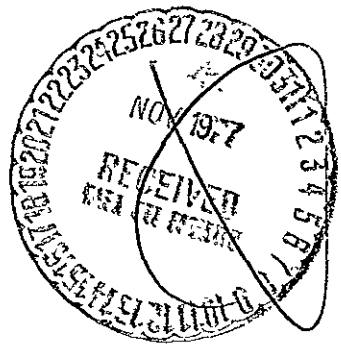
LARS Publication 090177

A Case Study Using ECHO☆ for Analysis of Multispectral Scanner Data

by Donna Scholz,
James Russell,
John Lindenlaub &
Philip Swain

☆Extraction and Classification of Homogeneous Objects



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

1977

LARS Publication 090177

A CASE STUDY USING ECHO
(EXTRACTION AND CLASSIFICATION OF HOMOGENEOUS OBJECTS)
FOR ANALYSIS OF MULTISPECTRAL SCANNER DATA

BY

DONNA SCHOLZ

JAMES RUSSELL

JOHN LINDENLAUB

PHILIP SWAIN

THE LABORATORY FOR APPLICATIONS OF REMOTE SENSING

PURDUE UNIVERSITY, WEST LAFAYETTE, INDIANA 47906

The work was supported by the National Aeronautics and Space Administration under contract NAS9-14970.

T-1314/4

STAR INFORMATION FORM

1. Report No.	2. Government Accession No	3. Recipient's Catalog No
090177		

4 Title and Subtitle	5. Report Date
A Case Study Using ECHO (Extraction and Classification of Homogeneous Objects) for Analysis of Multispectral Scanner Data.	September 1, 1977
	6. Performing Organization Code
7. Author(s) Donna Scholz, James Russell, John Lindenlaub, and Philip Swain	8. Performing Organization Report No. 090177
	10. Work Unit No.
9. Performing Organization Name and Address Laboratory for Applications of Remote Sensing Purdue University West Lafayette, IN 47907	11. Contract or Grant No NAS9-14970
	13. Type of Report and Period Covered Technical Report
12. Sponsoring Agency Name and Address NASA Johnston Space Craft Center Houston, TX	14 Sponsoring Agency Code

15. Supplementary Notes

This document is part of the series of materials making up the LARSYS Educational Package. Its primary use will be in conjunction with persons having access to the LARSYS software system.

16. Abstract This Case Study is a component of the LARSYS Educational Package. It is designed to introduce the ECHO processing function which may be implemented on a variety of computer systems. The typical steps in the analysis of remotely-sensed data using ECHO are illustrated through discussion, an illustrative example and exercises. The exercises have been written for implementation on a computer using LARSYS: however, they may be modified for other analysis systems.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
Remote Sensing; Multispectral Analysis; Object Seeking; Boundary Seeking; Context; Image Partitioning; Sample Classification	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified		

*For sale by the National Technical Information Service, Springfield, Virginia 22161

NASA — JSC

TABLE OF CONTENTS

A Case Study Using ECHO (Extraction and Classification of Homogeneous Objects) For Analysis of Multispectral Scanner Data

by Donna Scholz, James Russell, John Lindenlaub, and Philip Swain

PREFACE

This Case Study is a component of the LARSYS Educational Package. It is designed to introduce you to the ECHO processing function which may be implemented on a variety of computer systems. This publication illustrates the typical steps in the analysis of remotely-sensed data using ECHO through discussion, an illustrative example and exercises for you to complete at a terminal. The exercises have been written for implementation on a computer using LARSYS; however, they may be modified for other analysis systems.

Prerequisites

The material presented herein is based upon the assumption that you have mastered the instructional objectives in the first five units of the LARSYS Educational Package and either Unit VI or VII;[1]

Unit I	An Introduction to Quantitative Remote Sensing
Unit II	LARSYS Software System: An Overview
Unit III	Demonstration of LARSYS on the 2780 or Data 100 Remote Terminal
Unit IV	The 2780 or Data 100 Remote Terminal: A "Hands-On Experience"
Unit V	LARSYS Exercises
Unit VI	Guide to Multispectral Data Analysis Using LARSYS
or	
Unit VII	A Case Study Using LARSYS for Analysis of Landsat Data

Purpose

The purpose of this case study is to train you to use the ECHO analysis procedures. However, you should not expect to be a proficient ECHO analyst after completing this case study. It should give you the necessary fundamentals. Proficiency must come with additional experience.

As you probably are aware, the experience of the analyst is a very important factor in the man-machine analysis of multi-spectral scanner data as described in this case study. Donna Scholz who did the analysis of the Grand Rapids area used as the example has three years of experience with computer-aided analysis of multispectral data.

---

[1]The LARSYS Educational Package may be obtained from the System Services Manager, Laboratory for Applications of Remote Sensing, 1220 Potter Drive, West Lafayette, Indiana 47906.

General Objective

The analysis of a set of multispectral scanner data can be broken down into a sequence of steps. When you have completed this case study, you should be able to list the steps of the analysis sequence in the proper order. Furthermore, for each step in the analysis sequence, you should be able to:

1) EXPLAIN the significance of the analysis step with respect to the entire analysis sequence.

2) NAME and DESCRIBE any software tools useful in carrying out the analysis step.

3) APPLY the analysis principles to a specific analysis problem by writing control commands, running the program and interpreting the results of the functions used in the step.

Student-Tutor Interaction

While this case study 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 ECHO. This is especially true when you begin carrying out the Exercises. You are encouraged to discuss your progress periodically with your tutor.

Format

Each section of the case study follows this format: the instructional objectives are stated, followed by a discussion of the purpose, philosophy, and analysis techniques associated with that step in the data analysis sequence; then there is an example showing control cards, computer output, and an interpretation of the results. Exercises are provided at the end of the Case Study to test your mastery of the basic analysis procedures.

The material is presented in this format so that a person wishing to become adept in the analysis of multispectral data using ECHO can proceed through this case study and learn what the analysis steps are, why each step is important, how each step is carried out, and gain practice in using these analysis techniques.

References

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

references should be in your site library. Be sure you have them available before you begin the case study. See your tutor if you need help locating them.

Acknowledgements

The authors wish to thank the following LARS personnel: Barbara Davis, Research Statistician; Forrest Goodrick, Data Analyst; and James Kast, Program Developer who served as a committee of subject-matter advisors and provided valuable input for this case study. Numerous others who reviewed rough drafts of this material made important comments and suggestions which were incorporated into the case study.

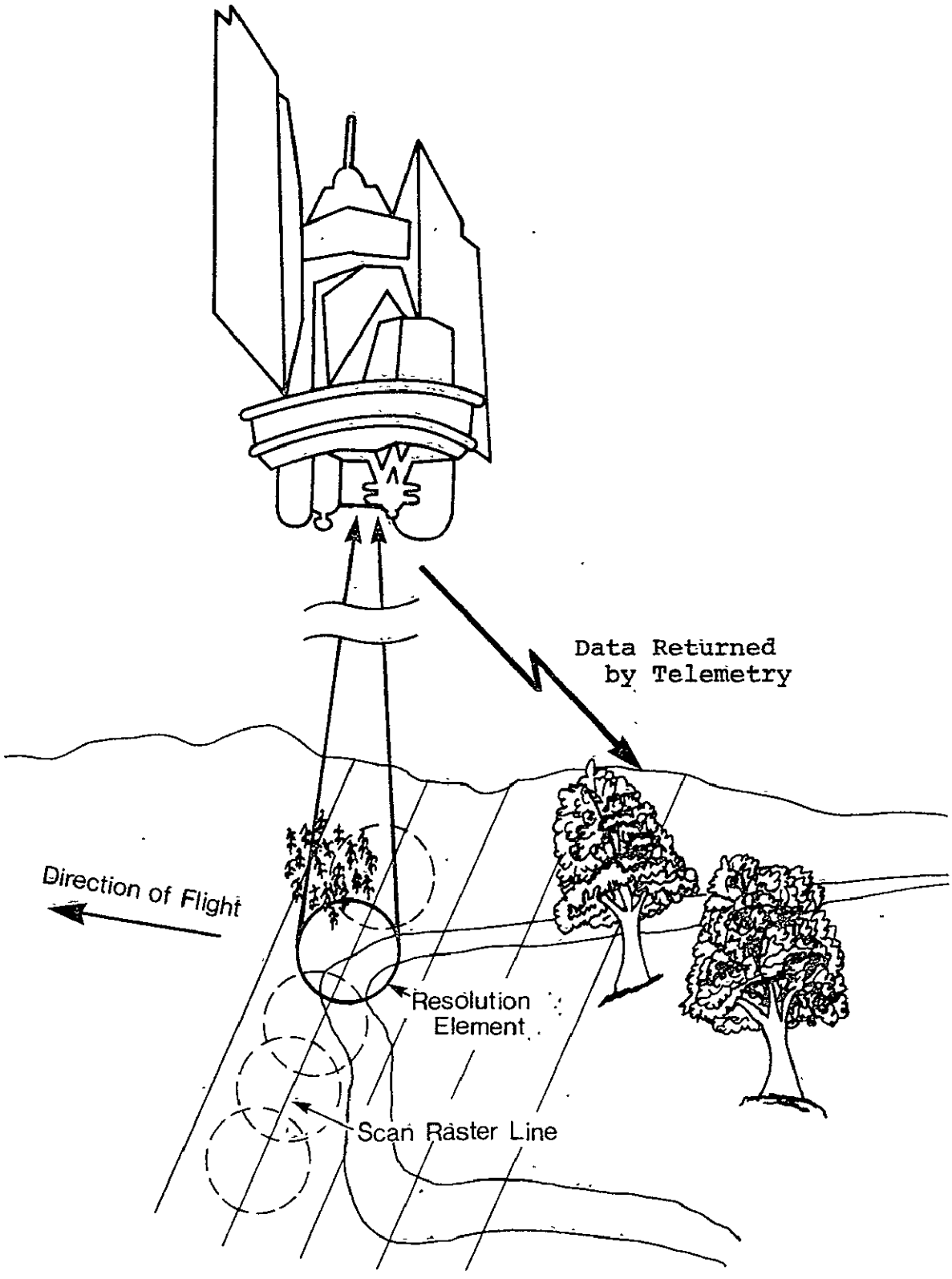


Figure 1. Multispectral scanner data collection system.

INTRODUCTION AND BACKGROUND

ECHO is an acronym for Extraction and Classification of Homogeneous Objects. This case study will provide an opportunity to study and work with the supervised ECHO classification algorithm. ECHO, like many other computer implemented classification schemes, is based on the principles of pattern recognition. In order to gain an understanding of both the capabilities and limitations of the algorithm and to place it in context with other classification algorithms, we will begin with a brief review of pattern recognition principles as they apply to remote sensing.

Figure 1 shows a schematic of a multispectral data collection system as it scans a portion of the earth's surface. Each scene element (picture element or pixel) results in a data vector, X, as the output of the sensor. The scanning action of the instrument results in the scene being represented by an array of picture element data vectors as shown in Figure 2. The data vectors serve as input to the pattern recognition system. The outputs of the pattern recognition process include a classification map and tabular summaries of the classification.

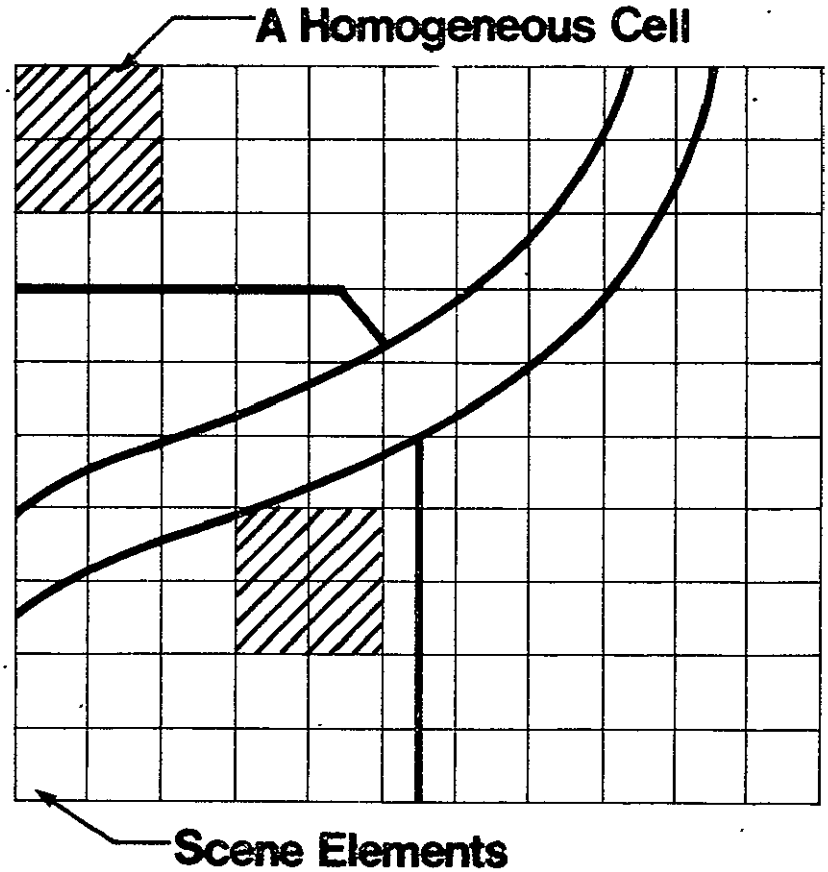


Figure 2. Array of scene elements. Heavy lines indicate boundaries of "objects" within the scene. Cross-hatched areas indicate two examples of homogeneous cells (i.e. groups of data vectors).

There are several approaches to classifying multispectral scanner data using pattern recognition [1,2]. Each picture element (pixel) can be classified individually or picture elements that have been determined to have similar spectral characteristics can be classified as a group. In the former case, called "per point" classification, a classification decision must be made for each individual measurement vector. In the latter case, called "sample" classification (from statistical sample), a classification decision is only required for each group of measurement vectors (each sample or, in the context of Figure 2, each object in the scene). One advantage of sample classifiers is that for a given scene fewer classification decisions need to be made. This results in a savings in computation time. This savings is partially offset by the fact that it requires some computation time to determine the objects or samples comprising the scene.

ECHO is a sample classifier which uses a conjunctive object finding algorithm to determine the objects (samples) in the scene to be classified [3]. A conjunctive algorithm begins with a very fine partition of the scene and simplifies it by progressively merging adjacent cells that are found to be similar according to certian statistical criteria. The objects are then classified using a maximum likelihood sample classification algorithm [2]. In this manner ECHO benefits from the spatial correlation of picture elements belonging to the same object.

ECHO is a useful tool for numerical data analysis of a scene consisting of areas large in areal extent relative to resolution element size. For LANDSAT data, the smallest area which can be resolved is about .45 hectares (1.1 acres). When using the point-by-point algorithm, each individual data point (.45 hectares) is classified. In the ECHO analysis procedure, homogeneous groups of data points are classified as a group. The resultant classification map will contain less 'noise' (isolated single data points) and the areas will have more distinct boundaries. (See Figures 4 and 5). The differences between the two approaches becomes most dramatic when the homogeneous areas are significantly larger than the resolution of the sensor.

---

[1] Lindenlaub, J.C. and J.D. Russell. "An Introduction to Quantitative Remote Sensing." LARS Information Note 110474, Purdue University, 1974.

[2] Swain, P.H. "Pattern Recognition: A Basis for Remote Sensing Data Analysis." LARS Information Note 111572, Purdue University, 1972.

[3] Kettig, R.L. and D.A. Landgrebe. "Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects." LARS Information Note 062375, Purdue University, 1975.

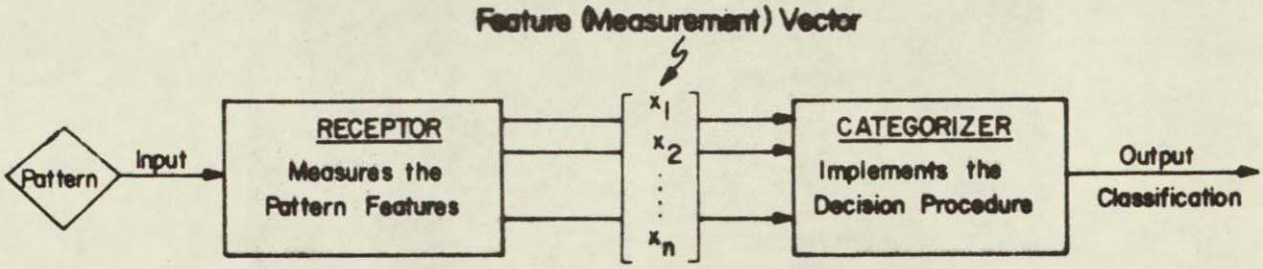


Figure 3. Model of a pattern recognition system. The categorizer makes classification decisions based upon inputs in the form of measurement vectors.

Classification algorithms can also be categorized as being supervised or unsupervised. In the supervised case the analyst must provide the computer with information about the spectral character of each class of interest. This is usually done by specifying the coordinates of data points known to represent the classes of interest. In the unsupervised case the computer examines the spectral characteristics of the area to be classified and divides (or classifies) the data into groups having similar spectral characteristics. Unsupervised classification algorithms operate without the benefit of a description of the spectral characteristics of the cover types of interest. They classify data strictly on the basis of structure inherent to the data. The generalized steps in classifying multispectral data are shown in Figure 6. This figure points out the difference between supervised and unsupervised classification. Both supervised and unsupervised ECHO classification algorithms have been developed [4]. This case study, however, will only make use of the supervised ECHO classifier.

The analysis of earth resources data is a dynamic process requiring close coordination between the user and the analyst and close interaction between the analyst and the data processing system. Although Figure 6 shows the steps in a linear or sequential fashion, the steps are actually quite interrelated and interpretation of the results at any one step might necessitate going

---

[4]Landgrebe, D.L. "Test of Boundary Finding Per Field Classification," Final Technical Report, Volume I, NASA Contract NAS9-14970. Laboratory for Applications of Remote Sensing, Purdue University, May 1977.

