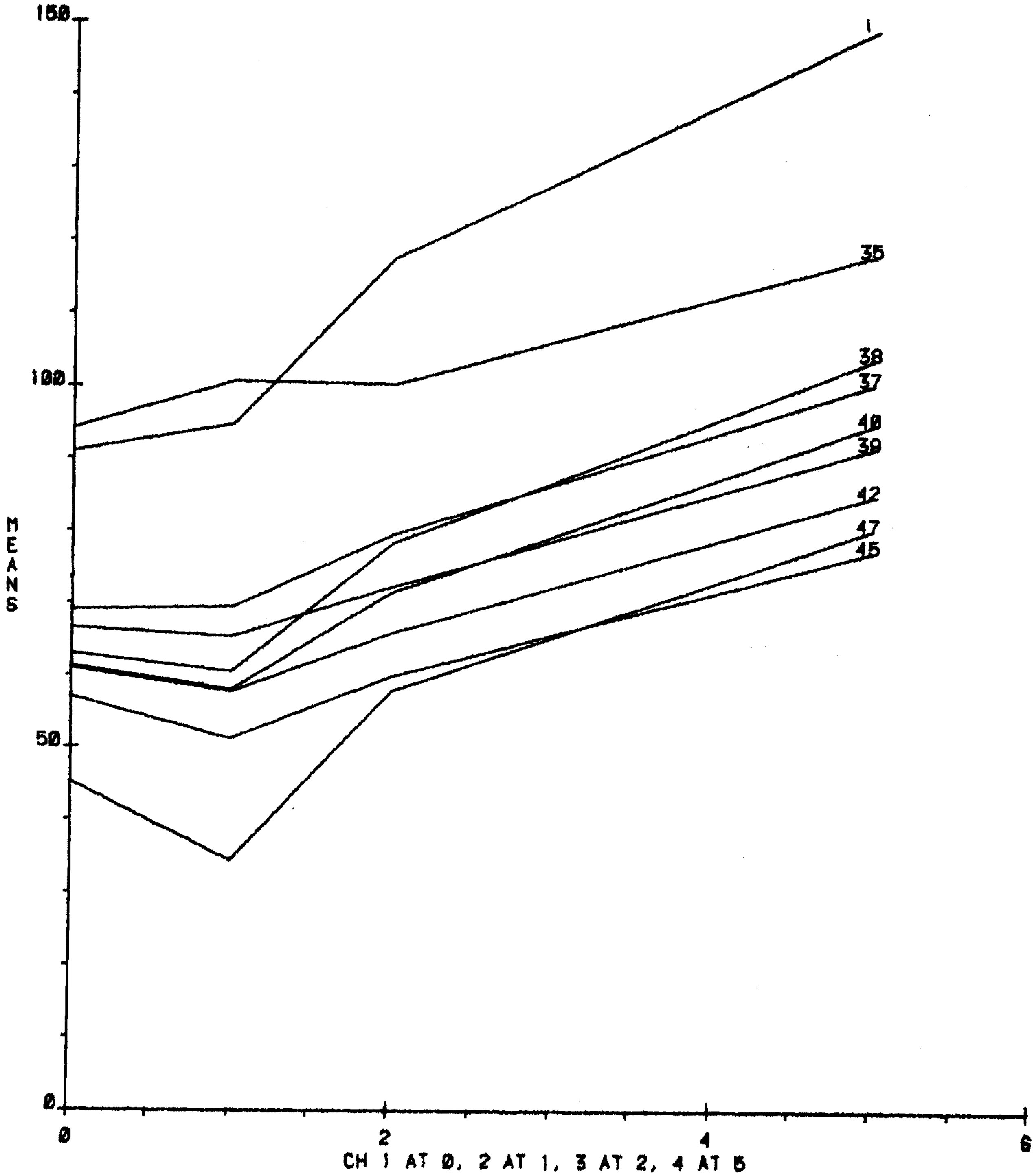


Figure III-20a. Calibrated means for a portion of the candidate training classes from the third clustering, classes 35 through 49. (Class 1 is retained for scaling purposes.)

