LARS Information Note 012376

NASA-CR- 147663

(NASA-CR-147663) A FORESTRY APPLICATION SIMULATION OF MAN-MACHINE TECHNIQUES FOR ANALYZING REMOTELY SENSED DATA (PURDUE UNIV.) 75 P HC $4.50 CSCL 05B

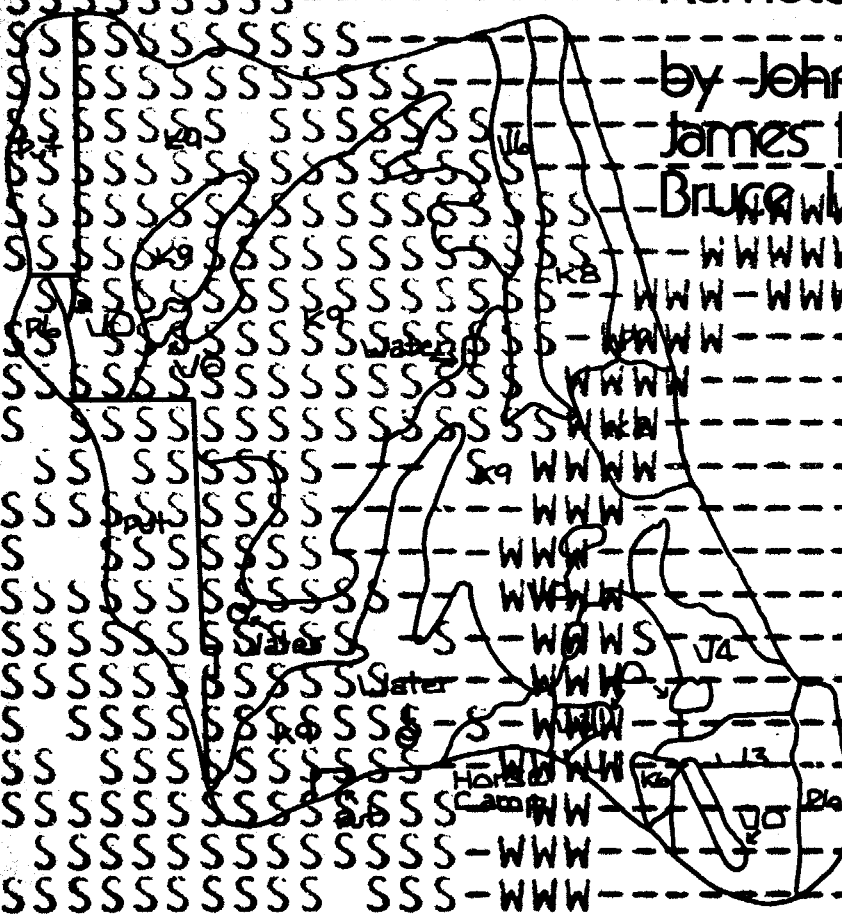
N76-25627

UNCLAS

G3/43 28268

A Forestry Application Simulation of Man-Machine Techniques for Analyzing Remotely Sensed Data

by John Berkebile, James Russell & Bruce Lube



The Laboratory for Applications of Remote Sensing

Purdue University — West Lafayette, Indiana

1976

LARS Information Note 012376

A FORESTRY APPLICATION SIMULATION OF MAN-MACHINE TECHNIQUES FOR ANALYZING REMOTELY SENSED DATA

by

John Berkebile, Prin. Invest?

James Russell and

Bruce Lube

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

This work was supported by the National Aeronautics and Space Administration under contract NAS9-14016.

T-1039/4

A FORESTRY APPLICATION SIMULATION OF MAN-MACHINE TECHNIQUES FOR ANALYZING REMOTELY SENSED DATA

by John Berkebile, James Russell and Bruce Lube

This publication is designed as a simulation to carry you through the typical steps in the analysis of remotely-sensed data for a forestry applications example. The example uses numerically-oriented pattern recognition techniques and emphasizes the man-machine interaction.

PREREQUISITES

The intended audience for this simulation is persons who have experience in forestry and a basic background in remote sensing. The remote sensing background can be gained by means of the following educational materials <u>or their equivalent</u>:

<u>LARSYS Educational Package</u>

Unit I An Introduction to Quantitative Remote Sensing

Unit II LARSYS Software System: An Overview

<u>Fundamentals of Remote Sensing Minicourse Series</u>

Remote Sensing: What is it?

The Physical Basis of Remote Sensing

Applications of Remote Sensing in Forestry

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

PREFACE

The purpose of this simulation is NOT to train you to be an analyst, but instead to give you an overview and understanding of how forestry data are analyzed. Our purpose here is analogous to teaching you how your automobile operates, but not teaching you how to repair it.

It should be pointed out that the experience of the analyst is a very important factor in the man-machine interaction described in this simulation. John Berkebile who generated this analysis has had a forestry background and over 2 1/2 years of experience with computer-aided analysis of multispectral data.

GENERAL OBJECTIVE

Upon completion of this simulation, you should be able to describe the sequential process of analyzing remotely-sensed forestry data using numerical analysis techniques. Your description should include the nature of the interaction between man (analyst) and machine (computer), and the product (results) of each step in the process.

ACKNOWLEDGEMENTS

The authors wish is to thank Forrest E. Goodrick, Professional Forester, Richard P. Mroczynski, Manager of Forestry Applications Research at LARS, and John C. Lindenlaub, Program Leader for Technology Transfer at LARS, who served as a committee of subject-matter advisors and provided valuable input for this simulation. Numerous others who reviewed rough drafts of this material made important comments and suggestions which were incorporated into the simulation. A special thanks goes to Tina K. Cary and John C. Lindenlaub whose Case Study Using LARSYS for Analysis of LANDSAT Data served as a model for this publication.

TABLE OF CONTENTS

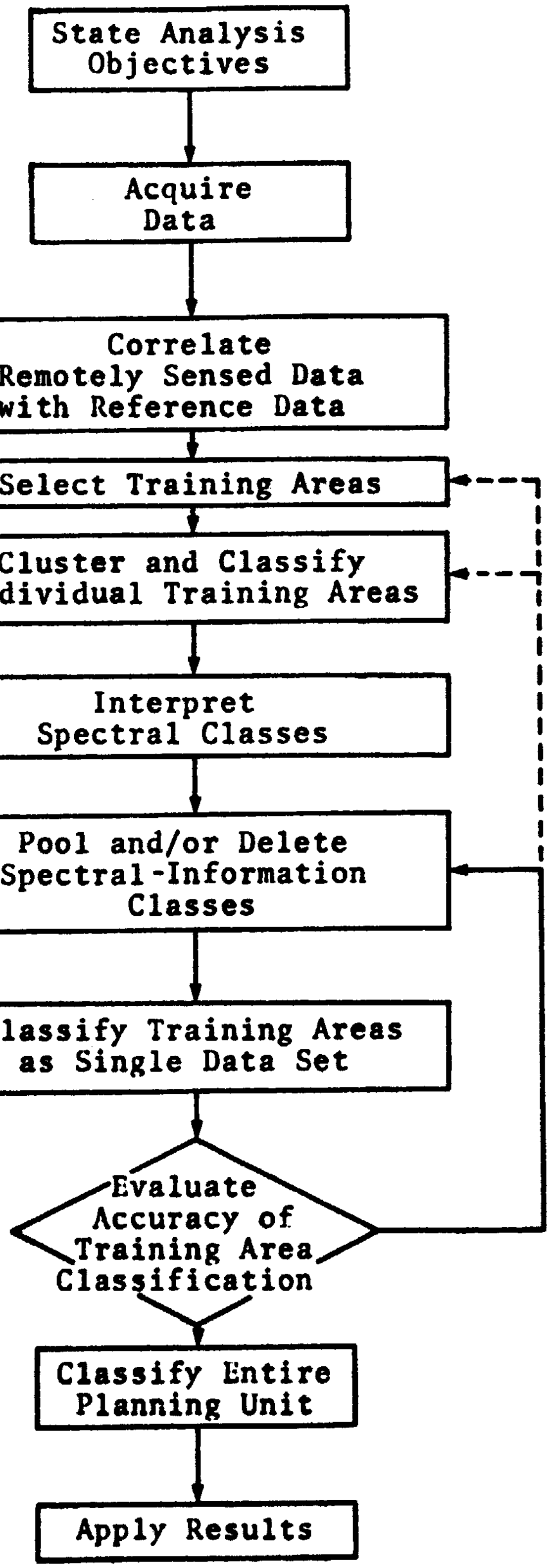


Figure 1. Typical Numerical Analysis Flowchart for Forestry Applications

OVERVIEW

The numerical analysis of remotely sensed data is a dynamic process which requires an interaction between man (analyst) and machine (computer). The process is both an art and a science, relying upon judgements and insights by the analyst as well as a documented technology of remote sensing analysis. A typical analysis sequence is shown in Figure 1, facing page. Even though it is shown here as basically a linear process, all of the steps are interconnected. At any step in the analysis, interpretation of the results of that step can lead the analyst to conclude that he should go back to a previous step and revise his procedure. For simplicity only the most commonly followed analysis sequence is shown.

Remote sensing techniques allow you to "survey" large areas with a minimum amount of time and cost. The computer can be "trained" to produce general land use maps as well as general forest cover maps. Even finer breakdowns of cover types may be achieved, such as timber stand maps, although mapping reliability is lower for these relative to general land-use maps. Foresters have used computer-aided analysis of multispectral scanner data to delineate the areal extent and boundaries of recent forest fires (Hitchcock and Hoffer, 1975)[1], to map detailed timber categories like aspen, ponderosa pine and Douglas fir (Fleming, et. al., 1974), to detect tree stress (Heller and Alrich, 1969), and many other applications.

The first step is to state the analysis objectives. To do this, you must determine the geographic area of interest, the general cover types and the nature of the application to which the results will be applied. An additional component which is often included in the analysis objective is the expected classification accuracy for initial estimates of timber resources. An example would be to "determine the percentage of Hoosier National Forest in each of these cover types: conifers, hardwoods and other with 85% accuracy".

Next, the remotely-sensed data are correlated with the available reference data. The multispectral-scanner data may be from aircraft or satellites, such as LANDSAT. The reference data might include USGS topographic maps (quad maps), stand compartment type maps and related information, aerial photographs, U.S. Forest Service land-use maps and actual ground observations. Each LANDSAT satellite covers the entire earth every eighteen days, so the analyst can most generally choose the time of the year most suitable for mapping the cover types of interest. The analysis sequence described in this simulation uses LANDSAT data.

---

[1]See Bibliography in Appendix C, page 73.

The training areas are then selected. The training areas contain typical examples of each cover type of interest and are supplied to the computer in order to "train" it to classify unknown data points. There are some general selection criteria to aid the analyst in choosing training areas, but successful training area selection relies heavily on the analyst's previous experience and knowledge of the areas being studied.

When training areas have been selected, the next step is to use a computer processor (algorithm) called CLUSTER on each of the training areas individually. The CLUSTER processor uses information from more than one channel, or wavelength band, to produce a single computer-generated image. Since more information is used, the boundaries of ground features and cover types are usually more distinct on images produced by the clustering process than on a single-channel image.

After clustering, obtaining statistics, and classifying each of the training areas, the anlayst looks at the output to see what each spectral class of the training areas represents. The spectral classes are groups of data points with similar spectral values (brightness levels). Aerial photographs and other reference data aid the analyst in making these associations between spectral classes and various cover types.

On the basis of the spectral separabilities and the known cover type information, the spectral-information classes may be pooled (merged together) or deleted. The spectral classes that are informationally and numerically similar (i.e. spectrally inseparable) are combined, while the spectral classes that are a mixture of two cover types (such as pasture and forest) may be deleted. The analyst should go back to his analysis objective(s) to help him decide which classes to combine and which to delete.

To check how well he did in the pooling and deleting of spectral classes, the analyst then classifies all the training areas together as a single unit. He then looks at the classification maps and compares them with other reference data. This step along with the output of the computer allows him to predict the probable accuracy to expect when he classifies the total planning unit.[2]

With the output from classifying the training areas as a single data set, the analyst must predict if the training areas selected are going to allow him to meet his objective(s) when he classifies the total area under consideration. Will the classification yield the stated accuracy? Are all cover types

---

[2] Each National Forest is divided into several planning units which exhibit uniformity of elements considered to be locally important to resource production or protection. Such units usually vary between 40,000 and 150,000 acres in size.

adequately represented? If not, he must go back to previous steps as shown by the arrows on the right side of the flowchart in Figure 1. If possible, he may merely go back to "pool and/or delete" again. In some cases, he might go back and reselect the training areas. He may even need to go back to the beginning and restate his analysis objectives.

When he is satisfied with the classification data from the combined training areas, the analyst instructs the computer to classify the total area. Using pattern recognition algorithms, the spectral responses of each data point are "compared" to the training sample for each class, and the point is assigned to the "most likely" or most similar class. The output after this step can be maps and data tables showing acreages (hectares) for the mapped cover types.

As indicated earlier, numerical analysis of multispectral scanner data is a <u>dynamic</u> process with each step providing feedback to the previous step. For simplicity, the process is shown here as a linear sequence. In reality, the analyst has all steps in mind before he actually begins an analysis. He may also refer back to previous steps and modify his procedure as the analysis continues.

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

Figure 2 A forester in the initial stage, writing his analysis objectives for a particular planning unit.

SECTION I STATING ANALYSIS OBJECTIVES

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

1. List the four usual components of an analysis objective.

2. Write an analysis objective of your choice incorporating these four components.

The first and one of the most important steps in the numerical analysis process is stating the analysis objective(s). What is the purpose of using the remotely-sensed data? What are you interested in doing? What information do you need?

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

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

The output will aid in deciding which stands are top priority for intensive timber mangement. If he's interested in watershed planning, the forester might want to:

> "Generate type maps which aid watershed management in the Wenatche National Forest by giving better estimates of water runoff and sedimentation rates."

This information is valuable in locating and protecting impoundments and wetlands. Providing the proper environment for wildlife might be the forester's concern:

> "Locate suitable cover types and wetlands (habitat diversity) for wildlife species in the Superior National Forest as an aid in managing wildlife openings and waterholes."

Forest fire protection might be aided if one can:

> "Produce a potential fire hazard map which integrates the slope and aspect in the Sequoia National Forest to assess the rate a fire will spread (fuel type and quantity) and the proportion of the drier aspects present with 80% accuracy.