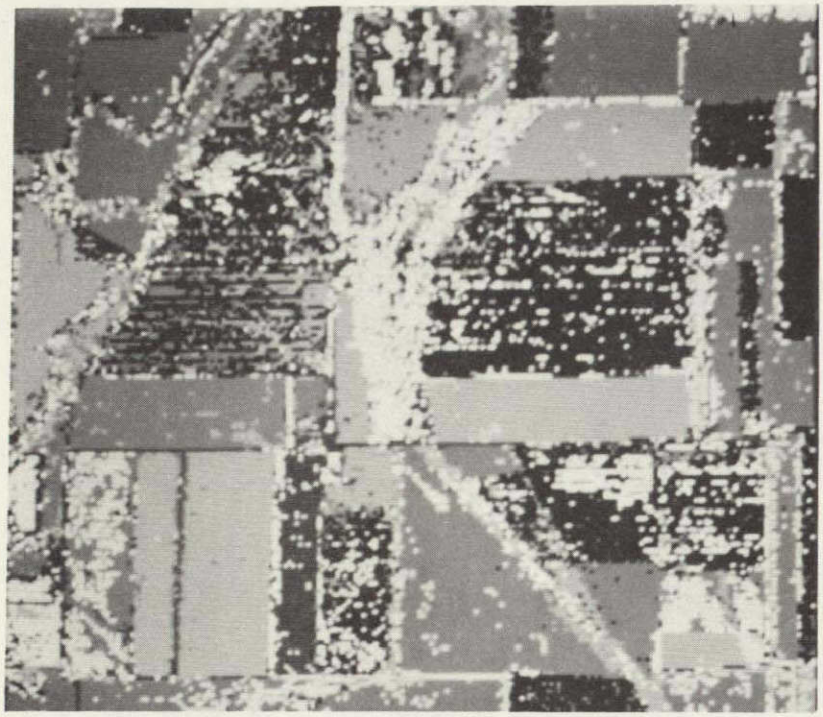


Figure 4. Gray-Scale-Coded Classification Map
Produced Using Per Point Classification
From Kettig and Landgrebe

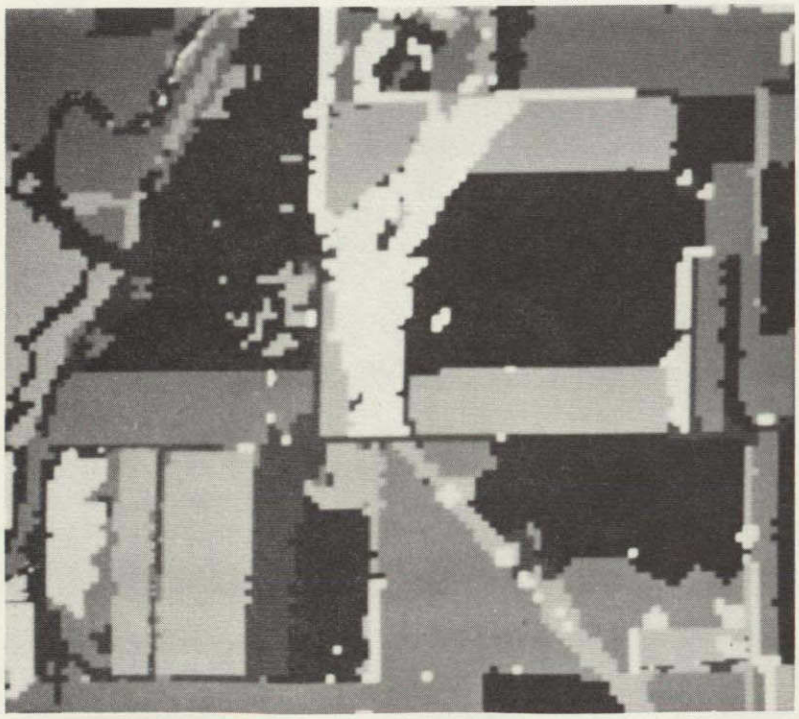


Figure 5. Gray-Scale-Coded Classification Map
Produced Using ECHO.
From Kettig and Landgrebe

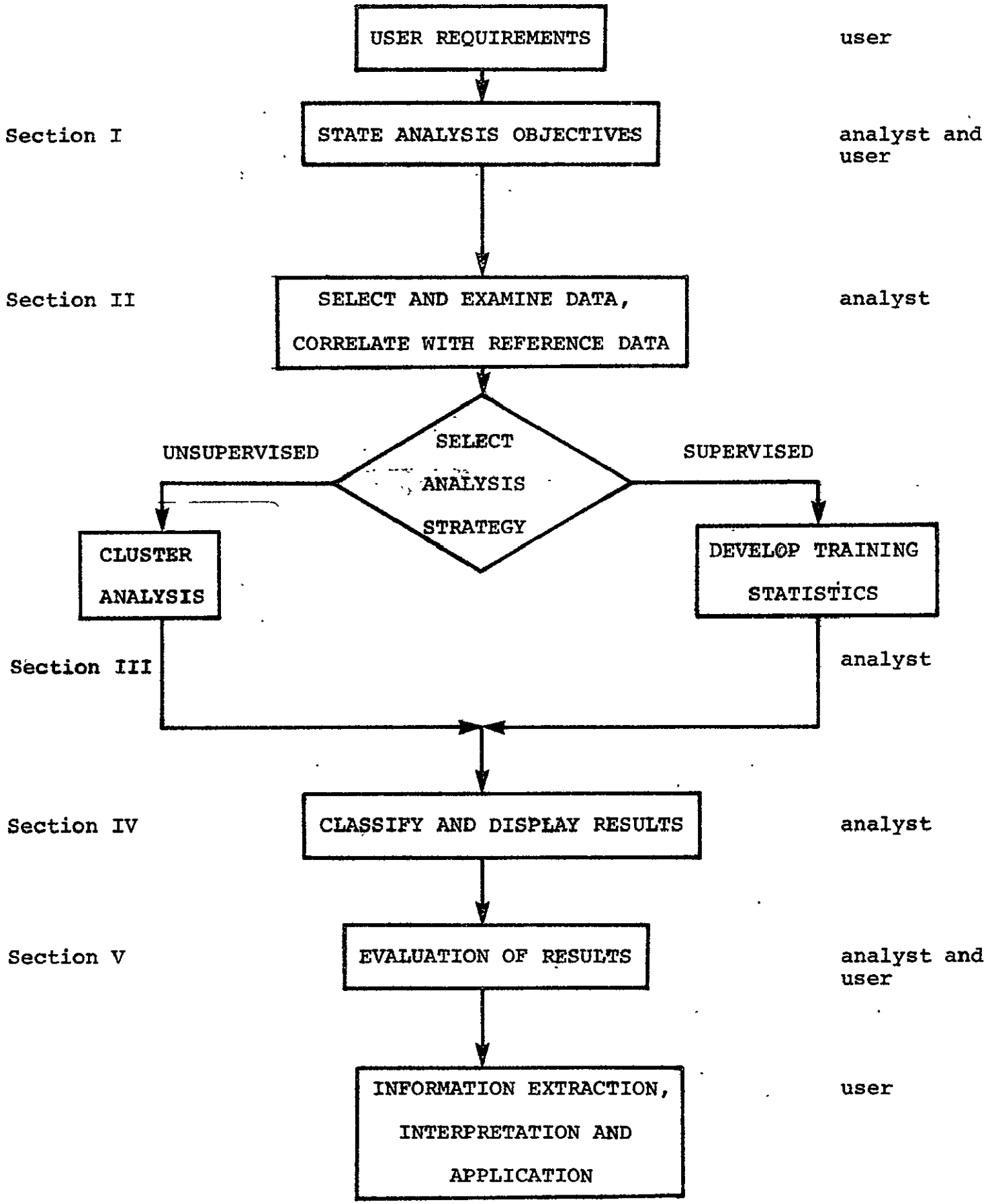


Figure 6: Generalized Steps in Classifying Earth Resources Data

back to earlier steps and revising the analysis. As an example, the user states his requirements and works jointly with the analyst to formulate the analysis objectives. After proceeding through the analysis and evaluation of the results, the conclusion might be reached that it is impractical to achieve the original analysis objectives. This could result from an objective which was too ambitious for the data available or a lack of understanding of the capabilities of the technology. In any case, a restatement of the objectives and additional analysis, i.e., a reiteration of the steps shown in Figure 6, might lead to a satisfactory set of results.

This case study will concentrate on those steps in Figure 6 in which the analyst plays an active role and will stress the interaction of the analyst and the data processing system.

As suggested above it is best if the analyst can work with the user in formulating the analysis objectives. The user knows his requirements, the analyst is familiar with the technology and together they can formulate a reasonable set of objectives for the analysis. The user can provide information about the location of the analysis study site, the types of ground cover of interest, how the analysis results will be used, and the degree of classification accuracy desired. This information will serve as a guide for the analyst at decision points during the classification procedure.

The analyst then selects a suitable data set to use for the analysis. In doing this, consideration is given to the type of data collection system, how readily available the data is, and time of year data was collected.

As might be expected, the quality of the data set to be analyzed has a direct impact on classification performance. Factors such as clouds or bad data lines may adversely affect the accuracy. Therefore, it is necessary to examine a prospective data set for these factors.

Correlation of reference information with the multispectral scanner data is a task that is important in helping the analyst relate to the ground cover types and their distributions in the scene. This may be accomplished by using reference materials such as maps, photographs, ground observations, computer gray scale printouts or digitally-derived imagery. Becoming familiar with features in the scene will facilitate the analyst's efforts in selecting training areas which are representative of the variations in the data. Usually at this point, the analyst is able to make at least a tentative decision on the type of analysis strategy appropriate for the job at hand.

After selecting and becoming familiar with the data set and assembling reference data the next step in the supervised analysis procedure is to develop training statistics. This process is begun by selecting candidate training areas. These training

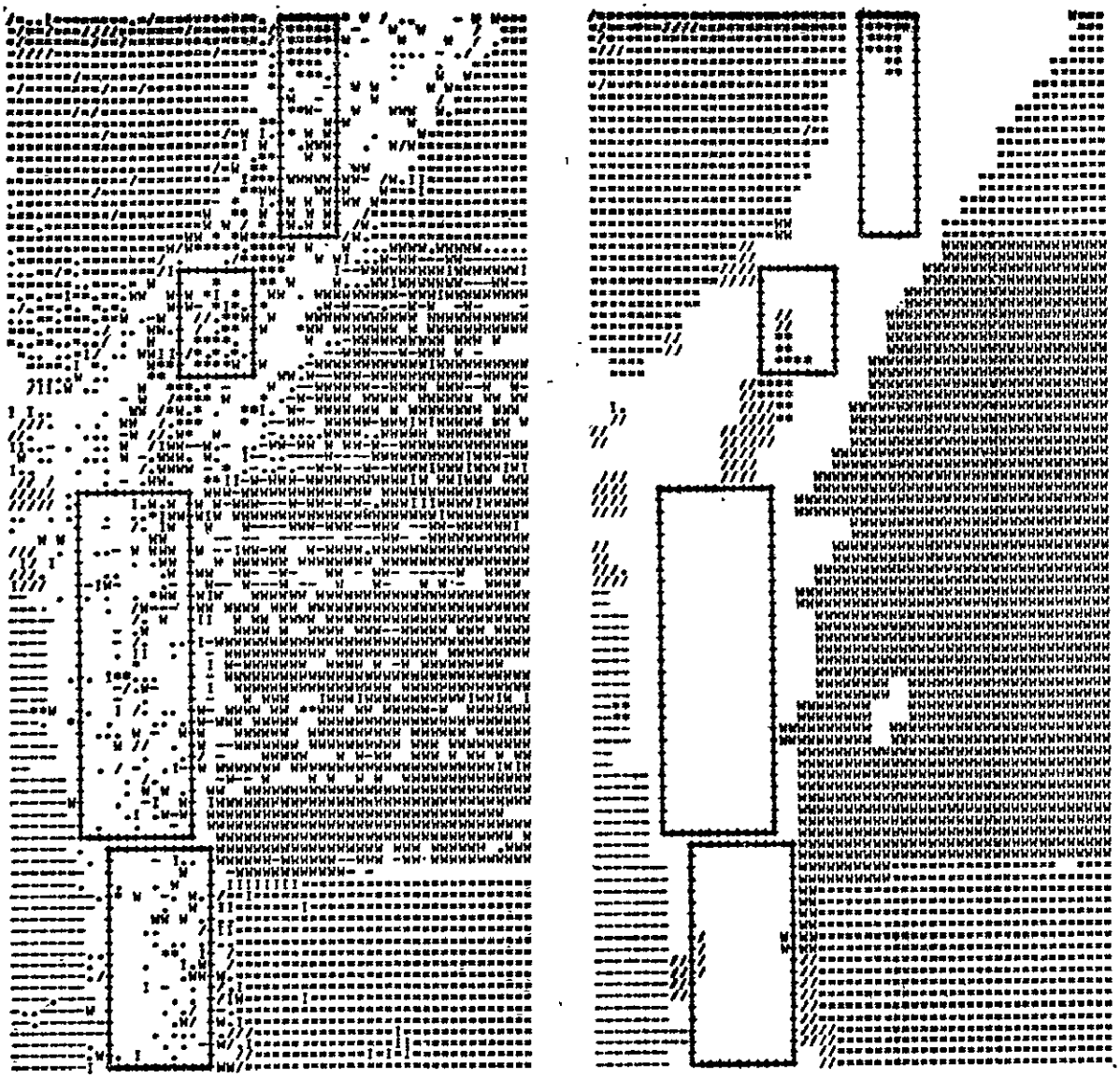
areas are groups of contiguous points within the scene which contain examples of all the classes present in the area to be classified. These areas will be used to "train" the computer to recognize specific classes of interest.

Each of the training areas is usually selected to contain more than a single ground cover type. It is necessary to determine the spectral subclasses associated with each of these cover types. The technique utilized for this task is called "cluster analysis." Cluster analysis groups data measurement vectors into groups or clusters representing inherent or natural structure within the data by utilizing spectral reflectance values in different portions of the electromagnetic spectrum. These clusters are then related to specific ground cover types and identified as a specific cover type or information class (i.e. water, forest, bare soil, etc.).

Since most of the ground cover types occur in more than one training area, some of the cluster classes from separate areas are statistically similar and only one of the multiple cluster classes is necessary to characterize a ground cover type. For this reason, it is then appropriate to calculate the similarity among all the cluster classes from all the training areas. Calculating the separability between the classes gives the analyst a "feel" for the probability of getting a correct classification using the statistical characterization of the classes he has selected. Once the training classes have been defined on a spectral basis and refined to a state that all are distinct and separable from one another, the resultant means and convariance matrices may be used as input to the ECHO classification algorithm.

The supervised ECHO algorithm is then used to classify the area of interest. Supervised ECHO examines blocks or cells of data consisting of several data points and tests them for homogeneity. If all points in the cell are similar, the ECHO classifier assigns that entire cell to the spectral class it most nearly matches. If a cell does not contain similar data points, the cell is subdivided into its individual data points and each point is classified independently. In this manner, similar portions of the study area will be identified as homogeneous blocks or "objects". (See Figure 7).

The classification results may now be displayed for the analyst to examine. The results might be displayed on a CRT screen as levels of gray or color coded, or a map might be generated on a computer line printer whereby each classified point was assigned a symbol corresponding to the class in which the point was placed by the ECHO classifier. The type of output product is best determined by the needs of the user as specified by the analysis objectives.



Per Point ECHO

SYMBOL	CLASS	SYMBOL	CLASS
W	Wheat	I	Idle
=	Lespedeza	*	Forest
-	Pasture	.	Corn, Soy, Rye, Hay
blank	Wooded Pasture	/	Non-Farm

Figure 7. Classification Maps. The four rectangular areas are wooded pasture, classified much more accurately by ECHO. From Kettig and Landgrebe, Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects.

The final step in the data analysis sequence is to evaluate the classification results. This can be accomplished by calculating the performance for a set of representative test fields (areas of known covertype within the study site, distinct from the training areas). If performance is judged to be below the level required, it might be necessary to go back in the analysis sequence and select additional training areas or otherwise refine the training statistics.

The analysis objectives should be reviewed with the user at this point to see that they have been met. If procedural decisions were made with the analysis objectives in mind, the resultant classification should be the type and quality of product the user desired.

With this general description of the supervised ECHO classification algorithm and a discussion of how it fits into the sequence of analysis steps, we are ready to begin the case study.

SECTION I STATEMENT OF ANALYSIS OBJECTIVES

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

1. IDENTIFY the four components of an analysis objective.
2. PREPARE an analysis objective incorporating these components.

The first and one of the most important steps in the quantitative analysis process is stating the analysis objectives. Some of the questions which your analysis objectives should answer are: What is the problem to be solved? What information do you need to solve the problem? What are you going to do with the results of the analysis? The statement of analysis objectives is often done jointly by the analyst and user of the analyst's results.

For example, the purpose of the analysis could be for agricultural applications:

> "Map the agricultural soil types of Kearny County, Kansas, with 85% accuracy. The map will be used in conjunction with other data to make recommendations for a county-wide drainage system."

> "Prepare a crop inventory (acerage and crop type) during June of 1977 for the state of Illinois. It is desired to estimate crop yield within ± 10%."

Land use management and future highway construction can also be facilitated by using ECHO analysis of Landsat data.

> "Locate with 70% accuracy suitable cover types and wetlands (habitat diversity) for wildlife species in the Superior National Forest as an aid in managing wildlife openings and waterholes."

> "Generate a soil type map with at least 80% accuracy of the Point Comfort (Texas) quadrangle to aid in the selection of the most desirable right-of-way for a new highway."

> "Identify with 90% accuracy fire hazard areas such as timber sale areas, brushland, and broomsedge fields in the Hoosier National Forest in order to make decisions concerning fire crew size and placement during the major fire seasons."

"Determine the land use (urban, agricultural, forest and water) in Tippecanoe County (Indiana) for locating alternative routes for Highway 43. 80% accuracy is acceptable."

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 hectares using airborne multispectral scanners) or a relatively large area (thousands or millions of hectares using multispectral scanner data from satellite-borne systems).

Covertypes. What types of ground cover are of interest? Are woodlands, agriculture, rangeland, pasture, barren land or marshes the ground covers of interest? Are finer divisions, such as individual crop types, needed? Do the classes of interest occur as areas considerably larger than the resolution of the sensor to be used (larger than 0.47 hectares using Landsat)? The ECHO processor assumes the areas to be large compared to the sensor resolution.

Applications. How will the analysis output be used? To detect changes in an urban area? To evaluate alternate highway routes? To determine the proportion of an area (i.e., county or state) used for agricultural or industrial purposes?

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

The analysis objective will serve as a guide by which decisions are made during the analysis procedure. Therefore, it is essential to include all critical information when stating the objective(s).

As you proceed through this ECHO case study the example used to illustrate each analysis step will include a description of what analysis decisions and procedures an experienced data analyst took at that step of the analysis sequence. These example steps are intended to suggest typical courses of action in an analysis using ECHO.

Self-Check[1]

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

I-B. Write an analysis objective that you might use in solving a problem in your area of interest.

[1]Answers to all Self-Checks are given in the Appendix.

EXAMPLE

In this example, the analyst met with a planning consultant employed by the Grand Rapids, Michigan, city zoning board which would be the final user of the ECHO classification results. At this meeting it was mutually decided that the classes of interest would include:

Bodies of water
Agricultural lands
Commercial areas
Old residential areas
Newer residential locales

Further discussion between the consultant and the analyst brought out the fact that the zoning board was interested in areas of land use about 2 hectares in size and not the more finely detailed aspects of Grand Rapids land use. At this point the analyst was able to tentatively decide that the ECHO processor's capability to smooth out small scale variations in land use cover types would make ECHO a data analysis tool well suited to the zoning board's analysis needs.

The zoning board wanted the results to be a geometrically correct computer map coded for the land uses of interest and scaled to 1:24,000. The area was to include Grand Rapids proper and peripheral agricultural lands covered by six U.S.G.S. 7½ minute topographic maps.

The following objective was agreed upon by the consultant and the analyst:

> "Produce a general land use map of the Grand Rapids, Michigan,area locating the broad land use categories of commercial areas, older residential neighborhoods, newer residential sites, agricultural and recreational lands and bodies of water including lakes and rivers. These land uses should be located and identified with an accuracy of at least 80%. The land use map should be at a scale of 1:24,000 and suitable for use as a base map by the Zoning Board for zoning planning."