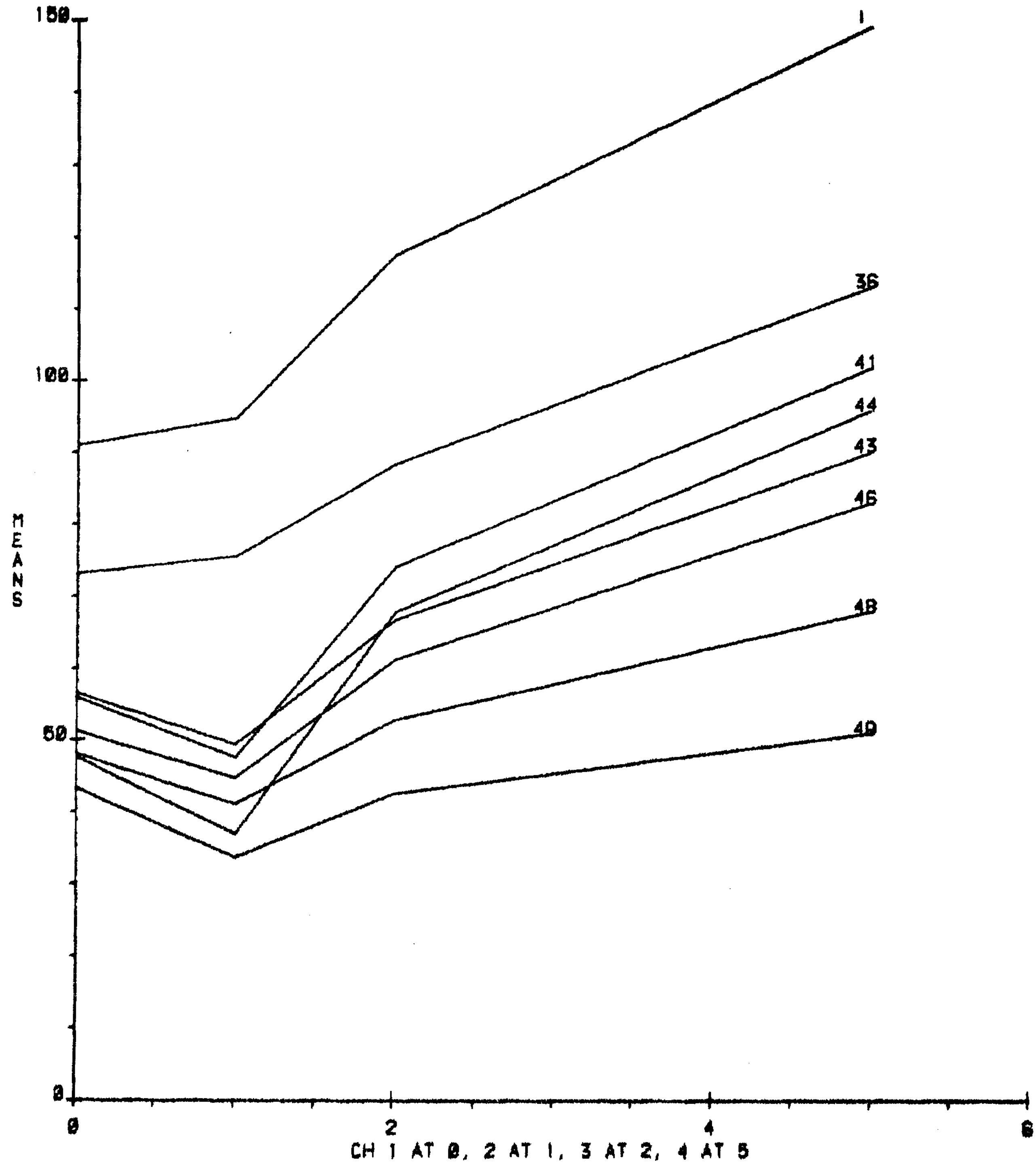


Figure III-20b. Calibrated means for the remaining candidate training classes from the third clustering, classes 35 through 49. (Class 1 is retained for scaling purposes.)

