LARS Publication 010582

Monroe County Case Study

for use with

Flexible Workshop on
Numerical Analysis of
Multispectral Image Data

by
James C. Tilton

Page ii is blank

TABLE OF CONTENTS

Preface.....v

Chapter I. Introduction.....I-1

Monroe County Case Study - Part I.....I-5

Chapter II. MSS Data Selection, Correlation with Reference Data and Training Sample Selection.....II-1

Selecting the Data Set.....II-2
Locating a Study Area and Correlating it with Reference Data.....II-5
Selection of the Training Sample.....II-7

Monroe County Case Study - Part II.....II-11

Selecting the Data Set.....II-11
Locating a Study Area and Correlating it with Reference Data.....II-22
Selection of the Training Sample.....II-25

Chapter III. Statistical Definition of (Spectral) Training Classes.....III-1

Clustering the Training Sample.....III-2
Associating the Candidate Training Classes with Information Classes.....III-7
Augmenting the Candidate Training Classes.....III-9
Visual Representation of Candidate Training Classes.....III-11
Calculating Statistical Distances Between the Candidate Training Classes.....III-13
Refining the Spectral Training Classes.....III-16

Monroe County Case Study - Part III.....III-27

Clustering the Training Sample.....III-27
Associating the Candidate Training Classes with Information Classes.....III-29
Augmenting the Candidate Training Classes.....III-30
Visual Representation of Candidate Training Classes.....III-35
Calculating Statistical Distances Between the Candidate Training Classes.....III-39
Refining the Spectral Training Classes.....III-40

Chapter IV. Classification of the Entire Study Area.....IV-1

Monroe County Case Study - Part IV.....IV-5

Chapter V. Pictorial and/or Tabular Display of the Classification Results.....V-1

Monroe County Case Study - Part V.....V-7

Chapter VI. Evaluation of the Classification Results.....VI-1

Monroe County Case Study - Part VI.....VI-5

Chapter VII. Closing Remarks VII-1

Pages I-1 to I-4 not in file

MONROE COUNTY CASE STUDY* - PART I

The case study analysis featured in this workshop is a LARSYS-based analysis of Monroe County, Indiana. Monroe County contains a small city (Bloomington, Indiana), a large reservoir (the Monroe Reservoir) and large areas of forested and agricultural lands. A county map is provided in Figure I-2. We would like to make a land use map of Monroe County distinguishing among four major cover types present: agriculture, forest, urban and water.

One way to obtain information about a multispectral scanner (MSS) data set with the LARSYS system is to use a program called IDPRINT. IDPRINT prints the identification record from a multispectral image storage tape.

Study the output from IDPRINT (see pages 1 and 2 of the computer printouts) and note:

1. the Run Number for this data set. Each data set has a unique Run Number.

2. the Date and Time on which 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 Bands for each of the four channels of data. Which portion of the spectrum do these bands fall in?

5. that Wavelength Bands 4,5,6,7 (as identified by the EROS Data Center) are now referred to as Channels 1,2,3,4.

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

* This case study description was authored by James C. Tilton based on materials presented in Workshop Series on Numerical Analysis of Remotely Sensed Data by Ronald K. Boyd and John C. Lindenlaub and on workshop notes provided by Joan S. Buis. Contributions to the case study analysis were made by Luis A. Bartolucci, Michael D. Fleming and Joan S. Buis.

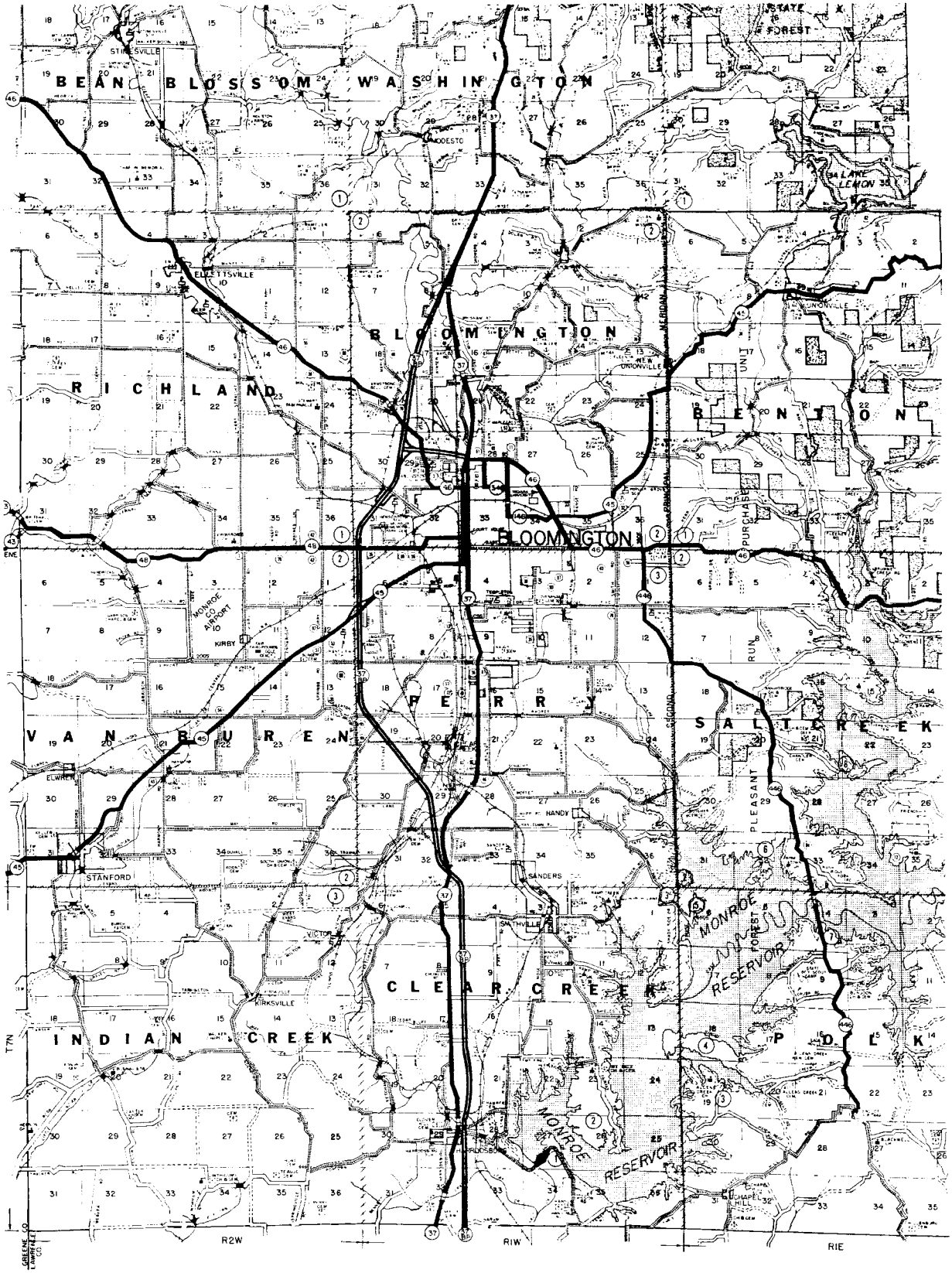


Figure I-2. Monroe County, Indiana.

under C0 and C1 can still be used to convert the data into radiometric units (in this case C0 happens to be 0.0 for each band).

We can print the numerical values associated with specific locations (pixels) in the scene by using the LARSYS program TRANSFERDATA. Pages 4, 5 and 6 of the computer listings show the output of TRANSFERDATA for three selected pixels. Notice that each pixel is identified by its line and column address and has four data values (a data vector) associated with it. The data values indicate the amount of energy returned or the brightness of that spot on the earth's surface as measured by the scanner system in each wavelength band. Data values for Channels 1, 2 and 3 of the Landsat satellite multispectral scanner have a possible range from 0 to 127. Zero is dark, signifying that no energy is being returned, and 127 is bright, indicating saturation of the scanner detector. Data values for Channel 4 may range from 0 to 63. The three data vectors printed by TRANSFERDATA are representative of the many thousands of data vectors within the scene. Later, when the computer classifies the data, it will make the classification based upon these numerical values.

The data vector associated with a known pixel of forest is shown below:

		CHANNELS			
		1	2	3	4
LINE	COL	DATA			
115	326	26.0	15.0	58.0	36.0

Compare these to the values shown on pages 4, 5 and 6 of the printouts. Which page shows data values which probably also represent forest? How did you decide? If you were only allowed to use Channel 3, could you have decided as easily? How about using just Channels 2 and 3 together?

MONROE COUNTY CASE STUDY - PART II

This portion of the case study demonstrates one possible way in which the first step in the analysis of MSS data might be accomplished. We will use a Landsat data set collected on June 9, 1973.

SELECTING THE DATA SET

The first thing we want to do with our data set is to examine it in order to assess data quality and cloud coverage. For this purpose, printer-plotter images from all four bands were generated and are included in this manual as Figures II-4 through II-7.

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

The process used to generate the gray scale maps is level slicing and contrast stretching. If you look on the back of the Channel 3 image, for example, will see a table labeled "THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE." The second line in that table indicates all pixels in the area plotted with data values between 24.5 and 50.5 in Channel 3 are represented by gray level 2. Level 2 is the second block from the right end of the gray bar at the bottom of the page. The remainder of the table specifies the limits for determining which gray tone is used to represent any given pixel value. These limits are calculated so as to fully utilize the 16 gray scale levels available from the plotter.

The range limits are calculated on the basis of the response in Channel 3 of all the pixels specified in the histogram block, shown just above the gray level table. The histogram block describes by line and column the group of pixels used to determine the levels. Note that in this case the pixels in the total area displayed are to be used, except with an interval of 2 (i.e., every other line and every other column). Thus one-fourth of the pixels displayed were used to determine how all of them should be displayed. Such a systematic sample of the MSS data is often used to determine how the entire image is to be displayed, saving valuable computer time in the process. The interval to be

Figures II-4 through II-7 (following eight pages). Printer-plotter images of the Monroe County, Indiana, Landsat MSS data. The channel number of each image is given in the table on the reverse of each image along with other pertinent information.

