LARS Publication 062579

Workshop Series on Numerical Analysis of Remotely Sensed Data

Ronald K. Boyd
John C. Lindenlaub

Laboratory for Applications of Remote Sensing
Purdue University West Lafayette, Indiana 47906 USA

LARS Publication 062579

Workshop Series
on Numerical Analysis
# of Remotely Sensed Data

by

Ronald K. Boyd
and
John C. Lindenlaub

Workshop Series
on Numerical Analysis
# of Remotely Sensed Data

by

Ronald K. Boyd and John C. Lindenlaub

PREREQUISITES

The intended audience for this workshop are persons who have a basic background in remote sensing. The remote sensing background can be gained by means of the following educational materials or their equivalent:

LARSYS Educational Package*

Unit I: An Introduction to Quantitative Remote Sensing

Fundamentals of Remote Sensing Minicourse Series**

Remote Sensing: What Is It?

The Physical Basis of Remote Sensing

Spectral Reflectance Characteristics of Vegetation

Spectral Reflectance Characteristics of Earth Surface Features

The principles and techniques described in this workshop apply to numerical analysis procedures in general. LARSYS is an example of a numerical analysis system and is used as the software system for data analysis in this workshop.

*The LARSYS Educational Package may be obtained from the Support Services Manager, Laboratory for Applications of Remote Sensing, Purdue University, 1220 Potter Drive, West Lafayette, Indiana 47906.

**The Minicourse Series may be obtained from G. W. O'Brien, Continuing Education Administration, 116 Stewart Center, Purdue University, West Lafayette, Indiana 47907.

PREFACE

People in need of information about the earth's surface now have more types of data and techniques for extraction of information from such data from which to choose than ever before. A source of data introduced in recent years called a multispectral scanner makes precise measurements of the energy being refelected and emitted from the earth's surface. Various techniques have been devised to extract information from such data. In this workshop statistical pattern recognition techniques will be studied by simulating an analysis of a common form of data, Landsat multispectral scanner data.

The purpose of this workshop is NOT to train you to be an analyst but instead to give you an overview and understanding of how multispectral scanner data are analyzed. During the workshop, you will be asked to make the same types of decisions an analyst must. You will not interact directly with a computer, but you will be working with computer output products.

STUDENT-INSTRUCTOR INTERACTION

While this workshop series attempts to summarize the experiences of a great many multispectral data analysts, there is no real substitute for talking to someone who is already familiar with the numerical analysis of Landsat data. This workshop series presumes you will study under the guidance of a tutor with analysis experience. You are encouraged to discuss your progress periodically with your tutor.

MATERIALS

A set of printouts entitled "One Mans Analysis of Run 73033802" is required to complete this workshop. In addition your tutor will have other handouts required at various points in the workshop. The textbook "Remote

Sensing - The Quantitative Approach", edited by Philip H. Swain and Shirley M. Davis and published by McGraw-Hill International Book Company is referred to in several activities as the source of technical details. Although the text is not required for completion of this workshop, its use can considerably broaden and deepen understanding of the material presented.

OBJECTIVES

Since the method of obtaining information about the earth's surface described in this workshop is only one of many, an overall study objective is to understand the strengths and weaknesses of the numerical approach and how it compares with conventional or alternative methods.

The numerical analysis of a set of multispectral scanner data can be broken down into a sequence of steps. By the time you have finished this workshop, you should be able to list the steps of the analysis sequence in the proper order. Furthermore, for each step in the analysis you should be able to do the following:

1) give a brief explanation of the significance of the analysis step with respect to the whole analysis sequence,

2) name and briefly describe any software tools available to carry out the analysis step,

3) describe the output or products from each step in the analysis sequence, and

4) describe the tasks listed at the beginning of each activity.

ACKNOWLEDGEMENTS

The workshop is based upon materials presented in A Case Study for Analysis of LANDSAT Data by Tina K. Carey and John C. Lindenlaub. Some of the early conversion was done by James D. Russell. Contributions to the analysis example were made by Luis A. Bartolucci and Michael D. Fleming. The authors wish to express their sincere thanks to these present and former

LARS staff members.

FORMAT

Each activity in the workshop follows this format: The instructional objectives are stated, followed by a discussion of the purpose, philosophy, and analysis techniques associated with that step in the data analysis sequence; samples of the computer output at each step are provided along with an interpretation of the results; and exercises are provided to test your mastery of the section's instructional objectives.

The material is presented in this format so that you can proceed through this workshop at your own rate and learn what the analysis steps are, the importance of each step, the inputs to each step, and the outputs expected from each step.

Time Required

Approximately 8 hours.

INTRODUCTION

The numerical analysis of remote sensing data is a dynamic process which requires interaction between man (analyst) and machine (computer). The process involves meshing the experience and insights of the analyst with appropriate computer programs to extract the maximum amount of information from the data. Numerical analysis techniques have been shown to be cost effective in many cases and allow detailed study of digital data. A typical analysis sequence is shown in Figure 1. Even though it is shown here as basically a linear process, all of the steps are interrelated. At any step in the analysis, interpretation of the results of that step can lead the analyst to conclude that he should go back to a previous step and revise his procedure.

When an analysis problem in remote sensing is conceived, certain steps are followed by the analyst. The first step is to state the analysis objectives. To do this, the analyst must determine the geographic area of interest, the general cover types present and the nature of the application to which the results will be applied. An additional component which is often included in the analysis objective is a statement of the desired classification accuracy.

Remotely sensed data can be acquired via aircraft or satellite. The platform used and type of data collected will be determined in part by the analysis objectives. The usefulness of these data may be affected by atmospheric conditions (haze, cloud cover) and system problems such as striping. Since each Landsat satellite covers the entire earth every eighteen days, the analyst can generally choose the time of the year most suitable for mapping the cover types of interest.

Aircraft data may be affected by sun angle and or view angle effects, and by system noise. There is no ongoing broad based program of aircraft data collection. Therefore it is the responsibility of anyone desiring aircraft data to arrange for its collection.

The goal of most analyses is to make a classification of the multispectral scanner data, which is done by comparing the spectral data values of the points to be classified to the spectral characteristics of known points. All of the steps leading up to making the classification are geared toward locating and identifying known points and assuring that the groups of known points called training classes are representative and sufficiently different to prevent confusion among them.

To begin the analysis the analyst locates the area of interest in the multispectral scanner data by comparing patterns in it to those in available

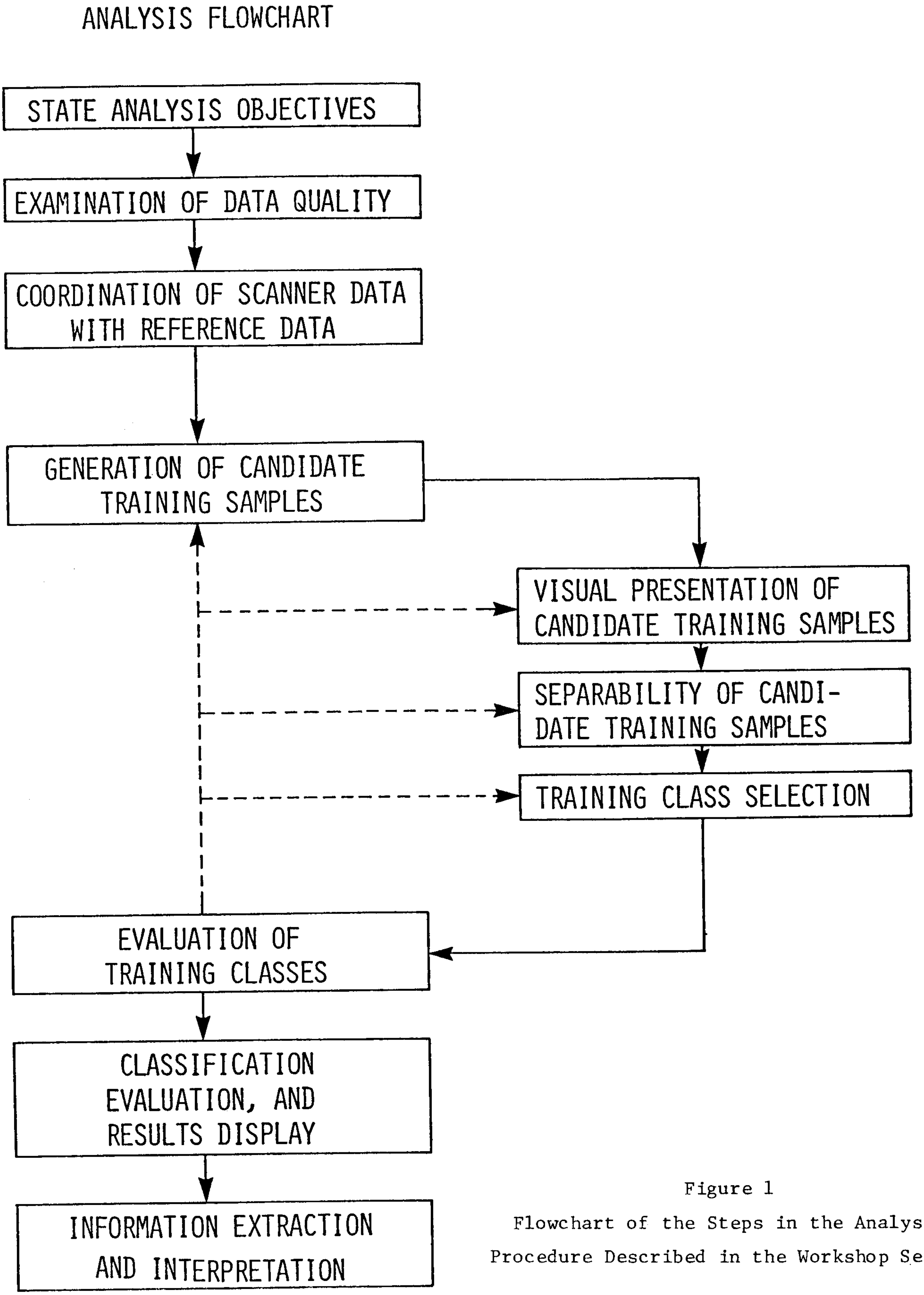

ANALYSIS FLOWCHART
STATE ANALYSIS OBJECTIVES
EXAMINATION OF DATA QUALITY
COORDINATION OF SCANNER DATA WITH REFERENCE DATA
GENERATION OF CANDIDATE TRAINING SAMPLES
VISUAL PRESENTATION OF CANDIDATE TRAINING SAMPLES
SEPARABILITY OF CANDIDATE TRAINING SAMPLES
TRAINING CLASS SELECTION
EVALUATION OF TRAINING CLASSES
CLASSIFICATION EVALUATION, AND RESULTS DISPLAY
INFORMATION EXTRACTION AND INTERPRETATION

Figure 1

Flowchart of the Steps in the Analysis Procedure Described in the Workshop Series reference data, such as USGS topographic maps, aerial photographs, and actual ground observations. This step is referred to as coordination of scanner data with reference data and is usually carried a step further to include general location of the cover types of interest. Then the analyst selects training areas, groups of points that separately contain several of the cover types of interest and collectively represent all the classes present in the area to be classified. Although guidelines exist for making training area selection, experience, familiarity with the scene, and the amount and quality of reference data strongly influence the representativeness of the areas selected.

Next each training area selected is supplied to a clustering processor which uses the spectral response values of each point in the training area to define groups of spectrally similar points. These groups of points are then identified as to what they represent on the ground. Ultimately, a subset of the groups will be used to train the classifier to recognize unknown points as being specific cover types. In some cases it may be necessary or desirable to generate additional candidate training classes using other methods.

As stated above, recognition is done on the basis of the spectral characteristics of the known classes, which are calculated within the clustering processor. Before making the classification, the statistics are used by the separability processing function to determine pairs of classes which would be confused with each other in the classification if left separate. In order to circumvent such confusion, a separability diagram is constructed and used to decide which classes to eliminate or combine with others. Statistics and separability are then recalculated for the newly selected training classes. If problems still exist, the classes are re-evaluated.

In some cases it is necessary to return to previous steps.

Once training classes have been selected and verified, the classification is made. Each point to be classified is classified into the class to which it has the highest probability of belonging. This probability is calculated on the basis of the spectral response values of each unknown point and the spectral characteristics of the training classes, which are described by the statistics calculated in previous steps.

Typically analysis problems include an evaluation of the accuracy of the classification. This is usually done by examining the class assignment of pixels comprising fields of known identity. Results at this stage may point out some previously unnoticed confusion or lack of representation that could lead to repetition of many of the previous steps. When the accuracy is acceptable, the results can be examined and interpreted to obtain the desired information.

As indicated earlier, numerical analysis of multispectral scanner data is a dynamic process with each step providing feedback as to the appropriateness of previous steps. In reality, the analyst has all steps in mind before he actually begins an analysis. He may also return to previous steps and modify his procedure as the analysis continues.

Now that you have looked at an overview of the entire process, let's go back and look at each step in more detail. You will want to refer frequently to Figure 1 to keep in mind exactly where you are during the discussion of the numerical analysis process.

BE SURE TO CONTACT YOUR TUTOR IF YOU HAVE ANY QUESTIONS

ACTIVITY 1 STATING ANALYSIS OBJECTIVES

Upon completion of this activity, you should be able to:

1. Name the four components of an analysis objective.
2. Write an analysis objective incorporating these four components.

The first and one of the most important steps in the numerical analysis process is stating the analysis objectives. What is the problem to be solved? What information do you need to solve the problem? What are you going to do with the results of the analysis?

For example, the purpose of the analysis could be to assist in choosing recreational sites in an area along the coast of Texas. Information on present land use is needed to make such a decision. In this case, the analysis objective might be:

> "Determine the land use in the Texas Coastal Zone with 80% accuracy in order to select potential sites for the location of new recreational areas."

Or analyst may desire an estimate of timber production. If so, his objective might be:

> "Estimate wood production with 85% accuracy in the Mark Twain National Forest using the following types of maps: cover type maps which inventory the present stand, slope-aspect maps which improve the accuracy of production estimates since slope and aspect strongly affect production, and density class maps which aid in estimating percent stocking."

The four essential components of an analysis objective are:

Location

What portion of the earth's surface is of interest? It may be a relatively small area (several hundred hectares) or a relatively large area (thousands or millions of hectares).

Cover Types

What types of ground cover are of interest to you? Are you interested in woodlands, agriculture, rangeland, pasture, barren land and marshes? Or are you interested in deciduous versus coniferous, soybeans vs wheat, etc.?

Applications

How will the analysis output be used? To map drainage patterns? To locate potential port or harbor locations? To inventory a specific area and prepare a cover type map?

Classification Performance

How accurate must the classification be in order to be of help to you? Would a classification performance of 65% be acceptable or do you need to have approximately 90% accuracy? The level of accuracy that can be obtained depends upon many factors: the level of detail desired, the time of the year data were collected, the analyst's training and skill, the particular region being mapped, and other variables. An indication of the accuracy required by the user is very useful to the analyst as it provides a point of reference to which he can compare and assess his analysis.

WRITE an analysis objective that might be useful in an area of interest to you.

The analysis objective that you will be using in this workshop is: "Produce a land cover of map of Monroe County, Indiana (see Figure 2, reduction of county map). The cover types to be distinguished in the map are urban, forest, agriculture, and water. The accuracy of the classification should be at least 85%. The results of the classification will be used to prepare study of the feasibility of locating a wildlife preserve in the area."

Self-Check*

Name the four components of an analysis objective

*After you have completed the self-check see your instructor.