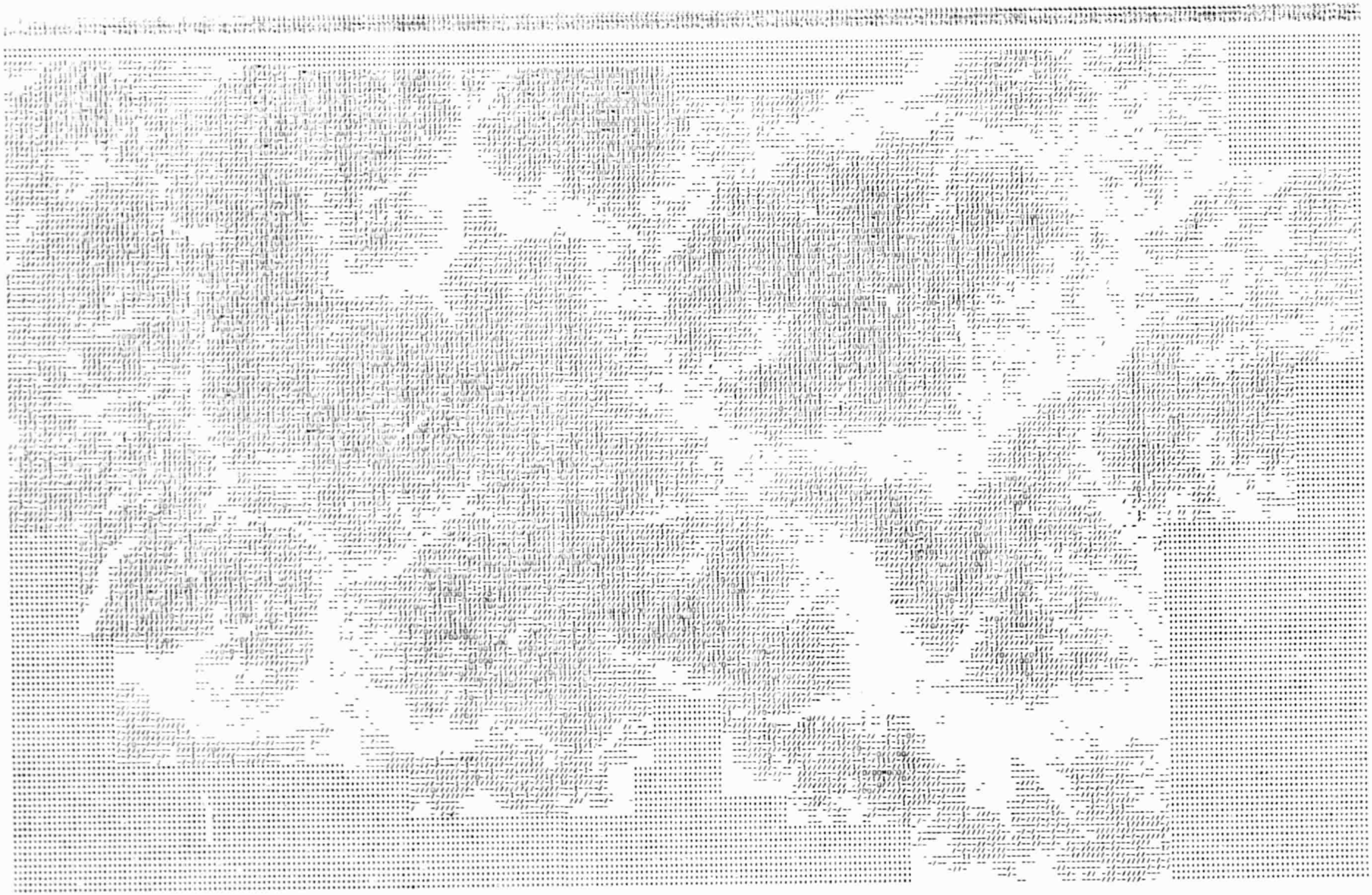


Figure 3 A grayscale printout of LANDSAT data showing the Brownstown District of the Hoosier National Forest.

This information will aid in making decisions as to crew size and placement during the fire season.

The essential components of an analysis objective are:

Location What portion of the earth's surface is of interest? It may be a relatively small area (several hundred acres using airborne multispectral scanners) or a relatively large area (thousands or millions of acres using multispectral scanner data from satellite-borne systems).

Cover Types What types of ground cover are of interest to you? Only forest types such as deciduous, conifers, or brush land? Or are you interested in water, agriculture, pasture, barren land and snow cover?

Applications How will the analysis output be used? To predict volume of growth? Extent of disease? A cover type map for acquistion planning? Fire control planning? Or recreational and wildlife potential?

Classification Accuracy How accurate must the classification be in order to be of help to you? Is 65% close enough or do you need to have approximately 90% accuracy? The level of accuracy will depend upon the level of mapping detail, time of the year data were collected, analyst training and skill, particular region being mapped, and other variables.

There are often two levels of objectives. The management objectives are more general and provide the overall purpose of the forester's study of a given area. These objectives primarily have to do with management decisions that affect the resource base and guide our use of the land. The ultimate goal is to achieve the proper balance of the following multiple uses: (1) timber, (2) water, (3) wildlife, (4) forage and (5) recreation.

"Locate fire hazard areas such as timber sale areas, brushland, and broomsedge fields and non-burnable areas such as cropland and pasture in order to make decisions concerning fire crew size and placement during the fire season."

The analysis objectives are more specific and express the extent of the information and data needed for a specific project. Here are two analysis objectives which will be used in this simulation:

"Produce a detailed classification of the Brownstown District of the Hoosier National Forest[1] using computer-assisted analysis of LANDSAT-1 data. The cover types to be mapped are: various water classes, pasture, cropland, brushland, slope and bottomland deciduous, ridgetop deciduous, sparse deciduous, conifer and cultural (urban-suburban)."

"Produce a general land use classification of the same area using the following classes: water, pasture, cropland, brushland, conifer, hardwood and cultural."

Self Check[2]

I-A. Name the four components of an analysis objective.

I.B. Write an analysis objective that would be useable for you in solving a forestry problem.

---

[1]A grayscale of LANDSAT data from the Brownstown District of the Hoosier National Forest is shown in Figure 3, page 6.

[2]Answers to all Self-Checks are given in Appendix B.

SECTION II ACQUIRE DATA

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

1. Identify two sources of multispectral scanner data for forestry analysis.
2. State the importance of high quality multispectral scanner data.
3. Describe three data idiosyncrasies which might hinder analysis.

Once the analysis objectives have been stated, you must acquire the remotely-sensed data which will be used for numerical analysis to meet your objective(s). There are basically two general types of multispectral data -- satellite and aircraft. The data distribution center for LANDSAT data is the EROS Data Center in Souix Falls, South Dakota. The data are available in the following formats:

Photo products such as
70mm negatives and positives,
9 x 9 negatives, positives and color composites
various size prints in black & white and color
Computer Tapes

If aircraft data are desirable, there are companies such as Environmental Research Institute of Michigan (ERIM), Park Aerial Surveys, Photographic Surveys, Mark Hurd Aerial Surveys, and Aero Service which contract to provide such data. Aircraft data, though generally more expensive than satellite data, provide the advantage of being able to select a specific data format (type and scale) at a time when there are no clouds to interfere with the data collection over the area of interest.

In all cases the type of data desired and the time of the year for data collection are determined on the basis of the stated analysis objectives. For example, when classifying certain types of ground cover or different timber types, you may find that they are most distinct (spectrally) at specific times of the year. For many areas, you would not attempt to classify deciduous forest cover types during the winter months.

When acquiring data, you must be concerned about data quality. A preliminary evaluation of digital data can be made by inspecting imagery created from the data tapes. This type of imagery can be obtained from the data distribution centers, such as EROS, which supply digital data tapes and images.

Figure 4 A computer operator mounting a digital tape containing satellite gathered, multispectral data covering the entire planning unit.

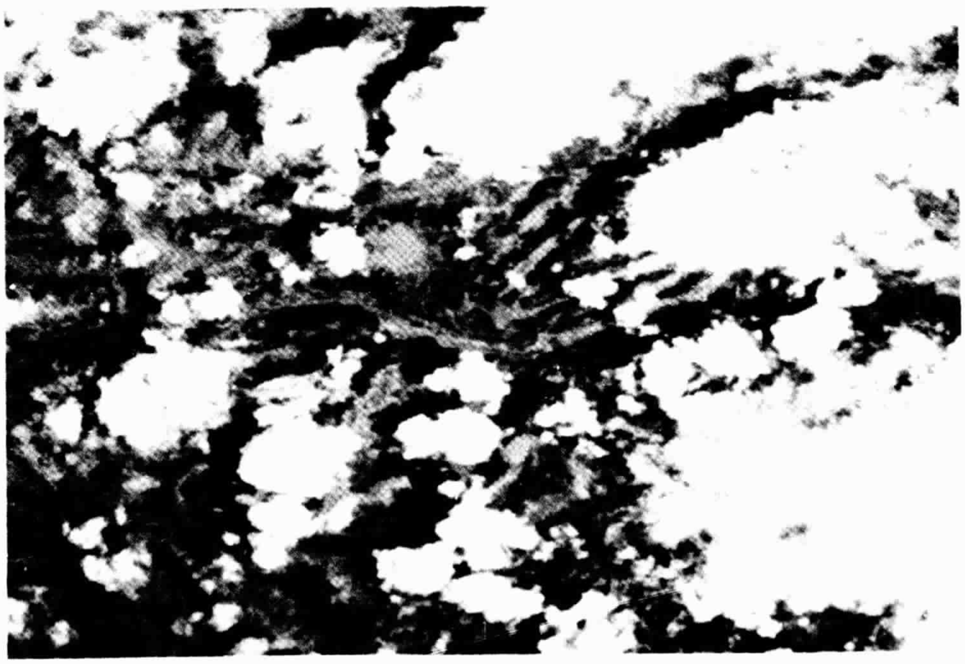


Figure 5 An example of clouds and their shadows.

Gross data characteristics, including haze, cloud cover and snow cover, will be apparent in digital display images or grayscale printouts. See Figure 5. Clouds can significantly decrease the usefulness of a data set. The presence of snow can be a limitation in data analysis if the cover types of interest are under the snow.

Sometimes "striping" will occur in the image. This undesirable effect is due to a defect in the scanner system. In the LANDSAT scanner system, six lines are scanned in each wavelength band every time the mirror oscillates. A separate set of detectors is used for each of these scan lines. If these detectors and their electronics are not properly matched or calibrated, the striping effect is noticeable in the imagery. A dramatic example is shown in Figure 6.

Figure 6 Striping effect in imagery.

The following table shows the mean and standard deviation for the output of each six detectors in channel 1 over the whole frame.

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

Notice that the mean value for detector 3 is very low compared to that of the other detectors. Apparently a malfunction accurred in the detector electronics, resulting in the striping shown in Figure 6. Sometimes a bad data line may go across the image just once (one stripe). It is caused by data collection, transmission or data processing errors but affects only one or even just part of a data line. These are just a few examples of the data quality problems which may face the analyst when he uses remotely-sensed data.

Self-Check

II-A. Name two sources of multispectral scanner data available for forestry analysis.

II-B. State why high quality multispectral scanner data is important to an analyst.

II-C. List three data quality problems which could cause problems for the analyst.

Figure 7 Part of an infrared aerial photograph taken at approximately 60,000 feet showing a portion of the Monroe reservoir, the shoreline and part of the Hoosier National Forest. The white patch in the upper left corner is a cloud. The area inside the lines is a training area of interest in this simulation.

SECTION III CORRELATE REMOTELY SENSED DATA WITH REFERENCE DATA

Upon completing this section you should be able to:

1. Name three types of reference data that can be correlated with remotely-sensed multispectral data.

2. Select one or more of the reference data types to meet your analysis objective(s) stated in the previous section.

Data collected by multispectral scanners are either stored on tape or transmitted to the ground by telemetry. This is "raw" numerical information which can be used to train the computer to recognize certain data values as specific cover types which are of interest. There are several tools available to assist the analyst in accurately training the computer.

One aid that is extremely useful, if available, is aerial photography. Four types of film are commonly used: Black-and-white, black-and-white infrared, color and color infrared. Each type of film has unique characteristics which would make it more or less useful. Color infrared is often preferred for data collection at high altitudes because of its haze penetration quality. Both infrared films are useful for enhancing differences between vegetation types. Photographs are usually taken at altitudes below 60,000 ft. (See Figure 7) Naturally, the nearer the earth's surface the better the scene resolution, but the smaller the area covered by each photographic frame. A scale compromise should be made at this time producing sufficient detail for complete interpretation, but not so much as to make correlation with LANDSAT data difficult (i.e., it would take too many photographic frames to cover a large area).

A second aid which can provide valuable information are USGS quadrangle maps, especially those containing U.S.F.S. data such as compartment maps. If the numerical output and the quad maps are similar in scale, a simple overlay technique can be used. (See Figures 8-10, pages 17-19) This process involves placing the computer printout of the area containing the cover type(s) of interest on top of the quad map and penciling in approximate borders for the training areas.

A third type of reference data involves direct ground observation by someone trained to observe the proper ground features which are of interest to the analyst. A fourth is type maps produced by a combination of the first three previously mentioned reference data materials.

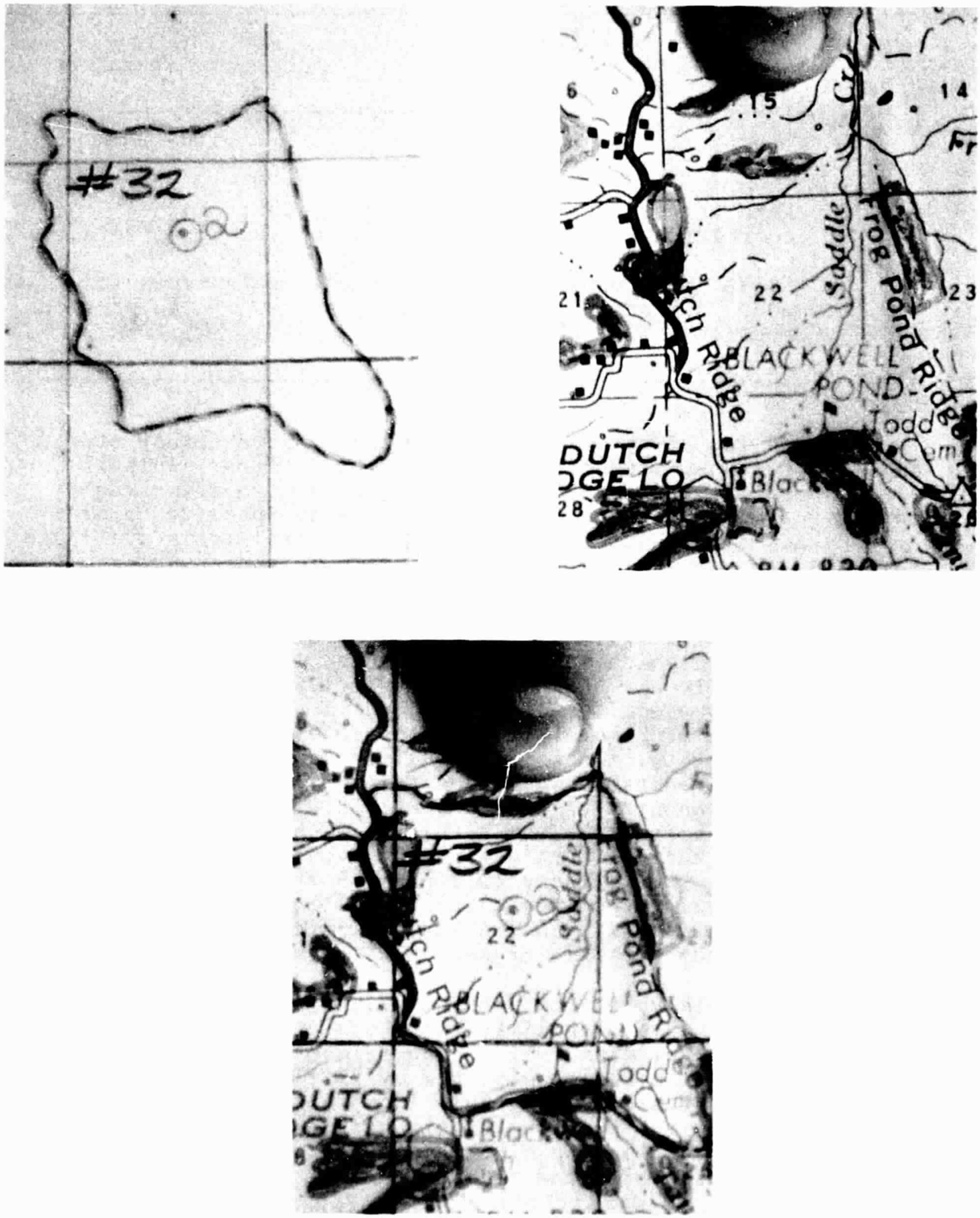
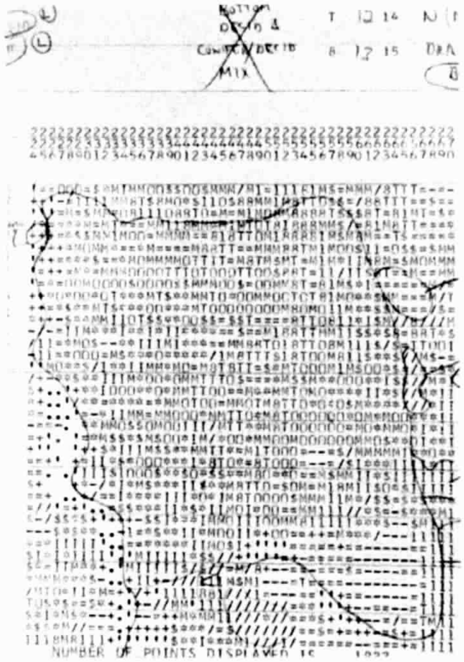
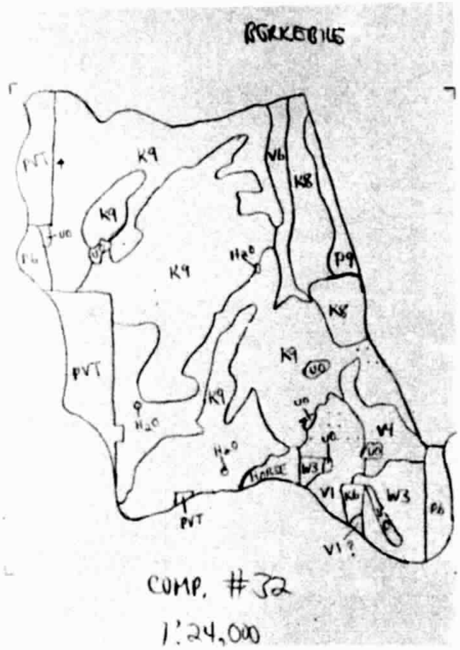


