LARS Publication 073183

POLK COUNTY (IOWA) CASE STUDY

by

Shirley M. Davis

For use with

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

August 1983

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 the output from the analysis steps.
2. Topographic maps, 1:250,000 and 1:24,000.
3. 9 x 9 color transparencies, obtained May 2, 1978, at 1:74,500

It is also advantageous to have one or two assembled grayscale lineprinter maps and an assembled classification map for an overview of the area. Although these printouts are included in the binder, it is difficult to gain a perspective on the entire area when they are in separate strips.

ACKNOWLEDGEMENTS

The author wishes to acknowledge a number of people who contributed to this project:

Dr. Luis A. Bartolucci, for directing the analysis and guiding the development of the documentation.

Carlos Valenzuela, Fabian Lozano, Jose Valdes, and Jeff Madden for carrying out the analysis.

Ober Anderson, the County Extension Agent of Polk County, Iowa for local information.

Personnel of the ASCS Office in Polk County for maps and aerial photographs.

Daryls MacDonald and Marilynn Klepfer for word processing expertise.

Special acknowledgement is given to Dr. Dan Cotter, Mr. Bert Chapman, and NOAA for their support and sponsorship of the development of this documentation.

 Pages I-1 to I-4 are missing

PART I - INTRODUCTION

The case study analysis featured in this workshop is a LARSYS-based analysis of Thematic Mapper (TM) data collected on September 3, 1983 over Polk County, Iowa, U.S.A. The principal features of the scene are the city of Des Moines (population 191,000), Saylorville Reservoir, the Des Moines River with its 1 to 3-mile wide river valley, and forested and agricultural areas. Approximately 84% of the county is in farms, with corn and soybeans the principle crops grown.

Polk County lies between latitudes 41°25'N and 41°50'N and from longitude 93°20'W to 93°25'W. The general topography is flat to undulating, with some steep area around the streams and rivers. The geology of the area consists mainly of a Wisconsin Glacial Till with some loess deposits in the southern portion. The entire area is underlain by a shale bedrock of the Des Moines group. A map of the region is provided in Figure I-2.

The overall objective of the analysis is to create a land-cover map of the area represented by one million pixels, an area of 812.25 square kilometers. The analysis will seek to classify the data into the following earth surface materials: water (pure and turbid); urban (industrial, older residential, newer residential); agricultural with varying crops; forest; and highways. The goal is to achieve a classification with 85% accuracy.

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

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

1. the run number for this data set. Each data set has a unique run number; the first two digits designate the year of data acquisition.

2. the date and time of day when the data were collected. (Don't confuse this date with the reformatting date.)

---

* This case study description, authored by Shirley M. Davis, is based on an analysis performed by Carlos R. Valenzuela, Luis A. Bartolucci, D. Fabian Lozano, and Jose A. Valdes.

3. the number of lines and the number of data samples (columns) in this data set.

4. the wavelength range of each of the seven channels of data. Label each one with the portion of the spectrum it covers: visible, near-infrared, middle-infrared, thermal (far-infrared).

5. the calibration pulse values.

The calibration pulse values, shown in columns C0, C1, and C2, are the signals recorded by the sensor for calibrating the data. Historically, aircraft scanner systems recorded three calibration signals for each data channel: a value from the black interior of the scanner, one from an incandescent lamp, and a value for solar irradiance. With Landsat, only two calibration signals are recorded: from the interior of the scanner and from an incandescent lamp. These values, shown in mWatts/cm2-sr, appear in columns C0 and C1 and indicate the minimum and maximum radiances for each channel. As we shall see later, these values are used to convert the relative reflectance data into radiometric units.

Let's take a closer look at the data values themselves. In order to find out what the numerical data values are for any location in the scene (for any pixel), we can use another processor, TRANSFERDATA. Pages 3 through 8 of the computer listing show the output of TRANSFERDATA for five individual pixels, selected to represent different ground materials.

Notice that each pixel is identified by its line and column address and has seven data values associated with it. This set of values is often 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 area as measured in each wavelength band by the scanner system. Data values may range from 0 to 255 in each channel. A data value of zero is dark, signifying that no energy is measured by that detector for that pixel; a value of 255 is the brightest, indicating saturation of the scanner detector.

The five data vectors printed by TRANSFERDATA (pages 3-8 in the printout book) 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 defined class on the basis of these numerical values.

Figure I-2. Map of Polk County, Iowa, showing the major topographic features of the area described in the case study. Note that Saylorville Lake is defined only by short dashed lines overlaying the river basin.

The data vector below is for a pixel known to represent forest:

		CHANNELS						
		1	2	3	4	5	6	7
LINE	COL				DATA			
456	451	58.0	21.0	16.0	81.0	51.0	13.0	127.0

Compare these values to the ones shown on printout pages 4 through 8. Which page shows a data vector which probably also representes forest? How did you decide? If you had been allowed to use only Channel 4 data, could you have decided as easily? Would it have been easier if you had Channels 3 and 4? Why or why not?

From your understanding of the spectral characteristics of materials found on the earth's surface, can you decide what materials the other data vectors represent? Possible answers include: agriculture, clear water, turbid water, and bare soil.

 Pages II-1 & II-10, are blank missing

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

This portion of the case study demonstrates one possible way to carry out the first step of the analysis of digital multispectral data. The data set used in this case study was obtained by the Thematic Mapper aboard Landsat 4 on September 3, 1982. At this time of year, the major earth-surface features to be located (i.e., agriculture, forest, urban, and water) could be discriminated on the basis of spectral response in the TM wavelength bands.

ASSESSING DATA QUALITY

Our first task with the data is to examine it to assess the quality of the data, with regard to both cloud cover and radiometric characteristics. 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. Analysts therefore used a digital image display device, the COMTAL Vision One/20, to display an overview of the scene, and from this they selected the specific study area that included Des Moines and surrounding rural areas.

An alphanumeric printout of the study area was obtained for several of the channels. A portion of that printout appears in the book of printouts on pages 9-82. The symbols on the printouts represent the relative values of the data in each channel; the darker symbols (X, Z, $) represent areas of lower response and the lighter symbols (-, =, blank) represent higher values. These printouts are cumbersome to handle, but they allow you to look at every pixel individually in all channels.

The process used to generate the gray scale line-printer map is a combination of level slicing and contrast stretching. First, the computer determines the overall range of the data in each channel; the data to be used in this evaluation are specified by the analyst with the block card, and the coordinates of the area are shown in the printouts under the heading "Data Block(s) Histogrammed" on page 10 of the printout. Data in Channels 3 and 4 are displayed.

Use the printout from PICTUREPRINT to answer the following questions about displaying the data:

1. What are the line and column designations of the area histogrammed?

2. What line interval and sample intervals are used? (The interval is the third number in each set of parentheses. An interval of 1 means that every pixel is used; an

interval of 2 means that values in every other line and/or sample are used.) What percentage of the data was used to create these histograms?

Once the processor determines the maximum and minimum values in the data from each channel and assesses the spread of the data through histogramming, the data are divided into ten subsets or "bins," from the darkest values to the brightest.

Use the printout from PICTUREPRINT to answer the following questions:

1. What values fall into the "darkest" bin? What symbol is used to display values in this bin?

2. Compare some features in the first 100 lines, Columns 280-390, of the printouts of Channel 3 and Channel 4 data:

   The linear features in that section are highways. In Channel 3, are highways brighter or darker than the surrounding agricultural fields? What are the data ranges possible for each? How does the Channel 4 image depict the roads and agricultural fields? What data ranges are possible for each?

   Several agricultural fields within the first 30 lines of data are relatively dark (represented by X) in Channel 3 and very bright (blank) in Channel 4 data. We can therefore probably assume that they have the same type of crop. Can you find any other fields with the same spectral characteristics in both channels within the first 125 lines of data?

PICTUREPRINT provides a way to look at the individual data values in the scene, but in order to assess data quality and study the reflectance characteristics of the scene, it is necessary to have a black-and-white representation of the scene in each channel. These images, Figures II-4 through II-10, were created by photographing the screen of a digital image display system (COMTAL Vision ONE/20) on which the higher values are represented as light tones and the lower ones as dark tones. Prior to display, the data are histogrammed and appropriate functions calculated to display the data in an optimal way.

