LARS Information Note 062874

Guide to Multispectral Data Analysis Using LARSYS

by
John C. Lindenlaub

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

1974

GUIDE TO MULTISPECTRAL DATA ANALYSIS USING LARSYS

by

John C. Lindenlaub

Professor of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

Table of Contents

This work was supported by the National Aeronautics and Space Administration (NASA) under grant number NAS 9-14016 through the Laboratory for Applications of Remote Sensing (LARS).

LARSYS 3.1 Version

PREFACE TO THE STUDENT

Prerequisites

As indicated on the flow chart for the LARSYS Educational Package shown on the next page, this Guide is the last component of the formal instructional sequence. It is assumed that you have mastered the objectives of the previous units.

The analysis of a set of multispectral data can conveniently be broken down into a sequence of steps. By the time you finish studying this guide and the recommended references and working the case study (i.e. carrying out a detailed analysis yourself), you should be able to list the steps in the analysis sequence in their proper order. Furthermore for each step in the analysis you should be able to:

Instructional Objectives

a. give a brief explanation of the significance of the analysis step with respect to the whole analysis sequence,
b. discuss what analytical and/or software tools are available to carry out the analysis step, and
c. apply the analysis principles to a specific problem.

Included in this last objective is the ability to write the control card statements, run the programs and interpret the results of the LARSYS functions used in the analysis sequence.

References

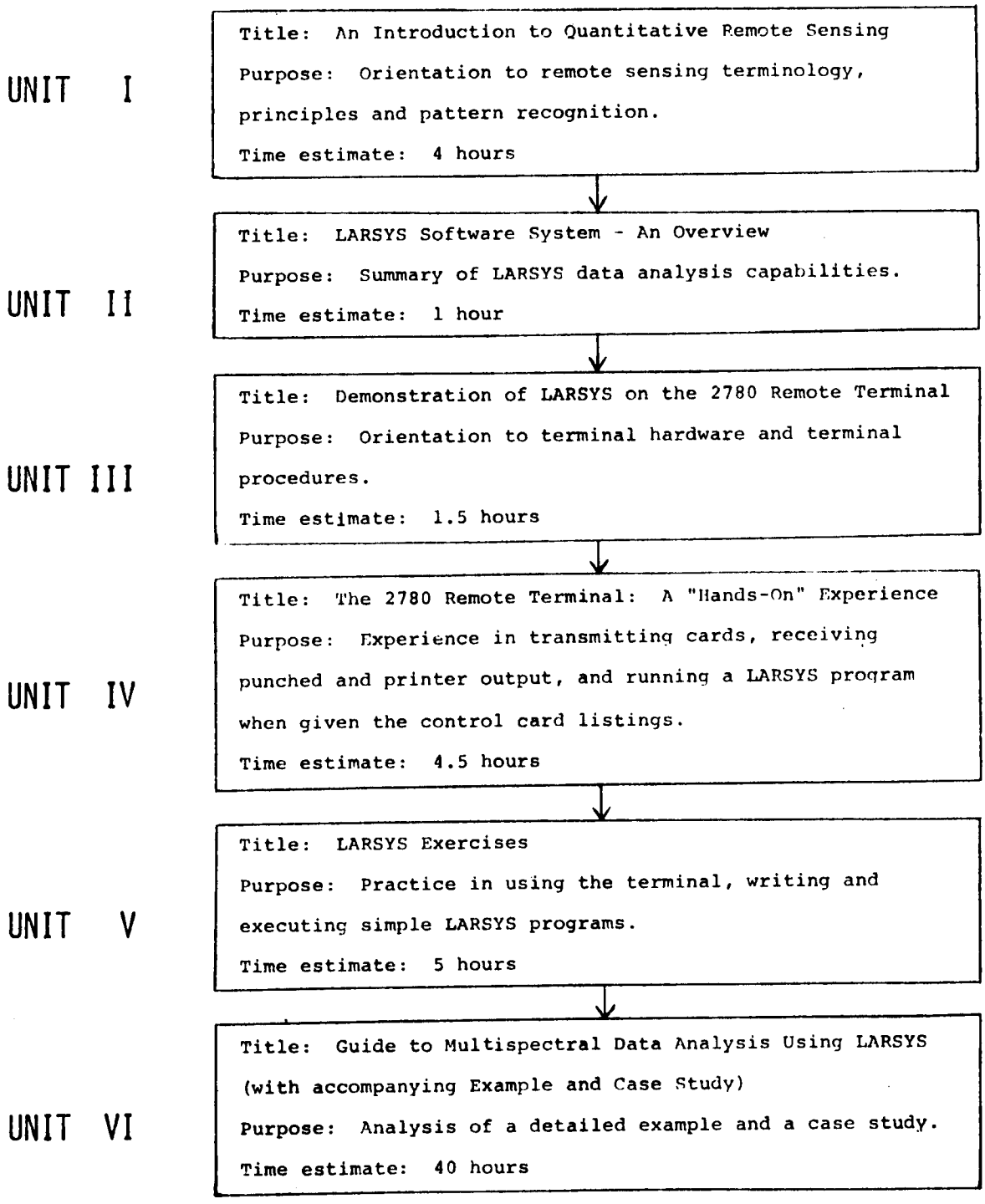
Throughout this guide references will be made to other written materials. The most commonly referenced sources are: LARSYS User's Manual, edited by T. L. Phillips and Pattern Recognition: A Basis for Remote Sensing Data Analysis by P. H. Swain (LARS Information Note 111572).* These references are considered to be part of this unit of instruction.

Student-Instructor Interaction

While this guide 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 use of the LARSYS programs. You will find this to be especially true when you begin working on the

*Subsequent references to this work appear as Swain, 1972.

THE LARSYS EDUCATIONAL PACKAGE



flightline analysis case study. It is recommended that you determine who is available for consultation while you are working on this study.

After the introduction, each section of this guide is divided into three distinct parts: (1) a discussion of the purpose, philosophy and analysis techniques associated with that step in the data analysis sequence plus an example showing computer control cards, computer output and an interpretation of the program results; (2) exercises designed to test your mastery of the section's instructional objectives and (3) problems associated with the case study flightline analysis. It is intended that a person wishing to become adept in the analysis of multispectral data using LARSYS will proceed through this guide, studying the descriptive material (which will explain why each step is important) and the example (which will show how each step is carried out and working both the exercises and the case study (which will provide familiarization and practice with techniques).

In addition to its primary function as an instructional tool, the example portion of this guide will serve as a handy reference for carrying out subsequent analyses.

To assist you in planning your work, the following time estimates are given for each step in the analysis sequence.

Study Time Estimate

Short title	Estimated time required in hours: Text and Example	Exercises	Case Study*
Examination of Data Quality	1	1/2	1 1/2
Coordination with Ground Observation	1/2	1/2	9
Candidate Training Samples	1 1/2	1/2	3
Refinement of Training Samples	2	1	5
Statistical Characterization	1/2	1/2	2
Feature Selection	1	1/2	3
Classification	1	0	3
Results Analysis	1/2	1/2	1 1/2
Total	8	4	28

*Total time, including punched card preparation and on-line execution of LARSYS runs.

INTRODUCTION

Although LARSYS is a rather sophisticated data processing system for analyzing multispectral data, the analysis process is by no means completely automatic. LARSYS provides a facility for machine-assisted data analysis. The quality of the results obtained depends on the difficulty of the analysis problem, the nature and quality of the data being analyzed, and the experience and ability of the analyst.

The procedures described in this guide are based on several years experience gained by a number of remote sensing researchers working with relatively low altitude (3000 to 10,000 feet) data taken with a scanner having a 3-milliradian instantaneous field of view. The resultant ground resolution is of the order of from 9 to 30 feet. The majority of the experience was gained analyzing agricultural regions in the midwestern United States.

The type of analysis described here is known as supervised classification. Data points corresponding to known types of ground cover are used for training samples in the classification algorithm. If it turns out that the several classes of interest are also spectrally distinct, the classification will be "successful." If it turns out that two or more classes are spectrally similar, the classification algorithm will not do a good job of distinguishing between these classes.

A sequel to this guide will present an approach to unsupervised classification. In this approach, the analyst first determines spectrally distinct classes (without regard to the actual ground cover type), performs a classification and then attempts to draw a correspondence between spectrally distinct classes and cover types.

The same set of LARSYS functions is used in both supervised and unsupervised classifications, but the sequence in which the algorithms are used differs. The experience of LARS researchers has shown that the supervised approach is most easily applied to data collected from relatively low altitudes over regularly patterned terrain (agricultural areas) whereas the unsupervised approach is often the best approach for the data collected over terrain which has not been under man's influence.

As a new analyzer of remotely sensed multispectral data, you should be aware that these two different approaches exist. Although this guide concentrates on the supervised approach, the understanding and insights gained should provide a basis from which variations can be made with relative ease.

The steps in the supervised analysis are:

to examine data quality
to coordinate imagery with ground observations
to select candidate training fields and classes
to refine training fields and classes
to obtain statistical characteristics of training samples
to select features
to do the classification
to extract information and analyze the results.

It should be pointed out that during the course of an analysis it is usually necessary to repeat one or more times a number of steps. This will be illustrated later in the analysis example.

Section 1

EXAMINATION OF DATA QUALITY

Instructional Objectives for this Section

By the time you complete the reading of this section, work the exercises and begin the case study you should be able to:

a) state why one needs to examine the quality of the data being considered for analysis.
b) state at least two sources of data quality information.
c) name at least four data idiosyncrasies which might hinder data analysis.
d) use LARSYS processing functions to look for evidence of a gain change in a particular set of data.

Examination of Data Quality

One of the first things a remote sensing data analyst should do is examine the quality of the data to be analyzed. This step serves to determine whether the data is good enough for the analysis to continue or at least gives insights into possible limitations that might result from less than excellent data quality.

A good first source of data quality information is the log book maintained at LARS by the data preprocessing and reformatting group. Typical data log information is shown in figures 1-1 and 1-2. The first figure is a typical log of ERTS satellite data. The second figure is an example of an aircraft data log sheet. Indications of data quality are likely to appear in the "Run Conditions and Comments" portion of the form. Some of the same basic information from the data log is available from the runtable. Other information can be requested from LARS at Purdue.

Examination of the imagery often provides clues to overall data quality. A series of photographs (1-3 to 1-10) shows various kinds of data idiosyncrasies. A reference is also given to a LARSYS run number which shows the same or similar characteristics. You are encouraged to obtain gray scale printouts and/or video displays of these runs in order to observe firsthand the various effects that can degrade the quality of multispectral scanner data.

FORM - 17A

DATA STORAGE TAPE FILE

RUN NUMBER..... 73088100
DATE TAPE GENERATED... NOV 29,1974
TAPE NUMBER..... 2023 FILE... 1
LINES OF DATA..... 2320
SECONDS OF DATA..... 28.41 SEC
AREA E-W 99 NM N-S 99 NM
LINE RATE..... 81.68 LINES/SEC
TIME DATA WAS TAKEN..... 1546 (GMT)
SUN ELEVATION..... 48 DEGREES
SUN AZIMUTH..... 137 DEGREES
REVOLUTION NUMBER..... 5702
DAY SINCE LAUNCH..... 409
SCENE/FRAME ID..... 1409-1546500
FRAME ID..... 190F0G00
STRIP ID..... 0000

FLIGHTLINE ID.....140915465 OHIO
DATE DATA TAKEN..... 9/ 5/73
TIME DATA TAKEN..... 0946 (LST)
PLATFORM ALTITUDE..... 3062000
GROUND HEADING..... 190 DEGREES
FIELD OF VIEW 11.43 DEG 0.1995 RAD
DATA SAMPLES/LINE/CHANNEL..... 3232
SAMPLE RATE 0.0617 MILLIRADIANS
LAT. AT FRAME CENTER..... 40 D 21'N
LONG. AT FRAME CENTER... 083 D 44'W
LAT. AT NADIR..... 40 D 18'N
LONG. AT NADIR..... 083 D 33'W
RUN CENTER.... 83D 44'W/ 40D 21'N
AQUISITION SITE..... GODDARD

SUN CALIBRATION DATA.....
HI GAIN BAND 1.....
LINE LENGTH ADJUST..... *
DIRECT DATA..... *
CALIBRATION WEDGE.....

HI GAIN BAND 2.....
RECORDED DATA.....
COMPRESSED DATA..... *
DECOMPRESSION..... *
CALIBRATION..... *

SPECTRAL BAND LIMITS IN MICROMETERS

CHAN	LOWER	UPPER	CHAN	LOWER	UPPER	CHAN	LOWER	UPPER
(1)	0.50	0.60	(2)	0.60	0.70	(3)	0.70	0.80
(4)	0.80	1.10	(5)	----	----	(6)	----	----
(7)	----	----	(8)	----	----	(9)	----	----
(10)	----	----	(11)	----	----	(12)	----	----

RUN CONDITIONS AND COMMENTS-- LINES 1 - 2340/1. COLUMNS 7 - 3232/1.

Figure 1-1. Log of ERTS data.

031571 LARS - 17

Aircraft Data Storage Tape File

Run Number: 70006802 Flightline Identification: Purdue FL LN 36

Date Tape Generated: July 10, 1972 Date Data Taken: August 13, 1970

Tape Number: 607/1005 Time Data Taken: 1547 hours

File Number: 1/4 Aircraft Altitude: 2000 feet

Lines of Data: 5400* Ground Heading: 225 °

Seconds of Data: 450. Field of View: 1.4714 radians

Miles of Data: 17.3 Data Samples per Channel Per line: 228

Line Rate: 12.00 lines per sec. Sample Rate: 6.6278 milliradians

Spectral Bandwidth in Micrometers:

Chan	Lower	Upper	Chan	Lower	Upper	Chan	Lower	Upper
(1)	.40	.44	(2)	.46	.48	(3)	.50	.52
(4)	.52	.55	(5)	.55	.58	(6)	.58	.62
(7)	.62	.66	(8)	.66	.72	(9)	.72	.80
(10)	.80	1.00	(11)	1.00	1.40	(12)	1.50	1.80
(13)	2.00	2.60	(14)	4.50	5.50	(15)	8.00	14.00
(16)	8.00	14.00	(17)			(18)		
(19)			(20)			(21)		
(22)			(23)			(24)		
(25)			(26)			(27)		
(28)			(29)			(30)		

Data Run Conditions:

Discontinuities in data, lines 1830, 2395, 5103.

*Original data 6512 lines - tape capacity exceeded

Data Tape Comments:

Sampling Rate = .37975 degrees/sample. Thermal IR overlay. Run 70006801 (3 chan thermal IR) was overlayed on 70006800. CO for chan 16 does not exist.

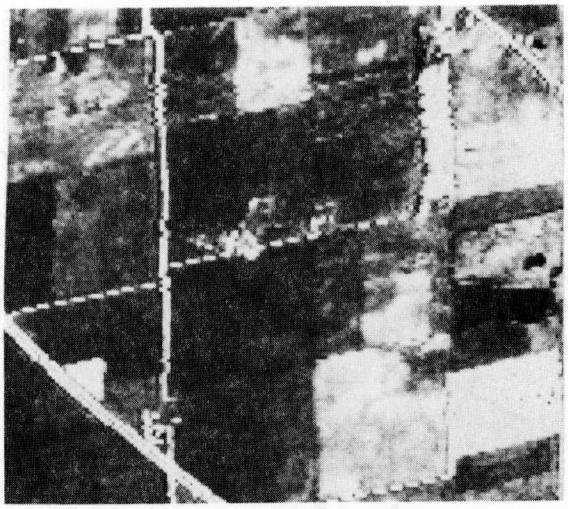
Figure 1-2. Log of aircraft data.

Figure 1-3 shows an example of aircraft data with a geometric distortion known as crabbing or skew. Notice that roads and field boundaries appear to cross vertical roads at an angle of about 10 degrees; however, the accompanying photograph shows these roads to be perpendicular to the direction of flight of the scanner aircraft. The distortion arises in data collected by aircraft-borne scanners when heavy cross winds require the pilot to "angle into the wind" in order to maintain direction along the flightline. A similar phenomenon occurs in ERTS data due to the rotation of the earth. A rectangular image on the ground appears as a nonrectangular parallelogram, the top edge of the image being shifted with respect to the bottom edge by approximately 5% of the height of the image. In addition the ERTS orbit is not oriented due north and south. This results in a rotation of the imagery (about 12° at 40° north latitude). While the rotation may not be pleasing to a person used to working with conventionally oriented maps, the rotation in itself is not distortion in the same sense as the aircraft crabbing effect shown in figure 1-3. Further examples of crabbing may be observed in run 71062701, an aircraft scanner example, and in run 72032804, an ERTS example.

Sun-angle/view-angle effects may cause "shading" in the imagery, resulting from goniometric and/or shadowing effects. Figure 1-4 illustrates a sun/scanner geometry which might lead to shading effects in the scanner imagery. The severity of the sun-angle effect depends upon time of day, time of year, flightline direction and type of ground cover. Figure 1-5 is an example of imagery exhibiting these angle-variable effects. Note that the right side of the picture is much darker than the left side. Run 71062700, especially in channel 6, also exhibits sun-angle effects.

Clouds can also degrade data quality significantly, as shown in figure 1-6, an example of data from a satellite-borne multispectral scanner. Heavy cloud cover can make a particular data set useless for analysis purposes. Additional cloud pattern effects may be observed in runs 72033000 and 72051400. In run 72033000 the segment bounded by lines 376 and 450 and columns 832 and 942 is particularly interesting to look at.

Occasionally you will encounter a noisy image which may be the result of a noisy detector, a noisy data channel, a telemetry problem or some combination of effect. Figure 1-7 is a series of images, each progressively noisier. (These images were produced by adding noise artifically to a good quality data set. For further details see LARS Information Note 102670, Random Noise in Multispectral Classification by Steve Whitsitt.)



Multispectral Image with Crabbing

Air Photo

Figure 1-3. Comparison of a multispectral image where crabbing is present, and a photograph of same area.

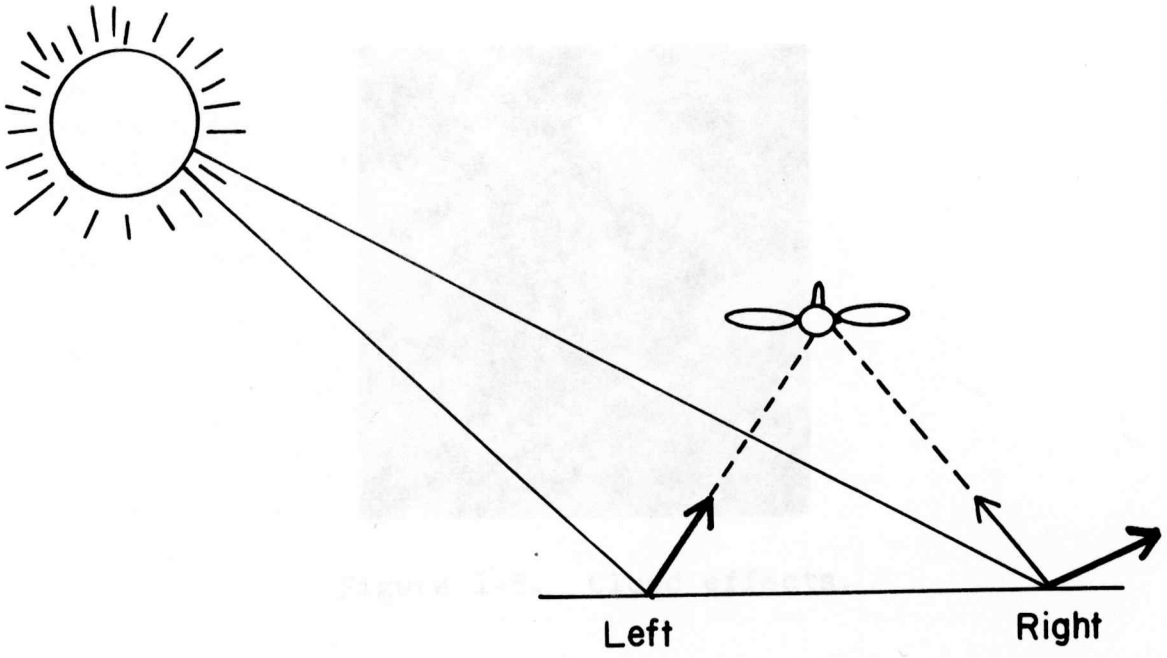


Figure 1-4. Bidirectional reflectance geometry for aircraft scanner - a cause of shading effects.

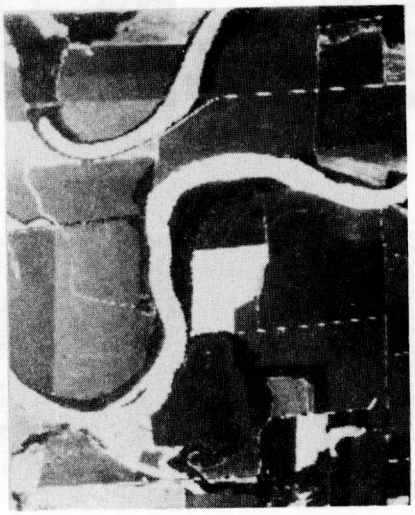


Figure 1-5. Sun angle effects on imagery.

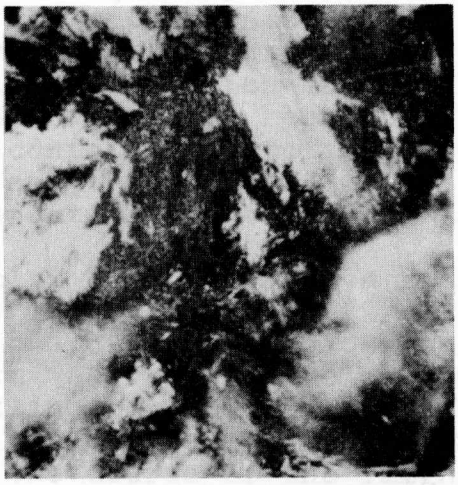
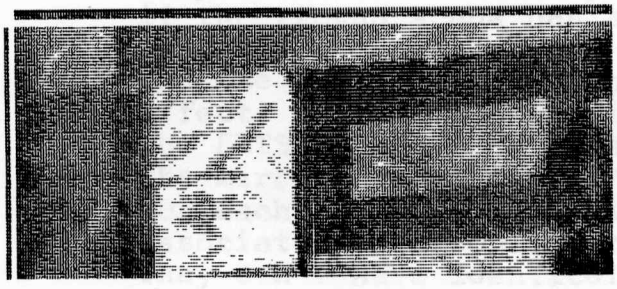
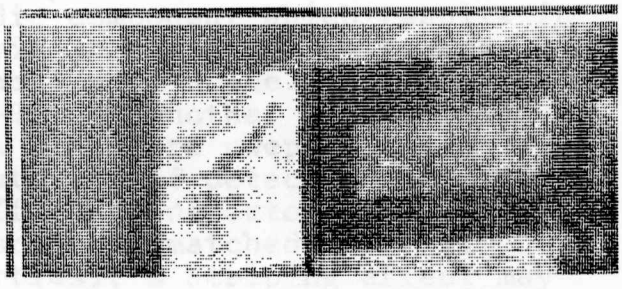


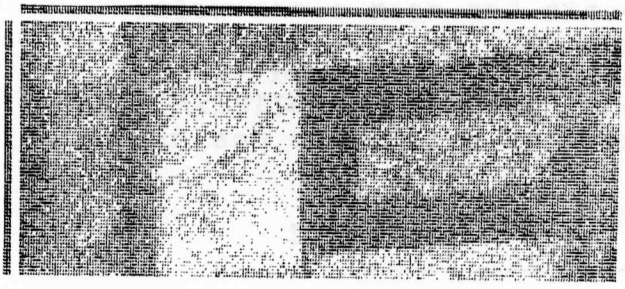
Figure 1-6. Cloud effects.