Figure 8 The border of a Forest Service stand compartment map drawn on transparent cellophane, "A". A Forest Service Land Use and Cultural Feature map of compartment #32, "B". An overlay of the two, "C".

Figure 9 Using the Forest Service Land Use and Cultural Feature map, "A", an analyst can also overlay aerial photograph centers, "B", to combine reference data for correlating with computer printouts, "C".





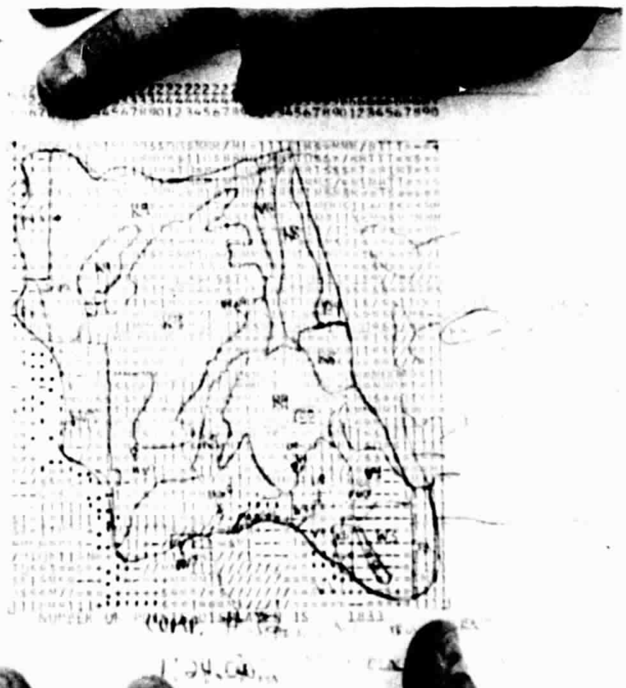


Figure 10 Illustration "A" shows the outline of stand compartment #32 on a computer printout. (At this point in the analysis, either a cluster map or printresults map would be used.) "B" is a forest cover type map of compartment #32. "C" is an overlay of "B" on "A".

Any of these singly or in combination, can be used to establish representative information for training the computer to recognize the numerical scanner data as specific cover types. See Figure 11. For more information concerning correlating importance, see LARS Information Note 120371, The Importance of "Ground Truth" Data in Remote Sensing, by Roger M. Hoffer.

```
         11111111111111111111111111111111111111111111111
         55555555666666666677777777778888888888999999999
         23456789012345678901234567890123456789012345678

279        =-    =    II/--  //-=--=---=---=/     --  -//
280       -     -- --=II/--  -/--= =----=---I-------=   --
281      =-/-   --  I*I=--    ==   -/=-/--- -//------=- - -
282      /=//-  /* **-        -- -=--/-   =- //-===--=/----
283      /=/---=Z*ZI=--  -/=-------  -   /- --==- --///=
284      Z//--/ZZZ/- -  -===/----  /     --  =I=--   --=
285      X*I/*ZXZZ/  /-//-// -/- - /- -  -  =-  ----  -
286      XZIZXZII - =III--=  -- ----/--- -///- - --   --
287      ZZZXZII-- ==I/--I=- =-  ==/  ---=/I/- --//--  -
288      XZZ$ZI-- I=-  =- =  --  //-    -/I/ =----/=--*I
289      XZZZZ-//ZZI--//--/   - -//-   ///--=- -----I/*I
290      XXXXZ*IZZ*//I//= // --/=       /I=  /=   -- -IIII
291      X$$XZZZZX-/--=/-/I/- -/=----=-/=-- -/=-=- //II=-
292      $$X$XZZZZI- -= -/I//-=/I--=///II//== --  //I/-=
293      XXX$$XXZII//=--  =*IIII/=--/==IIIII==----/***---
294      X$$XXXZZI/-=-  /****III/I====II***I=---I/Z*/==-
295      $ZX$XXZZI///- --=III/III*II*I**III=/II/-=*ZZ/-=
296      ZZXXXX$XZZI=I/-I/I*Z*/II*ZZ*II/==/II-=/I=ZZZI--
297      ZZXXXXXZZZI/=--==III*II*ZZZ*I// -/I/**ZZZXZ*=I/
298      Z$X$X$$XXZZ**--/IIII**I*ZZ**I-/*****I//ZZZZ7*II
299      Z$X$$$$$XXXXX**Z*I**ZZZZXZZZ*******III**ZZZZZ*I
300      XXZXXXXXXXXXXZZZ***ZZXXXZZ***IZII**I*ZZZXZZ/*Z
301      XX$$XXX$X$XXXXXZZ*****Z$$ZZZZZ**I******ZZZZZZZZ
302      ZZZZXXXXX$XX$$$$Z**I*Z$X$XZZ*III*II*****ZXZZZXZ
303      ZZZZXX$XXXX$XXXXX****Z$$$XZ***I*I/*ZXZ*ZZZZZXXX
304      Z$$XXM$$$$XXX$XXXXZZXZX$$ZZII*II=-I*XZZZZIZ7$$$
305      *Z$X$$$$$$$$$$$$$$$XXXX$XZZ****III==/ZXXXZZ*Z$$$
306      ZZZ$$$XX$$$$$$$M$$M$XZ$XXZZZ****I/**Z$XX$XXXXXX
307      Z*XXX$$$$XX$X$$M$$$$ZZX$Z*****ZZZZ*ZZ$XX$$$X$$$
308      IIZZXZZZ$$$$$$$$$$$XZZXMZZ*****ZX$XZXXX$$$XXXXXX
309      -I///Z$ZXXX$X$$$$$$ZXMXXZZ*****Z$$$M$$XXXX$XXX
310      /I/- *ZZXX$$X$X$X$$XXM$ZZ*Z*Z$$X$$M$$$$X$$$$M$X
311      /I/  -ZM$$$M$$XX$XX$$$X$$XZ$$X$$$$$$$$$$$XMMMMMM
312      =I/  *$$XM$$$$$$$$M$$$$X$$$XX$M$$$$$$$$$$$$$$$$$
```

Figure 11 This is the spectral responce of the area outlined in Figure 7. The analyst in this case specified the number of lines (279-312) and columns (152-198) which approximate the coordinates of the area shown in Figure 7 and requested a grayscale printout of one channel (0.8 - 1.1 um). The scales have been adjusted to allow the analyst to overlay the printout on the aerial photograph to indicate how the spectral response for the various cover types has been recorded for this one channel.

Self-Check

III-A. List at least three types of reference data available for someone doing a forestry analysis.

III-B. Select at least one types of reference data and describe how it relates to your analysis objectives.

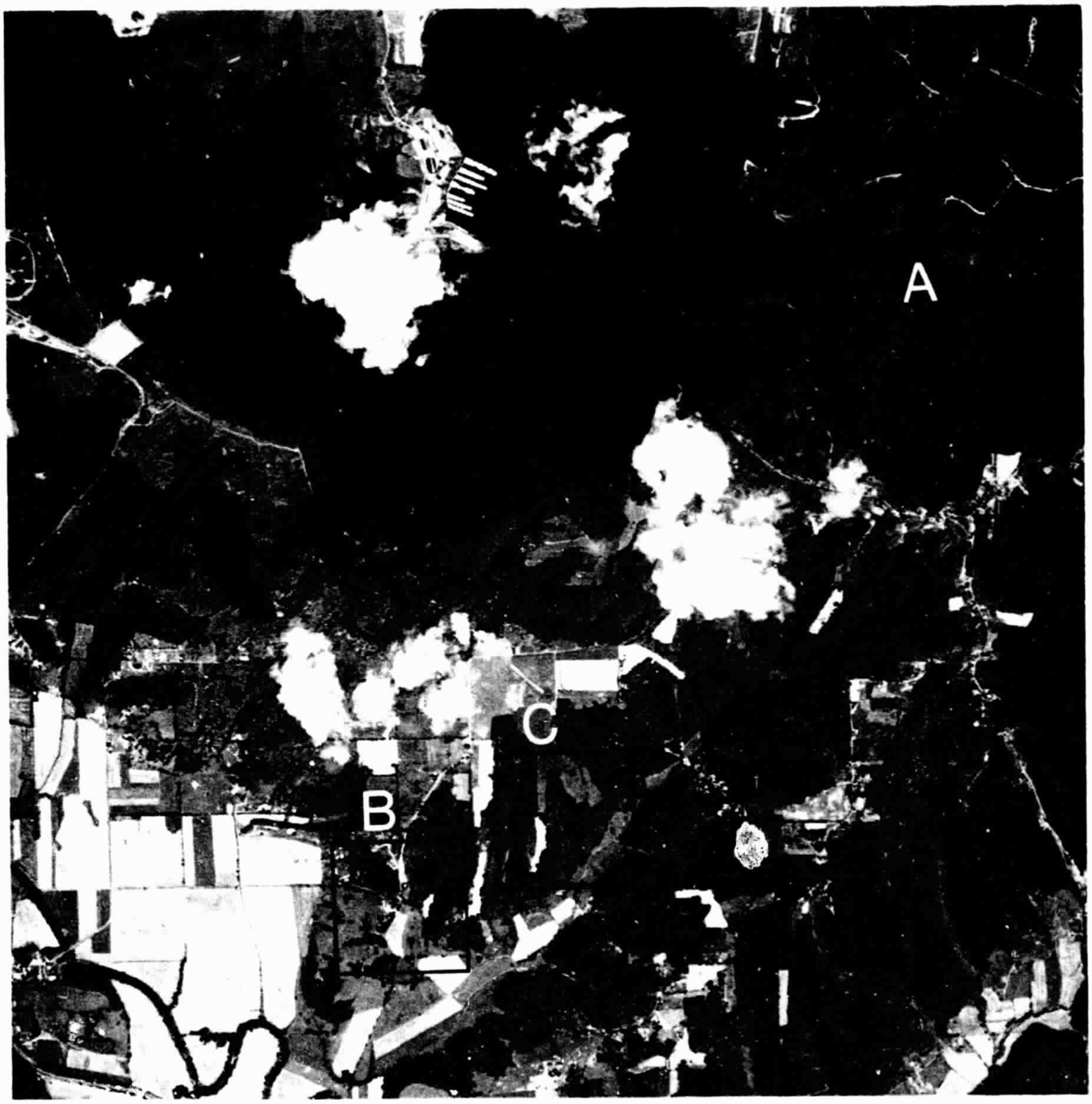


Figure 12 This aerial photograph indicates three types of training area selections. Area "A" is too uniform in cover type and therefore would not be a good area to select as a training site. Area "B" has so many potentially different cover types that it would be difficult to sort out the details. Area "C" has a minimum number of cover types (about 3) which are distinguishable and can be used to accurately train the computer.

SECTION IV SELECT TRAINING AREAS

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

1. Name at least two considerations when selecting training area sites.

2. List at least one size range for a training area and any specific reasons associated with that size selection.

3. State two methods of determining the number of training areas to be selected and identify criteria which should be considered.

4. Select training areas in a planning unit and justify your location, size and number of training areas.

Once you have located and obtained useable reference data, selecting the areas used to train the computer begins. Several questions arise at this time. The answers and their sequence depend upon the objective(s), skill, experience and, in some cases, preference of the analyst. This phase of data analysis is a mixture of art and science closely tied to man-machine interaction. Some general guidelines and an example are given. Your task will be to apply those parameters that relate to your planning unit and your objective.

You might ask, "Where in the planning unit should I select my training areas?" One obvious, yet important consideration is to choose areas for which you have some a priori knowledge through reference data. Diversity is another desirable trait for training areas. However, if your area is too complex, you may not be able to separate the major cover types. A rule of thumb might be to choose an area which has at least two distinct cover types per area. See Figure 12. The maximum number will depend on the size of your training area and how distinct the boundaries are between cover types. See Figure 13, page 24.

Other considerations might include scattering training areas throughout the planning unit if this is possible. Also, if aerial photographs are used as reference data, the center of the photograph will have less distortion due to lens geometry.

