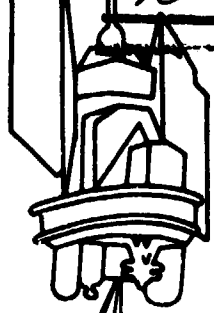
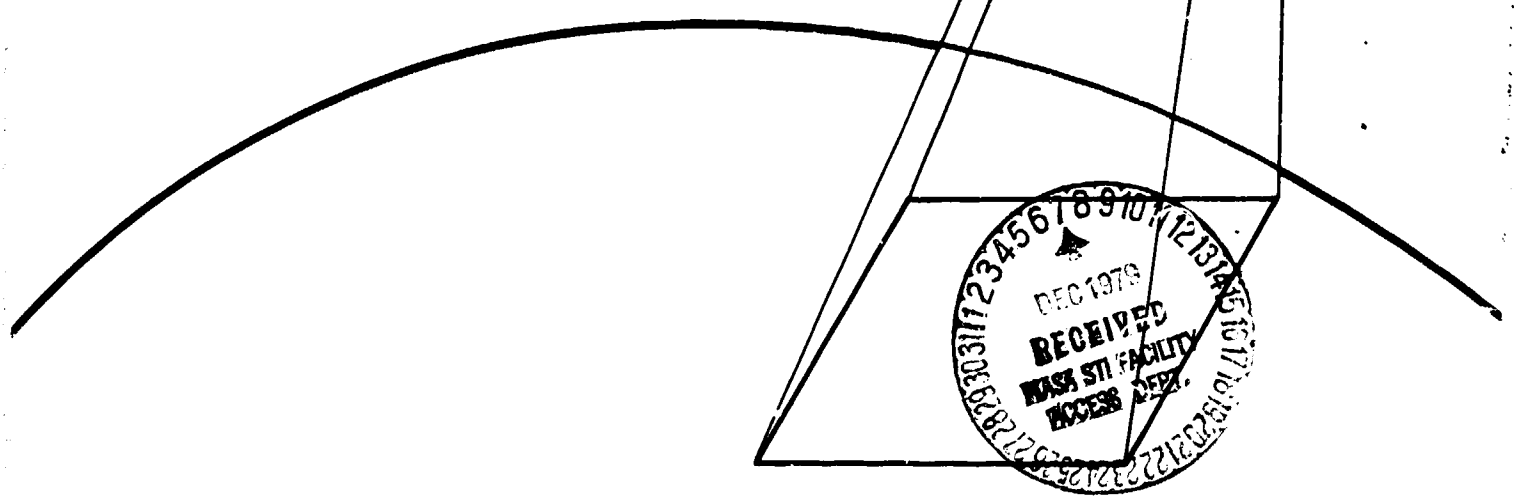
LARS Technical Report 102679



Computer-Aided Processing of Landsat MSS Data for Classification of Forestlands

Ross F. Nelson
Roger M. Hoffer



Laboratory for Applications of Remote Sensing
Purdue University West Lafayette, Indiana 47906 USA
1979

COMPUTER-AIDED PROCESSING OF LANDSAT MSS DATA FOR CLASSIFICATION OF FORESTLANDS

by

Ross F. Nelson

and

Roger M. Hoffer

October 1979

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 - INTRODUCTION

Background

At 11:05 PDT, July 23, 1972, the first Earth Resources Technology Satellite (ERTS-1) left the Western Test Range at Vandenburg Air Force Base near Lompoc, California atop a two stage Thor-Delta rocket. Approximately one hour later ERTS-1 was inserted into its final orbit 915 km (570 miles) above the earth's surface, travelling in excess of 26,000 km/hr (16,500 mph).

The satellite circuits the earth 14 times a day in a near-polar orbit, allowing almost complete coverage of the earth's surface once every 18 days. ERTS-1 was to provide multispectral scanner (MSS) and Return Beam Vidicon (RBV) camera imagery, and collect remote ground station data. The numerical information was telemetered to ground stations in Fairbanks, Alaska, Goldstone, California, and Greenbelt, Maryland. Later, receiving stations were built in Canada, Brazil, and Italy.

The satellite "was designed to demonstrate the feasibility of mapping and monitoring earth surface features from space." (Reeves, 1975, pg 569). The spacecraft"s design life was one year; in that year NASA expected the satellite would provide multispectral scanning and RBV camera imagery for research and evaluation in a variety of application disciplines.