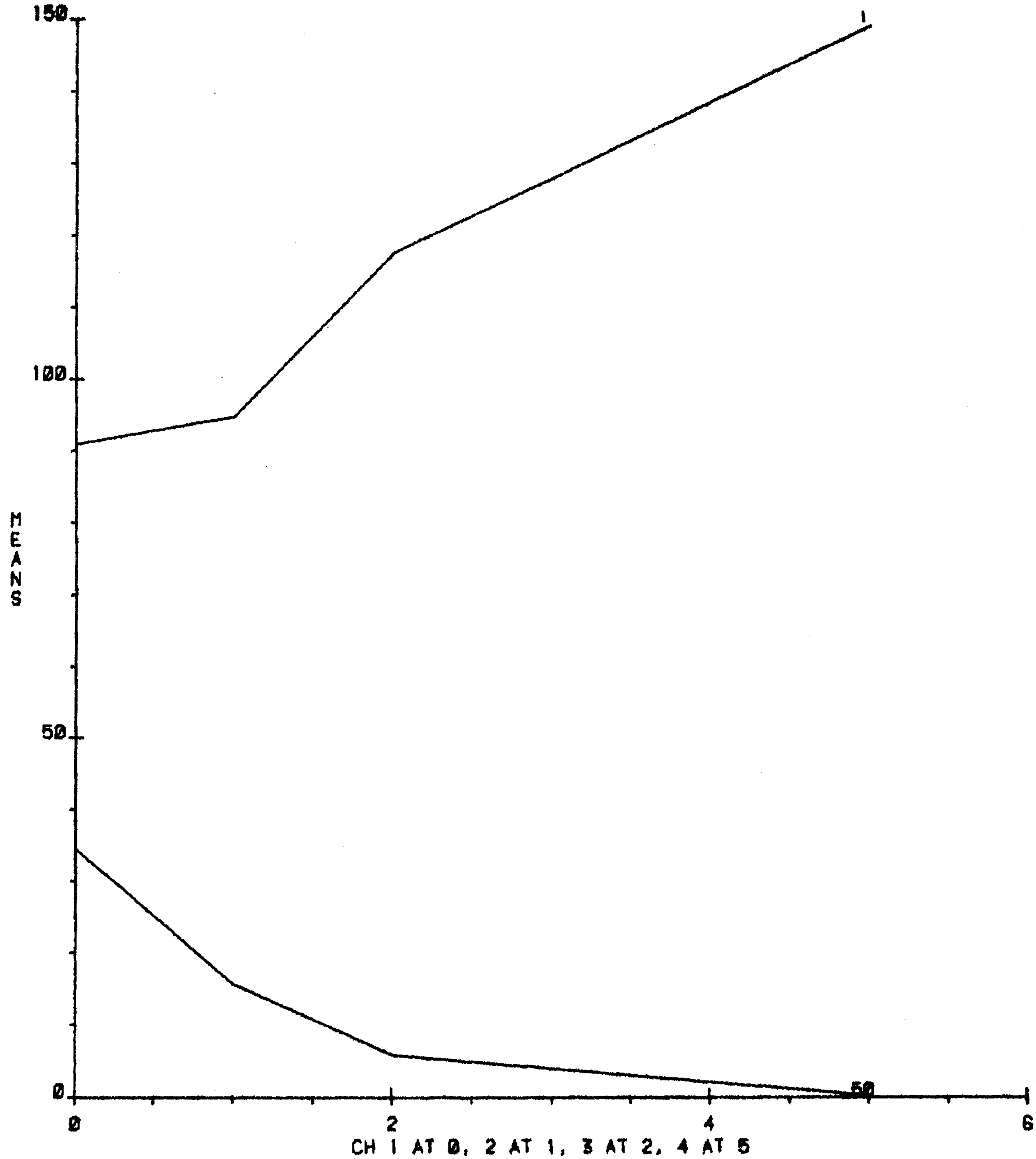


Figure III-21. Calibrated means of candidate training class 50. (Class 1 is retained for scaling purposes.)