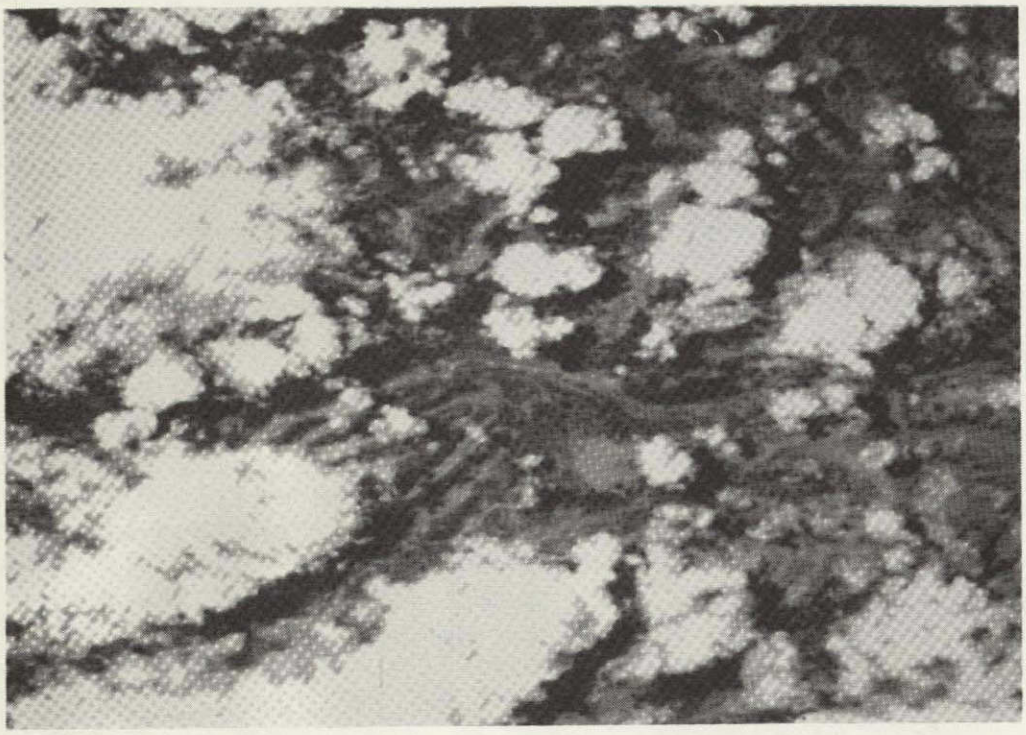


Figure 8. An example of clouds and their shadows.

Figure 9. Example of striping of data.

SECTION II SELECT AND EXAMINE DATA, CORRELATE WITH REFERENCE DATA, AND SELECT ANALYSIS STRATEGY

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

1. DESCRIBE two characteristics of data which may decrease usefulness.

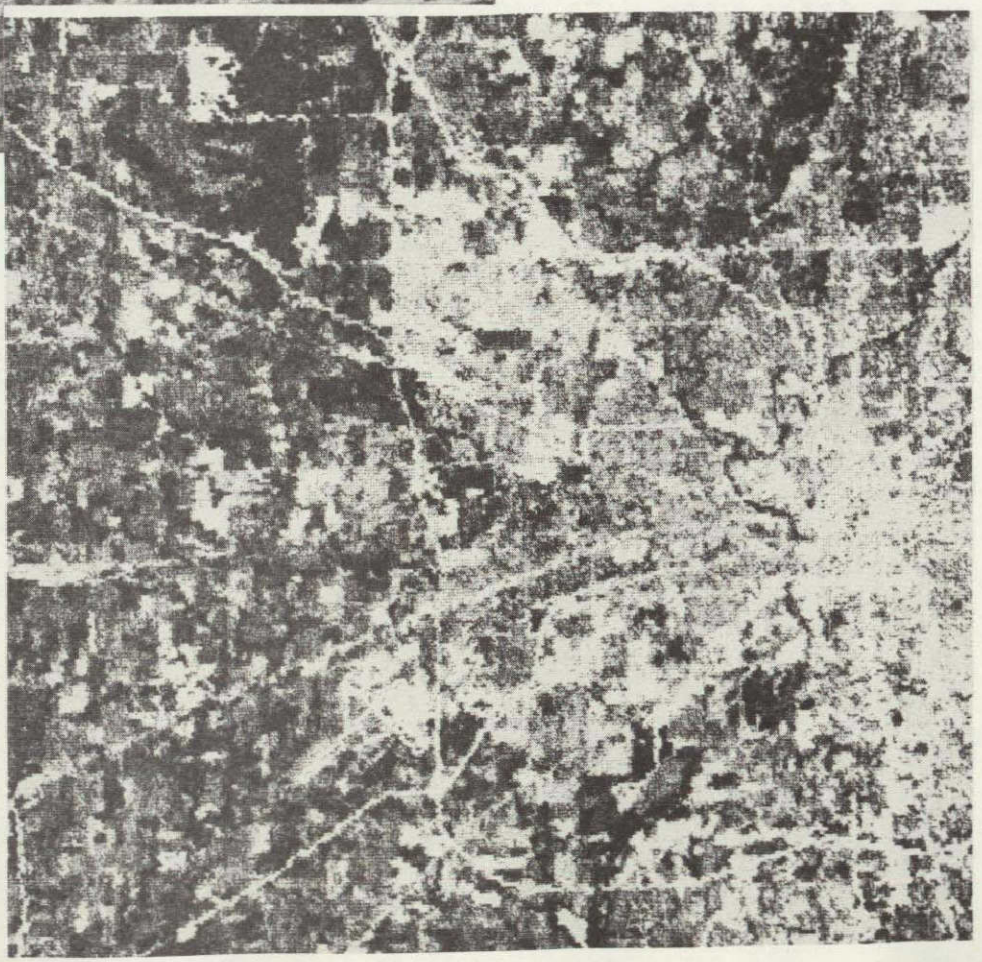
2. IDENTIFY two types of preprocessing that aid the analysis of LANDSAT data.

3. NAME three types of reference data.

After the analysis objectives have been stated, data suitable for the analysis must be selected. The analyst will want to consider such factors as data availability, spatial resolution, and dates when data were collected before selecting a data set. In carrying out an analysis, serious consideration must be given to data quality. In general, a greater level of analysis accuracy is possible when the original data is of the highest quality. A preliminary evaluation of digital data can be made by inspecting imagery created from the data. If, upon examining the imagery, one finds the area of interest totally obscured by clouds, meaningful analysis of that data set will not be possible. Gross data characteristics which may decrease the usefulness of a data set include haze, clouds, and snow cover. Data sets can be screened for these characteristics by examining digital display images or grayscale computer printouts. (See Figure 8).

Sometimes systematic idiosyncrasies such as "striping" will occur in the image. In the scanner system for Landsat 1 and 2, six data lines are simultaneously recorded in each wavelength band each time the scanner mirror oscillates. A separate detector is used for each channel of each of these scan lines. If any of these detectors and their electronics are not properly matched or calibrated, a striping effect is noticeable in the imagery of that channel. A dramatic example is shown in Figure 9. Even though striping appears to have seriously degraded the data, analysis may still be possible since striping usually occurs in only some of the channels. For example, since Landsat 1 and 2 have four multispectral scanner channels, it may be possible to get meaningful analysis results when only two or three of the four channels are available.

Original



Geometrically
corrected

Figure 10. Comparison of original and geometrically corrected and rotated Landsat imagery over Grand Rapids.

Once a data set has passed an initial screening, the analyst may request some preprocessing of the data to make it easier for him to correlate the remotely sensed data with reference data, such as maps. Preprocessing can facilitate the analyst's ability to interact with the data and produce output at the desired scale. One type of preprocessing (geometric correction) involves rotating and deskewing the image so that the vertical columns are aligned in a north-south orientation. Imagine yourself trying to locate corresponding points between a skewed LANDSAT image and an aerial photograph or map. The task would be much easier if the two images were oriented in the same direction. In another geometric correction, the data is rescaled. The scale of the data can be adjusted to allow the data to be overlaid on the reference data, such as a highway map or U.S.G.S. topographic map.

Illustrations of the effects of the rotation and geometric correction operations are shown in Figure 10.

Next, multispectral scanner data is correlated with available reference data. This not only allows the analyst to gain familiarity with the geographic region being analyzed, but it will aid him in developing good training statistics and in evaluating the classification results later in the analysis sequence.

Reference data is any information which aids the analyst in his task. There are several types of reference data which may be available. Typical examples are aerial photographs, maps, previous analysis results and on-site observations.

Aerial photography is very useful. Four types of film are commonly used: black-and-white, black-and-white infrared, color, and color infrared. The type of film selected is determined by the nature of the application and the conditions under which it will be used. Infrared camera systems are often preferred for high-altitude photography because of their haze penetration quality. Infrared films are also useful for enhancing differences between types of vegetation.

Maps are often used as reference data. A USGS topographic map may prove to be a valuable piece of reference data. Remotely sensed data can be preprocessed so that a computer grayscale "map" has the same scale as the "topo" map. Then a simple overlay technique could be used to correlate the MSS data with the USGS map. By overlaying the grayscale map of the area of interest on the quadrangle maps of the same area, features of interest to the analyst can be located in the data and their row and column numbers noted.

A third type of reference data involves direct ground observation by someone trained to observe relevant characteristics of ground features of interest to the analyst. Ground observations have been recorded which range from biological assemblages to water resources.

Self-Check

II-A. What are two data characteristics which may decrease the usefulness of your Landsat data?

II-B. Name two types of geometric preprocessing which aid in the analysis of Landsat data.

II-C. What are three types of reference data which can be correlated with remotely-sensed multispectral data?

The analyst obtained a grayscale computer map of bands 5 (.6 - .7μm) and 7 (.8 - 1.1 μm) of the Landsat frame previously selected for this study. It was discovered immediately upon examination that the data contained striping in at least one of the channels. Consequently, that particular frame was judged to be below the quality required for the project and another Landsat frame (scene I.D. 1411-15581) was selected.

An alphanumeric grayscale map was generated for this new data set using the control cards shown in Figure 11. No apparent striping or bad data lines were in evidence in the data, nor were excessive clouds in the scene. A graytone map of this same data was printed on a Varian plotter and is shown in Figure 12.

In the analysis of the Grand Rapids, Michigan, area the analyst had available the six 1:24,000 scale USGS topographic maps as well as color infrared aircraft photography. Using these pieces of reference data and the grayscale computer printout the analyst oriented himself and began identifying major features on the grayscale map.

In this step of the analysis the analyst was not interested in locating specific objects. Rather he wanted to establish a familiarity with the scene: Where, in general, is the urban area? Where is the river or agricultural areas? This general information will be of use to the analyst in the next step of the analysis.

```
-COMMENT PICTUREPRINT OF GRAND RAPIDS MICH FOR ECHO CASE STUDY
-RUNTABLE
DATA
RUN(72054806),TAPE(454),FILE(2)
END
*PICTUREPRINT
DISPLAY RUN(72054806),LINE(340,660,1),COL(90,360,1)
CHANNELS 6,8
END
```

Figure 11. Control card listing for generating a grayscale map of the Grand Rapids study site.

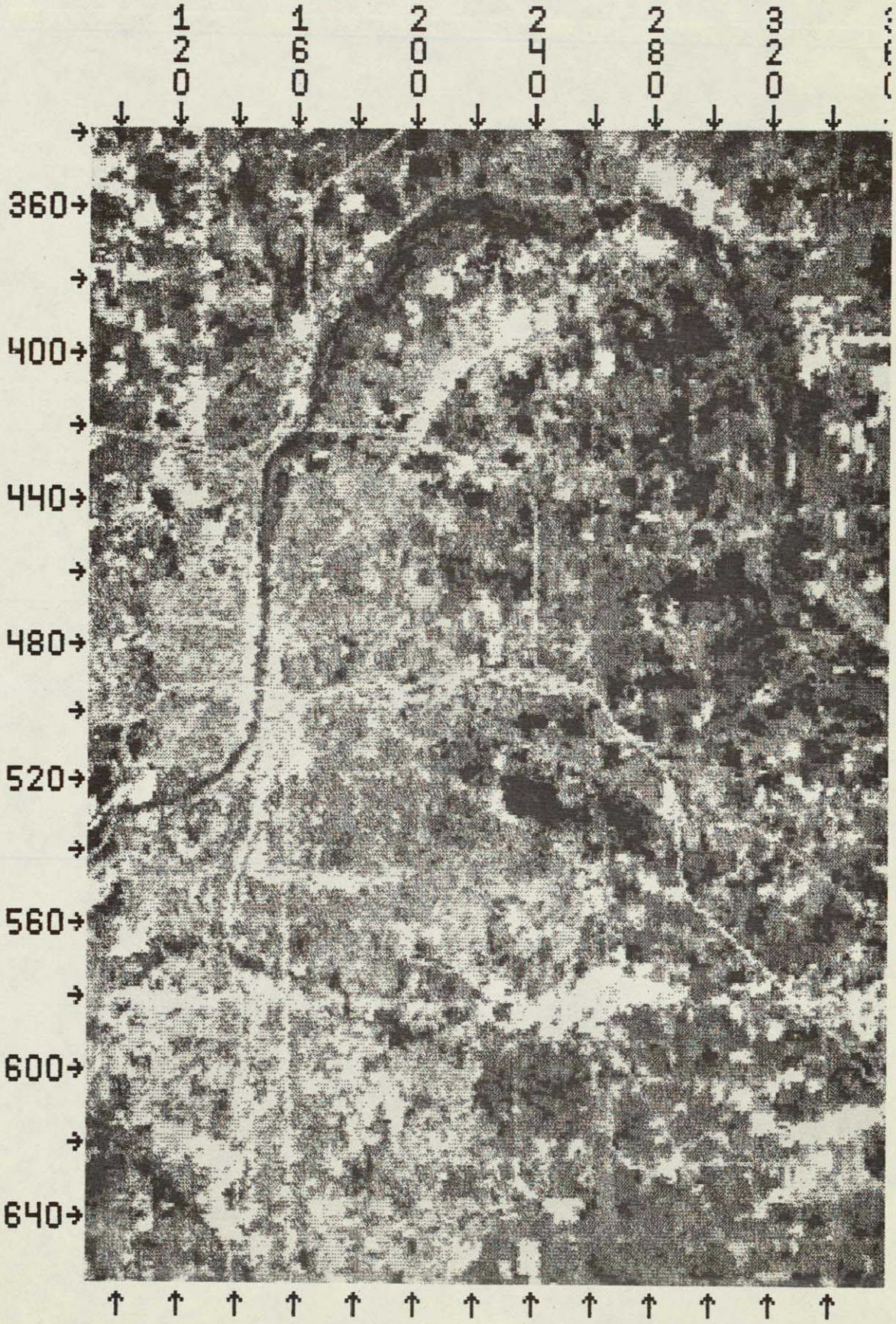


Figure 12. Grayscale map of band 5 of the Landsat data used in the Grand Rapids analysis.

SECTION III DEVELOP TRAINING STATISTICS

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

1. DESCRIBE the procedure for selecting training areas. Your description should include guidelines as to location, number and size.

2. DISCUSS clustering and how the analyst uses clustering to aid in the development of training statistics.

3. DESCRIBE why cluster classes are associated with information classes.

4. DEFINE, with the aid of simple drawings, the concept of statistical distance.

5. STATE two reasons why the statistical distances between clusters are calculated.

The next step in carrying out a supervised classification is the development of training statistics. The overall purpose of the training statistics is to "train" the computer to accurately classify the entire area of interest using pattern recognition techniques. Pattern recognition is a useful tool for identifying and classifying the ground scene. First, multispectral sanner data representing known ground cover types are located with the help of reference data. These patterns of known classification constitute the training patterns against which the patterns of unknown ground cover types are compared in the ECHO classification algorithm.

Developing the training statistics for the area under investigation is the most critical and time consuming step for the analyst. This step requires more interaction of man and machine than any other part of the analysis. The classification performance is highly dependent upon how well the training data represents the entire area and how well the ground cover types can be distinguished.

It is convenient to break the development of training statistics into four substeps as follows:

- Selection of training areas
- Cluster analysis of training areas
- Association of cluster classes with information classes
- Calculation of statistical distance between clusters

In the process of carrying out these substeps the analyst uses a variety of processing functions. As you better understand the capabilities and limitations of these processes and gain experience using them, you will undoubtedly formulate your own procedures for developing training statistics, specializing your procedures to the remote sensing applications of particular interest to you.

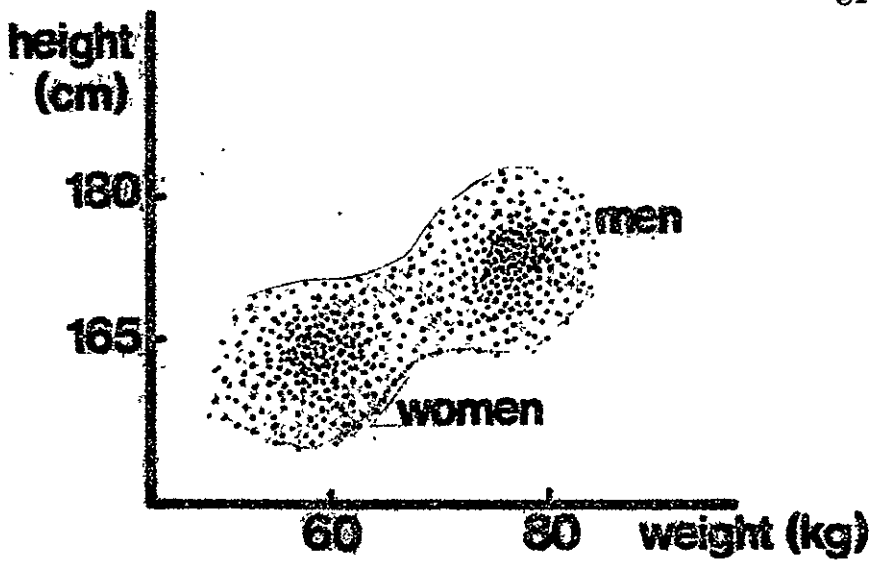
Selection of Training Areas

The analyst begins by determining the location, number and size of the training areas he is going to use. Several questions arise at this time. The first question is, "Where in the area to be analyzed should the training areas be selected?" One obvious yet important consideration is the location of areas for which some reference data is available. In addition, the training areas should represent all of the variability present in the total area. This generally means that the training areas should be distributed throughout the entire area being analyzed. In an effort to obtain representative training statistics, the analyst should select areas which include different ground surface features and sample the variations in the different environments.