ERTS-1 (later called Landsat 1) immediately developed problems with the RBV cameras, but the multispectral scanning system remained fully operational until the fall of 1977 when MSS band 4 (0.5-0.6 um) became inoperable. Complete instrument failure occurred January 6, 1978, about five and one half years after launch (Rohde, 1978). Since the summer of 1972, two more earth resources satellites have been placed in orbit. Landsat 2 joined Landsat 1 on January 22, 1975, and Landsat 3 was launched March 5, 1978. Landsat 2 and 3 currently provide repetitive MSS coverage of almost every spot on the earth's surface once every nine days.

In it's first year of operation, Landsat 1 provided "complete, cloud free coverage of the United States; cloud free coverage of a sizeable percentage of the remaining land surface, polar, and coastal areas of the earth; and repetitive coverage ... which shows significant temporal changes in the United States and other land areas of the world." (Reeves, 1975, pg 570).

Much research has been conducted in the last six years documenting the abilities and limitations of manual interpretation and computer-aided analysis of the Landsat data. NASA has sponsored investigations of potential imagery uses in the United States and fifty two other countries. Landsat data has been used to delimit, inventory, and/or detect changes on forests, agricultural lands, rangelands, wetlands, surface and near-surface marine environments, urban and rural areas. Agricultural applications include the use of Landsat MSS imagery to inventory and predict crop acreages and yields. Extensive soil surveys have been compiled in Indiana and Missouri. Rangeland uses include identification and monitoring of areas prone to overgrazing and fire. Forestry applications include incorporation of Landsat data into a multistage sampling system which provided estimates of timber volume, total timber resources present, and timber stand conditions on a site in northern California. Landsat has helped to define the extent of clear-cutting in Oregon, and Canadian foresters have used the imagery to delineate burned areas in northern Saskatchewan. Spacecraft data have been used by the emerging third world nations to produce extensive initial inventories of their undeveloped, unexplored lands. Brazil is taking advantage of Landsat's repetitive coverage to study and control the development of the Amazon forests. Many studies have found the data useful in delineating flooded areas, locating bodies of water, and mapping snowpack in order to predict runoff. Landsat data have been used to locate near-surface groundwater sites by interpretating linament patterns, surface lithology, vegetation, and geomorphic properties of an area.

The Landsat MSS imagery lends itself to small scale cartographic projects since it provides a synoptic view over a wide area from a near

vertical angle and maps may be produced soon after data acquistion. The U.S. Bureau of Census has mapped rural and urban boundaries near densely populated areas, and monitors the boundaries for significant changes. The scanner's ability to penetrate clear water to depths up to 20 meters (65 feet) enabled cartographers to chart shallow underwater features in the Caribbean.

In short, the MSS data hold a wealth of information if interpreted properly and if the limitations of the data are recognized. Interpretation may be facilitated by taking advantage of computer-aided analysis techniques (CAAT).

Computer-Aided Analysis Techniques

Most classification procedures demand that training statistics be developed. LARS paesonnel have found that a procedure combining clustering (unsupervised) and supervised techniques is the most effective method of formulating the training statistics. This approach, called the Multicluster Blocks (McB) has been shown to be more accurate and cost effective than the commonly utilized supervised method of training (Fleming and Hoffer, 1977). The McB approach formulates training statistics by clustering relatively small, heterogeneous areas (approximately 1600 to 3600 pixels per block). The spectral classes formed in each block are identified using photointerpretive techniques. The statistics are merged to form the final training deck used by the classifier. Classification performance depends heavily on the analyst's ability to correctly identify the spectral classes and to properly merge those classes. Spectral class identification in turn depends on the photointerpretive capabilities of the analyst or on the availability of ground information (such as type maps) on the training blocks.

A software system (EODLARSYS) was recently developed which has the capability of reducing analyst involvement in the development of training statistics. A subset of this system, a collection of processors called Procedure 1, more fully automates the classification process. In Procedure 1, once training points of known identity are selected and defined to the computer, various processors are employed to 1. compile the training point information; 2. cluster the scene; 3. label the clusters by comparing cluster statistics with the dot information; 4. classify

the scene using the labelled cluster statistics; and 5. display the scene and estimate classification performance (see Figure 1.1).

The Multicluster Blocks approach has been successfully used to develop training statistics to accurately classify forested areas (Hoffer, 1975; Fleming and Hoffer, 1977). Analysts have demonstrated the abilities of the Procedure 1 processors on agricultural areas (MacDonald, 1976) but the processors have not been tested in a forest-land situation.