RUN NUMBER..... 73033802 DATE DATA TAKEN... JUNE 9,1973

FLIGHT LINE... 132115595 IN TIME DATA TAKEN..... 0959 HOURS

DATA TAPE/FILE NUMBER.. 445/ 1 PLATFORM ALTITUDE..3062000 FEET

REFORMATTING DATE. FEB 8,1974 GROUND HEADING..... 180 DEGREES

CHANNEL 1 SPECTRAL BAND 0.50 TO 0.60 MICROMETERS CALIBRATION CODE= 1 C0 = 0.0

HISTOGRAM BLOCK(S)

RUN NUMBER	LINES	COLUMNS	CALIBRATION CODE
73033802	(30, 400, 2)	(112, 474, 2)	1

THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE

LOWER LIMIT	UPPER LIMIT	LEVEL NUMBER	SAMPLE COUNT	PER CENT OF TOTAL SAMPLE
<	26.5	1	1933	5.7
26.5	26.5	2	0	0.0
26.5	26.5	3	0	0.0
26.5	28.5	4	8852	26.1
28.5	28.5	5	0	0.0
28.5	28.5	6	0	0.0
28.5	30.5	7	6061	17.9
30.5	30.5	8	0	0.0
30.5	32.5	9	3668	10.8
32.5	32.5	10	0	0.0
32.5	32.5	11	0	0.0
32.5	34.5	12	6511	19.2
34.5	34.5	13	0	0.0
34.5	38.5	14	3718	11.0
38.5	42.5	15	1307	3.9
42.5	>	16	1802	5.3

THE TOTAL NUMBER OF SAMPLE POINTS... 33852

THE AVERAGE NUMBER OF SAMPLE POINTS ASSIGNED PER GRAY LEVEL... 2115.750

THE STANDARD DEVIATION OF THE NUMBER OF SAMPLE POINTS PER GRAY LEVEL... 2849.958

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

Figure D-4 Legend

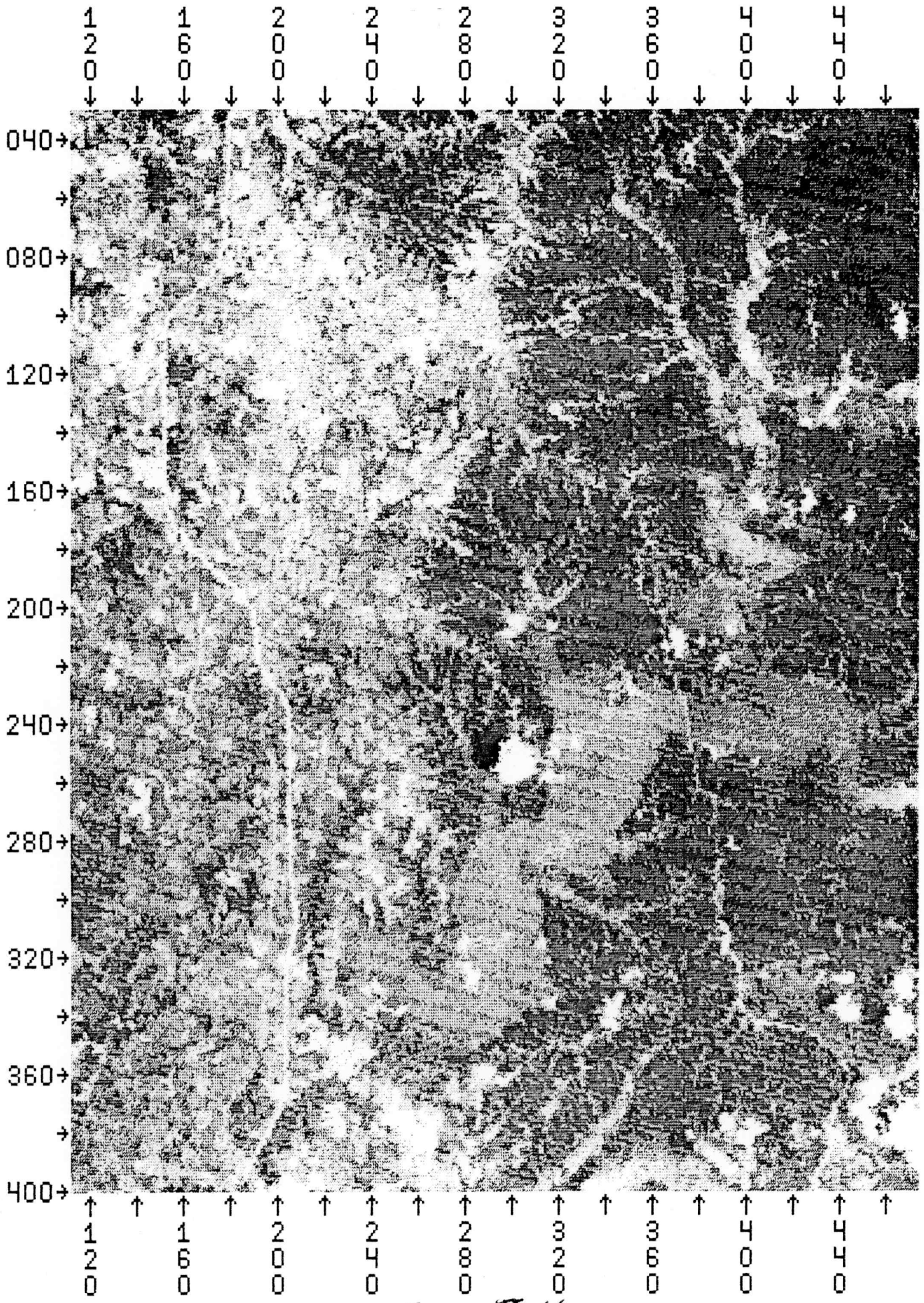

120
160
200
240
280
320
360
400
440
040
080
120
160
200
240
280
320
360
400

Figure II-4

RUN NUMBER..... 73033802 DATE DATA TAKEN... JUNE 9,1973

FLIGHT LINE... 132115595 IN TIME DATA TAKEN..... 0959 HOURS

DATA TAPE/FILE NUMBER.. 445/ 1 PLATFORM ALTITUDE..3062000 FEET

REFORMATTING DATE. FEB 8,1974 GROUND HEADING..... 180 DEGREES

CHANNEL 2 SPECTRAL BAND 0.60 TO 0.70 MICROMETERS CALIBRATION CODE= 1 C0 = 0.0

HISTOGRAM BLOCK(S)

RUN NUMBER	LINES	COLUMNS	CALIBRATION CODE
73033802	(30, 400, 2)	(112, 474, 2)	1

THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE

LOWER LIMIT	UPPER LIMIT	LEVEL NUMBER	SAMPLE COUNT	PER CENT OF TOTAL SAMPLE
<	14.5	1	285	0.8
14.5	16.5	2	7869	23.2
16.5	16.5	3	0	0.0
16.5	16.5	4	0	0.0
16.5	16.5	5	0	0.0
16.5	18.5	6	5775	17.1
18.5	18.5	7	0	0.0
18.5	20.5	8	2422	7.2
20.5	20.5	9	0	0.0
20.5	22.5	10	5724	16.9
22.5	22.5	11	0	0.0
22.5	24.5	12	3141	9.3
24.5	26.5	13	1465	4.3
26.5	30.5	14	2786	8.2
30.5	38.5	15	2344	6.9
38.5	>	16	2041	6.0

THE TOTAL NUMBER OF SAMPLE POINTS... 33852

THE AVERAGE NUMBER OF SAMPLE POINTS ASSIGNED PER GRAY LEVEL... 2115.750

THE STANDARD DEVIATION OF THE NUMBER OF SAMPLE POINTS PER GRAY LEVEL... 2473.791

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

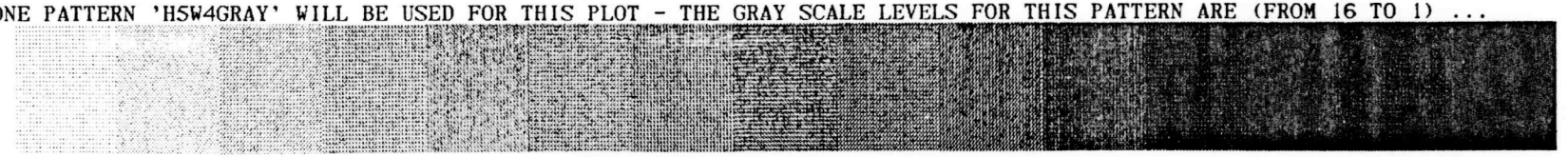


Figure 2 Legend

II-5

II-15

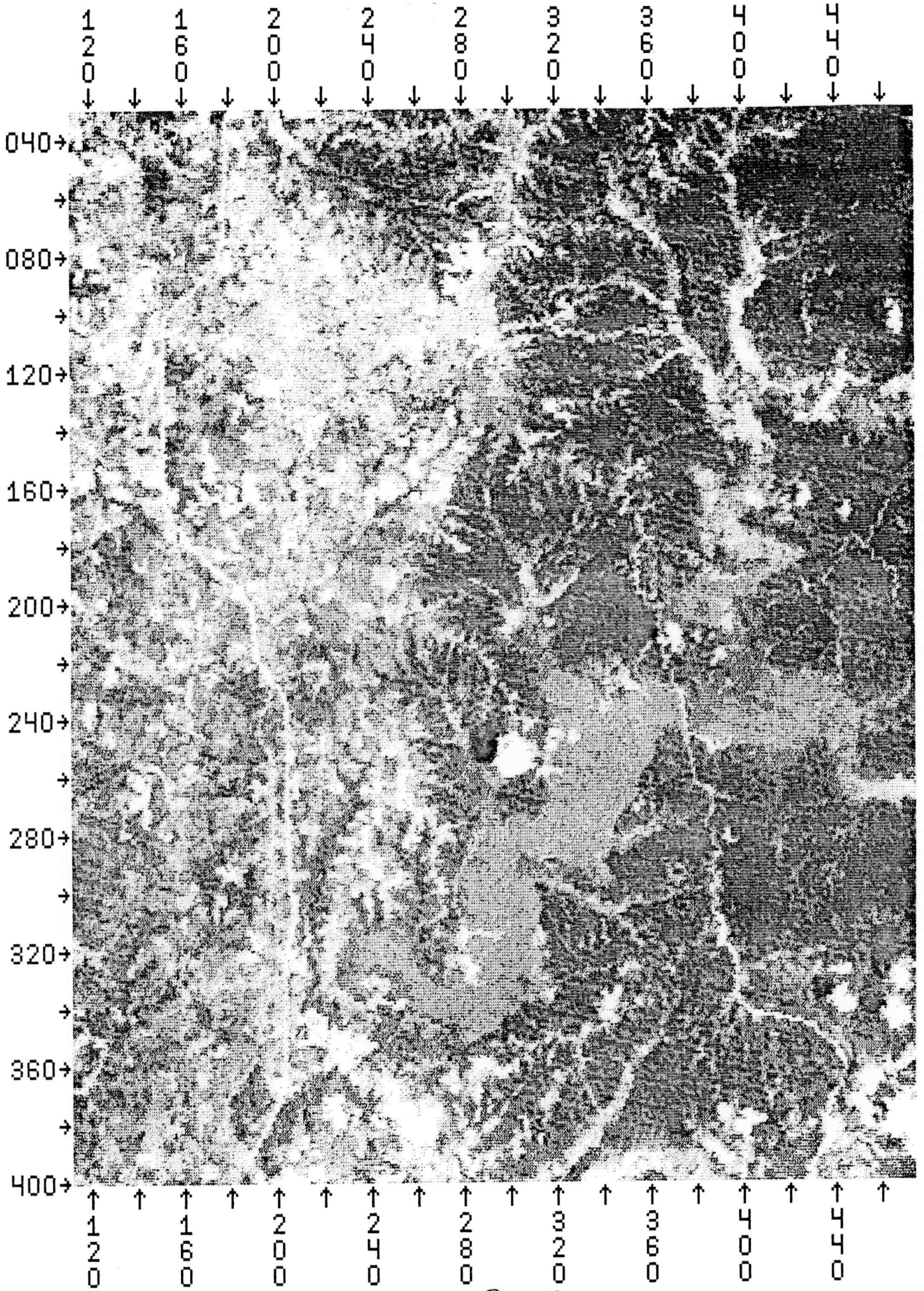

120
160
200
240
280
320
360
400
440
040
080
120
160
200
240
280
320
360
400

Figure II-5

RUN NUMBER..... 73033802 DATE DATA TAKEN... JUNE 9,1973

FLIGHT LINE... 132115595 IN TIME DATA TAKEN..... 0959 HOURS

DATA TAPE/FILE NUMBER.. 445/ 1 PLATFORM ALTITUDE..3062000 FEET

REFORMATTING DATE. FEB 8,1974 GROUND HEADING..... 180 DEGREES

CHANNEL 3 SPECTRAL BAND 0.70 TO 0.80 MICROMETERS CALIBRATION CODE= 1 C0 = 0.0

HISTOGRAM BLOCK(S)