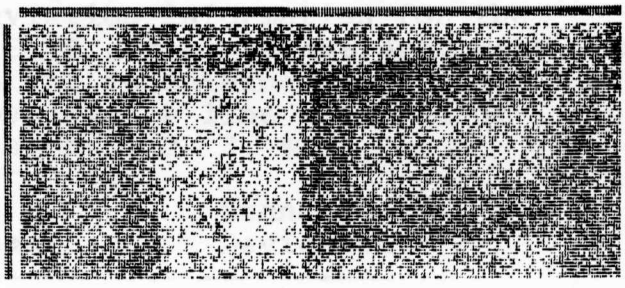
no noise



sigma = 2



sigma = 10



sigma = 20

Figure 1-7. Effect of noise on imagery.
(Sigma is a measure of the amount of noise added to the original data set.)

The LARSYS system of programs provides other opportunities to examine data quality. In the airborne scanner system the rotation of the mirror enables the detectors to look inside the aircraft during part of the cycle. This opportunity is used to provide calibration data. The format of the Multispectral Image Storage Tapes requires the last 6 data values of each line to contain calibration information. By using the COLUMNGRAPH and TRANSFERDATA processing functions, you can examine this data and obtain some information on data quality.

As an example, figure 1-8 is a graph of the calibration lamp output (C1) for channel 11 of run 70003600. This particular run was one of those used during the 1971 Corn Blight Watch Experiment. A graph of the calibration lamp output for the whole flightline provides a mechanism for determining whether a gain change was introduced in any of the recording channels during the course of the flight. Also, the variance of the C0, or dark reference sample, provides a measure of the noiseiness of the data. An exercise is given later for examining data quality by looking at calibration values.

Striping of multispectral scanner imagery can arise from many sources. For example, Moiré patterns occur in the data if the ground scene has a periodic component which results in a beat frequency between the periodic sweep of the scanner and the periodic component in the ground scene. Moiré pattern effects are visible in figure 1-9. As another example, in the ERTS scanner system six scan lines are swept out 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 (i.e. if they don't have identical properties), a striping effect may be noticeable in the imagery. A dramatic example may be seen by examining Channel 1 of run 72044401, an ERTS frame of the Lafayette, Indiana area (figure 1-10). The table below shows mean and standard deviation information for the output of each of the channel 1 ERTS detectors averaged over the whole frame. Such information might be obtained from the STATISTICS processing function by using a line interval of six and successive starting lines of 1, 2, 3, 4, 5, 6.

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

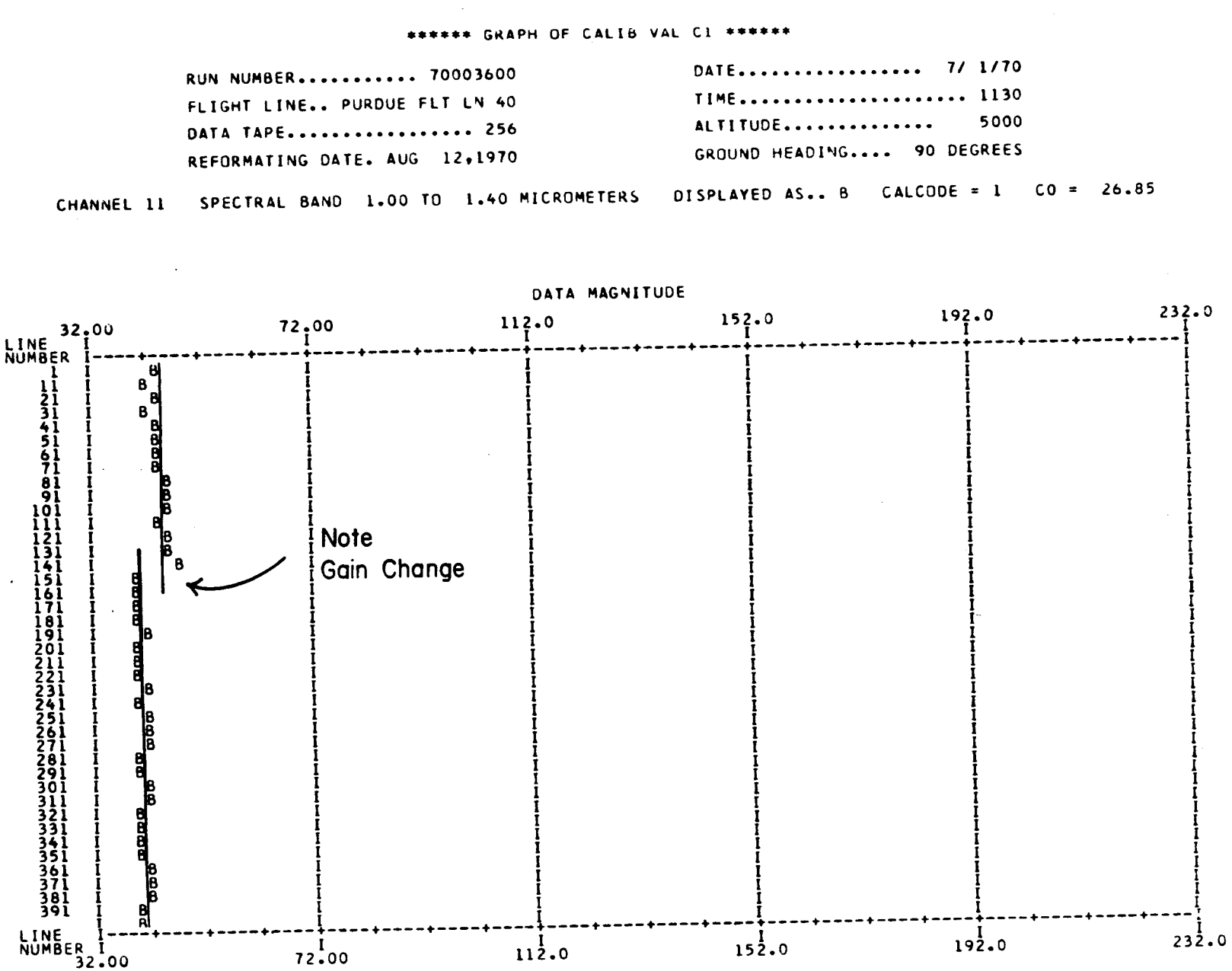


Figure 1-8. Graph of calibration lamp output showing a gain change near line 141.

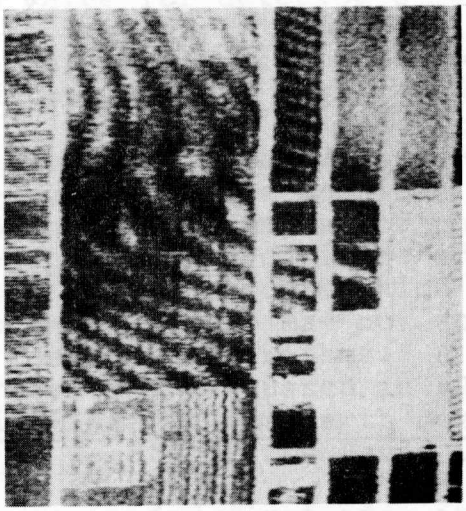


Figure 1-9. Moiré patterns.

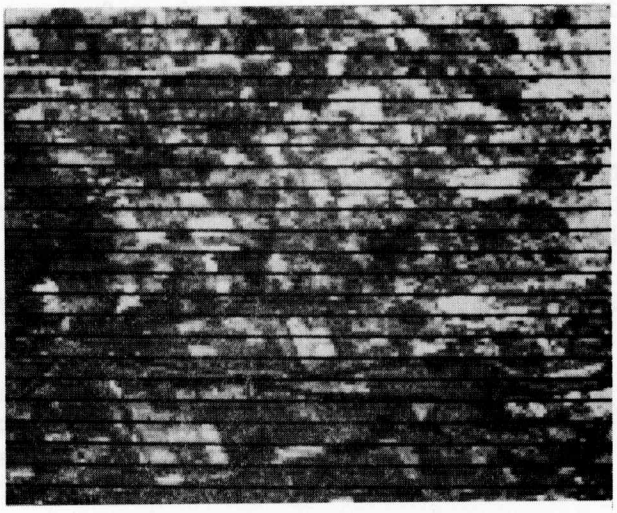


Figure 1-10. Striping effect in imagery.

Notice the mean value for detector 3 is very low compared to that of the other detectors. Apparently a malfunction occurred in the detector electronics which resulted in the striping shown in figure 1-10.

Several examples of data idiosyncrasies have been illustrated to alert you to these possible degradations in data quality. Although these examples of poor quality data have been illustrated by showing an image of the data, it is important to point out that data which might appear to be of poor quality to the observer may not appear to be of poor quality to the computer algorithms. A dramatic example of this is illustrated and discussed in LARS Information Note 062273, Analysis Research for Earth Resource Information Systems: Where Do We Stand? by David Landgrebe. (see page 3)

A highly recommended first step in the analysis of multispectral data is to examine the data to get a general evaluation of its overall quality. LARSYS processing functions were used to produce the illustrations of data idiosyncrasies shown above. The example and exercises that follow will give you an opportunity to use LARSYS processing functions to examine data quality.

References to LARSYS User's Manual

a) Section 4 (volume 1) of the LARSYS User's Manual, pages 4-1 to 4-3, gives a general description of LARSYS Control Commands. The remaining pages in Section 4 describe the individual Control Commands in detail. It is suggested you review the REFERENCE RUNTABLE Control Command.

b) Section 6 (volume 2), pages 6-1 to 6-3, gives a general description of LARSYS Processing Functions. The remaining pages in Section 6 describe the individual Processing Functions in detail. It is suggested you review the IDPRINT, COLUMNGRAPH and PICTUREPRINT Processing Functions. In particular note the last paragraph on page PIC-7.

c) Read pages IV-1 to IV-4 of Appendix IV (volume 3) to familiarize yourself with the format of Multispectral Image Storage Tapes. Note especially that the last six data values on each line represent respectively:

C0 calibration value
Variance of C0 calibration value
C1 calibration value
Variance of C1 calibration value
C2 calibration value
Variance of C2 calibration value.

Example

The examples given in conjunction with each step of the analysis include representative control card listings, computer printouts and interpetations drawn form an anlysis of flightline C1, run 66000652.

To examine data quality and obtain an overall impression of the data to be analyzed, the analyst requested the ID record of the run. The listing includes identifying information about the run (run number, flight line number, date recorded, etc.) as well as a table of the spectral bands and calibration values for all channels recorded on the tape.

The ID record printout gives the number of lines in the run. This information can also be obtained using the REFERENCE RUNTABLE Control Command. The analyst typed at the terminal:

```
reference runtable 66000652
```

and the computer typed back:

RUN NO.	TAPE	FILE	LINES	CHAN	SAMP	FLIGHTLINE ID
66000652	1001	1	950	12	228	PURDUE FLT LN C1

The number of lines of data is 950 and the number of columns (or samples per line) is 228. The last six samples contain calibration information, leaving 222 data samples per line.

If the analyst wished to check calibration for gain changes he would use the following LARSYS run, plotting, say, two channels at a time:

```
*COLUMNGRAPH
PRINT RUN(66000652), LINE(1,950,10), C1
CHANNELS 1,12
END
```

To check all twelve channels he could run this six times specifying different pairs of channels each time.

The analyst then wanted to obtain gray scale printouts in those channels that would give him the greatest distinction between fields so he could outline boundaries. To do this he requested a sample printout from each channel. The cards needed were:

```
*PICTUREPRINT
DISPLAY RUN(66000652), LINES(200, 500, 2), COL(1,222,2)
CHANNELS 1,2,3,4,5,6,7,8,9,10,11,12
BLOCK LINE(200,500,4), COL(1,222,4)
END
```

On the basis of the sample output, the analyst decided which channels gave best distinction to field boundaries. He then acquired a complete run from those channels.

```
*PICTUREPRINT
DISPLAY RUN(66000652), LINE(1,950,2), COL(1,222,2)
CHANNELS 9,11
END
```

EXERCISES

1. Explain in your own words why it is important for the analyst to examine the data quality before undertaking any extensive analysis.
2. Name at least two techniques which are available to the analyst of remote sensing data for examining data quality.
3. Name at least four types of data idiosyncrasies the analyst may find.

FLIGHTLINE ANALYSIS CASE STUDY

As you progress through this guide you will be asked to carry out an analysis of Segment 210, Mission 43M of the 1971 Corn Blight Watch Experiment (run 71053900). We begin by examining the data quality of the run.*

1. Use LARSYS to graph calibration parameter C1 for run 71053900. Do any of the channels show noticeable gain changes?
2. The case study involves the analysis of only part of the available data, lines 200 through 1055. Obtain gray scale printouts for the .61-.70, 1.0-1.4, and 2.0-2.6 μm channels. These gray scale printouts will be used in the next step of the analysis.

---

*A copy of this run has been made for your remote terminal site. Consult your instructor for the proper tape and file number.

Section 2

COORDINATION OF IMAGERY AND GROUND OBSERVATIONS

Instructional Objectives for this Section

By the time you read the text material, work the exercises and complete the next step in the case study you should be able to:

a) give reasons for the necessity of ground observations and for correlating multispectral imagery with ground observations.
b) list at least two sources or techniques for obtaining ground observations.
c) correlate the location of ground features apparent on the multispectral imagery with those on an aerial photograph of the same area.

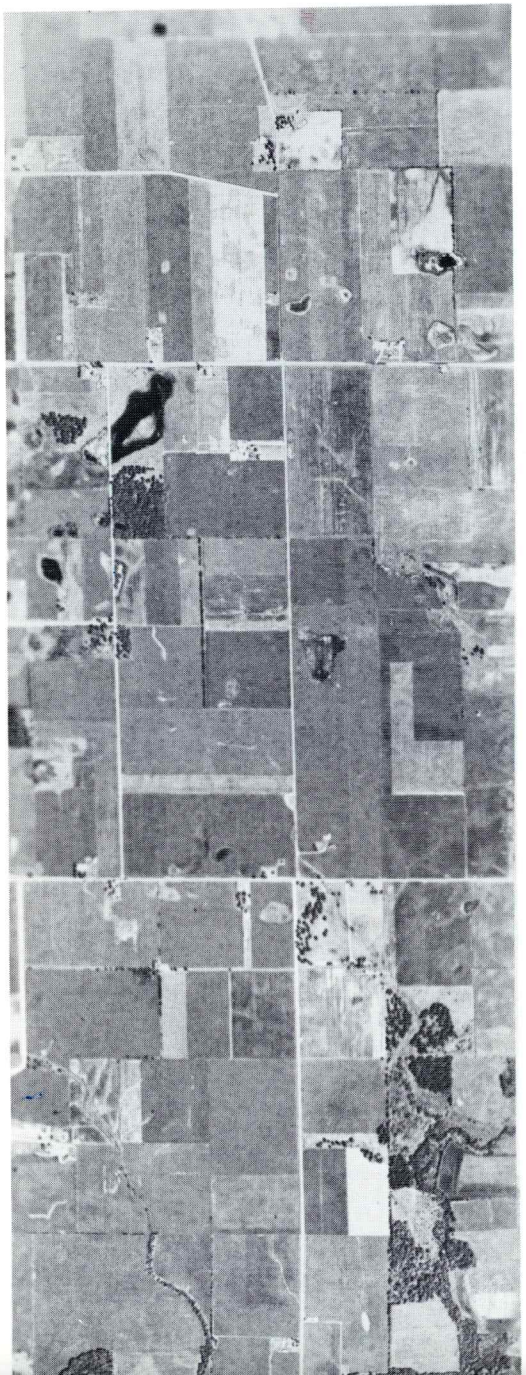
Coordination of Imagery and Ground Observations

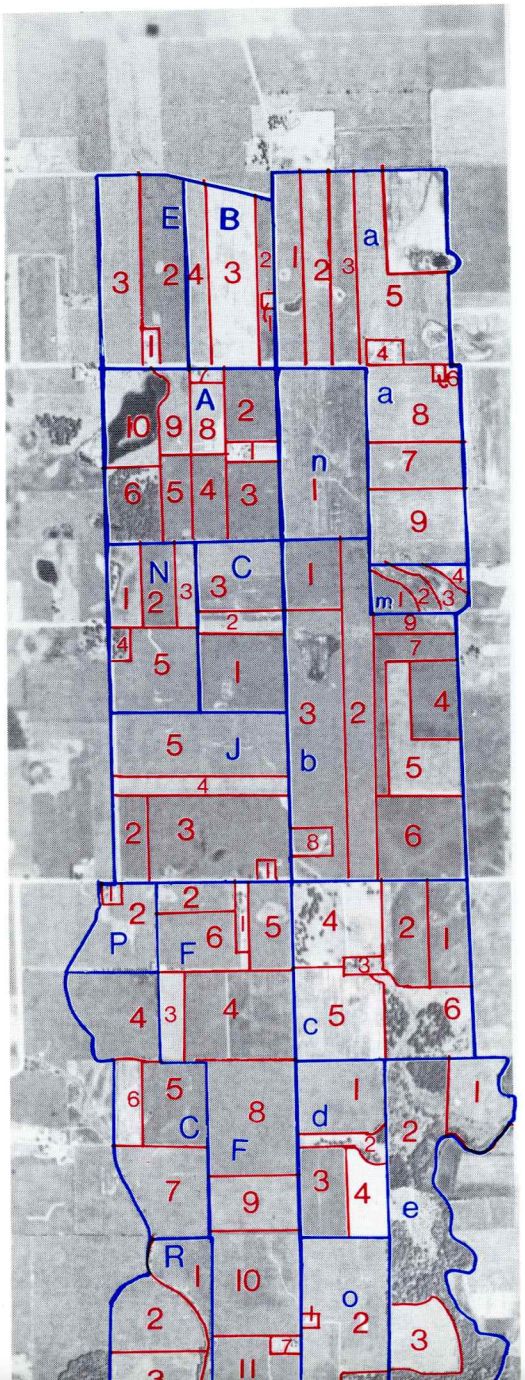
It is necessary to coordinate multispectral imagery with known features on the ground in order to determine the row and column coordinates of training data. (The need for training data is discussed in the next analysis step.) Sources of ground observations include on-site visits, interpreted aerial photographs and maps. The importance of ground observations is discussed in LARS Print 120371, The Importance of Ground Truth Data in Remote Sensing, by R. M. Hoffer. This information note should be read at this time.

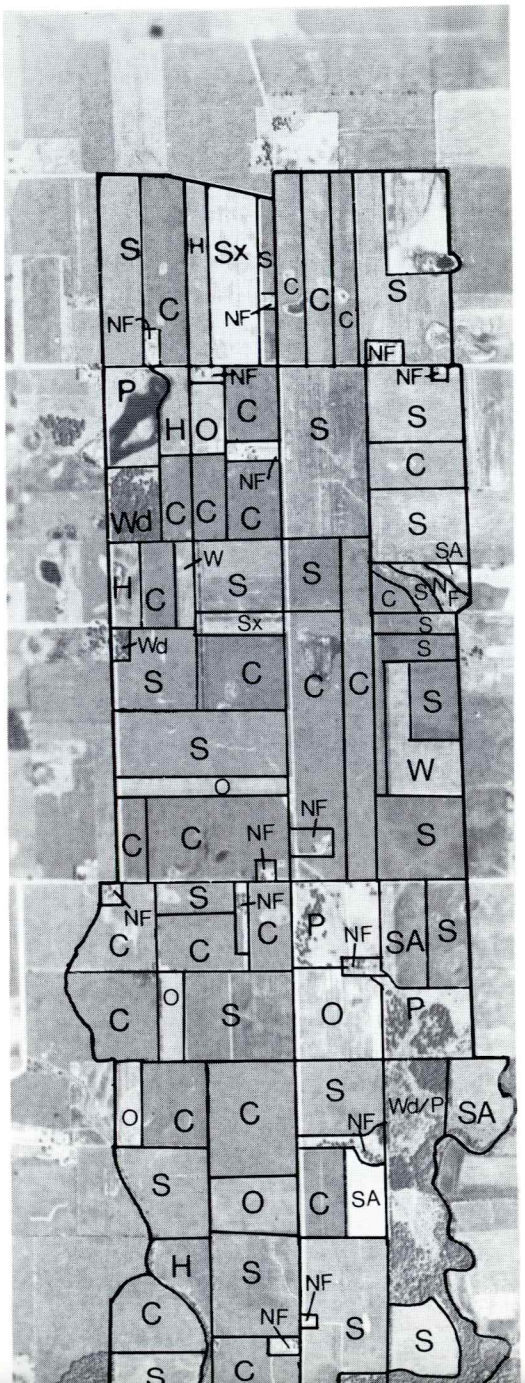
When you are dealing with agricultural data, the annotated aerial photograph provides a convenient method for correlating multispectral imagery with ground observations. Figure 2-1 shows aerial photographs of Segment 210 of the 1971 Corn Blight Watch Experiment. Field and tract information has been superimposed on the middle photograph. Each tract, i.e., land under the control of one operator or owner, is outlined in blue, and fields within the tracts are outlined in red. Tracts are designated by letters (upper and lower case) and fields within a tract are designated by numbers. Thus a tract-letter/field-number combination uniquely identifies each field. For example, field U4 is near the bottom of the flightline; field E2 is near the top.