The interest in Procedure 1 stems from the fact that it can directly utilize forest inventory data that may currently be available on many private, state, and national forests. For instance, Washington and Minnesota use a statewide network of field checked points to inventory their forest lands. This readily available gorund information may be located in the Landsat scene. Once this inventory information is input, demands on the analyst are minimal; the classification procedure is fully automated. Multicluster Blocks does not have this capability. McB may use the inventory information to help identify the spectral classes in each training block, but the procedure remains heavily dependent on the abilities of the analyst. This dependency is most acute in the statistics merging phase where the analyst actually constructs the final training statistics deck. Those who have had to merge statistics decks know that formation of the final deck may be viewed as an art, one which mirrors the analyst's experience and intuition. Procedure 1 removes this source of analyst interaction.

The overall objective of this study then is to compare the Multicluster Blocks approach to Procedure 1. The methods use different clustering processors to obtain training statistics. These statistics are used by the individual classification processors to categorize the area of interest into informational classes. The clustering processors and classification processors used in McB and P-1 are compared. Figure 1.2 depicts the experimental design of the study.

To summarize, the overall objective of the study is to compare the McB approach for developing training statistics to Procedure 1's approach. This comparison is made keeping in mind that P-1 has inherent operational advantages which make it attractive in situations where

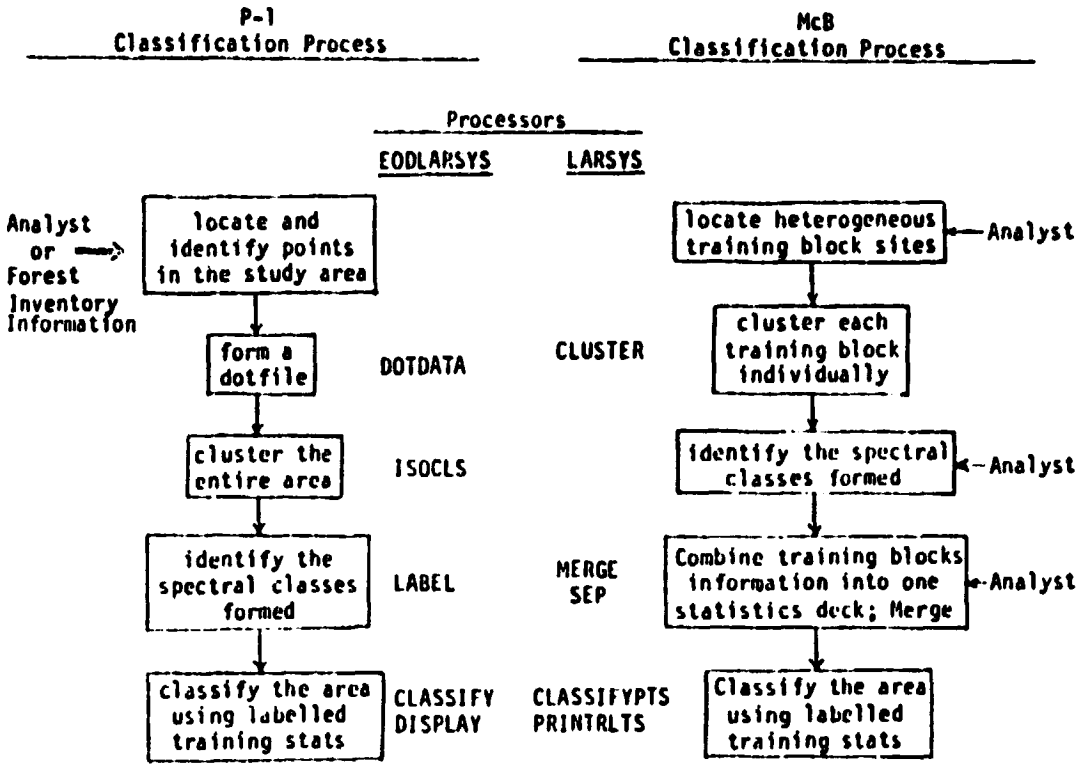


Figure 1.1 Comparison of the Procedure 1 and Multicluster Blocks approaches to classification of forest lands.