RUN NUMBER	LINES	COLUMNS	CALIBRATION CODE
73033802	(30, 400, 2)	(112, 474, 2)	1

THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE

LOWER LIMIT	UPPER LIMIT	LEVEL NUMBER	SAMPLE COUNT	PER CENT OF TOTAL SAMPLE
<	24.5	1	2131	6.3
24.5	50.5	2	2404	7.1
50.5	52.5	3	1306	3.9
52.5	54.5	4	2531	7.5
54.5	56.5	5	4398	13.0
56.5	56.5	6	0	0.0
56.5	56.5	7	0	0.0
56.5	58.5	8	4393	13.0
58.5	58.5	9	0	0.0
58.5	60.5	10	4244	12.5
60.5	60.5	11	0	0.0
60.5	62.5	12	5535	16.4
62.5	62.5	13	0	0.0
62.5	64.5	14	2973	8.8
64.5	66.5	15	2401	7.1
66.5	>	16	1536	4.5

THE TOTAL NUMBER OF SAMPLE POINTS... 33852

THE AVERAGE NUMBER OF SAMPLE POINTS ASSIGNED PER GRAY LEVEL... 2115.750

THE STANDARD DEVIATION OF THE NUMBER OF SAMPLE POINTS PER GRAY LEVEL... 1845.010

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

II-17

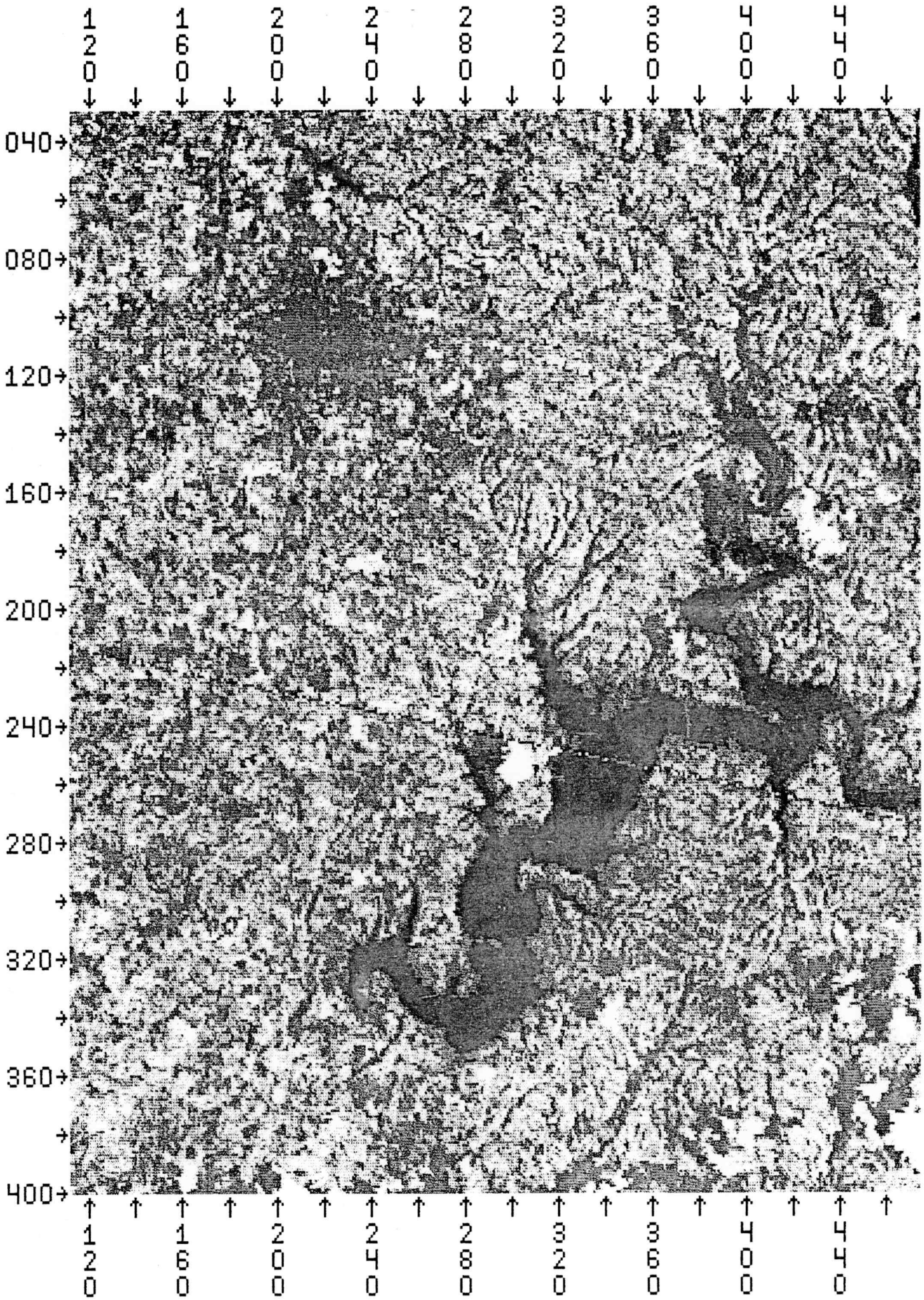

120
160
200
240
280
320
360
400
440
040
080
120
160
200
240
280
320
360
400

Figure II-6

RUN NUMBER..... 73033802 DATE DATA TAKEN... JUNE 9,1973

FLIGHT LINE... 132115595 IN TIME DATA TAKEN..... 0959 HOURS

DATA TAPE/FILE NUMBER.. 445/ 1 PLATFORM ALTITUDE..3062000 FEET

REFORMATTING DATE. FEB 8,1974 GROUND HEADING..... 180 DEGREES

CHANNEL 4 SPECTRAL BAND 0.80 TO 1.10 MICROMETERS CALIBRATION CODE= 1 C0 = 0.0

HISTOGRAM BLOCK(S)

RUN NUMBER	LINES	COLUMNS	CALIBRATION CODE
73033802	(30, 400, 2)	(112, 474, 2)	1

THE DATA RANGES ASSIGNED TO THE GRAY LEVELS ARE

LOWER LIMIT	UPPER LIMIT	LEVEL NUMBER	SAMPLE COUNT	PER CENT OF TOTAL SAMPLE
<	10.5	1	2155	6.4
10.5	26.5	2	2715	8.0
26.5	26.5	3	0	0.0
26.5	30.5	4	4907	14.5
30.5	30.5	5	0	0.0
30.5	30.5	6	0	0.0
30.5	34.5	7	9340	27.6
34.5	34.5	8	0	0.0
34.5	34.5	9	0	0.0
34.5	34.5	10	0	0.0
34.5	34.5	11	0	0.0
34.5	38.5	12	11088	32.8
38.5	38.5	13	0	0.0
38.5	38.5	14	0	0.0
38.5	38.5	15	0	0.0
38.5	>	16	3647	10.8