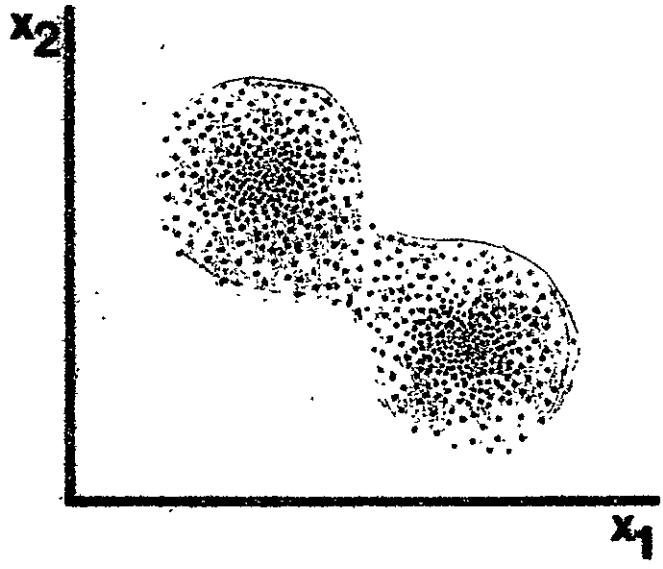
To select these training areas, an analyst begins by reviewing the analysis objectives. In stating the objectives, the cover types of interest are listed. These cover types are called "information classes." Training areas are selected in such a way that every information class is represented in at least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

A common procedure for selecting the training areas is to identify in the available reference data some well distributed areas that contain the information classes. These areas are then located on grayscale printouts of the multispectral scanner data. From these areas, training areas are selected. Experience has indicated that a reasonable starting point for Landsat data is to select training areas which are 25 to 100 lines and 25 to 100 columns in size, and each containing at least two cover types.

To help assure obtaining representative training data, the training areas should be distributed uniformly throughout the area to be classified, but this may not be possible if adequate reference data is not available. Usually, representative training data for all information classes can be obtained by selecting from four to eight training areas. In summary, each area



(a) A distribution of adult heights and weights



(b) Multimodal data from wheat fields
(hypothetical illustration)

Figure 13. Examples of data which tend to cluster. (from Swain and Davis, eds., Remote Sensing: The Quantitative Approach)

should be 25 to 100 lines by 25 to 100 columns, each area includes more than one cover type, and every cover type is included in at least one (preferably two or more) candidate training area.

·Cluster Analysis of Training Areas

The next step in the development of training statistics is to determine the spectral subclasses within each of the training areas. The analyst accomplishes this by using a technique known as "cluster analysis." Cluster analysis groups data vectors into clusters such that the vectors within any given cluster all have similar spectral reflectance characteristics. This technique takes advantage of the fact that there is often an inherent natural structure to multispectral scanner measurement vectors. Figure 13a illustrates this concept with a hypothetical sample of human height and weight occurrences. Note that the natural structure to this set of measurement vectors reveals that women are typically lighter and shorter than men. A set of reflectance measurement vectors for a wheat field, Figure 13b, illustrates the natural structure of reflectance values of a wheat canopy in channels x_1 and x_2. The clustering algorithm attempts to define classes by minimizing the distances between points within each class while maximizing the distances between the classes. Since the classes thus defined are inherently spectrally different, they are often referred to as "spectral classes."

Clustering is basically a computer operation to which the analyst must supply only information which tells the computer which data set to use, the lines and columns defining the area in the data to be clustered, and the number of clusters desired. The number of clusters requested by the analyst is influenced by the cover types of interest and their estimated spectral variability. The experience of the analyst plays an important role in his assessment of anticipated spectral variability.

The clustering algorithm requires that the analyst specify the number of clusters to be found. Experience has indicated that most cover types contain at least two clusters, and a rule-of-thumb is to request the number of clusters to be 1.5 to 2.0 times the number of expected information classes. If an analyst requests an insufficient number of clusters, the cluster variances will tend to be quite high and the resulting classes will be difficult to separate. If too many clusters are requested, cluster variances will be small but the resulting cluster classes may be difficult to correlate with the information classes. When choosing the number of clusters, the analyst should keep in mind the fact that the ECHO classifier tends to misclassify data points into those classes with high variances. Therefore, cluster classes with extremely high variances should generally be avoided. In choosing the number of clusters, the analyst must try to obtain a balance between having few enough clusters to be able to relate them to information classes and having enough clusters so that the variances of any one cluster are not disproportionately high in comparison to the other clusters. A "good"

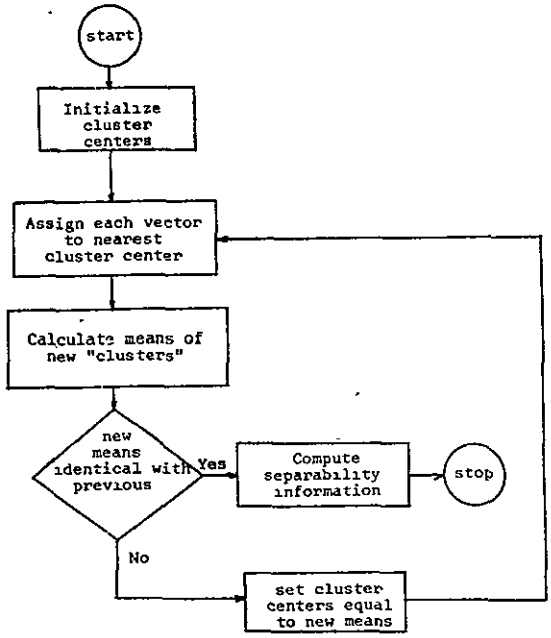


Figure 14. Flowchart of clustering algorithms.

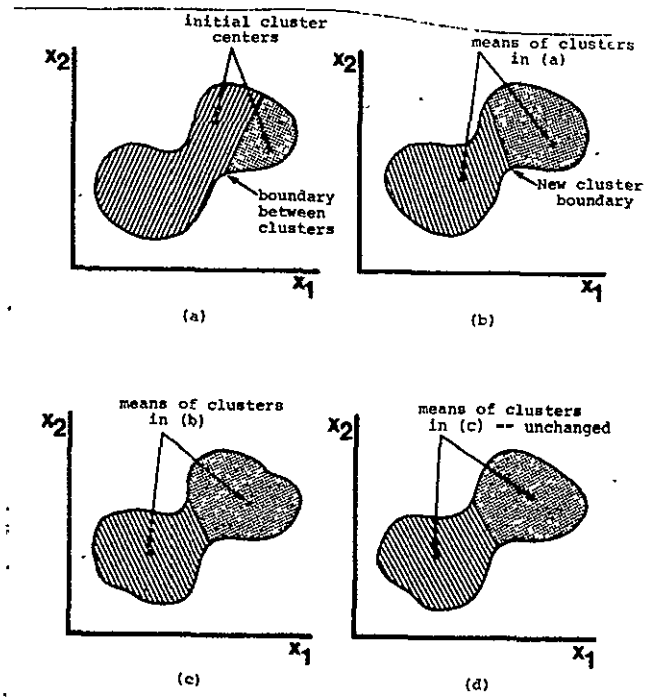


Figure 15. A sequence of clustering iterations (from Swain and Davis, Remote Sensing: The Quantitative Approach).

number of clusters such as is suggested by the "1½x" rule will tend to optimize these trade-offs. However, in some cases after examining the output it will be evident that a different number of clusters are needed, in which case the cluster processing should be repeated.

A flowchart of the LARSYS cluster algorithm is given in Figure 14 and its logical operation is described in the following text. The computer assigns locations in the feature space as the initial centers of each cluster (see Figure 15). It then calculates the distance between each of the data points and each cluster center and assigns each point to the nearest cluster. Next, new cluster centers are determined by calculating the mean vector for the data points assigned to each cluster. The computer then proceeds to re-calculate the distance between each data point and the new cluster centers and reassigns each sample to the closest newly defined cluster center. The computer continues the cycle of calculating the cluster centers and reassigning data points until all the data points have been assigned to a cluster and do not change allegiance with a further iteration of the cluster cycle. Once the cluster cycle is complete, the means and covariances for each cluster class can be punched directly onto cards. Cluster maps which pictorially represent each area that was clustered can be printed, and a summary table of the number of points assigned to each cluster produced. Information about mean reflectances and variances in each channel for each cluster class are also listed in tabular format. Also available are tabular listings of separability information for all possible class pairs and suggested cluster groupings for similar classes.

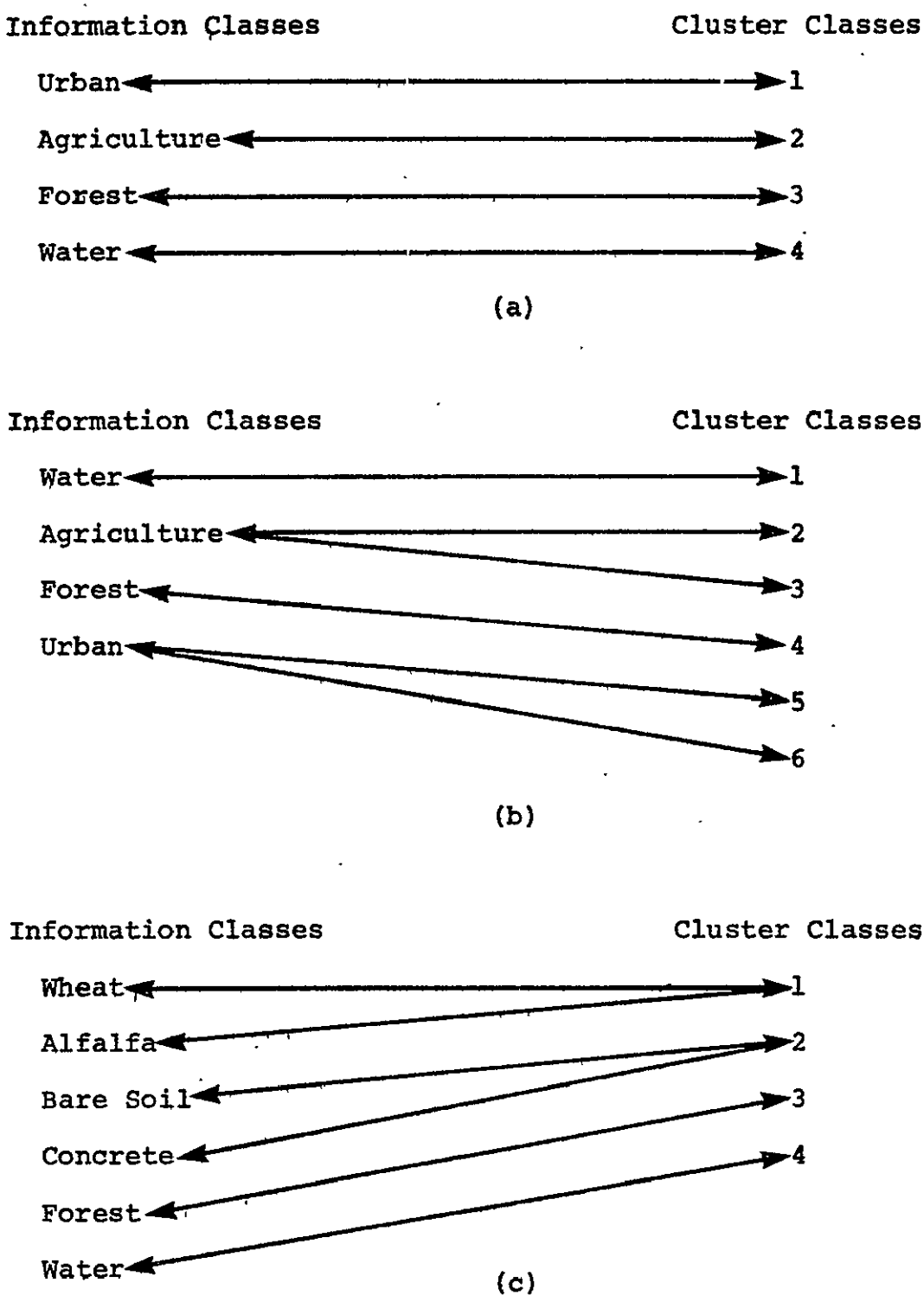


Figure 16. Examples of relationships between different information classes and their cluster classes.

Association of Cluster Classes With Information Classes

After obtaining the cluster analysis output the analyst examines the tables of cluster class means and variances. It is especially important to note any classes having unusually high variances, as it may be necessary to delete these classes to avoid biasing the ECHO classification results toward these classes.

The analyst then begins the process of associating each cluster class identified in the previous step with one of the information classes of interest (i.e., agriculture, urban, water, forest). The purpose of this step is to determine which spectral classes will be used to represent the information classes in the ECHO classification process. It should be pointed out that there will not necessarily be a one-to-one correspondence between the information classes and the cluster classes. Remember, an information class is a distinct cover type of interest, while a cluster class is a group of data points which are spectrally similar. As shown in Figure 16a, there may be a one-to-one correspondence between the two, but this is often not the case. It is possible that several cluster classes will represent the same cover type (information class) as shown in Figure 16b. Sometimes several information classes will be associated with the same cluster class, Figure 16c. In this case, the cover types are spectrally similar and cannot be differentiated using these data.

To carry out this step of the analysis, maximum use is made of all available reference data, so that the cluster classes can be reliably identified. If errors occur in this step of the analysis the training data supplied to the classifier will not be representative of the information classes. The association of cluster classes and information classes is sometimes difficult and time consuming, but this step is most important for insuring that the classifier is correctly trained.

Calculation of Statistical Distances Between Clusters

The analyst is now familiar with the variety of spectral responses occurring in his training areas. He should be so familiar with the data and study site at this point that he might decide to combine those cluster classes within each individual training area that seem to be spectrally similar and represent similar information classes. It is unnecessary to keep cluster classes apart as unique spectral classes when in fact cluster separability analysis suggests they should be combined. It is also desireable to assemble all the statistics from the separate training areas into a single statistics deck. This single deck will facilitate calculation of class separabilities between classes from different training areas in the next step of the analysis sequence.

At this point in the analysis sequence, the statistical characteristics of each of the cluster classes have been calculated. It would be possible to use all of these cluster classes to train the classifier, but this is usually not done. The number of clusters available at this point is normally greater than the number of classes needed to adequately train the classifier. This has occurred because some of the same classes were present in more than a single training area, which has led to the same cover type being represented separately in two or more sets of cluster statistics. An analyst would like to combine the training patterns in such cases because a smaller number of classes saves computer time and simplifies interpretation of results.

Also the analyst would like to have some indication of the probability of correct classification in advance of using all these classes in doing the actual classification. If there appears to be confusion between classes, an analyst may do more clustering on the areas already considered, asking for a different number of clusters, or the analyst may decide to select alternative training areas in an effort to get good distinction between classes.

A processing function that calculates the separability of cluster classes can help the analyst determine which cluster classes are similar, and it can serve as an indicator of probability of correct classification. To explain how this can be accomplished, the concept of *statistical distance* must first be discussed. Figures 17 and 18 show two cases of one-dimensional density functions. In each instance, you feel intuitively that the "distance" between the density functions is greater in case b than in case a.

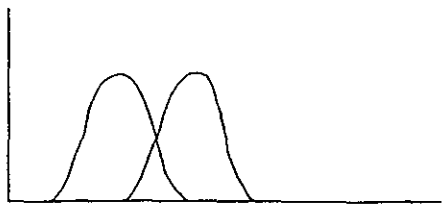


Figure 17a.

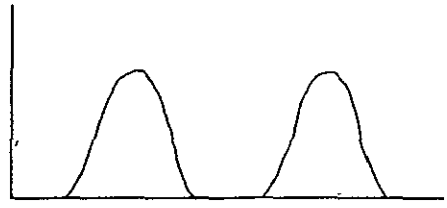


Figure 17b.

Figure 17. Two pairs of one-dimensional density functions. The statistical distance between the density functions in part b is greater than in part a.

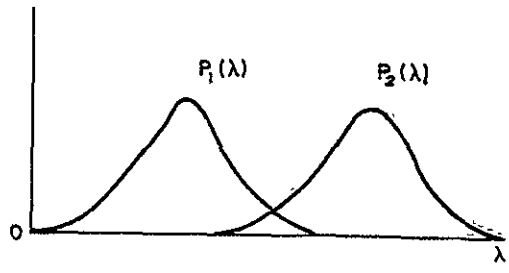


Figure 18a.

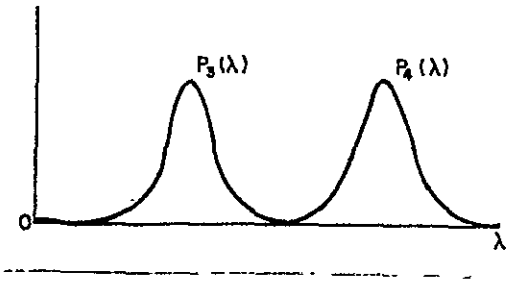


Figure 18b.

Figure 18. Each pair of distribution functions shown above has equidistant means, but the smaller variance in $P_3(\lambda)$ and $P_4(\lambda)$ cause them to have a larger statistical distance.

There are a number of ways of defining *statistical distance*.[1] One approach is to implement a distance measure called *transformed divergence*, and assume that the density functions are Gaussian. The distance between two Gaussian probability density functions depends not only on the ordinary (Euclidean) distance between the mean values but also on the "spread" of the data. Figure 18 illustrates this point. The Euclidean distances between the mean values are equal in both of the cases shown, but the smaller variances (smaller "spread") in (b) results in a larger statistical distance between the two density functions.

The class separability information discussed above can be utilized to combine those cluster classes from the different training areas which the reference data identifies as belonging to the same information class and whose statistical distance infers spectral similarity. It is important that both separability values and information class identity be considered in deciding whether to combine cluster classes as a single spectral/informational class so that each resultant class will have a unique spectral as well as informational identity. It might be necessary during this important analysis phase to delete some of the original cluster classes from the statistics deck if these classes appear to act as confusion classes between otherwise separable spectral/informational classes. This step is recommended only when the confusion class appears to offer no more significant spectral or scene information than is already represented by other classes defined in the analysis.

When trying to relate transformed divergence to probability of correct classification, we can assume that a greater statistical distance between density functions will result in greater classification accuracy. But the relationship between transformed divergence and classification accuracy is not a linear one. A graphic presentation of their relationship is shown in Figure 19; accuracy is given in terms of probability of correct classification. Note that very separable classes (probability of correct classification near 100%) have transformed divergence values of 2000. If all final spectral/informational classes have separability values greater than 1000 the analyst might expect his classification to have an accuracy of approximately 84%. This type of accuracy prediction could be very important if a level of required accuracy was specified in the analysis objectives. When examining the class separability information and beginning the process of cluster class merging the analyst should be aware of the fact that the ECHO algorithm performs best when the minimal number of spectral classes required to adequately represent the cover types in the study site is used. Therefore to reduce the number of spectral/

---

[1]Wacker, A., and D. Landgrebe. The Minimum Distance Approach to Classification.

informational classes it might be necessary for the analyst to calculate statistical distances between the classes he created in the previous step to identify which of those new classes could be combined. The resultant statistics deck can then be used in performing the ECHO classification described in the next analysis step.