We are now ready to select the final training classes and to make decisions about which candidate training classes should be deleted, which should be pooled, and which should be kept as they are. While making these decisions, we need to keep in mind the information class labels, the normality of the histograms, the number of pixels in each candidate training class and the statistical distance between each pair of candidate training classes.

The two major goals we must seek to satisfy when forming our final training classes are:

1. As a group, the final training classes should represent everything in the scene.

2. The training classes should be spectrally separable from each other.

We can make use of two different strategies when refining our training classes: first, the pooling approach combines classes that are spectrally similar, i.e., that have a transformed divergence value that is less than a threshold value set according to the overall accuracy desired; and second, the deleting approach, in which classes that are on the borders between information classes are deleted. The pooling approach is more generally directed toward accomplishing the first goal above, representation, while the second helps achieve the second goal, separability. Experience has shown that the optimal approach is a combination of these two strategies.

Perhaps you noticed that classes of similar materials tend to plot into the same portion of the graph. This, of course, is to be expected since similar materials would have similar spectral responses. As this plot is set up, soils tend to fall in the upper right and water on the lower left, forming an almost diagonal line, with various mixture classes toward the center. The most dense forests fall toward the upper left, with agriculture between forests and soil.

From the solid lines you have drawn on your bi-spectral plot, you may observe that several of the classes are quite separable and have specific identities associated with them. These classes will certainly be retained as training classes: seawater (T_2), water (D_2), water edge (C_2), mangrove (=), Agriculture (B).

Now the more difficult decisions need to be made: what to do with the groups of classes that are not spectrally separable. Look at the three classes indicated by S_2 , R, and B_2 identified as Mangrove, Burned Charral, and Marsh. We know from the transformed divergence values that the classifier can not distinguish between them very well. One useful strategy to use at this point is to keep the center class, R, and delete the other two. Because of the similarity in spectral response, most of the pixels that are members of the S_2 and B_2 classes will be classified into the class represented by the symbol R. Another approach would be to pool classes 18 (R) and 32 (B_2) since they represent two different information classes. The choice of whether to delete or pool is the analyst's. In this case the analyst pooled the two classes. Whichever technique is used, in order to make the class identification more valid, we will broaden the class name of class R to accommodate the cover types we know will be classified into it and call it "marsh and burned charral."

Let's turn our attention now to the upper right portion of the plot, where the soil classes occur. Many of these classes are not separable, and yet when we look at the graph of the calibrated means, we observe that the shapes of the curves are quite different. Compare, for example, the curves for the six classes: class 1 (A), class 4 (D), class 6 (F), class 19 (S), class 20 (%), and class 35 (E_2). Classes 1 and 19 are quite similar, but from the calibrated means we see that class 4 has a very obvious vegetation component. While class 35 is similar in many respects to class 1, there is a cross-over in the responses which could indicate a significantly different type of soil. Classes 6 and 20 are similar to each other, but are not separable from other classes. In fact when we look at the transformed divergence values between 1500 and 1700 for class pairs with F, we will see that F is also not well separated from D, G, or S. Class T we already know to be very similar spectrally to six other classes, that is it has a transformed divergence value of less than 1500 with these classes. There is little need to look further to determine for which classes it would have a value of 1500 to 1700. In this case, the analyst's solution was to keep classes 1 (A), 4 (D), and 35 (E_2) and to delete classes 6 (F), 19 (S), and 20 (T). Even though, as we noted before, the histograms for class 1 in the two visible channels are not Gaussian, the analyst felt that this class better represented the soils of this type than any other classes available.