THE TOTAL NUMBER OF SAMPLE POINTS... 33852

THE AVERAGE NUMBER OF SAMPLE POINTS ASSIGNED PER GRAY LEVEL... 2115.750

THE STANDARD DEVIATION OF THE NUMBER OF SAMPLE POINTS PER GRAY LEVEL... 3539.787

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

Figure II-7 legend

II-19

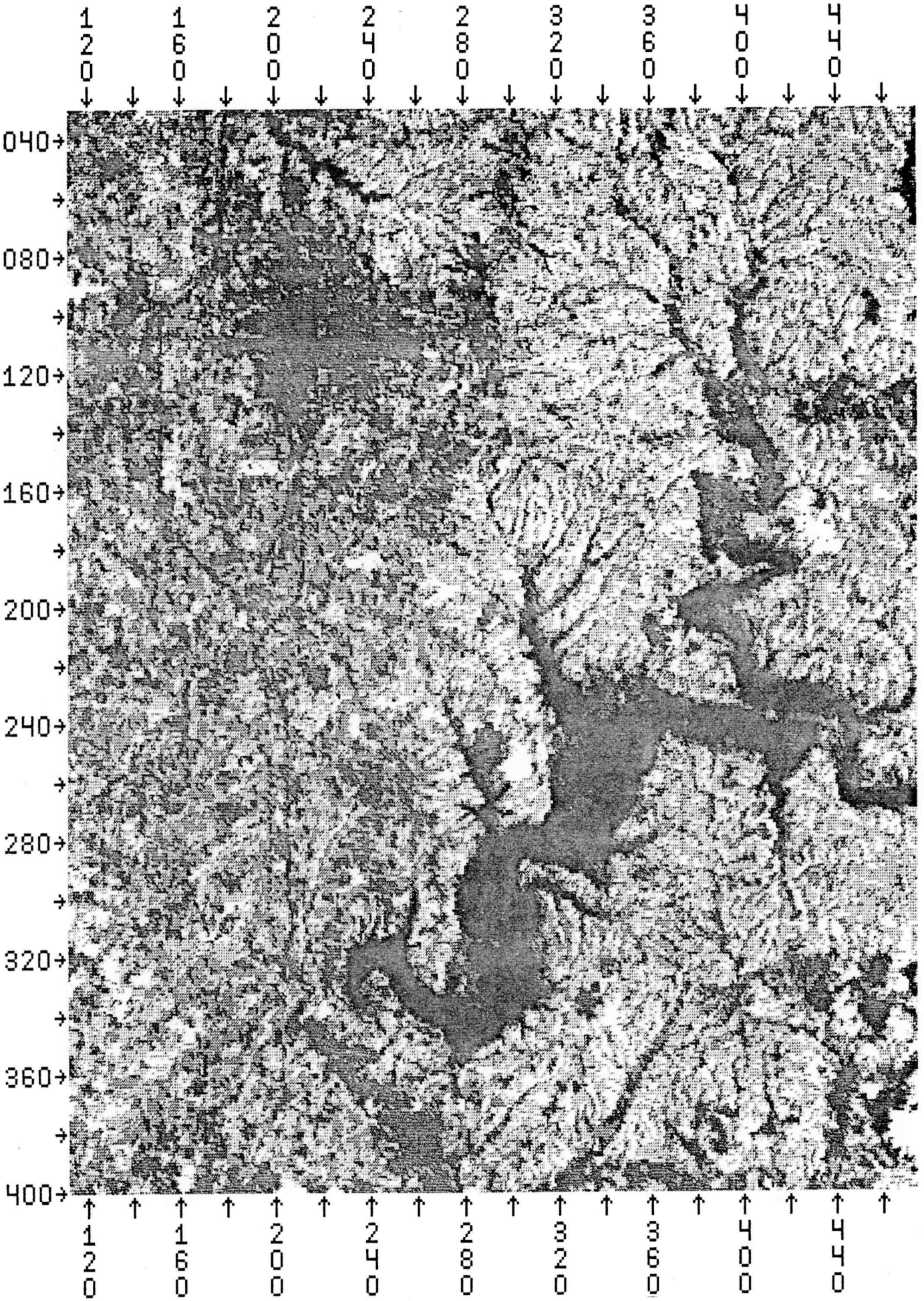

120
160
200
240
280
320
360
400
440
040
080
120
160
200
240
280
320
360
400

Figure II-7

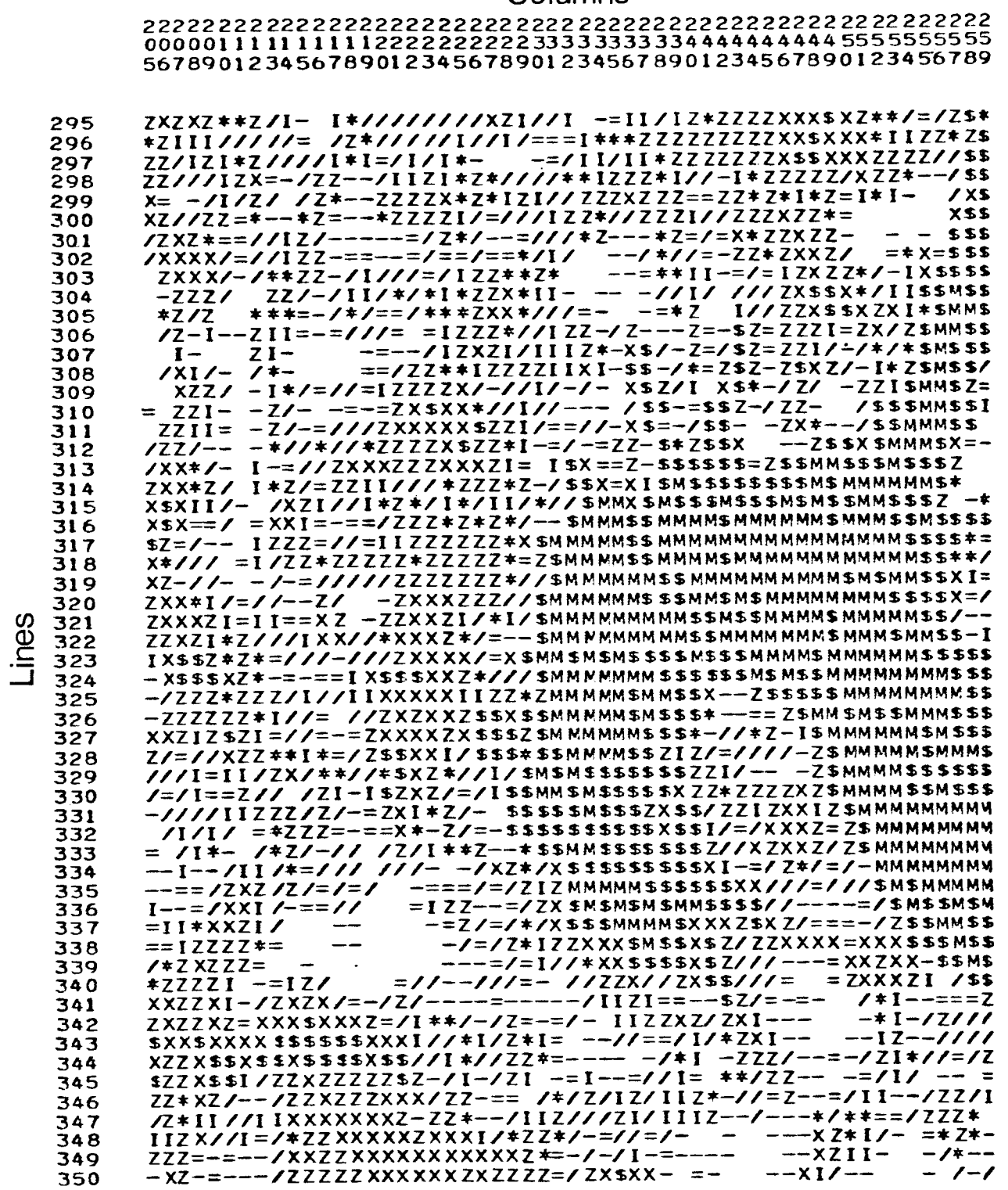


Figure II-8. Gray scale map of a portion of the Monroe County, Indiana, Landsat MSS data set, channel 4.

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 for a more complex image of the same size.

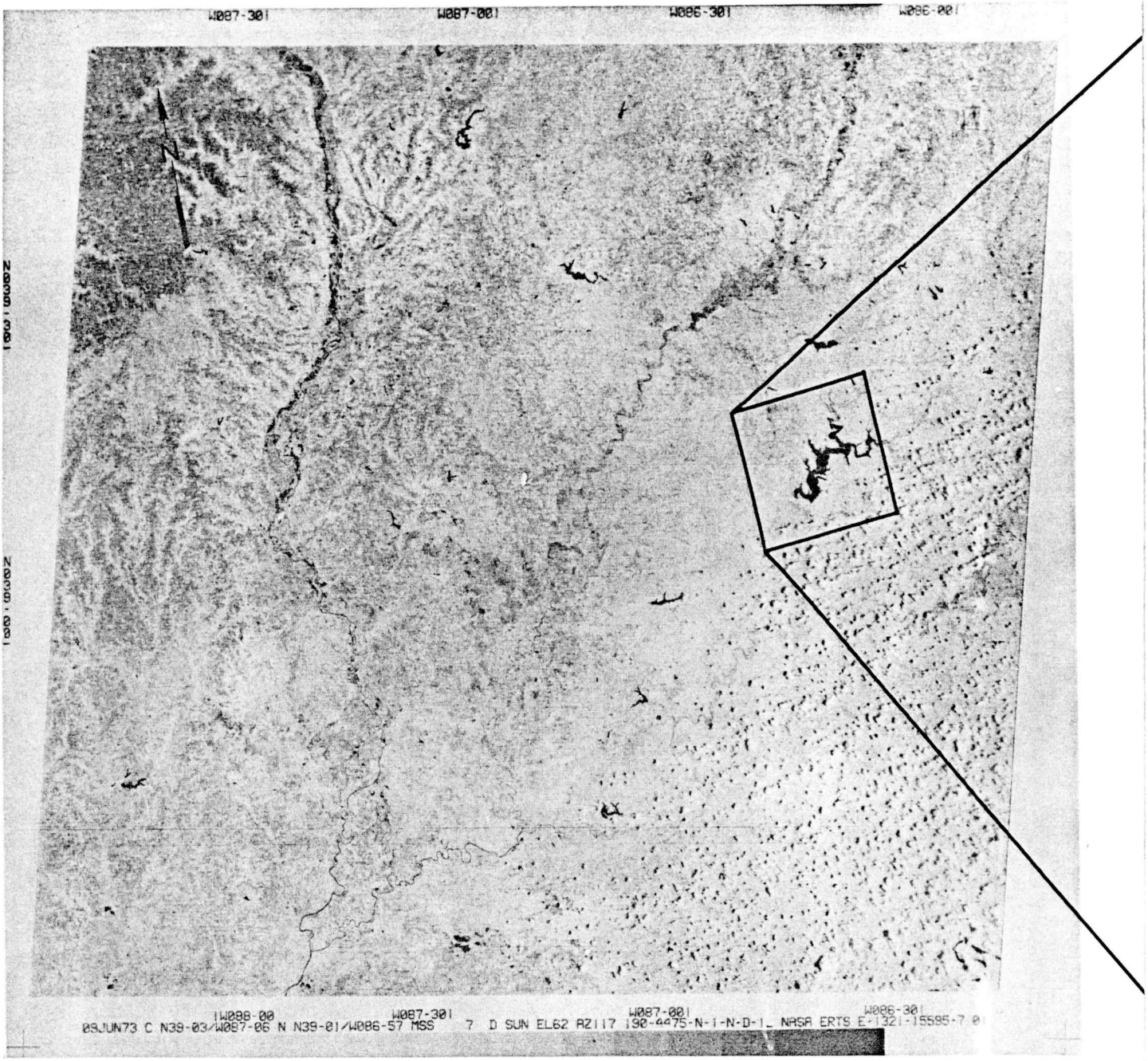
> Identify from the information on the backs of the printerplotter images, which image is Channel 1, 2, 3 and 4. Examine the gray scale images. Why are the images from Channels 1 and 2 similar to each other? Why are the images from Channels 3 and 4 similar to each other but different from those of Channels 1 and 2? What can you identify on the gray scale images? What data quality problem is apparent in all the images (especially Channels 1 and 3)? How much cloud coverage is apparent in the imagery? Note the bad data line in the Channel 3 image.

LOCATING A STUDY AREA AND CORRELATING IT WITH REFERENCE DATA

This data set has already been geometrically corrected and rescaled. You can check this by noting that the ground heading listed by IDPRINT for this data set is 180 degrees. An uncorrected data set would have a ground heading of about 190 degrees. As we noted earlier while studying IDPRINT, there are only 493 lines (number of lines) and 480 columns (number of data samples) in this data set. Because of the cost of geometrically correcting large amounts of data, it is wise to request correction of only the portion needed for the analysis. Figure II-9 (the foldout) shows two Landsat images and a print of an aerial photograph taken at 60,000 feet on the same day. The aerial photograph, taken over the area of interest, was used to select the portion of the data to be geometrically corrected, shown outlined on the image of Band 7 (Channel 4). (The term "ERTS" mentioned in the foldout captions is the original name for Landsat. The satellite was renamed "Landsat" in 1975.)

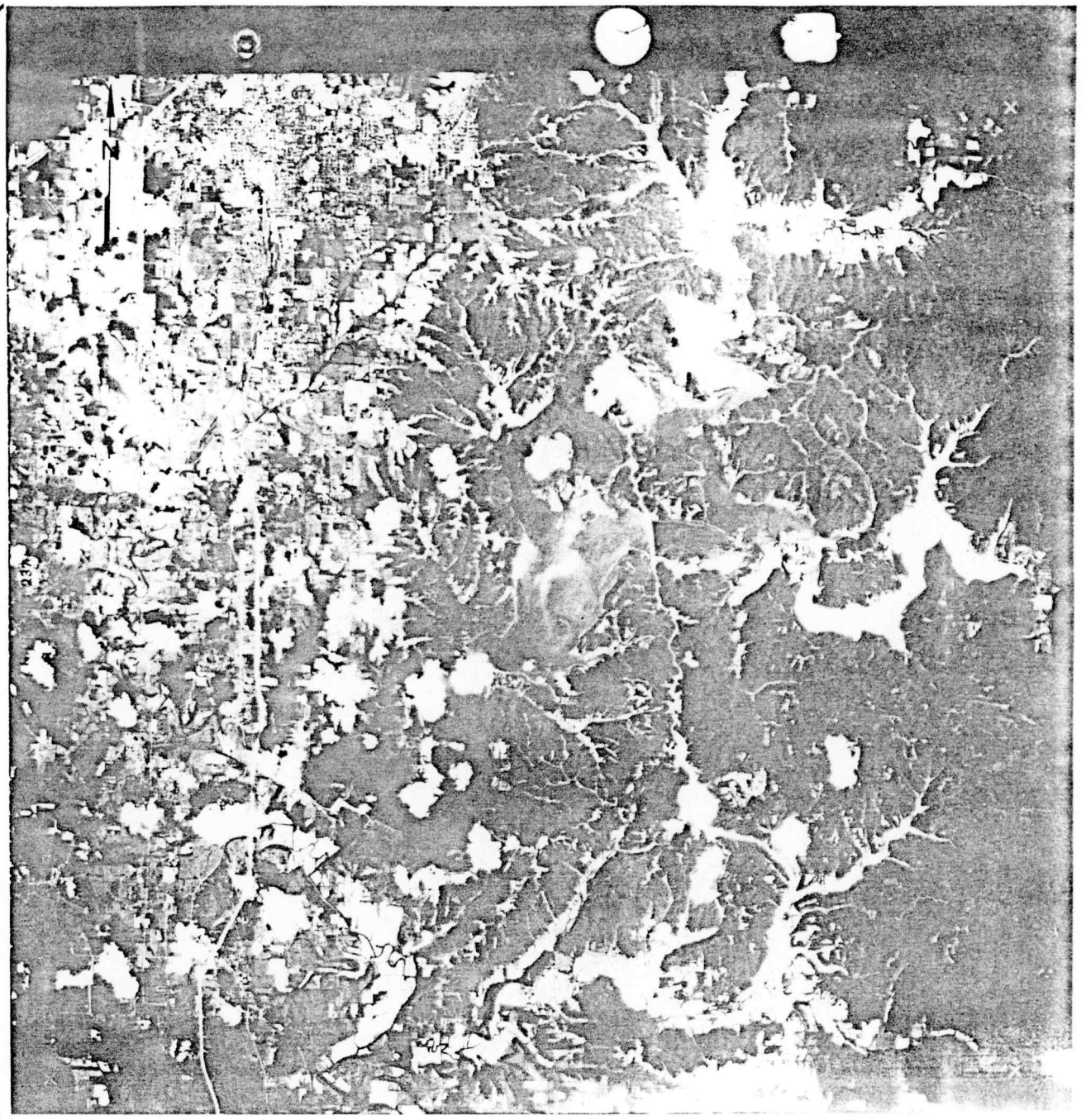
The printer-plotter images we looked at earlier were made from this geometrically corrected Landsat data set. Looking at the reverse side of these printer-plotter images, we can see that the analyst chose to display only lines 30 through 400 and columns 112 through 474. This portion was displayed since it was noted from the aerial photography that the area of interest was located within these line and column ranges.