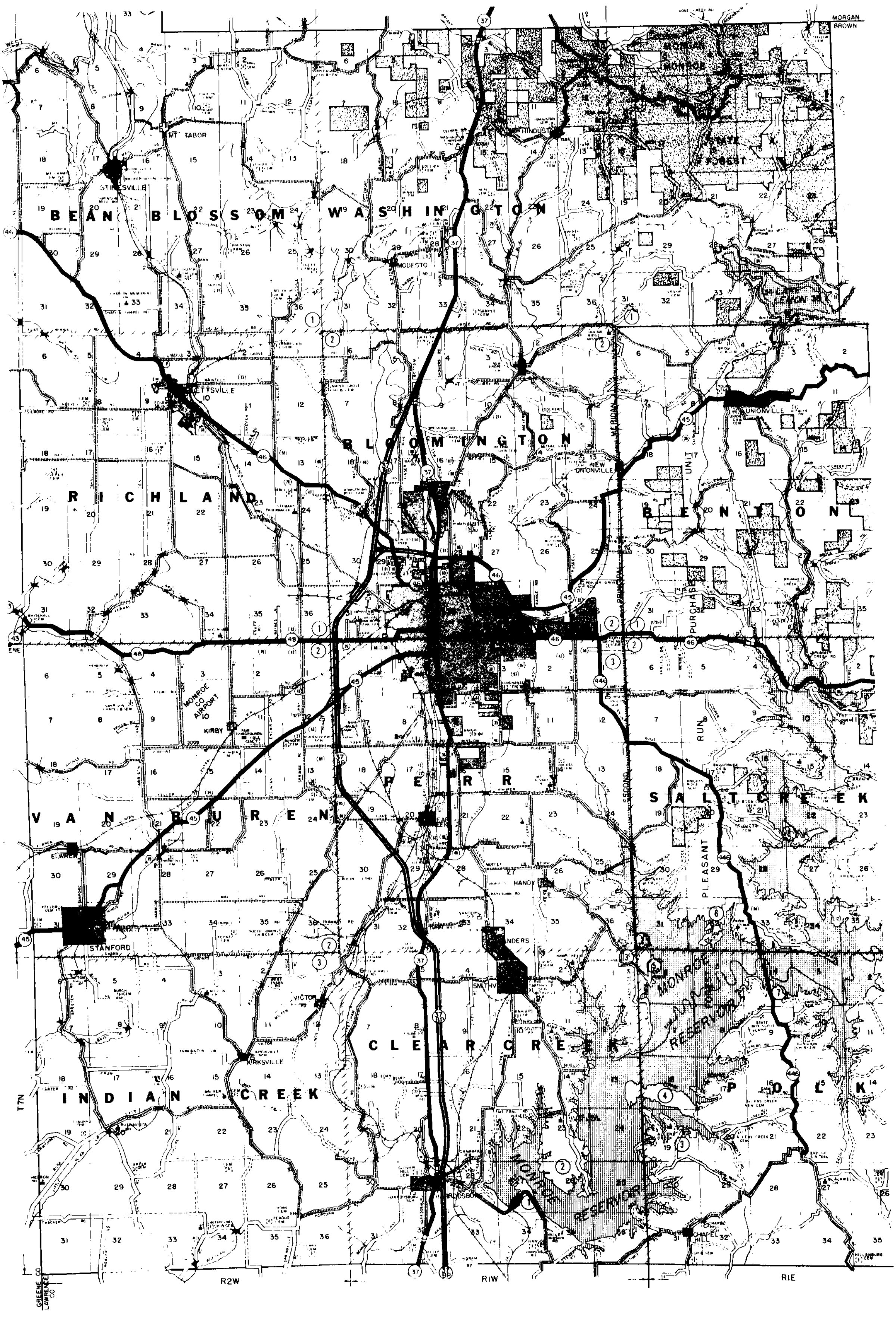


Figure 2. Monroe County Indiana.

ACTIVITY 2 EXAMINATION OF DATA QUALITY

Upon completion of this activity, you should be able to:

1. State at least one reason why the quality of the data being considered for analysis must be evaluated.
2. Name at least two sources of data quality information.
3. Name at least three data idiosyncrasies which might hinder analysis.
4. Describe two types of geometric corrections that might aid the analysis of Landsat data.

After the analysis objectives are stated, a data set must be selected. Landsat satellites with eighteen day repetitive coverage provide a wealth of data over any area. Computer listings of data available for any specified geographic area within any specified cloud coverage and quality limits may be requested from EROS Data Center, Sioux Falls, South Dakota. From this data, an analyst will choose the data of acceptable quality taken at the time of year most suitable for the cover types of interest. Although time of year is the most important factor in choosing a data set, the presence of clouds or spectral bands exhibiting marginal quality cannot be ignored.

A preliminary evaluation of data can be made by inspecting photograph-like imagery created from the digital data. This kind of imagery can be obtained from the data distribution centers, such as EROS Data Center, which also supply the digital data tapes. Gross data characteristics, including cloud cover and snow cover, will be apparent in these products. Figure 3 shows an example of this kind of imagery, with cloud cover on the right side of the scene.

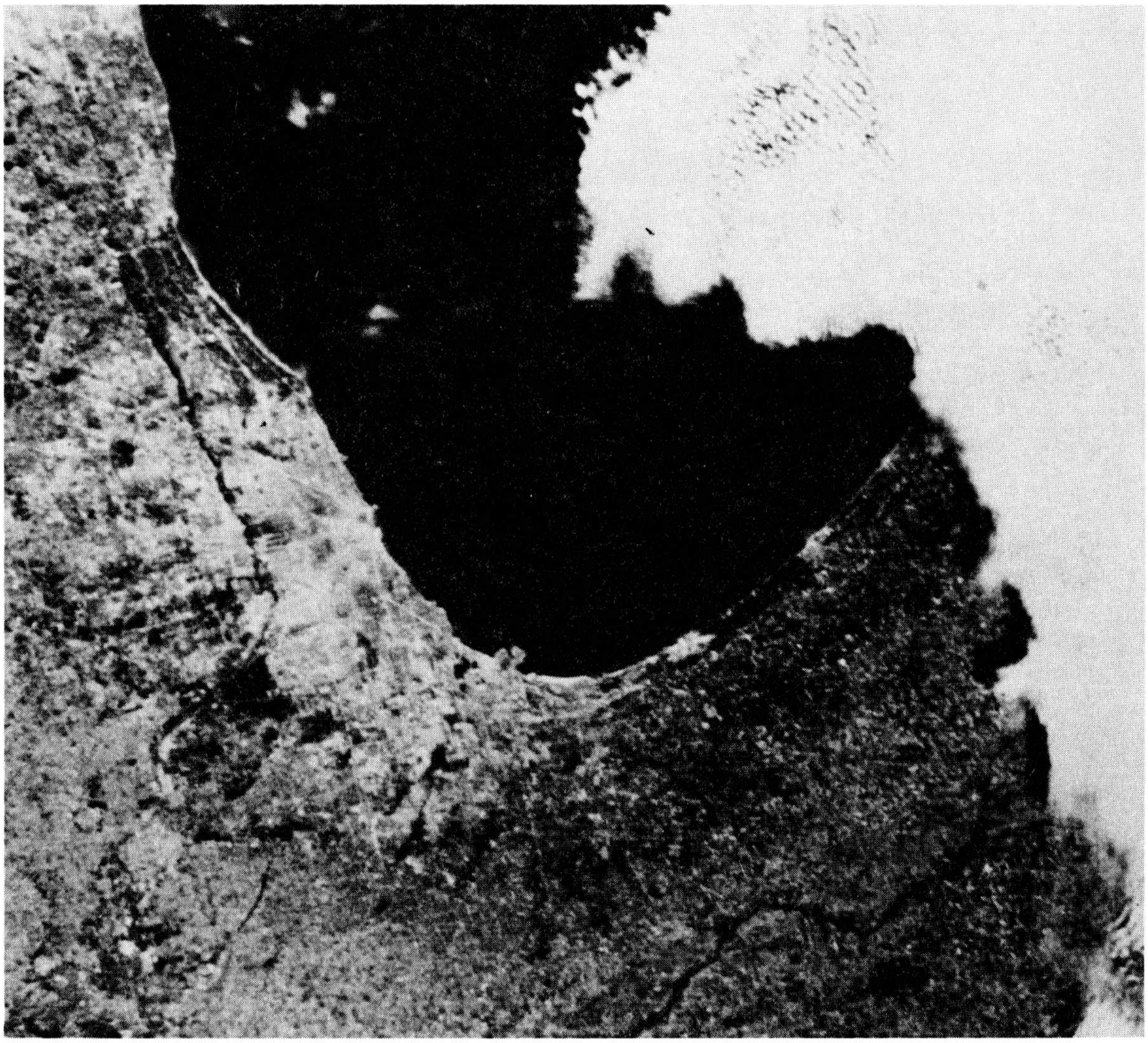


Figure 3. Scene number 1070-16041 over Chicago and the surrounding area has clouds obscuring the east side of the scene.

The data shown in Figure 4, over the mountains of Colorado, has quite a bit of snow cover. The presence of snow can be a limitation in data analysis if it obscures the cover types of interest, such as vegetation or soils. But if the purpose of the analysis were to determine the real extent of snow cover, the presence of snow is essential. Normally presence of both clouds and snow in the same data is undesirable since they are spectrally similar in Landsat bands. Of course, if the purpose of the analysis was to compare the responses of clouds and snow, even this data set would be appropriate. This example points up the necessity of clearly formulating analysis objectives and keeping the objectives in mind.

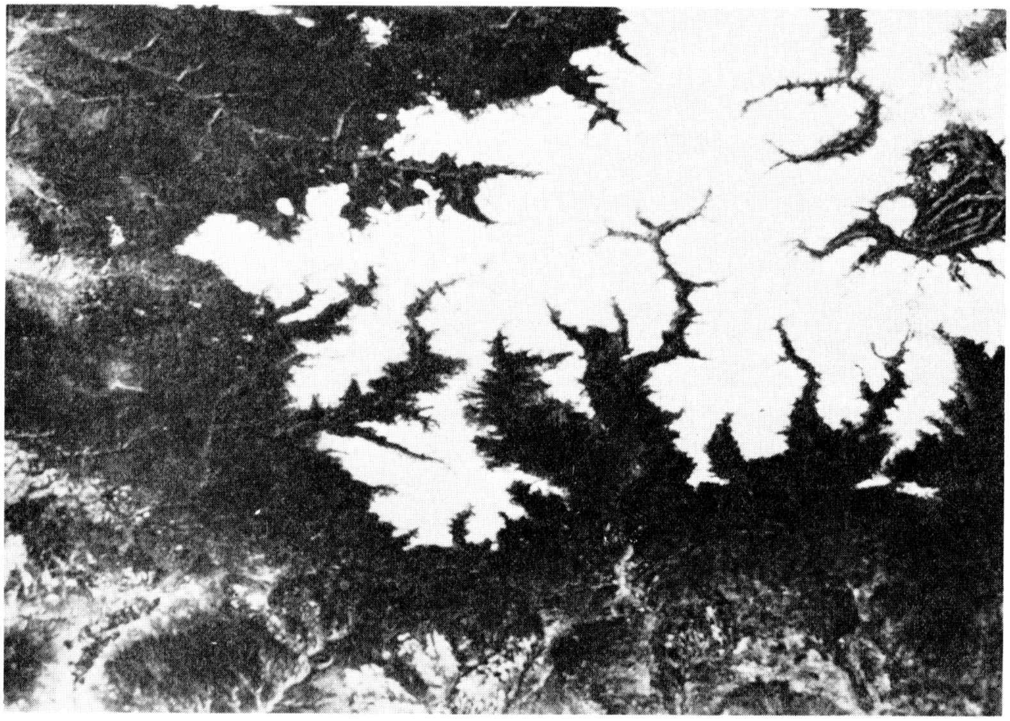


Figure 4. Snow covers the higher elevations in the moutains of Colorado.

Another idiosyncrasy which can occur in the Landsat data appears as stripes in the image. In the Landsat scanner system, six scan lines are collected in each wavelength band each time the mirror oscillates. A separate set of detectors is used for each of these scan lines. If these detectors and their associated electronics are not properly matched or calibrated, a striping effect may be noticeable in the imagery of one or more bands. A dramatic example can be seen in Figure 5.

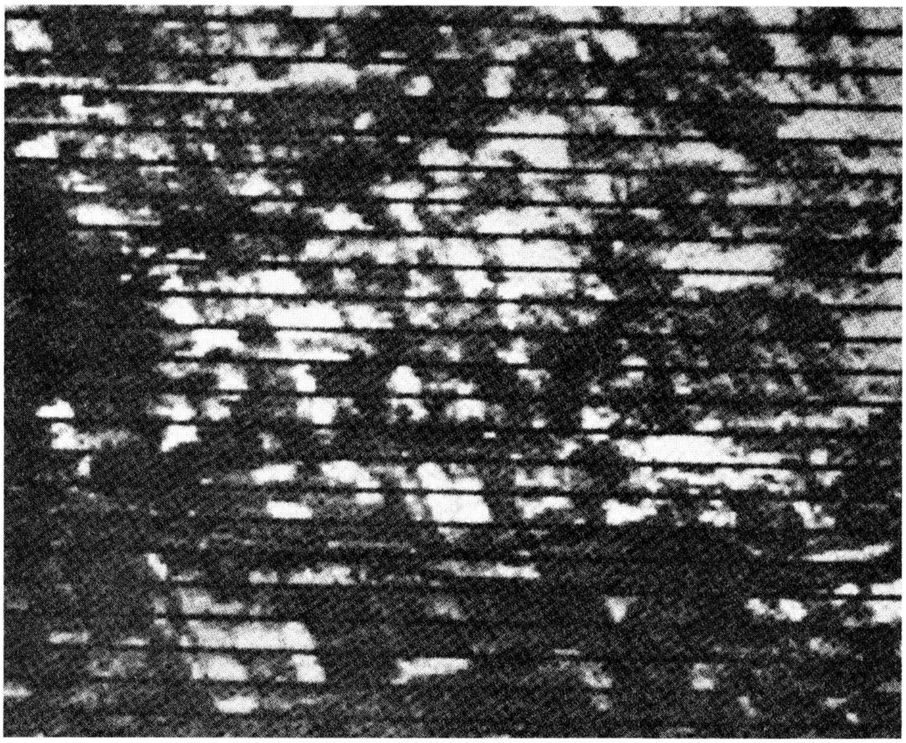


Figure 5. Striping effect in imagery.

The table below shows the mean and standard deviation of the output of each of the six channel 1 detectors over the whole frame:

Detector	Mean	Standard Deviation
1	21.9	3.21
2	21.8	3.07
3	7.0	1.52
4	21.5	3.13
5	20.9	3.11
6	21.9	3.03

Notice that the mean value for detector 3 is very low compared to that of the other detectors. Apparently a malfunction occurred in the detector electronics, resulting in the striping illustrated in Figure 5.

Figures 4 and 5 were produced by photographing an image of the data displayed on a television screen. Alternatively striping in the data could be identified by displaying an image of each wavelength band of the data on a line printer or printer-plotter. Other indications of data quality may be obtained from catalogues of available Landsat data published by NASA and in the computer search listings obtained from EROS Data Center. In the NASA catalogues the quality of each band is indicated as poor, fair, good, or excellent. In the EROS Search Listings quality ranges from 0-9 with 9 indicating highest quality. Estimates of percent cloud coverage are also shown in the catalogues and listings.

The data in Figure 6a show a phenomenon which occurs in Landsat data. Rectangular objects on the ground appear as skewed parallelograms in the original image, the top edge of the image is shifted to the right with respect to the bottom edge by approximately 5% of the height of the image. In addition the Landsat orbit is not oriented exactly over the north pole. This results in a rotation of the imagery which varies with latitude (this rotation is about 12 degrees at 40 degrees north latitude).

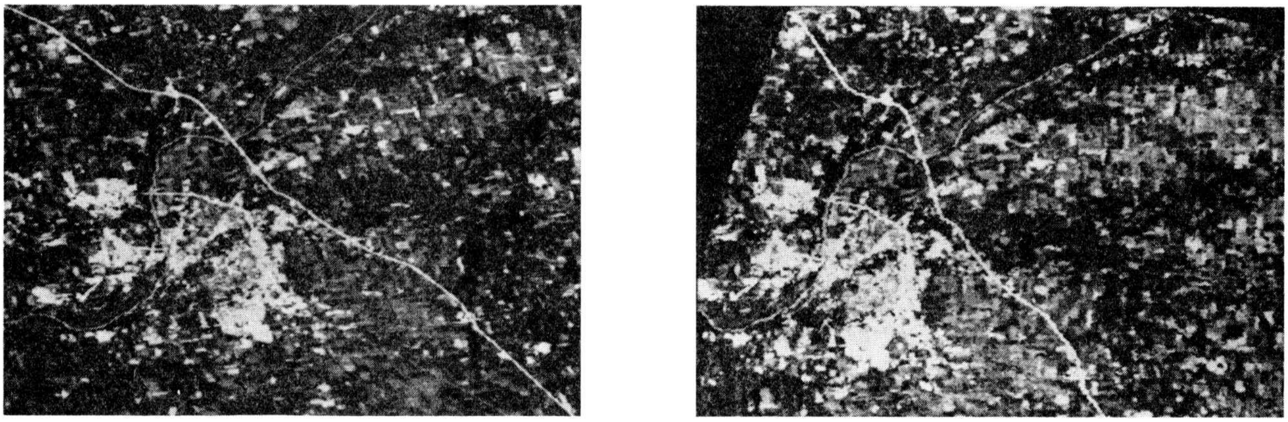


Figure 6. Landsat data before and after processing to remove effects of the earth's rotation.

Pre-processing can facilitate the analyst's ability to interact with the data and can produce output at the desired scale. Figure 6b shows geometrically corrected data, processed to remove skewing and rotation. The vertical columns are now aligned in a north-south orientation. If you imagine yourself trying to locate a corresponding point between a Landsat image and an aerial photograph or map, you can see that this task would be much easier if the two images are oriented in the same way. During geometric correction, the data may also be rescaled. The scale of the Landsat data can be changed to correspond to USGS topographic maps (1:24,000) or any other scale desired to allow the Landsat data to be overlayed on the reference data. In the analysis discussed in this simulation, the data was rescaled to match 7 1/2 minute topographic maps of the area.