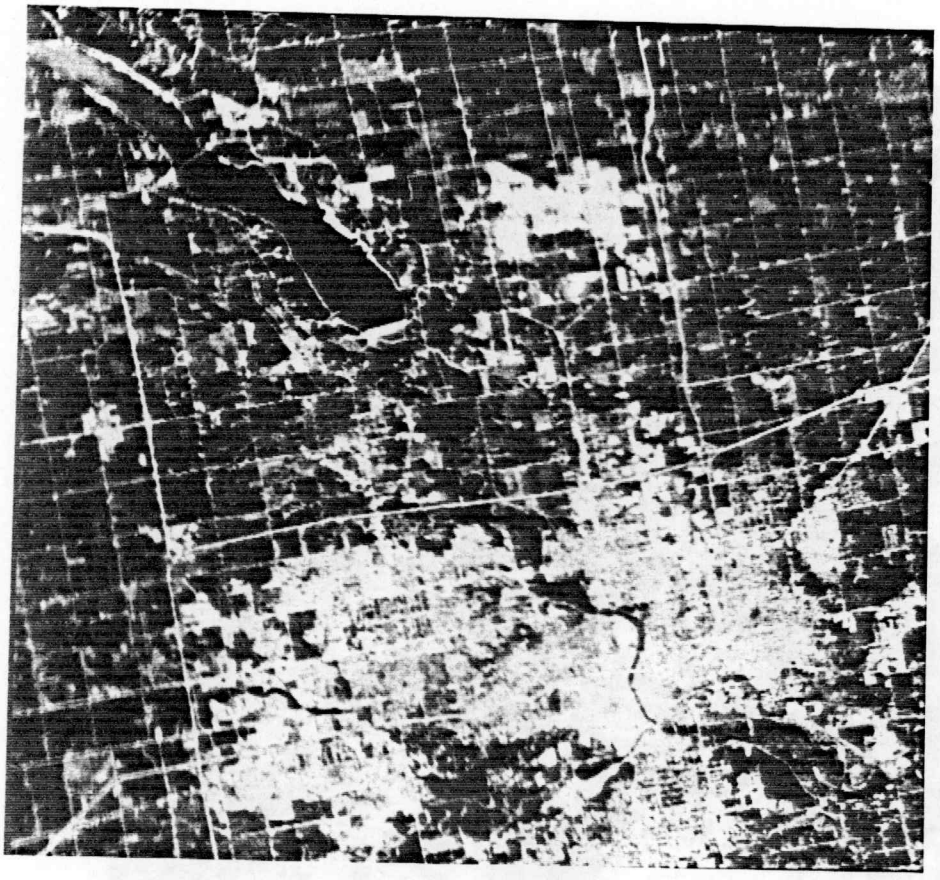

Figure II-4. Channel 1

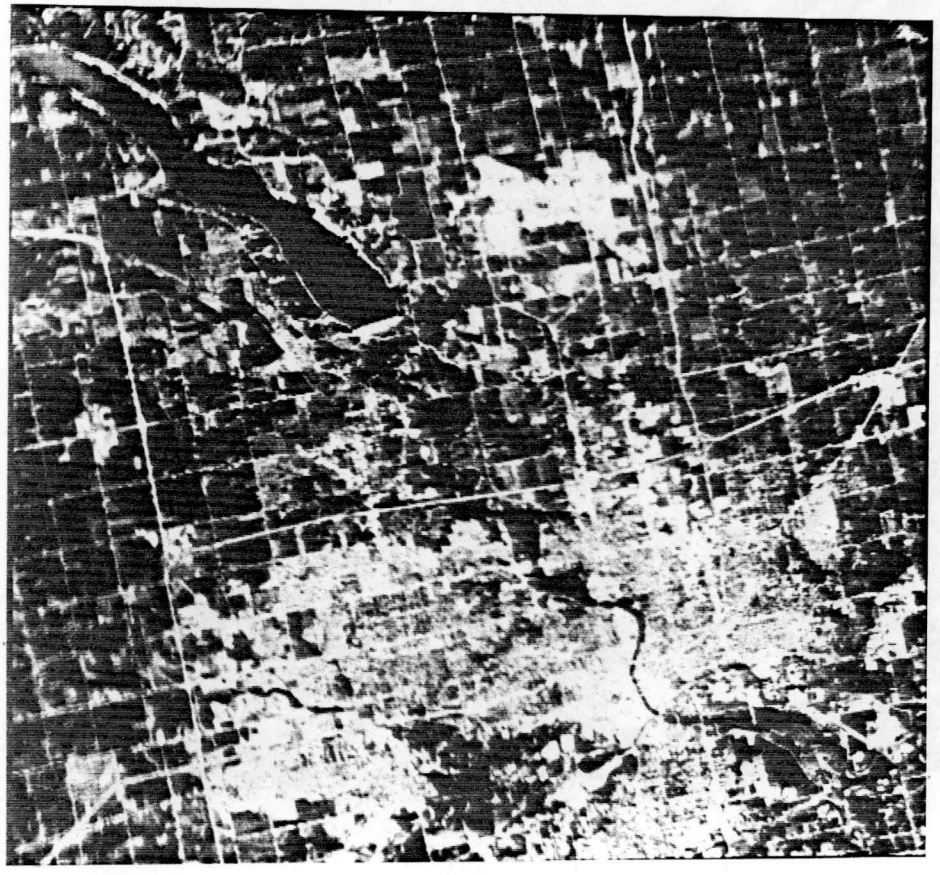

Figure II-5. Channel 2

Figure II-6. Channel 3

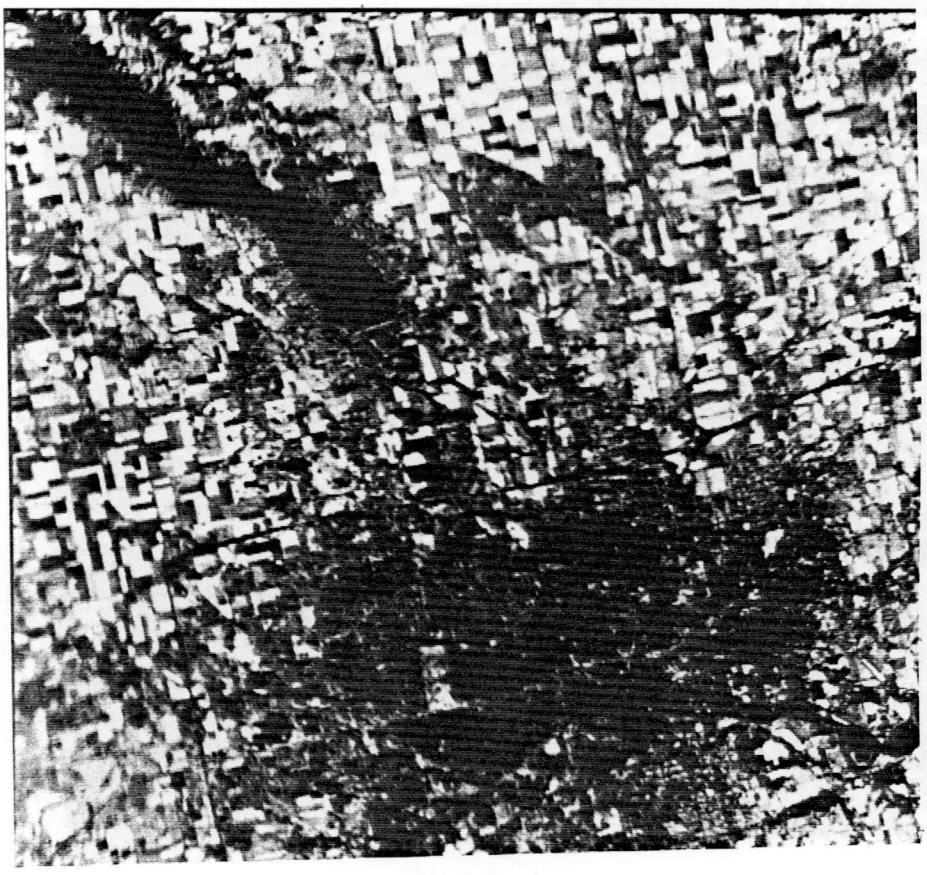

Figure II-7. Channel 4

Figure II-8. Channel 5

Figure II-9. Channel 6

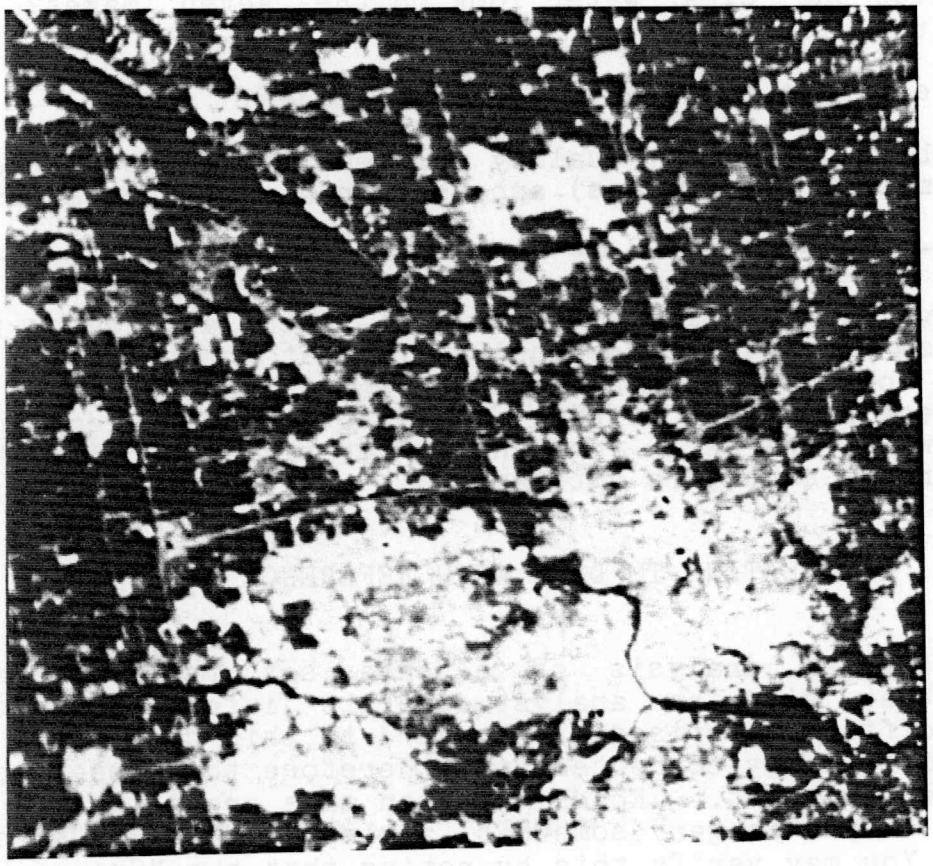

Figure II-10. Channel 7

Examine the images in all seven channels in order to answer the following questions:

1. Why are the images from Channels 2 and 3 similar to each other?
2. Do any other pairs of images appear to be as similar?
3. How does the clear water look on the seven images? How does the city look?
4. What else can you identify on the images?
5. Are there any data quality problems (such as clouds or radiometric distortions) apparent on any of the images?

The image of Channel 7 data has a different appearance from the others because the spatial resolution in that channel, the thermal channel, is one fourth that of the other channels. The data in Channel 7 were resampled so that the data could be geographically coordinated with the data from the other six channels, but the blurry appearance cannot be altered.

CORRELATING TM DATA WITH REFERENCE DATA

Completing an analysis of a data set requires maps and photographs that give the analyst accurate information about the area. One difficulty arises with regard to this data set: it has not been geometrically corrected and therefore the scene appears to slant to the northwest. Roads that on maps are aligned north to south appear on the imagery somewhat angled, about 10 degrees away from north. You may verify this by noting that the "Ground Heading" shown on the IDPRINT output is 191 degrees. A corrected Landsat data set would have a ground heading of 180 degrees.

The scale of the data has been set so that line printer printouts of the scene are at a scale of approximately 1:12,000. This information is useful when selecting and using reference data.

Reference data are available for the study area in several formats:

a) U.S.G.S. topographic map at a scale of 1:250,000 prepared in 1954, revised 1972 (Des Moines Quadrangle)

b) nine U.S.G.S. topographic maps, 7.5 minute series, at a scale of 1:24,000, some prepared in 1956, revised 1976, others prepared in 1972: Granger, Polk City, Elkhart, Grimes, Commerce and Des Moines NE, NW, SE, and SW

c) four 9x9 color infrared transparencies obtained on May 2, 1978, at a scale of approximately 1:74,500

d) low-altitude 35-mm color aerial photographs for selected agricultural areas obtained within one week of the Landsat overpass

e) plot maps at a scale of approximately 1:63,360 for selected agricultural areas

f) photo-based section maps with field identification numbers for selected agricultural areas at a scale of approximately 1:15,840

g) color and color-IR oblique aerial photographs obtained from approximately 5,000 feet on May 2, 1983.

It is not always possible for an analyst to have so many different types of reference data available for a study area, but each type contains different information which the analyst can use synergistically.

Using the images in Figures II-4 through II-10, spend some time correlating these images with the locations of several major geographic features: Saylorville, Des Moines urban and residential areas, the Des Moines River, the small town of Ankenny, the river patterns made up by the Des Moines and Raccoon Rivers, the major highways, and the agricultural areas.

1. On Figure II-7, the image for Channel 4 (near infrared), find and outline with a pen two or three examples of each major cover type of interest: agriculture, forest, water and urban.

2. With the aid of the low-altitude aerial photographs, find and outline an example of bare soil.

3. Do you have enough information to divide the city into commercial-industrial and residential? Which reference data helped the most?

SELECTING THE TRAINING AREAS

As you will recall, the Flexible Workshop manual describes three approaches to selecting the sample of data points that will be used to train the classifier: the supervised approach, the unsupervised approach, and a hybrid approach that draws on both. If you were using a supervised approach, the fields you outlined on Figure II-7 would be a good starting point for training the classifier. The procedure that will be demonstrated in this case study will be a hybrid approach that includes elements of both the unsupervised and the supervised approaches.

To begin this process, the analyst needs to select candidate training areas, areas that can provide the data needed to train the computer to recognize the classes of interest. Recall 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 seven-channel data we are using, you should select four to six training areas each containing up to 6241 points. For example, fields could be as large as 76 lines by 79 columns. Training areas of at least 50x50 pixels are easier to correlate with reference data than areas that are much smaller. It is helpful, too, to include some easily identifiable landmark such as a highway intersection, an irregular coastline, or other highly visible features.

With the aid of the reference data, select six training areas according to the guidelines given above. Make sure that every cover type of interest (agriculture, water, forest, urban, and bare soil) 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 3 image. 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 to be used in the classification of the Polk County study area. We will continue demonstrating the hybrid approach in this analysis, i.e., a combination of supervised and unsupervised approaches.

CLUSTERING THE TRAINING SAMPLE

No two analysts could 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 choices may be equally valid. In this case study, analysts have chosen the areas outlined on Figure III-17. Together they were chosen to represent all the major cover types:

Area 1: agricultural, river valley, highway
Area 2: river valley, highway
Area 3: city center, highway, river valley
Area 4: suburban, new residential
Area 5: agricultural, highways
Area 6: highways, agricultural
Area 7: several water classes.

> Locate the seven training areas outlined on Figure III-17. Become familiar with the major features of each cluster area.

The training samples for water (Area 7) are a combination of several small rectangles. These were chosen because the analyst wanted to be able to identify several classes of water in the reservoir. This selection serves to represent another approach that is possible when clustering: using the processor to identify subclasses of a single, basic cover class.