Your instructor will provide you with the following reference data: U. S. G. S. 7 1/2-minute topographic quadrangle maps covering Monroe County and a color-infrared photograph taken over Monroe County (available as a 35 mm slide or a 9x9 inch print). As noted earlier, the data set with which you are working has

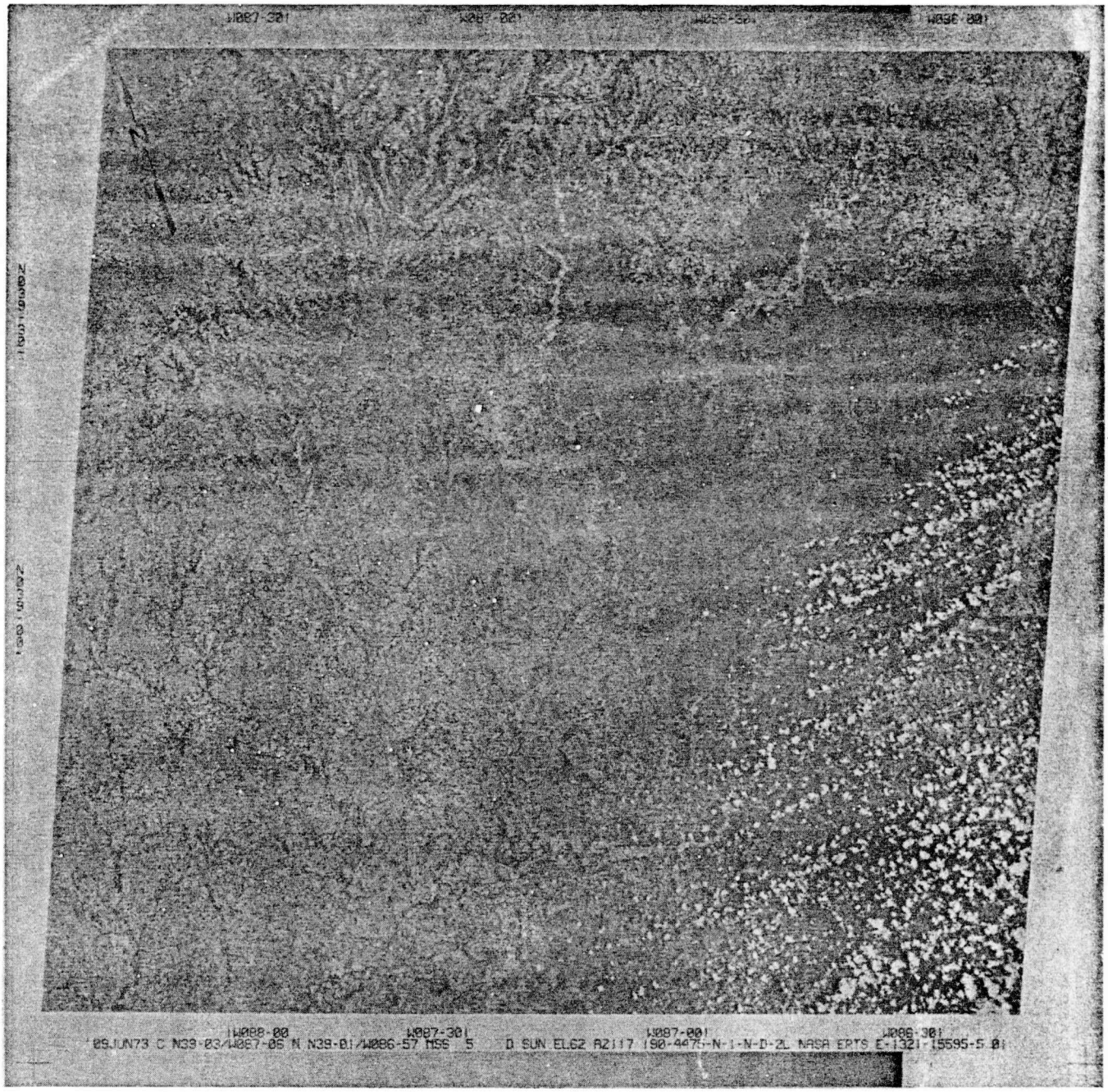


This print shows channel 7 (.8-1.1μm) of the same ERTS scene. The area outlined corresponds to the frame of aerial photography on the next page. This area includes Monroe Reservoir and Bloomington, Indiana.

II-23 b



This print was made from a 9 x 9 color infrared photograph collected at an altitude of 60,000 feet at 11:40 a.m. the same day.



This is a print of ERTS scene 1321-15595, channel 5 (.6-.7μm), collected June 9, 1973 at 9:59 a.m. Note that north is displaced 13° from vertical.

been geometrically corrected and rescaled. The rescaling was done such that a line printer image of the data would have a 1:24,000 scale, which matches the scale of the U. S. G. S. 7 1/2-minute topographic series. The scale of the printer-plotter images is approximately the same as the aerial photography reference data.

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

SELECTION OF THE TRAINING SAMPLE

We will use the hybrid approach to selecting the training sample in this workshop. Remember that each training area selected should include more than one cover type, and 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. If you select only three training areas make them on the larger side of the range. If you select more than three training areas, the training areas can be correspondingly smaller.

Select three training areas which are representative of the scene. Use the available reference data, the gray scale printouts and the guidelines described earlier. Make sure that every cover type of interest (urban, agriculture, forest and water) is included in at least one of the training areas. Outline your areas 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.

Pages III-1 to III-26 not in file

MONROE COUNTY CASE STUDY - PART III

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

CLUSTERING THE TRAINING SAMPLE

The analyst chose three training areas in the previous step of the analysis. The location of each training area is outlined in Figure III-17. Each training area was clustered separately. The cluster output for the first training area is shown on pages 8-43 of your computer output.

On page 8, the phrase "OPTIONS MAXCLAS(15), CONV(98.5)" indicates that the analyst requested 15 clusters and a convergence value of 98.5%. Six information classes (water, forest, pasture, bare soil, urban and emerging crops) were identified in this training area giving 12 as the suggested number of clusters by the "2X" rule. To allow for various transitions between the information classes, the analyst requested 15 clusters rather than 12.

The cluster means and variances for the first training area are given on page 10. Also listed are the number of pixels ("POINTS") assigned to each cluster. In general, the first cluster has larger mean values than the last cluster. The general trend is from bright (larger mean values) to dark (smaller mean values). The variances indicate the spread or dispersion of the data in each channel.

The cluster map is shown on pages 11 and 12. Comparing the cluster map with the gray scale imagery, we can readily determine that cluster 15 (designated by $ on the map) represents water. The less obvious associations of cluster classes with information classes will be made in the next section.

Pages 13-42 show histograms of the data values in each channel for each cluster class. The histograms show for each channel the distribution of the data values of the pixels grouped into each cluster by the clustering algorithm. Since the classification algorithm we will use assumes that each training class can be represented by a Gaussian density function, we must check each histogram to see how closely each cluster class can be approximated by a Gaussian density function. Note the rather non-Gaussian character of the channel 3 histogram of cluster 3 (page 17). This deviation is not great enough, however, to require reclustering. We will, however, take this deviation into account later when we make decisions about which candidate

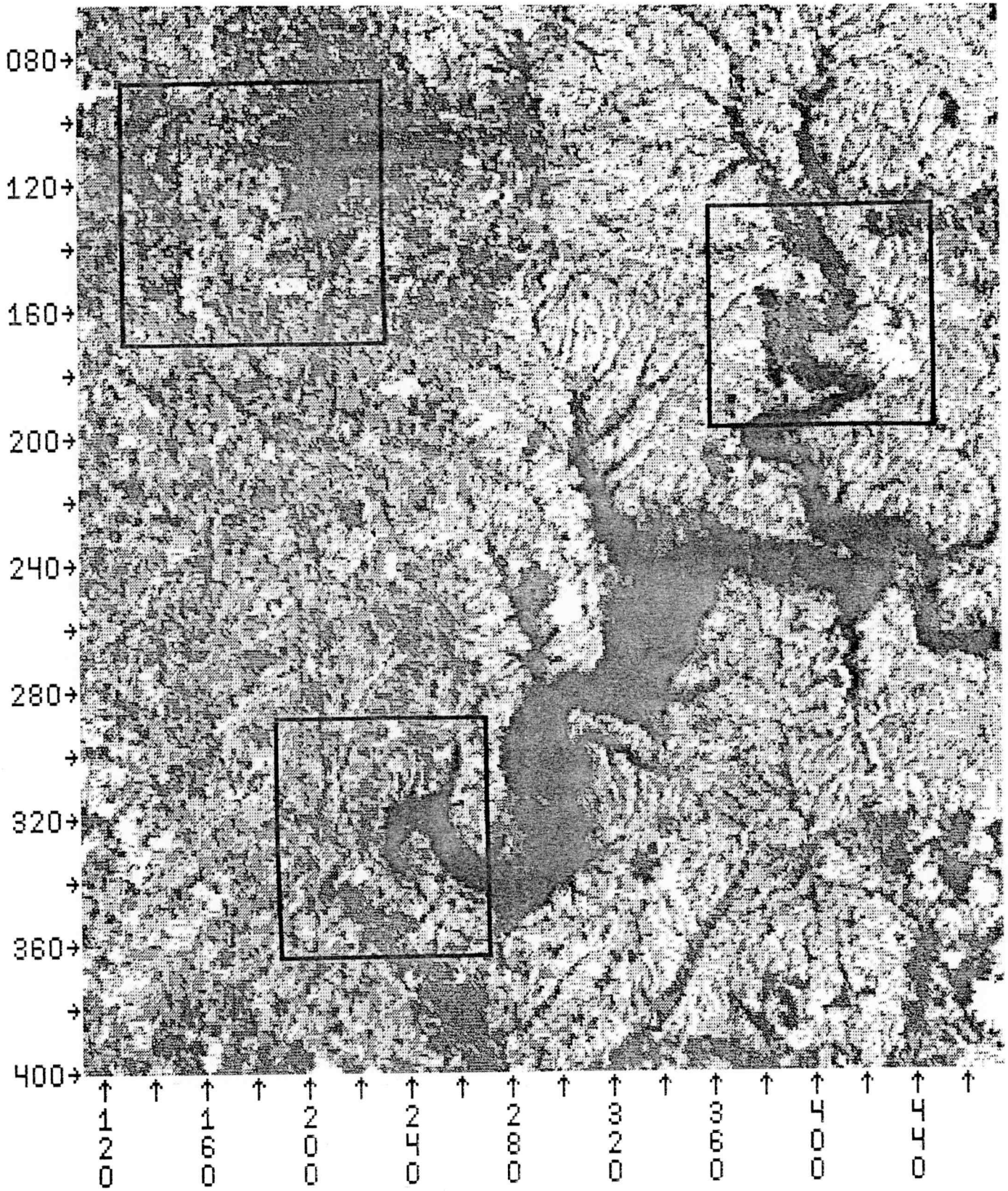


Figure III-17. Monroe County case study training areas.

training classes will make up our set of final training classes. Note also that the large variances indicated for certain clusters (page 10) are reflected in the histograms by the way the data is spread out.

Examine the remaining output from the CLUSTER processor (training areas 2 and 3, output pages 44-115).

1. Look at the mean values for each cluster on pages 46 and 84. Try to determine a general identity for each cluster (vegetation, bare soil or water) by comparing the relative values in each band to known spectral reflectance characteristics of 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 any unusually high or low values.