> Take a close look at the candidate training classes that appear in the "green vegetation" portion of the coincident bi-spectral chart, specifically classes 3 (C), 5 (E), 10 (J), 12 (L), 13 (M), 23 (W), and 41 (K_2). As you can see from your chart, these classes are in turn associated with other classes, but it helps to isolate a group of classes to study in detail. Use the bi-spectral plot, the separability information (transformed divergence values), the graphs of the calibrated class means, and the output from CLUSTER to make your decisions about which classes to keep, which to delete, and which to pool.

After deleting and pooling some of the original 50 candidate training classes, our analyst came up with 18 final training classes. They are listed in Table III-3. Our analyst used the deleting approach more than the pooling approach. Note that every final training class is made up of either one or two candidate training classes.

> Compare your set of training classes to the 18 classes in Table III-3. Did you pool more classes? Did you delete more?
>
> Your final training classes are not necessarily better or worse than our analyst's set. Remember that this step, more than any other in the analysis, is more an art than a science.

Now that we have selected our 18 training classes, we need to evaluate them in order to get an indication of the probability of correct classification which would result from using these training classes. To do this in LARSYS, we can again run the SEPARABILITY processing function. The SEPARABILITY output for this second running of the processor begins on page 366 of the computer printouts. Again the analyst asked for the separability using all four channels as well as the best combinations of three or two channels and the best single channel. Lists of class pairs with transformed divergence values of less than 1800 were requested this time; this allows us to check all new class pairs against a slightly higher threshold. The list based on separability using all four channels of data appears on page 372.

> Add the new separability information and class identifications to the bi-spectral plot of the 18 final training classes, which can be found on pages 361-362 of the computer printouts. Use a dashed line to connect pairs with values between 1500 and 1800. This will indicate class pairs where some confusion may occur.

Since there are no transformed divergence values of less than 1500 for this set of training classes, it is reasonable to expect, on the average, at least an 85% correct classification for the most spectrally similar classes. (Refer to Figure III-12 to see how this estimate was obtained.) The average transformed divergence value is now 1978, and so with all the evidence available in the printouts, we can conclude that this is an appropriate set of training classes and we will adopt them as the final training classes. The final SEPARABILITY output also indicates that separability declines rapidly when fewer than four channels of data are used.

Table III-3. List of 18 final spectral training classes for the Tempisque Valley Case Study.

Pool Name	Cluster Area	Cluster Number	Class in Separability Study	Class Identity
Soil 1	1	1/18	1/50	Bare soil; very dry grass
Agric. 1	1	2/18	2/50	Sugar cane
Agric. 2	1	4/18	4/50	Partially green pasture
Agric. 3	1	5/18	5/50	Partially green pasture
Forest 1	1	12/18	12/50	Sparse forest (green)
Forest 2	1	13/18	13/50	Dense forest (green)
Burned Cane	1	17/18	17/50	Burned sugar cane; marsh
Marsh, Burned Charral	1+2	18/18+14/16	18/50	Marsh; burned thicket; topographic shadow
Soil/Veg.	2	6/16	24/50	Burned sugar cane; moist soil
Mangrove	2	11/16	29/50	Mangrove; topographic shadow
Water edge	2	15/16	33/50	Turbid water
Water	2	16/16	34/50	Water
Soil 2	3	1/15	35/50	Bare soil; very dry grass
Soil 3	3	3/15	37/50	Dry pasture
Charral	3	9/15	43/50	Dry deciduous forest, thicket
Sparse Vegetation 1	3	11/15	45/50	Dry deciduous forest, thicket
Sparse Vegetation 2	3	13/15	47/50	Thicket
Sea Water	-	-	50/50	Sea water

PART IV - CLASSIFICATION OF THE ENTIRE STUDY AREA

The Tempisque Valley study area was classified with the LARSYS CLASSIFYPOINTS processor using the 18 training classes developed in the previous step. This processor is based on the maximum likelihood decision rule. The maximum likelihood discriminant functions for each training class are evaluated for the data values of each pixel. Each pixel is then classified into the class associated with the discriminant function yielding the highest value.

The principal output from this process is the classification results, including the class number that each pixel was assigned to and a probability ranking that indicates the probability that the pixel was correctly classified. These results are stored on the results tape.