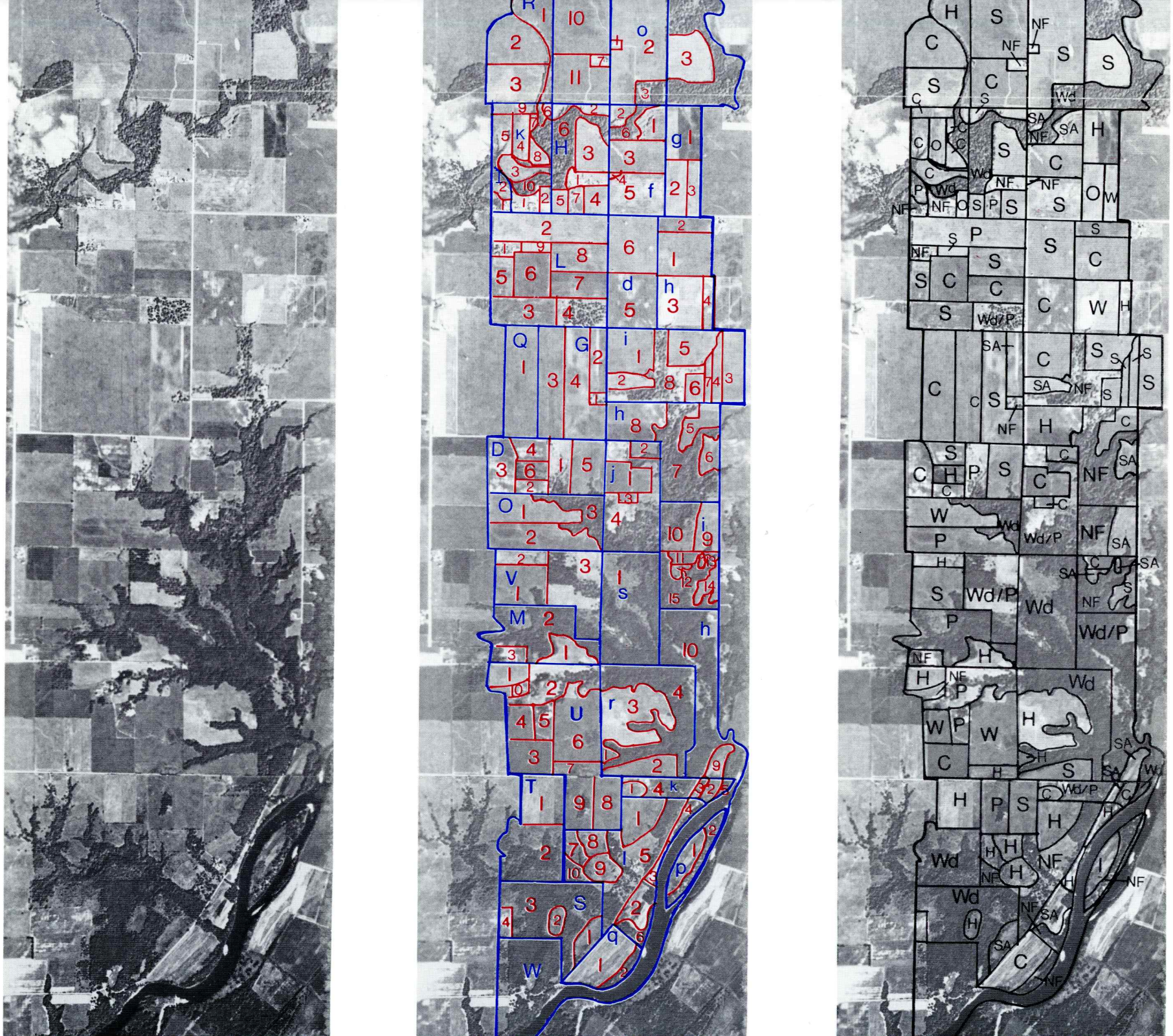
The right-hand photo column has ground cover information superimposed on the fields. The key to the ground cover annotations is given in the figure caption. By comparing the center and right-hand photos, we can determine that field E2 is a corn field while U4 is a wheat field. The ground observation information contained on these photographs was obtained by on-site visits.

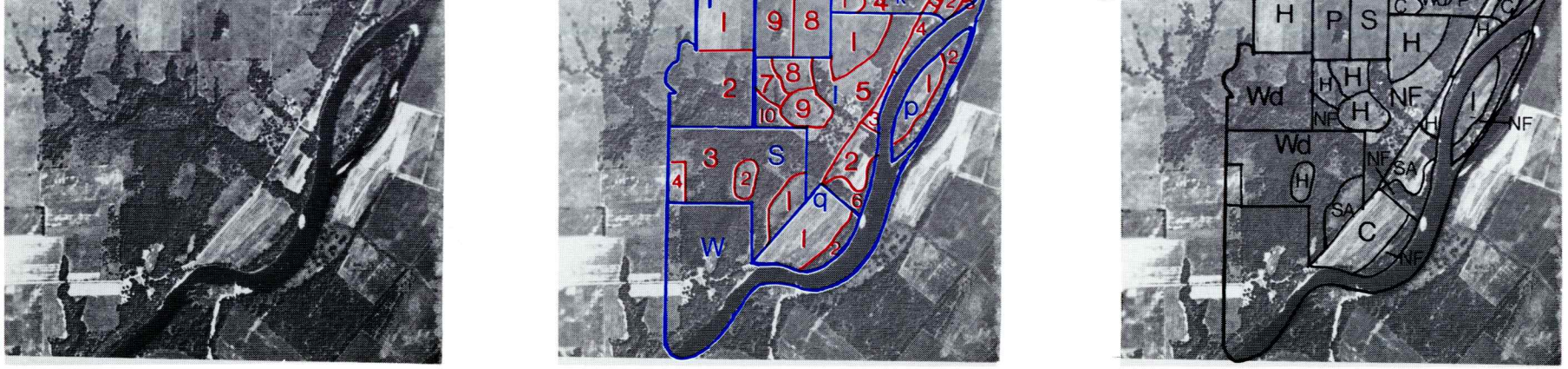
Figure 2-1 Color foldout showing aerial photographs of segment 210 of the 1971 Corn Blight Watch Experiment follows.











Aerial Photograph and Ground Observations for
Agricultural Area in Indiana (Seg. 210 - 1971 CBWE).

C - Corn	Wd - Woods
S - Soybeans	W - Wheat
H - Hay	P - Pasture
O - Oats	Sx - Sudex
NF- Non-Farm	SA - Set-Aside
I - Idle	

By working with multispectral imagery, an aerial photograph, and ground observations, the analyst can correlate points on the multispectral image with corresponding ground observation points. Comparison of the two images helps the researcher locate specific field boundaries. When you are working with agricultural and other man-made scenes, it is often useful to outline with a colored pen as many roads, field boundaries and other recognizable features as possible on a gray scale printout of the area. Experience will show you that it is useful to have printouts of several channels available; features that don't show up well in one channel may show up better in another.

Example

Continuing with the example analysis of run 66000652, the analyst used an annotated photograph of the area to draw in field boundaries on a gray scale printout. A portion of this printout is shown in Figure 2-2. The solid lines denote field boundaries, and the letters denote the type of ground cover within each field. The significance of the "candidate training samples" will be explained in the next section.

EXERCISES

1. State in your own words the necessity for ground observations and for correlating multispectral imagery with ground observations.
2. State at least two sources or techniques for obtaining ground observations.

FLIGHTLINE ANALYSIS CASE STUDY

With the aid of the annotated photograph of segment 210 of the 1971 Corn Blight Watch Experiment (Figure 2-1), outline and annotate on a gray scale printout all of the fields lying between lines 200 and 1055. You may do this directly on the gray scale printouts you obtained earlier or you may generate new printouts. You may find it desirable to use double-width printouts, i.e., every line and column, and tape the two halves together.

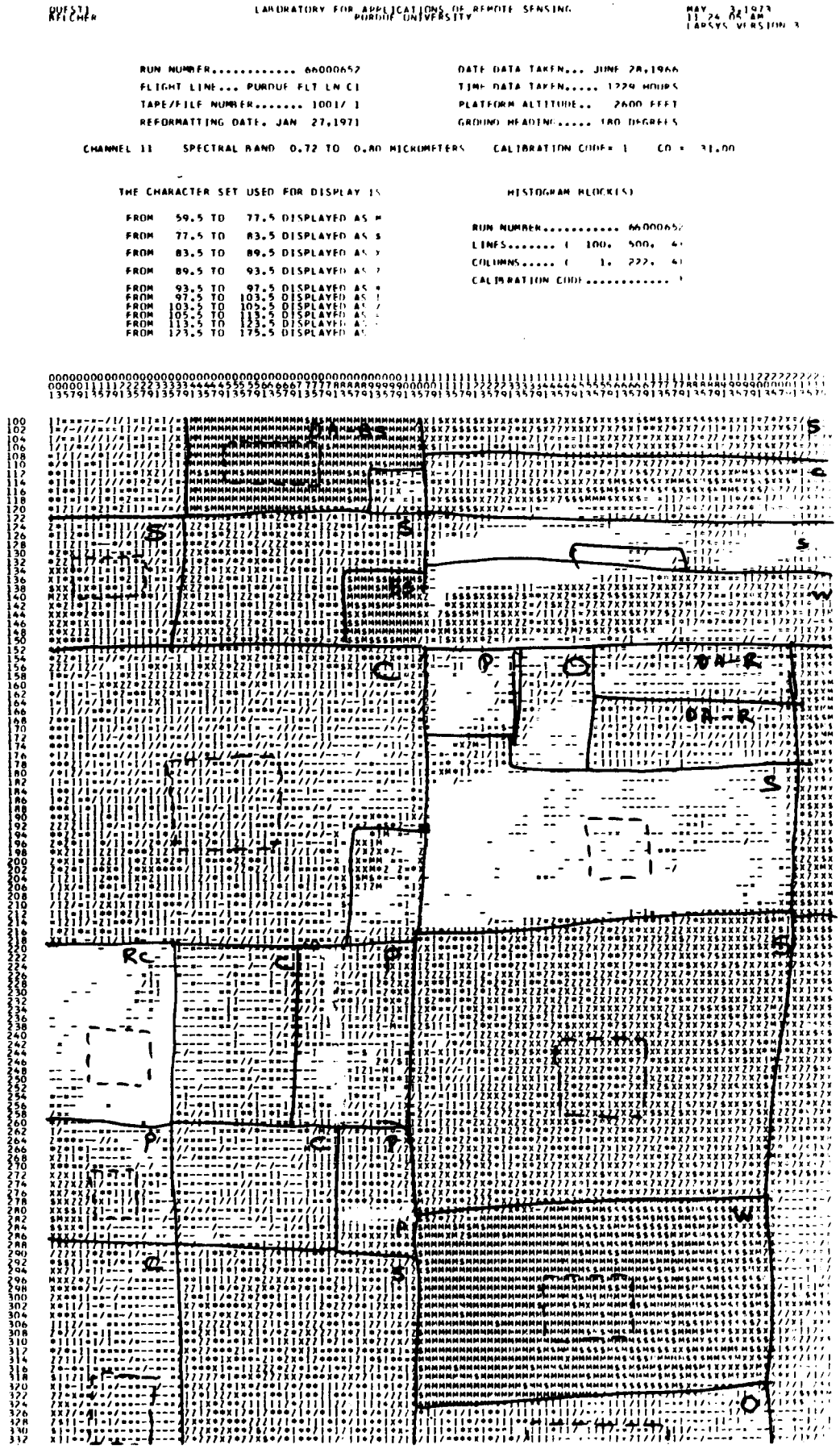


Figure 2-2. Example of gray scale printout with field boundaries (solid lines) and candidate training samples (dotted lines) outlined.

Section 3

SELECTION OF CANDIDATE TRAINING SAMPLES

Instructional Objectives for this Section

By the time you complete reading the text, consulting the references, studying the examples of this section and completing the next step in the case study analysis, you should be able to:

a) state in your own words why it is necessary to select training classes and training fields.
b) name at least two considerations that might go into the selection of training classes.
c) determine the practical lower limit for the number of training samples needed per class for a given set of multispectral data.
d) describe the use of test fields as distinct from training fields.
e) actually carry out the process of selecting training classes and specifying, by means of Field Description Cards, training fields and test fields for each class.

Selection of Candidate Training Samples

The next step in the analysis of multispectral data is the selection of candidate training samples. We shall begin by explaining what training samples are and why they are needed.

The basis of remote sensing data analysis using LARSYS is pattern recognition (Swain, 1972). The pattern recognition algorithms 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. Later, when the classification operation is being carried out by the pattern recognition algorithms, each data point (or group of data points in the case of sample classification) is "compared" to the training samples, and the point (or group of points) is assigned to the "most likely" or most similar class. The mathematical basis for pattern recognition and the classification algorithms have been discussed in detail by Swain (Swain, 1972).

We speak of candidate training samples because experience has shown that it is wise to examine one's first choice of training samples in detail to see if they truly appear to be representative of the desired class. As an example, an analyst using ground observations (such as the annotated photograph you used earlier) might choose a particular corn field as a training field. It may be that early spring flooding of one corner of the field has resulted in data points from this area being distinctly different from those in the rest of the field. These points should be discarded since they are not representative of the class corn.

There are two aspects of this step in the analysis: the selection of training classes and the selection of training samples (sets of data points) representing each class. In general there is an underlying reason for wanting to classify the data into certain classes. The reason may reflect an economic interest, a scientific inquiry, or a feasibility study. The important point is that at the outset one often can not be sure whether the classes of interest are distinguishable, i.e., whether they are "spectrally" distinct. The degree to which you can actually separate the classes you are interested in will not be known until much further along in the analysis. It may be necessary later to redefine the classes and to repeat the analysis steps.

When selecting training classes one should draw on one's background and experience. For instance, an agronomist will know that corn and soybeans are both row crops. He might suspect that early in the spring, when there is a good deal of bare soil visible, the two ground covers might be difficult to distincuigh spectrally. On the other hand, later in the growing season when the corn has tasseled, a spectral difference would be expected. A good understanding of the interaction between solar energy and matter can also be helpful in selecting training classes. For instance, Figure 3-1 shows the reflectance properties of bare soil, green vegetation and water. From the differences shown in these spectral signatures, one would strongly suspect that the separation of earth surface types into these three classes could be done rather successfully.

The method used to choose training classes involves gathering together information on mission objectives, ground observations and multispectral imagery. Then, based on the results of the previous analysis step (coordination of ground observations with imagery) and past knowledge and experience, candidate training classes are designated.

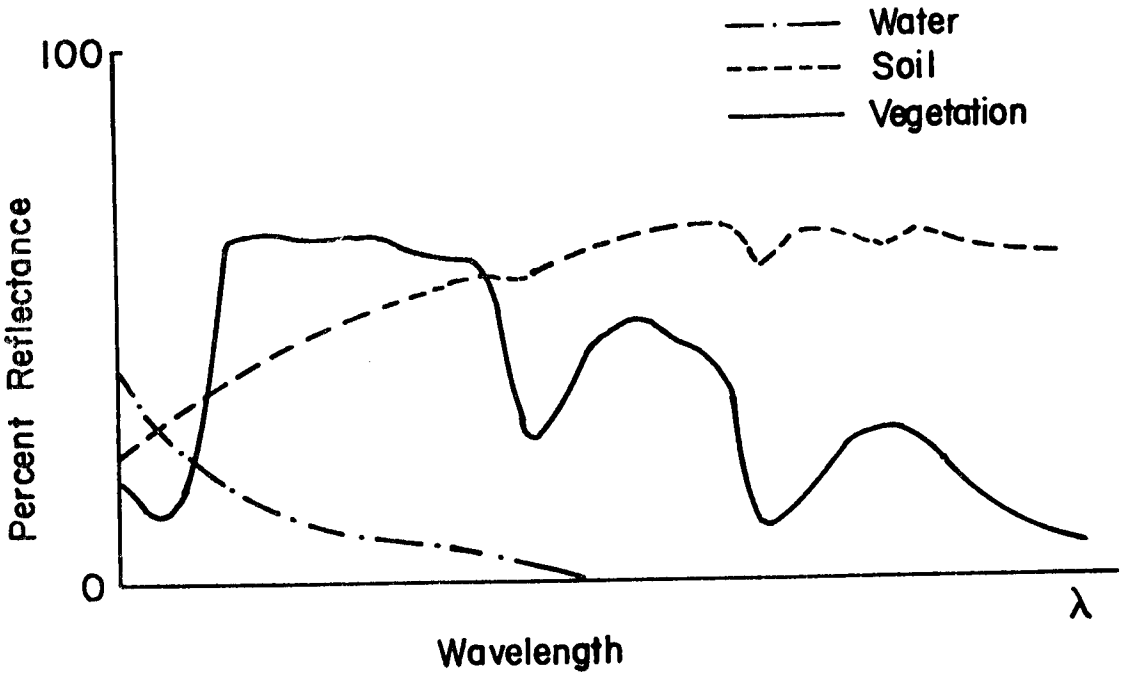


Figure 3-1. Typical reflectance properties of bare soil, green vegetation and water.

Once candidate training classes are selected, the next step is to select training samples representative of each class and to specify these to the computer by means of Field Description Cards. The key word here is representative. The aim is to select training samples which are representative in that they must effectively tell the classification algorithm what typical members of the class "look like." Without a good description of the classes the classification program cannot be expected to do a very good job of classifying.

How can you be sure samples are representative? There is no single answer to this question, but there are some techniques which have proved useful. If the physical size of the field from which you are selecting data points is large enough, it is a good idea to stay away from the physical boundaries of the field. Figure 3-2 shows a portion of a gray scale printout with the physical field boundaries drawn in with a pen. Well within these physical boundaries are rectangular areas outlined by dotted lines. The points contained within these dotted lines were used as training samples. The reason for avoiding the field edges is that these regions may be non-typical due to fence lines, ditches, access roads, etc. If the scale of the imagery is such that the physical fields contain only a few resolution elements, it may be difficult to take this precaution. Training field areas are identified to the LARSYS processing functions by the beginning and ending line and column numbers. The results is that training fields are rectangular and oriented in the direction of the flight path. If the natural field boundaries are not rectangular or if they have a different orientation, it may be necessary to define the desired training area by a number of small rectangular fields.

If ground observations are available over much of the flightline, a reasonable approach is to "scatter" training fields somewhat uniformly over the flightline.* This scattering would tend to minimize any effects caused by changes in geography, agricultural practices or climatic conditions. Figure 3-3 shows some examples of how training fields might be selected.

How many data points are needed for training? Before giving a direct answer to that question we'll go into a little more depth on how the training samples are used by the classification algorithm. The algorithm is based on the assumption

---

*An exception to this rule would be if the objective of the analysis were to determine the extent to which training samples chosen from one area could be used to classify data from another area.

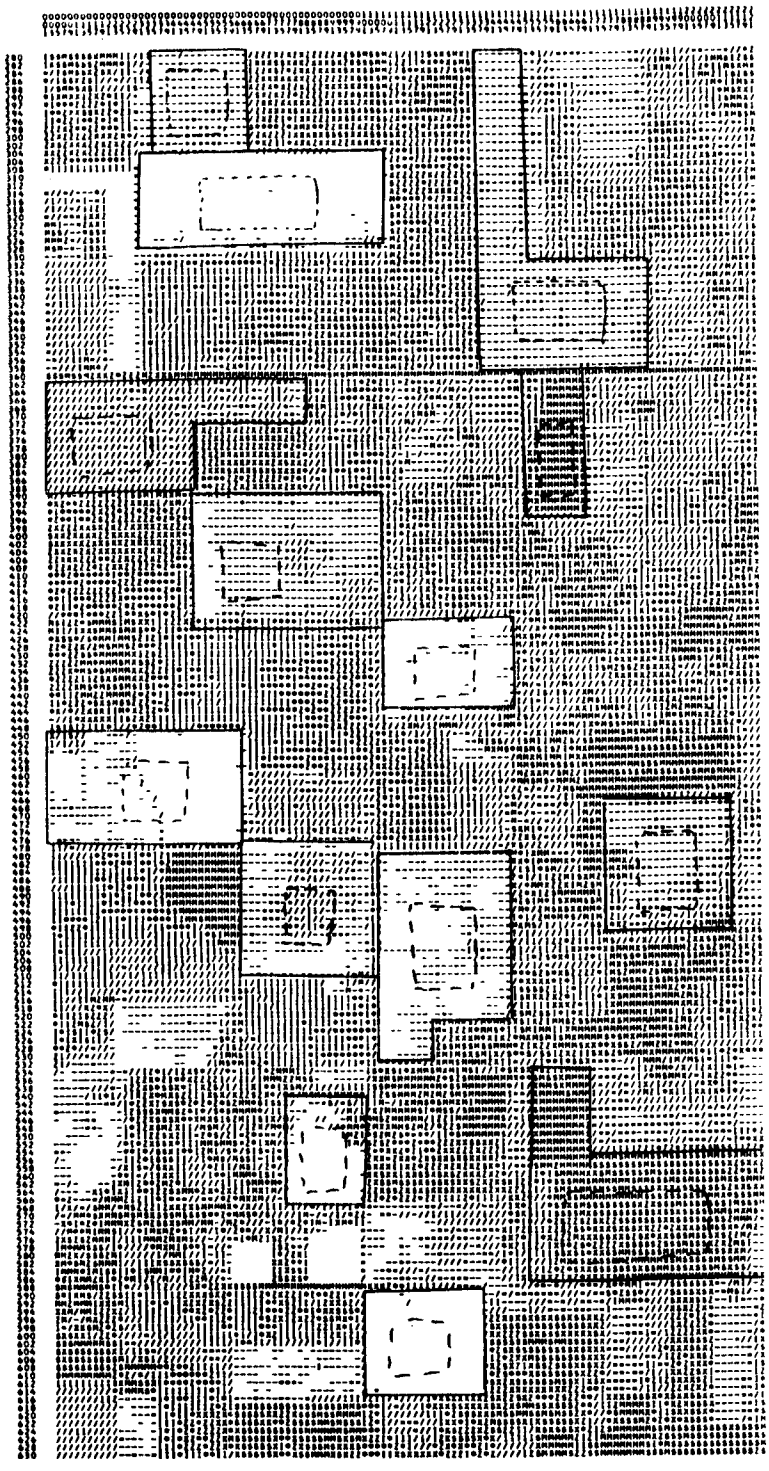


Figure 3-2. Portion of gray scale printout with some physical fields (solid lines) and training fields (dotted lines) outlined.

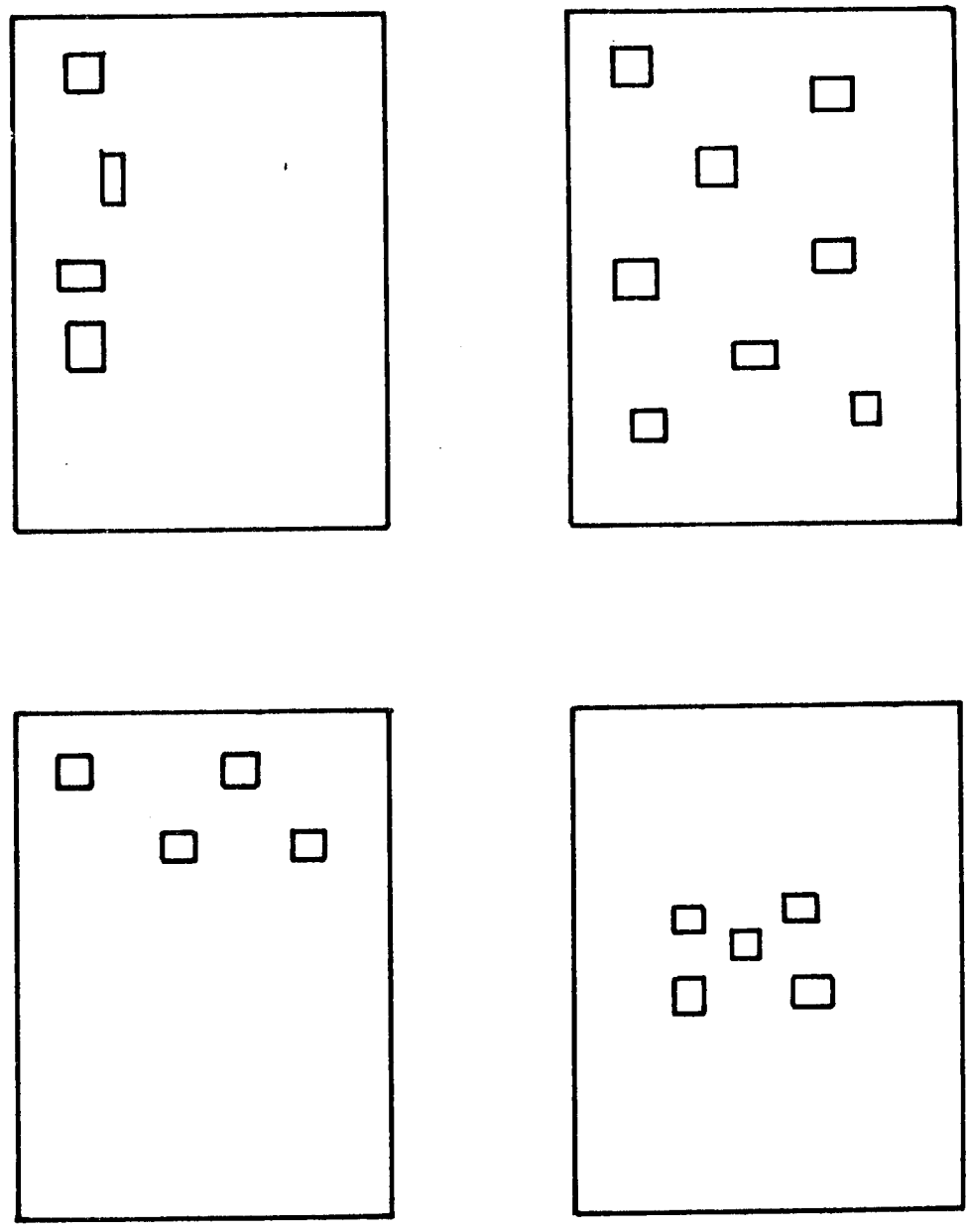


Figure 3-3. Various ways to "scatter" training fields along a flightline. Which scheme do you think is best? Can you construct circumstances where each scheme would be best?

that each of the classes can be characterized by a multidimensional Gaussian probability density function. Each density function is in turn specified by its mean vector and covariance matrix. The classifier requires estimates of the mean vector and covariance matrix for each class from the training samples. In general, the accuracy of the estimate tends to increase as the number of data points used for training increases. This suggests that you should use as many as possible**. Theoretically a lower bound on the number of training data points for any class is n + 1 where n is the dimensionality of the data vector (number of channels) used by the classifier. Fewer than n + 1 points leads to a singular covariance matrix which the classifier cannot use. A practical lower limit is about 10n, but 20n to 100n is desirable if enough ground observations are available.