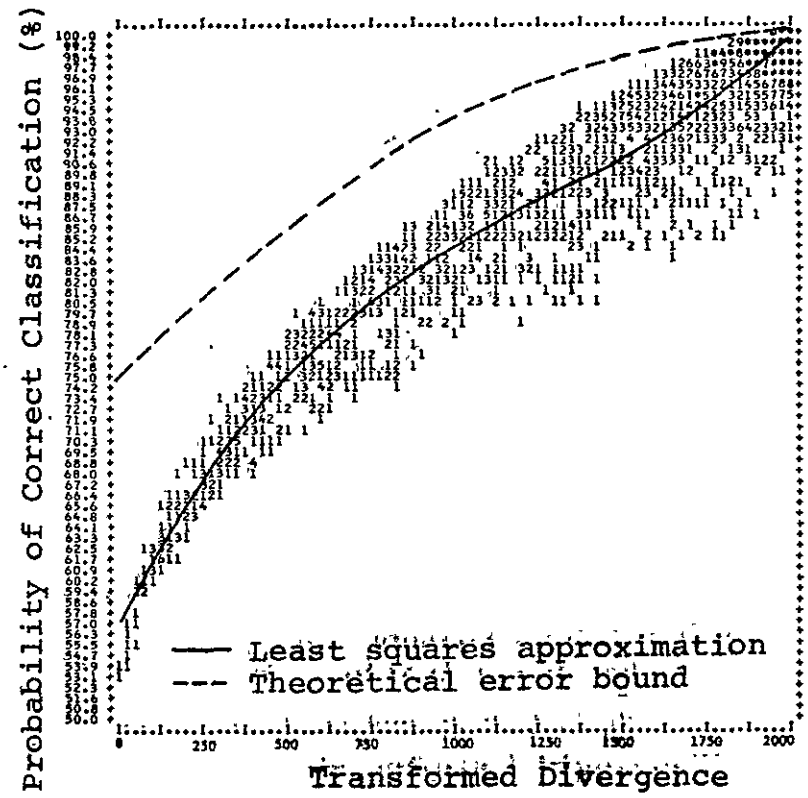


Figure 19. Observed values of probability of correct classification versus transformed divergence. From Swain and King, Two Effective Feature Selection Criteria for Multispectral Remote Sensing.

Self-Check

III-A. How should you select training areas? Where should they be located? How many should there be and how large should they be?

III-B. How does the analyst use clustering to aid in developing training statistics?

III-C. Why are cluster classes associated with information classes?

III-D. What is "statistical distance"? (Drawings may help to illustrate it.)

III-E. Why are the statistical distances between clusters calculated?

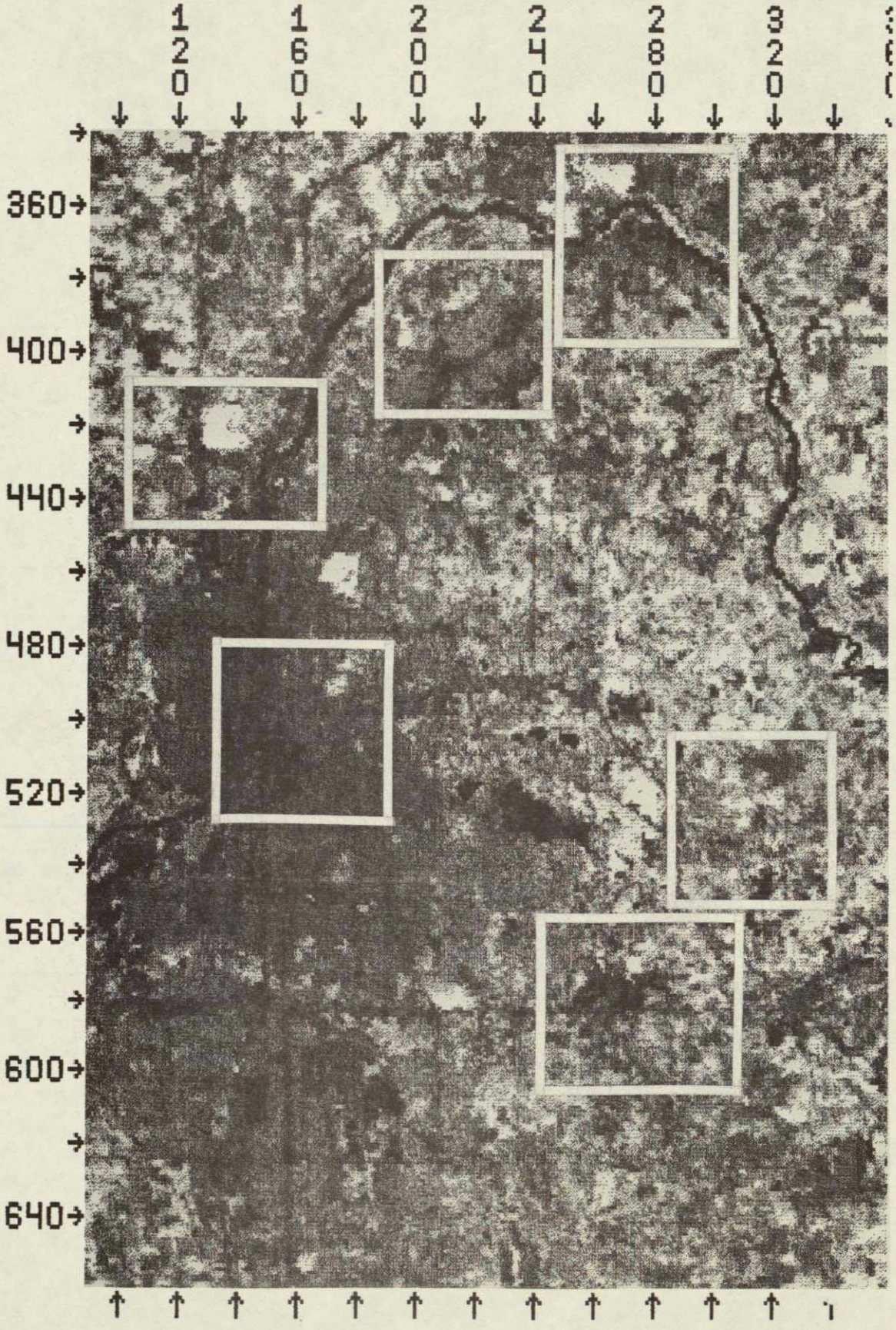


Figure 20. Grayscale map of band 4 Landsat data showing the six training areas selected for the Grand Rapids analysis.

The analyst examined the data and available reference materials for the Grand Rapids area and selected portions of the data to be used as training areas. He selected six areas in the scene which typified covertypes found in the commercial, residential, and rural sectors of the Grand Rapids study site. The locations of these training areas are shown in Figure 20.

The analyst then looked more closely at the first of his six training areas and a corresponding set of color infrared photography. He determined that this training area had four distinct covertypes (water, old residential, commercial, and grassy areas). Using the rule-of-thumb for determining the number of clusters, the analyst calculated that he should specify eight cluster classes (2 times the number of expected covertypes). The control cards he submitted in performing the cluster analysis on this first training site are given in Figure 21.

The analyst examined the cluster analysis computer output for his first training area which was located at the intersection of an interstate highway and the Grand River. He paid particular attention to the table of cluster class variances as shown in Figure 22. The analyst knew that it is important to avoid including a class with unusually high variances since the ECHO algorithm tends to favor such classes during classification.

The analyst then began to study the cluster map (see Figure 23) and associate each cluster class identified in the previous step with an information class (i.e., commercial, residential, water, grass). It should be remembered that there is not necessarily a one-to-one correspondence between the information classes and the cluster classes. Analyst notations are included in Figure 23 to show the associations that were made between the information classes and the spectral classes. As noted above, an information class is a distinct covertype of interest, while a cluster class is a group of data points which are spectrally similar. As shown in Figure 24, it is possible that several cluster classes will represent the same covertype (information class as also shown in Figure 24). Sometimes it is even possible for several information classes to be associated with the same cluster class. In that case the covertypes are spectrally similar and cannot be differentiated using these data.

The analyst had now numerically identified the natural spectral structure of the Grand Rapids area as represented in his training areas. He then examined the cluster class grouping table (Figure 25). He noted that the table indicated that some of the cluster classes within a single training area were spectrally similar enough to combine into a single class. If any one class had contained less than 40 points (ten times the number of channels used) that class would have been deleted by the analyst because there were too few points from which class statistics could be estimated. Preferably a class should have more than the minimum number of points for accurate statistical representation of a covertype.

```
-COMMENT CLUSTER OF I-96 & GRAND RIVER INTERCHANGE TRAINING AREA
-RUNTABLE
DATA
RUN(72054806),TAPE(454),FILE(2)
END
*CLUSTER
OPTIONS MAXCLAS(8),CONV(99.5)
PUNCH STATISTICS
CHANNELS 5,6,7,8
DATA
RUN(72054806),LINE(410,477,1),COL(104,168,1)
END
```

Figure 21. Control card listing used to cluster the first training area in the Grand Rapids study area.

CLUSTER	POINTS	MEANS CH(5)	CH(6)	CH(7)	CH(8)
1	261	47.81	49.38	58.62	28.17
2	381	30.05	20.63	65.70	40.34
3	840	36.13	30.55	54.60	29.68
4	741	28.61	19.82	55.54	33.06
5	880	31.98	25.37	46.11	24.93
6	659	38.53	34.31	48.18	24.20
7	420	40.62	38.16	39.77	17.54
8	238	24.23	17.70	16.49	6.13

CLUSTER VARIANCES

	CH(5)	CH(6)	CH(7)	CH(8)
1	29.75	59.37	42.13	15.82
2	6.81	15.02	14.27	11.13
3	6.25	10.79	8.62	4.25
4	6.50	12.13	9.78	6.29
5	7.88	11.33	15.86	7.22
6	5.97	11.41	7.86	3.90
7	13.58	26.23	18.36	5.56
8	9.08	10.87	41.53	18.70

← HIGH

Figure 22. Sample cluster analysis tables of class means and variances for the first training area in the Grand Rapids example.

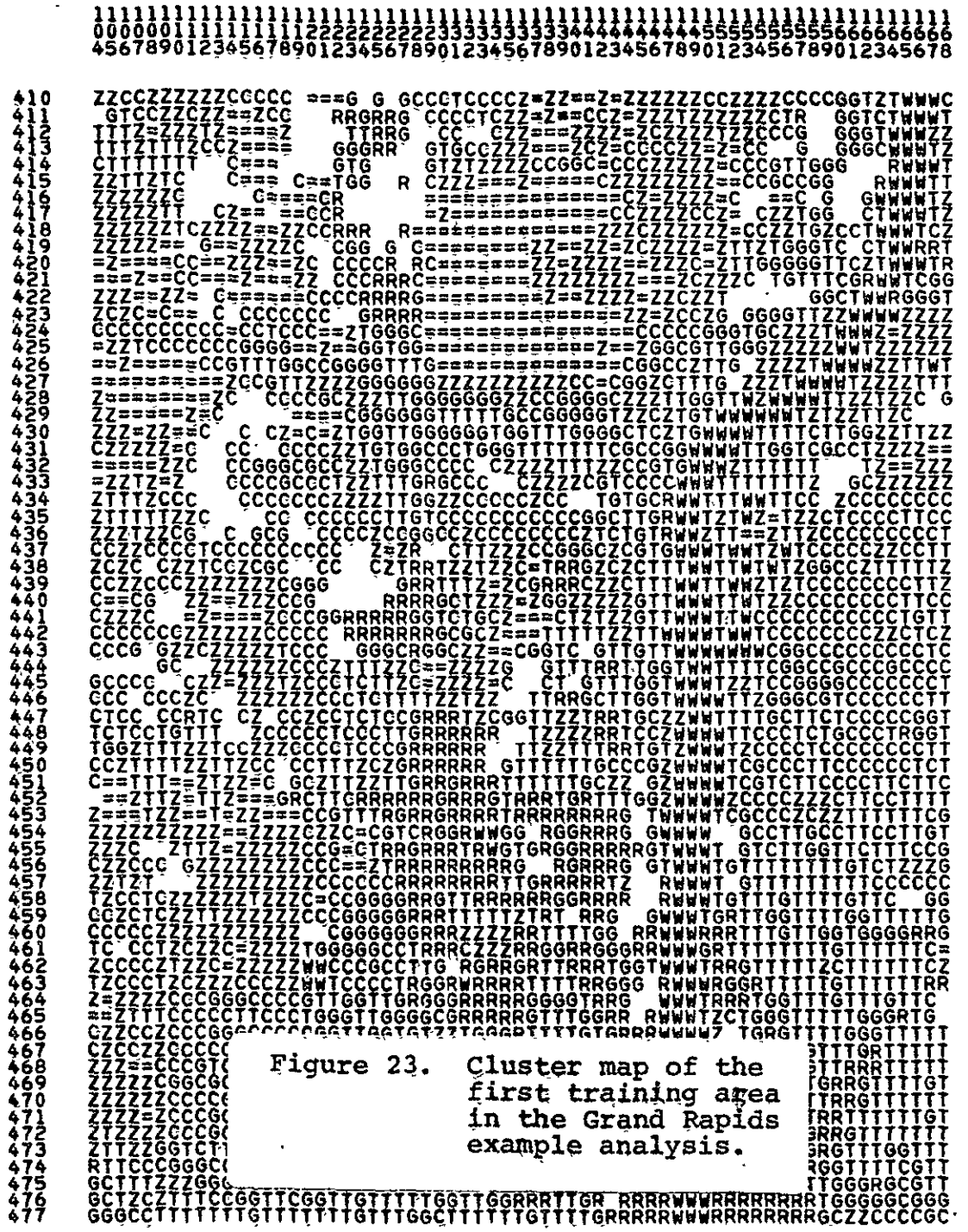


Figure 23. Cluster map of the first training area in the Grand Rapids example analysis.

NUMBER OF POINTS PER CLUSTER

CLUSTER	1 Commercial & Subdivisions	2 Golf & Orchard	3 Subdivision	4 Trees	5 Residential	6 Grass	7 Commercial	8 River
SYMBOL	=	C	Z	T	G	R	-	W
POINTS	261	381	840	741	880	659	420	238

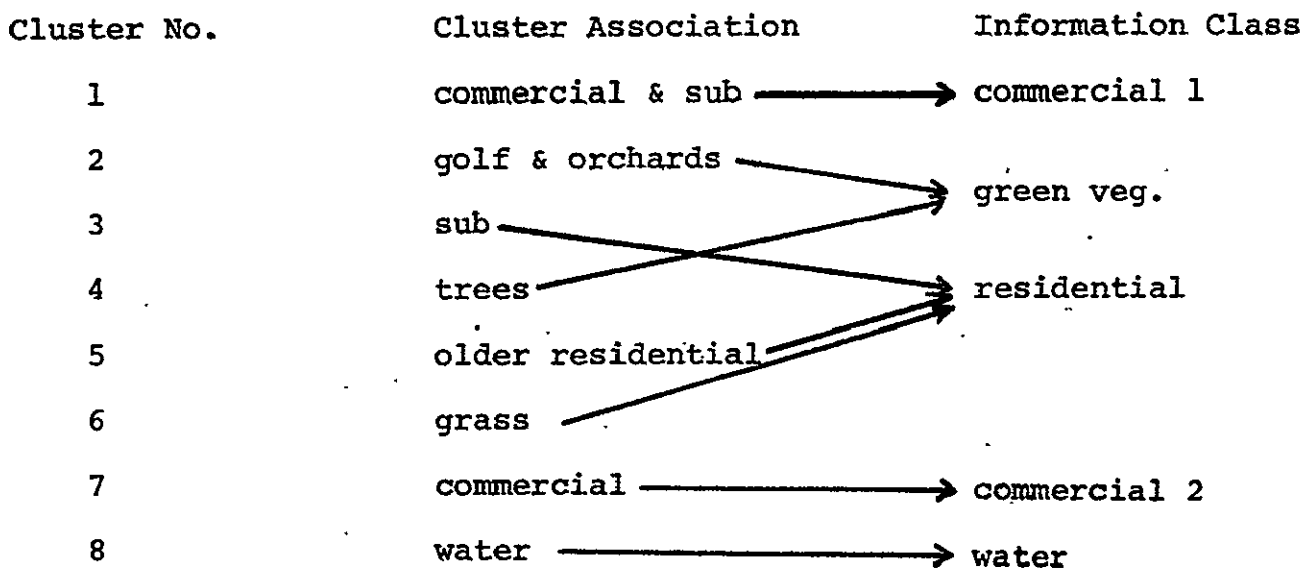


Figure 24. Relationships between different information classes and their cluster classes in the first training area of the Grand Rapids Analysis.

RESULTS OF CLUSTER GROUPING

THRESHOLD = 0.750

GROUP	CLUSTERS	NO. PTS.	
1	1	261	Commercial 1
2	2	381	Golf & Orchards } Grn Veg
	4	741	Trees }
3	3	840	Trees
	5	880	Residential } Residential
	6	659	Grass
4	7	420	Commercial 2
5	8	238	River

Figure 25. Cluster Grouping Table with the analysts annotations for the training area located at the intersection of I-96 and Grand River in the Grand Rapids analysis study.

Using the MERGESTATISTICS function, the analyst combined the statistics of the similar cluster classes that he had detected earlier in this analysis step and at the same time he merged the six statistics decks from the cluster analyses of the six training areas. The MERGESTATISTICS processor allows the user to combine the statistics from two or more separate statistics decks into a single deck without referring back to the original multispectral data. At the same time, it is also possible to pool the statistics of any two or more classes from those decks into the statistics for a single class. This involves computing an appropriately weighted average of the original means; the new covariance matrix depends on the original covariances and means and the number of points in each of the original classes. It is also possible to delete unwanted classes and to label all resultant classes with an informational class name. The control cards that the analyst used for this merging and labeling of his six statistics decks are given in Figure 26. Note that the analyst asked for a copy of the merged statistics deck to be punched onto cards so he could use it in the next analysis substep.

The analyst then ran the SEPARABILITY processor to determine which clusters from the various separate training areas could be combined. He used the control cards shown in Figure 27.

The analyst intended to use all four of the available Landsat channels (rather than selecting a subset of features) and therefore the parameter 4 appears on the required COMBINATIONS card. The DIV (1000) on the PRINT card would result in a summary listing of all class pairs whose pairwise distance was less than or equal to 1000 and therefore were candidates for further merging. See Figure 28.

This pairwise distance would also be used by the algorithm to compile a class grouping table at the end of the SEPARABILITY output (see Figure 29). This grouping table was examined to see if, based on the information class to which the analyst had assigned each spectral class, any "unusual" groupings had been recommended. This step is a good check to see if the analyst had mislabeled any spectral classes at the cluster stage of the analysis sequence.

The analyst next utilized the separability grouping table (Figure 29) to determine which spectral classes were similar enough to warrant combining them. He used the MERGESTATISTICS function to accomplish the desired class combinations and generated an output statistics deck containing 13 spectral classes from all six training areas with all classes spectrally distinct from one another. This statistics deck was punched onto cards and saved for use as input to the ECHO classifier in the next analysis step.