While the "uncorrected" data is adequate for some analysis tasks, "corrected" data simplifies the analyst's job of locating ground features since it can more easily be compared with reference data (maps, aerial photography).

Before you can use the data to obtain more information about the Monroe Reservoir area, you will need to become more familiar with the data. A printout showing identification information (IDPRINT) for the data over the Monroe Reservoir area is shown on page 2 of your computer printouts.

STUDY the IDPRINT (see pages 1 and 2 of computer printouts) carefully and note:

a) Date and Time on which data was taken. (Don't confuse date data taken with reformating date.)

b) The Spectral Bands for each of the four channels of data. Which portion of the spectrum do these bands fall in?

c) Wavelength bands 4,5,6,7 (as identified in EROS Data Center) are now named channels 1,2,3,4.

d) The calibration pulse values. Historically, aircraft scanner systems recorded three calibration signals for each channel to be used to remove inconsistencies in the scanner electronics during data collection and to convert the data into radiometric units. In the case of Landsat only one calibration source is used. However the values shown under C1 can still be used to convert the data into radiometric units, facilitating band-to-band comparisons.

For more information about the Landsat Scanner system see section 2.8 of Swain and Davis.

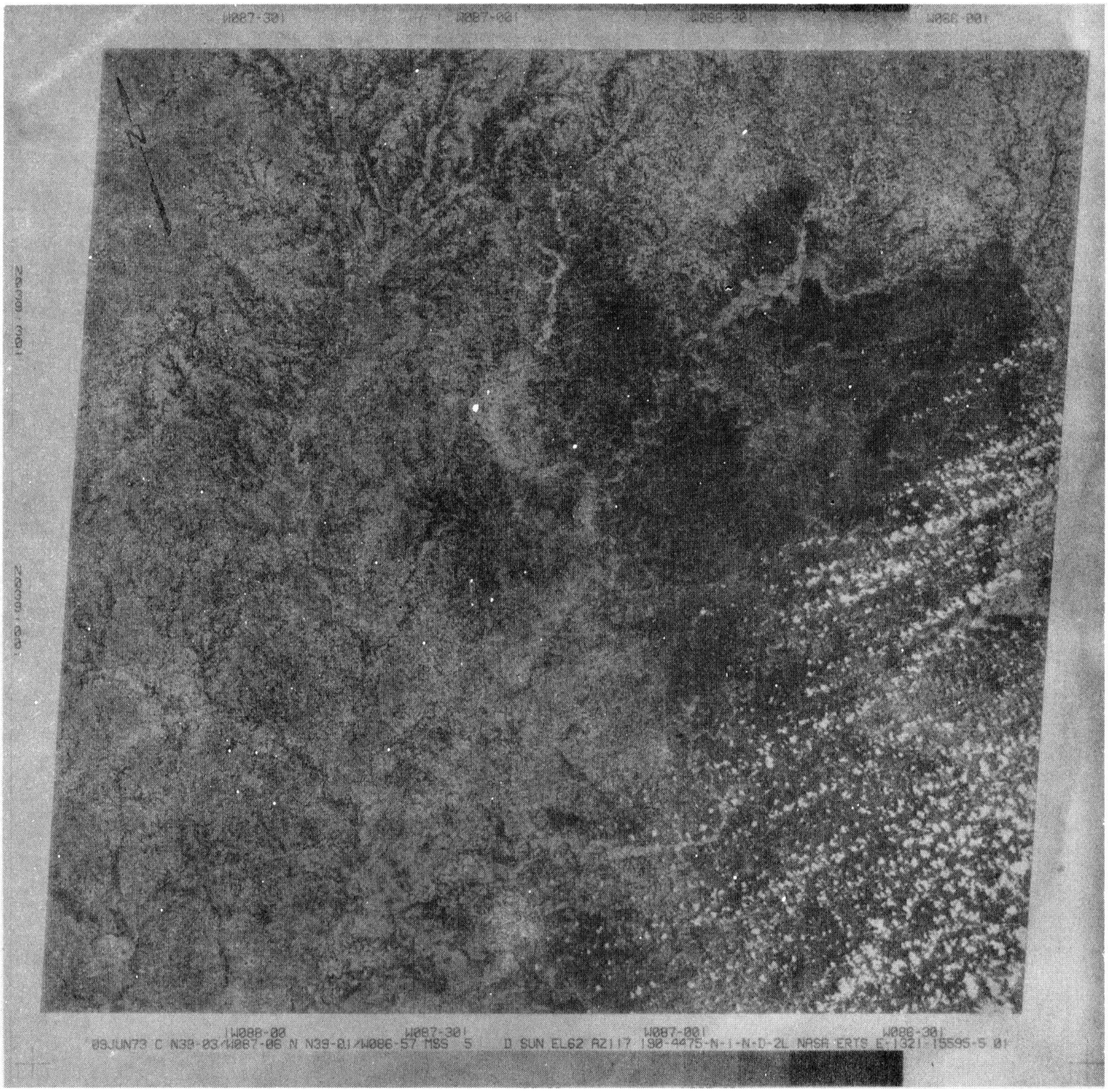
Pages 4, 5 and 6 of the computer printouts show the numerical values associated with three specific data vectors (pixels) in the scene. Notice that each pixel is identified by its line and column address and has four data values or "numbers" associated with it. The numbers indicate the amount of energy 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 range from 0 to 127. 0 is "black," signifying that no energy is being returned, and 127 is "white," indicating saturation of the scanner detector. Data values for channel 4 range from 0 to 63. These three data vectors are representative of the many hundreds 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 values associated with a known pixel of forest are shown below.

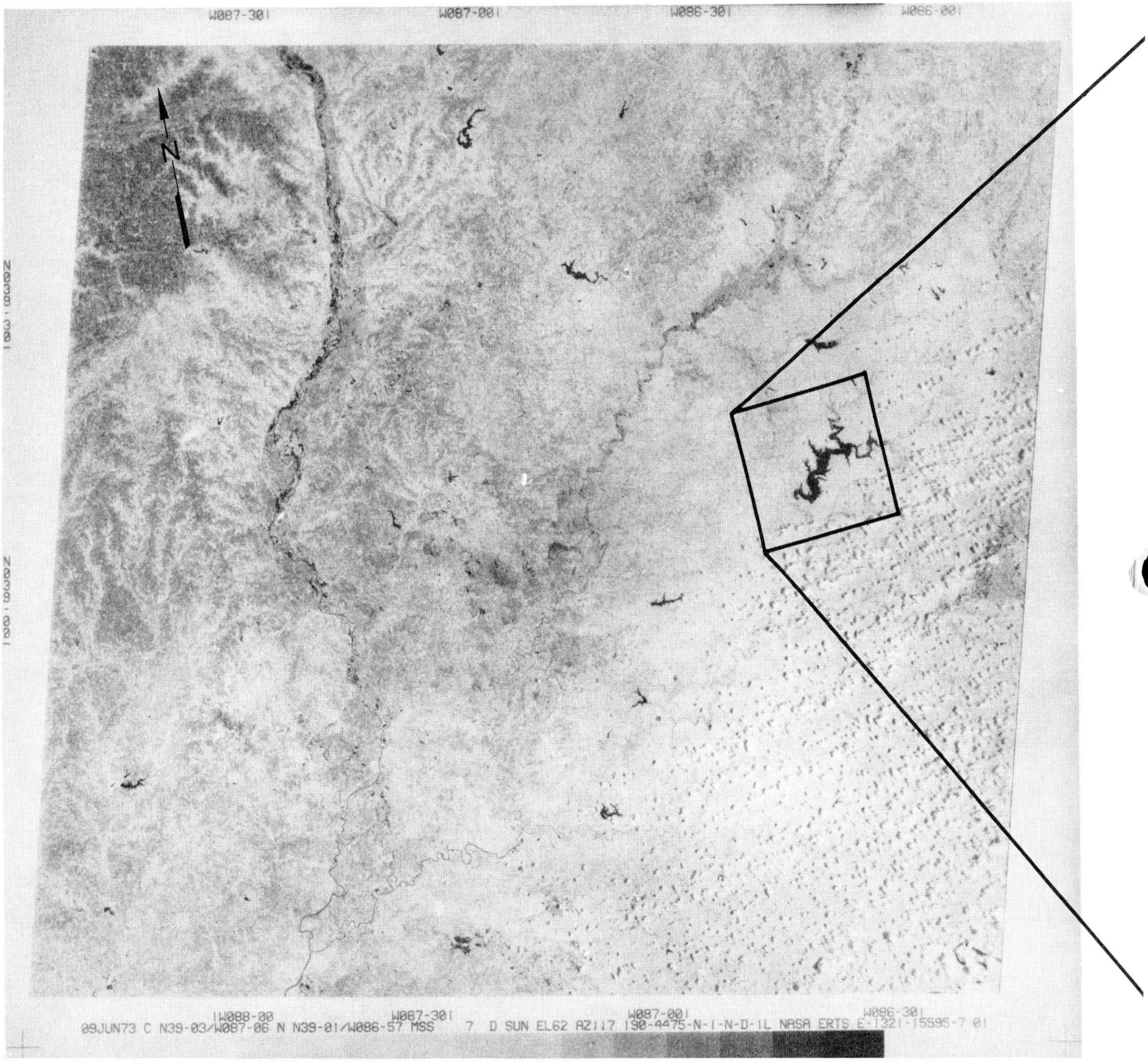
LINE	COL	1	2	3	4
115	326	26.0	15.0	58.0	36.0

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 de-

Pages 21, 22, 23 are fold out
3 images
No Page 24

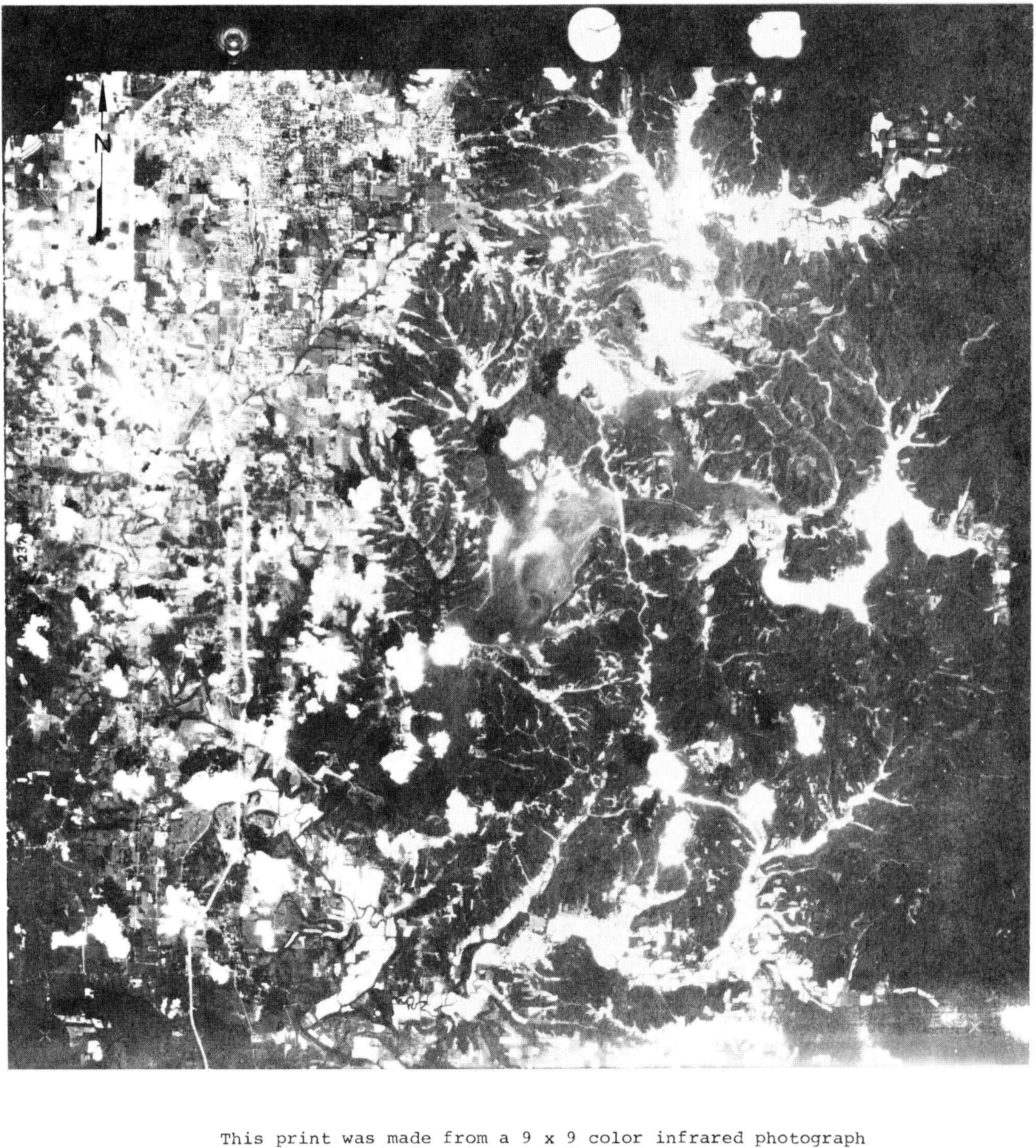

W087-30
W087-00
W086-30
W086-00
N
N39-30
N39-00
W088-00
W087-30
W087-00
W086-30
09JUN73 C N39-03/W087-06 N N39-01/W086-57 MSS 5 D SUN EL62 AZ117 190-4475-N-1-N-D-2L NASA ERTS E-1321-15595-5 01

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.



W087-30
W087-00
W086-30
W086-00
N
N039-00
N038-30
W088-00
W087-30
W087-00
W086-30
09JUN73 C N39-03/W087-06 N N39-01/W086-57 MSS 7 D SUN EL62 AZ117 190-4475-N-1-N-D-1L NASA ERTS E-1321-15595-7 01

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.



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.

the run to display the analyst used the aerial photography to determine that the area of interest was located between lines 30 and 400 and columns 112 and 474. These gray scale images were produced on an electrostatic printer-plotter using dot patterns of varying darkness to represent the gray level each pixel was assigned to. The printer-plotter creates an image which is picture-like in nature, enhancing the spatial features of the data. This output was also chosen for use in the workshop because its size and scale are similar to those of the aerial photography you will be using. The analyst can also request gray maps to be printed on a line printer using symbols of varying darkness to represent the gray levels. This creates a large map in which each pixel can easily be seen, a situation which is useful in many cases. You may want to ask your tutor to show you the line printer gray map so that you can appreciate the differences more fully.

EXAMINE the IMAGES:

Why are the images from channels 1 and 2 similar to each other?

Why are the images from channels 3 and 4 similar but different from those of 1 and 2?

What can you identify on the gray images?

The process used to generate the gray scale maps is "level slicing." If you look on the back of the channel 3 image, you 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 displayed with data values between 24.5 and 50.5 in channel 3 were 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 was used to represent any given pixel. The method of representing pixels by gray tones instead of line printer symbols allows greater ease in visualization of spatial features.

The ranges in the table are calculated on the basis of the response in channel 3 of the pixels specified in the histogram block, shown just above the gray level table. The 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 would be displayed. If you have further questions on this process, please ask your tutor.

Self-Check

1. Explain in your own words why examination of data quality is necessary

2. Name at least two ways in which the remote sensing analyst can examine data quality.

3. Name at least three types of data idiosyncrasies an analyst might find in Landsat data.

4. What two types of geometric correction aid in the analysis of Landsat data?

ACTIVITY 3 COORDINATION OF MULTISPECTRAL SCANNER DATA WITH AVAILABLE REFERENCE DATA

Upon completion of this activity, you should be able to:

1. State one reason for correlating multispectral scanner data with reference data.

2. List at least four kinds of reference data.

3. Correlate the location of ground features apparent on multispectral scanner data with the location of those features on an aerial photograph.

In this step of the analysis, the analyst correlates the multispectral scanner data with available reference data. This operation not only allows him to gain familiarity with the geographic region being analyzed but aids him in developing good training statistics and in evaluating the classification results later in the analysis process. What tools can an analyst use to obtain information about the ground scene?

Aerial photography is one source of information. Photography can be collected with the different systems at various altitudes, resulting in reference data at a range of scales. In general, as a plane flies higher or as focal length decreases, each photograph will cover a larger area, but less detail will be discernible. Another variable in aerial photography is film type. Black-and-white panchromatic or infrared film, color film, and color infrared film all provide different kinds of information about a ground scene, and can serve as reference data for an analyst who understands how to interpret photographic images.

Aircraft multispectral scanner data can also serve as reference data for an analyst working with satellite data, aircraft data can often provide more detailed information about the spectral characteristics of portions of a scene than the satellite data. For example, there may be more spectral bands available, at greater spatial and spectral resolution.