The computer output from CLASSIFYPOINTS is on pages 394-415.

> Examine page 396 of the output from CLASSIFYPOINTS. The analyst chose to have the correlation matrix printed out in order to have a record of class statistics.
>
> Study the mean values for the first class, Soil 1. Do they follow the spectral characteristics you would expect a soil class to have?
>
> A perfect correlation, such as one channel with itself, is given a correlation value of 1.00. Between which two different channels does the highest correlation occur? Between which two, the lowest?

Page 414 gives the information that verifies which portion of the data was classified (in lines and columns) and which tape file the results are stored on.

PART V - PICTORIAL AND/OR TABULAR DISPLAY OF THE CLASSIFICATION RESULTS

A classification map, as produced by the LARSYS PRINTRESULTS processor, is shown on pages 418-447 of your computer printouts. The control statements on page 416 show the map symbol we chose for each of the eighteen classes; this information is presented in a more convenient form on page 417, where class names are associated with the symbols. The map that follows is several printout sheets wide. Ask your instructor to show you a map which has been assembled from these sheets.

> Look over the classification map to identify some ground features that you are familiar with. Would these features be easier to locate if different symbols had been chosen? Suppose you were interested only in the locations of the forested land. What symbol set might you have assigned to the eighteen classes?

The classification map can be displayed in several ways. Figure V-5 shows one in which graphic symbols are used to represent each class or group of classes. This map covers a portion of Training Area 1. The actual classification used to create this map is the same as was used to create the alphanumeric printout; there are, however, two differences in the display: 1) similar classes, like Forest 1 and Forest 2, are represented by the same symbol, and 2) the symbols chosen carry with them a suggestion of the class type, making it unnecessary to refer repeatedly to the symbol table.

The classification results are also displayed in tabular format on pages 449-451. The table on page 449 lists the number of points (pixels) that were assigned to each class, the area in acres and hectares that the class covers and the percentage of the entire area.

PART VI - EVALUATION OF THE CLASSIFICATION RESULTS

In the last chapter we looked at the classification map of the Tempisque Valley area. Now we will test our classification to see how accurate it is. We will do this by selecting test fields for each major cover type in the data set (bare soil, agriculture, forest, sparse vegetation and water) and checking to see how accurate our classification is in these test fields.

> Use the gray scale maps (Figures II-5 through II-8), photographs and topo maps to select two test fields for each major cover type (bare soil, agriculture, forest, sparse vegetation and water). The test fields should be distributed throughout the area but should not include any of the training areas. They should also be as large as possible while still being "pure."

How many pixels from each cover type are in the group of test fields you chose? To avoid a bias in the estimate of classification accuracy, we should chose fields so that the number of pixels in each major ground cover category is roughly proportional to the number of pixels in each class in the entire classified area. If we discover a bias, we can choose additional test fields for the ground cover types that are under represented in the original group of test fields.

Another approach to selecting test fields, one that is more likely to yield unbiased estimates, is to select test fields randomly. To use this method, we could divide the entire scene into 3x3 cells and randomly choose 10% of them with a random number table. We would then identify the ground cover type in each test cell, eliminating those test cells that contain more than one major cover type.

Whichever method is used for selecting test fields, the field coordinates and the field type identification are entered into the PRINTRESULTS processor to produce an accuracy/error matrix for the points in those fields. If the test field selection is not biased, these results should represent reasonably well the overall accuracy of the classification.

> Analyze how the pixels in each test field you chose were classified; then fill out the error matrix below.

Field Type	No. of Pixels	Number of Pixels Classified Into Baresoil	Agri.	Forest	Sp.Veg.	Water
1. bare soil						
2. bare soil						
3. agriculture						
4. agriculture						
5. forest						
6. forest						
7. sparse vege.						
8. sparse vege.						
9. water						
10. water						

Study the matrix to see if you can account for any gross examples of misclassification. Are there any changes you would make in training class refinement, in class identification, or in any other aspect of the analysis to improve the classification?