Referring again to Figure 6 (page 5), the analyst had at this point completed the development of his training statistics.

```
-COMMENT MERGING OF THE 48 CLASSES FROM THE 6 TRAINING AREAS
*MERGE
OPTIONS NOFIELDS,COSPEC
PUNCH
CLASSES ENTIRE(1,2,3,4,5,6)
POOL      COMERC1(1/1/),GRNVEG(1/2,4/),RES1(1/3,5,6/),BS1(1/7/)
POOL WATER1(1/8/),COMERC2(2/1,2/),GRASS(2/3/),RES2(2/4,5,6/)
POOL COMERC3(2/7/),WATER2(2/8/),GRAV1(3/1/),GRNVEG2(3/2,5/)
POOL RES3(3/3,4/),BRUSH1(3/6,7/),WATER4(3/8/),GRAV2(4/1/)
POOL COMERC4(4/2/),RES4(4/3,4/),GRNVEG3(4/5,6/),BRUSH2(4/7/)
POOL WATER7(4/8/),BS(5/1,2/),RES5(5/3/),AG1(5/7/),GRNVEG3(5/4,5,6/)
POOL MUCK(5/8/),COMERC5(6/1,3/),COMERC6(6/2/),RES6(6/4/)
POOL AG2(6/5,6,7/),WATER7(6/8/)
DATA
        *****STATISTICS FROM TRAINING AREA 1*****
DATA
        *****STATISTICS FROM TRAINING AREA 2*****
DATA
        *****STATISTICS FROM TRAINING AREA 3*****
DATA
        *****STATISTICS FROM TRAINING AREA 4*****
DATA
        *****STATISTICS FROM TRAINING AREA 5*****
DATA
        *****STATISTICS FROM TRAINING AREA 6*****
END
```

Figure 26. Control card listing for merging statistics from the six Grand Rapids training areas.

```
-COMMENT SEPARABILITY OF GRAND RAPIDS MERGED TRAINING CLASSES
*SEPARABILITY
COMBINATIONS 4
CARDS READSTATS
PRINT DIV(1000)
DATA
        *****MERGED STATISTICS DECK FROM THE 6 TRAINING AREAS*****
END
```

Figure 27. Control card listing used to calculate the separability of the combined training (cluster) classes in the Grand Rapids example analysis.

CLASSES THAT MAY BE COMBINED-MAX DIV. = 1000

AK	604.
AP	341.
AV	621.
A$	392.
BC	964.
BM	788.
CG	674.
CH	834.
CM	723.
CR	892.
CS	795.
CY	987.
C/	529.
DN	763.
DT	714.
DX	885.
EU	513.
EZ	612.
FQ	265.
F+	305.
GL	562.
GS	185.
GY	222.
G/	73.
HR	990.
HT	877.
HX	982.
H=	897.
KP	373.
K$	371.
LS	277.
LY	166.
L/	538.
MR	279.
MV	300.
M=	520.
NS	934.
NT	168.
NX	112.
N/	939.
P$	501.
Q+	132.
RV	560.
R=	135.
SY	52.
S/	107.
TX	80.
UZ	283.
V=	810.
Y/	189.

END OF LIST

Figure 28. Separability pairwise listing showing class pairs whose separability distance is 1000 or less on a scale of 0 to 2000.

RESULTS OF SEPARABILITY GROUPING

THRESHOLD = 1000.

GROUP	CLASSES	SYMBOL	NO. PTS.	
1	2 Comerc 1	B	787	Comerc 1
2	3 Res 1	C	1300	Res 1
	8 Res 2	H	1533	
3	4 BS 1	D	103	Ag
	14 Brush 1	N	820	
	20 Brush 2	T	207	
	24 Ag 1	X	272	
4	5 Water 1	E	123	Water 1
	21 Water 7	U	92	
	26 Muck	Z	45	
5	6 Comerc 2	F	539	Comerc 2
	17 Comerc 4	Q	196	
	28 Comerc 6	+	264	
6	9 Comerc 3	I	408	Comerc 3
7	10 Water 2	J	124	Water 2
8	1 Comerc 1	A	157	Comerc 4
	11 Grav 1	K	113	
	16 Grav 2	P	124	
	27 Comerc 5	$	297	
9	7 Grass	G	122	Grn Veg 1
	12 Grn Veg 2	L	1169	
	19 Grn Veg 3	S	784	
	25 Grn Veg 3	Y	1312	
	30 Ag 2	/	2143	
10	13 Res 3	M	824	Res 2
	18 Res 4	R	991	
	22 BS	V	313	
	29 Res 6	=	604	
11	15 Water 4	O	148	Water 3
12	23 Res 5	W	496	Res 3
13	31 Water 7	A	42	Water 4

Figure 29. Separability grouping table for the 31 class merged training area statistics with the analyst's annotations.

SECTION IV CLASSIFY AND DISPLAY RESULTS

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

1. DESCRIBE the ECHO classification process using the flowchart in Figure 30.

2. LIST three formats in which the classification results may be displayed.

Having developed the training statistics for the classes (and subclasses) of interest, we are ready to classify the area and display the classification results. Most of the "work" in this step of the analysis is done by the computer. The supervised ECHO classification algorithm aggregates data points into "objects" (groups of points which are spectrally similar) and classifies the objects into classes defined by the training statistics. Classification begins the last major analysis phase in deriving useful information from remote sensing data. Once the classification has been completed and the results stored in a computer file, the analyst can display the results in a variety of map-like and tabular forms. Also, if a set of test fields are available (areas in the scene having known covertypes) the accuracy of the classification can be assessed.

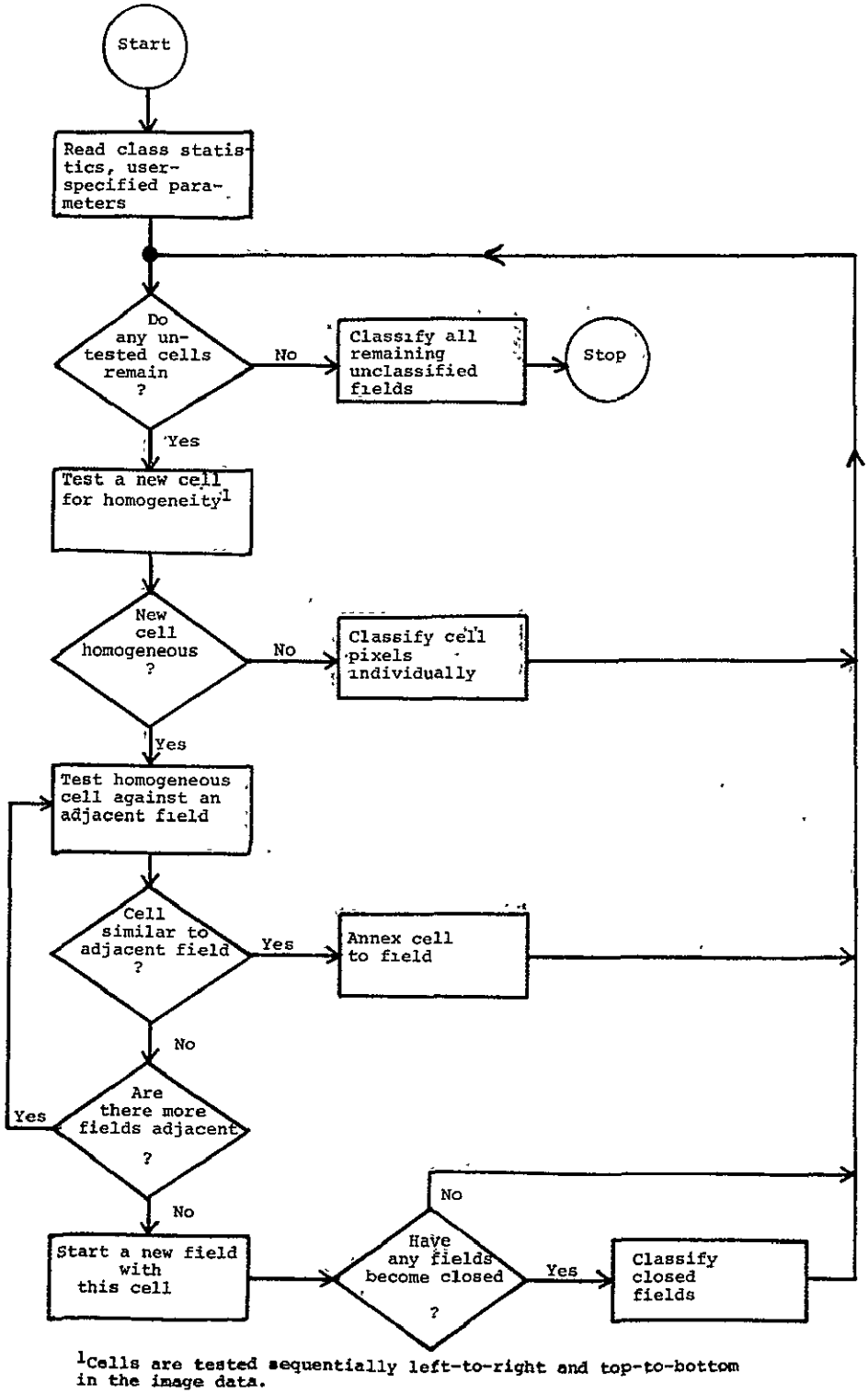


Figure 30. Flow Diagram of the ECHO Classifier.

·Classification of the Study Area

A discussion of the basic ideas behind the ECHO classifier will aid in the understanding of how the classifier works and provide guidance to the analyst for specifying the parameters required by the ECHO classifier. The phrase from which the acronym ECHO is derived, Extraction and Classification of Homogeneous Objects, describes the operation of the algorithm. The ECHO algorithm is a two-phase process.[1,2] The first phase involves "object finding," the detection of spectrally homogeneous regions in the data. The second phase accomplishes the classification of the objects based on a maximum likelihood decision rule. (See Figure 30)

The object finding is itself a two-step operation. Each square block or "cell" of data points to be classified is tested for similarity of its data points based on a homogeneity threshold specified by the user. When a cell fails to pass this test, its individual data points are each classified independently. Adjacent homogeneous cells are then tested against each other based on another homogeneity threshold and combined, when sufficiently similar, into "fields" or objects.

There are three parameters the analyst must specify:

1.) The cell size or cell width (number of pixels on each side of a square area) to be considered when partitioning the data (the parameter name is CELW),

2.) a threshold value for accepting or rejecting the cell as a unit based on the similarity of the data points within the cell (the parameter name is THRESHOLD),

3.) a threshold value for the algorithm to use when making a decision on whether to annex adjacent cells into a larger unit ("field") based on the similarity of the adjacent cells (the parameter name is ANNEX).

During the first phase a statistical test for homogeneity is performed on a group of pixels (a square array called a cell) to determine whether or not the pixels comprise part of an object. That is, the test determines whether the measurement vectors are similar in a statistical sense. The analyst must specify two parameters used in this test: the cell width and a threshold value for accepting or rejecting the hypothesis that the cell is homogeneous.

---

[1]Landgrebe, D.A., Final Report NASA Landsat NAS9-14970. Volume I, May 1977.

[2]The ECHO Users Guide is strongly recommended as a reference for this portion of the case study.

$(CELW)^2 \times NCHAN$	almost always break up cell					seldom break up cell
8	7	16	26	39	52	78
12	11	22	32	49	65	98
16	15	27	39	58	78	117
18	17	29	42	63	84	126
20	19	32	45	67	90	135
24	23	37	51	76	102	153
27	26	40	55	83	110	166
28	27	42	56	85	113	170
32	31	46	62	93	124	187
36	35	51	67	101	135	203
45	44	62	80	120	160	240
48	47	65	84	126	168	252
50	49	68	86	130	173	259
54	53	72	91	137	183	275
63	62	82	103	155	206	310
64	63	84	104	157	209	314
72	71	93	114	172	229	344
75	74	96	118	177	237	355
80	79	102	124	187	249	374

Figure 31. Table of threshold parameters (THRESHOLD values) to input into ECHO phase 1 processing.

The cell width parameter is chosen on the basis of the analysis objectives, spatial resolution of the scanner system and expected size of objects within the scene. One tries to balance the constraints of choosing the cell width so that the cell size is approximately equal to the smallest "mapping unit" of interest but small compared to the average object size (see Figure 2, page 1). It should be possible to determine the minimum size area of interest from the objective(s). If agricultural fields of tens of hectares are being classified, larger cell widths (CELW) would be used than if urban areas were being classified. In Landsat 2 data each data point (pixel) represents .45 hectares or 1.1 acres. Therefore, a cell size of three (CELW = 3, a three by three square or a total of 9 data points) would represent about 40.5 hectares (10 acres).

The threshold parameter establishes a criterion for deciding if a cell is homogeneous or not. The larger the threshold value the more likely the decision that the cell is homogeneous. Assuming the cell width has been chosen so that the cells are small compared to object sizes, it is usually desirable to choose the threshold so as to achieve a relatively low rejection rate. To determine the appropriate homogeneity threshold value (THRESHOLD) the analyst can use the information in Figure 31. To use this table he must know the number of channels to be used in classifying the data, the cell size (CELW) to be used, and the relative degree of homogeneity he wants a cell to have in order to be classified as a whole cell. It is necessary for the analyst to calculate an entry point to the table by using a formula involving cell width (CELW) and the number of channels to be used in classifying (NCHAN). The formula is $(CELW)^2 \times NCHAN$. The analyst then selects a homogeneity value along the appropriate row depending upon the degree to which he wants the ECHO processor to classify cells as a whole or to break them apart for per point classification. If he selects a value near the left side of the table, the cells will often be broken apart and a per point classification used. On the other hand, a value selected near the right side of the table will result in the cells usually being classified as a whole.

In the second object-finding phase, cells which pass the homogeneity test are compared to an adjacent "field", which is simply a group of one or more connected cells that have previously been annexed. If the two samples appear statistically similar then they too are annexed. Otherwise the cell is compared to another adjacent field or becomes a new field itself. By successively annexing adjacent cells, each field expands until it reaches its natural boundaries. The field is then classified using a maximum likelihood sample classifier[3] and the resulting classification is assigned to all pixels in the cell. Cells that fail the homogeneity test

---

[3]Swain, P.H. Pattern Recognition: A Basis for Remote Sensing Data Analysis. LARS Information Note 111572, 1972.

in phase one are termed "singular" cells and often overlap a boundary between fields. The cells are split and the individual data points are classified independently using a per point maximum likelihood classification rule.

In the annexation phase it is necessary for the analyst to specify the third processing parameter, the annexation threshold value (ANNEX). This parameter commonly has values ranging from 0 to 4.6. A value of zero will result in no cells being combined into fields. A value of 4.6 could lead to the entire area being classified as only a few very large fields.

In summary, supervised ECHO classification is carried out in several phases. Tha analyst specifies a cell width parameter to determine cell size and a threshold parameter which essentially determines how similar the cell pixels have to be in order to judge the cell as being homogeneous. Next, homogeneous cells are annexed to adjacent fields if they pass an annexation (similarity) test. An annexation parameter specified by the analyst determines how spectrally "similar" the cell and the field have to be before annexation takes place. Finally, each field or object resulting from annexation is classified as a unit into one of the classes defined by the training statistics.

·Displaying Results

After classification, the analyst displays the results for purposes of evaluation or to produce an end product for the user. Depending upon the proposed method of testing or evaluating the ECHO classification's "correctness", it might be desirable to present the classification results in any one or more of several possible formats. This also holds true for the final classification product which the user desires for his application needs.

Depending on the output devices available on the system, several types of map-like products can be produced. Examples are alphanumeric computer line printer maps, gray scale printer-plotter maps and black and white or color photographs generated from a film writer or CRT display screen. Specific outputs are shown in the example portion of this section (see pages 58, 59, and 60).

In addition to map-like products, tabular quantitative results may also be obtained and may constitute a desired end product. An example of the type of table which might be produced is given in Figure 32. This particular type of end product might be of interest to users concerned with area estimates of the covertypes in the study area.

GROUP	POINTS	ACRES	HECTARES	PERCENT
COMERCL	9528	10957.2	4436.1	11.0
OLD RES	14727	16936.0	6856.7	17.0
NEW RES	21537	24767.5	10027.3	24.9
NONURBAN	38077	43788.5	17728.2	44.1
WATER	2531	2910.6	1178.4	2.9
TOTAL	86400	99359.9	40226.7	100.0

EACH DATA POINT REPRESENTS 1.15 ACRES
0.47 HECTARES

TOTAL POINTS IN CLASSIFICATION = 86400

Figure 32. Tabular results product generated from the ECHO classification of Grand Rapids, Michigan. Classes have been grouped into the five covertypes of interest with areas tabulated in acres, hectares and total number of data points for each.

a.)

```
-COMMENT ECHO PHASE ONE OF GRAND RAPIDS ANALYSIS
-RUNTABLE
DATA
RUN(72054806),TAPE(454),FILE(2)
END
*SECHO
RESULTS TAPE(6),INITIALIZE
OPTIONS CELW(2)
THRESHOLD(58.8)
CHANNELS 5,6,7,8
DATA
RUN(72054806),LINE(340,659,1),COL(90,359,1)
END
```

b.)

```
-COMMENT ECHO PHASE TWO OF GRAND RAPIDS ANALYSIS
*SECHO
ANNEX(1.0),CELW(2)
INPUT TAPE(6),FILE(1)
DATA
RUN(72054806),LINE(340,659,1),COL(90,359,1)
END
```

c.)

```
-COMMENT COPYING ECHO FINAL RESULTS FROM COMPUTER STORAGE TO A TAPE
*COPYRESULTS
FROM DISK
TO TAPE(444),FILE(6)
END
```

Figure 33. Control card listing for ECHO processor as set up for Grand Rapids example classification.

Deriving training statistics enabled the analyst to spectrally characterize all the covertypes of interest in the Grand Rapids study area. He was thus ready to utilize the two-phase ECHO algorithm by selecting the appropriate ECHO processing options to perform the classification. The analyst had to determine:

1.) cell size (CELW)

2.) threshold value for cell homogeneity test (THRESHOLD)

3.) threshold value for annexation test (ANNEX)

The analyst reviewed the analysis objective from the zoning board to determine the appropriate cell size. The size area on the ground that the zoning board was interested in was about 2 hectares. Details smaller than 2 hectares were not necessary for the zoning board's purposes. The analyst knew that a single Landsat data point represented about .45 hectares (1.1 acres) and therefore a cell size of 2 data points by 2 data points (CELW = 2) would give a gound area of 1.8 hectares (4.4 acres).

To determine the homogeneity threshold value (THRESHOLD) he used the table in Figure 31 (page 50). Since he selected a cell width of 2 and wished to use 4 channels of Landsat data, the table entry point is 2^2 x 4 or 16. He wanted to select a homogeneity threshold which would tend to cause the ECHO processor to classify whole cells at a time, breaking them apart for per point classification only when the cells were distinctly heterogeneous. Therefore he chose a value on the row for $(CELW)^2$ x NCHAN = 16 which was closer to the right side of the table, deciding to use 58.88. Using these parameters (CELW = 2 and THRESHOLD = 58.88), the analyst submitted the ECHO phase one control cards given in Figure 33a. Note that the analyst specified that the results from phase one should be placed on Results Tape Six; any tape available to the analyst could have been used for this purpose.

For phase two of the ECHO classification process, the analyst specified an annexation threshold value of 1.0 since he only wanted cells to be annexed if there was a reasonable degree of similarity between them. (An ANNEX value of zero will result in no cells being annexed and a value of 4.6 will lead to most cells being combined.) The control cards assembled to run phase two of the ECHO processor are shown in Figure 33b. Note that the analyst specified where the intermediate results from phase one were located so that phase two could use them to complete the final classification.