Maps (county highway maps or U.S. Geological Survey maps, for example) and historical records (past crop yields or weather patterns, for instance) can be useful to an analyst by helping him know more about an area and its characteristics.

Another source of information includes observation "at the scene" by the analyst or other personnel. They may include soil moisture samples, crop identification, biomass determination or other detailed measures. These observations can provide the key to successfully relating the spectral responses in the data to the cover types on the ground.

In this part of the workshop we are <u>not</u> interested in locating specific fields. Rather the objective is to establish a familiarity with the scene; where, in general, is the urban area, where is the forest? This general information will be of use in the next step of the analysis.

OBTAIN THE FOLLOWING REFERENCE DATA FROM YOUR TUTOR:

U.S. Geological Survey 7 1/2 minute topographic quadrangle maps of the area

Color-infrared aerial photograph of the area (available as 35 mm slide or 9 x 9 inch print)

The data set with which you are working in this workshop has been preprocessed by geometric correction. Line printer images have been rescaled to 1:24,000, which matches the scale of the U.S. Geological Survey 7.5 minute topographic series. The scale of the printer-plotter images is approximately the same as the available aerial photography reference data.

ASSOCIATE the REFERENCE DATA with the PRINTER-PLOTTER IMAGES

Using the channel 1 and 4 gray scale images, find and mark with a pen the general location of the cover types 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.)

Note: Some features are more apparent on one image than the other.

Self-Check

1. State one reason for correlating multispectral scanner data with reference data.

2. List at least four kinds of reference data.

ACTIVITY 4 SELECTION OF CANDIDATE TRAINING AREAS

Upon completion of this section, you should be able to:

1. State why training areas must be selected.
2. Name at least two considerations that affect the selection of candidate training areas.
3. Select candidate training areas.
4. Compare supervised and nonsupervised approaches to generation of training data.

The next step in this analysis of multispectral scanner data is the selection of candidate training areas. To explain what training samples are and why they are needed, some pattern recognition concepts should be reviewed. The pattern recognition algorithms we will deal with require that examples of typical data from each class of interest be supplied to the computer programs. These data, called training samples, are used to set certain parameters for the pattern recognition algorithms, in effect "training" the computer to recognize the classes of interest. When the classification operation is carried out by the pattern recognition algorithms, each data point to be classified is "compared" to the training sample for each class, and the pixel is assigned to the class it "most likely" belongs to, that is, to the most similar class.

There are two major methods of obtaining training samples. The first, referred to as the supervised approach, involves locating individual, pure fields of pixels, each of which represents a single cover type. In this case all fields identified as belonging to the same cover type are grouped and used as the training samples. The second, referred to as the nonsupervised approach, does not utilize reference data to select and group pixels to form the training samples. Rather a systematic sample of pixels is selected from the study area and processed to dileneate groups of pixels within the sample that are spectrally similar. A hybrid approach involves locating blocks of pixels, referred to as candidate training areas, that contain several cover types. Each candidate training area is then processed as in the nonsupervised approach to identify groups of pixels that are similar on the basis of their spectral characteristics. For each block processed in this way a map is then printed showing into which group each pixel was assigned. In this hybrid approach the cover type identity of each group of spectrally similar pixels is then determined by comparing the map of each block to the reference data.

The nonsupervised approach has as its strength the ability to assemble into a group spectrally similar pixels regardless of their spatial positions. This can be advantageous when working with a heterogeneous scene in which the likelihood of observing several adjacent pixels of the same cover type is low. Therefore, when working with a mountainous terrain, for example, in which there are many slope, aspect, elevation, vegetation combinations, this nonsupervised approach is often utilized to group pixels which are spectrally similar and therefore probably belong to the same ground cover type, but which are not necessarily next to each other.

A common mistake is to assume because points are spatially adjacent and belong to the same covertype that they are spectrally similar. As the song goes, "It ain't necessarily so." In the agricultural scene if we think of a corn field, experience tells us that within that field there may be low spots where the corn is greener and has greater ground cover and high spots where there is slight moisture stress and the corn is therefore thinner and less green. There are many other reasons for spectral variability within a cover type, and it is safe to say that almost all cover types contain at least two groups of pixels that are specially different. The nonsupervised approach can be used to identify such spectrally similar pixels which may represent a special condition or variety of a ground cover type.

The hybrid approach allows an analyst to take advantage of the strengths of the nonsupervised approach while still establishing the identity of each training sample or group.

The first step in obtaining training samples in the hybrid approach described here is selection of candidate training areas. Experience has indicated that when candidate training areas are from 50 to 100 lines by 50 to 100 columns in size, the computer time required to arrive at the spectral groups of points will be minimal and the task of identifying each group will be less time consuming.

To select these candidate training areas, an analyst begins by reviewing the analysis objectives. In stating the objectives, the cover types of interest are listed. These cover types are called information classes. Candidate training areas are selected in such a way that every information class is represented in at least one of the areas. When possible, each information class is included in more than one candidate training area. This increases the likelihood that the training data will be representative of all the variations in cover types in the scene being analyzed. When training data is representative the classification of pixels is more likely to be correct.

A common procedure for selecting the candidate training areas is to identify in the available reference data areas that contain the information classes. The analyst also locates these areas on images of the multispectral scanner data. You did this in activity 3. With that information in mind the analyst selects candidate training areas, following the guidelines indicated previously: each area is from 50 to 100 lines by 50 to 100 columns, each area includes more than one cover type, and every cover type is included in at least one (preferably two or more) candidate training areas. To help insure that the training data is representative, the candidate training areas should be distributed somewhat uniformly throughout the area to be classified; however, this may not be possible if reference data is limited. Usually, representative training data for all information classes can be obtained by selecting from three to six candidate training areas.

SELECT three candidate training areas which are representative of the scene. Use the available reference data, the gray scale printouts and the guidelines described above. Make sure that every cover type of interest (urban, agriculture, forest and water) is included in at least one of the candidate training areas. Outline your areas on the channel 2 gray scale image with a felt tip pen. Then CONSULT your tutor and be able to justify your selection of candidate training areas.

Self-Check

1. State why training areas must be selected.
2. Name two considerations that should go into the selection of candidate training areas.

ACTIVITY 5 CLUSTERING CANDIDATE TRAINING AREAS

Upon completion of this section, you should be able to:

1. Describe at least two tasks the CLUSTER processing function can perform.

2. State the rule-of-thumb used to determine the number of clusters to request and the reason behind it.

The concept of "training samples" has been discussed, and candidate training areas chosen. The progression from candidate training areas to training samples is a complex process as well as a crucial one. The process often is somewhat circuitous, as indicated in Figure 8.

The CLUSTER processor can use information from more than one channel or wavelength band (four channels in this workshop) to produce a single image. When multiple channels are used, only pixels whose data values in those wavelength bands are similar are grouped. After the clustering process is finished a map is printed showing to which cluster each pixel was assigned. Pixels assigned to the same cluster are represented by the same symbol. Thus, spectrally homogeneous (in the wavelength bands used) areas within the data will also be displayed with the same symbol. In this way the cluster processing function accomplishes boundary enhancement, in that such spectrally homogeneous areas can be more easily identified.

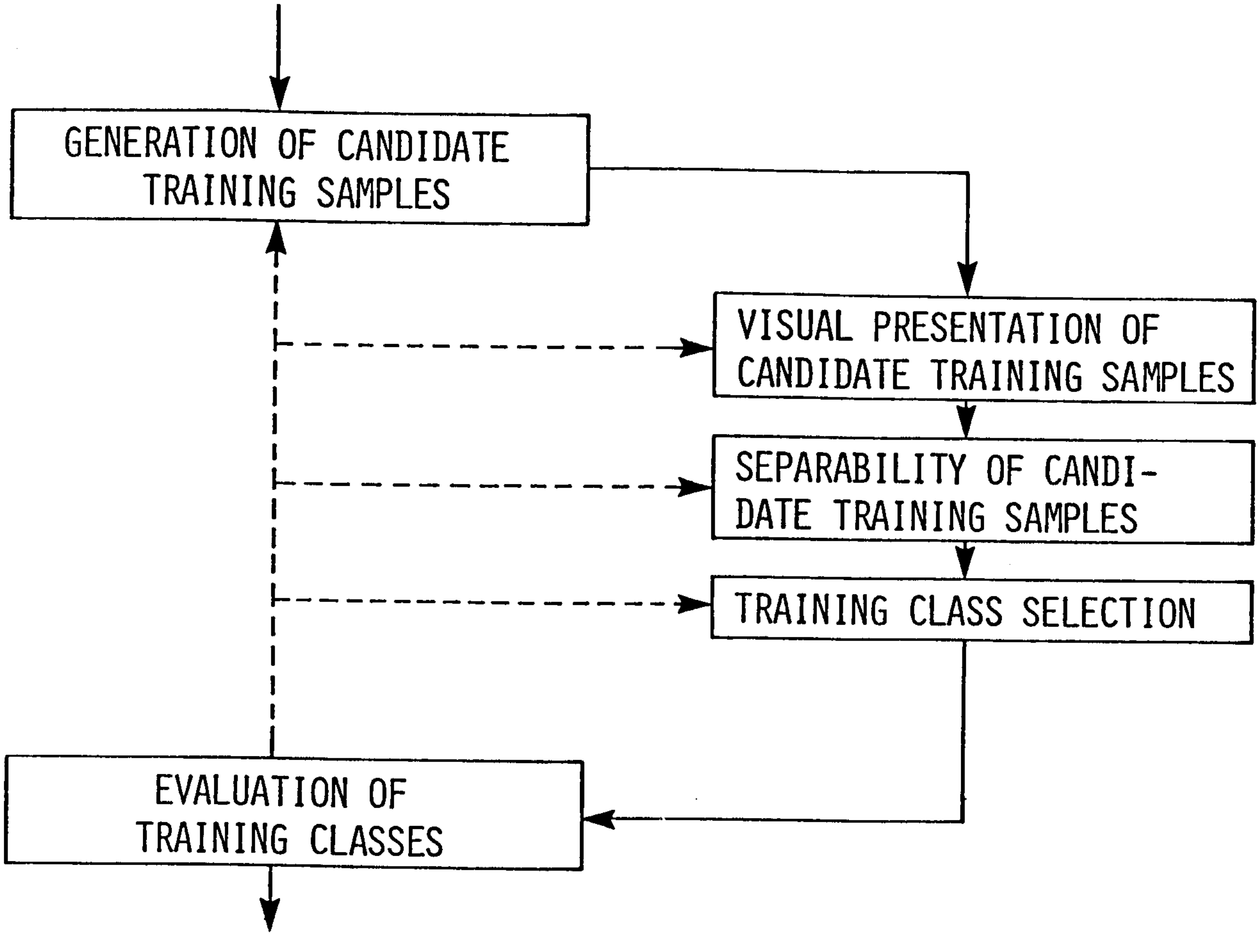

GENERATION OF CANDIDATE TRAINING SAMPLES
VISUAL PRESENTATION OF CANDIDATE TRAINING SAMPLES
SEPARABILITY OF CANDIDATE TRAINING SAMPLES
TRAINING CLASS SELECTION
EVALUATION OF TRAINING CLASSES

Figure 8. Flow chart indicating the steps involved in refinement of candidate training areas. Dashed lines indicate potential iteration loops. This is a portion of the flow chart shown in the Introduction, Figure 1.

The clustering algorithm is called a nonsupervised classifier because it groups ground points strictly on the basis of multi-channel data values associated with each of the ground points in the training area. Neither the location of the pixels relative to one another (spatial information) nor ground cover type is considered in determining the clusters. Rather, it groups those pixels with similar response values in the multiple channels. These natural groupings in the data are called cluster or <u>spectral classes</u>.

Thus, another task the CLUSTER processing function can accomplish is to determine spectral classes within a data set.

When data from a ground scene is clustered there is a tendency for the data points within each cluster class to be distributed in a Gaussian fashion. Figure 9 shows a typical Gaussian function in one dimension -- commonly called a "normal curve." Figure 10 shows a two-dimensional Gaussian density function. The fact that clusters in remotely sensed data tend to be Gaussian is important because several of the classification algorithms used are based upon a Gaussian assumption, i.e., that the distribution of the data values within each of the classes to be classified can be approximated by Gaussian density function.

Often more than one Gaussianly distributed cluster is necessary to represent an information class. As an example, an agricultural crop might exhibit a multimodal distribution (more than one peak) due to different soils, moisture content, planting dates, crop density, seed varieties, or a combination of these factors. If necessary to satisfy the Gaussian assumption the multimodal non-Gaussian density function in Figure 11 could be decomposed into two Gaussian components by clustering, as shown in Figure 12. These components are commonly referred to as subclasses. The subclass concept is an important one as it allows the analyst to use a classification algorithm based upon a Gaussian assumption even though the information class distributions may be non-Gaussian.

In this step of the analysis, the CLUSTER processing function is used to subdivide each training area into cluster classes. Enhanced boundaries on the cluster maps will be used later to help establish associations between the cluster classes and information classes.

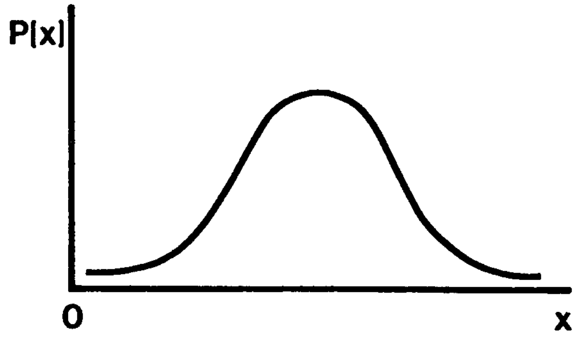


Figure 9. Gaussian density function in one dimension.

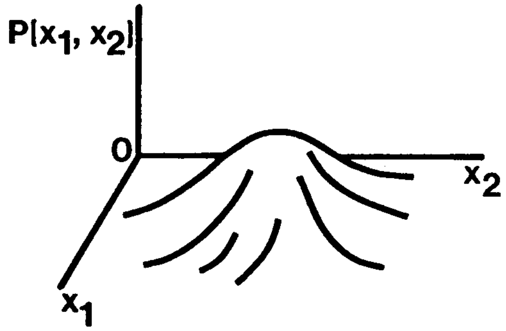


Figure 10. Gaussian density function in two dimensions.

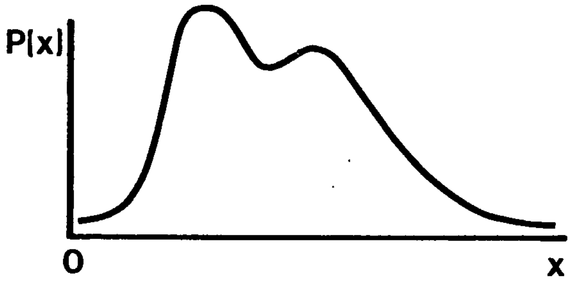


Figure 11. Multimodal non-Gaussian density function.

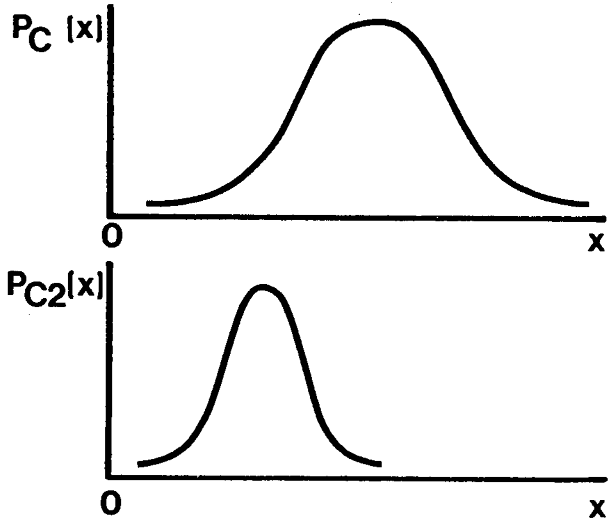


Figure 12. Multimodal function decomposed into two Gaussian components.

The clustering algorithm requires that the analyst specify the number of clusters that the data is to be grouped into. Experience has indicated that most cover types contain at least two subclasses, and a rule-of-thumb is for the analyst to request twice the number of information classes present. If an analyst requests an insufficient number of clusters, the cluster variances will be unusually high and the resulting classes will be difficult to separate. If "too many" clusters are requested, cluster variances will be small but the resulting classes may be difficult to identify.

However, the situation of having too many clusters can be remedied by pooling clusters whereas when too few clusters are obtained the entire-clustering process must be repeated. A "good" number of clusters, such as is suggested by the "2x" rule, will optimize these trade offs. However, in some cases it will be evident after examining the output that a different number of clusters is needed.