Each of the seven training areas (including Area 7 made up of several small rectangles of data) is analyzed separately by the CLUSTER processor. Once the sample areas are chosen, the analyst has two parameters to set when running the processor: the number of clusters that the sample should be divided into and the convergence level. The analyst selected the following parameters for the seven areas:

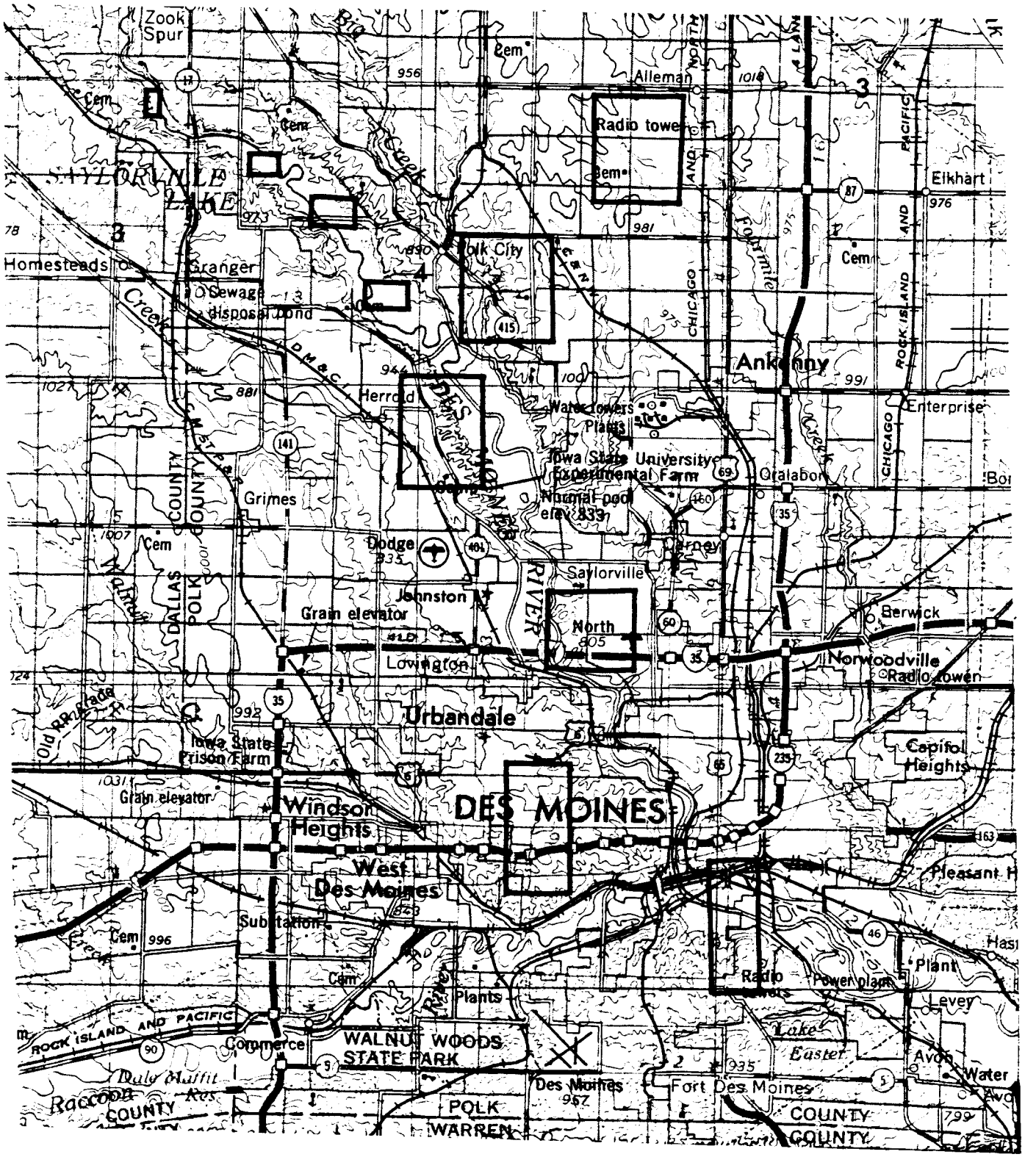

Figure III-17. The seven areas outlined above provided the data analyzed by CLUSTER.

Area	Number of Classes	% Convergence
1	18	98.5
2	16	98.5
3	10	98.5
4	12	98.5
5	15	98.5
6	13	98.5
7	7	99.5

In general, when an area is more complex, that is, when it contains a wider variety of spectral classes, a greater number of classes is requested. For areas that are more homogeneous, such as Areas 3 and 7, fewer classes are requested. It is common practice to ask for two times the number of clusters as there are information classes, plus a few to allow for transitional zones.

Notice that the Area 7 clustering is carried out to a convergence of 99.5%, instead of 98.5 as are all the others. Because all of the pixels clustered are water samples and this is the smallest sample clustered, using the higher convergence tends to make the seven classes more distinct from each other. While the convergence could have been set higher for the other areas, the larger number of samples would greatly increase the processing time, and when classes are spectrally very dissimilar, experience shows that a higher convergence level does not significantly improve the final classification.

Output from the CLUSTER and SEPARABILITY processors falls on pages 83 to 207 of the printout book, with the output for Area 1 on pages 83-147. The CLUSTER output provides important statistical information about the spectral classes that the sample pixels were divided into. First, the mean value of each cluster class from the first area and the class variance of each area are shown on page 85. Also listed is the number of pixels ("POINTS") assigned to each cluster. In general, the first clusters listed have higher mean values than the latter ones, resulting in a general sequencing from bright (higher mean values) to dark (lower mean values). The variance values indicate the spread or dispersion of the data in each channel and serve as a guide to the spectral purity of each class.

The "cluster map" obtained from clustering the data in Area 1 is shown on pages 86 and 87. When you compare the cluster map with the topographic maps and the photography, you can start to identify the ground classes that some of the spectral classes represent. For example, Cluster 18 (designated on the map by the symbol W) represents water (it's a coincidence that the symbol W was assigned!), and Cluster 7 (designated by I) represents forest. More detailed identification of the classes will be a major task later in this section.

Pages 88 through 141 contain histograms of the data values from each of the seven channels for each of the 18 clusters, that is, the channel-by-channel distribution of the data values from the pixels that were grouped into each cluster by the clustering processor. 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 that its distribution is approximately normal. Note that distributions for Cluster Class 1, Channels 1, 2, and 3, are all approximately normal, but that histograms for Cluster Class 2, Channels 1, 2, and 3 (page 91), do not follow the normal pattern. If you look at the list of class sizes shown on page 85, you will note that this class with its 90 points is one of the smallest of the eighteen classes. It is not uncommon for small classes to have bi-modal histograms. While this deviation from normal is not great enough to require reclustering, we will take this distribution and the class size into consideration later when we make decisions about which cluster classes will make up the final set of training classes.

Another aspect of class distribution can be seen in the histograms: the relative variance of the data. Note that the histograms for Class 2, pages 91-93, appear wider than many of the others. This same fact is reflected in the variance values recorded for each channel and class on page 85. A smaller variance indicates that the data fall relatively closer to the mean of the class in that channel. Generally, classes with low variances are easy to separate from other classes that are spectrally similar.

Following the CLUSTER output for Area 1 is the output from SEPARABILITY for the same data. SEPARABILITY provides information about the statistical distance between every possible pair of the spectral classes (clusters) that were extracted by CLUSTER when the data in all seven channels are used. The distance measure used is transformed divergence; reread page III-13 for an explanation. Turn to page 143. By reading across the lines, you can learn the transformed divergence value for every pair of classes. The value 2000 indicates complete separability; lower values indicate a greater similarity between classes and, hence, a greater chance of misclassification between the two classes. Class A, the first of the 18 clusters, is clearly different from every other class (B through R) except for Class E, Cluster 5. The distance between Class B and the remaining 16 classes is shown by the following transformed divergence values shown on the printout.

The results of clustering the data in areas two through seven and measuring the distance between the resulting spectral classes in each area are shown on the printout pages 148-207. The output follows the same format as the CLUSTER and SEPARABILITY output just discussed: class mean value for each channel, class variance for each channel, number of points assigned to each class, a cluster map for each area, with separate "maps" for each of the small rectangles selected for Area 7. Because of the bulk of the output, the histograms have been omitted for all areas after the first. The transformed divergence values follow each CLUSTER results.

Examine the output from the CLUSTER and SEPARABILITY processors for Training Areas 1 through 7, pages 83 through 207.

1. Identify as many of the cluster classes in Area 1 as you can as to general cover material: vegetation, water, urban, or bare soil. Use the mean values on page 85 to assist you. Make the identification by comparing the relative values in each band to the known spectral reflectance characteristics of basic earth surface features. (The response values shown here have not been calibrated to allow accurate band-to-band comparisons to be made; however, general patterns can still be seen.) Write your conclusions in the first column of Table III-1.

2. Note any clusters with fewer than 70 points, 10 times the number of channels.

3. Examine the variances associated with each cluster class and note on the table any unusually high values, i.e., above 50.

4. Check the cluster map for any obvious groups of symbols that correspond with features appearing on the reference data. Outline these features on the cluster map and, if needed, modify or refine the class identifications you previously entered on Table III-1.

5. Examine the 126 histograms for Area 1. Note any obviously non-Gaussian distributions. Select a cluster class with generally narrow histograms and compare its variance values with those from a class with wider histograms. Note non-Gaussian classes on Table III-1.

6. Analysts must go through the identical steps for the cluster results from the remaining six areas. Select one or more areas, as your time permits, and enter your results on Table III-1.

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

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

ASSOCIATING CANDIDATE TRAINING CLASSES WITH INFORMATION CLASSES

In the previous section, you used mean values and cluster maps to do a preliminary identification of the spectral (cluster) classes resulting from clustering. To refine this identification as completely as possible, use the reference data to work through the classes one by one to identify them to the next level of detail. For example, you should now be able to separate the vegetation classes into forest and crops; in fact, you may be able to distinguish between various types of crops and between crops and grasses.

Remember that the correspondence between information classes and spectral classes is not necessarily one-to-one. Most often, more than one spectral class can represent a single information class, as in the case of a single crop at various stages of maturity; occasionally (we hope, rarely) more than one information class is associated with a single spectral class. When this occurs, dual class names must be given to the spectral class, such as "bare soil and emerging crop" or "bare soil and highway."

Use all of the reference data available (maps and aerial photographs) as well as the statistical information to refine the identification labels you have given to the eighteen classes in Area 1. Record your conclusions on Table III-1, replacing the more general identifications you made earlier.

Note that these TM data have not been geometrically rotated to a N-S direction, which complicates somewhat your work in correlating the printouts with the reference data. After a short time of working with the maps, you will become accustomed to this distortion. As you are making comparisons, keep in mind the month or season during which each type of data was obtained in order to account for seasonal variations in appearance.

As time permits, perform the same refinement of identification labels for one or more of the other areas.

AUGMENTING THE CANDIDATE TRAINING SAMPLES

Now that you have assigned class names to the spectral classes produced by clustering, it is important to consider whether the training sample is truly representative of the scene. Are there any surface features that are not included in the training sample?

When completing this analysis, analysts felt they had represented all classes adequately. Therefore, instead of augmenting the samples at this point in the procedure, they followed an alternate strategy, which you'll read about later: they did a test classification of a small portion of the data. From that they discovered the need to add additional samples. Procedures for augmenting training samples will, therefore, be discussed later in this description.

VISUAL REPRESENTATION OF CANDIDATE TRAINING CLASSES

The next step in the process gives us a set of tools that are extremely helpful in the task that follows: we will look at a number of visual or graphic representations of the spectral classes and use this information to help select the final training classes.

Previously you assigned information class names, to the detail possible, to each of the 91 spectral classes created by clustering (or as many of them as you had time to complete). Table III-2 lists the labels assigned to each class by the analysts working on this study. Because there were no reference data available for Area 6, analysts were unable to assign class names.

As you look across the list, you'll notice that many of the same materials occur in more than one of the training samples; for example "forest" occurs as Classes 5, 6, 7, and 8 in Cluster Area 1, as Classes 10, 11 and 12 in Cluster Area 2, and as Classes 8 and 9 in Cluster Area 3. In order to reduce the number of candidate training classes from 91 to a number that reflects the actual number of classes we wish to identify, we need to delete and/or combine some of these redundant classes.