The classification which was produced by the ECHO classifier in this analysis step was stored as a permanent record on a computer storage tape (Figure 33c). To display the classification the analyst first requested an alphanumeric printout of the results from a line printer using the PRINTRESULTS control cards shown

in Figure 34. This type of output product is particularly good for large scale grayscale type maps and for use as base maps in many applications. He assigned a letter or symbol to each class so that when a class occurred on the results map it would be designated by a character having both informational and grayscale values. In this case, for example, the analyst assigned roads a bright symbol (blank), and water a very dark symbol (see Figure 35).

An even more graphic presentation of classifications might be obtained by displaying results in a black-and-white graytoned map. One example of this type of map product is shown in Figure 36. A third method of displaying classification results involves a photographic image. This could be a black-and-white (Figure 37) or color image. This type of display is most commonly generated as a final product for use by the individual or organization requesting the classification results.

```
-COMMENT GRAND RAPIDS, MI ECHO RESULTS CELW=2 THRESHOLD=58.8 ANNEX=1.0
*PRINTRESULTS
RESULTS TAPE(444),FILE(6)
SYMBOLS &,*,+,I,-,=,.,V,M,M,M,M
END
```

Figure 34. Control card listing used to print an alphanumeric computer map of the ECHO classification results in the Grand Rapids analysis.

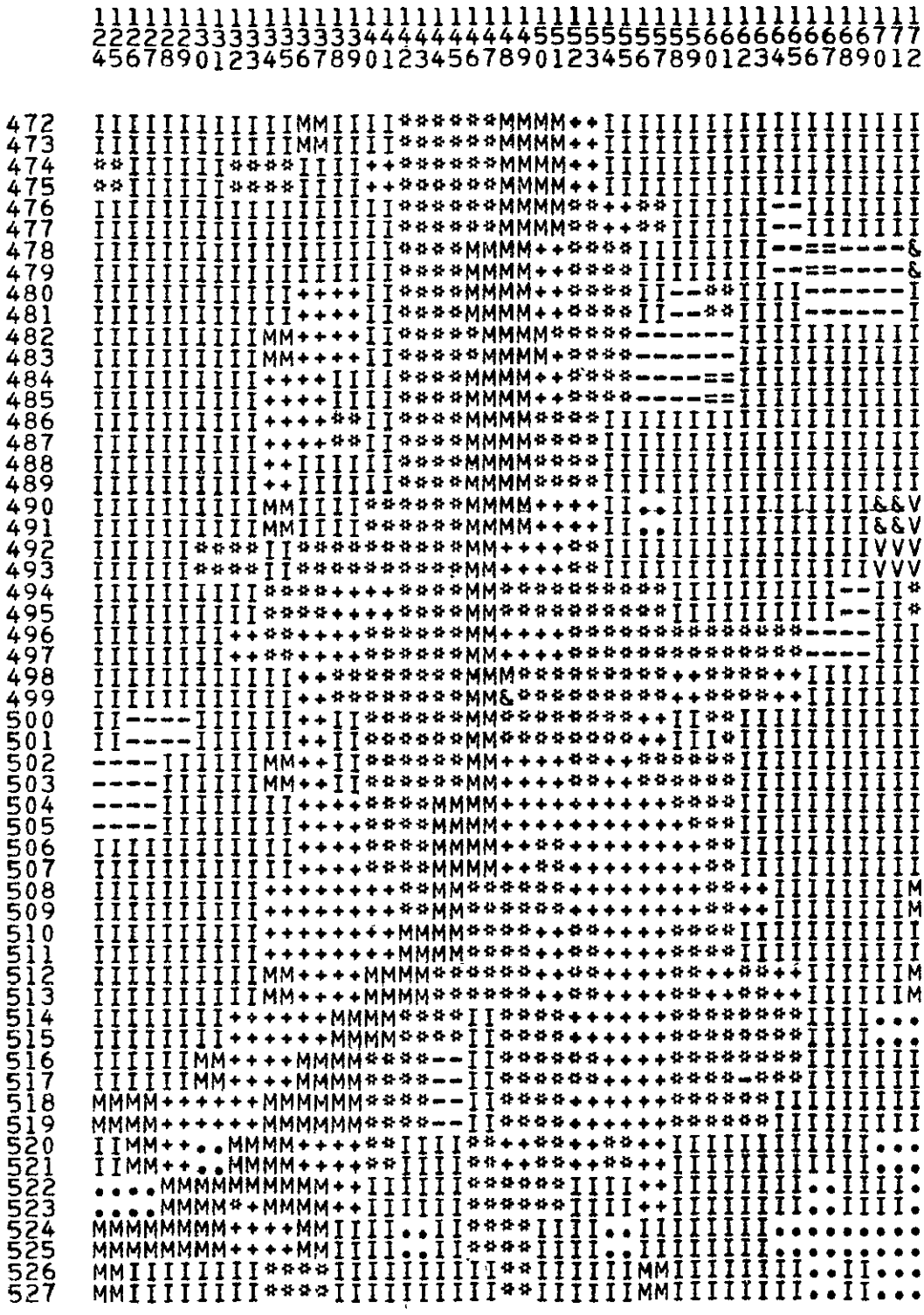


Figure 35. Alphanumeric line printer results map (scale 1:24,000) of a portion of an ECHO classification of downtown Grand Rapids, MI.

Key: ·Agriculture
V Green Vegetation
M Water
& * + Commercial
I - = Residential

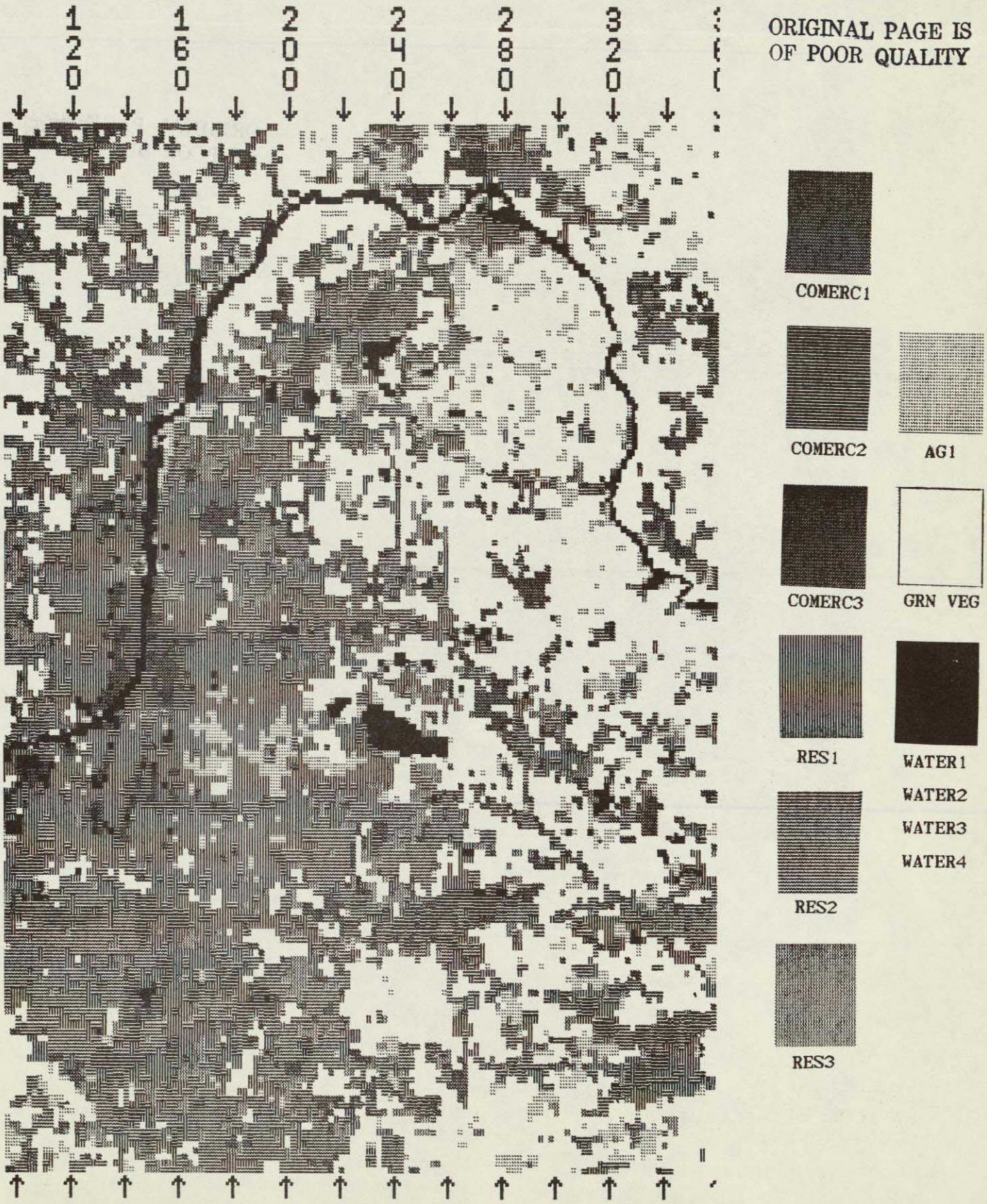


Figure 36. EHCO classification results map of the Grand Rapids study area using ECHO parameters CELW = 2, THRESHOLD = 58.8, ANNEXATION = 1.0. Map was produced on a Varian plotter.

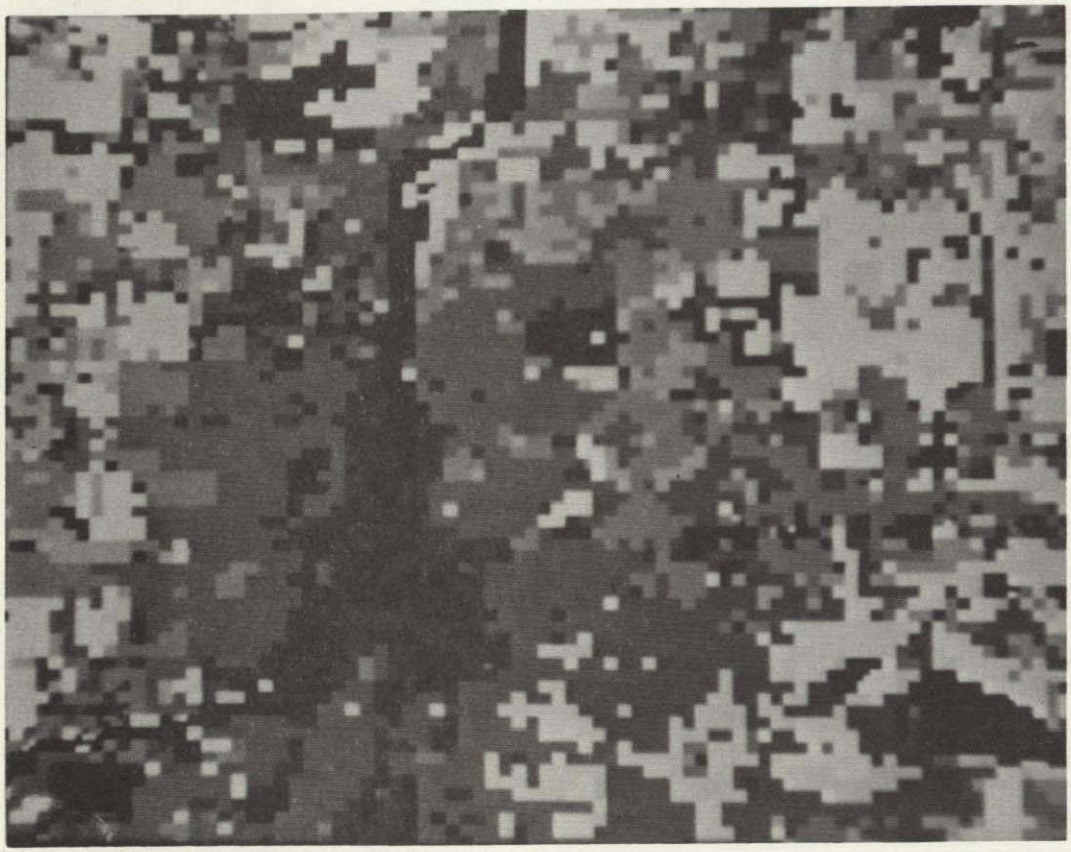


Figure 37. Black and white photographic product generated from an ECHO classification results of Grand Rapids, Michigan. The area shown is the left central portion of the study area.

Self-Check

IV-A. Referring to the flowchart on page 48 (Figure 30), briefly describe the classification process which is used by ECHO.

IV-B. What are three different formats in which the ECHO classification results may be displayed?

SECTION V EVALUATION OF RESULTS

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

1. IDENTIFY three criteria which should be met in selecting test fields.

2. DESCRIBE a procedure for evaluating the classification performance of the ECHO processor.

Quantitative evaluation of classification performance is an important step in the analysis sequence. The decision to be made is whether or not the classification of the covertypes for the entire area is adequate. The first step in obtaining an accurate assessment of the results is to select a set of representative test fields for all covertypes and their variations in the study area. Optimally these test fields should be selected using random or unbiased selection techniques. The test fields should also be representative of the covertypes in the analysis area and of sufficient size for performing a statistical evaluation.

To meet these criteria the following guidelines may be used:

1. Test fields should be selected randomly such that the number of data samples for each class are proportional to that covertype's occurrence in the analysis site.

2. The full range of variation within each covertype should be included in the test fields for that class.

3. A sample size of 100 to 500 data samples for each covertype being tested should be selected.

These guidelines are actually general rules of thumb providing a reasonable starting point for test field selection. The details of an optimal test field selection procedure depend considerably on the use to which the analysis results will be put. Some considerations along these lines may be found in the reference.[1]

These test fields are then classified and the results evaluated. In order to assess and evaluate classification results, the computer can provide an alphanumeric printout and tables containing

---

[1]Hoffer, Roger, "Computer-Aided Analysis of SKYLAB Multispectral Scanner Data in Mountainous Terrain for Land Use, Forestry, Water Resources and Geologic Applications "

quantitative information about a classification. The analyst must specify the coordinates of the test fields. The computer then examines and tabulates the classification decision for each data point within these test fields and prints out a summary by fields, classes, or both, as specified by the analyst. An example of the tabular results for classes is shown in Figure 38. Such a table is called a test class performance matrix.

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

Figure 38. Test class performance matrix.

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

The diagonal elements of the matrix can be summed, and that total divided by the total number of samples. The result is called overall performance. For Figure 38, the overall performance is $(65 + 87 + 69 + 53 + 28)/316 \times 100\% = 95.1\%$.

If evaluation indicates the results to be below the standard stated in the analysis objectives, it may be necessary to determine those classes with low accuracy in the performance tables and to select new training areas for those classes and develop new training statistics. It should be pointed out that there may be situations in which the analysis objectives simply cannot be met. This could happen if the requirements are too stringent (i.e., a very high accuracy required), if there is insufficient reference data to do adequate training or if the data quality is poor (a high proportion of cloud cover or a lot of bad data lines, for example).

Self-Check

V-A. What three criteria should be met in selecting test fields?

V-B. How can you evaluate the classification performance of the ECHO processor?

A set of representative test fields was selected from the data for use in evaluating the ECHO classification results. A set of candidate fields was selected at random and their covertypes identified utilizing available aerial photography and topographic maps. A candidate field was accepted as a test field if it was found to display a homogeneous cover of one of the classes of interest. The locations of these test fields are shown in Figure 39.

The analyst submitted these fields to the PRINTRESULTS processor with control cards given in Figure 40. Note that the analyst included a control card which requested tables to be generated for the test fields he had included in the deck (Test (F)). He also requested that those fields be outlined on the alphanumeric map (Outline Test (F)).

The analysis objective had stated that an overall accuracy of 80% was adequate for classification performance. When the analyst examined the performance table output of the PRINTRESULTS processor he noted his overall accuracy was 85% and therefore more than adequate for the Grand Rapids Zoning Board requirements (See Figure 41).

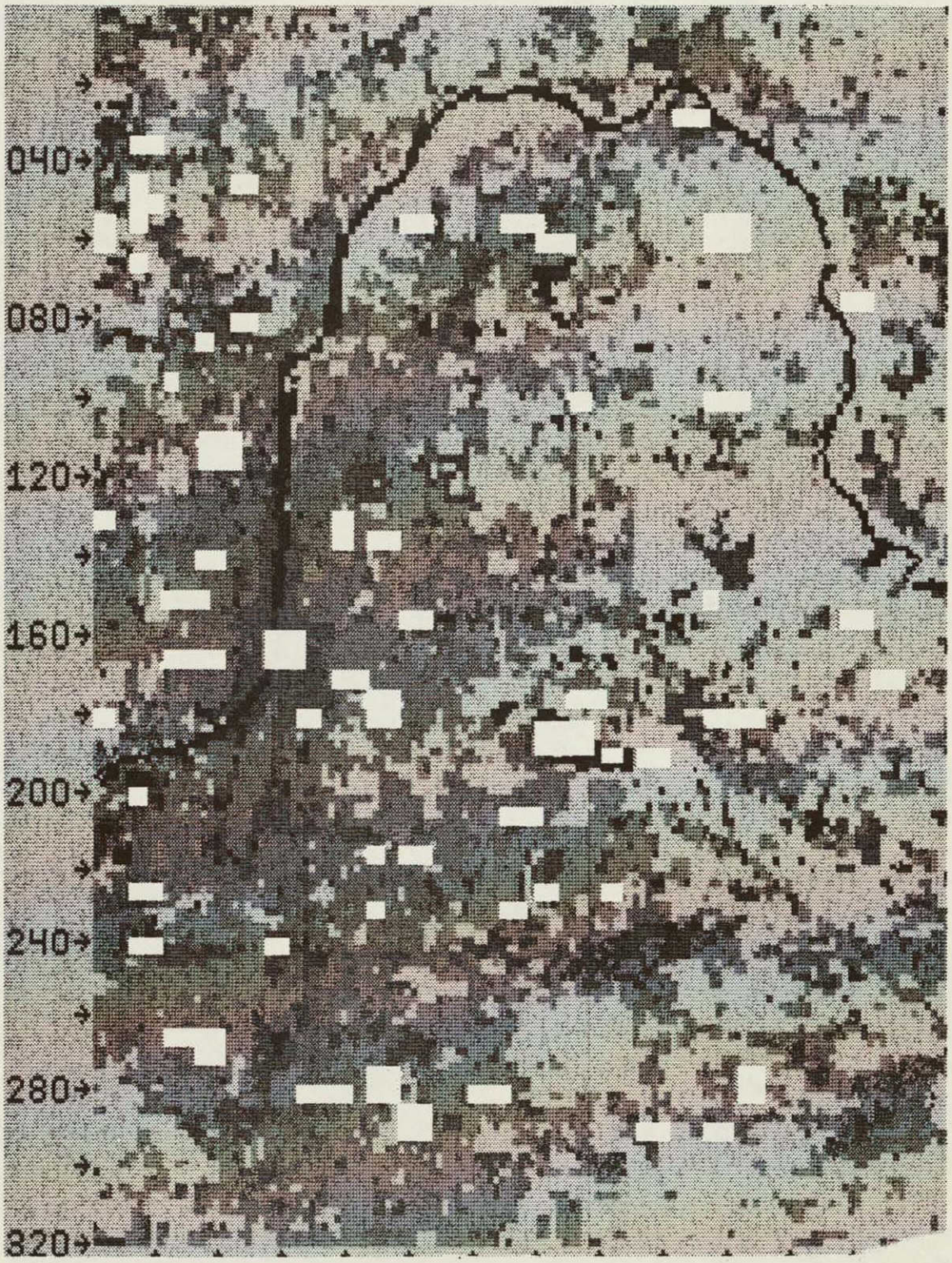


Figure 39. Classification results map of the Grand Rapids study area showing locations of test fields used in evaluating classification performance.

```
-COMMENT EVALUATION OF GRAND RAPIDS ECHO CLASSIFICATION
*PRINTRESULTS
RESULTS TAPE(444),FILE(6)
PRINT OUTLINE(TEST)
SYMBOLS &,*,+,I,-,=,.,V,M,M,M,M
GROUP COMERCL(1/1,2,3/),OLD RES(2/4/),NEW RES(3/5,6/)
GROUP NONURBAN(4/7,8/),WATER(5/9,10,11,12/)
DATA

TEST 1
          ********* TEST FIELDS FOR COMMERCIAL *********

TEST 2
          ********* TEST FIELDS FOR OLDER RESIDENTIAL *********

TEST 3
          ********* TEST FIELDS FOR NEWER RESIDENTIAL *********

TEST 4
          ********* TEST FIELDS FOR NON URBAN *********

TEST 5
          ********* TEST FIELDS FOR WATER **********

END
```

Figure 40. Control card listing used to generate a performance table for test fields and to outline those field locations in the Grand Rapids analysis area.