What size training area should be used? The size question has almost as many answers as there are analysts. In our example the analyst is using a Bausch & Lomb Zoom Transfer Scope (ZTS) and he has determined that an area of about 47 lines by 34 columns (with LANDSAT data scaled at 1:24,000) fits the ZTS's field of

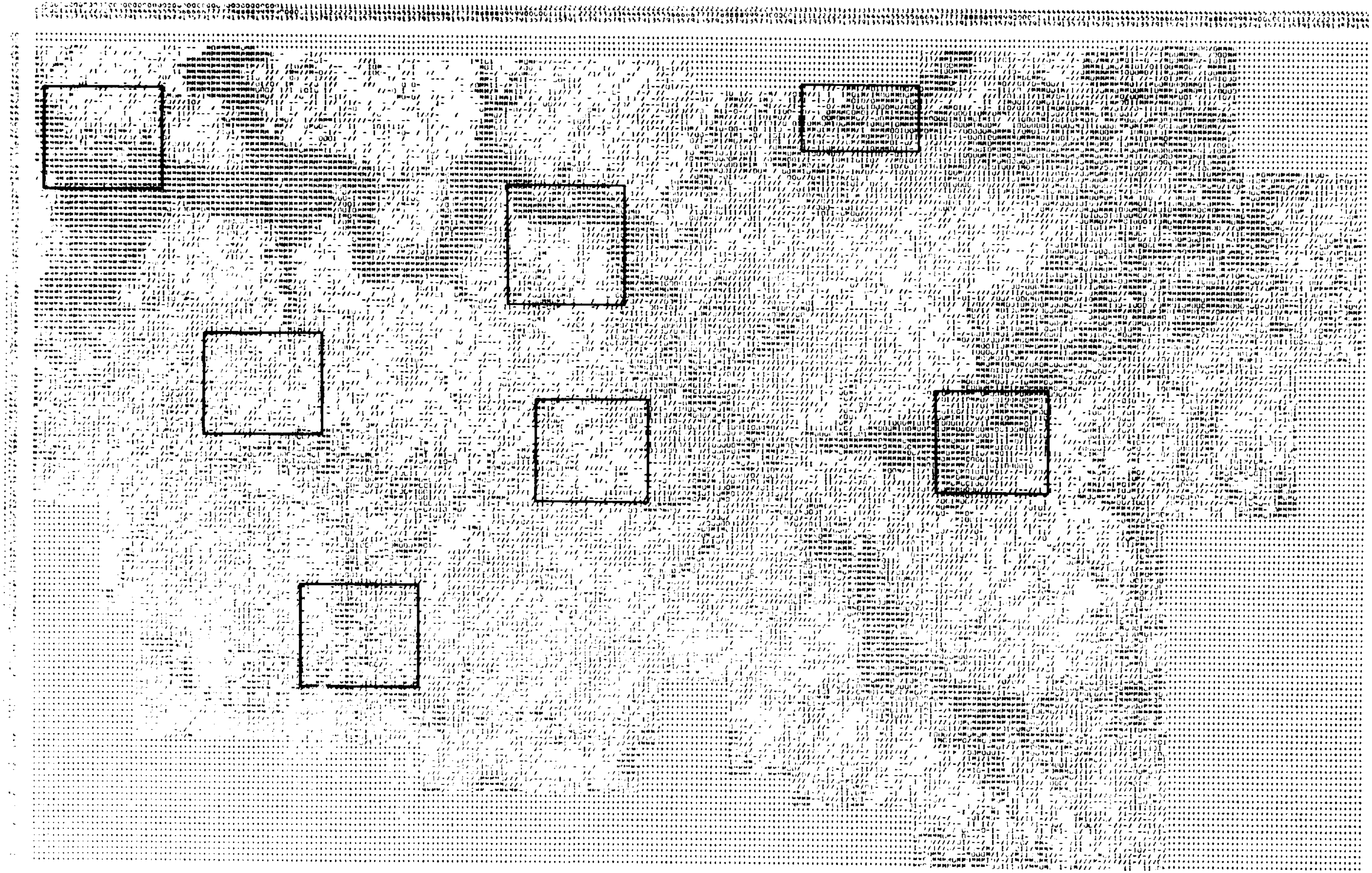


Figure 13 A grayscale printout of the entire planning unit and the seven training sites selected.

view. See Figure 14. When not using a ZTS, other analysts recommend using from 40 to 100 lines by from 40 to 100 columns for the size of the training areas. The number, size, and clarity of cover types being examined should be considered since all have an impact on this decision. Each cover type should be represented in at least one and preferably more than one training area.

Figure 14 The forester using a ZTS to correlate reference data (aerial photographs and a USGS quad map).

At this point, you might be asking how many training areas do I need? No set answer exists, but there are criteria for helping make the decision. Some analysts say four to eight areas. Others use a percentage figure ranging from 1-15% of the total planning unit. Factors to keep in mind here are ruggedness and complexity of topography as well as number and/or size of the training areas. In our example, we are examining a planning unit with relatively flat noncomplex topography and want our classification accuracy to be at least 85%.

In summary, you should note that the computer will classify every data point in the planning unit by means of the training statistics. An attempt should be made to clasify every spectrally distinct cover type even if it is not of direct interest to you. For example, if you were examining some characteristics of vegetation in a forested region containing a large area in pasture, you may not be interested in the pasture in the training area. However, if it is spectrally distinct and the computer is not told to recognize it as a separate class, the data values will be added to another cover type. The result would be a large area being misclassified.

A more thorough understanding of what training areas are and why they are needed can be found in LARS Information Note 110474, An Introduction to Quantitative Remote Sensing by John Lindenlaub and James Russell, and in LARS Information Note 111573, Pattern Recognition: A Basis for Remote Sensing Data Analysis, by Philip H. Swain.

Self-Check

IV-A. Name two factors to be considered when choosing training area sites in a planning unit.

IV-B. Describe a size range of a training area and state a rationale for choosing those size parameters.

IV-C. State one criterion for selecting a given number of training areas.

IV-D. Select training areas in a planning unit of your choice and list why you selected the size, number, and specific location of these training areas.

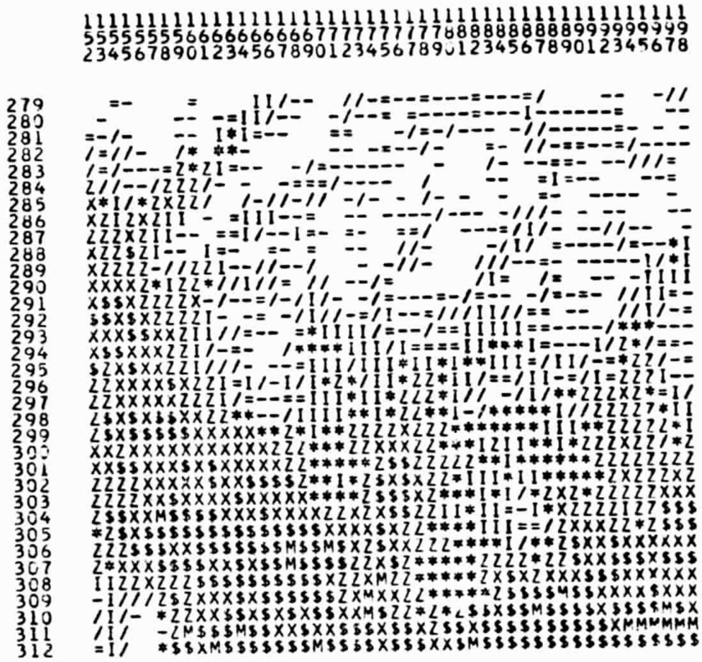


Figure 15 Grayscale printout of a single channel (channel 8, 0.8 - 1.1 µm).

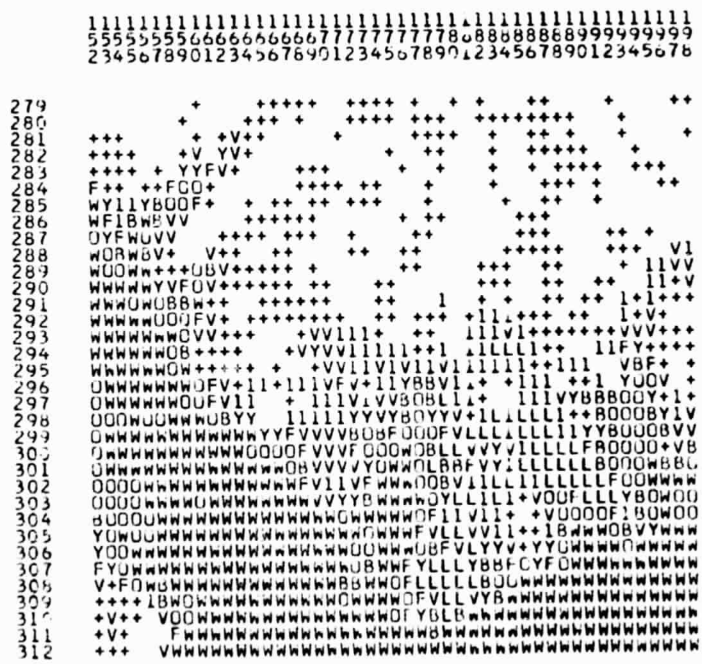


Figure 16 Cluster map of same area using four channels and ten cluster classes.

cluster classes, clustering takes more time and costs more. Experience has shown that clustering with more than twenty cluster classes often results in a single cover types being broken down. They would have to be recombined later, thus wasting time and money. We decided to use nineteen cluster classes. See Figure 17.

As a general rule, in more diverse (spectrally complex) areas, more cluster classes are requested. On the other hand, if the area is less complex, fewer can be used. This decision depends to a large extent on the experience of the analyst. See Figure 18. Additional, more technical information on the clustering process can be found in <u>Pattern Recognition: A Basis for Remote Sensing Data Analysis</u> by Philip H. Swain (LARS Information Note 111572, pp. 27-36).

```
         11111111111111111111111111111111111111111111111
         55555555666666666677777777778888888888999999999
         23456789012345678901234567890123456789012345678

279        - -   +  - ++/-+  ++++------+-  -+   ---   ++
280       -    -   -//+ -- -+-++ ++-- +---+---   +    --
281      ++/--  -- +Y/+-    +- ---++-+-----++ ----+ -- -
282      ++/+-- 1L *Y--    --   +  +-- -+--+++++---+---
283      +++----*YF/+--  -++  ----  -   +  --++----+++--
284      F-11++FOO+---   ++++ +-  -+   ---  +++ -------+-
285      Q*LL*$QOF+- + ++-++ -+----+--- ----+-- -       -
286      QXLBQB//-  -///+-+  --  ---+---- +++  - - -
287      OXKWOY/  -++/+ -/+  +- --++  -  -//+ -- ++--   -
288      QQBW$/-- /+---+- -  --  ++----  +++-+ ---++1 LL
289      WQOWW-++OB/--++- +-  - -++-    +//--+- 1--+  71V/
290      QWWQQ*/FOY++++++-++--  ++-     +++  ++  ---177//
291      QQQQQOBBW+/--+/-+/+-   1+- 1111+----++ +- 11L/+-
292      QQQQQOOOF/---+--+/++111L-++111LL1++1-+-- 11L+ -
293      WQQQQWWO//+/- -  +/YL111--+111LLLLL+11-+11YYY---
294      WWWWQWQB++-+--  1YYYY11L1L11111L7S71+1 1L1F*+---
295      WWWWWQOW+++++-111//L11LLLLLLLLL7LL7+1LL111Y$F+-+
296      QQWWWWWQOFY1LL1LL1YXY1L7*BKL1111-17L-11L1*UO/  -
297      OQQWQQQQOXL11  -111L1YLLV$QKV711111L1Y*$$BQO*1L+
298      QQQQQQWQQQ$X*  11LLLL**L*BOV*L1777777L/1$QOQ$*L/
299      QQWWWWWWWWWWQ**XYLYLBOBXQQOXVSSS77S77L**$QQQBLL
300      QQQQWQQWWQWQQQQQOXLLVXOQQQQKSSLL*L7777SXBQQQO/YB
301      QWQQQQWWWWWWQWWQ$YLVL*OWQOS$$XLV77SSSSSBOOQW$$O
302      QQQQWWWWWWWWWWWWFVLLVKQWWQQ$YL777777SSVVXQQWWQW
303      OQQQWWWQQWWWWWWWVVVV$WWQWO*7S77L1YOQKSSSXKOQQQ
304      $QQQQQQQQWWWWWQWWWQQWWWWQX77Y7L11+YQQQQKLKOQQQ
305      *QQQQWWWWWWWWQWQWWWWWQQWWBV7SYYLL11L$QWQO$LXWQQ
306      *QQWWQWWWWWWWWWWWWWWWQOQWWOBXVS**L/**OWWWWQWQQQW
307      F*QWWWWWWWWWWWQWWWWWWQ$QWKVVSVV$BXQ*XOWQWWWWWWW
308      //XOW$WWWWWWWWWWWWWW$$QWQX7SSSV$QQQWWWWQQWWWWWWW
309      -/++1BWQWWWWWWWWWWWWQWWWWOXVS7L*$WWQWWQWWWWWWWWW
310      +/+-  YOQWWQWWWWWWWWWWWWWOKVKV$QWWWQWWWWWWWWWWWWQ
311      +//     FQQWQWWWWWWWWWWWWWWWWKWWWWWWWWWWWWWWWWWWW
312      -/+   LWQWWWWWWWQWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW
```

Figure 17 Nineteen cluster classes from Training Area 1.

Figure 18 The **forester** determining the number of cluster classes he will use for his planning unit.

Once the training areas have been chosen and clustered, it would be possible to use all of the cluster classes to train the computer. However, all of the cluster classes are not used for a number of reasons: First, the number of cluster classes available at this point is normally greater than the number of classes needed to adequately train the computer. We usually reduce the number of cluster classes to save computer time and to simplify interpretation of the results. Secondly, some of the cluster classes may have too few data points to give good statistical characterization of the spectral classes. Thirdly, by combining spectrally similar clusters from several training areas, you can do more clustering, ask for a different number of clusters or even chose to go back and select additional training areas in an effort to get good spectral definition between cluster classes.

The SEPARABILITY processing function[2] allows you to determine which cluster classes are similar and what the probability of correct classification is. The computer is able to calculate the statistical distance between all pairs of cluster classes. This distance, called "transformed divergence" is a measure of the distance between two cluster classes based upon their mean vectors and covariance matrices. The "transformed divergence" is a number between 0 and 2000, where 2000 is "complete" separability. See Figure 19.

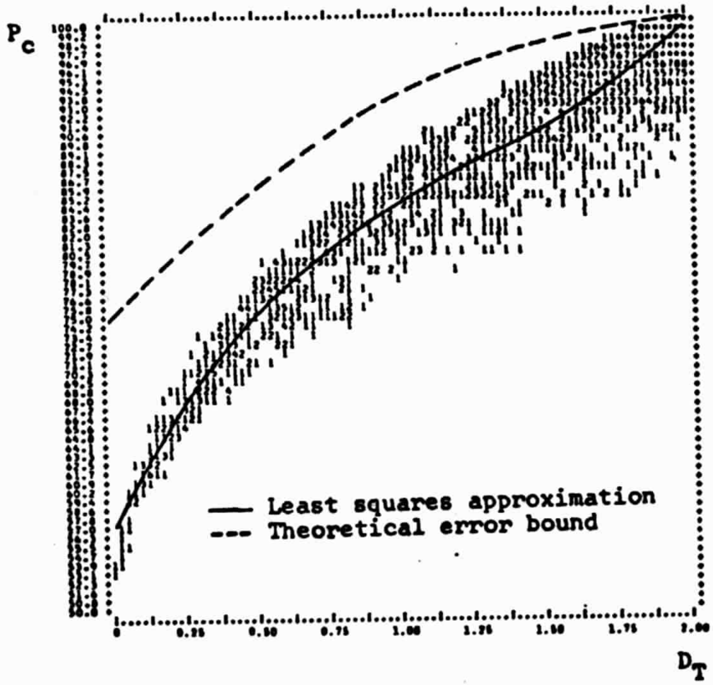


Figure 19 Observed values of probability of correct classification versus transformed divergence. To get transformed divergence values as they are printed by SEPARABILITY, multiply the x-axis by 1000. (from LARS Information Note 042673, Two Effective Feature Selection Criteria for Multispectral Remote Sensing, By Swain and King).