The Concept of Test Fields

As described above, training fields are used by the classification program to establish a basis for assigning each data point to one of the classes. To assess the success of the classification, a second set of fields, known as test fields, is used. A more detailed explanation of test fields will be given later. Briefly they are used in the following manner: after the classification has been completed, the computer is given additional information about the actual ground cover type for a set of test fields. The system then compares the classification results with the known cover type and tabulates the number of correct and incorrect classifications. An example of the output is shown in Figure 3-4.

The concept of a test field is brought up at this point because the selection of test fields can conveniently be made at the same time training fields are selected. Test fields should also be representative of the classes because they are used to estimate the overall accuracy of the classification. Working with ground observations and multispectral imagery, the analyst usually outlines as many field boundaries as he can, and then for each class chooses a subset for training and another, usually larger, subset for testing. The two subsets must be distinct in order to avoid biased results.

**One must not get carried away however. In the analysis which you are carrying out in conjunction with this study, you have ground observations for all fields. It would not be reasonable to choose data points from all fields as training samples.

CLASSIFICATION STUDY 311772103 CLASSIFIED. APR 27,1973

CHANNELS USED

CHANNEL 5	SPECTRAL BAND	0.50 TO	0.52 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 6	SPECTRAL BAND	0.52 TO	0.55 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 8	SPECTRAL BAND	0.58 TO	0.62 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 12	SPECTRAL BAND	0.80 TO	1.00 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00

CLASSES

	CLASS	GROUP	THRES PCT
1	OATS1	OATS	0.05
2	OATS2	OATS	0.05
3	OATS3	OATS	0.05
4	CORN1	CORN	0.05
5	CORN2	CORN	0.05
6	CORN3	CORN	0.05
7	CORN4	CORN	0.05
8	WHEAT1	WHEAT	0.05
9	WHEAT2	WHEAT	0.05
10	WHEAT3	WHEAT	0.05
11	SOY B1	SOYB	0.05
12	SOY B2	SOYB	0.05
13	SOY B3	SOYB	0.05
14	GRASS1	GRASS	0.05
15	GRASS2	GRASS	0.05
16	GRASS3	GRASS	0.05
17	GRASS4	GRASS	0.05

TEST CLASS PERFORMANCE

	GROUP	NO OF SAMPS	PCT. CORCT	NUMBER OF SAMPLES CLASSIFIED INTO OATS	CORN	WHEAT	SOYB	GRASS	THRSHOLD
1	OATS	66	98.5	65	0	0	0	0	1
2	CORN	93	93.5	0	87	0	5	1	0
3	WHEAT	69	100.0	0	0	69	0	0	0
4	SOYB	57	91.2	0	0	0	52	0	5
5	GRASS	31	83.9	0	5	0	0	26	0
	TOTAL	316		65	92	69	57	27	6

OVERALL PERFORMANCE(299/ 316) = 94.6

AVERAGE PERFORMANCE BY CLASS(467.1/ 5) = 93.4

Figure 3-4. Table showing classification results.

Summary

This step in the analysis logically breaks down into two parts. First, working from your own background experience and the results of coordinating the multispectral imagery with ground observations, you choose candidate training classes. Second, you specify training fields and test fields for each class and check to determine whether a sufficient number of data points has been included in each training class.

References

Field Description Cards are described on pages 2-27 and 2-28 (volume 1) of the LARSYS User's Manual. When reviewing this material, pay particular attention to the second format. This is the format usually used for specifying training and test fields.

Example

Refer again to Figure 2-2 (page 17). The analyst of flightline 66000652 designated the cover types, as obtained from ground observations, by mean of letters in the upper right-hand corner of each field. Five classes were chosen as candidate training classes: Oats, Corn, Wheat, Soybeans and Grass. Grass actually is a catch-all, including spectrally similar red clover, hay, rye, pasture, and diverted acres. The analyst felt that the number of data points available for each of these five cover types was inadequate for specifying training and test fields; thus he combined them into one class, Grass. The area under study also contained water, roads, bare soil and towns. No classes were designated for these items; the majority of them, if they were spectrally dissimilar from the training classes (a reasonable assumption), would be "thresholded" in the classification results. Thresholding will be described later.

After deciding what initial classes to use, the analyst specified training and test fields for each class. The boundaries, identified with dashes on Figure 2-2, delimit the areas used for training and test fields.

EXERCISES

1. State in your own words why it is necessary to select training fields and training classes.
2. What is a practical lower limit on the number of training points needed for a given class.
3. Name at least two factors that go into selecting training classes.
4. What are test fields used for? How are they used as compared to training fields?

5. Assume that training classes have been selected. Describe a technique that might be used to select training and test fields for the classes.

FLIGHTLINE ANALYSIS CASE STUDY

1. The first step in this phase of the analysis is the selection of candidate training classes. By now you should have gained some familiarity with the data in run 71053900. Examine the ground observation information given in figure 2-1 and select a set of candidate training classes.

2. Using the annotated gray scale printout you prepared earlier and the set of classes decided on above, select both candidate training and test fields for each class. Prepare Field Description Cards for each field. Be careful to keep your training fields separate from your test fields. The Field Description Card format is shown in figure 3-5.

FIELD DESCRIPTION CARD CODING SHEET Page____ of____

Run Number (1-8)	Field Designation (11-18)	First Line (21-25)	Last Line (26-30)	Line Interval (31-35)	First Column (36-40)	Last Column (41-45)	Column Interval (46-50)	Field Type (51-58)	Additional Information (59-72)

Figure 3-5. Field Description Card format.

Section 4

REFINEMENT OF TRAINING FIELDS AND CLASSES

Instructional Objectives for this Section

Upon finishing the reading, exercises and case study work associated with this section you should be able to:

a) explain in your own words why refinement of training fields and classes is desirable.
b) explain in your own words the reasons for subclasses and the conditions under which you would define them.
c) when given typical clustering program output:
-determine whether or not subclasses should be defined, and if so, properly define them
-alter training field boundaries to improve homogeneity of the training fields
-decide whether or not further clustering analysis is required.
d) carry out the refinement of an initial set of training data.

Refinement of Training Fields and Classes

The use of the word "candidate" in the previous step in the analysis implied that the initial selection of training fields would be followed by additional analysis to determine if the choices were good ones. This analysis involves the refinement of training fields and classes.

The question of why the training samples have to be refined needs to be answered in terms of the algorithm used for classification. This algorithm is based on the assumption that the data for each class can be described by a multidimensional Gaussian density function. The degree to which this assumption is true affects the accuracy of the classifier. The purpose of the refinement step in the analysis is to check the validity of this assumption.

Clustering

The analysis tool available for examining the statistical characteristics of the data is the clustering program. What clustering is and how the clustering algorithm works is described in pages 27 through 36 of Swain, 1972. It is suggested that you now read this material.

Satisfying the Gaussian Assumption - The Subclass Concept

How is clustering used to refine training field selection? Recall we will be using a classification algorithm which is based upon a Gaussian assumption, i.e., that the data can be approximated by a Gaussian density function. Figure 4-1 (a) shows a typical Gaussian function in one dimension while Figure 4-1 (b) shows a two-dimensional Gaussian density function. Clustering the training samples for each class gives an idea of whether or not the training samples tend to be Gaussian and, more importantly, provides a mechanism for dividing the training classes into approximately Gaussian subclasses if the original data is non-Gaussian. This latter idea is illustrated in Figure 4-2. Figure 4-2 (a) shows a multimodal non-Gaussian density function. Figure 4-2 (b) illustrates how this density may be broken into two components each of which has a Gaussian characteristic. The clustering program is used to determine whether or not the training samples tend to group themselves into distinct clusters. If they do, the original class is divided into subclasses corresponding to these clusters.

The subclass concept is further illustrated in Figure 4-3. This flow chart shows the progression of class and subclass formation and recombination in relation to the total multispectral data analysis sequence. When the analyst first selects classes, he can not be sure that these classes will be spectrally separable. It is also not clear that the class training samples will satisfy the Gaussian assumption. The concept of subclasses and grouping of classes provides a technique for dealing with these uncertainties.

The clustering program gives an indication of whether the training classes satisfy the Gaussian assumption. If the original training samples are represented better by two or three subclasses, then subclasses are specified. No effort is made to distinguish between subclasses in either the feature selection or classification steps of the analysis.

Illustrations of Clustering Program Output

Perhaps the best way to illustrate how the clustering algorithm is typically used to refine training fields is by example. Figures 4-4, 4-5, and 4-6 show cluster maps for three soybean fields selected as candidate training fields. The soybean training samples have been clustered into four clusters. Figure 4-4 shows that all samples from field L46 fell into cluster 4, and most samples from field

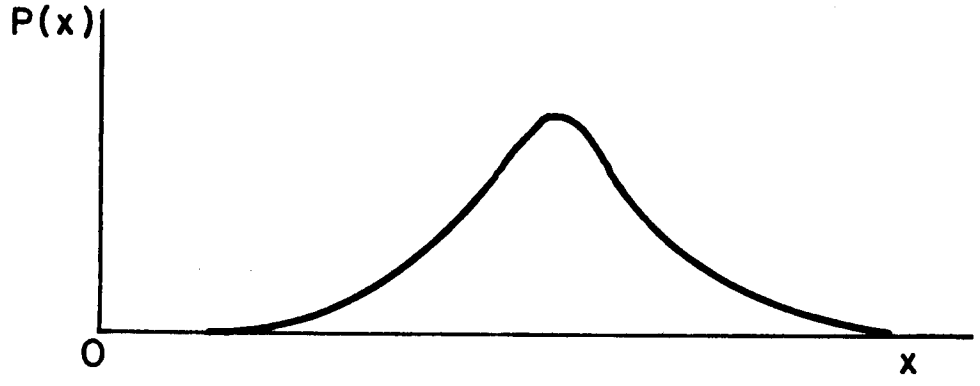


Figure 4-1a. Gaussian density function in one dimension.

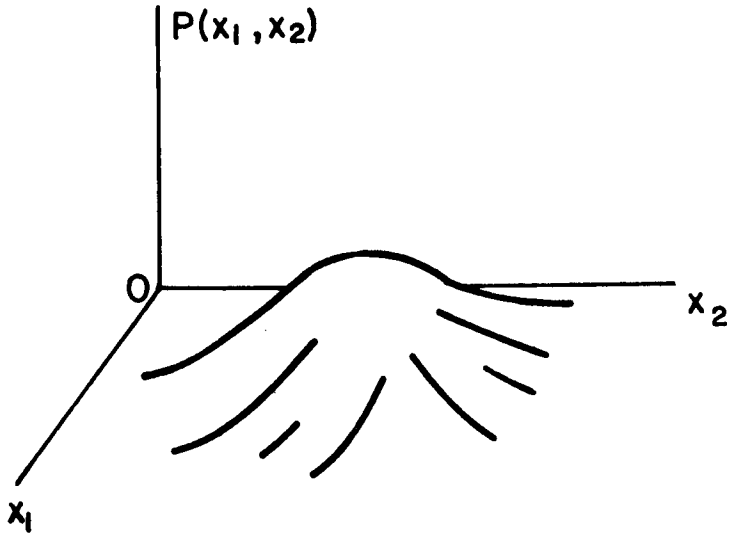


Figure 4-1b. Gaussian density function in two dimensions.

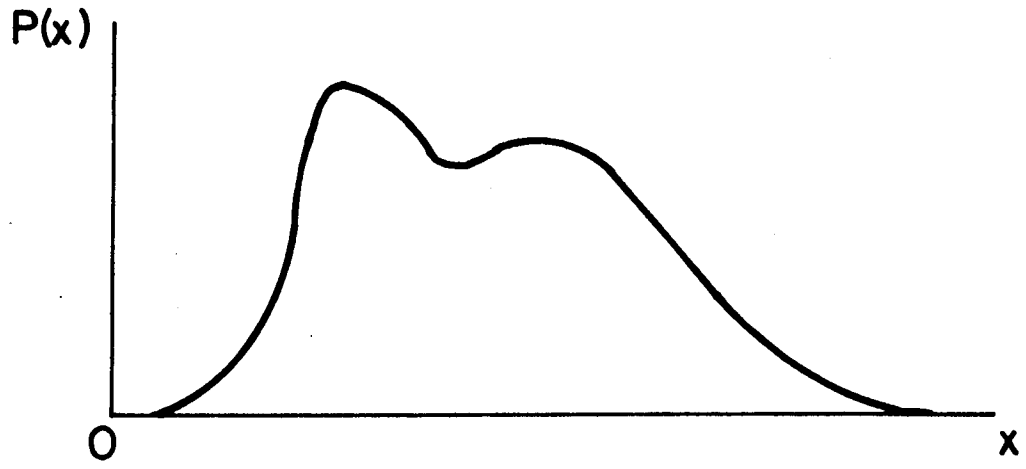


Figure 4-2a. Multimodal non-Gaussian density function.

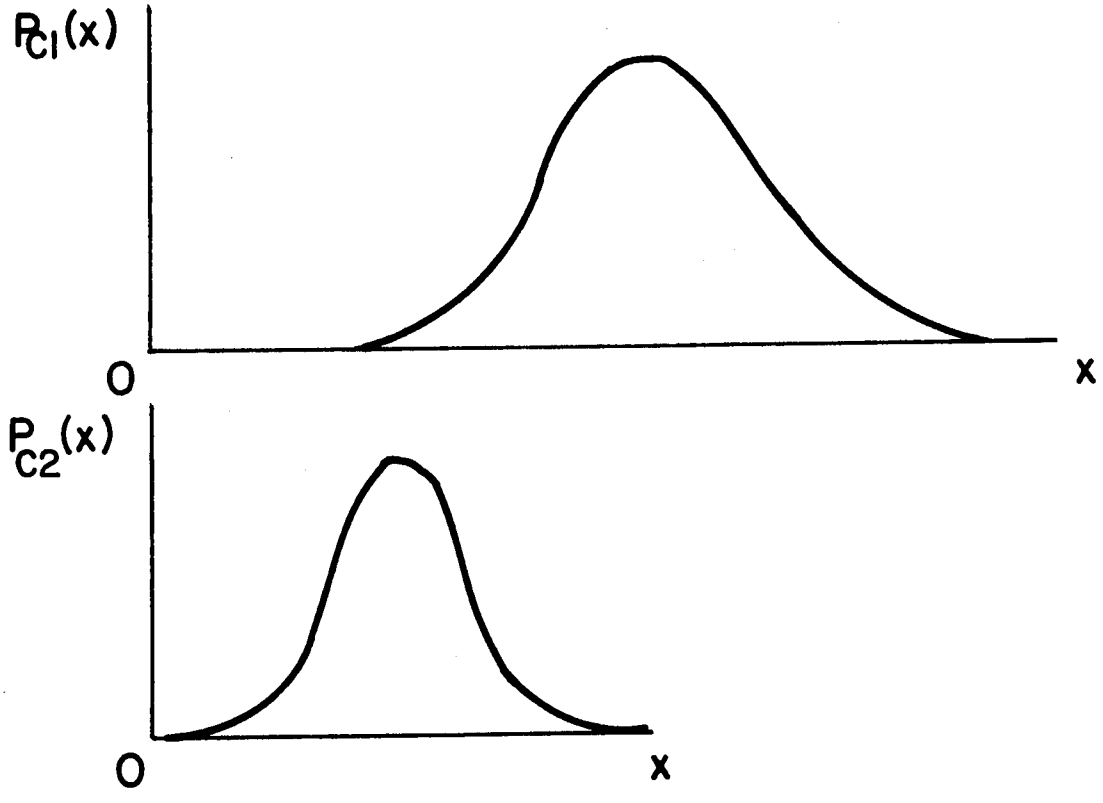


Figure 4-2b. Multimodal function decomposed into two Gaussian components.

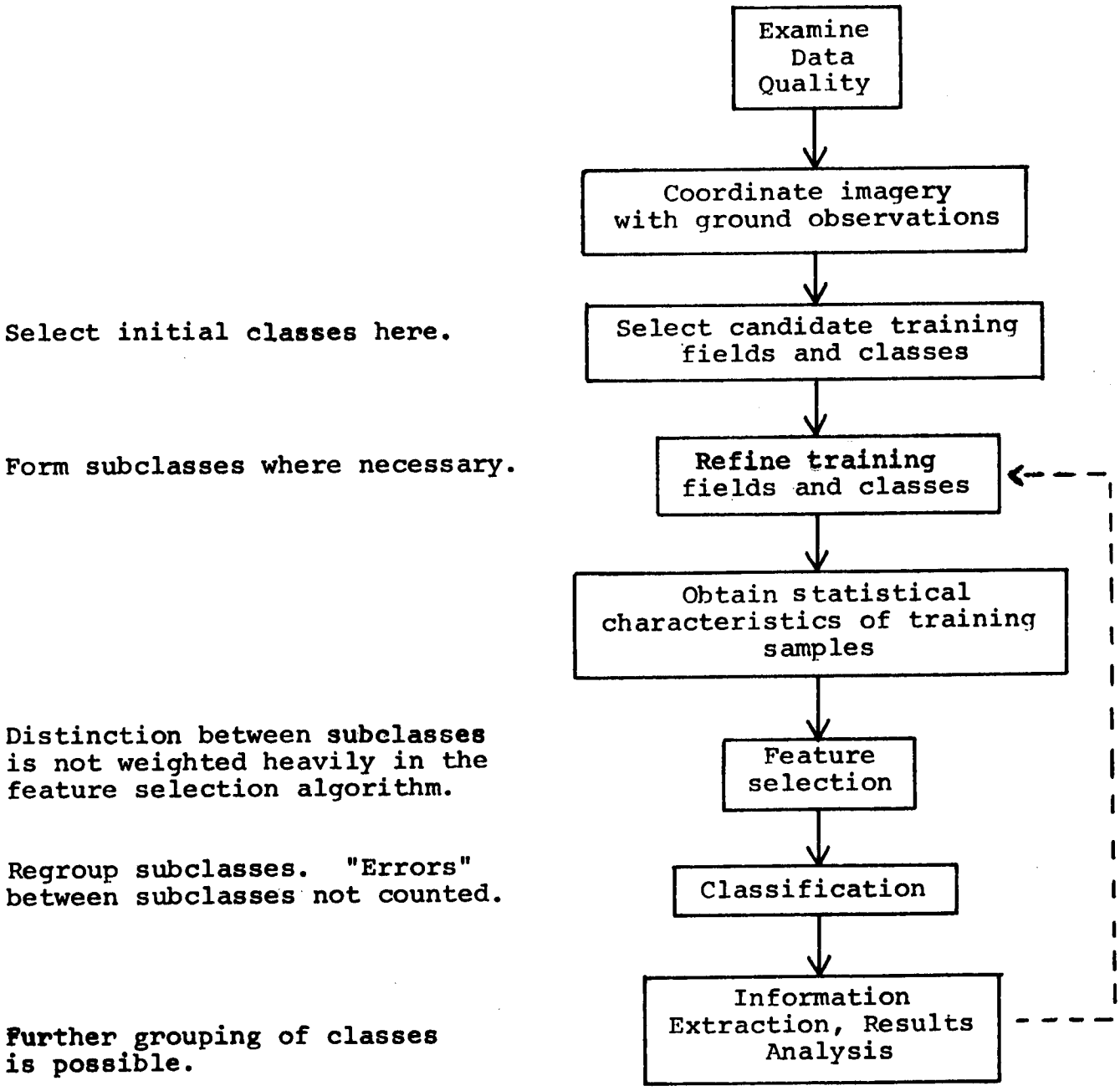


Figure 4-3. Flow chart showing the progression of class and subclass formation.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 14 12 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L46
RUN NO. 66000652
OTHER INFORMATION

TYPE SOYBEANS
NO. OF SAMPLES 189

LINES 291- 311 (BY 1)
COLUMNS 137- 145 (BY 1)

```
          111111111
          333444444
          789012345

291       MMMMMMMMM
292       MMMMMMMMM
293       MMMMMMMMM
294       MMMMMMMMM
295       MMMMMMMMM
296       MMMMMMMMM
297       MMMMMMMMM
298       MMMMMMMMM
299       MMMMMMMMM
300       MMMMMMMMM
301       MMMMMMMMM
302       MMMMMMMMM
303       MMMMMMMMM
304       MMMMMMMMM
305       MMMMMMMMM
306       MMMMMMMMM
307       MMMMMMMMM
308       MMMMMMMMM
309       MMMMMMMMM
310       MMMMMMMMM
311       MMMMMMMMM
```

NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	0	0	0	189

Figure 4-4. A candidate training field for class Soybeans. The data from this and the fields in the next two figures were processed together.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 30 32 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L40
RUN NO. 66000652
OTHER INFORMATION

TYPE SOYBEANS
NO. OF SAMPLES 270

LINES 740- 754 (BY 1)
COLUMNS 53- 70 (BY 1)

```
                  000000000000000000
                  555555566666666667
                  345678901234567890

740               I
741               II   I
742
743
744
745
746
747
748
749
750
751
752
753                         II
754                    II
```

NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	262	8	0	0

Figure 4-5. A second candidate training field with sample points clustered.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 30 31 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L28
RUN NO. 66C00652
OTHER INFORMATION

TYPE SOYBEANS
NO. OF SAMPLES 221

LINES 63- 75 (BY 1)
COLUMNS 67- 83 (BY 1)

```
     00000000000CCC0000
     66677777777778d88
     78901234567890123

63   EEEEEEEEEEEEEEEEE
64   EEEEEEEEEEEEEEEEE
65   EEEEEEEEEEEEEEEEE
66   EIIEEEEEEIIIIIIEEE
67   EEEEIIIIIEIIIIIEEE
68   EEEEIIIIIIIIIIEEEE
69   EEEEEEEEEEEIIIEIEE
70   IIEEEEEEIIIEEEEIEE
71   EEEEEEEEEEEIEEEEEE
72   IIEEEEEEEEEEEEEEEE
73   EEEEEEEEEEEEEEEEEE
74   EEEEEEEEEEEEEEEEIE
75   EEEEEEEEEEEEEEIEII
```

NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	0	42	179	0

Figure 4-6. A third candidate training field for soybeans.