The cluster algorithm groups individual data points into a predefined number of groups (clusters) specified by the analyst. The computer assigns a location in the multidimensional space as the initial center of each cluster (see Figure 13). It then calculates the multidimensional distance between each data point in the data be clustered and each cluster center and assigns each point to the nearest cluster. New cluster centers are then determined by calculating the mean vector for the data points assigned to each original center. The computer then recalculates the multidimensional distance between each data point and the new cluster centers and reassigns each sample to the closest newly defined cluster center. The computer continues the cycle of calculating the cluster centers and reassigning data points until the percentage of data points that are not reassigned to a new cluster center reaches a value known as "convergence," which is specified by the analyst. Specifying a convergence value less than 100%, for example 98.5%, can result in a significant saving of computer time without seriously affecting the clustering results. See section 3.10 of Swain and Davis for more details.

The cluster output for the first candidate training area is shown on pages 8-43 of your computer output.

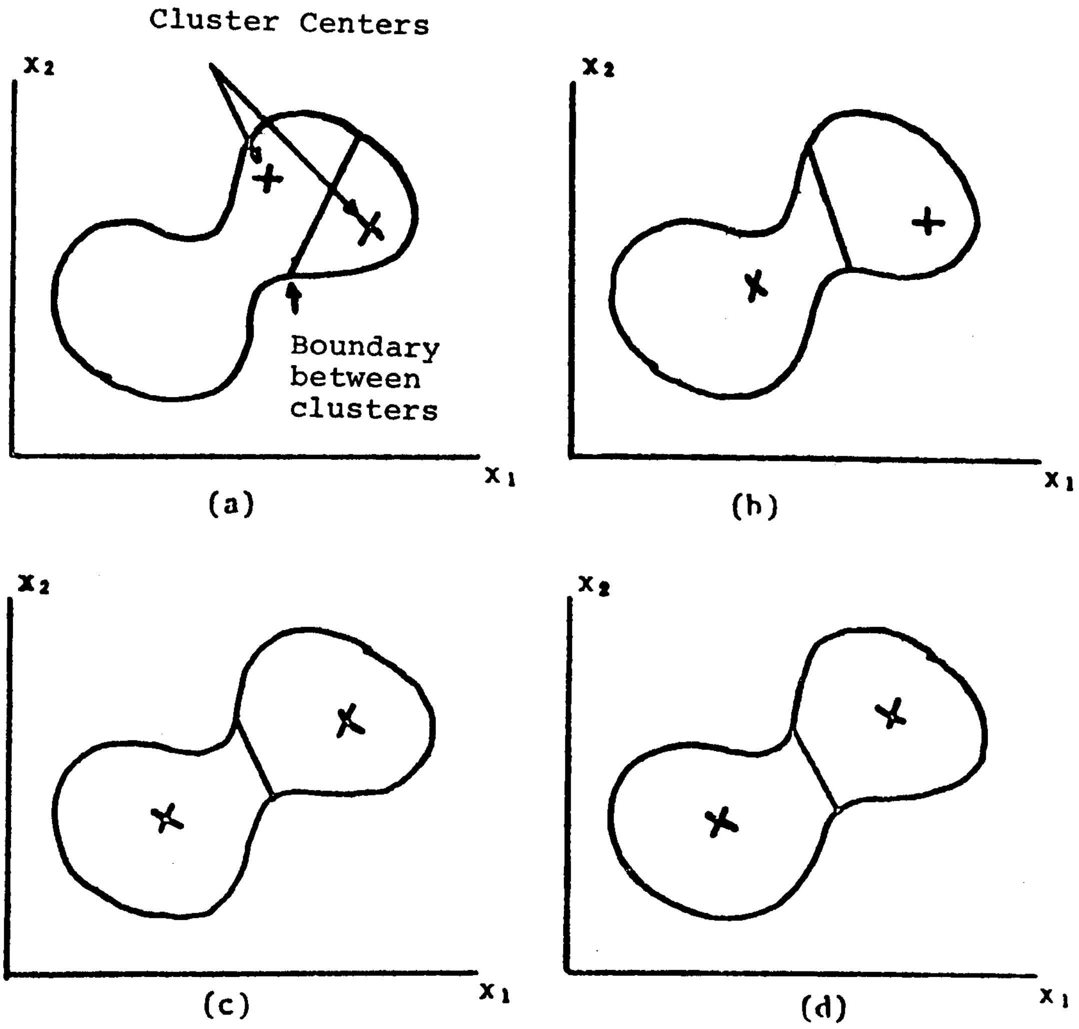


Figure 13. A sequence of Clustering Iterations (a) Initial Cluster Centers (b) (c) Intermediate Steps (d) Final Center Configuration.

Page 9 reveals the fact that although the analyst chose three candidate training areas, each area was clustered separately. This procedure saves computer time and tends to produce cluster maps with more distinct boundaries, which is very important for the next step of the analysis.

The cluster means and variances for the first candidate training area are given on page 10. Also listed are the number of data vectors (pixels) assigned to each cluster. Note that 15 clusters were requested and obtained. Six information classes, water, forest, pasture, bare soil, urban, and emerging crops were identified giving 12 as the suggested number of clusters. 15 clusters were requested to allow for the various transitions between the information classes. In general cluster 1 has larger mean values than the last cluster and 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 one 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 step of the analysis.

Pages 13-42 show histograms of the data values in each channel for each cluster class. Each histogram shows the distribution of the data values in one channel of the pixels grouped into each cluster by the clustering algorithm. A question at this point might be: How can each distribution be described to the classification algorithm? One precise method might be to count the number of times each possible response value occurs in each distribution. Although such an approach would be feasible when working in only one dimension, to describe the multivariate nature of each cluster would require a very large number of parameters. If we can assume that each distribution is in reality normal or Gaussian (bell shaped curve), we can then describe each distribution with only two parameters, the mean and the variance. In this case this is precisely what was done, and as you already saw the means and variances are shown on page 10. When using this approach, the analyst examines each histogram noting the degree to which the cluster class can be approximated by a Gaussian (normal) density function.* Note the rather non-Gaussian character of the channel 2 histogram of cluster 11 (pg 33). Large variances reported for certain clusters are reflected in the histograms by the way the data is spread out.

When calculating statistical parameters to describe classes, as was done for the cluster classes, care must be taken to insure that a sufficient number of training observations, in this case pixels, are available to base the calculations on. Theoretically for each cluster one more pixel than the number of wavelength bands being considered is required. However, to avoid problems due to the possibility of pixels having identical data vectors and to heighten the representativeness, at least 10 times as many pixels as the number of wavelength bands is usually suggested as the minimum number to characterize each cluster.

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

*A subtle point here is that the density functions shown are marginal (one dimensional) density functions and there is no guarantee that even if each marginal density is reasonably Gaussian the joint density function (four dimension density function in this case) is Gaussian. If a marginal density function is non-Gaussian, however, the joint density function is also non-Gaussian.

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 few than 40 pixels

6. A grouping table is shown at the end of the output for each training area (pages 43, 81 and 115). The groupings suggested there were based upon a measure of the similarity of the clusters, a concept which is discussed in ACTIVITY 9. Although the CLUSTER processor normally prints a grouping table at the end of the output the grouping table shown on pages 43, 81 and 115 are not the original ones and were not based upon the distance measure normally used by the CLUSTER processor.

Self-Check

1. Describe two tasks the cluster processing function can accomplish for you.

2. State the rule-of-thumb used to determine the number of cluster classes to request and the reasoning behind it.

ACTIVITY 6 ASSOCIATION OF CLUSTER CLASSES AND INFORMATION CLASSES

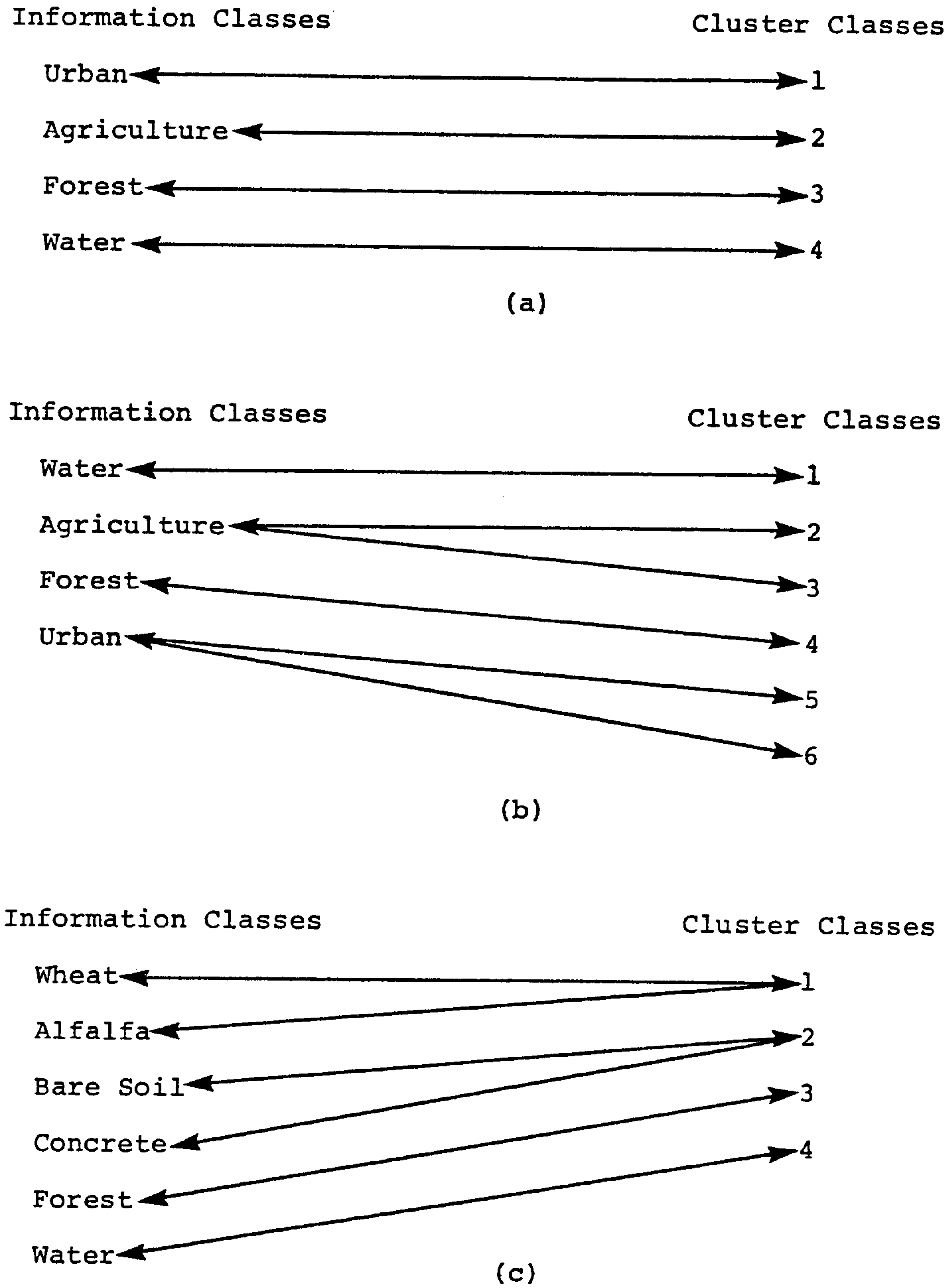
Upon completion of this activity, you should be able to:

1. Describe why cluster classes are associated with information classes.

2. Given printed cluster maps and available reference data, associate cluster classes with information classes.

Up to this point in the process of meeting the analysis objective, the Landsat imagery has been examined for data quality, the imagery has been correlated with reference data, candidate training areas have been selected and the data within each candidate training area has been clustered. Previous activities have introduced the concepts of information classes and cluster classes. Recall the information classes were defined during the process of stating the analysis objectives. Cluster classes for each candidate training area were determined by means of the CLUSTER processing function.

The purpose of this step in the analysis sequence is to associate each cluster class identified in the previous step with an information class (e.g., agriculture, urban, water, forest). It should be pointed out that there is not necessarily a one-to-one correspondence between the information classes and the cluster classes. Remember, an information class is a distinct cover type of interest as noted above, while a cluster class is a group of data points which are spectrally similar. As shown in Part (a) of Figure 14, there may be a one-to-one correspondence between the two. It is more possible that several cluster classes will represent the same cover type (information class) as shown in Part (b), Figure 14. Sometimes several information classes will be associated with the same cluster class (Part



(a)

(b)

(c)

Figure 14. Examples of relationships between different information classes and their cluster classes.

(c), Figure 14). The latter situation indicates that the cover types are spectrally similar. It may be possible to separate the classes into different clusters by requesting a larger number of clusters.

To identify the clusters obtained, maximum use is made of all available reference data, so that the cluster classes can be reliably identified. Note that if incorrect identifications are made in this step they will be carried through to the classification step, resulting in incorrect maps and acreage estimates. The association of cluster classes and information classes is difficult and time consuming, but this step is most important for insuring that the classifier is correctly trained.

Reference data often includes aerial photography. An overhead projector can be used to superimpose a 9" x 9" transparency on the printed cluster maps. By varying the projector-to-wall distance, it is possible to project the transparency to the scale of the printout, and if the data has been geometrically corrected, a match can be obtained. A better match can be achieved with a 2" x 2" slide and slide projector.

U.S. Geological Survey quadrangle maps of the area are also useful for location purposes. This is especially true when working with lineprinter output of data that has been geometrically corrected and rescaled to the scale of the quadrangle maps. Lineprinter output can be directly overlaid onto the 7 1/2 minute maps. A light table is useful for this purpose.

An instrument that makes this task even easier, a zoom transfer scope, uses a lens system to adjust the scale of two images or maps to match each other. By viewing a cluster map and aerial photography superimposed on one another, each cluster can be identified reliably and quickly.

There are several points to remember. One point is that if a single cluster appears to correspond to more than one information class, it should be identified that way.

Another point is that cluster classes from different training areas but displayed with the same symbol on the maps of each training area do not necessarily correspond to the same information classes. Since each candidate training area was clustered separately, the results of clustering describe the clusters from only that area. In this example, each candidate training area contained different information classes.

ASSOCIATE cluster classes with information classes. Using the computer-generated maps from clustering the data, 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 candidate training area. The Landsat data used in this example was geometrically corrected and rescaled so that when printed on the line printer 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 potential crops will 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.

After you have identified each cluster class in the first candidate training area, record your results in Table 1.

SEE YOUR TUTOR WHEN THIS ACTIVITY IS COMPLETED or if you have questions.

Self-Check

1. State why cluster classes must be associated with information classes.

Table 1. Association of Cluster Classes in the Information Classes

Cluster		Candidate Training Areas		
Number	Symbol	Training Area 1	Training Area 2	Training Area 3
1	.			
2	-			
3	=			
4	/			
5	+			
6	*			
7	I			
8	H			
9	Y			
10	L			
11	P			
12	U			
13	O			
14	X			
15	$			
16	M			

ACTIVITY 7 AUGMENTING THE CANDIDATE TRAINING DATA

Upon completion of this activity, you should be able to:

1. Describe a situation in which the analyst would want to augment his candidate training data.

2. Contrast the supervised and unsupervised approach to obtaining training samples.

It was mentioned earlier that generation of the training samples for the classifier is often an iterative process. Upon completion of the cluster-class/information-class associations, it is important to check that all information classes known to be in the scene are represented. In this example none of the candidate training areas contained clouds or cloud shadows. Since the classifier will be required to assign each data vector to one of the classes, not including cloud and cloud shadow classes would force the classifier to make a number of errors. Variations in water quality in the Monroe Reservoir also can be observed on the color IR photograph and is especially evident in the eastern portion of the reservoir. Based on this observation the possible advantages that might result from trying to identify several subclasses of water should be investigated.

The previous two activities used an unsupervised approach to the selection of training samples. The clustering processor was used to determine the inherent spectral similarities within each candidate training area and the cluster classes were related to information classes with the aid of reference data (color IR photo). A supervised approach can now be used to select training samples for clouds, cloud shadows and additional water classes. In the supervised approach areas within the scene which are known to contain particular cover types are identified. Because of their distinct geometric character and size it is relatively easy to select from the scene samples of clouds, cloud shadows and water.

Using the printer-plotter gray scale images and the color IR photo, specify the line and column coordinates for a number of cloud, cloud shadow and water 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 data.

In selecting training areas bear in mind the minimum number of data vectors needed to estimate the statistical properties of the training sample.

Upon completion of this activity see your tutor and discuss your selections.