Through experience, it has been found that if cluster classes with a transformed divergence greater than 1700 are combined, the result may be a combining of more than one cover type into the same spectral class. On the other hand, cluster classes with transformed divergence of less than 1500 should generally be combined because of the high probability of their being the same cover type due to their spectral similarity. This, of course, needs to be checked by the analyst. The lower the transformed divergence, the more spectrally similar they are.

[2]For brief description of SEPARABILITY, see Appendix A, page 67.

We decided to use a transformed divergence of 1600 as the threshold point. Thus cluster classes less than 1600 were combined; those greater than 1600 were kept separate. The results are shown in Figure 20.

RESULTS OF SEPARABILITY GROUPING

THRESHOLD = 1600

GROUP	CLASSES	NO. PTS.
1	1	174
2	2 3	233 175
3	4	50
4	5 7	91 82
5	6 8	41 27
6	9	26
7	10	32
8	11	30
9	12	22
10	13	12
11	14	13
12	15	20
13	16	28
14	17	46
15	18	171
16	19	325

Figure 20 Separability output using a threshold of 1600 which combines (pools) those spectral classes which are spectrally inseparable.

Using a computer process called MERGESTATISTICS[3], the computer can be directed to combine cluster classes with transformed divergence less than any specified value.

```
YOU HAVE SELECTED THE FOLLOWING POOLS

 1        DECK  1  CLASSES  1,
 2        DECK  1  CLASSES  2, 3,
 3        DECK  1  CLASSES  4,
 4        DECK  1  CLASSES  5, 7,
 5        DECK  1  CLASSES  6, 8,
 6        DECK  1  CLASSES  9,
 7        DECK  1  CLASSES 10,
 8        DECK  1  CLASSES 11,
 9        DECK  1  CLASSES 12,
10        DECK  1  CLASSES 13,
11        DECK  1  CLASSES 14,
12        DECK  1  CLASSES 15,
13        DECK  1  CLASSES 16,
14        DECK  1  CLASSES 17,
15        DECK  1  CLASSES 18,
16        DECK  1  CLASSES 19,
```

Figure 21 Output of MERGESTATISTICS showing the pooling of 19 cluster classes into 16 spectrally separable classes as defined by the SEPARABILITY grouping table shown in Figure 20.

After the statistics (means and covariances) of cluster class groups with transformed divergence less than 1600 have been combined, the computer is then used to classify each training area using CLASSIFYPOINTS[4]. The new statistics deck for each area is required as input to the classification program, and the output is a classification map of each training area. The classification map for one of the seven areas is shown in Figure 22. A photo of the same area is shown in Figure 7, page 14. Compare them and note the correspondence.

---

[3]For a brief description of MERGESTATISTICS, see Appendix A, page 65.

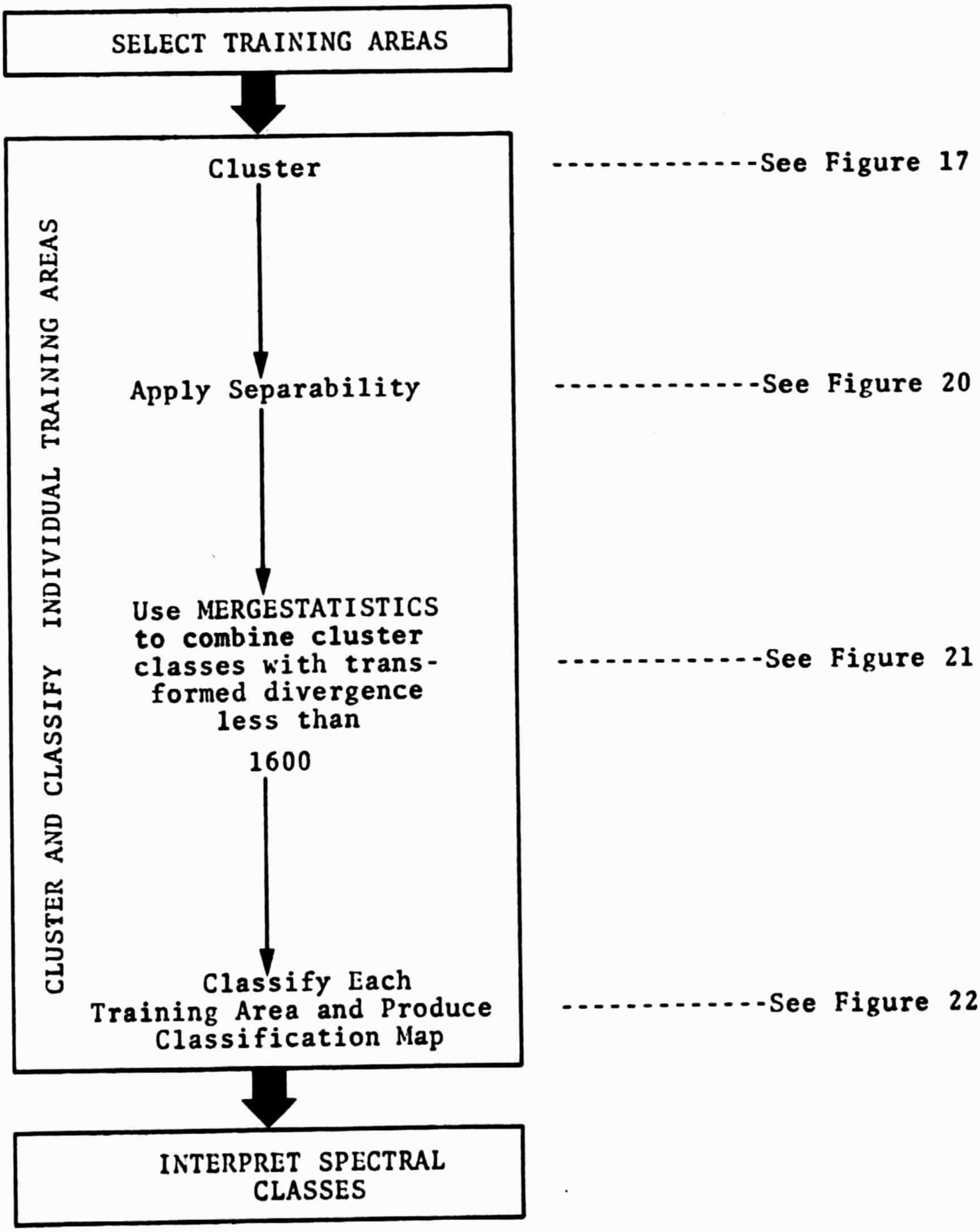
[4]For a brief description of CLASSIFYPOINTS, see Appendix A, page 63.

```
        11111111111111111111111111111111111111111111111
        55555555666666666677777777778888888888999999999
        23456789012345678901234567890123456789012345678

279         -   -  - +++--  ------------  -+   ---   --
280      -     -    -+- -- ----- ---- --------   -   --
281     --+--  -- -S+--    -  -------------- ------ -/ -
282     /-++-- /S LS--    --   -  --- ----------------
283     -------LS*+---  ---  ----   -   -  -------------
284     *-//--*KK----   ----  --  -+   ---  --- --------
285     BL//LFBK*-- - ----- ------+--- -------- -     -
286     BV/XBX++-   +++---  --  --------- ---   - - -
287     KVY$KS+  ---+- -+-  -- -----  -  -++- -- -+--  -
288     BBX$F+-- +------ -  --  -------  ------ -----////
289     $BK$$--/KX+-----/ +-  - ---    -++----- /---/1/7+
290     B$$BBL+*KS--------++-- ---    ---   --  ---/11++
291     BBBBBKXX$-+----+--+-- //-- ////-------- -- ///+--
292     BBBBBKKK*+---------+-+////---////1/--/----////-
293     $BBBB$$K++-+- -  ++S1///---///////7/-//--//SSS---
294     $$$$B$BX------ /SSSS/////////////711/-/ ///*L----
295     $$$$$BK$----/-///++//////////////11-///////SF*---
296     BB$$$$$BK*S////////SVS//1LXY/////////1/////1/LKKS -
297     KBB$BBBBKV/// -/////S//7FBY71/////////SLFFXBKL//+
298     BBBBBBB$BBBFVL ///////LL/LXK7L//111111/+/FKBBFL/+
299     BB$$$$$$$$$$BLLVS/S/XKXVBBKV711111111111LLFKBPXS/
300     BBBB$BB$$B$BBBBKV//7VFBBBBY11//L/111111VXKBBK/SX
301     B$BBBBB$$$$$$B$$BFS/7/LK$BK1FFV/711111111XKKB$FFF
302     BBBB$$$$$$$$$$$$$*7///7YB$$BBFS11111111111177VBB$$B$
303     KBBB$$$$BB$$$$$$$$7777F$$B$KL1111//SKBY111VYKBBB
304     FBBBBBBBB$$$$$$B$$$BB$$$$BV11S1///+SBBBBY/YKBBB
305     LBBBB$$$$$$$$$B$B$$$$$$BB$$X711SS/////FB$BKF7V$BB
306     LBB$$B$$B$$$$$$$$$$$$BKB$$KXV71LL/+LLK$$$$B$PBB$
307     *LB$$$$$$$$$$$$B$$$$$$BFB$Y77177FXVBLVK$B$$$$$$$$
308     ++VK$F$$$$$$$$$$$$$$FFB$BV11111FBBB$$$$BB$$$$$$$
309     -+--/X$B$$$$$$$$$$$$$B$$$$KV7117LF$$B$$B$$$$$$$$$
310     -++-  SKB$$B$$$$$$$$$$$$$KY7Y7FB$$$B$$$$$$$$$$$$$B
311     -++  /*BB$B$$$$$$$$$$$$$$$$Y$$$$$$$$$$$$$$$$$$$$
312     -+- /+$B$$$$$$$B$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$$
            NUMBER OF POINTS DISPLAYED IS     1598
```

Figure 22 Classification map (PRINTRESULTS) of training Area 1 using 16 spectrally separable classes defined by SEPARABILITY. An aerial photograph of this area is shown in Figure 7 on page 14.

It should be pointed out that there are alternate approaches to the process just described. The approach which we have just described is summarized below.



Self-Check

V-A. Describe the function of the CLUSTER processor.

V-B. State two reasons for using the CLUSTER processing function.

V-C. Discuss the function of the SEPARABILITY processor.

V-D. Describe the output after using the SEPARABILITY processor.

V-E. State what "transformed divergence" measures.

V-F. Describe the function of "MERGESTATISTICS".

Figure 23A

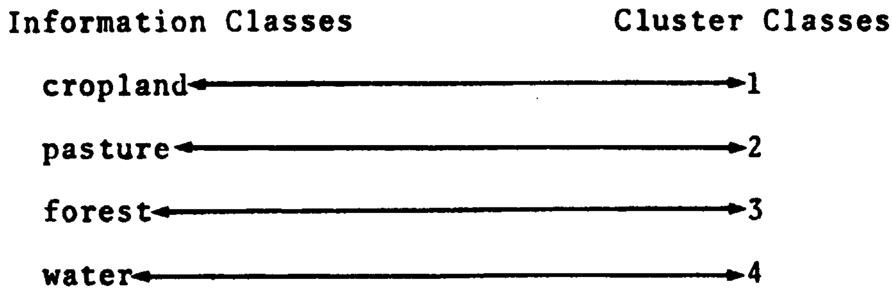


Figure 23B

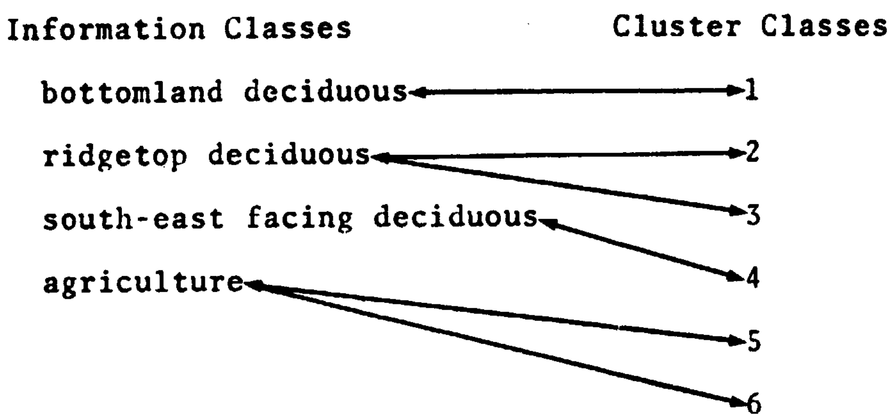


Figure 23C

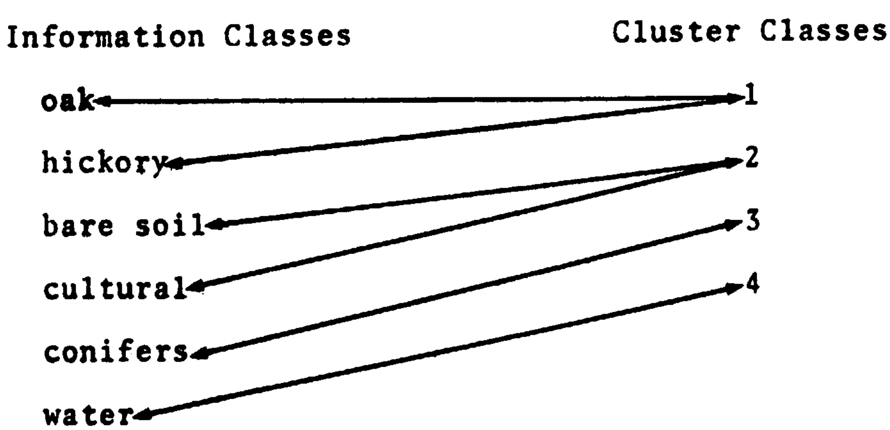


Figure 23 Examples of relationships between different information classes and their cluster classes.

SECTION VI INTERPRET SPECTRAL CLASSES

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

1. Describe the three possible correlations between information classes and spectral classes.

2. Describe the process by which the forester correlates computer-generated cluster classes with known ground information.

3. Define "mixture class".

The purpose of this step in the analysis sequence is to associate each cluster class identified in the previous step with an information class (i.e., agriculture, urban, water, forest). It should be pointed out that there is not necessarily a one-to-one correspondence between the information classes and the cluster classes. Remember, an information class is a distinct cover type of interest as noted above, while a cluster class is a group of data points which are spectrally similar. As shown in Figure 23A, there may be a one-to-one correspondence between the two. However, this is rarely the case. It is possible that several cluster classes will represent the same cover type (information class) as shown in Figure 23B. Sometimes several information classes will be associated with the same cluster class (Figure 23C). In this case, the cover types are spectrally similar and cannot be differentiated using these data.

In order to determine the association between the cover types and the cluster classes, we have chosen to compare the classification map generated in the previous section with an aerial photo or stand compartment map. This can be done by laying the two side-by-side on a table, but it is very difficult to make a direct comparison. To greatly faciliate this task, a Zoom Transfer Scope (See Figure 24) can be used. This optical device allows you to look at both the classification map and the aerial photo (or stand compartment map) while one is superimposed over the other. With the proper adjustment of the scale and light intensity, it is possible to see the two as a single image.

When the Zoom Transfer Scope (ZTS) has been properly adjusted, the analyst correlates the symbols on the classification map with the features shown in the aerial photography. For example, he may note that the letter "M" on the computer-generated classification map appears where there is water on the aerial photograph.