L40 fell into cluster 1. The sample points for field 28, Figure 4-6, fell into two clusters. Examining the separability information, Figure 4-7, it can be seen from the quotient column that clusters 2 and 3 are spectrally similar, that is the quotient is less than .75. Therefore, field L28 can be used as a training field as it is. Figure 4-7 also shows that the other clusters are spectrally distinct (QUOT values > 1).

References

Pages 27 to 36 of Swain, 1972, have previously been recommended as background reading on cluster analysis.

Pertinent pages in Section 6 (volume 2) of the LARSYS User's Manual are CLU-1 to CLU-16 for a discussion of the clustering processing function and pages CLU-17 to CLU-20 for a discussion of the algorithms used in the program.

Example

The analyst realized that the success of the classification depends on the careful selection and distribution of the training fields. Frequently significant spectral variation may be observed among fields containing the same crop. (For example, as you will see, the analyst found three significantly different subclasses of oats in this run. This could have been due to different soils, moisture, planting dates, crop density, and/or seed brands. To maximize accuracy the analyst needed to divide his classes into spectrally different subclasses.) The analyst used CLUSTER in the following way:

```
*CLUSTER
OPTIONS MAXCLAS(6)
CHANNELS 1,6,10,12
DATA
 [field description cards for Oats]
END
*CLUSTER
OPTIONS MAXCLAS(6)
CHANNELS 1,6,10,12
DATA
 [field description cards for Corn]
END
*CLUSTER
 :
 :
(continue until all classes are represented)
```

SEPARABILITY INFORMATION

I	J	D(I,J)	D(I)	D(J)	D(I)+D(J)	QUOT
1	2	16.640	6.465	6.813	13.278	1.253
1	3	22.092	6.549	3.476	10.025	2.204
1	4	44.382	4.610	3.561	8.171	5.432
2	3	5.760	6.149	3.691	9.840	0.585
2	4	30.204	8.584	3.396	11.980	2.521
3	4	27.395	3.233	3.418	6.652	4.119

AVERAGE QUOTIENT 2.686

Figure 4-7. Table showing relative separation between clusters formed from soybean training fields.

The "6" in MAXCLAS(6) caused six clusters to be formed. The analyst chose six initially by a rule-of-thumb estimate of "twice the expected subclasses."

Four channels were chosen for the clustering analysis. This number is a compromise dictated by the constraints of computer storage capacity and computation time. The processor can cluster slightly less than 40,000/n vectors, where n is the number of channels used. (See page CLU-6 of LARSYS User's Manual.) When choosing a subset of channels to use for clustering, it is usually a good idea to choose the channels so as to obtain a good representation of the spectral range covered by the multispectral scanner.

Figure 4-8 illustrates the output for the initial clustering of each of four oats training fields. (For the present, ignore the markings on the fields.) One way of determining subclasses is shown in figure 4-9. The analyst wrote the numbers 1 through 6 in a circle. He then connected the pairs of numbers that have small separability quotients (less than 0.8 in this case). The separability quotients between clusters 1 and 2, 2 and 3, and 1 and 3 are all less than 0.8. However the quotients between clusters 1, 2, and 3 and the remaining clusters are all larger than .9. Thus the analyst decided to lump clusters 1, 2, and 3 into one subclass. He then looked closely at the clusters 4, 5, and 6. The distance between clusters 4 and 6 is .55. The distance between 4 and 5 is .81 but the distance between 5 and 6 is .93. Clusters 4, 5, and 6 could be grouped as a second subclass but the analyst decided to be more cautious and make 4 and 6 the second subclass and 5 a third subclass. This procedure for grouping cluster is included in the CLUSTER processing function and the last page of CLUSTER output provides a table of suggested groupings. You should read the pages on CLUSTER in Volume 2 of the LARSYS User's Manual for more information.

The analyst went back to the clustered field printouts shown in figure 4-8. He marked representative fields for each of his subclasses and made new Field Description Cards. Just to see how "solid" his subclasses were, he reran the clustering for oats using the new Field Description Cards and MAXCLAS(4). The analyst might have chosen MAXCLAS(3) because the previous analysis had resulted in three distinct clusters but he was curious to see if requesting four clusters would still result in three distinct clusters. The resulting output is shown in figure 4-10 and 4-11. The subclasses were divided into different, highly separable clusters except for Oats 3 which combined two clusters. The two clusters, 1 and 2, are close to each other (separability quotient = 0.62) and as a pair are highly separable from the other two clusters. Thus three separable subclasses are maintained. Note the calculations on figure 4-11 to determine the approximate number of data points in each subclass. All subclasses contained sufficient data points for, say, a 4-channel classification (n=4),

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

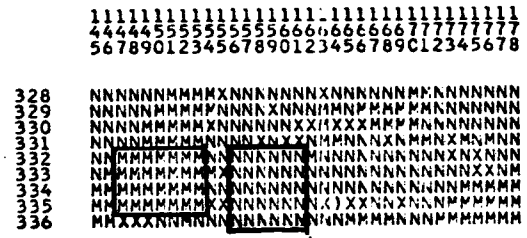
DEC 30, 1974
5 41 21 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD 25
RUN NO. 66000652
OTHER INFORMATION

TYPE DATA
NO. OF SAMPLES 306

LINES 328- 336 (BY 1)
COLUMNS 145- 178 (BY 1)



NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4	5	6
SYMBOL		)	S	X	N	M
POINTS	0	1	0	28	176	101

Figure 4-8a. Initial cluster map for field 25. Markings indicate analyst's interpretation of the output.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

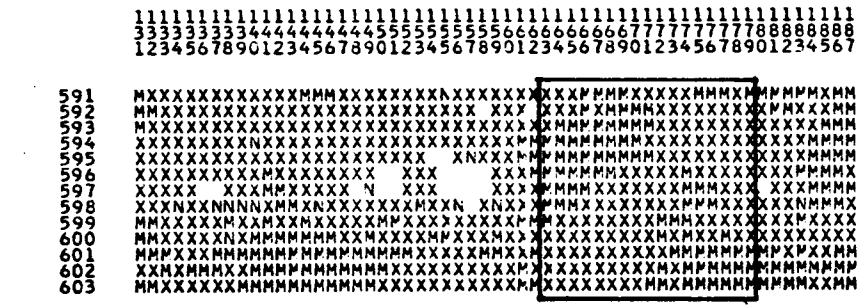
DEC 30,1974
5 41 22 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD 48
RUN NO. 66000652
OTHER INFORMATION

TYPE OATS
NO. OF SAMPLES 741

LINES 591- 603 (BY 1)
COLUMNS 131- 187 (BY 1)



NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4	5	6
SYMBOL		)	S	X	N	M
POINTS	21	0	0	473	14	233

combined

Figure 4-8b. Initial cluster map for field 48. Markings show refinement of training field boundaries.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

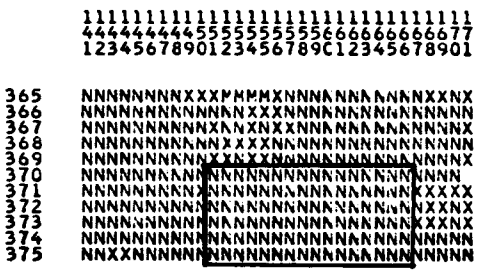
DEC 30, 1974
5 41 21 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD 28
RUN NO. 66C00652
OTHER INFORMATION

TYPE OATS
NO. OF SAMPLES 341

LINES 365- 375 (BY 1)
COLUMNS 141- 171 (BY 1)



NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4	5	6
SYMBOL		1	S	X	N	M
POINTS	1	0	0	39	297	4

Figure 4-8c. Initial cluster map for field 28. Refinement of boundaries for eliminating cluster 4 points is shown.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

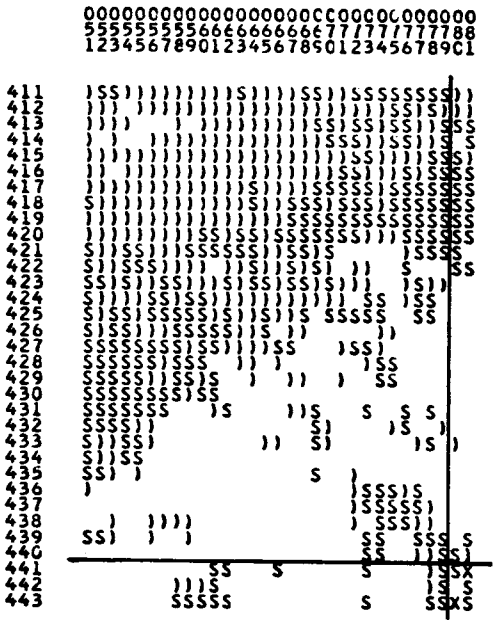
DEC 30,1974
5 41 22 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD 40
RUN NO. 66000652
OTHER INFORMATION

TYPE OATS
NO. OF SAMPLES 1023

LINES 411- 443 (BY 1)
COLUMNS 51- 81 (BY 1)



NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4	5	6
SYMBOL		)	S	X	N	M
POINTS	394	333	294	2	0	0

combined

Figure 4-8d. Initial cluster map for field 40.

XBLARSYS
JAMES RUSSELL

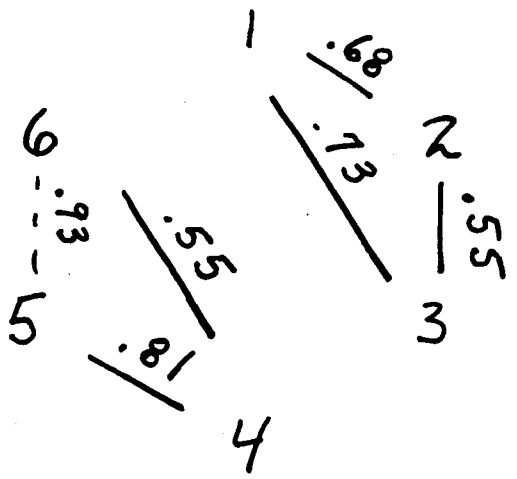
LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 41 25 PM
LARSYS VERSION 3

SEPARABILITY INFORMATION

I	J	D(I,J)	D(I)	D(J)	D(I)+D(J)	QUOT
1	2	7.035	5.398	4.908	10.306	0.683
1	3	8.656	5.502	6.302	11.804	0.733
1	4	11.059	6.020	4.446	10.467	1.057
1	5	11.559	4.878	3.835	8.713	1.327
1	6	18.162	5.802	5.969	11.771	1.543
2	3	6.680	5.865	6.180	12.046	0.555
2	4	11.519	5.525	3.783	9.308	1.281
2	5	11.823	4.188	3.422	7.610	1.554
2	6	16.348	6.589	4.374	10.963	1.491
3	4	7.650	3.778	3.361	7.140	1.071
3	5	13.537	4.822	4.260	9.082	1.491
3	6	11.343	4.160	4.076	8.236	1.377
4	5	10.676	6.765	6.418	13.183	0.810
4	6	8.168	6.545	8.229	14.774	0.553
5	6	16.065	8.931	8.312	17.243	0.932

AVERAGE QUOTIENT 1.097



The six clusters can be grouped into 3 subclasses.

I 1, 2, 3

II 4, 6

III 5

Figure 4-9. Separability information for initial clustering of OATS training samples.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 43 10 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L50
RUN NO. 66000652
OTHER INFORMATION

TYPE OATS1
NO. OF SAMPLES 28

LINES 332- 335 (BY 1)
COLUMNS 147- 153 (BY 1)

```
          1111111
          4445555
          7890123

332       EEEEEEE
333       EEEEEEE
334       EEEEEEE
335       EEEEEEE
```

NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	0	0	28	0

Figure 4-10a. Cluster map for refined OATS training field.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 43 10 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L51
RUN NO. 66C00652
OTHER INFORMATION

TYPE OATS2
NO. OF SAMPLES 30

LINES 332- 336 (BY 1)
COLUMNS 156- 161 (BY 1)

```
          111111
          555566
          678901

332       MMMMMM
333       MMMMMM
334       MMMMMM
335       MMMMMM
336       MMMMMM
```

NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	0	0	0	30

Figure 4-10b. Cluster map for refined OATS training field.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 43 11 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L53
RUN NO. 66000652
OTHER INFORMATION

TYPE OATS2
NO. OF SAMPLES 96

LINES 370- 375 (BY 1)
COLUMNS 151- 166 (BY 1)

```
                  1111111111111111
                  5555555556666666
                  1234567890123456

              370 MMMMMMMMMMMMMMMM
              371 MMMMMMMMMMMMMMMM
              372 MMMMMMMMMMMMMMMM
              373 MMMMMMMMMMMMMMMM
              374 MMMMMMMMMMMMMMMM
              375 MMMMMMMMMMMMMMMM
```

NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	0	0	0	96

Figure 4-10c. Cluster maps for refined OATS training field.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

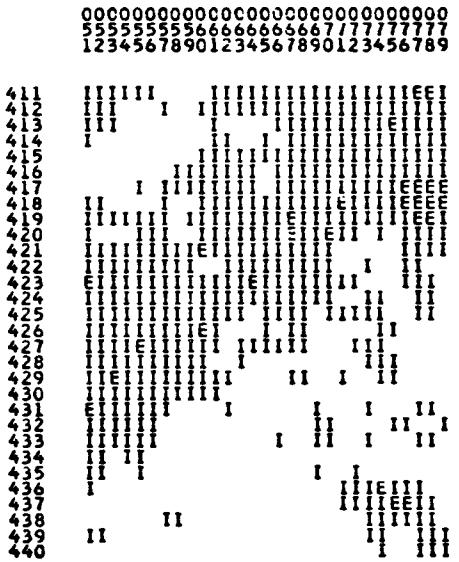
DEC 30, 1974
5 43 11 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L54
RUN NO. 66C00652
OTHER INFORMATION

TYPE OATS3
NO. OF SAMPLES 870

LINES 411- 440 (BY 1)
COLUMNS 51- 79 (BY 1)



NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	374	469	27	0

only 3% of total number of samples

Figure 4-10d. Cluster map for refined OATS training field.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 43 13 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L55
RUN NO. 66000652
OTHER INFORMATION

TYPE OATS1
NO. OF SAMPLES 221

LINES 591- 603 (BY 1)
COLUMNS 163- 179 (BY 1)

```
          11111111111111111
          66666667777777777
          34567890123456789

591       EEEEEEEEEEEEEEEEE
592       EEEEEEEEEEEEEEEEE
593       EEEEEEEEEEEEEEEEE
594       EEEEEEEEEEEEEEEEE
595       EEEEEEEEEEEEEEEEE
596       EEEEEEEEEEEEEEEEE
597       EEEEEEEEEEEEEEEEE
598       EEEEEEEEEIEEEEEEE
599       EEEEEEEEEEEEEEEEE
600       EEEEEEEEEEEEEEEEE
601       EEEEEEEEEEEEEEEEE
602       EEEEEEEEEEEEEEEEE
603       EEEMEEEEEEEEEEEEE
```

NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	0	1	219	1

Figure 4-10e. Cluster map for refined OATS training field.

XBLARSYS
JAMES RUSSELL

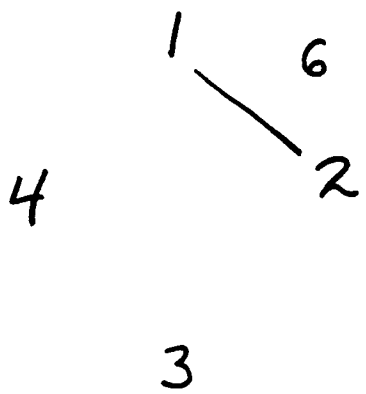
LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 43 15 PM
LARSYS VERSION 3

SEPARABILITY INFORMATION

I	J	D(I,J)	C(I)	D(J)	D(I)+D(J)	QUOT
1	2	7.005	5.700	5.584	11.284	0.621
1	3	15.089	5.744	7.766	13.510	1.117
1	4	11.907	4.341	3.470	7.811	1.524
2	3	10.735	5.878	5.837	11.715	0.916
2	4	13.145	4.557	4.110	8.668	1.517
3	4	15.397	9.043	7.855	16.897	0.911

AVERAGE QUOTIENT 1.101



3 subclasses are still evident

class	3 Oats 1	4 Oats 2	1, 2 Oats 3
# points	28	30	870
	221	96	
total	249	126	870

Figure 4-11. Analysis of separability information for second clustering run.

but if the analyst had been planning to use all 12 available channels (n=12), then the 126 points in subclass 3 would have been at best marginally adequate.

The analyst did a similar analysis on the other candidate training classes and arrived at a total of 17 subclasses for the entire run.

Note that in the card setup used by the analyst the IDNUMBER card was not used. Because of the number of iterations which are often required when refining training classes and fields, it is recommended that you use an IDNUMBER card to help identify your output.

Comment: The clustering algorithm is easily the least understood of the LARSYS analysis algorithms. The details of its use are sometimes very much problem and data-dependent. For instance, in the example above, clustering into "twice the expected subclasses" and use of 0.75 as the breakpoint for cluster separability are strictly rules of thumb, although they have been pretty well established for crop classification problems by extensive experience with agricultural data. The analyst may use these suggestions as a starting point for his work, but he is encouraged to be flexible and to experiment with the use of this algorithm as applied to his particular problem.

EXERCISES

1. Write a brief statement explaining why it is desirable to refine the training samples.
2. Prepare an outline from which you could give a three or four minute talk explaining the concept of subclasses and why they are used.
3. Figures 4-12, 4-13, 4-14 and 4-15 show clustering output for a corn training class. Analyze this output showing how you would refine field boundaries and select subclasses.

FLIGHTLINE ANALYSIS CASE STUDY

Use the clustering program to refine the training samples you selected earlier for run 71053900. Assume initially five or six clusters for each class. Analyze the cluster maps and separability information for each class and divide the training classes into subclasses where appropriate. Refine training field boundaries as needed. You may wish to take advantage of your instructor's experience at this point in the analysis. Consult him after you get your initial cluster maps and discuss refinement alternatives.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 15 27 PM
LARSYS VERSION 3

FIELD INFORMATION

FIELD L9
RUN NO. 66C00652
OTHER INFORMATION

TYPE CORN
NO. OF SAMPLES 260

LINES 270- 289 (BY 1)
COLUMNS 205- 217 (BY 1)

```
          2222222222222
          0000011111111
          5678901234567

270       MMMMIIIIIIIII
271       MEMMIMMIIIIIM
272       MMEMIIIIIIIII
273       MMEEIIIIIIEEI
274       MMEEIIIIIEEEE
275       MMEEEEEEEEEEE
276       IMMEEEEEEEEEE
277       MMEEEEEEEEEEE
278       MEEEEEEEEEEEE
279       MMEEEEEEEEEEE
280       MMEEEEEEEEEEE
281       MMEEEEEEEEEEE
282       IMEEEEEEEEEEE
283       MMEEEEEEEEEEE
284       MMEEEEEEEEEEE
285       IMEEEEEEEEEEE
286       MMEEEEEEEEEEE
287       MMEEEEEEEEEEE
288       MMEEEEEEEEEEE
289       IMMEEEEEEEEEE
```

NUMBER OF POINTS PER CLUSTER

CLUSTER	1	2	3	4
SYMBOL		I	E	M
POINTS	0	39	177	44

Figure 4-12. Clustering output for a corn training field.

```
XBLARSYS                    LABORATORY FOR APPLICATIONS OF REMOTE SENSING          DEC 30,1974
JAMES RUSSELL                        PURDUE UNIVERSITY                          5 16 43 PM
                                                                                LARSYS VERSION 3

                              FIELD INFORMATION

FIELD L7                          TYPE  CORN                  LINES    177-  189  (BY  1)
RUN NO.  66C00652                 NO. OF SAMPLES  260         COLUMNS   51-   70  (BY  1)
OTHER INFORMATION

                          000000C0000C00000C00
                          555555555666666666 7
                          123456789012345678 90

                    177   IIIIIIIIIIIIIIEIIIIIE
                    178   IIIIIIIIIIIIIIIIIIIII
                    179   IIIIIIIIIIIIIIIIIEIII
                    180   IIIIIIIIIIIIIIIIIIIII
                    181   IIIIIIIIIIIIIIIIIEEII
                    182   IIIIIIIIIIIIIIIIIIIII
                    183    IIIIIIIIIIIIIIIIIIII
                    184   IIIIIIIIIIIIIIIIIIIII
                    185   IIIIIIIIIIIIIIIIIIIII
                    186   IIIIIIIIIIIIIIIIIIIII
                    187   IIIIIIIIIIIIIIIIIIIIE
                    188   IIIIIIIIIIIIEEIIIIIII
                    189   IIIIIIIIIIIIIIII IIIE

                        NUMBER OF POINTS PER CLUSTER

   CLUSTER      1        2        3        4
   SYMBOL                I        E        M
   POINTS       2      249        9        0
```

Figure 4-13. Clustering output for a second corn training field.

```
XBLARSYS                    LABORATORY FOR APPLICATIONS OF REMOTE SENSING          DEC 30,1974
JAMES RUSSELL                        PURDUE UNIVERSITY                          5 18 10 PM
                                                                                LARSYS VERSION 3

                              FIELD INFORMATION

FIELD L11                     TYPE  CORN                  LINES    381-  395  (BY   1)
RUN NO.  66000652             NO. OF SAMPLES  195         COLUMNS    5-   17  (BY   1)
OTHER INFORMATION

                              0000000000000
                              00C001111111
                              5678901234567

                          381 MIIIII II
                          382 IIIII
                          383 IIII I
                          384 IIII
                          385 III
                          386
                          387 II
                          388
                          389 I
                          390
                          391
                          392
                          393
                          394
                          395 I

                        NUMBER OF POINTS PER CLUSTER

          CLUSTER     1      2      3      4
          SYMBOL             I      E      M
          ------------------------------------
          POINTS    166     28      0      1
```

Figure 4-14. Clustering output for a third corn training field.

XBLARSYS
JAMES RUSSELL

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

DEC 30,1974
5 19 34 PM
LARSYS VERSION 3

SEPARABILITY INFORMATION

I	J	D(I,J)	D(I)	D(J)	D(I)+D(J)	QUOT
1	2	14.174	8.151	6.820	14.972	0.947
1	3	28.687	7.993	8.597	16.590	1.729
1	4	29.524	9.128	11.678	20.806	1.419
2	3	15.301	7.363	8.425	15.788	0.969
2	4	18.759	7.960	10.532	18.492	1.014
3	4	18.871	8.179	9.892	18.070	1.044

AVERAGE QUOTIENT 1.187

Figure 4-15. Table showing relative separation between clusters.

Section 5

OBTAINING STATISTICAL CHARACTERISTICS OF THE TRAINING SAMPLES

Instructional Objectives for this Section

After you have read this section, examined the LARSYS User's Manual references, worked the exercises and completed the next part of the case study, you should be able to:

a) state what is meant by the statistics of a training class
b) explain why statistics are needed
c) be able to use the LARSYS processing functions to obtain the training statistics when you have been given the Field Description Cards for a set of training classes.

Obtaining Statistical Characteristics for the Training Samples

Once the training samples have been selected and refined, the next step in the analysis is to obtain the training sample statistics. Recall our earlier mention of the fact that the classification algorithm is based on the assumption that the various classes (and subclasses) can each be characterized by a multivariate Gaussian probability density function. These density functions are defined in terms of their mean vectors and covariance matrices. The training samples are used to estimate the class mean vector and covariance matrices.