Recall that we are working towards making a classification of the data into a set of predefined information classes. The classification will be done on the basis of the similarity of the spectral values of the points to be classified to the spectral characteristics of the training classes. At this stage we have obtained some cluster classes to which there have been associated informational names, and in addition some areas on the ground representing specific cover types (clouds, shadows, water) have been identified. However, the spectral characteristics of the candidate training classes have not been emphasized up to this point. The statistical description of the spectral characteristics of training classes may be obtained through the use of one of several computer processing functions. For instance, an analyst may obtain from the CLUSTER processing function a deck of cards which contains statistical information for each of the cluster classes. Specifically, the statistical descriptions of each training class consists of the mean vector, composed of the averages of the data values in the class in each wavelength band, and the covariance matrix. The covariance matrix is a multivariate generalization of variance and is used to characterize the multidimensional "spread" or dispersion of the data. These two statistical parameters provide all of the information needed to define a multivariate Gaussian probability density function. Several of the classifiers that will be used later are based upon an assumption that the classes of interest may be represented by Gaussian density functions.

Histograms of the training data in each spectral band is another output available from CLUSTER. We saw illustrations of this earlier and studied those histograms in order to identify any obviously nonGaussian characteristics.

At this point areas of cloud, cloud shadow and water have been located but the statistical characteristics of these training samples have not been obtained. Because of an interest in possibly determining subclasses of water, the five water training areas were clustered. Since the areas are all small and since they all contain the same cover type, the five areas were clustered together. This illustrates an alternative use of the CLUSTER processor. The computer output for that step is shown on pages 117 to 135. Five clusters were requested as two to three varieties of water were anticipated. 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 class variances are all small except for channel 3, cluster class 1.

As mentioned above, a statistical description of the various cluster classes may be obtained from the CLUSTER processor itself. A different processor was used to obtain the necessary statistical description of the cloud and shadow areas. This processor, known as the STATISTICS processor, requires as input the line and column coordinates of the areas for which statistics are to be calculated. The algorithm reads the data values of the pixels within the coordinates specified from the data tape and calculates the means and the covariances of those data vectors without clustering. The STATISTICS output for the cloud areas is shown on pages 138 to 140 of the printouts and the output for cloud shadows is shown on pages 141 through 143.

Examine the CLUSTER processor output for the water training areas and the STATISTICS processor output for the cloud and cloud shadow training areas. Note the following:

number of pixels in each class

histograms for each water subclass

correlation matrix for clouds

correlation matrix for cloud shadows

histograms for clouds and cloud shadows

Discuss the output with your tutor

Self-Check

1. Name the two sets of statistical parameters which define multivariate Gaussian distributions.

2. Explain why statistics are needed at this point in the analysis.

3. Name one difference between the supervised and nonsupervised approach to generation of candidate training classes.

ACTIVITY 8 VISUAL REPRESENTATION OF CANDIDATE TRAINING CLASSES

Upon completion of this activity you should be able to

1. State at least one method of visualizing the spectral characteristics of candidate training classes.
2. Given a bi-spectral plot identify the general location of water, green vegetation, and bare soil on the plot.

At this point in the analysis sequence, candidate training areas have been chosen and clustered, the cluster classes have been associated with information classes, the cluster classes have been augmented with water, cloud and cloud shadow training samples, and the statistical characteristics of all candidate training classes have been calculated. It would be possible at this point to use all of these candidate training samples to train the classifier, but that is usually not done for at least two reasons. First, the number of training samples (generated by CLUSTER or picked using the supervised approach) available at this point is normally greater than the number of classes needed to adequately train the classifier. For instance, one of the cluster classes in area 1 identified as forest may be spectrally similar to the cluster class of forest in area 3. An analyst would like to reduce the number of training classes in such cases, because this will save computer time and simplify interpretation of results. Also, some of the clusters may have too few data points to get good estimates of the mean vector and covariance matrix. By combining spectrally similar clusters, the number of data points used to calculate the mean vector and and covariance matrix for the resulting pools of training classes will be greater, and as long as the clusters being combined are spectrally similar, these combinations will generally lead to a better representation of the cover types.

To begin the analysis of which candidate training classes may be duplicates and which others are distinct, it is useful to visualize the spectral characteristics 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 of the mean values in the two Landsat near infrared channels for each candidate training class. On the other axis is plotted the average of the mean values in the two visible bands. By plotting the average of the means in the two near infrared bands versus the average of the two visible bands, information from the 4-dimensional measurement space can be displayed in two dimensions. The rationale for averaging the means in this way is based on the observation that responses in the two visible bands are highly correlated, as are responses in the two IR bands (this may be observed by comparing the two visible gray scale images and the two IR gray scale images). 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 of the means of the visible and near infrared bands, and the symbol used to represent each class. The first 15 classes may be recognized as coming from candidate training area 1, the next 16 from candidate training area 2, and the next 14 from candidate training area 3. The next set of five are the water training classes generated by clustering a supervised sample and the last two represent cloud and cloud shadow respectively.

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 thru 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 A_2.

See your tutor if you need help.

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.

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.

Ask your tutor to explain the "tassled cap" concept.

Self-Check

1. Describe one method of 'seeing' the spectral characteristics of candidate training classes.

2. Identify the regions of a bi-spectral plot belonging to major cover types.

3. State two reasons why spectrally similar cluster classes are combined.

ACTIVITY 9 CALCULATION OF STATISTICAL DISTANCES BETWEEN CANDIDATE TRAINING CLASSES

Upon completion of this activity, you should be able to:

1. Given two pairs of one-dimensional density functions, identify the pair which is separated by the larger statistical distance.
2. Name the two characteristics of Gaussian probability density functions which determine the statistical distance between the density functions.

Although at this point it is possible to visualize and compare all candidate training classes, the comparison is on the basis of the means of the classes only. Therefore before any decisions are made about which classes to pool or delete, it will also be useful to consider the amount of dispersion or variability each class has.

The second major reason for this step is to give the analyst some indication of the probability of correct classification in advance of doing the classification. If there appears to be considerable confusion among information classes, more clustering may be done on the areas already used, asking for a different number of clusters; alternatively additional candidate training areas could be selected in an effort to get improved distinction among classes, instead of trying to combine the cluster classes he already has.

Calculating the "separability" of the cluster classes can help determine which cluster classes are similar, and it can serve as an indicator of probability of correct classification.

To explain how this is accomplished, the concept of statistical distance must first be discussed. Figure 15 shows two examples of one-dimensional density functions. Intuitively you know that the "distance" between the density functions is greater in case B than in case A. The distance between two Gaussian probability density functions depends not only on the distance between the mean values but also on the "spread" of the data. Figure 16 illustrates this point. The distances between the mean values are equal in both of the cases shown, but the smaller variances (smaller "spread") in case B of Figure 16 result in a larger statistical distance between the two density functions.

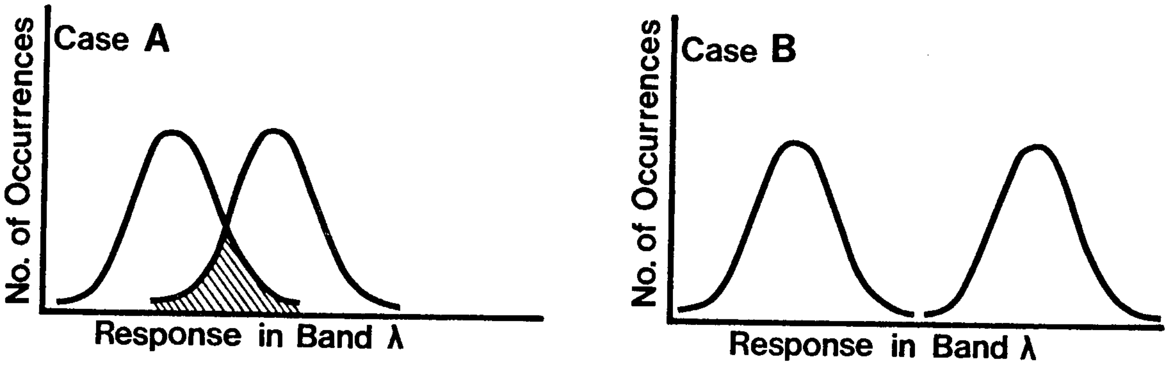


Figure 15. Two pairs of one-dimensional density functions. The statistical distance between the density functions in case A is less than in case B.

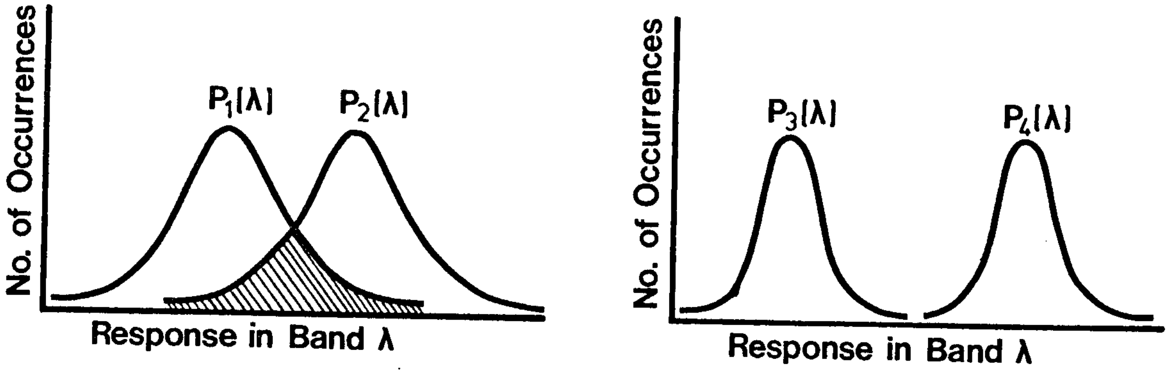


Figure 16. Both pairs of distributions shown above have equidistant means, but the smaller variances of P_3 (lambda) and P_4 (lambda) cause this pair to have larger statistical distance.

In two dimensions the density functions may be represented by ellipses (Figure 17). In three or more dimensions, as with four-channel Landsat data, the density functions are represented by ellipsoids, blimp-like surfaces of equal probability in the measurement space. As in the one dimensional case, the statistical distance in two or more dimensions is an estimate of the overlap of the density functions. As shown in Figure 17, the overlap depends not only on the size and shape of the ellipsoids but on their orientation as well. In case B of Figure 17, the overlap is greater due to the slight shift in the direction of one ellipse. Therefore, we would anticipate the classification accuracy in case B to also be less.

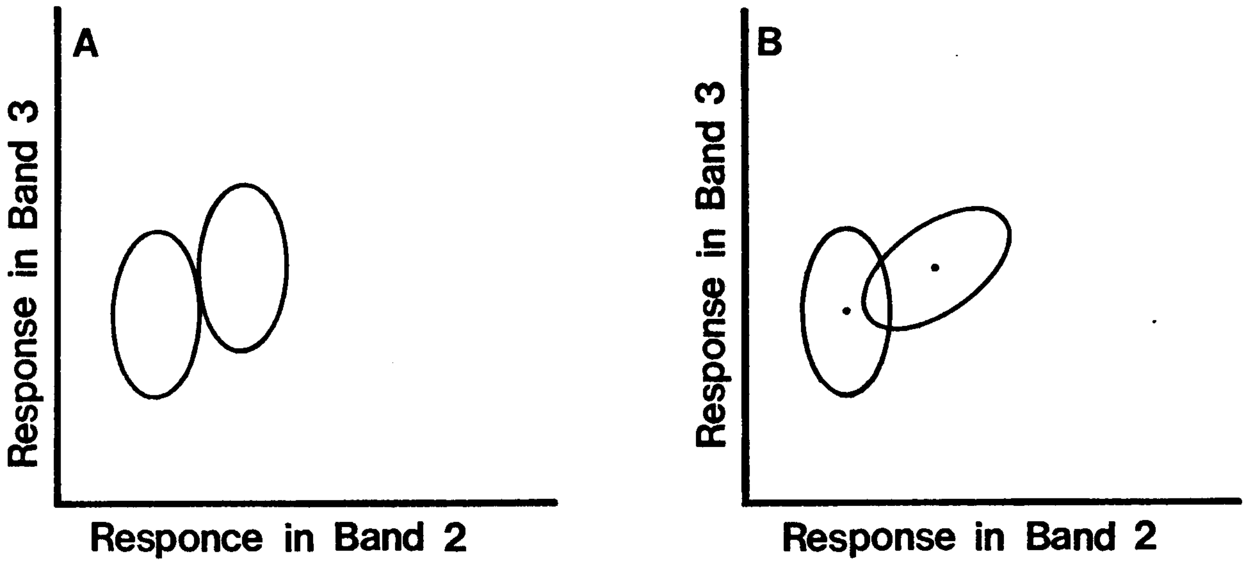


Figure 17. Ellipses representative of training classes.

There are a number of ways of mathematically defining statistical distance. One way which performs well in estimating the probability of correct classification between pairs of classes is transformed divergence. In particular, experimental results of plotting probability of correct classification versus transformed divergence for training data are shown in the graph in Figure 18. Notice that class pairs with larger transformed divergence values (D_T) also achieved a higher classification accuracy (P_C) although the relationship is not perfectly linear nor one-to-one. This graph can help in determining what the minimum acceptable transformed divergence value between pairs of classes should be. According to Figure 18 to achieve 85% accuracy, as is desired for the analysis being pursued, we see that the final training classes should have transformed divergence values between about 800 and 1800. Classes with transformed divergence of 800 would achieve 85% accuracy only infrequently. On the other hand at 1800, 85% accuracy could almost always be obtained. Unfortunately although one might be tempted to require a very high threshold, as higher and higher transformed divergence values are required, only more general classes will be remain distinct. Therefore, a balance between the level of detail desired and the minimum allowable transformed divergence must always be struck, one at which accept-

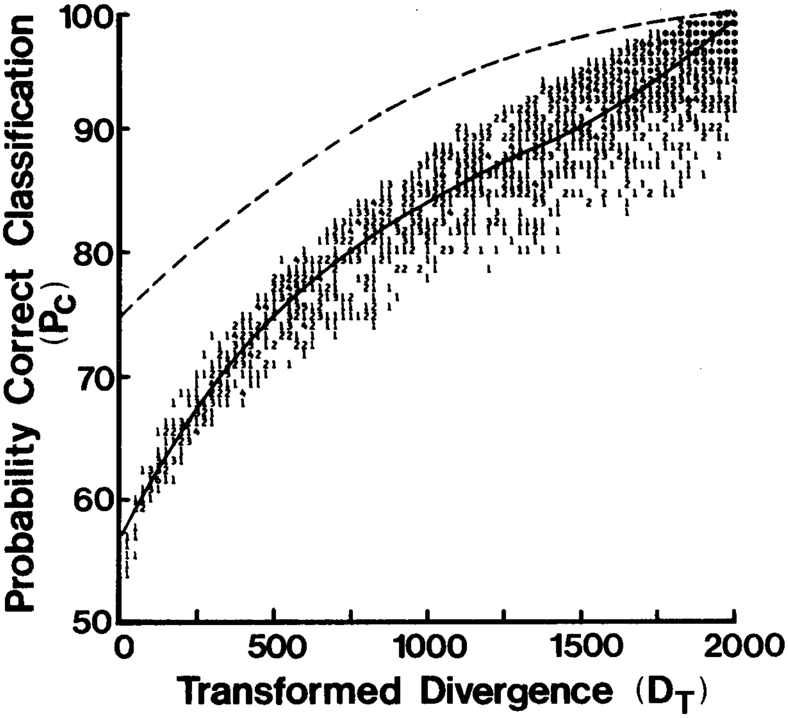


Figure 18 Observed values of probability of correct classification versus transformed divergence.*

*Taken from "Two Effective Feature Selection Criteria For Multispectral Remote Sensing" by P. H. Swain and R. C. King, presented at the International Joint Conference on Pattern Recognition, Washington, D.C., November 1973.

able accuracy will be obtained most of the time for classes of the desired informational content. More information about statistical distance measures can be found in sections 3.7 and 3.8 of Swain and Davis.

The SEPARABILITY processor was used to help determine which clusters could be combined. See pages 150 through 168 of the computer printouts for an example of SEPARABILITY output. 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 N2 etc.