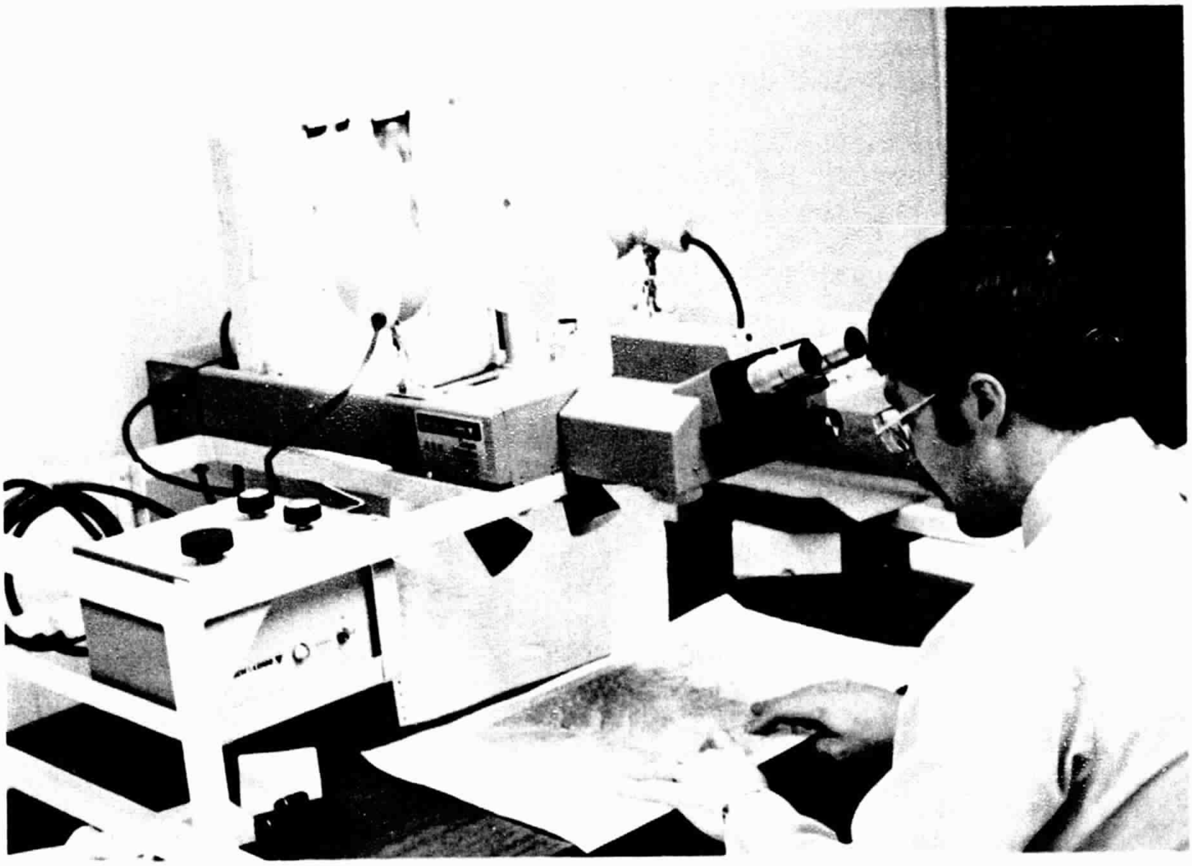


Figure 24 The forester comparing the classification map output with the aerial photographs using the ZTS.

The accuracy of correlating the computer symbols (clusters) with the cover types depends upon your photo-interpretation skill, if the photography rather than maps is the reference data. Often, scattered or boundary symbols will represent "mixture classes". A "mixture class" is a spectral class which represents more than one cover type, i.e., pine-hardwood. Hence, these cover types can't be mapped separately by spectral procedures and must remain grouped. One may attempt to spectrally separate these classes by additional training, but success will not always be attained. As indicated in Figure 24, it is possible to have the same cluster symbol represent more than one cover type.

Self-Check

VI-A. Discuss one method which can be used by a forester to correlate computer-generated cluster classes and known ground information.

VI-B. Define "mixture class".

SECTION VII POOL AND/OR DELETE SPECTRAL-INFORMATIONAL CLASSES

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

1. Describe the use of the SEPARABILITY processor function.
2. Describe the use of the MERGESTATISTICS processor function.
3. Describe the process of forming a final statistics deck for training the computer by means of the SEPARABILITY and MERGESTATISTICS processors.

If as a result of the previous section mixture classes are identified, they can be deleted by using MERGESTATISTICS unless they are of interest. Then one would label all distinct spectral classes.

Up to this point, each training area has been examined separately and various cover types of interest have been associated with a certain cluster of data points (cluster class). We are now ready to combine data from all the training areas. One reason for doing this is that a specific cover type may have a slightly different spectral response in a different part of the planning unit. This may be caused by differences in slope-aspect, substrate, moisture conditions or other variables. By combining the training data, we are essentially telling the computer to expect a certain amount of variance in the spectral response of a specific cover type as the entire planning unit is mapped.

After combining the data, the SEPARABILITY function is used once more to create a grouping table for the combined data from all the training areas. A threshold of 1000 (transformed divergence) was selected. It can be shown that spectral classes with less than 1000 will not be distinct and have a fairly high probability of being the same information class. We combine spectral classes in a two step procedure. The first step (transformed divergence of 1000) collects spectral classes which are similar. The second step requires a decision based upon the experience of the analyst. We chose to use a transformed divergence of 1600 which has proven to enable a compromise between informational classes which you would like to classify and spectral classes which you can classify.

At this point, the computer is programmed through a function known as MERGESTATISTICS to pool the cluster classes which have a transformed divergence of less than 1000. Unneeded mixture classes and those with small numbers of data points can be deleted as desired. Another SEPARABILITY run with a transformed divergence of 1600 shows further associations of closely related cluster pairs. The cluster groups remaining after the 1000 pass can again be either pooled or deleted.

The cluster classes can be combined for various transformed divergence values, such as 1400, 1500, 1600 and 1700. If you want greater information detail, you will have to sacrifice some classification accuracy. See Figure 25. On the other hand, if the analysis objective(s) can be satisfied with less detail, you can achieve greater classification accuracy. See Figure 26.

There are two types of classification errors that the analyst is concerned with -- omission and commission. As the term indicates, omission is omitting data points that should have been classified as a given cover type. An example in Figure 25 would be the 169 points called spruce-fir that should have been classified pine. When a data point is classified as a cover type when it should not have been, it is an error of commission. An example in Figure 25 are the 254 points which were actually pine, but were classified as spruce.

OMISSION FROM PINE

Group	No. of samples	Percent correct	Pine	Spruce-fir	Oak	Aspen	Pasture	Bare	Water	Cult. crop
1. Pine	1111	81.4	904	169	5	9	3	20	1	0
2. Spruce-fir	747	64.9	254	485	2	6	0	0	0	0
3. Oak	481	61.7	8	0	297	95	80	0	0	1
4. Aspen	204	78.4	5	0	33	160	6	0	0	0
5. Pasture	188	94.1	2	0	4	0	177	1	0	4
6. Bare	98	93.9	0	0	0	0	1	92	0	5
7. Water	240	100.0	0	0	0	0	0	0	240	0
8. Cult. crop	54	61.1	0	0	2	0	18	1	0	33
Total	3123		1173	654	343	270	285	114	241	43

Overall performance (2388/3123) = 76.5
Average performance by class (635.5/8) = 79.4

COMMISSION INTO PINE

Figure 25 A level II[1] test field performance matrix indicating an overall performance accuracy of 76.5%

[1]Anderson, J.R., Hardy, E. E., and Roach, J. T. 1972. A Land Use Classification System for Use with Remote-Sensor Data Geological Survey Circular 671.

Group	No. of samples	Percent correct	Conifer	Deciduous	Agricultural	Water	Bare
1. Conifer	1858	97.5	1812	22	3	1	20
2. Deciduous	685	85.4	13	585	87	0	0
3. Agricultural	242	95.9	2	6	232	0	2
4. Water	240	100.0	0	0	0	240	0
5. Bare	98	93.9	0	0	6	0	92
Total	3123		1827	613	328	241	114

Overall performance (2961/3123) = 94.8
Average performance by class (472.7/5) = 94.5

Figure 26 A level I[2]test field performacne matrix indicating an overall performance accuracy of 94.8%.

The result of performing MERGESTATISTICS and SEPARABILITY is a final statistics deck which contains a statistical description of the various cover types. The description includes the mean vector and covariance matrix for each of the training classes.

[2]Ibid.

Self-Check

VII-A. State the reason the SEPARABILITY processor is reused at this point in the analysis.

VII-B. The MERGE processor function is used to accomplish what task?

VII-C. What is the result of using MERGE and SEPARABILITY functions on a set of data?

VII-D. In Figure 25 in the aspen group (line 4) the 33 points which were actually oak were what kind of error?

____

VII-E. In Figure 25, 95 aspen samples were classified as oak (line 3) and therefore represent which type of error?

____

```
     11111111111111111111111111111111111111111111111
     55555555666666666677777777778888888888999999999
     23456789012345678901234567890123456789012345678

279  000C000000000000C000000000C0C000000000000000000
280  000000000C00000000000000C0000000000000000000000
281  00000C00000*00000000000000000000000C00000000000
282  00CCC000/0.*000CC000000000000000000000000000000
283  C000C00.*W00000CC000000000000000000000000000000
284  W00000WWWC0C000000000000C00C0000000000000000000
285  W.0-.WWWW000000000000000C00000000000C0000C00000
286  WW-WWW000000000CC0000000000000000000000000C00C0
287  WWWWW*000000000CC000000000000000000000000000000
288  WWWWW0000C00C0000000000C00000000000000000000//0
289  WWWWWC00WW0000000000000C00000000000C0000000-0.0
290  WWWWW.0WW*00000CC0000000000000000C0C0000000--C0
291  WWWWWWWWW000000CC00000000000//00000000000000000
292  WWWWWWWWW000000C0000000C0000/0--000000000-00000
293  WWWWWWWW000000000*/-0000C000/0--/0000C000/.*000
294  WWWWWWWWW000000C****-//-000//--...C000000W.00C0
295  WWWWWWWW0000000/C000-//-/0//0-.---000000/*WW000
296  WWWWWWWWWWW*C00C00.W.0--.WW/-/000/-0000-/.WW*00
297  WWWWWWWWWW/0/00000/0.-/.WWW.-000//-/*.WWWWW.000
298  WWWWWWWWWWW..0/CC--0../.WW../0-.....000WWWWW.C0
299  WWWWWWWWWWWWW..W////WWWWWWWW........--..WWWWW//
300  WWWWWWWWWWWWWWWWW//.WWWWWWW.../././-....WWWWWW0.W
301  WWWWWWWWWWWWWWWWW//./.WWWW.WWW/.........WWWWWWW
302  WWWWWWWWWWWWWWWWW.//.WWWWWWW.--..-......WWWWWWW
303  WWWWWWWWWWWWWWWWW....WWWWWW...-.-0*WWW...WWWWWW
304  WWWWWWWWWWWWWWWWWWWWWWWWWWW..*--000/WWWWW0WWWWW
305  .WWWWWWWWWWWWWWWWWWWWWWWWW...*/-0000WWWWWW/WWWW
306  .WWWWWWWWWWWWWWWWWWWWWWWWWWW....00..WWWWWWWWWWW
307  W*WWWWWWWWWWWWWWWWWWWWWWW.....WWWW.WWWWWWWWWWWW
308  00WWWWWWWWWWWWWWWWWWWWWWW.....WWWWWWWWWWWWWWWWW
309  00000WWWWWWWWWWWWWWWWWWWW.../.WWWWWWWWWWWWWWWWW
310  000CC*WWWWWWWWWWWWWWWWWWW.W.WWWWWWWWWWWWWWWWWWW
311  000000WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW
312  00000/WWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWWW

          NUMBER OF POINTS DISPLAYED IS      1598
```

Figure 27 Classification map of training area I showing six general land-use categories.

SECTION VIII CLASSIFY TRAINING AREAS AS A SINGLE DATA SET

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

1. Describe the purpose of classifying all the training areas as a single data set.

The data analyzed up to this point, have been punched on a series of statistics cards describing the mean vectors and covariance matrices of the training classes. This deck was used by the analyst as he grouped and deleted classes (Section VII). CLASSIFYPOINTS is used to identify all points in the various training areas as belonging to specific training classes. The CLASSIFYPOINTS function operates by taking each multispectral data point and assigning it to the class to which it is most closely related statistically.

The output for this simulation is seven separate classification maps. See Figure 27 for training area 1. By comparing Figure 27 with Figure 22 on page 35, you can see that fewer symbols are used and the boundaries are smoother and cleaner. The spectral classes shown in Figure 27 are representative of the entire planning unit. The seven maps are used in the next step to qualitatively evaluate classification performance.

Self-Check

VIII-A. In a couple of sentences describe the reason for classifying all the training areas as a single data set.

SECTION IX

EVALUATE THE ACCURACY OF TRAINING AREA CLASSIFICATION

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

1. Name the criteria used to determine if the proper level of accuracy has been achieved.

2. List and describe two reasons for revising training area analysis.

Available output at this point is a classification map which could be used with a Zoom Transfer Scope to visually compare the classification with some reference data such as an aerial photograph. This technique relies heavily on the judgement and photo interpretation skill of the analyst.

If you are satisfied with the classification and the level of accuracy gained, move to the final phases of data analysis. Satisfaction should be based on the analyst's objectives. If the results will not make it possible to accomplish the objectives, then revision in one of the previous steps will be necessary.

In some cases, entirely new training areas may have to be selected (See Part IV). This is a drastic change which rarely occurs if reasonably careful site selection of the training areas was initially done. A less drastic but still time consuming revision would be if you altered the number of classes asked for when the initial clustering of data in the training area was being done (See Part V). With experience and thought, these kinds of problems can be avoided. The most common revision needed is to return to the pooling and deleting phase of data analysis (See Part VII). After determining the level of accuracy if you decide that higher accuracy and fewer classes are more desirable, the MERGESTATISTICS processing function with its pooling capabilities would need to be used. If the objectives are reviewed periodically and careful decisions are made, no revisions will be needed and you can move to the next step.

PRECEDING PAGE BLANK NOT FILMED

Self-Check

IX-A. What is the key factor that indicates to the analyst that he has achieved the proper level of analysis accuracy?

IX-B. Describe two occasions that would prompt an analyst to make revisions in the training areas' analysis.

SECTION X CLASSIFY THE ENTIRE PLANNING UNIT

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

1. Describe five types of output which could result after the entire planning unit has been classified by the CLASSIFYPOINTS function.

The last major step in the computer-aided analysis is to use the refined training statistics to classify the entirepplanning unit. The CLASSIFYPOINTS function is again used to identify all the data points in the planning unit as one of the specified cover types. Output from this phase can assume several formats depending on your needs. PRINTRESULTS[1] provides a variety of outputs. The following is a partial list of output which could be produced as a result of the classification being done:

1. Acreage calculations for specific cover types, Figure 28.

2. CALCOMP maps, Figure 29.

3. Planning unit classification maps, Figure 30.

4. Color-coded image displays which produces an image where each color represents a specific cover type.

5. Estimates of classification accuracy, Figures 31 & 32.

These are products of the analysis--tools to be used to accomplish the final step in the process.