The processors that can assist in showing the relationships among all 91 classes are MERGESTATISTICS and SEPARABILITY. Because of difficulties involved in working with 91 classes, analysts first merged the statistics files from Areas 1, 2, and 3 and looked at those 44 classes together.. The output begins on page 208 of the printouts. The results of this process is that the three separate statistics files (from Clustering Areas 1, 2, and 3) are merged into a single statistics file containing 44 classes, listed on page 209. As before, SEPARABILITY was run on this file of 44 classes, to determine the transformed divergence value between every possible pair in the 44 classes when data in all seven channels are considered. The list of classes on page 211 identifies the 44 symbols used to represent the classes. Note that because of the limited number of symbols available, 14 symbols (A through N) are used twice in the following output. For example, the symbol "K" will represent the eleventh spectral class from Area 1 and the seventh spectral class from Area 3. You will find it useful here and later to add the subscript "2" to the second occurrance of each

Table III-2. Analysts' assignment of information class names to the 91 spectral classes created by CLUSTER.

Cluster	Area 1 Symbol ID	Area 2 Symbol ID	Area 3 Symbol ID	Area 4 Symbol ID	Area 5 Symbol ID	Area 6 Symbol ID	Area 7 Symbol ID
1	1 ¥ Agricul.	19 ¥ Soil	35 ¥ Concrete	45 ¥ Industrial	57 ¥ Soy	72 ¥ Bare Soil	85 ¥ Water
2	2 - Beach	20 - Agricul.	36 + Urban	46 . Urban	58 - Soy	73 .	86 / Water
3	3 + Bare Soil	21 = Substation	37 1 Urban	47) Urban/ Highway	59 = Soy	74 /	87 I Water
4	4 / Soil/Veg.	22) Quarry	38 L Industrial	48 7 Industrial	60 1 Soy	75 J	88 * Water
5	5 J Forest	23 7 Bare Soil	39 V Urban/ Highway	49 2 Residential	61 C Soy	76 I	89 E Water
6	6 C Forest	24 I Soil/Veg.	40 Y Urban/ Highway	50 V Residential	62 I Soy	77 Z	90 $ Water
7	7 I Forest	25 S Grass	41 F Residential	51 Y Residential	63 Z Soy	78 * Bare Soil	91 M Water
8	8 Z Forest	26 3 Soil/Veg.	42 B Forest	52 8 Grass (dry)	64 * Grass/ Farm	79 T	
9	9 3 Soil/Veg.	27 & Agricul.	43 O Forest	53 K Urban	65 T Road/ Farm	80 E	
10	10 & Drainage	28 X Forest	44 W Soil (wet)	54 N Residential	66 F Farm/ Grass	81 H	
11	11 T Beach	29 E Forest		55 D Highway/ Resid.	67 G Corn	82 $	
12	12 F Soil/ Highway	30 K Forest		56 W Residential	68 B Corn	83 D	
13	13 G Beach	31 N Marsh			69 R Corn	84 W Water	
14	14 H Beach	32 R Soil (wet)			70 Q Corn		
15	15 $ Drainage	33 Q Soil (wet)			71 W Wheat/ Residue		
16	16 O Soil (wet)	34 M Water					
17	17 Q Beach						
18	18 W Water						

symbol to keep the classes separate. The SEPARABILITY output extends from page 210-230.

Since we have merged classes that originally resulted from several CLUSTER operations, we can expect much lower transformed divergence values than we had for the original class pairs. These low values are important in locating class pairs that are too similar to be distinguished between. The analyst must therefore either delete one of the two classes or pool them. These steps will be discussed later.

In addition to the transformed divergence values, the analyst also has available several ways of displaying the 44 classes graphically. The BIPLOT processor has as one of its output products a two-dimensional plot; on one axis is plotted the mean value of each class in one channel, and on the other, the mean values in a second channel. Analysts working with TM data found that the most useful combinations were Channel 3 vs. Channel 4, because of its similarity to equivalent plots from the more familiar Landsat MSS data, and Channel 4 vs. Channel 5, because of the many spectral differences that are emphasized in Channel 5.

Pages 234-244 of the printout contain the bi-plots for the 44 candidate training classes in the first group of statistics. On page 232-233 is a legend giving the symbol used to represent each class on the plot and the class number from the original clustering.

Because of the large number of candidate training classes, some symbols appear twice on the bi-spectral plot. You will need to distinguish between the two A's, B's, etc.

Use the list on pages 232-233 and the previous CLUSTER output to locate the letters representing classes A (13/16) through N (10/10) on the plot and add a small subscript 2 to the symbols as they appear on the plot. For example, the second symbol A represents class 13 in cluster area 2. Page 150 of the CLUSTER output tells us that that class has a mean value in Channel 3 of 18.25 and in Channel 4 of 65.09. Locate that A on the plot on page 234-235 and label it A_2. Using these values, there should be no question in your mind which A is A_2. Do the same for classes 32 through 44.

To increase the usefulness of this plot, write the name of each class, as shown on Table III-2, next to the symbol for as many classes as possible. You will note that the soils tend to fall to the upper right, dense vegetation to the left of the major diagonal, and water in the lower left.

The final technique we will use for providing a graphic representation of the candidate training samples is shown in Figures III-18 and III-19, plots of the class mean values. Viewing the class means plotted in this way allows you to confirm the general cover type of each class. Because of the large number of classes, only selected classes have been displayed. Once you gain a familiarity with the basic shapes of the plots, you will be able to recognize vegetation, soil, water, and other surface materials and to have another way of observing spectral similarity.

Study the plots of the means for the eight classes represented in Figure III-18. The classes represented here depict different types of classes: water, beach (actually a very wet soil), soil and highway mixed, soil, concrete, forest, urban and highway mixed, and residential. Use the shape of the plots and your knowledge of spectral reflectance characteristics to identify each curve as one of the materials listed above.

___ water
___ beach
___ soil/highway
___ soil
___ concrete (a substation)
___ forest
___ urban/highway
___ residential

Use Table III-2 to verify your answers.

In the example above, plots were selected to demonstrate rather typical response patterns for major surface materials. You will recall, however, that there is in the statistics file information about several sub-classes within the various groups. As an example of the way families of classes may have similar characteristics, six urban/residential classes are plotted on Figure III-19. These plots indicate that while there are family similarities among the classes, there are also differences which, if significant enough, can be used to distinguish among some of the classes.

Study the six classes plotted in Figure III-19. Which plot represents the spectral class with the most vegetation? Which class would be related to concrete surfaces? What other sub-classes would you expect to find in the urban/residential complex?

Verify your answers with the information on Table III-2.

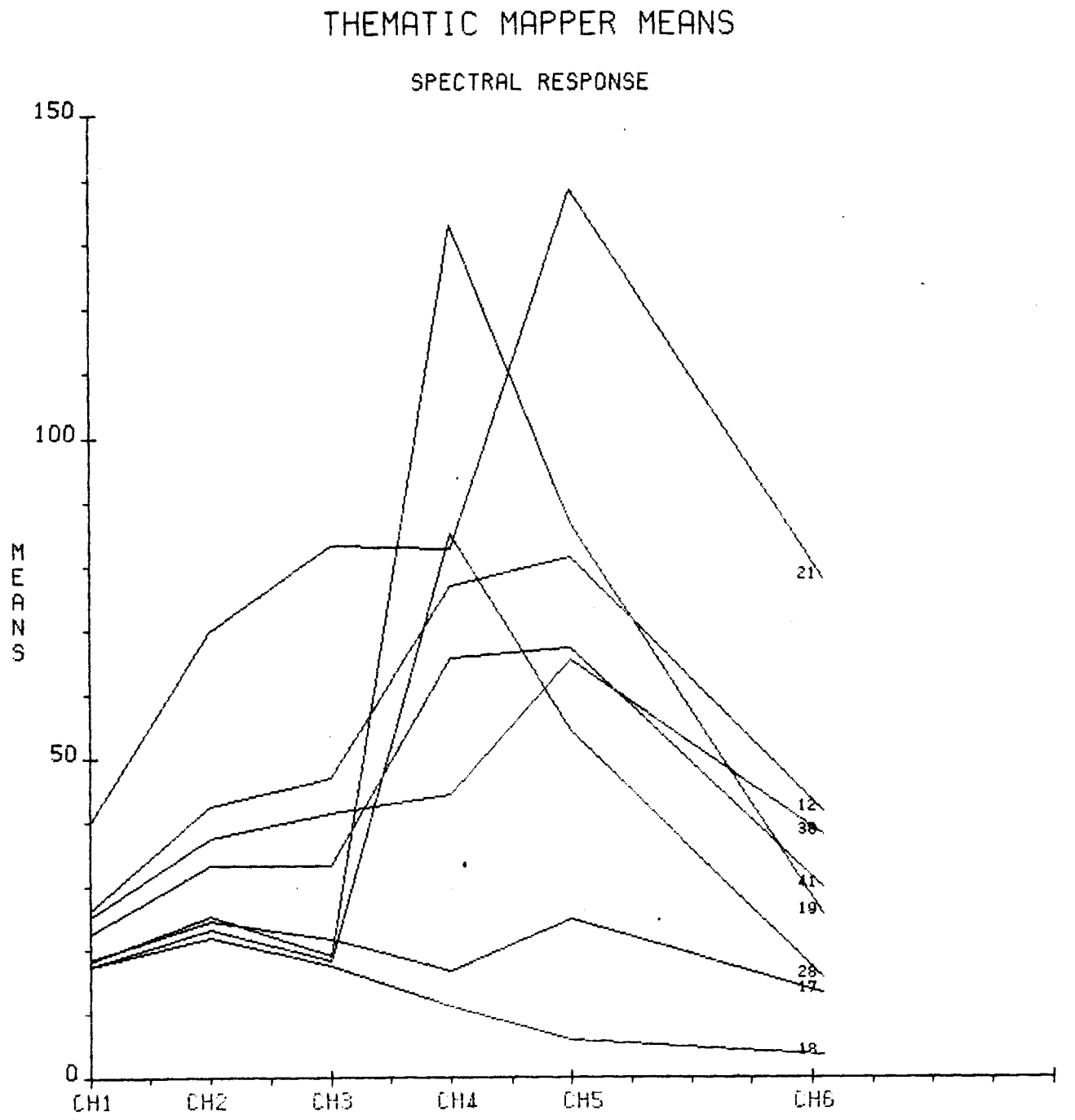

Figure III-18. Plots of class means for eight spectral classes.