3. Briefly check the cluster maps for spatial similarity with the reference data.

4. Examine the cluster histograms. Note any obviously non-Gaussian characteristics. Select a cluster having low variances and compare its histograms to a cluster having high variances.

5. Note any clusters having fewer than 40 pixels.

ASSOCIATING THE CANDIDATE TRAINING CLASSES WITH INFORMATION CLASSES

Now that we have generated a set of candidate training classes by clustering our three training areas, we need to identify which information class each candidate training class belongs to. We will use line printer cluster maps, aerial photography and U. S. G. S. quandrangle maps to assist in this identification. Remember that the correspondence between information clases and candidate training classes is not necessarily one-to-one. Most often more than one candidate training class is associated with one information class. Occassionally (hopefully rarely) more than one information class is associated with one candidate training class.

Associate cluster classes with information classes in the first training area. Using the computer-generated maps from clustering the data (printout pages 11-12), the color infrared imagery (format may be a 9" X 9" print or a 2" X 2" slide) and the topographic map, identify as accurately as possible each cluster class in the first training area. Record your identification in Table III-1. The Landsat data used in this case study was geometrically corrected and rescaled so that when printed on the line printer it could be overlaid on the topographic maps. Therefore, it may be advisable to refer to the topographic maps to aid in the association process. The color infrared photography will be helpful for making more detailed identifications, such as distinguishing fields of bare soil from green vegetation in agricultural areas or shopping centers from residential neighborhoods in the urban area. The aerial photography was acquired the same day as the Landsat data, about 2 hours later. The date is June 9, in southern Indiana; planting has just been completed, and the recently planted crops appear as bare soil (shades of gray) in the photography. The deep, intense red represents wooded areas whereas lighter reds and pinks probably represent wheat, pasture or grassy areas. Water and urban areas are more obvious and should be no problem to identify. If you have the time, also identify the candidate training classes in training areas 2 and 3. The cluster map for training area 2 can be found on pages 47-48 of the computer printout and the cluster map for training area 3 can be found on pages 85-86.

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 information classes in the scene. Clouds and cloud shadows are two information classes that do not appear in our training sample. We must now use the supervised approach to define cloud and cloud shadow training classes.

Using the printer-plotter gray scale images and the color IR photo, specify the line and column coordinates for a number of cloud and cloud shadow training areas. When trying to identify clouds, keep in mind that the Landsat data and aerial photograph were not collected at the same time of day. Therefore, clouds and their shadows are not in the same positions in the photography as they are in the MSS data. In selecting training areas bear in mind the minimum number of pixels needed to estimate the statistical properties of the training sample(see page III-7).

Table III-1. Student cluster-class/information-class associations.

SYMBOL	CLUSTER	AREA 1	AREA 2	AREA 3
.	1			
-	2			
=	3			
/	4			
+	5			
*	6			
I	7			
H	8			
Y	9			
L	10			
P	11			
U	12			
O	13			
X	14			
$	15	WATER		
M	16			

Our analyst chose four training areas for clouds and cloud shadows as outlined in Figure III-18, and submitted the line and column coordinates of these training areas to the STATISTICS processor. The STATISTICS output for the cloud areas is shown on pages 138-140 of the printouts, and the output for cloud shadows is shown on pages 141-143.

Examine the STATISTICS processor output for the cloud and cloud shadow training areas. Study the following for each class:

- number of pixels
- means and standard deviations
- correlation matrix
- histograms

Note the high variability of the cloud histograms and the saturation in channels 1, 2 and 3.

Our analyst noticed variations in water quality in the Monroe Reservoir in the color IR photograph and decided to attempt to identify several subclasses of water. In order to identify spectral subclasses of water, the line and column coordinates of these training areas were submitted to the CLUSTER processor. These training areas are outlined in Figure III-19. Since the areas are all small, and since they all contain the same cover type, the five areas were clustered together. The computer output from CLUSTER is shown on Pages 118-135.

Examine the CLUSTER processor output for the water training areas. Note the following for each subclass:

- number of pixels
- means and variances
- histograms

Note the similarity in mean value among all clusters on page 119. In particular note the low spectral response in channel 4. Compared to results seen earlier, the cluster variances are all small except for channel 3, cluster class 1.

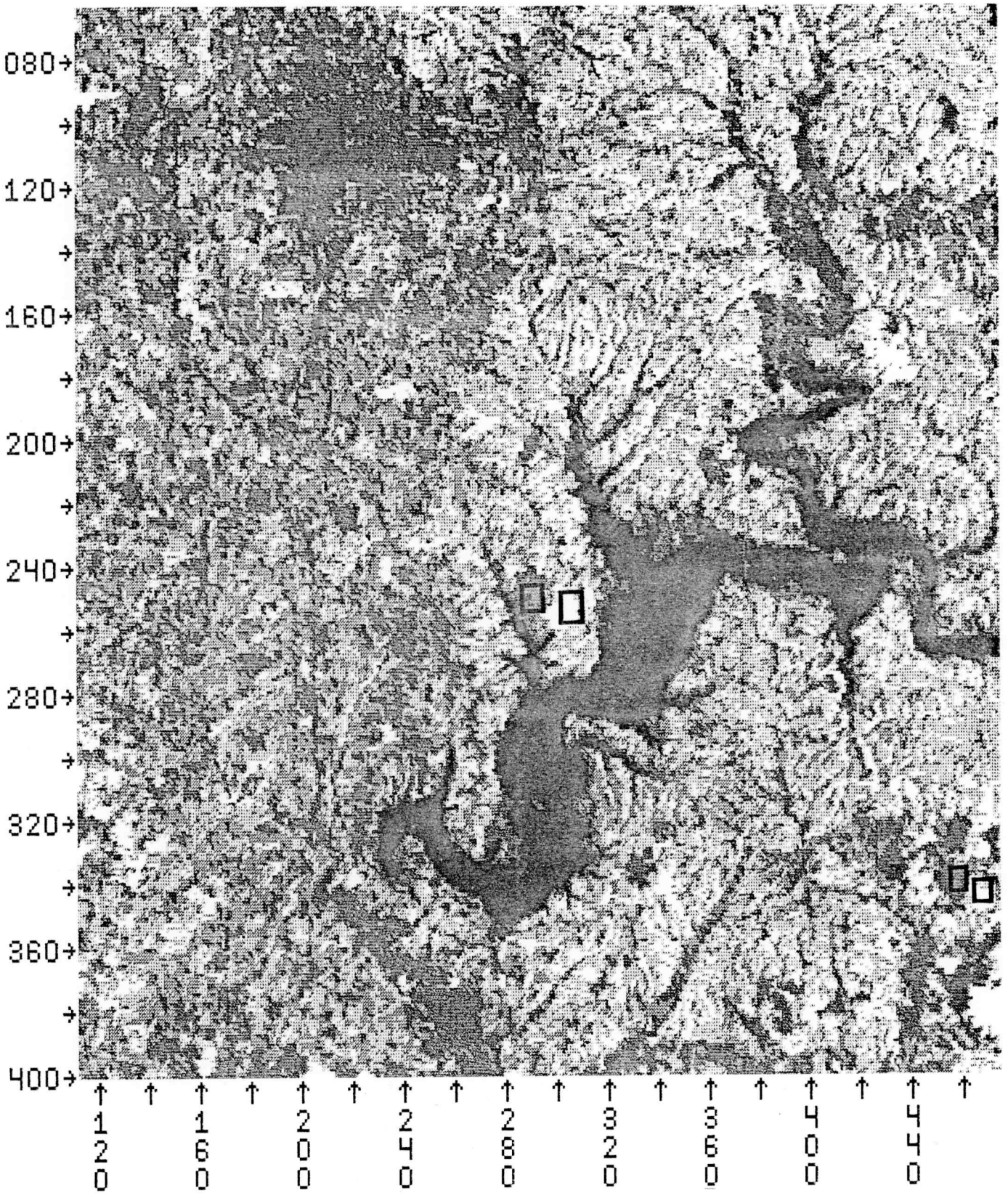


Figure III-18. Monroe County case study training areas for clouds and shadows.

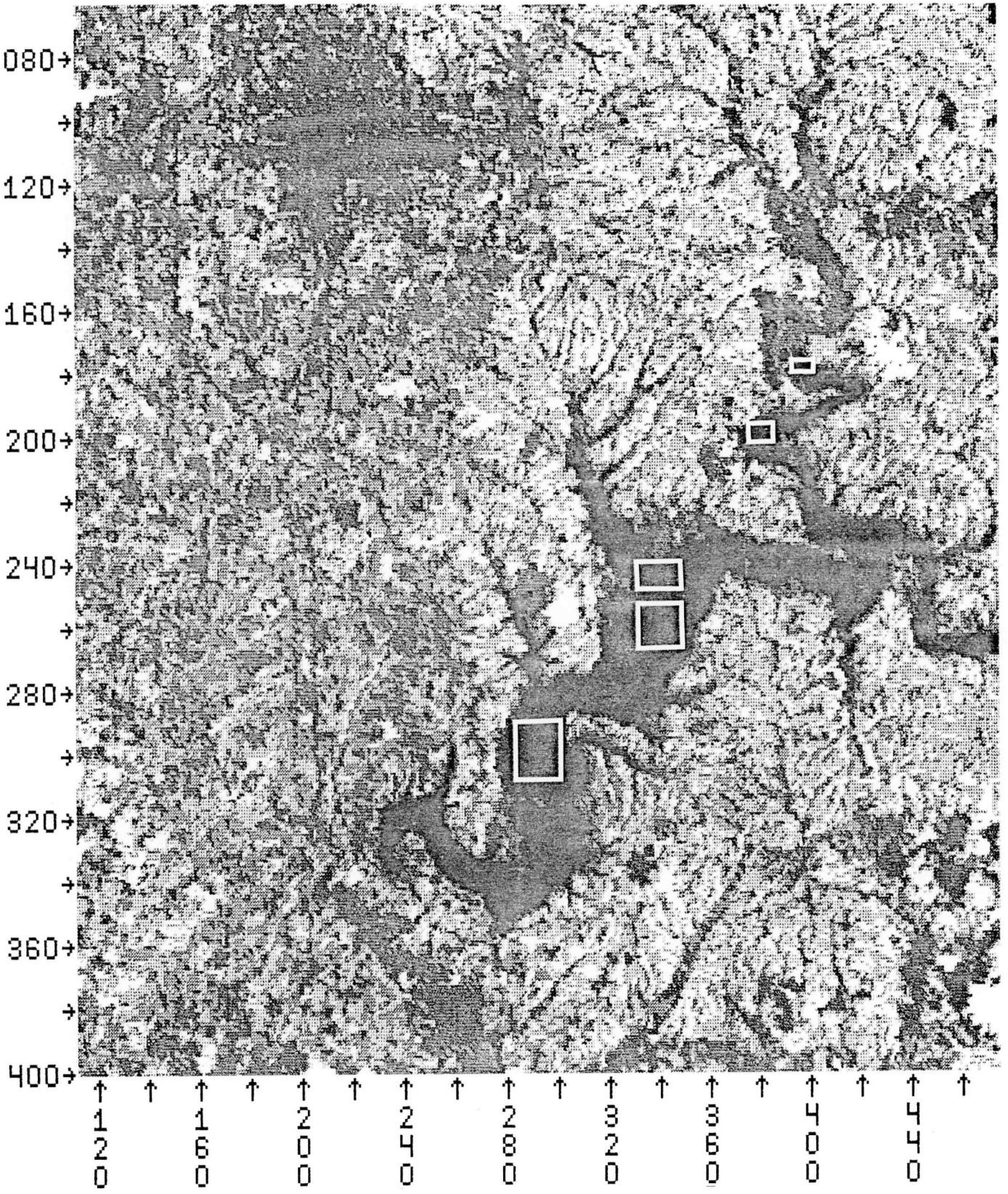


Figure III-19. Monroe County case study training areas for water.