---

[1]See Appendix A, page 66.

GROUP	POINTS	ACRES	HECTARES	PERCENT
HARDWOOD	66843	78294.4	31698.1	56.1
CONIFER	8761	10261.9	4154.6	7.4
BRUSH	13377	15668.7	6343.6	11.2
AG	23783	27863.3	11280.7	20.0
WATER	6278	7353.5	2977.1	5.3
TOTAL	119047	139441.8	56454.2	100.0

EACH DATA POINT REPRESENTS 1.17 ACRES
0.47 HECTARES

CLASSES NOT CONSIDERED	NO.PTS.
N	32741
TOTAL	32741

TOTAL POINTS IN CLASSIFICATION = 151788

Figure 28 Acreage calculations output from the planning unit examined.

(overlay)

LABORATORY FOR APPLICATIONS OF REMOTE SENSING
PURDUE UNIVERSITY

CALCOMP CLASSIFICATION MAP

CLASSIFICATION STUDY 515658670	CLASSIFIED
RUN NUMBER............ 73051708	DATE DATA TAKEN... 5 / 4 / 73
FLIGHTLINE.... 128516001 IN	TIME DATA TAKEN..... 1000 HOURS
DATA TAPE/FILE NUMBER.. 1970/ 1	PLATFORM ALTITUDE. 3062000 FEET
REFORMATTING DATE. NOV 15,1974	GROUND HEADING..... 180 DEGREES

CLASSIFICATION TAPE/FILE NUMBER ... 537/ 2

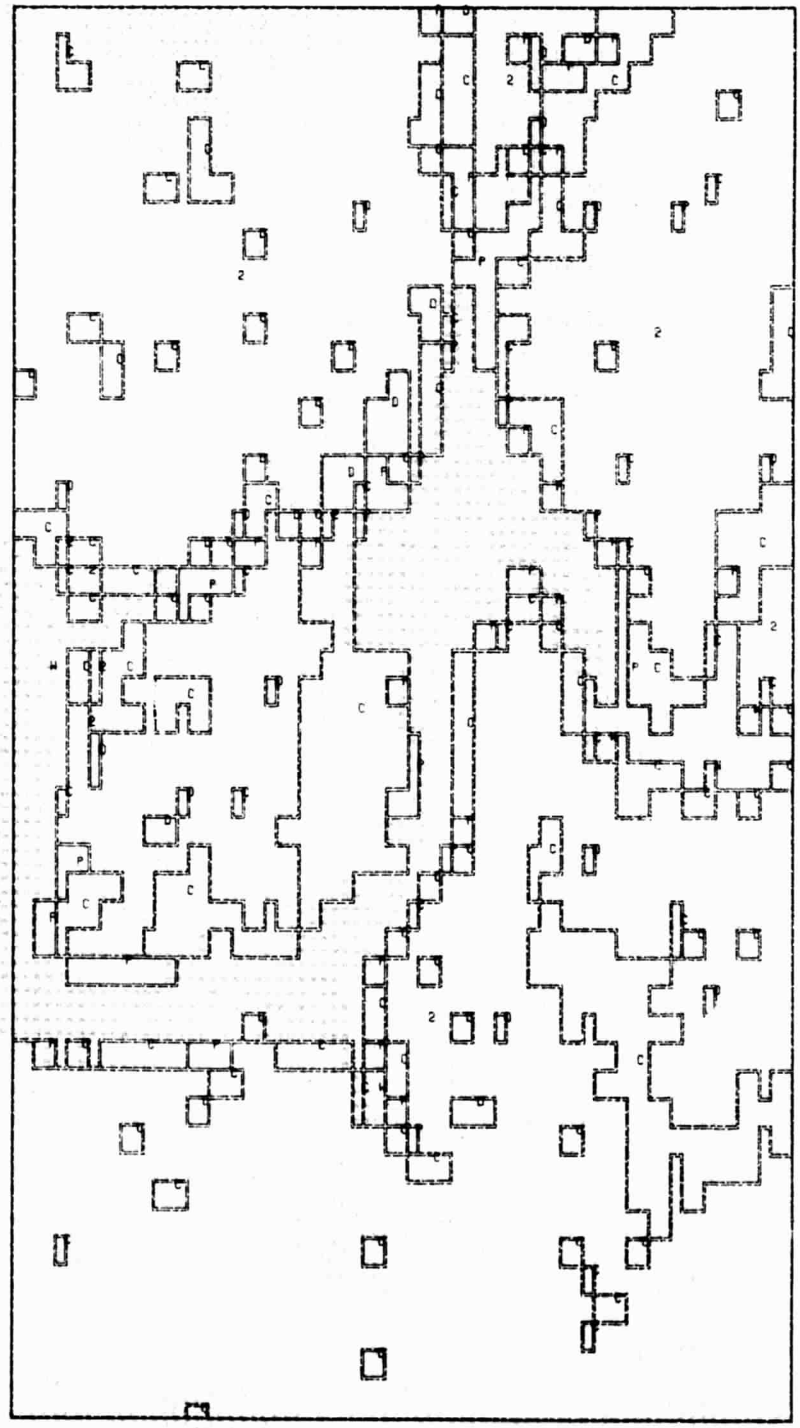


Figure 29A A CALCOMP map representing the planning unit.

P PASTURE
D DECIDE
2 DECIDE 2
W WATER
C CONIFER

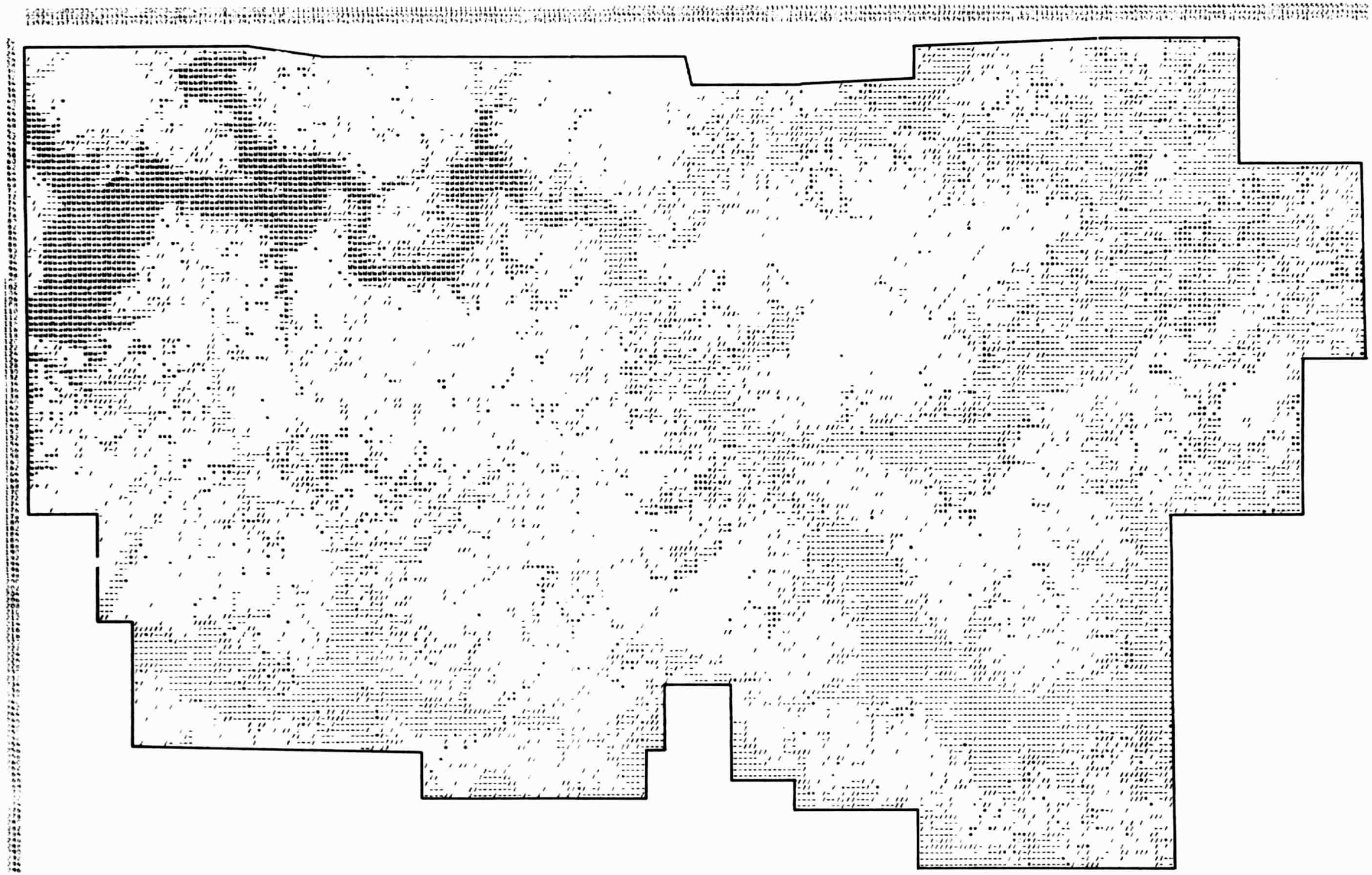


Figure 30 Planning unit classification map. The various symbols indicate the cover types classified. The blank area corresponds to the deciduous forest. The "*" represents coniferous forest, the "/" brushland, "-" agriculture, and the "w" water.

TRAINING FIELD RESULTS FOR JUNE CLASSIFICATION

OF HOOSIER NATIONAL FOREST

Cover Type	Number of Samples	Percent Correct	Deciduous Forest	Coniferous Forest	Brushland	Pasture	Cropland	Water
Deciduous Forest	4678	96.5	4515	7	152	3	1	0
Coniferous Forest	123	87.8	2	108	7	0	6	0
Brushland	360	81.1	19	11	292	20	18	0
Pasture	592	86.8	5	0	51	514	22	0
Cropland	982	96.7	0	18	5	7	950	2
Water	739	100.0	0	0	0	0	0	739

Overall Performance = (7118/7474) = 95.2%

Figure 31 A test field performance matrix indicating an overall classification accuracy of 95.2%.

Self-Check

X-A. With a sentence or two describe each of the following types of output available after the CLASSIFYPOINTS function has been used:

1. Acreage calculations -

2. CALCOMP maps -

3. Classification maps -

4. Color-coded images -

5. Accuracy estimates -

SECTION XI APPLY RESULTS

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

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

2. Name at least two types of applications that can be made from the results of a forest classification.

3. Describe an example of useful information extracted from analysis of remotely sensed data in your particular field of forestry.

A simple diagram might be appropriate to give an overall perspective of the sequence you have been involved.

DATA —analysis→ INFORMATION —analysis→ APPLICATION

So far we have concentrated on collecting data and analyzing it to obtain information. We are not ready to analyze the information in terms of the proposed application(s) . Several formats of information have been obtained as a result of analyzing the data. We have met our initial objectives by producing a general land use map and determining the extent of specific cover types. The final and perhaps most important step is the interpretation and application of this information.

Three applications for the information produced in this analysis include:

1. Determining sites for new recreation areas and more effective management of existing areas.

2. Stratifying timber production areas for more detailed ground studies of size-class/age distributions.

3. Determining potential fire hazard areas and either attempt to eliminate them or deploy fire crews accordingly.

This is not meant to be an exhaustive list of uses since each area is different and your objectives will undoubtedly vary. The advantage of computer-aided analysis can be seen when one considers the time savings in man hours spent surveying a large forested area. Repetitive surveys are also possible since LANDSAT data are collected over the same area every eighteen days.

Examples of results from multispectral data classifications with application to forestry may be found in several journals, including:

Forest Science

Journal of Forestry

Journal of Range Management

Journal of Soil and Water Conservation

Photogrammetric Engineering and Remote Sensing

Remote Sensing of Forestry

Remote Sensing of the Environment

Figure 33 A forester in the field checking the analysis results and planning the application of the results for his planning unit.

Self-Check

XI-A. Identify two types of information that can be extracted from a computer-aided numerical analysis of remotely sensed data.

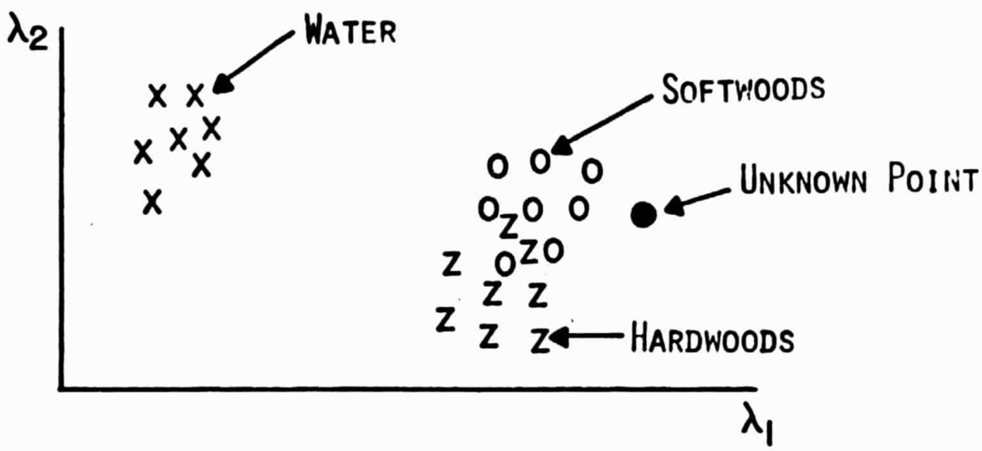
XI-B. List two applications of a forest classification study.

XI-C. Describe a specific use for you of a forestry classification study.

APPENDIX A

SUMMARY OF RELEVANT LARSYS PROCESSING FUNCTIONS

CLASSIFYPOINTS performs the maximum likelihood classification on a point-by-point basis over an area specified by the user.



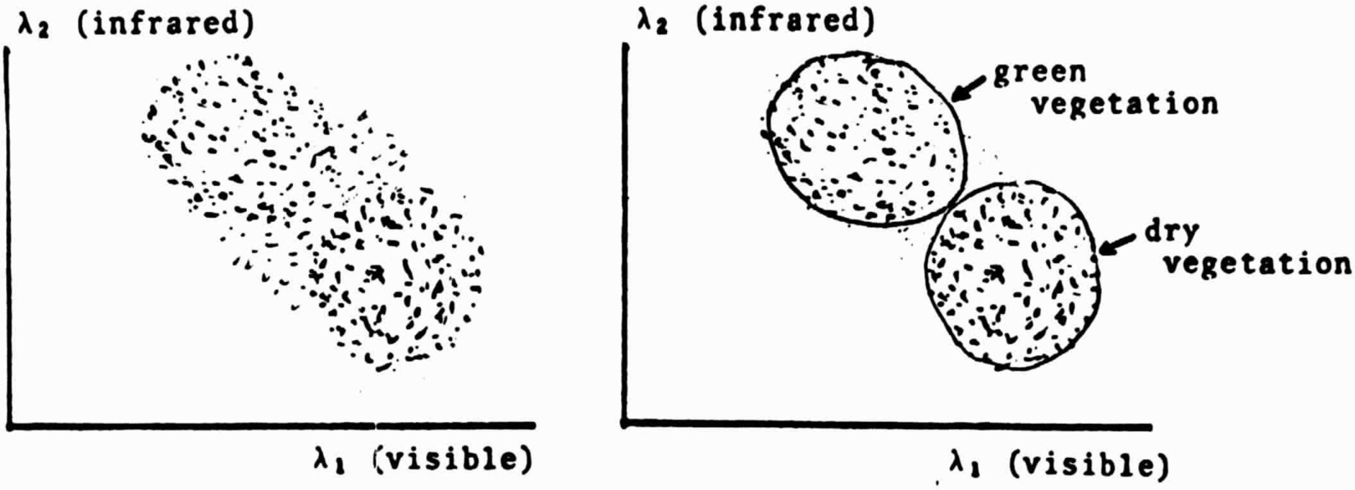
Input	•Data from Multispectral Image Storage Tape •Control Cards to select the processing and output options •Field Description Cards indicating area(a) to be classified.