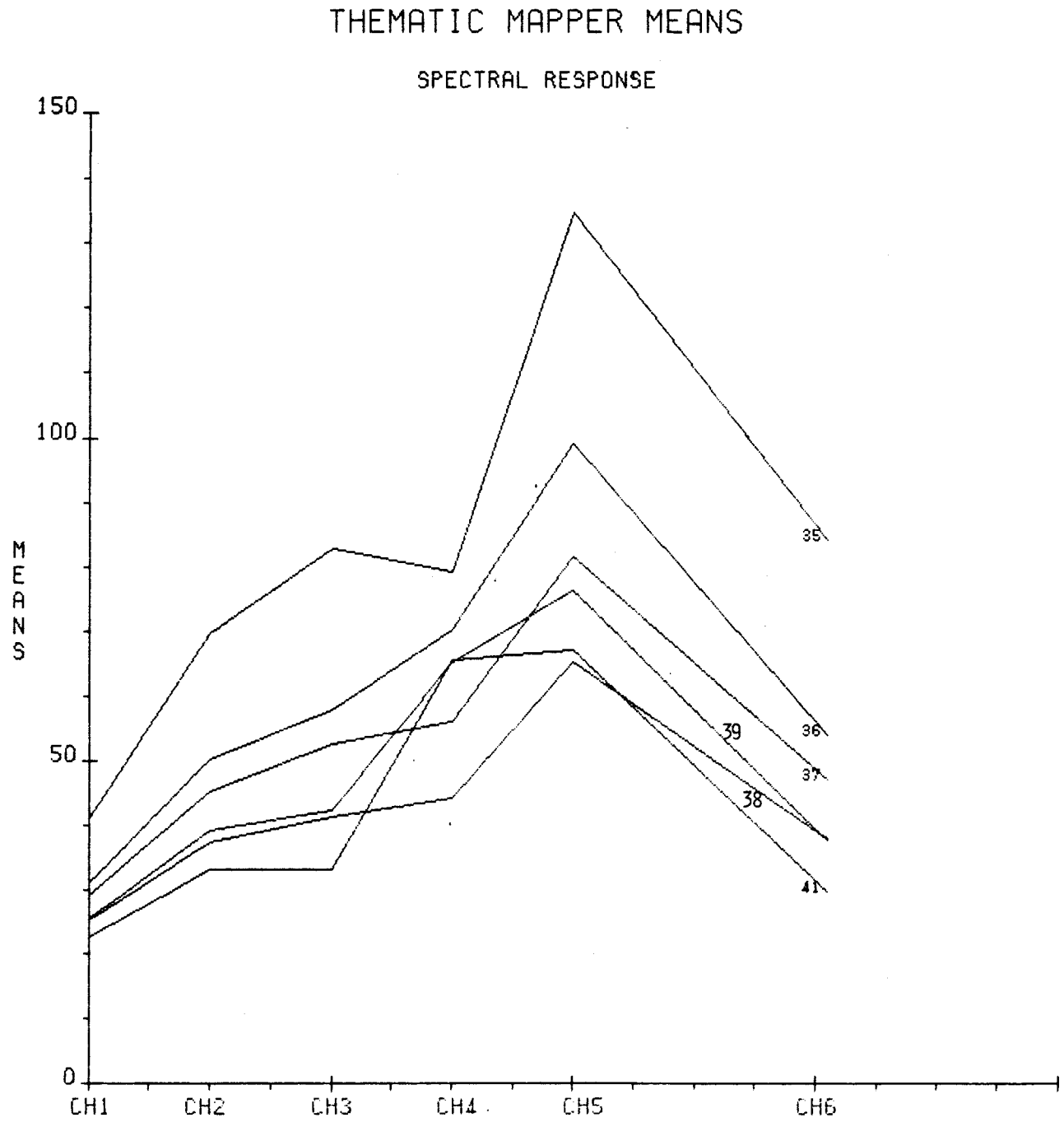

Figure III-19. Plots of class means for six spectral classes all in the urban/residential family.

The plots of the class means are introduced here because of their value as an analysis tool. Along with the output from CLUSTER, the graphic from BIPLOT, and the transformed divergence values from SEPARABILITY, the plots of the class mean values provide the analyst the tools needed for the next step in the analysis, refining the training statistics.

A DIGRESSION: OVERVIEW OF THE ANALYSIS PROCESS

The major task of the analyst at this stage of the process -- in fact, the step more dependent than any on the analysts' skill and understanding -- is the step to be studied now: refining the training classes. Before we move to that phase, however, it would be useful to take a look at the analysis steps we have already discussed in this chapter and the ones that are yet to come. The flowchart in Figure III-20 outlines the steps used in defining the spectral training classes.

First the seven-channel data in all seven candidate training areas were clustered, yielding seven statistics files. Because the total number of classes created was greater than the software can accommodate, the files from Areas 1, 2, and 3 were processed together and those from Areas 4, 5, and 6 were processed together. The discussion here has been limited so far to the files from Areas 1, 2, and 3. You will recall, for added information, the analyst ran SEPARABILITY for each statistics file, and then the three statistics files were merged into one statistics file with 44 classes. SEPARABILITY was run again on the 44 classes to determine which classes were not spectrally separable.

At this time in the analysis, we are ready to decide which classes to pool and which to delete, and in preparation for this, we have looked at some ways to represent graphically the spectral classes and their relationships to each other. As you look down the flowchart you will see that the analyst chose to delete nine of the 44 classes, leaving 35 candidate training classes from Areas 1, 2, and 3. We will look at the reasons behind these choices soon.

The same procedure was followed with the statistics files from Areas 4, 5, and 6: the three files were merged into a single 40-class file (printout pages 247-248). Upon study of the characteristics of these classes, sixteen of the 40 classes were deleted (pages 249-250), leaving 24 classes that were available to merge with the 35 classes resulting from Areas 1, 2, and 3 (pages 253-254). Further refinement of these 59 classes resulted in a statistics file with 37 classes (pages 255-256). As we shall see later, the last steps in developing the training statistics involve adding the water classes from Clsuter Area 7 and using the supervised technique to add three classes that were discovered not to be represented in the statistics. A final refinement resulted in the 42 classes used in the classification.

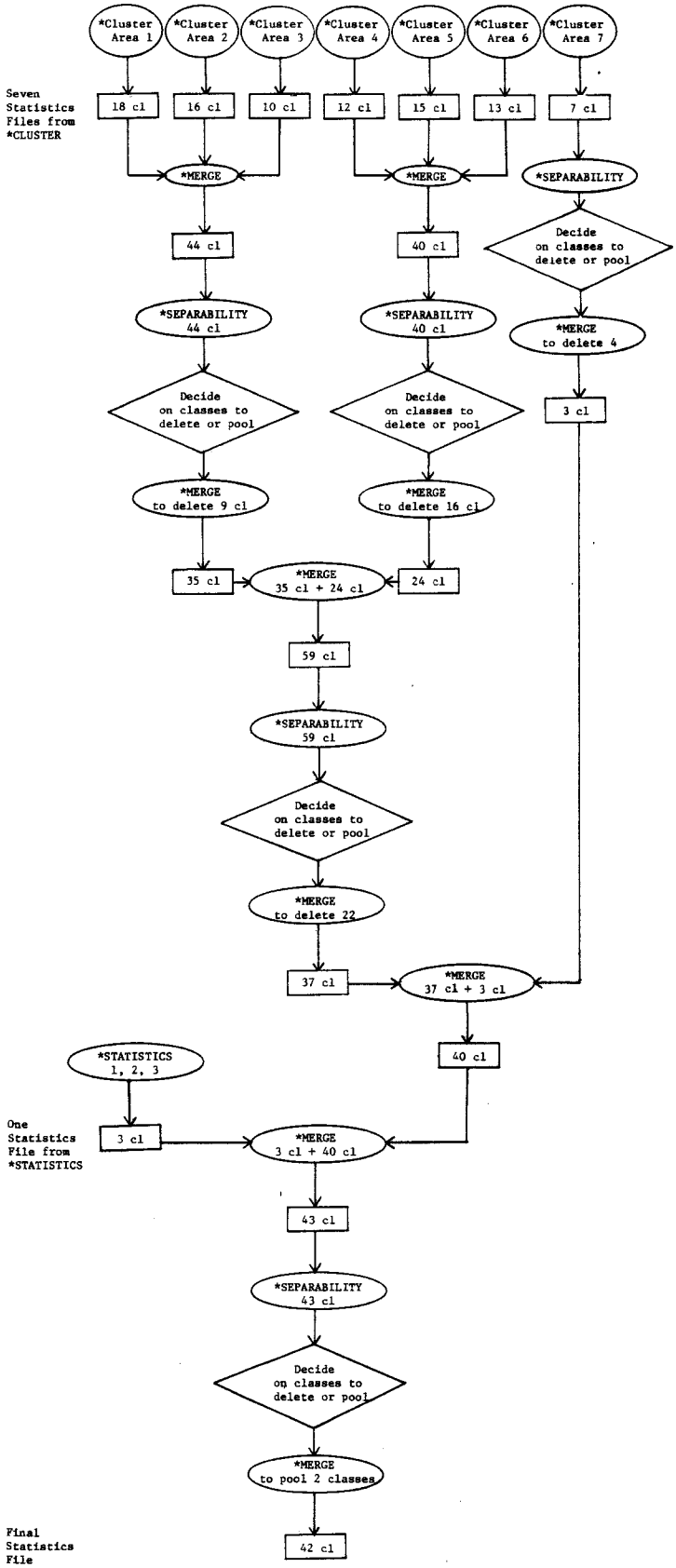

Figure III-20. Flow chart of the steps involved in the development of the final training statistics.

Now, back to the step-by-step discussion that was interrupted by this digression.

REFINING THE SPECTRAL TRAINING CLASSES

As aids in selecting the best spectral classes to retain as training classes, we now have transformed divergence information, two bi-plots (one labeled with class identities) that indicate pairs where the transformed divergence value falls below 1750, and the plots of the calibrated class means. In addition the CLUSTER output contains valuable information about class size and variances.

The two major goals we must seek to satisfy when selecting the training classes are:

1. As a group, the final training classes need to represent everything in the scene, and
2. The training classes need to be spectrally separable from each other.

We can make use of two different strategies when refining the training classes: first, the pooling approach, which 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 lie on the borders between information classes are deleted. The pooling approach more generally accomplishes the first goal, representation, while the second helps achieve the second goal, separability. Experience has shown that the best approach is a combination of these two strategies.

One of the outputs of SEPARABILITY (pages 227-228) is a list of all class pairs with a transformed divergence value of less than a selected level, in this case 1750. Looking down that list, you see the greatest problems, the lowest values. For example, look at the pair "DY," Class 4 from Area 1 (4) and Class 7 from Area 2 (25). With such a low transformed divergence value, it is not surprising that they both have the same identity: soil/veg. Figure III-21 verifies the strong spectral similarity of the two classes. In this case the analyst chose to retain Class 25 and delete Class 4; although the classes are approximately the same size (359 and 436 points respectively), the variances of Class 25 (given on the CLUSTER output) are lower than those of Class 4.

Let's look at another pair of classes with a low transformed divergence: Class 9 (Area 1) and Class 27 (Area 2). The transformed divergence value between them is 855, and the similarity of their mean values is shown graphically in Figure III-22.

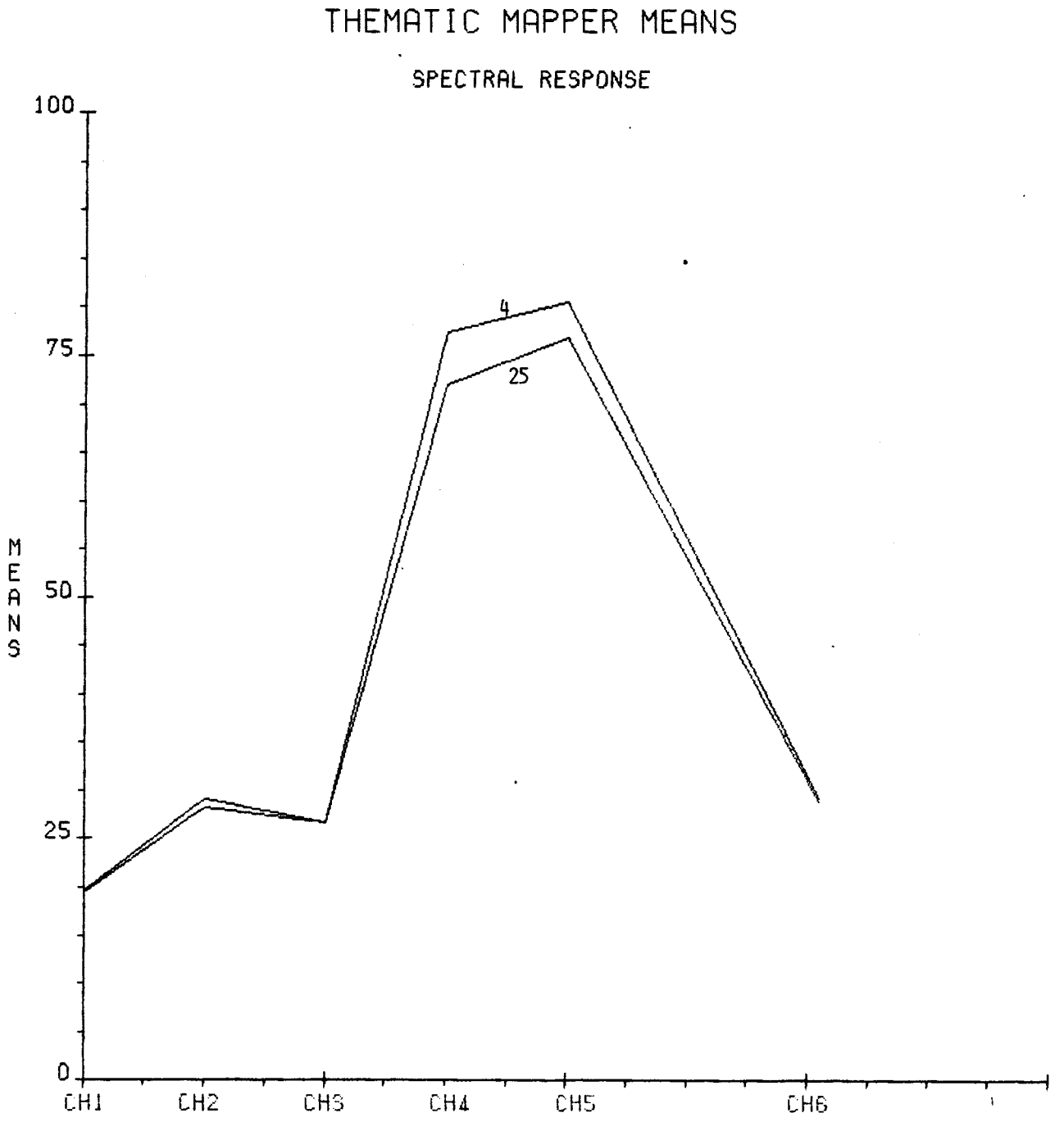

Figure III-21. Plot of class means for two classes both labeled soil/vegetation. Transformed divergence = 581.