VISUAL REPRESENTATION OF CANDIDATE TRAINING CLASSES

Previously you associated information classes with the cluster classes from the three training areas. Our analyst performed this association as shown in Table III-2. Look back at Table III-1 and compare your identifications with those given in Table III-2. When we add the cloud and cloud shadow classes and the five water subclasses to the cluster classes in Table III-2, we have 52 candidate training classes.

We would now like to reduce the number of training classes. We can do this by combining (pooling) some candidate training classes and deleting others. To begin the analysis of which candidate training classes may be pooled or deleted, it is useful to visualize the spectral charateristics of all candidate training classes at one time. The MERGESTATISTICS processing function has as one of its output products a two dimensional plot known as a coincident bi-spectral plot. Plotted on one axis is the average mean values in the two Landsat near infrared channels for each candidate training class. One the other axis is plotted the average mean values in the two visible bands. The output is a plot providing a visual comparison of the means of all candidate training classes.

Pages 147-148 of your computer printout show the bi-spectral plot for the 52 candidate training classes. Pages 148-149 list the classes, the average mean values in the visible and in the near infrared bands, and the symbol used to represent each class. The first 15 classes may be recognized as coming from training area 1, the next 16 from training area 2, and the next 14 from training area 3. The next set of five are the water training classes generated by clustering the special water training sample, and the last two are the cloud and cloud shadow classes obtained through the 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 148-149 add a subscript 2 to the symbols on the bi-spectral plot for classes 31 through 52. For example class 31, symbol A (second A), has a mean in the visible bands of 28.9 (y-axis) and 18.4 in the IR bands (x-axis). Locate and label this class A2. Do the same for classes 32 through 52. See Table III-3 for a listing of all the candidate training classes and the symbols used on the bi-spectral plot.

Examining the listing on page 149, you'll note that four candidate training classes do not appear on the bi-spectral plot because they fall outside of the range of the x-axis.

Table III-2. Case study analyst's cluster-class/information-class associations.

Cluster	Symbol	Training Area 1 South Reservoir	Training Area 2 Bloomington	Training Area 3 North Reservoir
1	.	Bare Soil	Bare Soil/Roof	Bottomland/Bare Soil + ?
2	-	Bare Soil/ Hiway	Bare Soil	Regeneration (sec. veg.)
3	=	Fields, Edges, Emerging Crops	Roof/Bare Soil	Deciduous Forest
4	/	Emerging Crop	Commercial	Forest/Grass
5	+	Pasture	Residential	Deciduous Forest
6	*	Forest	Older Resid.	Deciduous Forest
7	I	Forest	Residential/ Emerging Crop	Deciduous Forest (NW Slope)
8	H	Pasture (nat. veg.)	Older Resid. (more trees)	Emerging Crop/ Grass
9	Y	Emerging Crop	Resid. (most veg. & trees)	Grass Edge
10	L	Forest	Grass	Bare Soil (slight pink)
11	P	Emerging Crop (bare soil)	Grass/Pasture (nat. veg.)	Bare Soil (more pink)
12	U	Emerging Crop	Grass/Pasture (nat. veg.)	Emerging Crop/ Water Edge
13	O	Forest Edge/ Water	Grass/Pasture (nat. veg.)	Forest/Water
14	X	Water Edge	Small Trees (dense)	Turbid Water
15	$	Water	Older Resid.	
16	M		Water	

Table III-3. Detailed cluster class/information class associations.

Training Area	Class Number	Separability Symbol	Cluster Number	Class Identity	Number of Pixels
	1	A	1	Bare Soil	157
	2	B	2	Bare Soil/Highway	313
	3	C	3	Field Edges, Emergin Crops	340
	4	D	4	Emergin Crop	469
	5	E	5	Pasture	391
	6	F	6	Forest	542
	7	G	7	Forest	649
1	8	H	8	Pasture/Natural Vegetation	838
	9	I	9	Emerging Crop	673
	10	J	10	Forest	498
	11	K	11	Emerging Crop	230
	12	L	12	Emerging Crop	205
	13	M	13	Forest Edge/Water	80
	14	N	14	Water Edge	109
	15	O	15	Water	662
	16	P	1	Bare Soil/Roof	75
	17	Q	2	Bare Soil	292
	18	R	3	Roof/Bare Soil	392
	19	S	4	Commercial	255
	20	T	5	Residential	588
	21	U	6	Older Residential	590
	22	V	7	Residential/Emerging Crop	776
2	23	W	8	Older Residential (More Trees)	743
	24	X	9	Residential (Mostly Veg. & Trees)	807
	25	Y	10	Grass	541
	26	Z	11	Grass/Pasture	545
	27	$	12	Grass/Pasture	1063
	28	+	13	Grass/Pasture	622
	29	=	14	Small Dense Trees	472
	30	/	15	Older Residential	375
	31	A2	16	Water	45

Table III-3 (cont'd).

Training Area	Class Number	Separability Symbol	Cluster Number	Class Identity	Number of Pixels
3	32	B2	1	Bare Soil?	34
	33	C2	2	Regeneration	188
	34	D2	3	Deciduous Forest	645
	35	E2	4	Forest and Grass	455
	36	F2	5	Deciduous Forest	1272
	37	G2	6	Deciduous Forest	1126
	38	H2	7	Deciduous Forest (NW Slope)	625
	39	I2	8	Emerging Crop/Grass	469
	40	J2	9	Grass Edge	130
	41	K2	10	Bare Soil (Slight Pink)	157
	42	L2	11	Bare Soil (More Pink)	267
	43	M2	12	Emerging Crop/Water Edge	244
	44	N2	13	Forest/Water	273
	45	O2	14	Turbid Water	576
4	46	P2	1	Water	54
	47	Q2	2	Water	157
	48	R2	3	Water	335
	49	S2	4	Water	444
	50	T2	5	Water	106
5	51	U2	1	Cloud	136
	52	V2	2	Shadow	91

Plot in the margins of the bi-spectral plot the approximate location of the four candidate training classes which have been omitted from the plot.

CALCULATING STATISTICAL DISTANCES BETWEEN THE CANDIDATE TRAINING CLASSES

The bi-spectral plot studied in the last section shows which classes are spectrally related to each other. However, since the bi-spectral plot is based only on class means, it doesn't tell us anything about which classes overlap each other or how badly they overlap. To make reasonable decisions about which candidate training classes to pool or delete, we should also consider the spectral variability of each class and how much the classes overlap each other. We can do this by calculating the transformed divergence (D_T) between each class using the SEPARABILITY processor.

Pages 150 through 168 display the computer printouts from the SEPARABILITY processor for all 52 candidate training classes. Page 151 shows the symbol used to represent each class. Notice that the distances are given for pairs of classes (pages 152 to 166). For example, on page 152 the transformed divergence between classes A and B (classes 1 and 2 from training area 1) is 1676. Also note that the largest value appearing in the table is 2000, corresponding to maximum separability. A threshold of 1500 was chosen and a summary listing of all class pairs whose pairwise distance was less than or equal to 1500 is printed on pages 167 and 168. This condensed listing will be useful in constructing and analyzing a "separability diagram." This part of the procedure frequently involves two or more iterations, depending on how simple or complex the analysis problem is, and the threshold may change from one iteration to the next. 1500 has been a useful starting value for combining clusters in many problems.

Notice the legend accompanying the separability output on page 151. Because there are so many classes, some of the symbols are used twice. Ambiguity caused by the duplication was resolved before the printouts were generated. Thus on pages 167 and 168 each symbol is identified as A1 or A2, etc.

Using the list of class pairs with transformed divergence values less than 1500 (pages 167-168 of the computer printouts), add separability information to your bi-spectral plot (pages 147-148). Draw a solid line between

two classes if the transformed divergence between them is less than 1000. Draw a dashed line between them if the transformed divergence is between 1000 and 1500. No line is drawn if the transformed divergence is greater than 1500.

REFINING THE SPECTRAL TRAINING CLASSES.

As an aid for selecting our final training classes we now have a bi-spectral plot complete with separability information. One additional useful piece of information we would like to add to the bi-spectral plot before we consider pooling and/or deleting classes is the information class identity of each candidate training class.

A list of candidate training class identities is shown in Table III-3. The candidate training-class/information-class associations shown in Table III-3 were determined with the aid of a zoom transfer scope. Using this list, write an abbreviation of the identity of each class next to its symbol on the bi-spectral plot (pages 147-148 in the printouts).

We are now ready to make decisions as to which candidate training classes should be deleted, which should be pooled, and which should be used alone as final training classes. While trying to make these decisions, we need to keep in mind the information class labels, the normality of the histograms, and the number of pixels of each candidate training class, as well as 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. The pooling or lumping approach combines classes that are spectrally close to each other (i.e., have D_T less than a threshold value). The deleting or splitting approach deletes classes that are on the borders between information class groups on the bi-spectral plot. The pooling approach is generally more concerned with the first goal above (representation), while the deleting approach is more concerned

with the second goal (separability). Experience has shown that the optimal solution is a combination of these two strategies.

Let's consider refining the candidate training classes for bare soil. On the bi-spectral plot we notice that class P is not connected with any other class. We'll leave it alone. The other bare soil classes - A, Q, B, R and K2 - are connected by dashed and solid lines. It is pretty clear that we can combine classes A and Q, since they are spectrally very similar (D_T=178) and they unambiguously have the same information class label.

Even though classes B and R are close to each other spectrally (D_T=857), it is not so clear that they can be combined. It may be best to delete class B, since it is a borderline class (mixture of Bare Soil and Hiway) and is spectrally close to class T which is unambiguously Residential. Also, looking back at the histograms for class B (class 2, training area 1), we see that the distribution for channel 3 does not look very normal. On the other hand, the histograms for class T (class 5, training area 2) approximate a normal distribution very closely. Considering all of this, it is probably a good idea to delete class B to enhance the separability between the Bare Soil classes and the Residential class T.

Class K2 is identified as "slightly pink" Bare Soil. It is close spectrally to class S (Commercial) and class K (Emerging Crop). Looking at K2's histograms (class 10, training area 3) we see that the channel 2 histogram does not look normally distributed. On the other hand, all the histograms for class S (class 4, training area 2) and class K (class 11, training area 1) look much more normal. We also note that class K2 has fewer pixels than either class K or class S. Taking all of this into account, we would be justified in deleting class K2, especially since class K2 can be considered to be an "mixture" class of Bare Soil and Emerging Crop.

If we delete classes B and K2, we find that class R is separable from all of the remaining classes even though it is a mixture class (Roof/Bare Soil). This being the case, we can retain this class for this iteration of our training class refinement. Later we can decide if class R should be considered as Agricultural or Urban, or whether we should delete it.

> Select training classes using the bi-spectral plot augmented with the class names and separability information. Delete, pool or leave as is candidate training classes as appropriate to form your final training classes. At least one training class should be selected for each of the cover types present in the area. For the purposes of combining classes, use a transformed divergence threshold of 1500. Start with the water classes (water is generally the easiest cover type to work with).

After deleting and pooling some of the original 52 candidate training classes, our analyst came up with 19 final training classes. They are listed in Table III-4 and on page 172 of the computer printouts at the bottom of the second bi-spectral plot. 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 final training classes to the 19 classes in Table III-4. Did you pool more classes? Did you delete more?

Your final training classes are not necessarily better or worse than our analyst's set. For example, our analyst deleted class Q. Class Q could have just as well have been combined with class A. Remember that this step more than any other step in the analysis is more an art than a science.

Now that we have selected our 19 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. In LARSYS, this is done by means of the SEPARABILITY processing function. The SEPARABILITY output for this second running of the processor begins on page 174 of the computer printouts. In this case, our analyst also requested the mean vector and correlation matrix for each class. These are shown beginning on page 176. In order to be informed of pairs of the newly formed training classes that have a statistical distance only slightly greater than the previously used threshold, all class pairs with transformed divergence values less than 1750 were listed. This list appears on page 197.