The LARSYS statistics processing function is used to compute the training statistics. The analyst usually obtains the statistics in punched card form. The "statistics deck" will be used in later steps in the analysis. In addition to the mean and covariance information, the statistics processor can produce histograms of the data for the training fields and classes. Examples of histograms from channels 5, 8 and 12 are shown in figure 5-1. Typically the multispectral data analyst will request histograms for a representative set of channels for each class. A glance at these histograms serves as a partial check on whether the training samples are distributed in an approximately Gaussian manner. (This is only a partial check because the marginal density functions represented by the histograms do not necessarily reflect the nature of the multidimensional density function.) If a multimodal density function (two or more maximum points) appears, it is an indication of an non-Gaussian situation. Examination of the histograms for individual fields might reveal the reason for this condition.

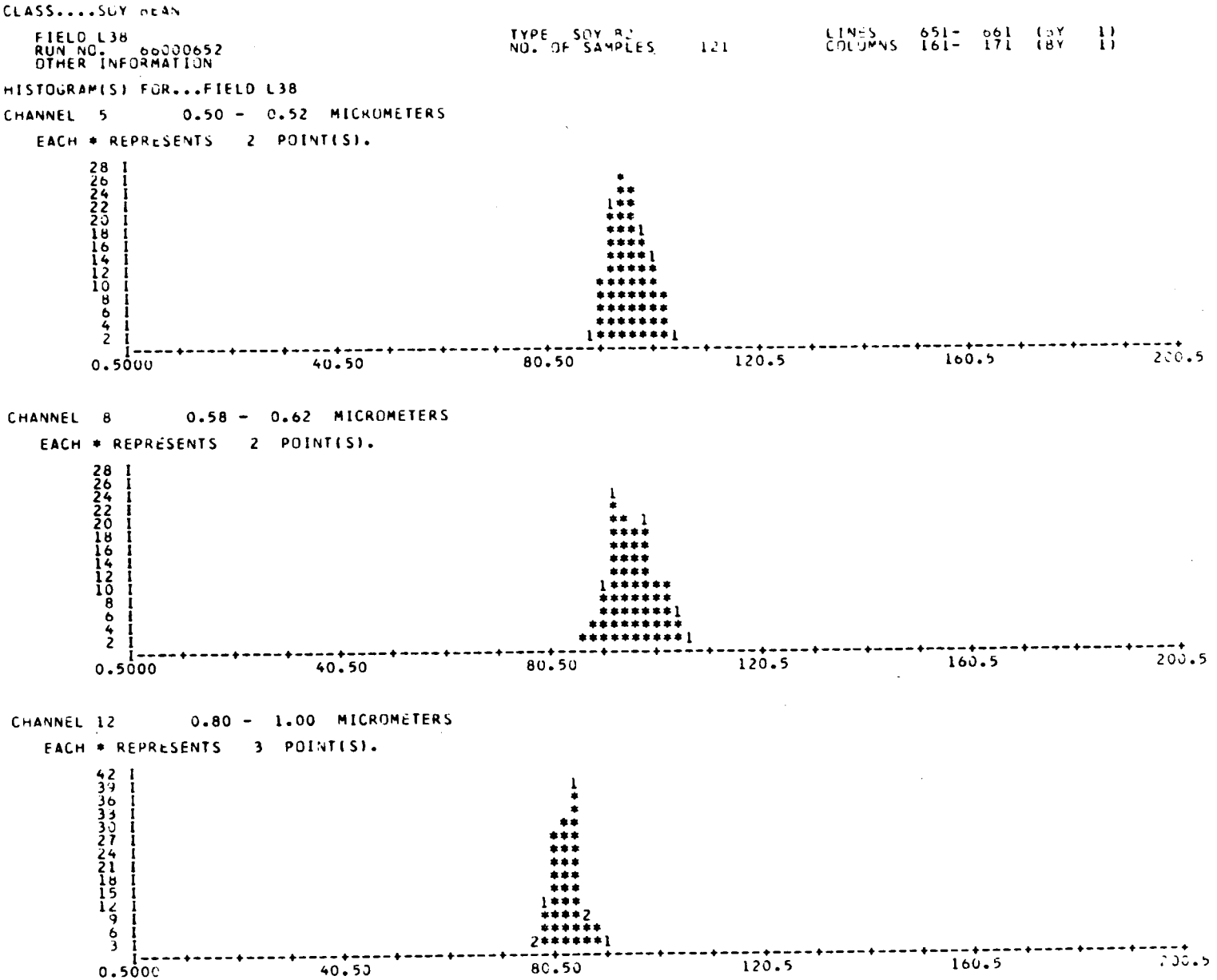


Figure 5-1. Histograms from channels 5, 8, and 12.

References

Pages STA-1 to STA-22 of the LARSYS User's Manual describe the STATISTICS processing function. A general description of the processing function is followed by a detailed discussion of the input control cards and examples of the printed output are given.

Example

After examining the candidate training fields and classes, selecting field boundaries and defining subclasses, the analyst was ready to obtain the statistical characteristics of the training samples. A necessary output at this stage is a punched statistics deck, needed later in the analysis. The analyst wanted to look at some histograms of his subclasses for a partial check on whether they were Gaussian in nature. The control card deck used was:

```
*STATISTICS
OPTIONS HIST(1,6,10)
PRINT HIST(C)
PUNCH
CHANNELS 1,2,3,4,5,6,7,8,9,10,11,12
DATA
CLASS OATS 1
 (Field Description Cards for Oats 1)
CLASS OATS 2
 (Field Description Cards for Oats 2)
  :
  :
(continued until all 17 subclasses were listed with appro-
priate Field Description Cards)
END
```

If the class histograms had been seriously multimodal, the analyst would have looked at his clustered output again to refine his fields further. He may have decided to rerun the clustering program with his refined fields to check for uniform subclass representation. For example, figures 5-2, 5-3 5-4 show the histograms of the subclasses Oats 1, Oats 2, and Oats 3. Since there were no seriously multimodal histograms, the analyst decided not to change his training fields further. He then moved on to the next step.

EXERCISE

State in your own words what is meant by the phrase "training class statistics" and why the statistics are required.

Using the training field description cards obtained in the refinement step of your analysis of run 71053900, use the statistics processing function to obtain a punched statistics deck for your training classes. Also obtain histograms for each class in a representative set of channels.

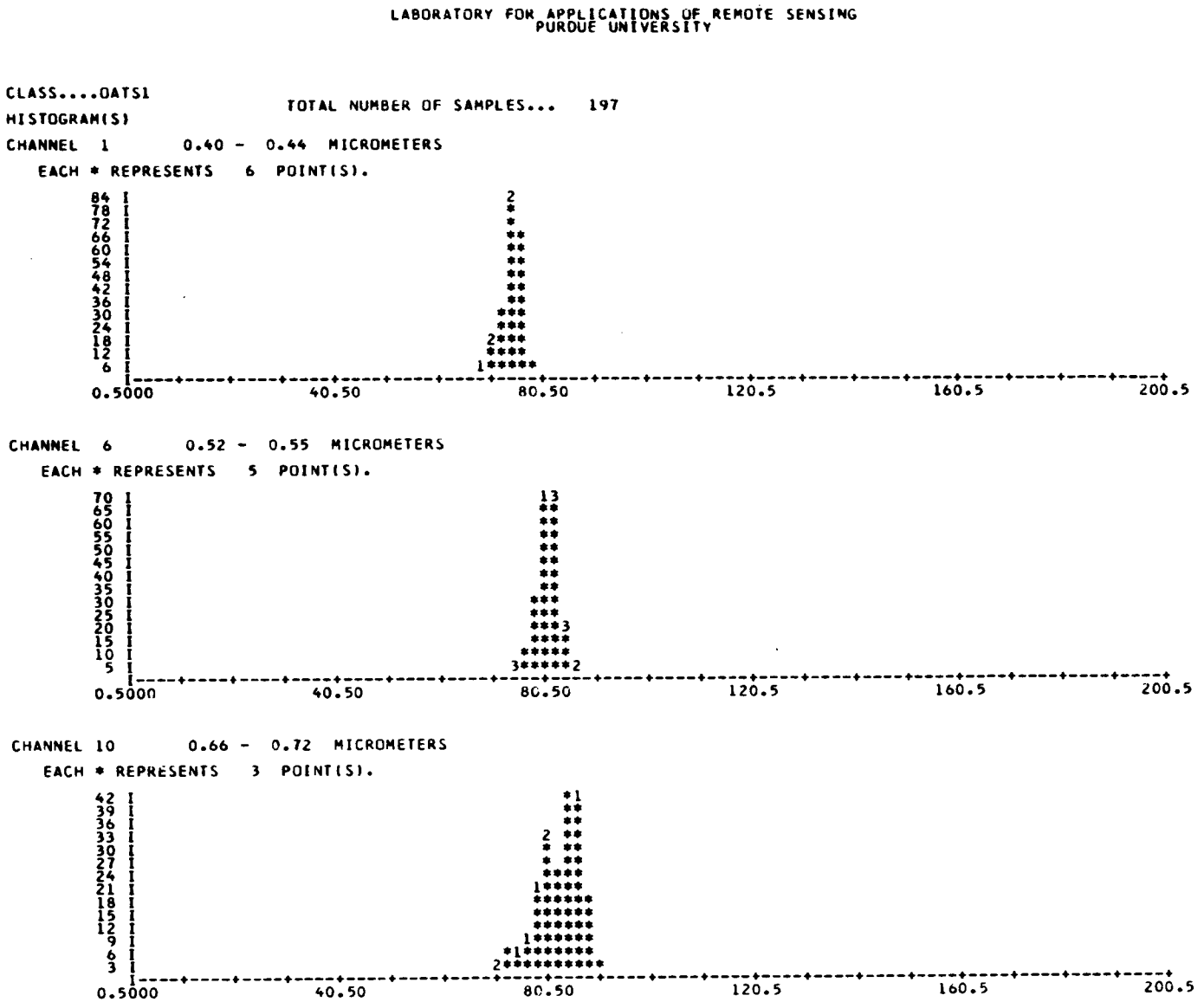


Figure 5-2. Representative histograms of subclass OATS1.

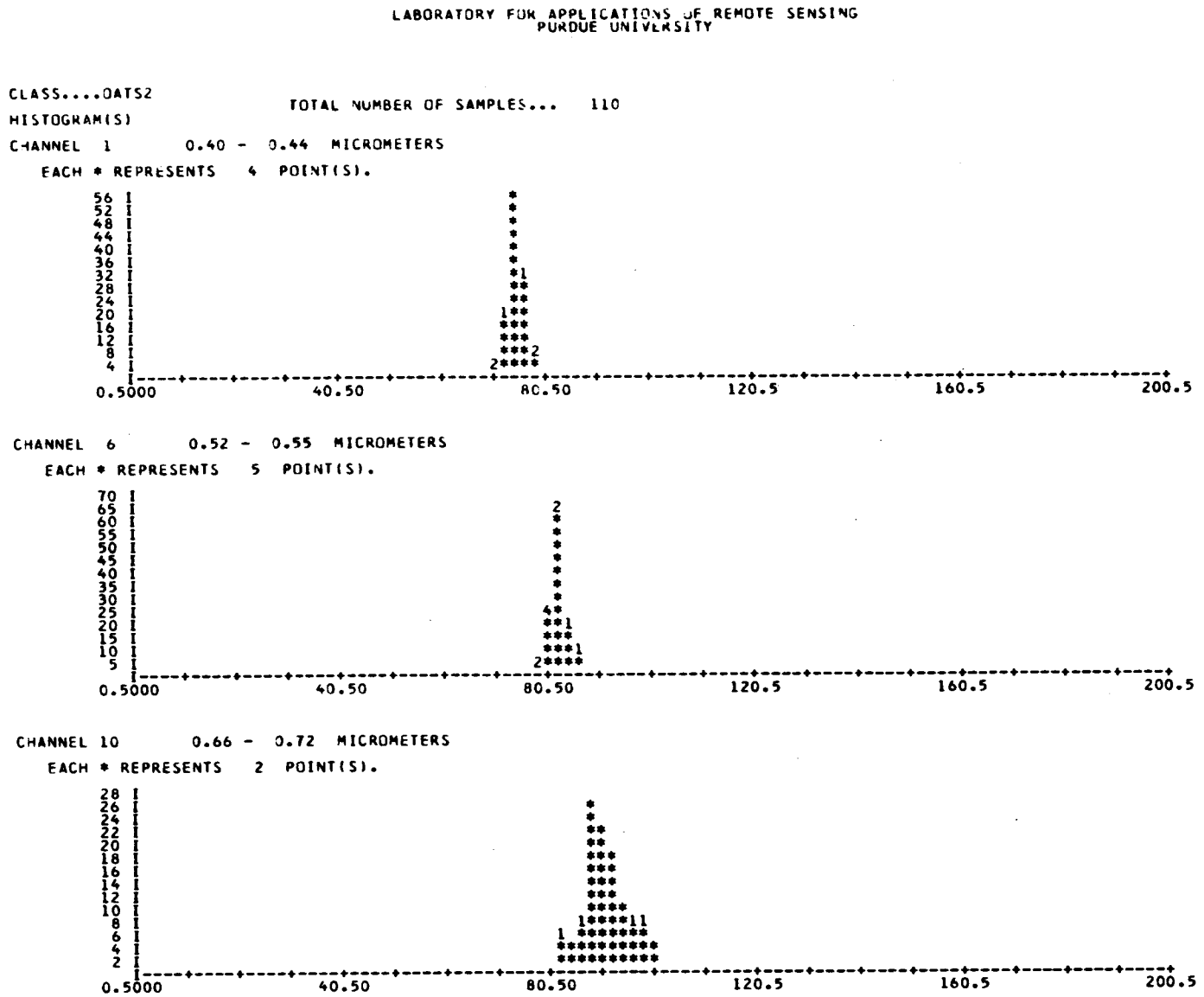


Figure 5-3. Representative histograms of subclass OATS2.

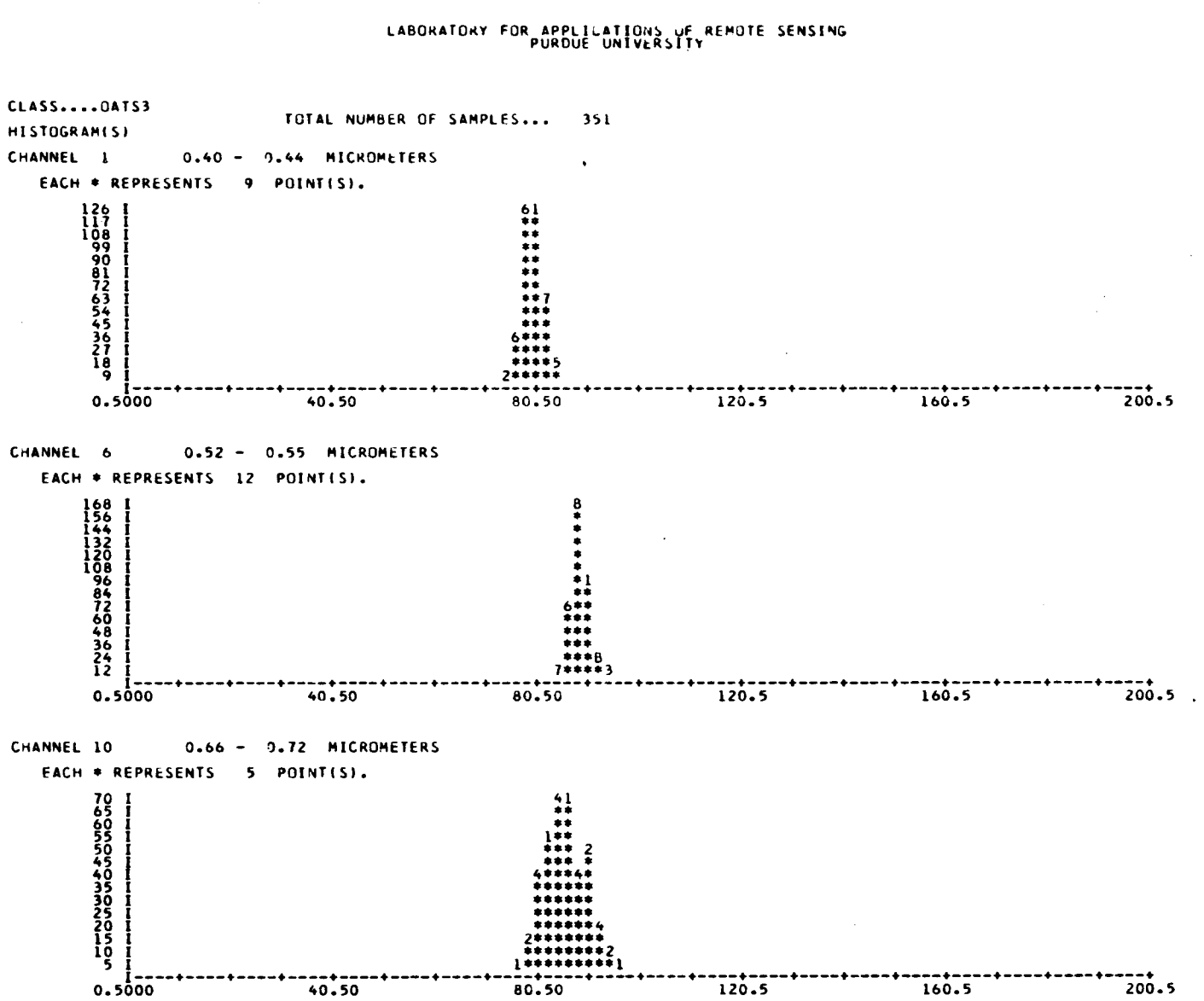


Figure 5-4. Representative histograms of subclass OATS3.

Section 6

FEATURE SELECTION

Instructional Objectives for this Section

Study of this section and its associated references, exercises and case study step should enable you to:

a) state, upon being shown two pairs of one-dimensional density functions, for which pair the statistical distance between the density functions is largest. Your choice should be supported by one or two reasons for making the decision you did.
b) state the general (not exact functional) relationship between statistical distance and probability of correct classification.
c) examine the output of a separability run and, based on this examination, select a subset of channels for use in the classification program, supporting your choice of channels with some sound reasons.
d) use the SEPARABILITY processing function of LARSYS to select a subset of channels for use by the classification algorithm when you have been given the statistics deck.

Feature Selection

The Introduction (pages 1 to 3) of Swain, 1972, provides a brief discussion of the role of feature selection in the overall sequence of multispectral data analysis. You should read this material at this time.

Swain points out that the feature selection operation in LARSYS may be described as the selection of a subset of the components of the measurement vector. One might ask the question, "Why not use all of the measurement vector components?" On the surface it would appear that the more features available, the better job one could do. A closer examination of the problem reveals that the computation time goes up substantially as more features are used. Furthermore, Marill and Greene* give examples which show that in some cases fewer features can be more effective than a larger number. The important point to recognize is that trade-offs exist between classification accuracy, computation time and the number of features. The purpose of the feature selection step in multispectral data analysis is to help optimize the trade-off between classification accuracy and computation time. For

*Marill, T., and Greene, D. M., "On the Effectiveness of Receptors in Recognition Systems," IEEE Transcations on Information Theory, January, 1963.

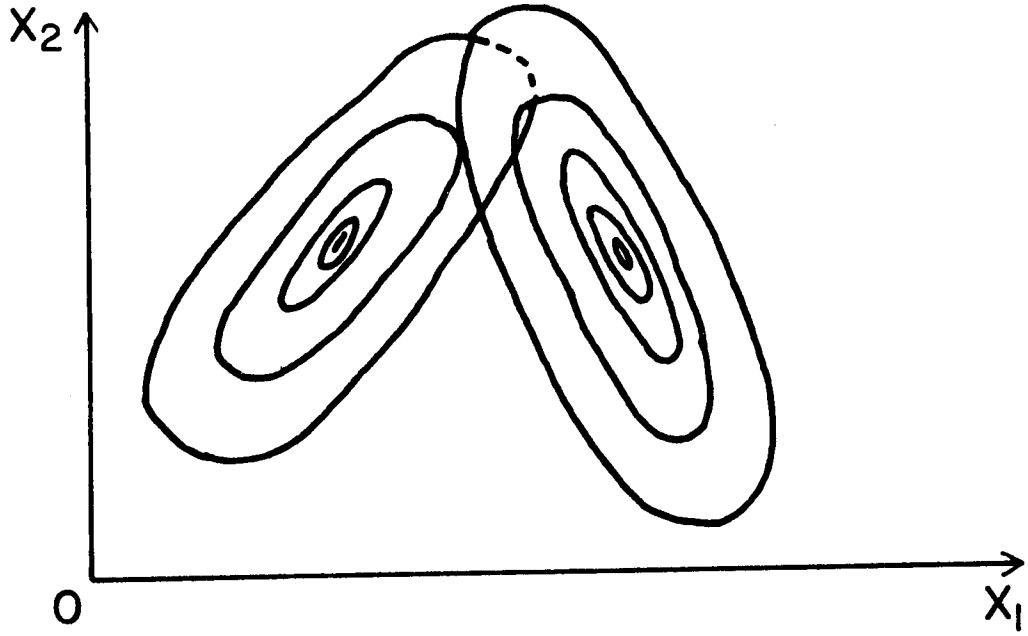
12-channel aircraft scanner data, studies have shown that as few as four or five channels can be used without seriously affecting classification accuracy. Thus, in the analysis of a typical aircraft scanner mission, the feature selection algorithm is used to determine the best subset of four channels.

In order to interpret the output of the feature selection program, one needs an understanding of the concept of "statistical distance." Recall that the classification algorithm is based on representing the classes in terms of multidimensional probability density functions. Two cases of two-dimensional density functions are shown in figure 6-1. It is obvious that the "distance" between the density functions in case b) is greater than in case a). One would also expect that the greater the statistical distance between density functions, the better the classification accuracy. This statement is true, but the functional relationship between accuracy and statistical distance is very complicated (see Swain, 1972 and LARS Information Note 020871, Comparison of the Divergence and β-Distance in Feature Selection by Swain, Robertson and Wacker).

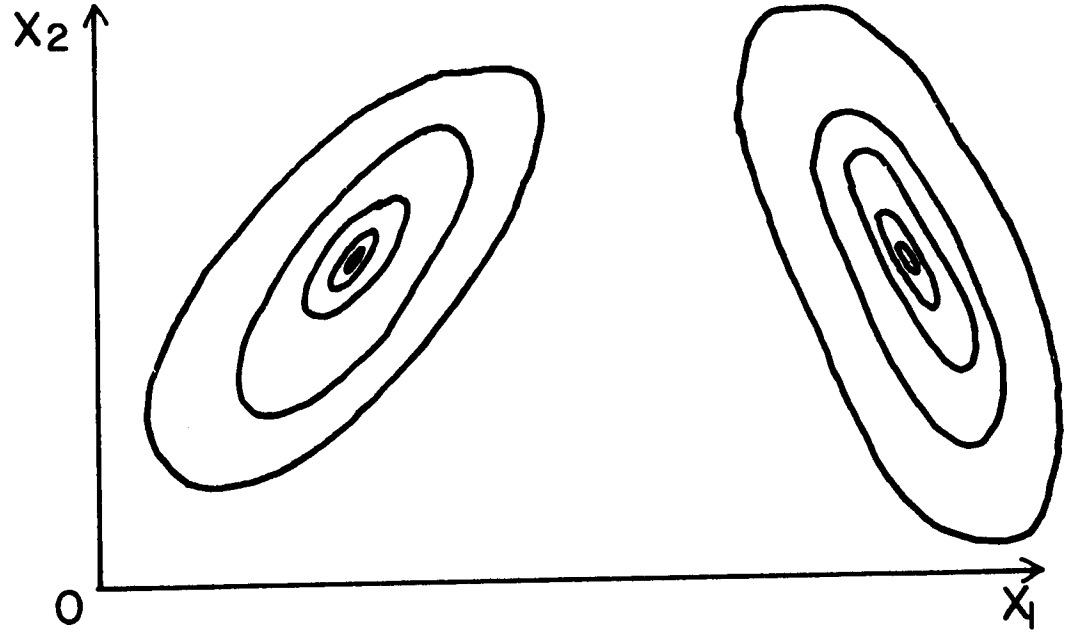
A number of statistical distance measures exist. We need not concern ourselves here with the mathematical definitions but should recognize that the distance between two probability density functions depends not only on the Euclidean distance between the mean values but also on the "spread" of the data. Figure 6-2 illustrates this point. In parts a) and b) of the figure, the Euclidean distance between the mean values are equal, but in part b) the smaller variances result in a larger statistical distance between the two density functions.