Process

The program uses the class mean vectors and covariance matrices and the data from each point to calculate the probability that the point belongs to each of the training classes. It then assigns the point to the most probable class.

Output	Classifiction Results File which is normally used as input to the PRINTRESULTS processing function to produce a variety of printed output for evaluation of the classification.

<u>CLUSTER</u> is a process that groups individual data points into a predefined number of groups (clusters) specified by the analyst.



Input

- Data from Multispectral Image Storage Tape
- Indication of the Area to be Clustered
- Number of Clusters Desired

Process

1. The computer assigns a location in the feature space as the initial center of each cluster.

2. It then calculates the distance between each data point and each cluster center and assigns the sample to the cluster with the minimum distance.

3. Next, new cluster centers are determined by calculating the mean vector for the data points assigned to each cluster. The covariance matrix is also calculated.

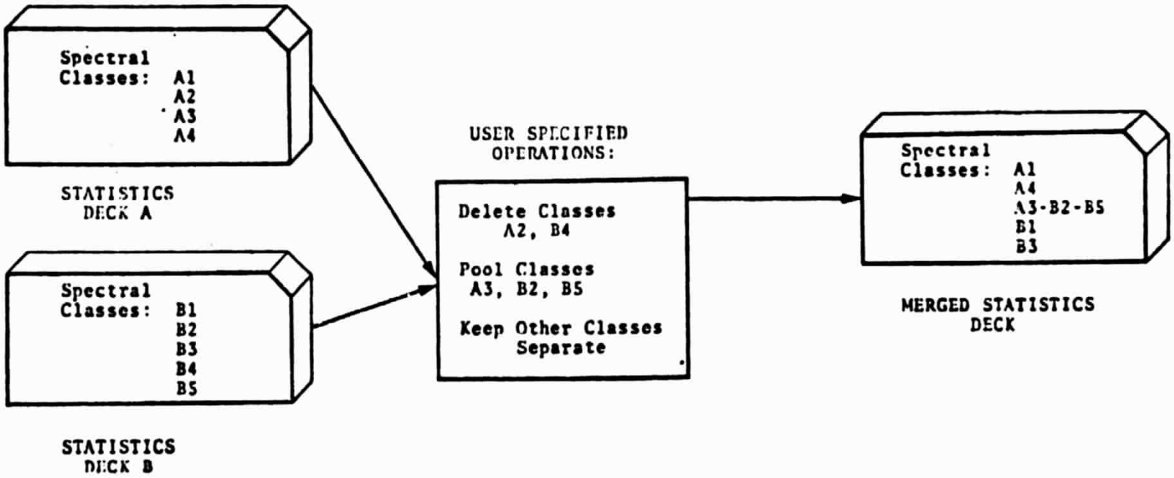
4. The computer then proceeds back to Step 2 and reassigns each sample to the closest newly defined cluster center.

5. The computer continues the cycle of calculating the cluster centers (Step 3) and reassigning data points (Step 2) until the percentages of data points that are reassigned to a new cluster center equals a value specified by the analyst.

Output

A Cluster Map which pictorially represents each area that was clustered and a summary of the number of points assigned to each cluster. Also available are tabular output, statistics and field description cards.

<u>MERGESTATISTICS</u> is a process that combines statistics decks or manipulates a single deck, according to user-defined specifications. Spectral classes may be kept separate, pooled (statistical combination of mean vectors and covariance matrices), or deleted in the merged deck.



Input

- One or more statistics decks
- User-specified operations for treatment of spectral classes.

Process

1. A new statistics deck will be created by removal of the "delete" classes.

2. For classes that are to be pooled, the computer calculates a weighted mean vector and covariance matrix from the original statistical parameters.

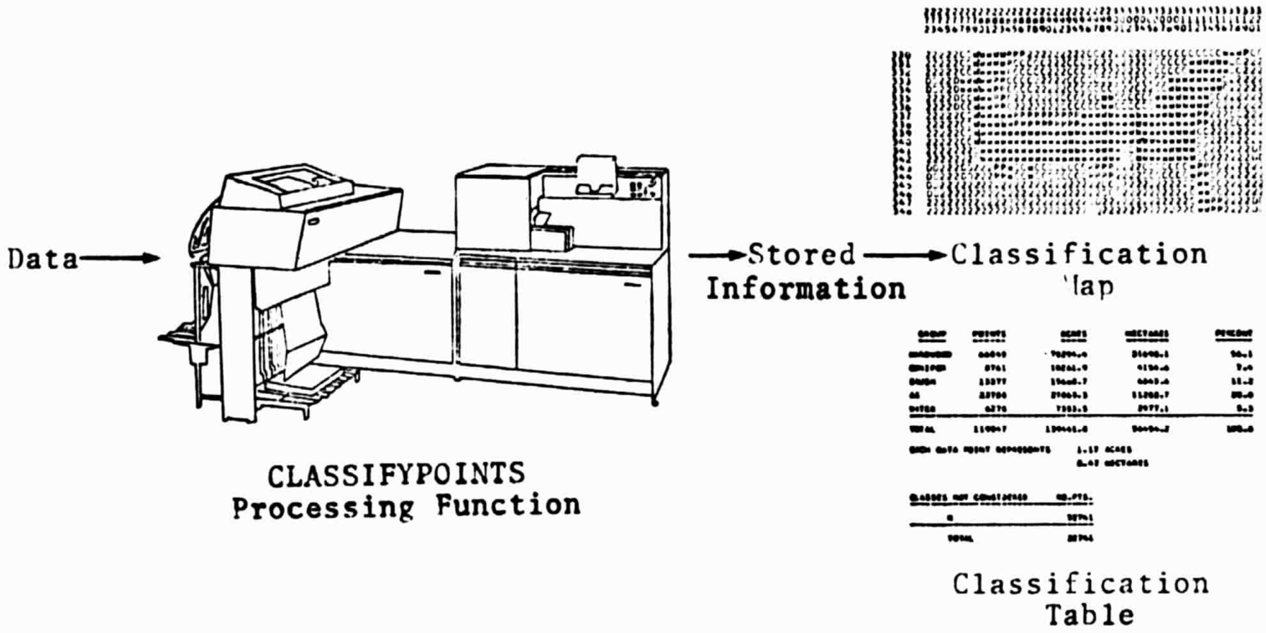
3. The computer renumbers all spectral classes according to their new, sequential order and assigns the user-designated names to the spectral classes.

4. If requested, the computer designs a coincident spectral plot that shows the percentage reflectance mean $\pm$ 1 standard deviation for each spectral class in each wavelength band.

Output

A single statistics deck with a new group of spectral classes. Also, a coincident spectral plot of all classes (if requested).

PRINTRESULTS produces a variety of printed outputs describing the results of a classification in the form of a map and/or tabular output.



Input

- Location (in computer file) of classification results to be printed.
- Symbols to be assigned to various classes.
- Area(s) to be used for map and tables.

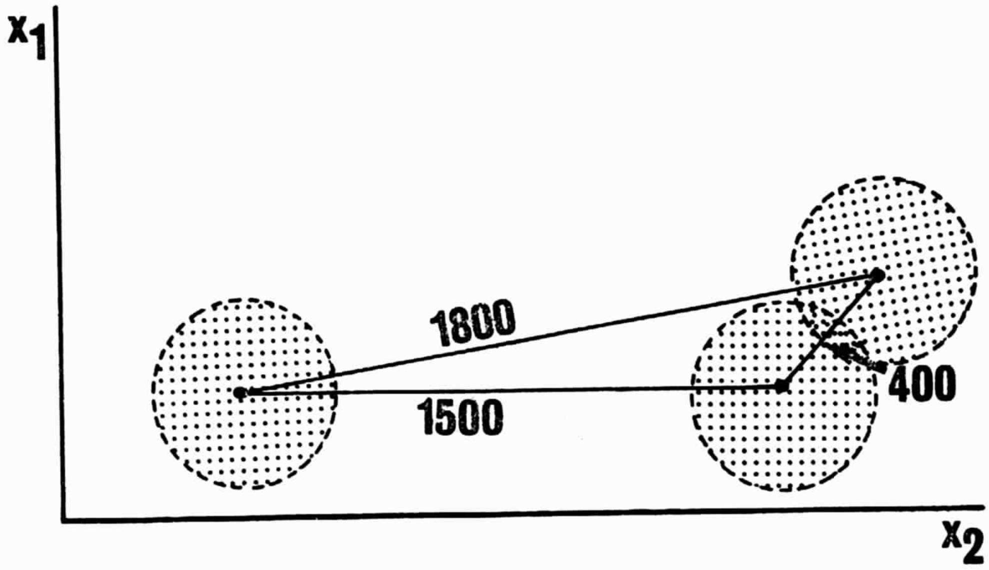
Process

The process prints out the information which is available in the computer on disk (or tape storage) that is a product of the CLASSIFYPOINTS processing function.

Output

- Classification map (using specified symbols) with outline of training and test fields (if requested).
- Tables showing training field and class performance.
- Tables showing test field and class performance.

SEPARABILITY is a process to calculate the "distance" between classes of interest (as a function of combinations of spectral bands).



Input

• Control cards to select processing and output options.

• Statistics files (cards or disk) from the Statistics function (mean vectors and covariance matrices) for each class.

Process

Calculate the "distance" (transformed divergence) between each indicated pair of classes based upon the mean vector and covariance matrices to determine how well the individual classes may be distinguished from one another.

Output

• Separability Results Listing indicates the transformed divergence between selected class pairs. The combinations are ranked according to degree of separability with the associated divergence for each class pair for each channel combination.

• Acreage estimates for each cover type in the entire planning unit or specified portions.

APPENDIX B

Answers to Self-Check Items:

I-A. 1) Specific Location 2) Exact cover type of interest 3) What applications will be made of the analysis results 4) What general level of accuract of classification results are needed.

I-B. Any response to this item could be correct. The only requirements is that if be useable by you and contain the four basic parts mentioned above.

II-A. 1) Aircraft collected multispectral data is one source 2) Satellite acquired data is another source available for forestry analysis

II-B. High quality is essential if high degrees of accuracy are desired. Data aberrations limit the classification capabilities of the analysis.

II-C. Haze, clouds, and in some cases, snow cover can be problems in data quality with which the analyst must contend. Mechanical problems can also cause data quality problems.

III-A. 1) If satellite data is being used correlated aerial photographs can be very useful 2) U.S.G.S. quadrangle maps are also useful reference tools 3) Direct ground observation can be a useful reference tool.

III-B. Any one or combinations of these reference tools can be used. Specific analysis objectives might make some reference data more desirable than others. The essential element is to select reference data which gives the best information available about the specific cover types of interest.

IV-A. 1) Select training areas from locations which have correlatable reference data. 2) The complexity of the ground scene in a potential area is also important. An approximate rule of thumb is to choose an area with at least two and usually no more that four distinct cover types. 3) A third consideration you might have used is to scatter the training areas throughout the planning unit to guard against missing a major cover type.

IV-B. Roughly 40 to 100 lines by 40 to 100 columns are used by analysts that do not have access to a Zoom Transfer Scope. Areas of about 47 lines by 34 columns are used if the Z.T.S. is available.

IV-C. 1) One might be to use your judgement to select between four and eight areas. 2) A second would be to select a percentage (from 1-15%) of the total planning unit as a guide for selection. The justification revolves around the complexity of topography, the diversity of cover types, the size of the planning unit, and the size of the training fields.

IV-D. Be sure to refer to your objectives and any reference data you have to aid you in deciding the size, number, and location of your training areas. If an instructor or co-worker is available, discuss your choices in detail with them.

V-A. The CLUSTER processor combines the spectral information from all the channels for the various cover types which results in increased <u>boundary enhancement</u>.

V-B. 1) Boundary enhancement 2) It determines cluster classes by combining data from all (four in this case) channels.

V-C. SEPARABILITY allows the analyst to determine the similarity between cluster classes and the probability of correctly classifying them.

V-D. The output of SEPARABILITY is a grouping table which identifies those spectral classes which are spectrally inseparable.

V-E. Transformed divergence is a number which indicates the relative distance between cluster classes based on their means and variances.

VI-A. Reference data can be placed beside generated information or overlain on a light table. A more desirable technique is to use the Z.T.S. which superimposes the two forms of data.

VI-B. A mixture class is one in which scattered symbols in a given area indicate spectrally inseparable cover types which must be grouped together.

VII-A. Since data from all training areas are being combined, a new grouping table is used to show which spectral classes could be combined from all the training areas.

VII-B. MERGE is used to pool these data at the various transformed divergence levels.

VII-C. A final statistics deck is created which describes the numerical characteristics of the various final training areas.

VII-D. This represents an error of commission since 33 points were committed to the aspen group when in fact they were points representing oaks.

VII-E. 95 points were actually aspen that were not included in the aspen group of points which represents an error of omission.

VIII-A. By taking the training statistics data and using it to classify each multispectral data point one at a time and assigning it to the class to which it is most statistically close the number of symbols is reduced and the boundaries between classes becomes more clearly defined.

IX-A. If the analyst is able to accomplish his analysis objectives, then the proper level of classification has been reached.

IX-B. 1) Changing the number of classes desired could require re-clustering 2) Changing the level of accuracy would cause re-examining the pooling process.

X-A. 1) Acreage Calculations gives the areal coverage of the different cover types in the planning unit.
2) CALCOMP maps are computer drawn maps that indicate the boundaries between cover types and in which symbols denote what class is found within that boundary.
3) Classification maps is a computer generated printout where symbols represent individual points (pixcels) that are classified as a specific cover type.
4) Color-coded image produced by the computer in which a color represents a certain cover type.
5) The accuracy of the classification map can be indicated by noting the accuracy of the classifier in the training fields.

XI-A. In this study the general land use map was one and determining the extent and location of specific cover types was another.

XI-B. 1) Forest management
2) Land use planning
3) Wildlife environment
4) Fire protection planning

XI-C. Many answers would be acceptable depending on your specific interest in forestry. Any of the above mentioned applications are suitable. The best answer is one that proves to be useful to you!

APPENDIX C

Aldrich, R. C. and R. C. Heller, 1969, "The Use of Multispectral Sensing Techniques to Detect Ponderosa Pine Trees Under Stress." University of California Publication.

Fleming, M. D., J. S. Berkebile, and R. M. Hoffer, 1975. "Computer-Aided Analysis of LANDSAT-1 MSS Data: A Comparison of Three Approaches, Including a 'Modified Clustering' Approach." Symposium Proceedings of Machine Processing of Remotely Sensed Data, June 3-5, 1975. 8 p.

Hitchcock, H. C. and R. M. Hoffer, 1974. "Mapping a Recent Forest Fire with ERTS-1 MSS Data." LARS Information Note 032674, Laboratory for Applications of Remote Sensing, Purdue University, 13 p.