Add separability information and class name identities to the bi-spectral plot of the 19 final training classes, which can be found on pages 171-2 of the computer printouts. Use a solid line for a transformed divergence value of less than 1500, and a dashed line for a value between 1500 and 1750. Note that the CLOUD class falls outside the scale of the plot. Plot this class in its approximate location in the margin.

Examine the completed bi-spectral plot. Are there any pairs of classes with a transformed divergence of less than 1500?

Table III-4. List of 19 final spectral training classes for the Monroe County case study.

Pool Name	Class in Separability Output		Cluster Number	Cluster Area
GRASS	5	E1	5/15	1
	25	Y1	10/16	2
DESIDFOR	34	D2	3/14	3
	36	F2	5/14	3
DESIDFOR	38	H2	7/14	3
REGENVEG	33	C2	2/14	3
EMGCROP	9	I1	9/15	1
RES/EMCR	22	V1	7/16	2
RESIDENT	20	T1	5/16	2
OLDRESID	30	/1	15/16	2
COMMERCL	19	S1	4/16	2
ROOF/BS	18	R1	3/16	2
BARESOIL	1	A1	1/15	1
BARESOIL	16	P1	1/16	2
CLOUD	51	U2	cloud	
CLDSHAD	52	V2	shadow	
WATEDGE1	44	N2	13/14	3
WATEDGE2	43	M2	12/14	3
WATER1	50	T2	5/5	4
WATER2	47	Q2	2/5	4
WATER	46	P2	1/5	4

From the transformed divergence information we can see that only one pair of classes has a transformed divergence value below 1500; classes H (OLDRESID) and P (WATEDGE2) have D_T=1424. We could decide that having even one pair of classes with D_T<1500 is unacceptable because we anticipate too many classification errors between these two classes and so redo our refinement of training classes. Or we could decide that it isn't so important to separate these two particular classes and leave the training classes as they are. We will leave them as they are and go on to the classification step.

Pages IV-1 to IV-4 not in file

MONROE COUNTY CASE STUDY - PART IV

The Monroe County study area was classified with the LARSYS CLASSIFYPOINTS processor using the 19 training classes developed in the previous chapter. 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 for each pixel. Each pixel is classified into the class with the largest discriminant function. For each pixel the class into which it is classified and the probability that the pixel was correctly classified is stored on a results tape. We will examine the results on this tape in the next chapter.

The computer output from CLASSIFYPOINTS can be found on pages 199-201 of your computer printouts.

 Pages V-1 to V-6 not in file

MONROE COUNTY CASE STUDY - PART V

A classification map as produced by the LARSYS PRINTRESULTS processor can be found on pages 204-227 of your computer printouts. The control statements on page 203 show the map symbols chosen by our analyst and the specified groupings. This information is presented in a more convenient form on page 204. Note that grass, soil and emerging crop are denoted by different symbols on the map but belong to the same group: agriculture. The map is several printout sheets wide. Your instructor can show you a map which has been assembled from these sheets.

A classification map printed with graphic symbols is shown in Figure V-5. This map covers a portion of training area 1.

The classification results can also be displayed in tabular form. On page 232 of your computer printouts you will find a table showing the number of pixels classified into each spectral training class, the number of acres and hectares covered by each class, and the percentage of the entire scene covered by each class.

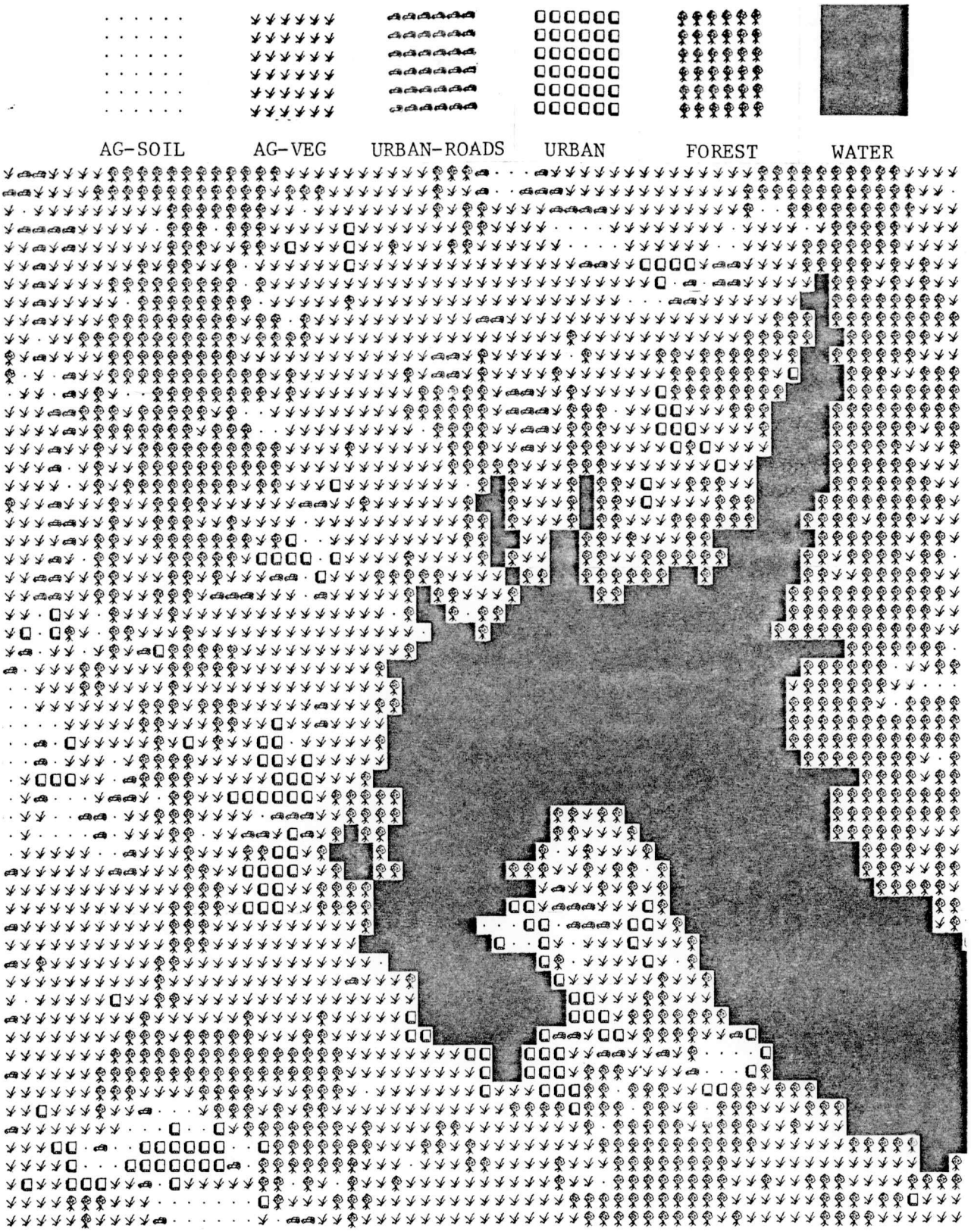


Figure V-5. Graphic symbols representation of the classification results from a portion of the Monroe County case study training area 1.

MONROE COUNTY CASE STUDY - PART VI

In the last chapter we looked at the classification map of the Monroe County 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 (agriculture, urban, forest, water and clouds) and checking to see how accurate our classification is in these test fields.

> Select two test fields on the gray scale maps for each major cover type (agriculture, urban, forest, water and clouds). The fields should be distributed throughout the area. Avoid the bad data line and the training areas. The test fields should 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 our classification accuracy estimate we have to be sure that the relative number of pixels in each major ground cover category is roughly the same in our group of test fields as in the entire classified area. What can we do to avoid this bias? For one, we could choose additional test fields for those ground cover types that are under representated in the original group of test fields. Or we could use an entirely different method of test field selection.

We are much more likely to obtain unbiased estimates of classification accuracy if we randomly select test fields. In using this method, our analyst divided the entire scene into 3X3 cells and randomly choose 10% of them by using a random number table. Our analyst then identified the ground cover type in each test cell and eliminated those test cells that contained more than one major cover type. The coordinates and cover types of the remaining test fields were then entered into the PRINTRESULTS processor.

> Examine the classification map on pages 204-227 of your computer printouts. Note that the 3X3 test fields are outlined with a "+". The Test Class Performance table is on page 228. For which classes were the most errors committed?

Out of the 441 test field pixels that were urban, only 244 were correctly identified as urban. This is a very poor

classification accuracy - 55.3% correct. The agriculture class accuracy wasn't quite as bad at 77.1%, but it still wasn't all that good. The forest, water and cloud accuracies were all acceptable with the water and cloud accuracies being very high.

Why is the urban accuracy so poor? If we look back at the pooling and deleting we did to form our spectral training classes, we will note that most of the deleted classes were urban or agricultural. Possibly the deletion of so many of the urban candidate training classes hurt our classification accuracy.

For comparison, let's look at a second attempt at defining the spectral training classes from the 51 candidate training classes. Comparing the training class definitions in Table VI-2 with those in Table III-4, we see that many more urban spectral classes are retained in Table VI-2. In this second attempt we have 27 spectral training classes compared to the 19 classes in the first attempt.

Using the same test fields, the results with these 27 training classes are listed in table VI-3 (compared to the results with 19 training classes). We see that the urban and agricultural class accuracies went up appreciably and the forest and water class accuracies went down slightly. The overall accuracy improved from 83.6% to 85.2%. With nearly 50% more training classes (and about 35% more money) we produced a classification with minimal overall improvement. However, this redefinition of training classes would be worthwhile if it were particularly important to accurately identify the urban class.

	PERCENT CORRECT	
CLASS	19 TRAINING CLASSES	27 TRAINING CLASSES
Urban	55.3	64.4
Agriculture	71.1	76.6
Forest	88.6	88.1
Water	97.2	96.9
Cloud	100.0	100.0
Overall	83.6	85.2

Table VI-3. Comparison of results obtained with two different sets of training classes.

Table VI-2. List of 27 alternate final spectral training classes for the Monroe County case study.

Pool Symbol	Pool Name	Class in Original Separability	Output	Cluster Number	Cluster Area
A	SOIL1	16	P	1/16	2
B	SOIL2	1	A	1/15	1
		17	Q	2/16	2
C	SOIL3	18	R	3/16	2
D	COMM	19	S	4/16	2
E	NEWRES	20	T	5/16	2
F	OLDRES1	21	U	6/16	2
G	VEGRES	30	/	15/16	2
H	EDGE	14	N	14/15	1
I	EDGE2	13	M	13/15	1
J	THNCLOUD	32	B2	1/14	3
K	GRASS1	22	V	7/16	2
		40	J2	9/14	3
L	OLDRES2	24	W	8/16	2
M	OLDRES3	24	X	9/16	2
N	EMRGCROP	12	L	12/15	1
		28	+	13/16	2
		39	I2	8/14	3
O	FORSLOPE	38	H2	7/14	3
P	GRASS2	25	Y	10/16	2
Q	GRASS3	27	$	12/16	2
R	FOREST1	7	G	7/15	1
		37	G2	6/14	3
S	FOREST2	36	F2	5/14	3
T	FOREST3	34	D2	3/14	3
U	GRASS4	5	E	5/15	1
		26	Z	11/16	2
V	REGNFOR	33	C2	2/14	3
W	WATER1	45	O2	14/14	3
		46	P2	1/5	4
X	WATER2	47	Q2	2/5	4
		48	R2	3/5	4
Y	WATER3	49	S2	4/5	4
		50	T2	5/5	4
Z	CLOUD	51	U2	CLOUD	
$	SHADOW	52	V2	SHADOW	