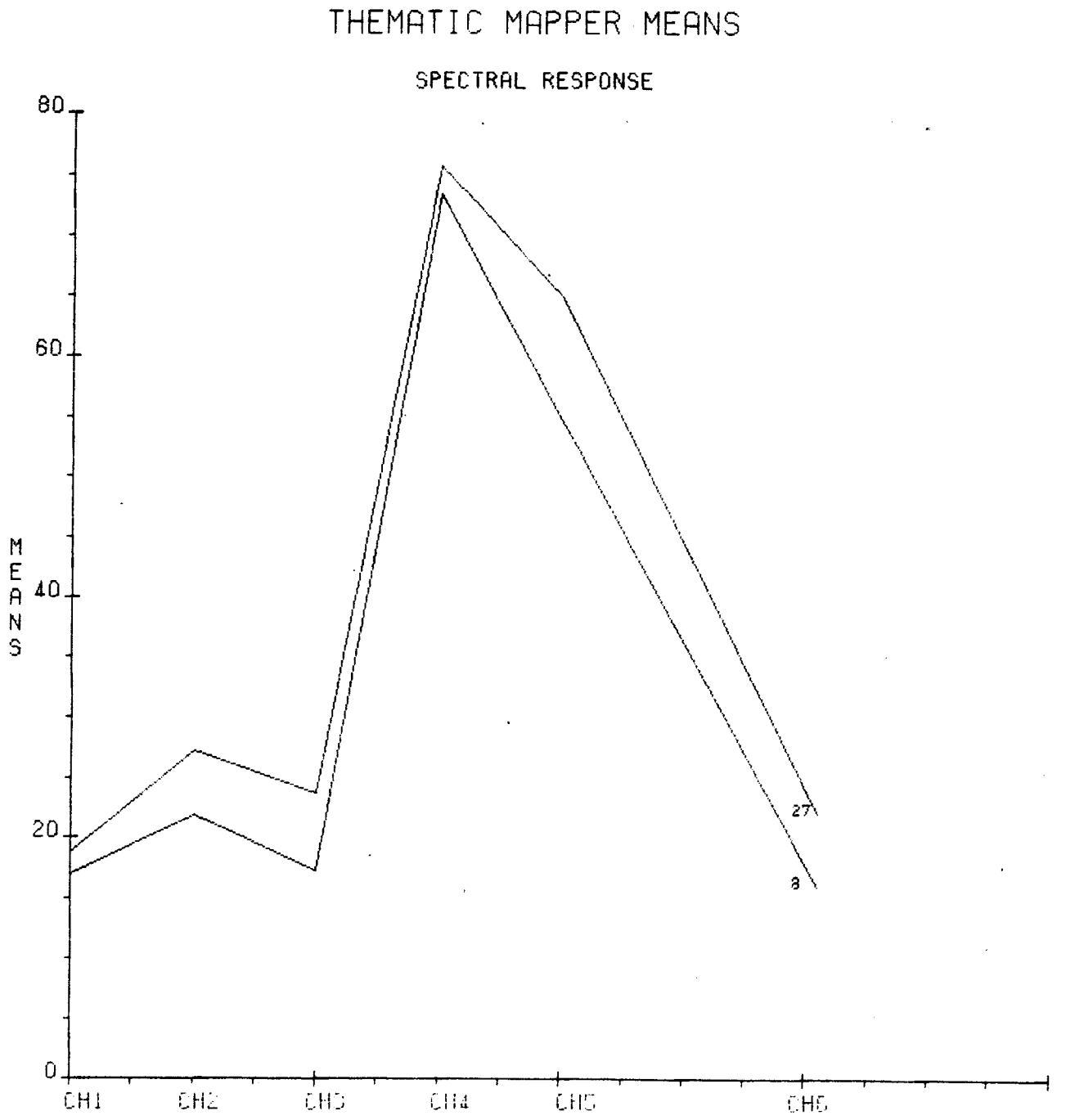

Figure III-22. Plot of class means for two similar classes. Transformed divergence = 855.

Study the available output to decide which of the two soil/vegetation classes you would delete. Why?

Look at the control cards for MERGESTATISTICS, page 245. Which of the two classes did the analyst delete?

Let's look at another example of a pair of classes in this group of 44 that are in the same family and are spectrally similar: Class 36 (Class 2, Area 3) and Class 40 (Class 6, Area 3). The former has been labeled "urban" and the second "urban/highway." The transformed divergence is 1547, not as low as the pairs studied earlier, but too low for discrimination. Plots of the class means are shown as Figure III-23. Because of the crossover of the lines on the plot and the spectral differences indicated, the analyst decided to keep both classes for the time being. As you will see later, however, both were deleted before the final classification.

As a final example of a pair of classes that are very similar, look at Class 7, Area 1 and Class 10, Area 2. Both are identified as forest and are spectrally very similar.

Look at the SEPARABILITY output beginning on page 210 and find out what their transformed divergence value is. Class 7, Area 1 is represented by "G" and Class 10, Area 2 by "+".

Study the output from the two CLUSTER runs (beginning on pages 83 and 147). Look at the number of points, variances, and spatial distribution of the two classes to decide which class to delete and which to keep.

The analyst who carried out this analysis noted that the transformed divergence value was 949 (page 216 of printout) and that the points in the Area 1 map (page 87) are much more concentrated than are those from Area 2 (page 151). As a result of this, he deleted Class 10, Area 2.

Classes 11 and 13 from Cluster Area 1 were deleted because of their small size, 17 and 10 points respectively.

The refining process continues, a slow process requiring close scrutinizing of information about the classes. In this phase the analyst eliminated nine classes, leaving 35 to merge later with statistics from Cluster Areas 4, 5 and 6. When classes are similar,

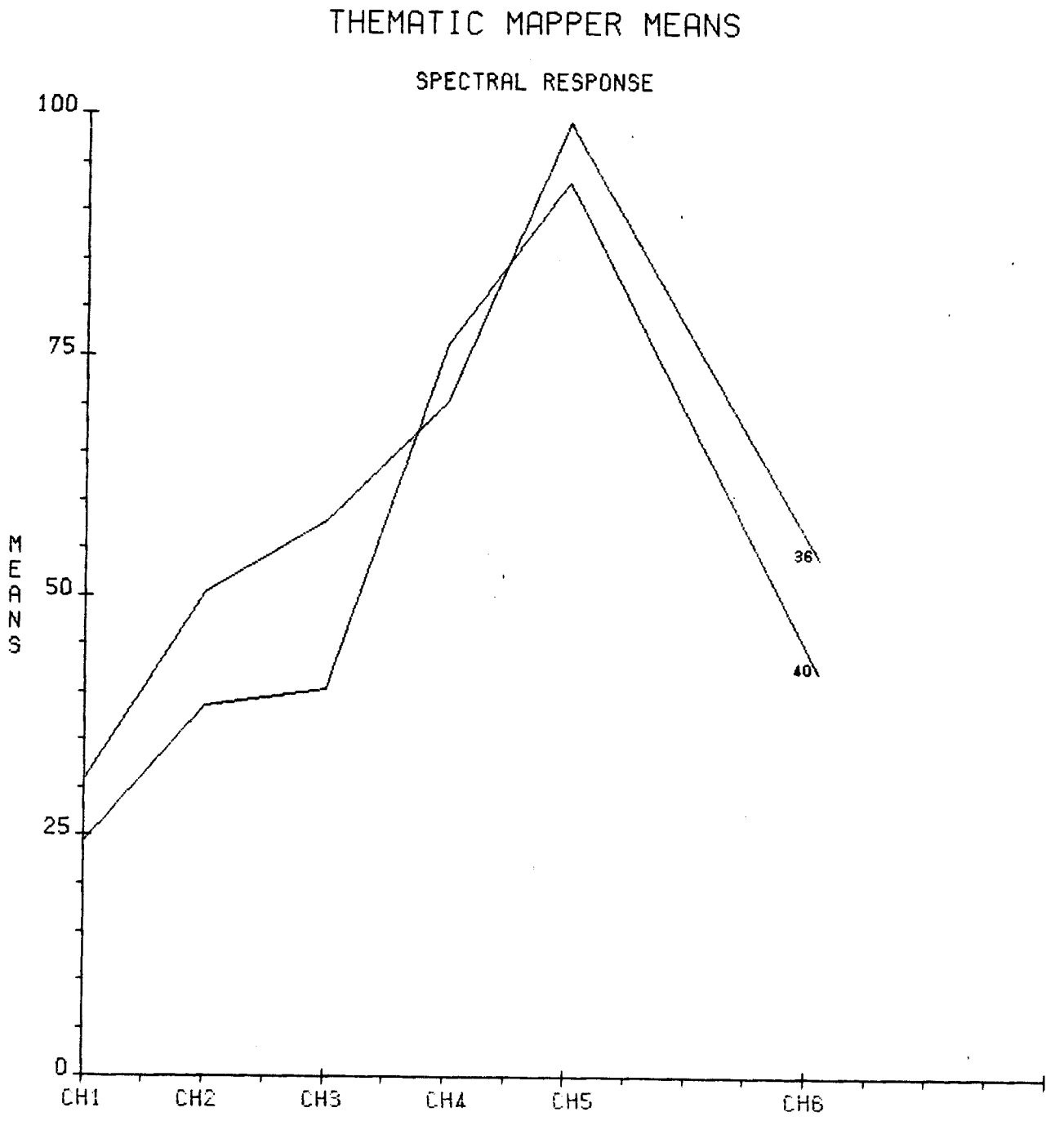

Figure III-23. Plots of class means for two classes of urban. Transformed divergence = 1547.

deletions are made for the following reasons: a small number of points, wide variances, and a lack of concentration of samples.

The flowchart (Figure III-20) demonstrates how the analyst proceeded. The statistics files from Areas 4, 5, and 6 contained 12, 15, and 13 classes respectively. The identity of many of the classes from Area 6 is unknown because of a lack of reliable reference data. The analyst merged the three files, creating a file with 40 classes. Through a similar process of close study of the classes, the analyst decided to delete 16 classes, leaving 24 to merge later with the 35 available from Areas 1, 2, and 3.

The printout on pages 253-254 shows the results of that merge; subsequently 22 of these classes were deleted (pages 255-256), leaving 37 classes.

It's time now to turn back to the seven water classes that came from clustering the four small rectangles in Area 7. As you will recall, the analyst asked for the data to be divided into seven classes. The results from this run of CLUSTER are shown on pages 197-203, with the SEPARABILITY information following.

Look at the output on pages 197-203 and the transformed divergence values to analyze the output and decide if any classes can or should be deleted. List here the classes you would delete and the reasons for your decisions.

Classes Deleted	Reason

The analysts carrying out this analysis chose to delete four classes, reducing the final number of water classes from this file to three, classes 1, 2, and 3. The reason for this decision was that he looked at these seven classes in conjunction with the water classes from Areas 1 and 2 (Classes 18/18 and 16/16). These are good classes that represent water with the lower response. The curves of the final 5 water classes are shown in Figure III-24.

The MERGESTATISTICS run on pages 251-252 accomplished the deletion of the four water classes and a new statistics file based on the three selected classes. This file is now ready to be incorporated with all the other statistics files from Cluster Areas 1 through 6, previously refined from 84 classes to 37 classes. The combination of the three water classes with the 37 classes brings the total classes to 40. The printout from this step is on pages 257-258.