The statistical distance concept is defined for a pair of distributions, but remote sensing applications usually involve more than two classes. Two methods of handling this situation are available in LARSYS. The first is to rank the feature subsets in terms of the average distance between all pairs of classes. The program provides the capability of weighting different pairs of classes differently in computing the average distance. This is a useful option because it allows the assignment of priorities according to the need to correctly distinguish certain pairs of classes. To illustrate this point assume a situation in which four training classes have been defined:

Bare soil	BSOIL
Water	WATER
Vegetation 1	VEG1
Vegetation 2	VEG2

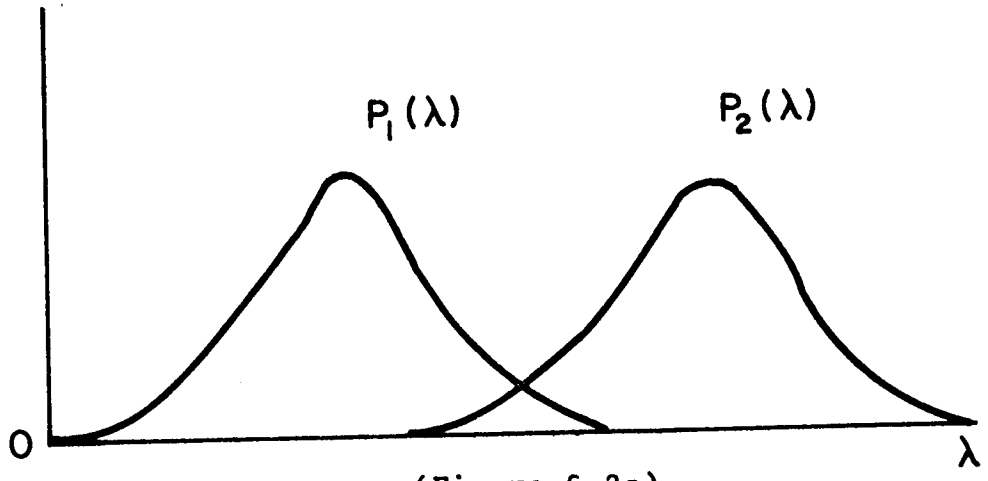


(Figure 6-1a)

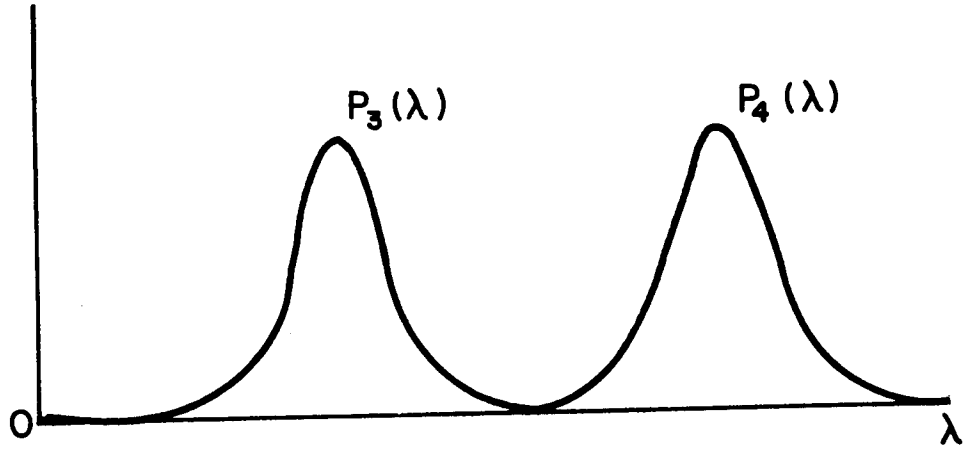


(Figure 6-1b)

Figure 6-1. Density functions which vary in "distance" from each other.



(Figure 6-2a)



(Figure 6-2b)

Figure 6-2. Each pair of the above distribution functions has equidistant means. But the smaller variance in $P_3(\lambda)$ and $P_4(\lambda)$ cause them to have a larger statistical distance.

Further assume that the initial desire was to classify the region into three classes: bare soil, water and green vegetation. Refinement of the training samples by means of the clustering program revealed that two vegetation subclasses existed, perhaps row crops and sown crops. In two-dimensional space the data might look like that shown in figure 6-3. Since the original analysis objective was to map three classes, a large separation between the subclasses VEG1 and VEG2 is unnecessary. A classification mistake between subclasses VEG1 and VEG2 is immaterial since both belong to the same class, green vegetation. In such a case the distance between these subclasses would be given a lower weight. A reasonable weighting scheme between the various subclasses is shown in figure 6-4.

One difficulty with just looking at the average separation is that one large term in the average can overshadow the other terms. As an example look at figure 6-5. Part a) shows a situation which would result in a larger average separation than figure 6-5 b), and yet better overall separation is possible in b). This suggests looking at the minimum pairwise separation as well as the average separation. We shall see in a moment how this may be accomplished with the separability processing functions.

To get a feeling for typical numerical values of separability we'll examine some processing output. The output from a typical run is shown in figure 6-6. Various combinations of four channels are listed along the left. These combinations have been listed according to decreasing order of average separability. The weighting coefficient for the various class pairs are given at the right, in parentheses below the symbols (letters) for each class pair. (In this case, the weight is ten for every pair.)

Notice that the largest separability appearing in the table is 2000. The statistical distance measure employed has the functional form shown in figure 6-7. The program has been written so that the saturation value is 2000. Generally speaking it has been observed that reasonably good classification accuracies will be obtained if the statistical distance is on the order of 1700 or larger.

References

Section 6 (volume 2) of the LARSYS User's Manual gives an extensive description of the SEPARABILITY processing function. Skim this material in order to familiarize yourself with what is covered in these pages. As you proceed through the example and exercises you may wish to study portions of the User's Manual in more detail.

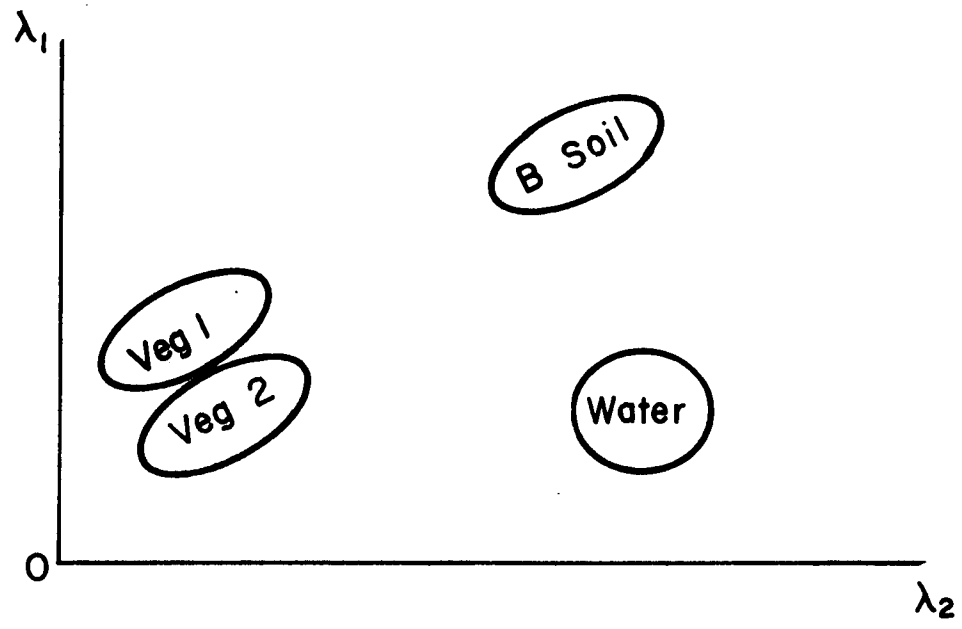
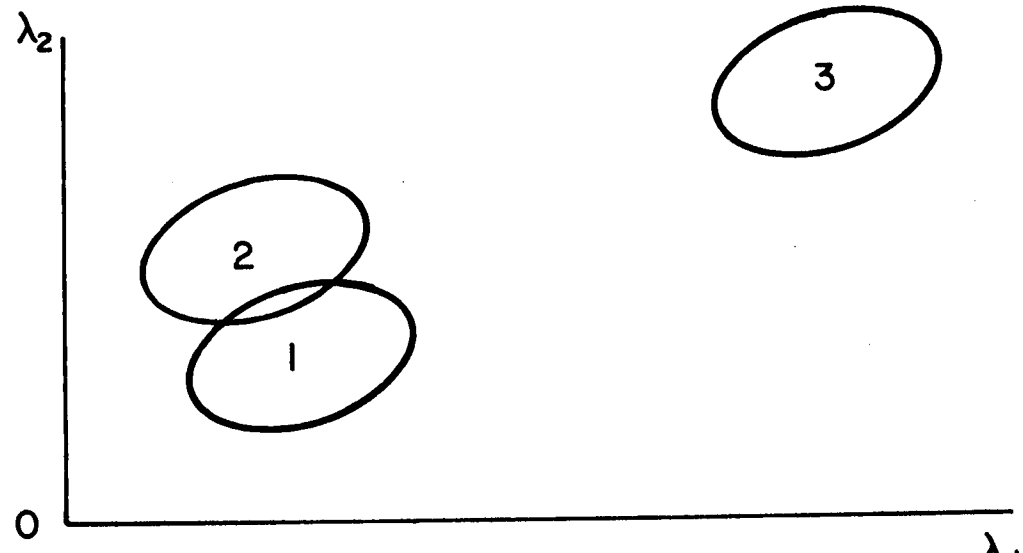


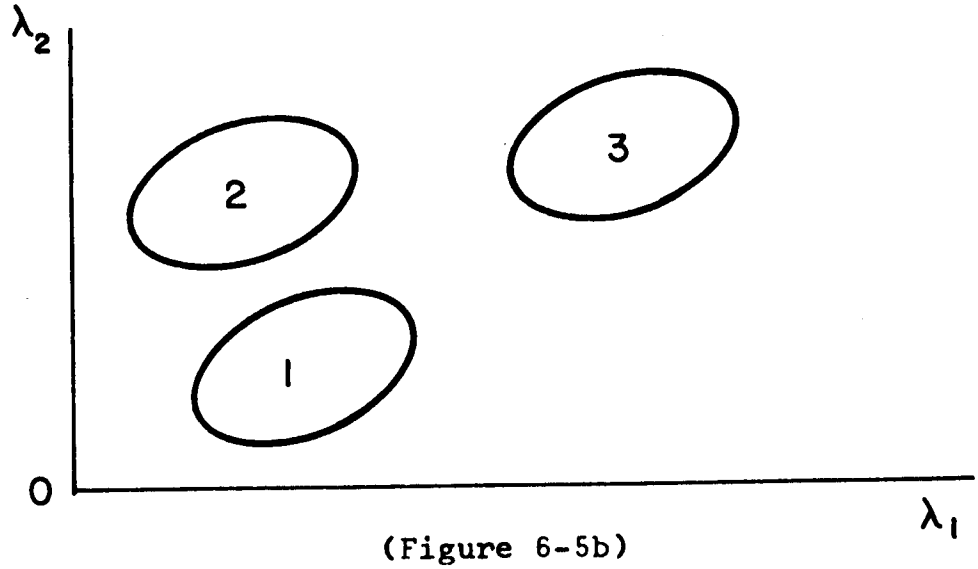
Figure 6-3. Data distributed into four clusters.

	B Soil	Water	Veg 1	Veg 2
B Soil	–	10	10	10
Water	10	–	10	10
Veg 1	10	10	–	0
Veg 2	10	10	0	–

Figure 6-4. Weighting scheme for computing average separability.



(Figure 6-5a)



(Figure 6-5b)

Figure 6-5. Although the average separation is larger in figure 6-5a, the distribution in figure 6-5b is more desirable from the standpoint of separating all three classes.

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

RETENTION LEVEL .. 495 MAXIMUM 30000
MINIMUM 0

DIVERGENCE **WITH** SATURATING TRANSFORM

	CHANNELS	DIJ(MIN)	D(AVE)	WEIGHTED INTERCLASS DIVERGENCE (DIJ)									
				AB (10)	AC (10)	AD (10)	AE (10)	BC (10)	BD (10)	BE (10)	CD (10)	CE (10)	DE (10)
1.	1 6 10 12	1585.	1867.	1978	1996	1986	1822	2000	1595	1585	2000	2000	1709
2.	1 6 10 11	1543.	1863.	1985	1998	1988	1873	2000	1543	1577	2000	2000	1665
3.	1 6 9 11	1456.	1837.	1964	1999	1975	1805	2000	1456	1493	2000	2000	1677
4.	1 6 8 12	1404.	1835.	1982	1999	1985	1813	2000	1502	1404	2000	2000	1668
5.	1 6 8 11	1427.	1835.	1986	1999	1986	1874	2000	1427	1447	2000	2000	1627
6.	2 6 10 12	1405.	1833.	1951	1993	1976	1726	2000	1590	1405	2000	2000	1690
7.	2 6 10 11	1406.	1832.	1957	1996	1973	1776	2000	1554	1406	2000	2000	1661
8.	1 7 10 11	1436.	1832.	1983	1999	1968	1853	2000	1436	1561	2000	2000	1518
9.	1 6 9 12	1428.	1830.	1950	1998	1975	1703	2000	1530	1428	2000	2000	1713
10.	1 7 10 12	1440.	1823.	1977	1996	1960	1797	2000	1440	1568	2000	2000	1496
11.	4 6 10 12	1352.	1817.	1952	1992	1978	1725	2000	1536	1352	1999	2000	1639
12.	4 6 10 11	1346.	1817.	1957	1995	1976	1769	2000	1511	1346	2000	2000	1615
13.	1 5 10 11	1353.	1809.	1978	1997	1964	1840	2000	1390	1571	2000	2000	1353
14.	1 7 9 11	1342.	1807.	1964	1999	1953	1790	2000	1342	1465	2000	2000	1560
15.	3 6 10 12	1304.	1806.	1933	1992	1969	1680	2000	1530	1304	1999	2000	1651
16.	1 6 9 10	1381.	1806.	1965	1940	1976	1769	2000	1381	1606	2000	2000	1420
17.	1 8 10 11	1320.	1805.	1985	1999	1958	1860	2000	1354	1577	2000	2000	1320
18.	3 6 10 11	1295.	1804.	1949	1996	1969	1754	2000	1484	1295	1999	2000	1595
19.	1 4 10 11	1294.	1800.	1979	1994	1967	1855	2000	1357	1554	2000	2000	1294
20.	1 5 10 12	1345.	1799.	1965	1997	1949	1773	2000	1389	1573	2000	2000	1345

Figure 6-6. First page of output showing combinations of four channels ordered by their corresponding average interclass divergence.

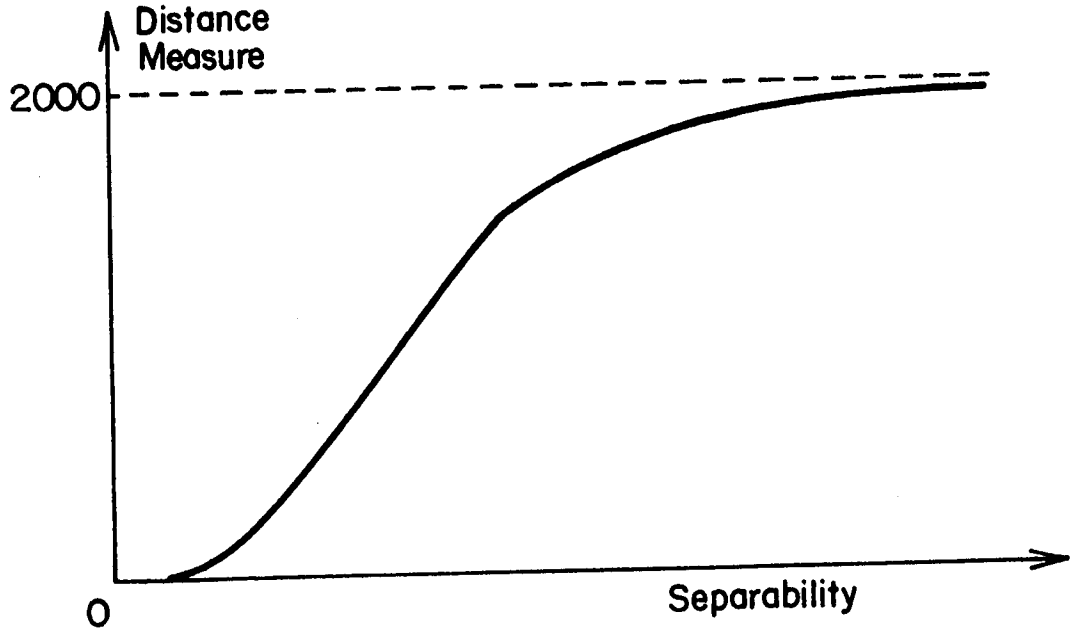


Figure 6-7. The statistical distance measure used in the separability processing function has the functional form shown above.

Example

The analyst was now ready to determine which combination of four channels out of the twelve available would give the best classification results. To do this he used his statistics deck and the SEPARABILITY processing function. The control cards used were:

```
*SEPARABILITY
COMBINATIONS 4
SYMBOLS A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q
WEIGHTS ABC(0), DEFG(0), HIJ(0), KLM(0), NOPQ(0)
CARDS READSTATS
OPTIONS SORT
DATA
   STATISTICS deck with twelve-channel statistics from
   previous STATISTICS run
END
```

The weights card was used to assign zero weights to subclasses of the same basic cover type.

The processor looks at all combinations of four channels out of the possible 12 and lists the top 30 combinations, ordered such that the first combination listed has the largest minimum divergence between classes (since OPTION SORT was specified). Based on the output of this run, a portion of which is shown in figure 6-8, the analyst chose channels 1, 6, 8, 12 to use in the classification program.

EXERCISES

1. Examine figure 1, page SEP-4, volume 2, of the LARSYS User's Manual and select the subset of three channels you would recommend for use with the classification program. Give reasons for your choice.
2. State the general relationship between statistical distance and probability of correct classification.

FLIGHTLINE ANALYSIS CASE STUDY

The statistics deck you obtained from the statistics processor may by used as input to the separability program.

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

RETENTION LEVEL .. 298 MAXIMUM 30000
MINIMUM 0

DIVERGENCE **WITH** SATURATING TRANSFORM

*** RESULTS ORDERED ACCORDING TO DIJ(MIN) ***

	CHANNELS	DIJ(MIN)	D(AVE)	WEIGHTED INTERCLASS DIVERGENCE (DIJ)									
				AB (0)	AC (0)	AD (10)	AE (10)	AF (10)	AG (10)	AH (10)	AI (10)	AJ (10)	AK (10)
1.	1 6 8 12	1476.	1975.			2000	1944	2000	2000	2000	2000	2000	2000
2.	1 6 9 12	1401.	1975.			2000	1942	2000	2000	2000	2000	2000	2000
3.	1 6 8 11	1394.	1976.			2000	1936	2000	2000	2000	2000	2000	2000
4.	1 6 10 12	1379.	1981.			2000	1971	2000	2000	2000	2000	2000	2000
5.	1 6 8 10	1378.	1969.			2000	1972	2000	2000	1995	2000	2000	2000
6.	2 6 9 11	1321.	1967.			2000	1909	2000	2000	2000	2000	2000	2000
7.	2 6 8 11	1294.	1964.			2000	1885	1997	2000	2000	2000	2000	2000
8.	2 6 10 12	1293.	1976.			2000	1925	2000	2000	2000	2000	2000	2000
9.	1 6 9 11	1291.	1976.			2000	1942	2000	2000	2000	2000	2000	2000
10.	1 6 9 10	1286.	1974.			2000	1970	2000	2000	2000	2000	2000	2000
11.	3 6 10 12	1278.	1974.			2000	1927	2000	2000	2000	2000	2000	2000
12.	3 6 10 12	1273.	1973.			2000	1908	2000	2000	2000	2000	2000	2000
13.	4 6 9 11	1251.	1965.			2000	1887	1999	2000	2000	2000	2000	2000
14.	6 8 10 11	1246.	1969.			2000	1941	2000	2000	2000	2000	2000	1999
15.	5 6 8 11	1244.	1963.			2000	1894	1999	2000	2000	2000	2000	2000

Figure 6-8. First page of SEPARABILITY output for example problem.

Write out the control cards and run the SEPARABILITY processing function to find the best combination of four channels for the segment of flightline 210 which you are analyzing. You should note that when you ran the STATISTICS processing functions an order was established for your classes and subclasses. When the statistics deck is used with the separability program, this same class order is preserved. Keep this in mind when preparing the SYMBOLS card.

Section 7

CLASSIFICATION

Instructional Objectives for this Section

After finishing this section and its associated exercises and case study work you should be able to:

a) name two classification algorithms implemented in LARSYS and give at least one distinction between them.
b) carry out a classification analysis (write control card statements, run the processors and interpret the results) when given the statistics deck and output from the feature selection step of the analysis.

Classification

In many respects this step in the analysis is the climactic step. Previous steps have been directed toward obtaining classification results, a substantial achievement in the process of reducing remote sensing data to useful information. It is possible that the first machine classification for a particular analysis task will not be satisfactory. It may be necessary to revise some decisions made in previous steps, perhaps even as far back as the selection of training classes. While this might seem to be a drastic revision, bear in mind that initial training classes are usually based upon what you would like to separate or distinguish. The classification process may reveal that some of the initially chosen classes are not spectrally distinct and that a revised set of classes needs to be defined in order to get maximally useful results. (As your analysis skill improves, you may be able to recognize this kind of difficulty earlier in the analysis sequence - perhaps in the training field and training class refinement step.)