COINCIDENT BI-SPECTRAL PLOT (MEAN) FOR CLASS(ES)

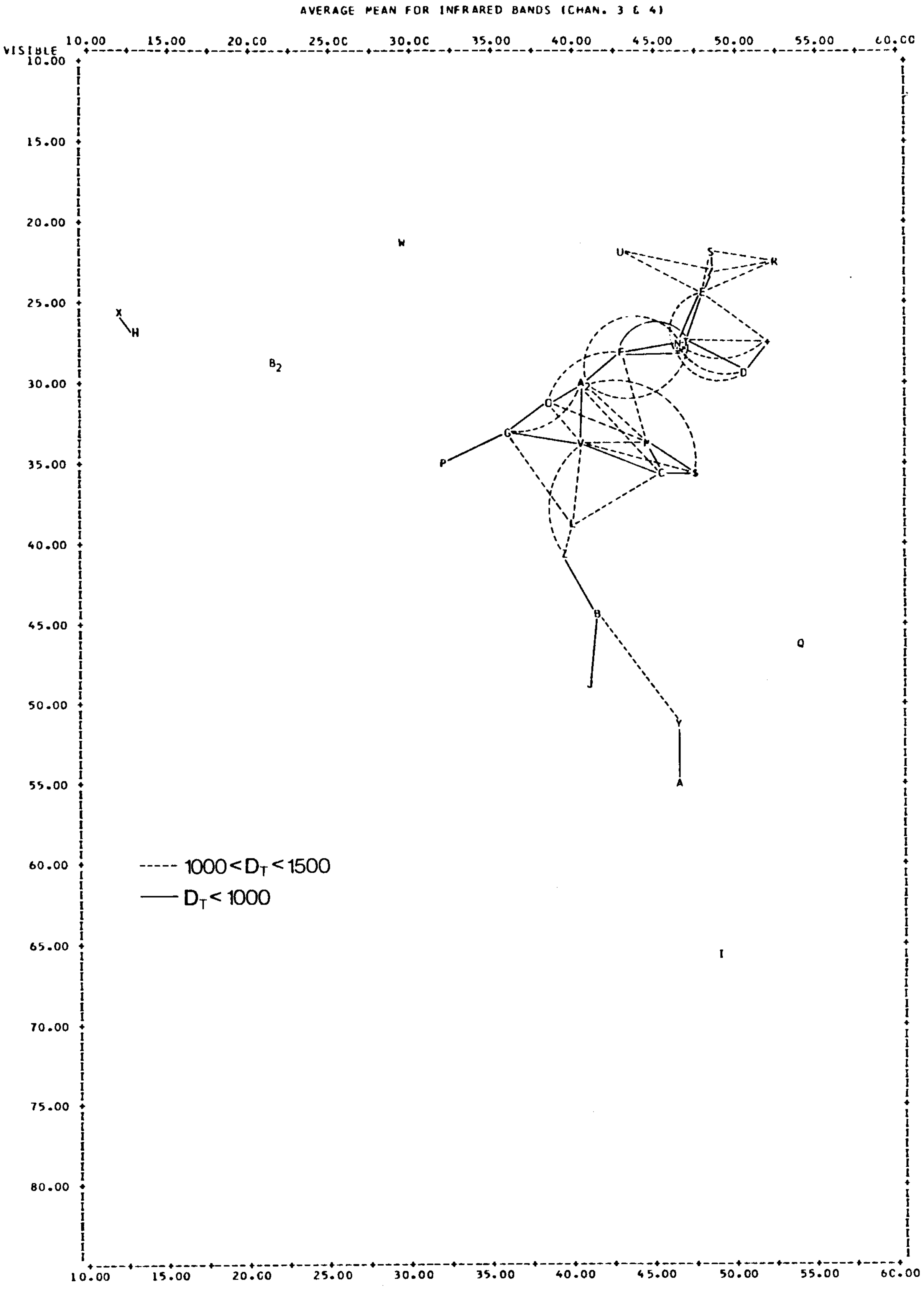


Figure 19. SEPARABILITY information added to bi-spectral plot.

An analyst typically combines information from the SEPARABILITY processor and the bi-spectral plot by adding the separability information to the bi-spectral plot. This is done as illustrated in Figure 19. A solid line is drawn between two classes on the bi-spectral plot if the transformed divergence between the pair of classes is less than 1000. A dashed line is drawn between the symbols if the transformed divergence is between 1000 and 1500. No line is drawn if the transformed divergence is greater than 1500.

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 (page 147-148) using the technique illustrated in Figure 19.

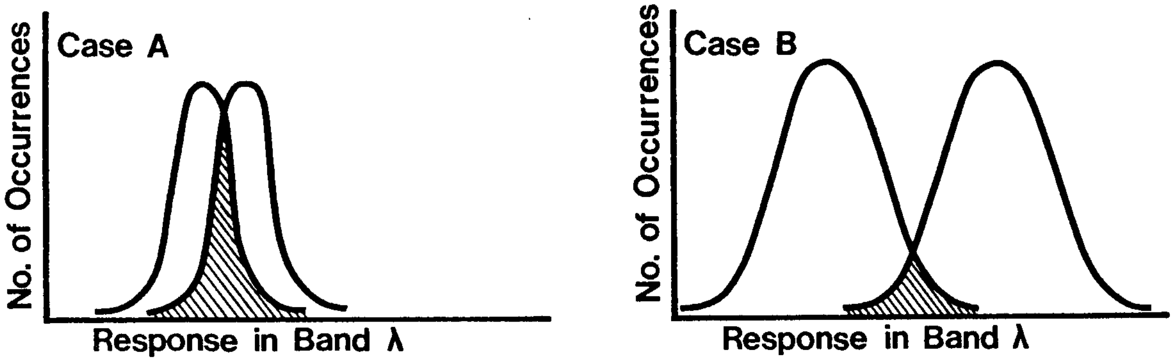


Figure 20. Two pairs of one-dimensional density functions.

Self-Check

1. Look at the two pairs of one-dimensional density functions shown in Figure 20. In which case is the statistical distance between density functions larger?

2. Name the two characteristics of Gaussian probability density functions which determine the statistical distance between the density functions.

3. What is the value of knowing the statistical distance between all possible pairs of classes?

4. How can the minimal allowable value of transformed divergence be determined?

ACTIVITY 10

SELECTION OF TRAINING CLASSES

Upon completion of this activity you should be able to:

1. Name two desirable characteristics of training classes.
2. Given a separability diagram and a list of the identities of each spectral class, select a set of training classes to use in the classification.

The final step in the process of refining the candidate training classes is to specify the training samples to be used by the classifiers. So far, refinement of candidate training areas has included clustering the areas, associating cluster classes with information classes (cover types) augmenting the cluster classes with water, cloud and cloud shadow samples, calculating statistics of the candidate training classes, running separability using transformed divergence to get a measure of the distance between clusters, and summarizing these results on the bi-spectral plot. Next, decisions will be made as to which candidate training classes should be grouped for training and which candidate training classes should be used alone. The bi-spectral plot and the cluster-class/information-class associations will be used to help make these decisions. Observations of the normality of class histograms and the number of points in each candidate training class will be used to help resolve difficulties.

A list of candidate training classes identities is shown on pages 67 and 68. This list is a summarization of the association cluster classes with information classes and augmenting the candidate training data steps. The candidate training-class/information-class associations shown on pages 67 and 68 were determined with the aid of a zoom transfer scope. Using that 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). Check with your tutor when finished.

Area	Class Number	Symbol ON Separability	Cluster Number	Class Identity	Number of Pixels
1	1	A	1	Bare Soil	157
	2	B	2	Bare Soil/Hiway	313
	3	C	3	Field Edges,Emerging Crops	340
	4	D	4	Emerging Crop	469
	5	E	5	Pasture	391
	6	F	6	Forest	542
	7	G	7	Forest	649
	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
2	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
	23	W	8	Older Res. (More Trees)	743
	24	X	9	Res. (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
3	32	B2	1	Thin Cloud	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

There are a number of ways to approach the task of selecting the final training classes from the available candidate training classes. The overall goal is to achieve training classes that in total are representative of the information classes present and which are separable from one another. Two divergent philosophies will be described and guidelines for handling difficulties will be presented before you are asked to apply one of or a mixture of those approaches.

The first philosophy or approach could be termed the pooling or "lumping" approach. In that approach candidate training classes whose interclass statistical distances are less than a chosen minimum acceptable value are grouped. Although the identity of each candidate training class is considered when deciding which classes to pool it is not required that identity of the classes pooled be exactly the same. For example, on Figure 21 those class pairs whose transformed divergence is less than 1000 have been connected by a solid line, these between 1000 and 1500 by dashed line. Two pools that might be chosen from that diagram are symbols R, S, 4, E, and U representing forest and 2, D, T, 3, N and F representing agriculture. Notice that although symbol N is labelled forest it might be pooled with agriculture since it is similar to and surrounded by classes labelled agriculture. In such cases it is advisable to recheck the identity of the seemingly anomalous class. In many cases the conclusion may be that such an

COINCIDENT BI-SPECTRAL PLOT (MEAN) FOR CLASS(ES)

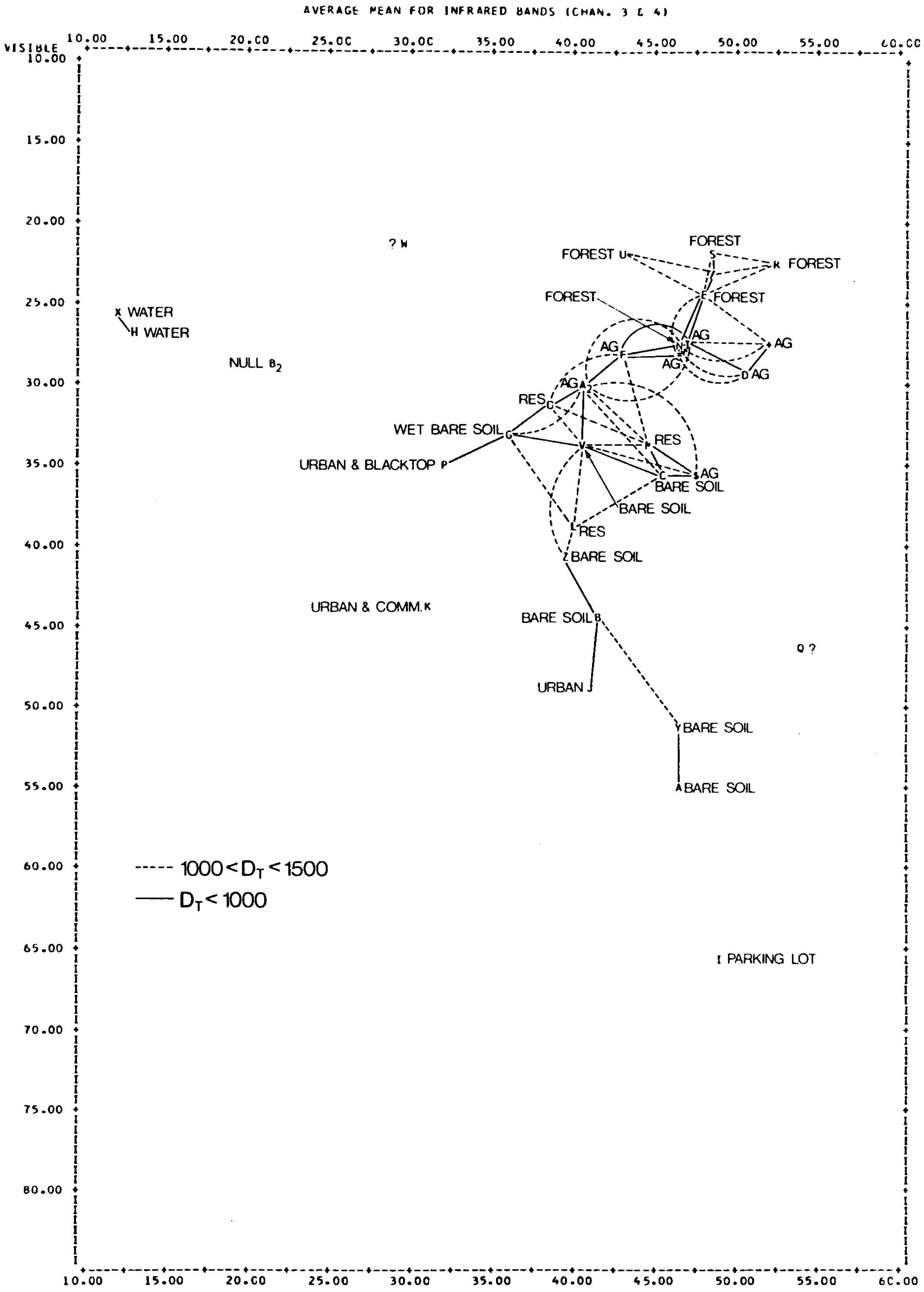


Figure 21. Separability diagram with identities of classes added.

"anomalous" class is really informationally more like the classes it was pooled with than its identity would suggest. To summarize then, the pooling approach tries to form all or nearly all of the candidate training classes into groups of classes connected to one another on the bi-spectral plot. The boundaries of such groups can be established on the basis of the bi-spectral plot alone but often the identities of the classes are used to indicate the portion of the plot (and thus the portion of the multidimensional space) that belongs to each information class.

The second major philosophy or approach to selecting training classes is often called the deleting or "splitting" approach. In the deleting approach groups of candidate training classes or more often individual candidate training classes are chosen in such a way as to completely separate the groups chosen from one another and to minimize the variance within each group. To accomplish those goals classes on the borders between information classes are deleted. Returning to Figure 21 and the previous example, using the deleting approach might result in classes R, S, 4 and U to represent forest and 2, D, T and 3 to represent agriculture. Classes E and N would be deleted. As a result the training classes for agriculture and forest would be more separate and distinct. Carried to extreme the deleting approach leads to larger numbers of more specific (but not necessarily more representative) training classes. For example classes R and S could be selected as a variety of forest and class U as another variety of forest. Classes 4 and E would be deleted.

As a point of comparison it could be said that the pooling approach is more concerned with representation and the deleting approach more with separation. In practice it is necessary to consider both and often analysts will use the two in combination to get the best trade off between representation and separability. For example the following logic can be adopted: because the spectral characteristics of any ground cover type are best described by a cloud of points rather than a single point, when that cloud gets broken into subclasses, by clustering, one or more of the subclasses will be composed of points near the edge of the cloud, which are not really very similar to the points in the center of the cloud that best represent the cover type. Therefore it is likely that such subclasses will be confused with some subclass(es) of a different identity. When this is the case, deletion of such subclasses from the training statistics is reasonable and valid.

Up to this point we have considered relatively straightforward nonproblemmatic examples. Unfortunately, however, almost every analysis has some problems, and rarely is the decision process simple. In the previous example of selecting training classes for forest and agriculture one basic decision to be made was whether or not class E should be retained or deleted. Under the pooling approach it would be retained and grouped with classes R, S, 4 and U on the basis that deleting it would degrade the representativeness of forest. The deleting approach on the other hand would delete the class to enhance the separability of the proposed groups. However, additional information could be brought to the decision process. For example how many pixels are in class E? How firmly is its identity established? How truly Gaussian is its distribution of data values (histograms)? How large are the variances in each channel? The answers to such questions give a measure of the quality of the class. Whereas one could feel justified deleting a class with a small number of pixels, large variances and whose identity was not completely clear, one would feel uneasy about deleting a class with a large number of pixels, tight variances, and unmistakable identity.

Even more challenging situations may arise, depending on the analysis objectives and the data set being used. It may happen that informationally different classes are extensively confused with one another. Again on Figure 21 Classes 1, M, C, V, O, L and G are all interconnected although O, M, and L are identified as residential and 1, C, 5, V, L and G as bare soil and agriculture. Again the first step in such situations is to double check the identities of the classes. Presuming that they are correct some effort could be expended toward refining the classes to enhance their separability. Such refinement would be especially appropriate if several of the classes have very non-gaussian distributions, large variances or were difficult to identify. The refinement could involve reclustering the same candidate training areas, requesting different numbers of clusters or could be pursued by selecting additional training areas and clustering those. The latter would be especially appropriate if the representativeness of the candidate training classes is in question. When problems are encountered some refinement should be attempted. However, one must recognize that if the digital brightness values of pixels from different cover types are very similar no amount of refinement will make them less similar, and therefore, one cannot expect the refinement process to always lead to greater separability of cover types.

When the refinement is not successful the analyst must examine his analysis objectives. Assume that he is interested in classifying an urban area and has discovered that a candidate training class identified as urban is similar to a candidate training class identified as agriculture. He could decide that, for his purposes, the error of classifying some agriculture data points as urban would not be too troublesome, while the error of classifying some urban points into agriculture would be disastrous. In that case, the analyst could choose to eliminate the class identified as agriculture from any subsequent processing.

It is appropriate to say that the process of selecting training classes is the process of deciding how portions of the multidimensional space will be classified. In the previous example the analyst chose to have the portion of the multidimensional space in question classified as urban. Another criterion that could be utilized would be to examine which information class is preponderant in the portion of multidimensional space in question. For example, if the urban class being confused with agriculture in the previous example had been surrounded by other agriculture classes, the decision might have been to reject the urban class and keep the agriculture class.

It should be noted that analysis objectives may lead the analyst to attempt to retain classes that have less than the practical minimum number of pixels. However, one must to think carefully about retaining such a class, primarily because it is unlikely that the statistics describing such classes do in fact accurately represent the informational class of interest. In addition there is a good chance that one component of the statistics describing such classes, the covariance matrix, will be singular. If so, the class cannot be used by the classifier, and eliminating it is unavoidable. A preferable alternative would be to pick some additional candidate training areas likely to obtain more points to represent the informational class of interest. If this is not possible, the results of the classification should be closely analyzed to be sure that the "ill-conditioned" class statistics have not resulted in peculiarities in the results.

One point which should be apparent by now is that there is no single correct way to progress through an analysis sequence. As you increase your understanding of the pattern recognition concepts and gain experience in analysis, you may even develop new procedures yourself.