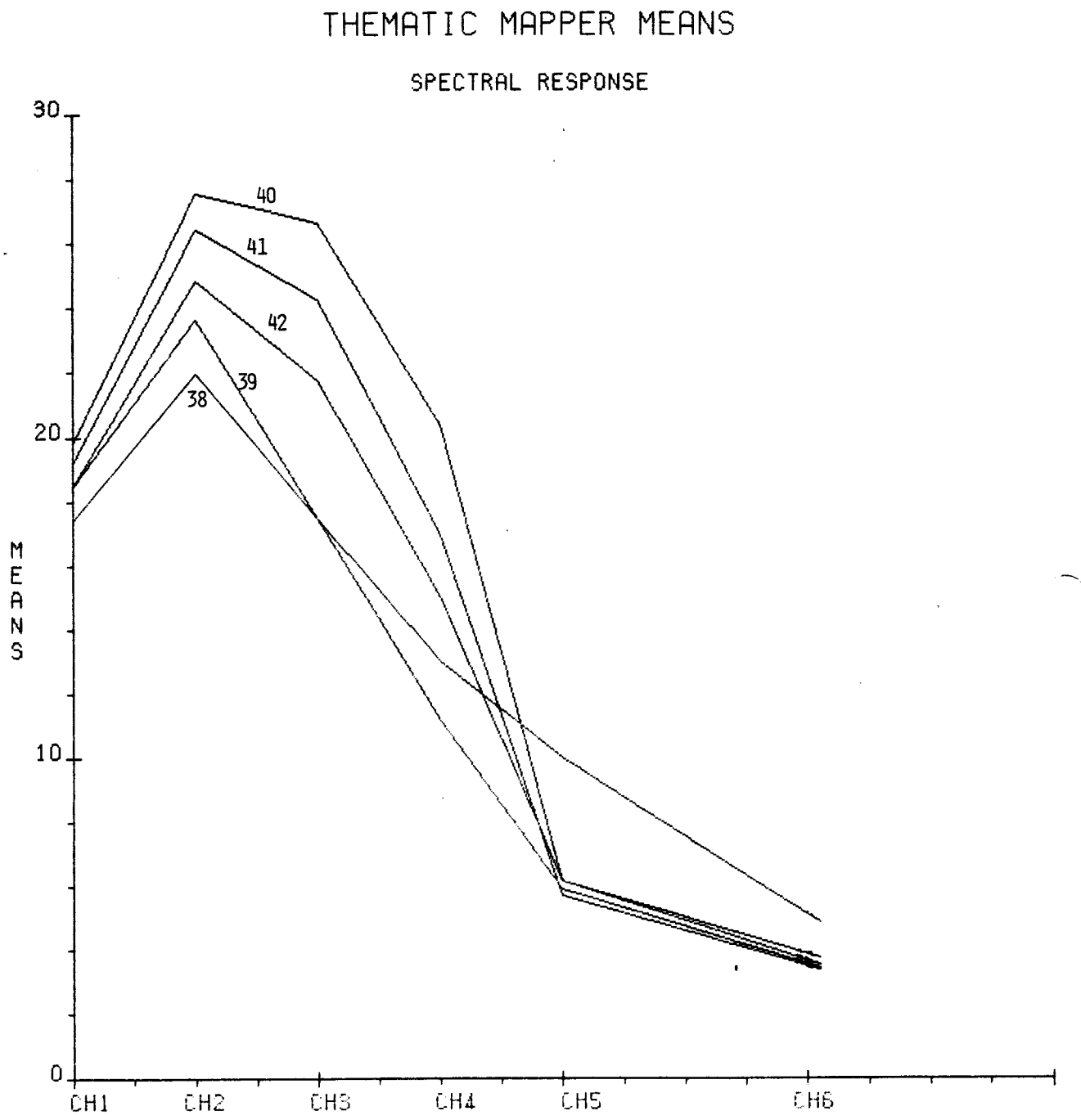

Figure III-24. Final 5 water classes selected.

These 40 classes were then considered the candidate training classes. The analyst knew that running the full classification of the 1000-by-1000 pixel area with seven channels and 40 classes would be costly, and so he ran a test classification on a 300-by-300 area to assess the accuracy that could be expected in the final classification. The area chosen was an area including some of the urban area and some agriculture, west of the center of the city.

The results from this test classification showed that a sludge pond near the downtown area was misclassified since there were no training statistics for the sludge pond. Other classification problems occurred showing that grasses and bare soil were not adequately represented in the training statistics. Although the field grasses were represented, the cultivated grass of golf courses and cemetaries was not. The problem that arose with the bare soil was that the fields in the test area that should have been classified as soil were classified as urban. The analyst recognized that the reason for the misclassification was that the soil was unusually bright, perhaps quite dry. Again the supervised approach enabled him to represent this field type in the statistics file.

For training statistics to be adequate, all the classes that exist in the scene must be represented in the statistics file. When classes are discovered which are not adequately represented, as is the case for the sludge pond, bare soil and grass, the supervised approach can be used to advantage. The analyst picked out some pixels that were known to contain these materials and submitted the line and column coordinates of these pixels to the STATISTICS processor. This processor is designed to determine the statistical properties of groups of single materials whose identity is known. The output from STATISTICS appears on pages 259-261 (for the sludge class) and pages 262-265 for the grass and soil classes.

The STATISTICS processor characterizes the training sample in terms of the mean and variance of the class in each channel. In this respect the statistical output parallels the output from CLUSTER.

Examine the output from the STATISTICS processor, pages 259-265. Note the following:

- Number of pixels in each class
- The class mean in each channel
- The standard deviation in each channel
  (values are the square roots of the variances)
- Correlation matrix
  (matrix showing the correlation of the data values between all pairs of channels)
- Histograms
- Coincident spectral plot.

Even though the statistics for these classes were developed using a supervised approach and the statistics of the 40 classes previously defined were developed through clustering (non-supervised approach), the two statistics files can be merged into a single file.

The final merging of classes, then, takes place in the MERGESTATISTICS shown on pages 266-267, resulting in 43 classes. As a final check on the classes, the analyst ran processors that check the inter-class relationships on the 43 classes: SEPARABILITY to determine the transformed divergence of all possible pairs of classes (pages 268-286) and BIPLOT showing Channel 3 vs. Channel 4 (pages 287-291) and Channel 4 vs. 5 (pages 292-296).

The result of this study was that there were still two classes that are very similar, both soybean classes, Classes 4 and 5 from Area 5. The spectral plots for the two classes are shown on Figure III-25. Since there was no need to distinguish between these two classes, and there was little possibility because of spectral similarity, the analyst had the choice of deleting one of the classes or pooling the two into one class. In this instance, as a demonstration, the analyst chose to pool the classes. In reality, since the analyst has a strong preference for deleting classes over pooling them, he would most likely have deleted one, except for the demonstration aspect of this procedure.

The MERGESTATISTICS output on pages 297-299 shows the final step in the refinement of the training statistics, pooling the two soybean classes and adding names to all 42 classes in the final file of training statistics. The final training statistics file is now ready to be used for classification. The 42 final classes and their identity are listed in Table III-3. Note that every final training class is either one of the original cluster classes, a combination of two cluster classes (when pooling occurred), or a class derived through the supervised approach.

As a representation of one of the final families of curves, Figure III-26 shows the final six soybean classes, derived from Areas 2 and 5.

Pages 300-630 contain the SEPARABILITY information for the final 42 classes when data in all seven channels are used and when data from only six channels, five channels, four channels, three channels, two channels and one channel are used. In this instance SEPARABILITY can be used to select the best channels to use for the classification if fewer than all seven channels were to be used.

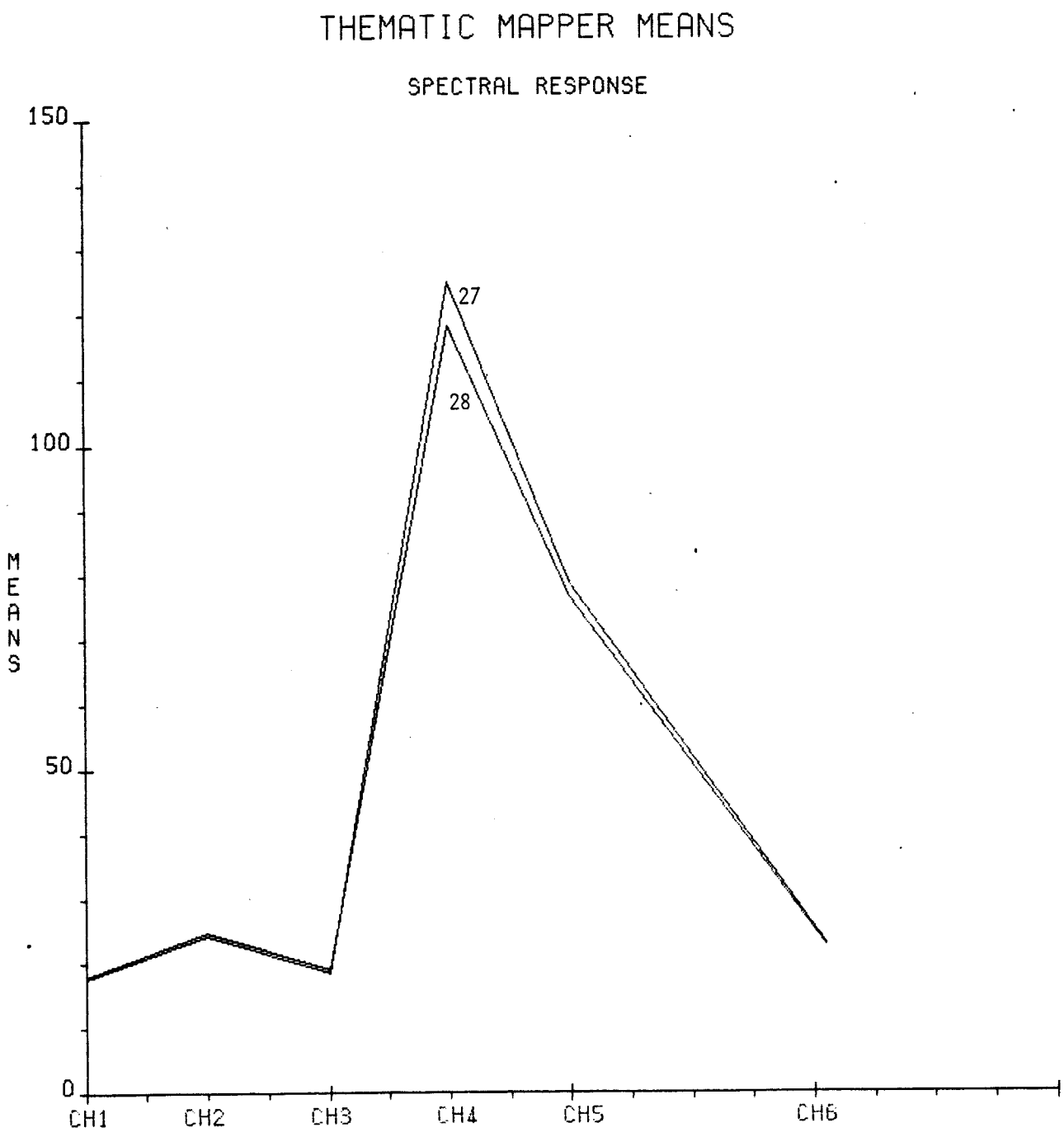

Figure III-25. Plots of calibrated mean values for two soybean classes that are pooled.

Table III-3. List of 42 final spectral training classes.