Two classification algorithms are available in LARSYS. They are known as "point classification" and "sample classification." The distinction between these two approaches is illustrated in figure 7-1 for a two-class, two-feature analysis. In the case of point classification, each data point is classified individually. The likelihood function is calculated and the point is assigned to the most likely training class. In sample classification a group of data points (a statistical sample) all assumed to represent the same type of ground cover is classified as a group. Clas-

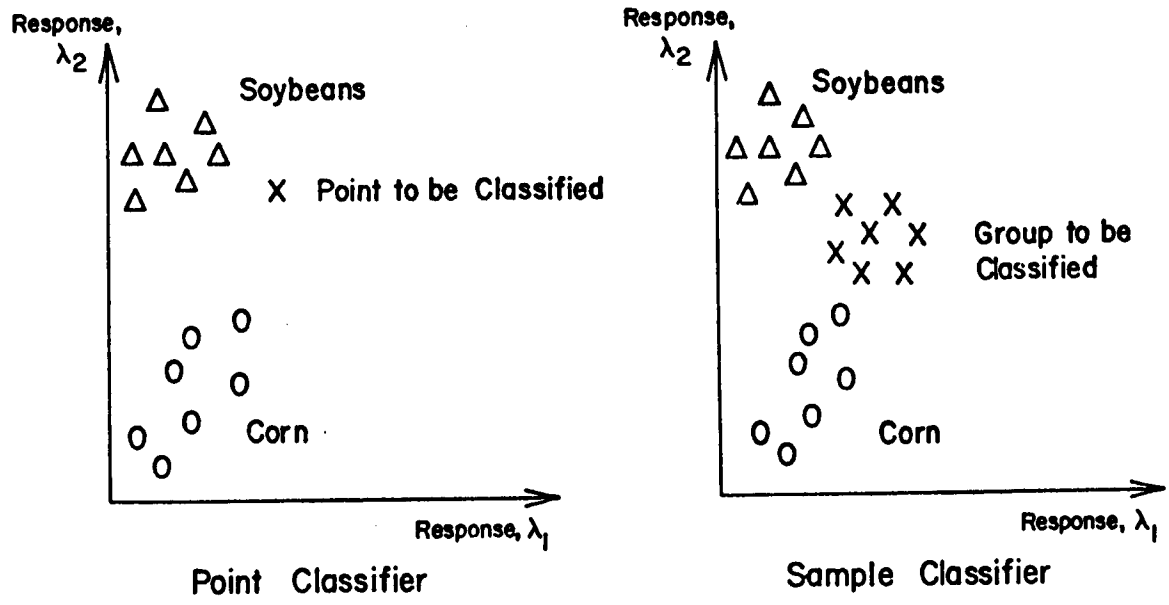


Figure 7-1. Point and sample classification schemes.

sification is based on using the group of data points to estimate the mean vector and covariance matrix of a Gaussian density function associated with the group and comparing the statistical distance between this density function and the density functions of each training class. The group of data points is assigned (classified) to the class which is closest.

Figure 7-2 summarizes the important features of the two classification algorithms which are available in LARSYS. While both algorithms assume that the training classes may be represented by multidimensional Gaussian probability density functions, the philosophy of each approach is quite different.

Both programs require a statistics deck as part of the input to the program. The statistics deck, obtained from a previous step in the analysis, specifies the mean vector and covariance matrix of each training class. Both classification programs also require that the channels to be used for classification be specified. The choice of channels is based on the results of the feature selection step in the analysis.

The area to be classified by the point classifier is specified on one of the input control cards. Test fields are used to estimate the performance of the classifier. You will recall that two sets of fields were specified using ground observation information. One set, designated "training fields," is used to train the classifier. The other set is called "test fields." After classification of the data points, the computer is given additional ground observation information about the test areas. The computer examines and tabulates the classification decisions for each test field and each test class. These tabulated results assist the analyst in assessing the reliability of the classification results. The point classifier also produces a classification map with each class represented by a different symbol.

The areas (fields) to be classified by the sample classifier are supplied to the computer by means of Field Description Cards. When ground observations are available, these same fields can serve as test fields for evaluating the performance of the classifier. Tables are printed to indicate how each field was classified. The sample classifier does not produce a classification map.

Point classification requires the use of two LARSYS processing functions. CLASSIFYPOINTS carries out the classification of each data point in the area specified and stores the results on the tape or disk file. PRINTRESULTS produces a classification map from the results file and tabulates training and test performance. Sample classification is accomplished by using the SAMPLECLASSIFY processing function alone.

	Point Classification	Sample Classification
Control Function Name	*CLASSIFYPOINTS and *PRINTRESULTS	*SAMPLECLASSIFY
Basic Philosophy	Each data point to be classified is compared to the training sample statistics. The data point is assigned to the "most likely" class. Each data point is classified individually.	A group of data points (a statistical sample) to be classified is compared to the training samples of each class. The entire group (sample) is assigned to the class whose statistics "most nearly resemble" the statistics of the sample to be classified.
Assumptions	Each class can be represented by a multidimensional Gaussian probability density function.	Each class can be represented by a multidimensional Gaussian probability density function.
Basic Program Inputs Required	Statistics Deck Channels to be used Area to be classified Test field specification	Statistics deck Channels to be used Test field specification
Output	Classification map Tabulation of training and test fields and/or class performances	Tables showing test and training field performance
Reference	For more precise explanation of the phrases "most likely" and "most nearly resemble" see Swain (LARS Information Note 111572).	

Figure 7-2. Summary of important features of the two LARSYS classification programs.

References

Swain, 1972, gives a deeper treatment of the theory behind both point and sample classification. Point classification is covered in pages 3 through 20; sample classification in pages 36 through 39. Consult this reference for a more detailed discussion of the LARSYS classification algorithms.

Three processing functions CLASSIFYPOINTS, PRINTRESULTS, and SAMPLECLASSIFY are used in the classification step of the analysis. The appropriate parts of section 6 (volume 2) of the LARSYS User's Manual should be consulted as required.

Example

After deciding which four channels to use, the analyst was ready to classify the data. In this example he used point classification, using the CLASSIFYPOINTS and PRINTRESULTS processing functions. The programs were run "back-to-back" with the following deck:

```
*CLASSIFYPOINTS
RESULTS DISK
CARDS READSTATS
CHANNELS 1,6,8,12
DATA
  STAT deck from previous STATISTICS run
DATA
RUN(66000652), LINES(1,950,2), COL(1,222,2)
END
*PRINTRESULTS
RESULTS DISK
PRINT OUTLINE(TRAIN,TEST), TRAIN(F,C), TEST(F,C,P)
SYMBOLS O,O,O,C,C,C,C,W,W,W,S,S,S,G,G,G,G
THRESHOLD 17* 0.1
GROUP OATS(1/1,2,3/), CORN(2/4,5,6,7/), WHEAT(3/8,9,10/)
GROUP SOYB(4/11,12,13/), GRASS(5/14,15,16,17/)
BLOCK RUN(66000652), LINES(1,950,2), COL(1,222,2)
DATA
TEST 1
  (Field Description Cards for Oats test fields.)
TEST 2
  (Field Description Cards for  Corn  test fields.)
  .
  .
  .
TEST 5
  (Field Description Cards for Grass test fields.)
END
```

There is a considerable amount of information given in the output of these processors. Samples follow.

Figure 7-3 is a section of the classified map. Each point has either been classified into one of the five major classes or thresholded (represented by a blank) as being very unlike any of the classes. The test fields and training fields have been outlined. Figure 7-4 tabulates the training field performance; figure 7-5 the test field performance; figure 7-6 the training class performance; and figure 7-7 the test class performance.

Although the results were generally good the analyst examined the weak areas. For example in figure 7-4 the field L6 (corn) was classified only 53% correct: 19 samples were classified as corn, 16 as grass and 1 as oats. The analyst then looked at L6 in the cluster program output. All the points in the field were from the same cluster. The other corn training fields were about 80% correct and all the test fields were above 80% correct. The relatively poor classification result of field L6 led the analyst to check his ground observation data to see if an error had been made in designating L6 as a corn field. No error was found. Had a mistake been made it would have been reasonable to delete L6 from the corn training data.

It is not uncommon to repeat the analysis sequence in order to refine a classification. In general there are several ways one may work to improve performance. One way is to further refine the training class definition by eliminating nonrepresentative, nonessential fields. A field is "nonessential" if, after it is eliminated, all training subclasses are still representated by at least 10n sample points, (where n is the number of channels used in the classification). Another way is to substitute new training fields for ones that are felt to be nonrepresentative. A third would be to add cards to either form new subclasses or give more sample representation to existing ones. In all these cases the STATISTICS, SEPARABILITY, CLASSIFYPOINTS and PRINTRESULTS processing functions would need to be rerun. The second and third approaches would also require use of the clustering algorithm.

FLIGHTLINE ANALYSIS CASE STUDY

1. Using products from previous steps, classify run 71053900 using the point classification algorithm. Set up the control cards so that you will get as output a classification map and tables showing the performance for all fields and classes, both training and test.

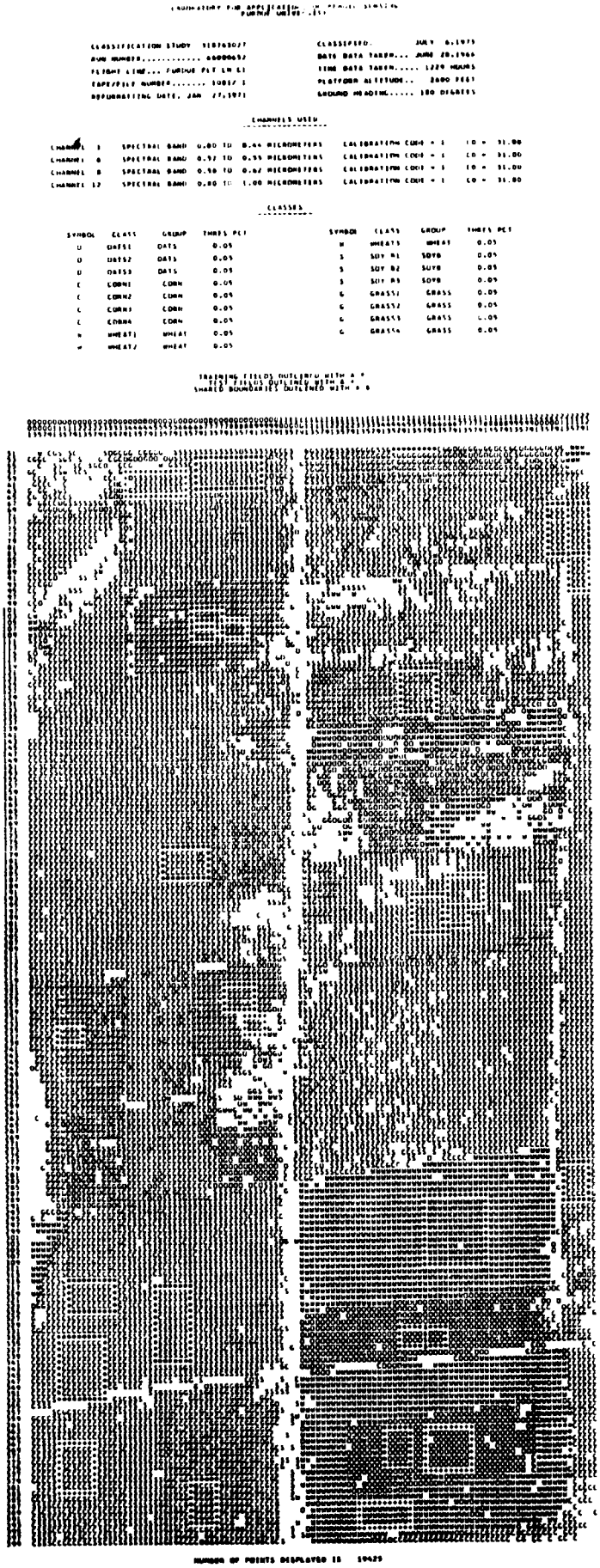


Figure 7-3. A portion of the classification map output.

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

CLASSIFICATION STUDY 324745114 CLASSIFIED. SEPT 4,1973

CHANNELS USED

CHANNEL 1	SPECTRAL BAND	0.40 TO 0.44 MICROMETERS	CALIBRATION CODE = 1	CO =	31.00
CHANNEL 6	SPECTRAL BAND	0.52 TO 0.55 MICROMETERS	CALIBRATION CODE = 1	CO =	31.00
CHANNEL 8	SPECTRAL BAND	0.58 TO 0.62 MICROMETERS	CALIBRATION CODE = 1	CO =	31.00
CHANNEL 12	SPECTRAL BAND	0.80 TO 1.00 MICROMETERS	CALIBRATION CODE = 1	CO =	31.00

CLASSES

	CLASS	GROUP	THRES PCT		CLASS	GROUP	THRES PCT
1	OATS1	OATS	0.10	10	WHEAT3	WHEAT	0.10
2	OATS2	OATS	0.10	11	SOY B1	SOYB	0.10
3	OATS3	OATS	0.10	12	SOY B2	SOYB	0.10
4	CORN1	CORN	0.10	13	SOY B3	SOYB	0.10
5	CORN2	CORN	0.10	14	GRASS1	GRASS	0.10
6	CORN3	CORN	0.10	15	GRASS2	GRASS	0.10
7	CORN4	CORN	0.10	16	GRASS3	GRASS	0.10
8	WHEAT1	WHEAT	0.10	17	GRASS4	GRASS	0.10
9	WHEAT2	WHEAT	0.10				

TRAINING FIELD PERFORMANCE

FIELD DESIG.	GROUP	NO OF SAMPS	PCT CORCT	NUMBER OF SAMPLES CLASSIFIED INTO: OATS	CORN	WHEAT	SOYB	GRASS	THRSHOLD
L50	OATS	8	100.0	8	0	0	0	0	0
L55	OATS	42	100.0	42	0	0	0	0	0
L51	OATS	6	100.0	6	0	0	0	0	0
L53	OATS	20	100.0	20	0	0	0	0	0
L54	OATS	48	95.8	46	0	0	0	2	0
L56	OATS	45	86.7	39	1	0	0	5	0
L5	CORN	49	100.0	0	49	0	0	0	0
L19	CORN	32	100.0	0	32	0	0	0	0
L9	CORN	20	100.0	0	20	0	0	0	0
L60	CORN	12	83.3	0	10	0	0	2	0
L6	CORN	36	52.8	1	19	0	0	16	0
L7	CORN	32	93.8	1	30	0	0	1	0
L10	CORN	32	100.0	0	32	0	0	0	0
L11	CORN	72	98.6	0	71	0	1	0	0
L13	CORN	18	72.2	0	13	0	5	0	0
L20	WHEAT	110	99.1	0	0	109	0	0	1
L22	WHEAT	20	90.0	1	0	18	0	0	1
L27	WHEAT	42	97.6	0	0	41	0	0	1
L25	WHEAT	24	100.0	0	0	24	0	0	0
L26	WHEAT	20	95.0	1	0	19	0	0	0
L28	SOYB	48	95.8	1	1	0	46	0	0
L36	SOYB	66	98.5	0	1	0	65	0	0
L40	SOYB	120	99.2	0	1	0	119	0	0
L33	SOYB	30	96.7	0	0	0	29	1	0
L39	SOYB	8	100.0	0	0	0	8	0	0
L46	GRASS	8	100.0	0	0	0	0	8	0
L49	GRASS	16	93.8	0	1	0	0	15	0
L45	GRASS	15	100.0	0	0	0	0	15	0
L52	GRASS	24	83.3	0	4	0	0	20	0
L43	GRASS	15	93.3	0	0	0	1	14	0
L51	GRASS	14	92.9	1	0	0	0	13	0
L44	GRASS	10	100.0	0	0	0	0	10	0
L50	GRASS	10	100.0	0	0	0	0	10	0
	TOTAL	1072		167	285	211	274	132	3

OVERALL PERFORMANCE(1020/ 1072) = 95.1

Figure 7-4. Training field performance.

CLASSIFICATION STUDY 324745114 CLASSIFIED. SEPT 4,1973

CHANNELS USED

CHANNEL 1	SPECTRAL BAND	0.40 TO	0.44 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 6	SPECTRAL BAND	0.52 TO	0.55 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 8	SPECTRAL BAND	0.58 TO	0.62 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 12	SPECTRAL BAND	0.80 TO	1.00 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00

CLASSES

	CLASS	GROUP	THRES PCT
1	OATS1	OATS	0.10
2	OATS2	OATS	0.10
3	OATS3	OATS	0.10
4	CORN1	CORN	0.10
5	CORN2	CORN	0.10
6	CORN3	CORN	0.10
7	CORN4	CORN	0.10
8	WHEAT1	WHEAT	0.10
9	WHEAT2	WHEAT	0.10
10	WHEAT3	WHEAT	0.10
11	SOY B1	SOYB	0.10
12	SOY B2	SOYB	0.10
13	SOY B3	SOYB	0.10
14	GRASS1	GRASS	0.10
15	GRASS2	GRASS	0.10
16	GRASS3	GRASS	0.10
17	GRASS4	GRASS	0.10

TEST FIELD PERFORMANCE

FIELD DESIG.	GROUP	NO OF SAMPS	PCT. CORCT	NUMBER OF SAMPLES CLASSIFIED INTO: OATS	CORN	WHEAT	SOYB	GRASS	THRSHOLD
T2	OATS	48	97.9	47	0	0	0	0	1
T3	OATS	18	100.0	18	0	0	0	0	0
T5	CORN	39	100.0	0	39	0	0	0	0
T16	CORN	12	100.0	0	12	0	0	0	0
T12	CORN	42	85.7	0	36	0	6	0	0
T23	WHEAT	27	100.0	0	0	27	0	0	0
T27	WHEAT	42	100.0	0	0	42	0	0	0
T29	SOYB	33	90.9	1	0	0	30	1	1
T33	SOYB	24	95.8	0	0	0	23	0	1
T58	GRASS	15	80.0	0	3	0	0	12	0
T45	GRASS	6	100.0	0	0	0	0	6	0
T44	GRASS	10	100.0	0	0	0	0	10	0
	TOTAL	316		66	90	69	59	29	3

OVERALL PERFORMANCE(302/ 316) = 95.6

Figure 7-5. Test field performance.

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

CLASSIFICATION STUDY 324745114 CLASSIFIED. SEPT 4,1973

CHANNELS USED

CHANNEL 1	SPECTRAL BAND	0.40 TO	0.44 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 6	SPECTRAL BAND	0.52 TO	0.55 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 8	SPECTRAL BAND	0.58 TO	0.62 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 12	SPECTRAL BAND	0.80 TO	1.00 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00

CLASSES

	CLASS	GROUP	THRES PCT		CLASS	GROUP	THRES PCT
1	OATS1	OATS	0.10	10	WHEAT3	WHEAT	0.10
2	OATS2	OATS	0.10	11	SOY B1	SOYB	0.10
3	OATS3	OATS	0.10	12	SOY B2	SOYB	0.10
4	CORN1	CORN	0.10	13	SOY B3	SOYB	0.10
5	CORN2	CORN	0.10	14	GRASS1	GRASS	0.10
6	CORN3	CORN	0.10	15	GRASS2	GRASS	0.10
7	CORN4	CORN	0.10	16	GRASS3	GRASS	0.10
8	WHEAT1	WHEAT	0.10	17	GRASS4	GRASS	0.10
9	WHEAT2	WHEAT	0.10				

TRAINING CLASS PERFORMANCE

	GROUP	NO OF SAMPS	PCT. CORCT	NUMBER OF SAMPLES CLASSIFIED INTO OATS	CORN	WHEAT	SOYB	GRASS	THRSHOLD
1	OATS	169	95.3	161	1	0	0	7	0
2	CORN	303	91.1	2	276	0	6	19	0
3	WHEAT	216	97.7	2	0	211	0	0	3
4	SOYB	272	98.2	1	3	0	267	1	0
5	GRASS	112	93.8	1	5	0	1	105	0
	TOTAL	1072		167	285	211	274	132	3

OVERALL PERFORMANCE(1020/ 1072) = 95.1

AVERAGE PERFORMANCE BY CLASS(476.0/ 5) = 95.2

Figure 7-6. Training class performance.

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

CLASSIFICATION STUDY 324745114 CLASSIFIED. SEPT 4,1973

CHANNELS USED

CHANNEL 1	SPECTRAL BAND	0.40 TO	0.44 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 6	SPECTRAL BAND	0.52 TO	0.55 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 8	SPECTRAL BAND	0.58 TO	0.62 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00
CHANNEL 12	SPECTRAL BAND	0.80 TO	1.00 MICROMETERS	CALIBRATION CODE = 1	CO = 31.00

CLASSES

	CLASS	GROUP	THRES PCT		CLASS	GROUP	THRES PCT
1	OATS1	OATS	0.10	10	WHEAT3	WHEAT	0.10
2	OATS2	OATS	0.10	11	SOY B1	SOYB	0.10
3	OATS3	OATS	0.10	12	SOY B2	SOYB	0.10
4	CORN1	CORN	0.10	13	SOY B3	SOYB	0.10
5	CORN2	CORN	0.10	14	GRASS1	GRASS	0.10
6	CORN3	CORN	0.10	15	GRASS2	GRASS	0.10
7	CORN4	CORN	0.10	16	GRASS3	GRASS	0.10
8	WHEAT1	WHEAT	0.10	17	GRASS4	GRASS	0.10
9	WHEAT2	WHEAT	0.10				

TEST CLASS PERFORMANCE

	GROUP	NO OF SAMPS	PCT. CORCT	NUMBER OF SAMPLES CLASSIFIED INTO: OATS	CORN	WHEAT	SOYB	GRASS	THRSHOLD
1	OATS	66	98.5	65	0	0	0	0	1
2	CORN	93	93.5	0	87	0	6	0	0
3	WHEAT	69	100.0	0	0	69	0	0	0
4	SOYB	57	93.0	1	0	0	53	1	2
5	GRASS	31	90.3	0	3	0	0	28	0
	TOTAL	316		66	90	69	59	29	3

OVERALL PERFORMANCE(302/ 316) = 95.6

AVERAGE PERFORMANCE BY CLASS(475.3/ 51 = 95.1

Figure 7-7. Test class performance.

2. Use the sample classifier to classify the same set of test fields as used in 1 above.

Section 8

INFORMATION EXTRACTION - ANALYZING THE RESULTS

Instructional Objectives for this Section

Upon completion of this section you should be able to list at least three types of information that can be extracted from the results of a classification analysis. You should also, by studying some of the references listed, be able to gain insights into information extraction techniques for particular application areas.

Information Extraction - Results Analysis

The last, and in some respects the most important step, is results analysis. What useful information can be extracted from the classification program output? The success of this final step is, at this point in time, very much dependent upon the background and training of the analyst. While a soil scientist may be able to extract useful soils mapping information from a multispectral data classification map, he is not likely to be expert at deriving watershed management information. Similarily a geologist analyzing multispectral data is not likely to be proficient in extracting crop yield information.

It is important to emphazise the point that classification results are seldom an end in themselves. Their usefulness is primarily dependent on whether or not the analyst can extract useful information from them. As research continues it is expected that some information extraction techniques will lend themselves to machine implementation.

Insight into information extraction and results analysis for some specific applications may be obtained by reading remote sensing articles in the journals listed below.

References

Examples of results analysis and the extraction of useful information from multispectral data classifications may be found in journals such as:

Remote Sensing of the Environment
IEEE Transactions on Geoscience Electronics
Remote Sensing in Ecology
Journal of Soil and Water Conservation
Photogrammetric Engineering
Agronomy Journal
Applied Optics

as well as in a number of LARS Information Notes, published proceedings of remote sensing conferences, etc.

EXERCISES

Check with your instructor on the availability of the above references at your location. Skim through one or more of these references.

FLIGHTLINE ANALYSIS CASE STUDY

Study your classification analysis results. What information can you glean from the results? Based on your results, are the cover type classes you initially selected sufficiently distinct spectrally to provide adequate classification accuracy? Would you consider it worthwhile to use these classes as the basis for a "real life" application of remote sensing?