TEST CLASS PERFORMANCE

Group	No. of Samps	Pct. Corct	Number of Samples Classified into: Comercl	Old Res	New Res	Non Urban	Water
1 Comercl	126	86.5	109	3	11	0	3
2 Old Res	225	80.4	3	181	39	1	1
3 New Res	117	85.5	0	12	100	5	0
4 Nonurban	225	89.3	12	1	11	201	0
5 Water	115	93.9	2	0	0	5	108
Total	808		126	197	161	212	112

AVERAGE
PERFORMANCE BY CLASS (435.7/5) = 87.1

OVERALL
PERFORMANCE (699/808) = 86.5

Figure 41. Classification performance table for the five covertypes of interest as stated in the analysis objectives for Grand Rapids, MI.

CONCLUSIONS

The description of the analysis techniques used in this ECHO case study should have increased your knowledge of the principles of digital analysis and given you an appreciation for the potentials of a processor which utilizes spatial as well as spectral information about objects on the ground. The ECHO processor shows interesting possibilities for scene noise reduction (See Figure 5 , page 4), object definition and greater operational power for digital production of maps and tables qualitatively comparable to those produced by other mapping techniques.

EXERCISE 1 STATE ANALYSIS OBJECTIVE

The Indianapolis (Indiana) Department of Transportation (DOT) has contracted your company to use satellite data for location of general urban uses on the western side of Indianapolis. The area of interest is defined by the four USGS quadrangle maps which may be obtained from your tutor. The DOT is interested in determining the major industrial, commercial, and residential areas so they can make transportation network and planning maps for newly developing areas.

Design and write an analysis objective suitable for the needs of the Indianapolis Department of Transportation as described above.

CHECK WITH YOUR TUTOR
AFTER YOU HAVE WRITTEN YOUR OBJECTIVE
OR IF YOU HAVE QUESTIONS.

EXERCISE 2 EXAMINE DATA QUALITY

Examine the imagery of Landsat frame 1411-1581 (September 7, 1973) in Figure 42. Note the location of the analysis area within the data. Make notes about any data quality features apparent from this imagery. Next generate grayscale maps for a representative set of channels of the data over the west side of Indianapolis (LARS Run 73001011*; channels 5,6,7,8; lines 241 to 610; columns 342 to 691). Examine this data set for quality as discussed in the example. The data set with which you will be working has been geometrically corrected so it has a scale of 1:24,000, which matches the scale of U.S.G.S. 7.5 minute topographic maps. Note any data problems that might affect your analysis efforts later on in the analysis sequence. Discuss your findings with your tutor.

*The DUPLICATERUN processing function has been used to make a copy of this run for your terminal site. Consult your tutor for the correct tape and file number.

Figure 42. Band 5 imagery of Landsat frame 1411-1581 showing location of the Indianapolis analysis area.

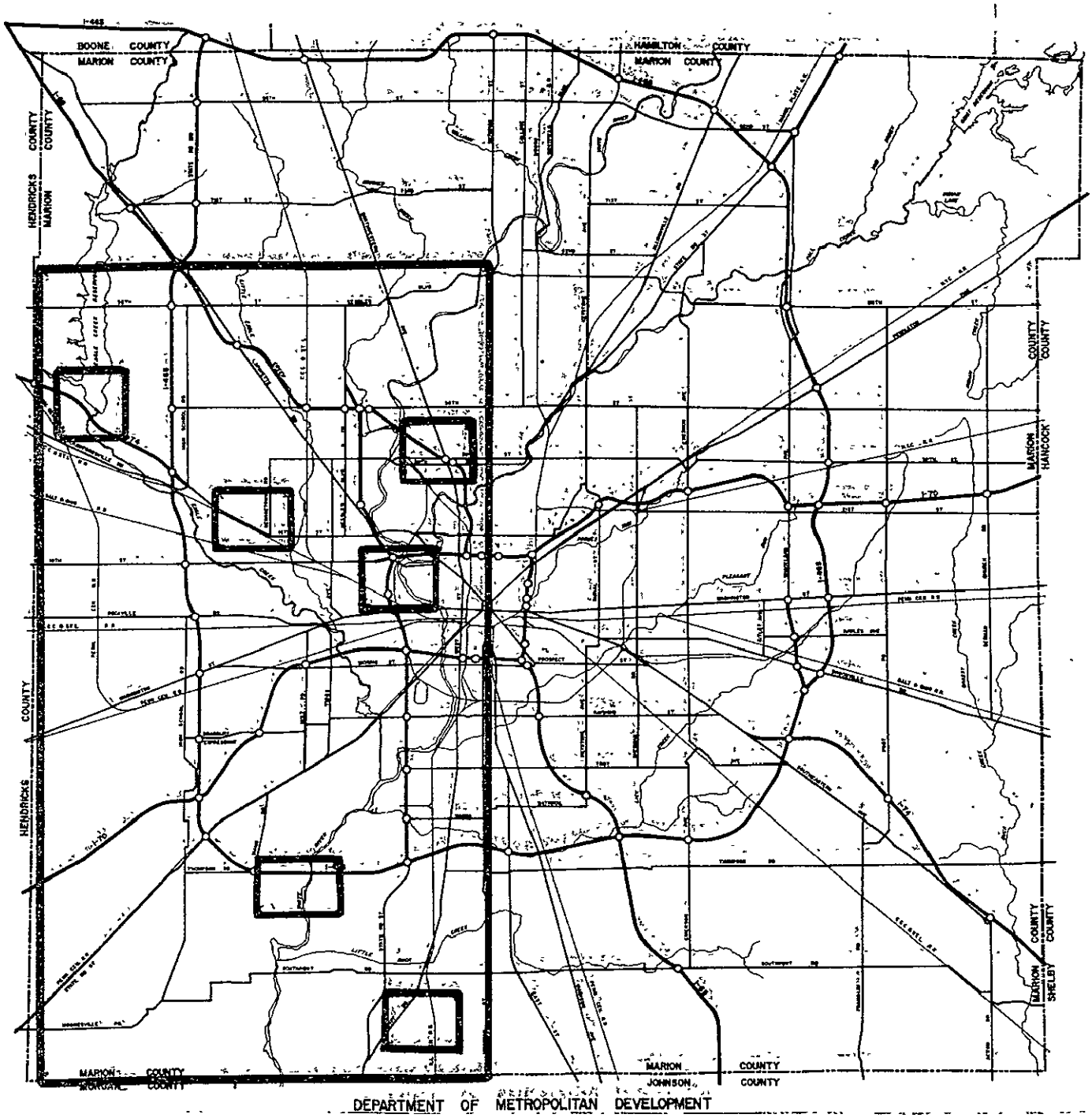


Figure 43. Transportation map of the Indianapolis metropolitan area showing the analysis area and the location of six color infrared aerial photographs to be used as reference data.

EXERCISE 3 CORRELATE REMOTELY-SENSED DATA WITH REFERENCE DATA

Obtain the following reference data from your tutor:

- U.S. Geological Survey 7½ minute topographic quadrangle maps of the area.

- 6 color infrared aerial photographs of portions of the area.

The reference data available for the Indianapolis area includes color infrared photography taken two weeks before the Landsat scanner data. The locations of the areas shown in these reference photographs are shown in Figure 43. One of the simplest ways of correlating these photographs to the cluster maps is for the student to refer back and forth between the two "images" to relate spatial similarities. These photographs can also be visually overlayed onto the printed cluster maps using an optical device like a Bausch and Lomb Zoom Transfer Scope.

The U.S. Geological Survey quadrangle maps of the area might be useful for determining general location in the data. Identify the covertypes as accurately as you can, making use of the color infrared photographs, and your knowledge of the area.

Associate the reference data with the grayscale maps and mark on the computer grayscales with a felt tip pen the general location of the covertypes of interest (agriculture, urban, forest, and water). Be sure to mark key features such as major roads, rivers, lakes, railroads, and parks. NOTE: Some features are more apparent on one printout than the other.

EXERCISE 4 SELECT TRAINING AREAS

Using the available reference data, the grayscale printouts and the guidelines described on page 23, select four to six candidate training areas which are representative of the scene. Make sure that every covertype of interest (commercial, residential, water, etc.) is included in at least one of the candidate training areas.

Let your tutor know how you're doing.

EXERCISE 5 CLUSTER TRAINING AREAS

Generate cluster maps and accompanying statistical output for each individual training area you selected in the previous step. Be sure to examine your color infrared photography and topographic maps to make a good estimate of expected covertypes in each training area.

Discuss your reasoning with your tutor.

EXERCISE 6 ASSOCIATE CLUSTER CLASSES WITH INFORMATION CLASSES

To carry out this step of the analysis, maximum use must be made of all available reference data so that the cluster classes will be reliably identified. If errors occur in this step of the analysis, the training data supplied to the classifier will not be representative of the information classes. The association of cluster classes and information classes is laborious and time consuming, but this step is most important for ensuring that the classifier is correctly trained.

There are several points to remember. One point is that if you feel a single cluster corresponds to more than one covertype, then it should be identified that way. Another point is that cluster class numbers from different training areas do not necessarily correspond to the same information classes. Each candidate training area was clustered separately, and the results of the clustering depend only on the data being clustered. Use your reference data from Exercise 3 to assist you in this Exercise. Check with your instructor after you have completed the table on the next page.

Indianapolis Area Cluster Identification

Cluster Number	Candidate Training Area					
	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						
7						
8						

Show Completed Table to Your Tutor

EXERCISE 7 CALCULATE STATISTICAL DISTANCE BETWEEN CLUSTERS

Examine the cluster class grouping tables at the end of each cluster output. Note that the processor has made some suggested groupings of cluster classes based on the class pair statistical distances. These suggested groupings are merely possible combinations and should be examined by you as the analyst using the information class identification you assigned in the previous exercise. Similar covertypes having small statistical distances are logical class combinations. Classes which are not paired with other cluster classes in the grouping table and which contain less than 40 data points should be considered for deletion in the next portion of this analysis sequence.

Using the cluster class grouping information you determined above it is now necessary to combine similar clusters from within each training area and to reassemble the statistics from all the training areas into a single statistics deck. Be sure to instruct the MERGESTATISTICS processor to punch the new statistics deck so that it may be used in the next analysis step. Check with your instructor if you have questions.

EXERCISE 8 CALCULATE DISTANCES BETWEEN CLASSES

Using your merged statistics deck from the previous analysis exercise, you may now use the SEPARABILITY processor to calculate which of the classes from your different training areas are spectrally similar and therefore should be combined into a new condensed statistics deck. Be sure to have these new statistics punched on cards so you can use them during the ECHO classification in the next step of the analysis sequence.

EXERCISE 9 CLASSIFY STUDY AREA

Determine the lines and columns of the area on the west side of Indianapolis that you wish to classify using ECHO. Select the appropriate cell width and homogeneity and annexation thresholds; and prepare your control cards for an ECHO classification of the study area. Discuss your parameter selection with your tutor. You will be assigned a tape number to use for the two phases of your ECHO classification as well as a tape to store your final classification results. When your tutor gives you these tape numbers note them in the space provided below.

ECHO phase 1 intermediate results	Tape ( )
Final ECHO results	Tape ( )

A set of phase one results of an Indianapolis analysis have been generated using different homogeneity thresholds and are available from your instructor. You are encouraged to utilize these intermediate results with different annexation parameters to explore the effects of varying this particular ECHO parameter.

EXERCISE 10 DISPLAY RESULTS

Produce a grayscale map of your Indianapolis ECHO Classification. The type of map (alphanumeric, gray tone, photographic image, etc.) you select should be usable for a preliminary visual evaluation of your analysis results. It is also important that the cost of generating this intermediate product be justifiable in terms of its value as an analysis tool in this particular analysis step. More elaborate and costly output products, such as hard copy color maps, are most often not required (if at all) until the final results product is generated for the user's needs.

Show your grayscale map

to your tutor.

EXERCISE 11 EVALUATE CLASSIFICATION RESULTS

A card deck of test fields is available from your instructor to help you evaluate your ECHO classification accuracy. Use these cards to measure your classification performance and repeat portions of the analysis as needed to produce a product which satisfies the analysis objectives of the Indianapolis Department of Transportation.

Let your tutor

know how you did.

REFERENCES

Hoffer, R., "Computer-Aided Analysis of SKYLAB Multispectral Scanner Data in Mountainous Terrain for Land Use, Forestry, Water Resources and Geologic Applications," LARS Information Note 121275, Purdue University, 1975.

Kast, J.L., ed., "ECHO Users Guide," LARS Information Note 083077, Purdue University, 1977.

Kettig, R.L. and D.A. Landgrebe, "Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects," LARS Information Note 062375, Purdue University, 1975.

Landgrebe, D.A., "Test for Boundary Finding Per Field Classification," Final Technical Report, Volume I, NASA Contract NAS-14970. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1977.

Lindenlaub, J.C., and J.D. Russell, "An Introduction to Quantitative Remote Sensing", LARS Information Note 110474, Purdue University, West Lafayette, Indiana, 1974.

Phillips, T.L. ed., "LARSYS Users Manual", Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1973.

Swain, P.H., "Pattern Recognition: A Basis for Remote Sensing Analysis", LARS Information Note 111572, Purdue University, West Lafayette, Indiana, 1972.

Swain, P.H. and S.M. Davis, eds., "Remote Sensing: A Quantitative Approach," McGraw-Hill International Book Co., Dusseldorf, 1978.

Swain, P.H. and R.C. King, "Two Effective Feature Selection Criteria for Multispectral Remote Sensing," LARS Information Note 042673, 1973, Purdue University, West Lafayette, Indiana.

Wacker, A.G. and D.A. Landgrebe, "The Minimum Distance Approach to Classification," Technical Report Tr-EE 71-37, The School of Electrical Engineering, Purdue University, 1972.

Appendix I

Self-Check Answers

Page 12

1-A Name the four components of an analysis objective.
1.) Location of the study area
2.) Covertypes to be mapped
3.) Applications for which the classification results will be used
4.) Classification accuracy desired

1-B Write an analysis objective that you might use in solving a problem in your area of interest.
Students' analysis objective should include the four components listed above and deal with a remote sensing problem of interest to him.

Page 18

II-A What are two characteristics which may decrease the usefulness of your Landsat data?
1.) cloud cover
2.) striping
3.) haze
4.) snow cover

II-B Name two types of geometric preprocessings which aid in the analysis of Landsat data.
a.) rotating
b.) deskewing
c.) rescaling

II-C What are the three types of reference data which can be correlated with remotely-sensed multispectral data?
1.) aerial photographs
2.) maps
3.) previous analysis results
4.) on-site observations

Page 24

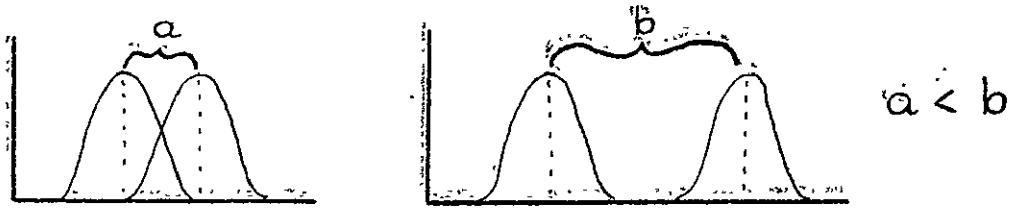
III-A How should you select training areas? Where should they be located? How many should there be and how large should they be?
LOCATION of training areas should be where there is some reference data available.
- Areas should contain representative samples of all the covertypes of interest in the study area.
- Training areas should be distributed throughout the entire study area.

SIZE of the training areas is dependant on the size of the study area, but as a guide should be from 25 to 100 lines by 25 to 100 columns each.
NUMBER of training areas varies with size and complexity of the study area, but usually 4 to 8 areas are sufficient for training.

III-B How does the analyst use clustering to aid in developing training statistics?
Clustering is used to determine the natural spectral structure in the data. The spectral subclasses within each training area have similar spectral reflectance characteristics.

III-C Why are cluster classes associated with information classes?
Cluster classes are associated with information classes to determine which spectral class represents each information class of interest before proceeding with the classification step.

III-D What is "statistical distance"?
Statistical distance is the distance between two density functions reflecting not only the ordinary (Euclidean) distance but also the "spread" of the data.



III-E Why are statistical distances between clusters calculated?
Statistical distances between clusters are calculated because statistical distances can be used by the analyst to determine which cluster classes are similar and to serve as an indicator of probability of correct classification.

Page 61

IV-A Referring to the flow chart on page 48 (Figure 29), briefly describe the classification process which is used by ECHO.
The data is subdivided into rectangular blocks of data points called cells. These cells are tested for homogeneity and kept as a unit if they meet a user specified threshold or if the cell is not homogeneous each of its individual pixels is classified independently.

The homogeneous cell is then tested against adjacent fields (one or more cells) for annexation based on a user specified threshold. If the cell is similar to an adjacent field it is annexed to that field. If it is not similar it becomes a field. After all cells have been tested for homogeneity and annexation all remaining fields are classified using the maximum likelihood classifier.

IV-B What are three different formats in which the ECHO classification results may be displayed?
1.) alphanumeric line printer maps
2.) tables
3.) black and white gray tone maps
4.) black and white or color photographs

Page 64

V-A What three criteria should be met in selecting test fields?
1.) random or unbiased selection
2.) sample full range of covertype variation
3.) sufficient size for statistical evaluation

V-B How can you evaluate the classification performance of the ECHO processor?
To evaluate classification performance the analyst might simply examine the map product for qualitative assessment. However to accomplish a quantitative evaluation the analyst should supply test fields from within the data for each covertype of interest, and calculate the performance of these fields.