Pool Name	Cluster Area	Cluster Number	Final Class Number	Pool Name	Cluster Area	Cluster Number	Final Class Number
Forest1	1	7/18	1/42	Substat	2	3/16	22/42
Forest2	2	11/16	2/42	Quarry	2	4/16	23/42
Corn1	5	13/15	3/42	Concrete	3	1/10	24/42
Corn2	5	14/15	4/42	Sludge	-	-	25/42
Soy1	2	1/16	5/42	Indust1	3	4/10	26/42
Soy2	5	1/15	6/42	Indust2	4	1/12	27/42
Soy3	5	2/15	7/42	Urb/Hiwy	4	3/12	28/42
Soy4	5	4/15 & 5/15	8/42	Soil/Hiw	1	12/18	29/42
Soy5	5	6/15	9/42	Reside1	3	9/10	30/42
Soy6	5	7/15	10/42	Reside2	4	7/12	31/42
Wheat/Re	5	15/15	11/42	Beach1	1	13/18	32/42
Grass1	4	8/12	12/42	Beach2	1	14/18	33/42
Grass2	-	-	13/42	Beach3	1	17/18	34/42
Grass3	2	7/16	14/42	SoilWet1	2	14/16	35/42
Soil/Ve1	2	6/16	15/42	SoilWet2	3	10/10	36/42
Soil/Ve2	2	8/16	16/42	Marsh	2	13/16	37/42
Soil/Ve3	-	-	17/42	Water1	1	18/18	38/42
Farm/Gra	5	10/15	18/42	Water2	2	16/16	39/42
Road/Far	5	9/15	19/42	Water3	7	1/ 7	40/42
Barsoil1	6	1/13	20/42	Water4	7	2/ 7	41/42
Barsoil2	6	7/13	21/42	Water5	7	3/ 7	42/42

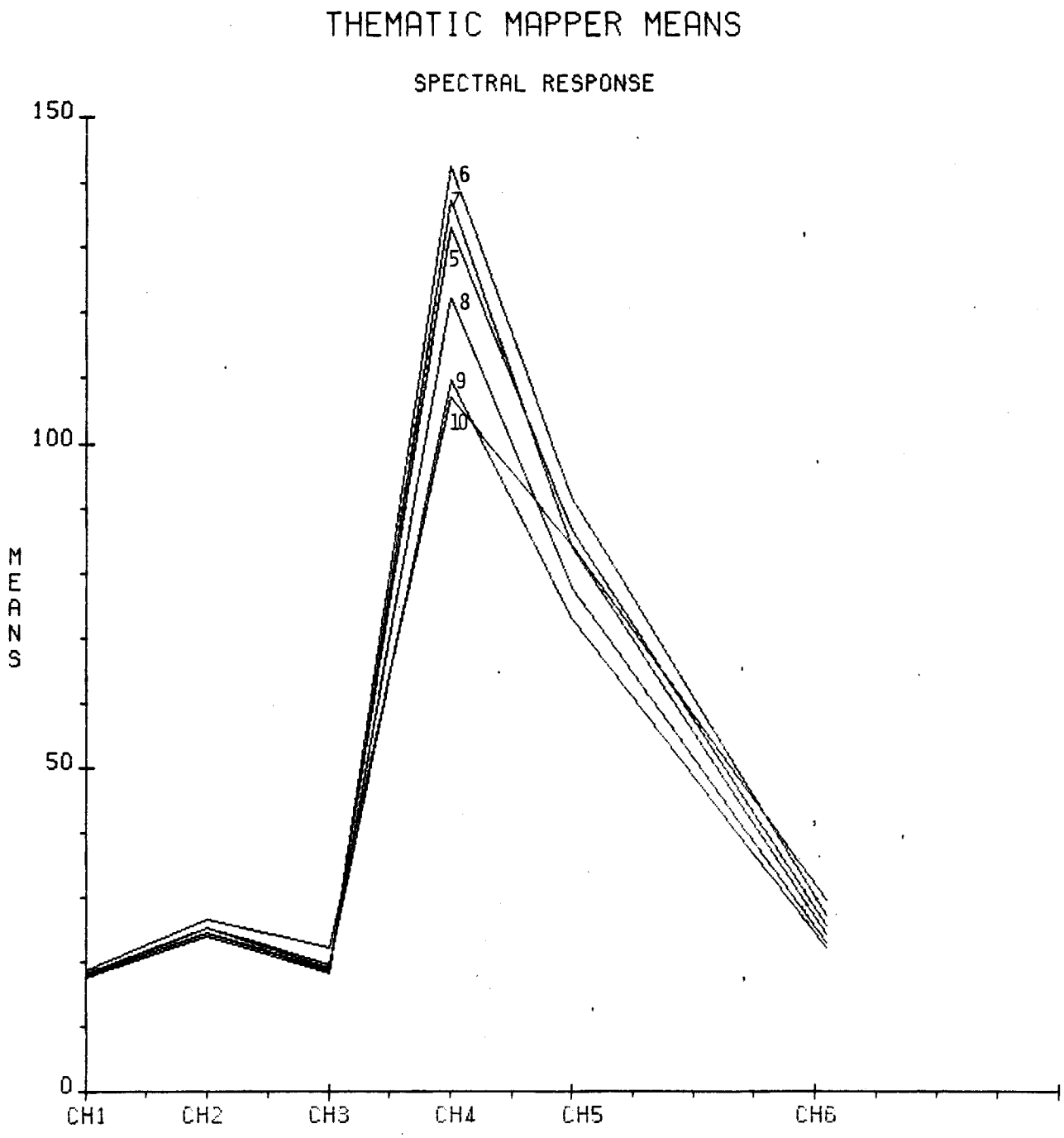

Figure III-26. Spectral plots of final six soybean classes.

Study the SEPARABILITY output starting on page 300 to answer the following questions:

1. If only one channel could be used for the classification, which one would be best? What is the average transformed divergence for one channel data?

2. Which channel is the next best single channel?

3. If three channels are used, which are the best three? What is the average transformed divergence value? (See page 393.)

4. If five channels are used for the classification, which are the best five? What is the average transformed divergence value? (See page 515.)

5. When all seven channels are used, what is the average transformed divergence for this set of classes? What is the minimum value? (See page 614.)

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 Bi-plots of the final 42 classes are shown for Channel 3 vs. 4 and Channel 4 vs. 5 beginning on pages 634 of the printout. Using the information included in the printout, add class names to one of the two bi-plots.

This concludes the refinement process. The 42 classes have been identified and tested. The next step is to classify the data in the entire 1000-by-1000 pixels study area, the subject of Section IV.

 Pages IV-1 to IV-4 are missing

Part IV - CLASSIFICATION OF THE ENTIRE STUDY AREA

The Polk County study area was classified with the LARSYS CLASSIFYPOINTS processor using the 42 training classes developed during the previous section. This processor is based on the maximum likelihood decision rule. The maximum likelihood discriminant functions for each training class are evaluated using the data values of each pixel, and then each pixel is classified into the class associated with the discriminant function yielding the highest value.

The principal output from this process is the classification results which includes the class to which each pixel was assigned (designated by class number)and a rating that indicates the degree of probability that the pixel was correctly classified. The results are stored on the results tape.

The computer output from CLASSIFYPOINTS is on pages 641-643, with page 642 listing the 42 classes by assigned name.

Page 643 verifies that the classification was completed and stored on tape. How many CPU seconds did the classification take? Express that in hours, minutes, and seconds.

The 42-class classification was completed for the 1000 by 1000 area using all seven channels. The large number of CPU hours represents the time to do the classification on an IBM 370/158 and would not be the same for other implementations of this processor. Nevertheless, the length of time required motivates analysts to look for other approaches that might save computer time. One approach would be to use the SEPARABILITY output for selecting the best channels so that the classification could be accomplished with half as many channels.

Refer to the SEPARABILITY output (pages 393-630) and list the average and minimum divergence values for the following combinations of classes.

	Average	Minimum
All 7 classes		
Best 6 classes		
Best 5 classes		
Best 4 classes		
Best 3 classes		

The results of this comparison show that while it is certainly best to use all seven classes, there is not much degredation in class separation when the best six classes are used, and average divergence values hold up fairly well down to four channels. Certainly there is a strong difference in the minimum value between using four and three channels. Experience in classifying four-channel MSS data into 20 classes for a 1000 by 1000 area provides evidence that on the same machine as described above this run would take about one hour of CPU. It would appear that some kind of feature selection procedure would be an appropriate step in the analysis of TM data.

Future classification studies will use the same statistics file to run the ECHO classifier and the LAYERED classifier, both available in LARSYS. Small area tests have shown that ECHO produces very effective classifications for the agricultural areas, but it does not help information extraction from the more heterogeneous urban areas. Use of the LAYERED classifier will reduce computer time because fewer channels are considered in each decision layer, sometimes as few as one or two.

Experimentation is continuing to quantify the extent of improvement in information content of the seven-channel TM data over the four-channel MSS data.

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 647-682 of the printouts. The control statements on page 644 show the map symbol chosen for each of the forty-two final classes; this same information is displayed in a more convenient format on page 645, where class names are associated with the symbols. Since there are so many classes, the analyst frequently chose to display similar classes with a single symbol. An example of this is the five soybean classes; three of them are displayed with a single symbol, 1, and the remaining three with the letter I.

Look over the classification map to identify some ground features that you are familiar with.

Would any of these features be easier to locate if different symbols had been used?

Suppose you were interested only in the locations of the forested land. What symbol set might you have chosen to enhance the display?

Once the classification is complete, as was done during the previous section, it can be displayed in many different ways to emphasize different aspects of the classification. Figure V-5 is an example of a printout created to emphasize the six different classes of soybeans, two classes of corn, and a wheat residue class in a small area.

Compare the printout in Figure V-5 with the same area in the binder of printouts, pages 647-648.

If you were interested in estimating potential soybean production, which printout would you prefer? Why?

Figure V-5. A portion of the classification displaying two classes of corn (I,1), six classes of soybeans (S, 8, 3, B, 5, and &), and a wheat/residue class (A). All other classes are displayed by blanks.

One of the great benefits of TM data over Landsat MSS data is that with the greater spatial resolution of the sensors and their increased spectral range, more sub-classes can be successfully discriminated in the data. Above you saw that six sub-classes of soybeans were differentiated. Five classes of water were also discriminated. Although the improvements in information content have not yet been quantified, it appears that the amount of sub-class detail possible to extract from TM data substantially exceeds that from Landsat MSS data.

The classification results are also displayed in tabular format on pages 700 to 702. The tables list 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. Each data point (pixel) represents .20 acres, and so it is a simple multiplication procedure to derive area estimates from the number of pixels. This class-by-class area estimate is often valuable in resource planning activities.

Pages VI-1 to VI-4 are missing

Part VI - EVALUATION OF THE CLASSIFICATION RESULTS

In the previous section, we studied the classification map of the study area. Now, we need to test the classification to evaluate its accuracy. This will be done by selecting test fields for each major cover material in the study area (forest, corn, soybean, soil/vegetation, water and urban) and checking to see whether the pixels in these test areas were successfully classified.

The fields that were used as test fields are outlined in the classification printout with the symbol "+". For example, there is a large water test area at the bottom of page 650 and several agricultural ones on pages 671 and 672. The entire list of test fields saved is given on page 683. You will notice that in choosing the test fields, the analyst selected different sub-classes of the major classes to avoid a bias in the test results. For example, test samples from both the reservoir and the pond below the reservoir were used.

Choose one of the six major classes that interests you the most: forest, corn, soybean, soil/vegetation, water, and urban. Use the list on page 683 to help you locate the test fields on the printout and visually examine the results. Were most of the pixels within the test areas classified correctly?

Turn to the tabular test results starting on page 685 and locate the correct page for the group of materials you choose. Compare the total number of test points with the number that were classified into the correct class (shown on the top line) and with the percent of points that were correctly classified.

What other class or classes were most often confused with the class of interest?

Note that test area estimates in acres and hectares are given as a further aid in evaluating the accuracy of the classification. All of the test field information is tabulated together on pages 698-699. By looking across the matrix, it is easy to see where the greatest confusion exists. Forest has the lowest percent correct, with some of the forest pixels classified into corn, grass, soil/vegetation, and soil. At this point it is impossible to know if these errors are the result of misclassification or if the test field selection could have been improved. Good ground reference data is essential for careful analysis of the classification accuracy.

All in all, the classification accuracy assessed in this way is 96.1%, an accuracy level that significantly exceeds the requirements of the analysis. Even the forest classification accuracy surpassed the goal of 85% correct classification.

The evaluation demonstrated here is a preliminary evaluation of a portion of the study area. When more ground reference data is available, the classification will be tested using a stratified random sample of test fields to avoid biases that may have been a factor in the initial test class selection.