Use the bi-spectral plot with the class names added (pages 147 and 148 in the printouts) to select training classes. A training class may be composed of one or several candidate training classes. Be sure to keep the analysis objective in mind -- you may want to refer back to the objective on page 12. At least one training class should be selected for each of the cover types present in the area.

Discuss your progress with your tutor. Go as far as you can on your own and then get his advice. Your tutor will want to discuss the process with you in detail when you have finished. For the purposes of combining classes, use a transformed divergence threshold of 1500.

Self-Check

1. What are two desirable characteristics for final training classes?

2. Is the pooling or deleting method of training class selection always preferable?

ACTIVITY 11 MERGING THE STATISTICAL CHARACTERISTICS OF THE TRAINING CLASSES

Upon completion of this activity, you should be able to:

1. Explain why statistics need to be merged at this point in the analysis.
2. Discuss the utility and limitations of statistical distance measures for predicting the result of classifying with any set of classes.

At this point the analyst has available to him the statistical descriptions of candidate training classes which appear on the bi-spectral plot. Before the classification can be performed, statistics must be obtained for the groups of classes decided upon in the last activity. Two of the classification algorithms to be used with the Monroe Reservoir data are based on the assumption that the training classes can be represented by multivariate Gaussian probability density functions, defined by mean vectors and covariance matrices for groups of training samples. The same information was available for each candidate training class. Therefore, the statistical description of each final training class can be calculated by 'pooling' the statistics of the classes grouped in ACTIVITY 10. Those statistics would then be input to the classifier to train it to perform the recognition.

After applying deleting and pooling to the original 52 candidate training classes, final training classes were formed. They are listed on page 172 of the computer printouts, at the bottom of the second bi-spectral plot.

Examine the second bi-spectral plot. Note the coordinates for any classes that fall outside of the scale of the plot. Indicate in the margin of the plot the approximate location of such classes.

After pooling the statistics an analyst will usually check their separability in order to get an indication of the probability of correct classification which would result from using these training classes. 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, the analyst also requested the statistics, 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 class whose statistical distance was 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.

It should be noted that satisfying a certain transformed divergence threshold only insures the degree to which the final training classes are spectrally different and that there is a certain probability of distinguishing among them. It does not insure that they are representative -- that must be done by the analyst. Representativeness is a function of the training areas selected, the accuracy of identification of cluster classes and the number and type of cluster classes that survive the training class selection step. Even when an analyst feels confident in the representativeness, separability, and accuracy of identification of his classes, misclassification can occur. One can only attempt to optimize all three, make the classification and, based upon the results, determine the acceptability of the classification.

In cases of unacceptable results previous steps may be repeated from the acquisition of scanner data through selection of training classes. Although the analysis process appears to be a straight line sequence of steps up to this point, in reality it is a very iterative sequence where the results of each step are closely scrutinized and often repeated before proceeding to the next step. The training class selection step is especially iterative in that as suggested in this activity the acceptability of the training class selections are tested by examining the separability of the newly formed training classes before proceeding to make the classification. Typically several attempts at making the training class selections are required before an acceptable set of training classes is arrived at.

Self-Check

1. Why are statistics needed at this point in the analysis?

2. If transformed divergence values of less than 1500 (or any threshold) between the newly formed training classes exist, must previous steps be repeated or can the analysis continue?

ACTIVITY 12 CLASSIFICATION, RESULTS DISPLAY, AN EVALUATION

Upon completion of this activity, you should be able to:

1. Name and briefly describe the decision rule implemented in the CLASSIFYPOINTS processing function.
2. Briefly describe the ECHO classifier algorithm.
3. Briefly describe the minimum distance classification algorithm.
4. Given an example of a class performance matrix, indicate points correctly classified, errors of omission, and errors of commission for a specified class.

Once a certain confidence in the training classes has been established, a step of special importance in the analysis sequence can be performed - classification. There may be only one classification algorithm available to the analyst or he may be in a situation where he may select one from among several available. For purposes of illustration and comparison, the Monroe Reservoir data set was classified using three different classification algorithms - CLASSIFYPOINTS, ECHO, and MINIMUM DISTANCE.

The CLASSIFYPOINTS processor is based upon the maximum likelihood classification rule. Each pixel to be classified is "compared" to each training class and assigned to the class it most likely belongs to. The concept of a classifier must be defined in a quantitative way so that the computer can do the work. This can be accomplished by means of a set of functions (mathematical expressions) corresponding to the training classes. These functions called <u>discriminant functions</u>, are defined so that when the data (brightness) values belonging to a pixel to be classified are substituted into all of them, the function having the largest value determines the class into which the pixel is to be classified.

The set of discriminant functions for the maximum likelihood classification rule is derived using statistical decision theory in such a way as to minimize the probability of making an erroneous classification (see Swain and Davis, section 3.6 for details). When the classes are assumed to be characterized by multivariate Gaussian density functions, the discriminant functions are defined in terms of the mean vectors and covariance matrices of the classes.

The classification which is produced is typically stored on tape. In order to access and evaluate classification results, another processing function is used. Such a processing function can provide an alpha-numeric printout, and it has a capability for providing quantitative information about a classification in the form of tables. As one means of evaluating the results, the analyst can specify the coordinates of other areas of interest, called "test fields. The size of each test field is such that it contains only one cover type. Since the identity of the pixels in each test field is known along with their addresses in the data a computer can then locate those pixels in the classification result (stored on tape) observe the class into which each pixel was classified, and compare that result with the ground truth verified identity of each pixel to determine whether the pixel was correctly classified. Typically several test fields are selected for evaluating the classification accuracy of each information class. The computer then examines and tabulates the classification decision for each pixel in each test field and prints out a summary by field, by class (all test fields chosen for an information class), or both, as specified by the analyst. An example of tabular results for testing classifier performance is shown in Figure 22. Such a table can be called a test class performance matrix.

	NO OF SAMPLES	PCT. CORRECT	OATS	CORN	WHEAT	SOYB	GRASS
OATS	66	98.5	65	0	0	1	0
CORN	93	93.5	0	87	0	6	0
WHEAT	69	100.0	0	0	69	0	0
SOYB	57	93.0	2	0	0	53	2
GRASS	31	90.3	0	3	0	0	28
TOTAL	316		67	90	69	60	30

Figure 22. Test class performance matrix.

What do the numbers in the performance matrix tell you about the classification? Look first at the 66 samples (pixels) of OATS. The table indicates that 65 of those points, or 98.5%, were correctly classified. Looking across that row, the table also indicates that one data point which the analyst knows to be oats was incorrectly classified as soybeans. That is, there was one error of omission for the 66 oats samples. Looking down for the column labelled OATS, there were two errors of commission for the class oats. That is, two samples were called oats that should not have been.

The five numbers on the major diagonal of the matrix can be summed and that total divided by the total number of samples is called overall performance. For Figure 21, the overall performance is (65 + 87 + 69 + 53+ 28)/316 = 95.1%.

Area estimates for each cover type in a classification can also be obtained in tabular format by examining and recording the identity of each pixel in the area classified. The output in that case will be a one-line table indicating the number of pixels classified into each cover type. The area represented by each data point multiplied by the number of data points per cover type will give the area per cover type. For example if one finds that 19,465 pixels were classified as deicidous forest in the total area classified and by knowing the appropriate conversion factor (1.145 acres/pixel for data geometrically corrected to 1:24,000, line printer aspect) one can calculate the area covered by forest in the study area (22,287 acres in this example).

Often test fields are selected on the basis of what is readily visible in the gray scale images; however, selecting the center of the larger, more homogeneous areas as the basis for testing a classification forces a bias into the test accuracies obtained. Therefore a more statistically valid method has been defined. Using this method one first divides the area classified into blocks either two or three pixels on a side, such that a grid is formed over the data. Then, using a random number table, the desired number of blocks are selected at random. The number of blocks is usually selected such that the blocks selected correspond to a percentage (commonly 5%) of the total data set. Each block is then identified as to cover type by locating those pixels in the reference data; mixed blocks are rejected. Blocks of the same cover type are grouped and used to evaluate the classification accuracy, in the same way the "supervised" test fields were used. Although this method of selecting test fields requires more extensive reference data and personnel resources it is more apt to provide a representative sample on which to base the evaluation.

Two additional guidelines for selecting test fields suggest that test fields should not fall inside of training areas and that they should together contain pixels for testing each cover type in roughly the same proportion as the proportion of those cover types in the scene (the latter is relatively assured be the random selection method). For example, in the Monroe Reservoir area having 90% of your test cells in water would give a biased estimate of the overall classification accuracy.

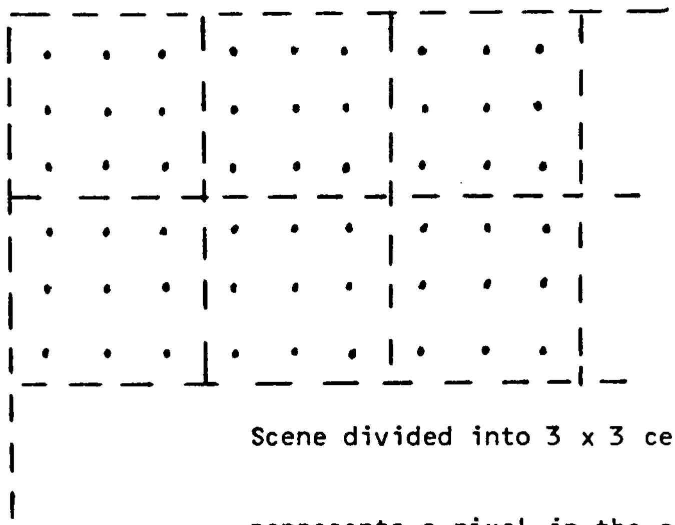
Try to select two test fields on the gray scale maps for each major cover type mentioned in the analysis objective. The fields should be distributed throughout the area. Avoid the bad data line and the training areas. The test fields should be as large as possible but still be "pure."

See your tutor to discuss the results of your selection.

The result map for the CLASSIFYPOINTS maximum likelihood classification is shown on pages 204-227. The control statements on page 203 show the map symbols chosen by the analyst and the groupings that were specified for computing tabular results. This information is presented in a more convenient form on page 204. Note that the same symbol (M) is used for all of the water classes; this is appropriate since the analysis objectives did not call for identifying water subclasses. Grass, soil, and emerging crop are denoted by different symbols on the map, but belong to the same group, agriculture. Thus misclassification errors among those subclasses of agriculture will not be counted in the performance tables. The map is several printout sheets wide. Your tutor can show you a map which has been assembled from these sheets.

Examine the CLASSIFYPOINTS classification map and the associated Test Class Performance table (page 228 of computer printouts). Discuss the performance table with your tutor.

The second classifier used was the ECHO classifier. ECHO is an acronym for Extraction and Classification of Homogenous Objects. ECHO uses spatial as well as spectral information in making classification decisions. When processing a data set, ECHO first divides the area to be classified into rectangular cells. The cell size is chosen by the analyst, usually on the basis of the average object size expected in the scene. Figure 23 illustrates a portion of the data set divided into 3 x 3 cells. The ECHO classifier examines the data vectors within a cell, say the upper left most cell in Figure 23, and performs a statistical test to determine whether or not the data vectors (pixels) within the cell are statistically similar. If they are, they are judged to belong to the same "object" in the scene. ECHO then examines the next cell to the right. If the data vectors making up this cell are statistically similar, a test is performed on the two cells. If they both have similar statistical properties the cells are combined to form an object. This process continues, cells are annexed to the object, until a



Scene divided into 3 x 3 cells. Each dot represents a pixel in the scene.

Figure 23. Data set divided into n x n cells by ECHO. Here n = 3.

cell is encountered which is not statistically similar to the cells comprising the object. This cell is then declared to belong to a different object. Once the objects have been identified all of the data vectors (pixels) comprising the object are classified as a group by comparing the estimated probability distribution of the data vectors in the object to the probability distributions of each of the training classes. In that way all of the pixels in the object are classified at one time as belonging to the most similar training class. If a non-homogeneous cell is encountered (the cell data vectors fail the statistical similarity test), each data vector in the cell is classified individually using the maximum likelihood (CLASSIFYPOINTS) decision rule. ECHO requires the same kind of training class statistics as does CLASSIFYPOINTS, namely, the mean vector and covariance matrix of each training class.

The minimum distance classifier is a "point" classifier like CLASSIFYPOINTS in the sense that each pixel in the scene is classified individually. The classification algorithm is simpler in both concept and implementation than either of the previous two classifiers used on the Monroe Reservoir data set. For each pixel in the scene, the Euclidian distance between the data values belonging to the pixel and the mean vector of each training class is computed. The pixel is assigned to the nearest training class. While computationally more efficient, the minimum distance classifier is inherently less powerful than the maximum likelihood classifier. However, the minimum distance classifier tends to yield approximately the same results as the maximum likelihood classifier when the training classes are generated in a way that leads to a large number of training classes representing the full range of spectral characteristics in the scene. The clustering algorithm used in this example yields classes with these characteristics. Better performance from the maximum likelihood classifier is most noticeable when the classes of interest have very similar or overlapping spectral characteristics.

Your tutor has a copy of the ECHO and MINIMUM DISTANCE classification maps, test class performance data, and a record of the amount of computer time required by each processor. Examine these data and discuss the similarities and differences with your tutor.

Self-Check

1. Name and briefly describe the kind of decision rule implemented in CLASSIFYPOINTS processing function.

2. With the aid of Figure 23 briefly describe how the ECHO classifier works.

3. Contrast the minimum distance classification algorithm with the maximum likelihood algorithm.

4. In Figure 22 indicate the points correctly classified, errors of omission, and errors of commission for soybeans.

ACTIVITY 13 INFORMATION EXTRACTION AND INTERPRETATION

Upon completion of this activity, you should be able to:

1. Name at least two kinds of information that can be extracted from a classification.
2. Give an example of useful information extracted from multispectral classifications in your discipline.

The final and most important step is interpretation of results. The classification results themselves are not usually the real product of interest. Instead, the objective of an analysis is usually to gain information for use in such operations as forest management or land use planning. For instance, the objective usually involves learning where specific cover types are located or what proportion of the area belongs to each cover type, so that management and planning decisions can be made.

To complete the analysis, the original objectives must be reviewed, and the desired information extracted.

Examples of results analysis and the extraction of useful information from multispectral data classifications may be found in several journals, including those listed here:

Remote Sensing of the Environment

IEEE Transactions on Geoscience Electronics

Journal of Soil and Water Conservation

Photogrammetric Engineering and Remote Sensing

Agronomy Journal

Applied Optics

Samples can also be found in a number of LARS Publications, published proceedings of remote sensing conferences, etc.

Study the classification analysis results. Refer back to the analysis objective (page 12). Was the objective met by this analysis? What information can you extract from the results? Based upon the results, would you say that the cover types initially selected were sufficiently distinct spectrally to provide adequate classification accuracy? Why or why not?

Discuss your conclusions with your tutor.

Self-Check

1. Name at least two kinds of information that can be extracted from a classification.

2. Check with your instructor on the availability of the listed references and skim through one or more of them. Then give a example of useful information extracted from multispectral classification.

SUMMARY

Numerical Analysis of Remotely Sensed Data Workshop Series

The activities of this workshop series have taken you through the analysis of a portion of a Landsat satellite scene. Several points should be emphasized with regard to this analysis example.

a) The analysis procedure used in the case study is a typical procedure based upon the experience of LARS researchers. However, it must be emphasized that the workshop has introduced you to an analysis procedure not the ultimate procedure. The analysis procedure used in the workshop was developed by making imaginative and intelligent use of available data processing algorithms. Different applications, perhaps your application, may well require a somewhat different approach.

b) The time constraints of the workshop necessitated a relatively "clean" case study. We did not lead you down any dead ends. Analysts often find that they must repeat portions of the analysis sequence several times before achieving satisfactory results. As an example, if you felt it important to display Highway 37 on the classification map, you would need to go back and define a highway class, select training samples and repeat the classification.

c) The LARSYS software system was used as a vehicle for explaining the steps and operations that go into the analysis of remotely sensed data. In this regard LARSYS may be considered a prototype software system. Other software systems exist and are under development. Basic to all remote sensing software systems are the abilities to display and manipulate data, compute training statistics, select features, carry out supervised and/or unsupervised classifications and summarize and display results.

Summary of Analysis Procedure

- State Analysis Objectives
- Examine data quality and coordinate multispectral scanner (or other remotely sensed) data with reference data.
- Select candidate training areas. Clustering was used to aid in the refinement of these areas.
- With the aid of reference data associate the cluster classes with information classes.
- Augment or refine the candidate training areas as required.
- Compute the statistical distance between the candidate training classes. Analyze these results and refine training class definitions.
- Compute training statistics and classify the data set.
- Evaluate and interpret results.

It is often necessary to iterate portions of the analysis procedure.