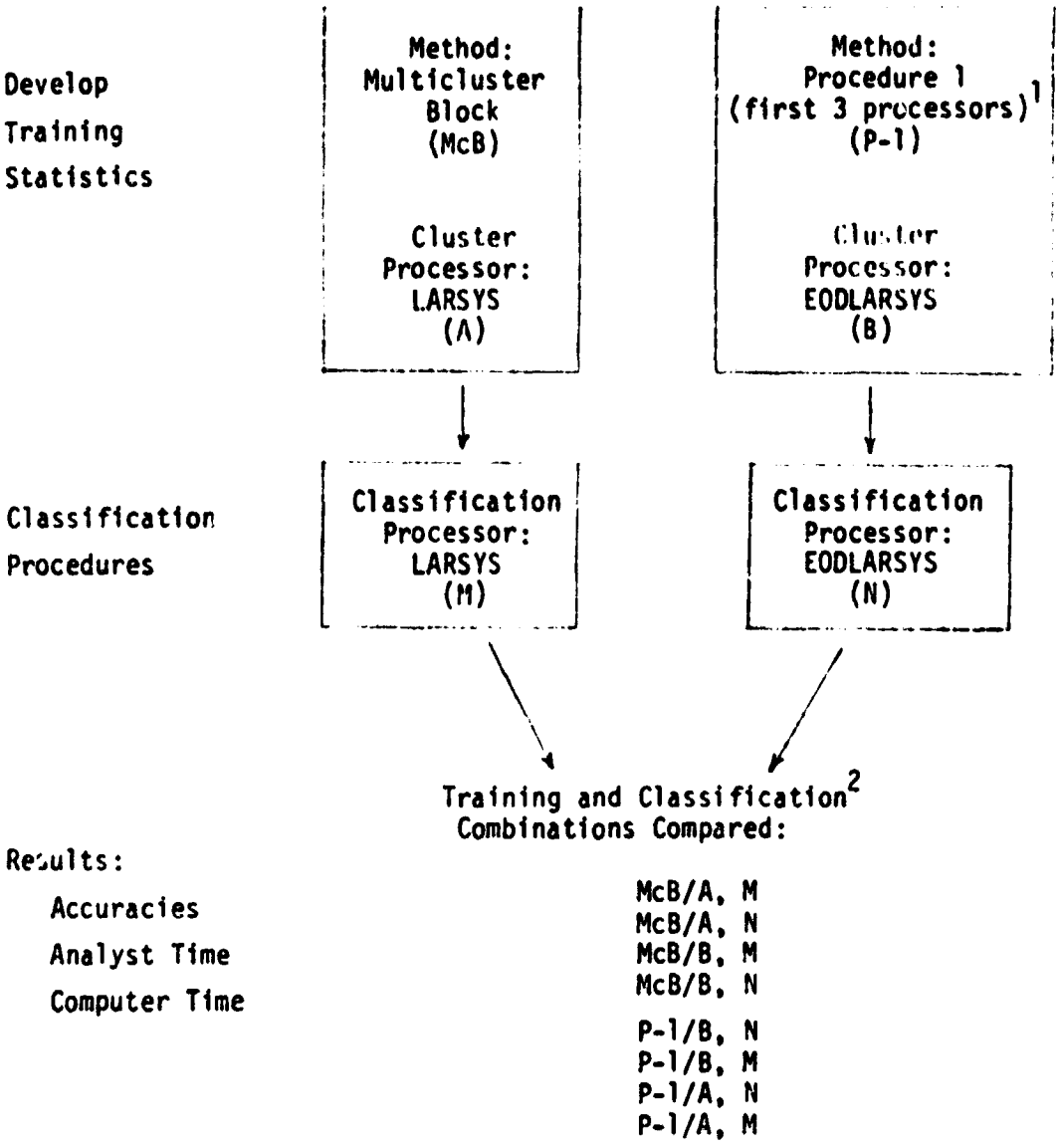


Figure 1.2 Results obtained which enable comparison of 1. McB vs P-1; 2. the clustering processors; 3. the classification processors.

---

1. The first three P-1 processors are used to develop training statisitcs - DOTDATA, ISOCLS, and LABEL.

2. This figure represents only a partial listing of the results obtained in this study. A more thorough explaination is given in the Procedures section.

forest inventory data is available. Within this framework, a number of subobjectives may be listed.

1. A parameter study is required in order to formulate a parameter set which allows efficent clustering of Landsat MSS forestland data. The study is necessary because little work has been done with P-1's clustering processor ISOCLS to discern the effects of its control parameters in a forested situation.
2. The clustering processors, ISOCLS and CLUSTER (used in P-1 and McB respectively), are compared to determine which works most efficently and produces the most accurate set of training statistics using Landsat MSS forestland data and forest inventory data.
3. Two classification processors are compared, one a LARSYS routine associated with the McB approach, the other an EODLARSYS processor associated with P-1.

CHAPTER 2 - LITERATURE REVIEW

Current Forest Survey Methods

The necessity for accurate, up-to-date resource information is highlighted by a number of federal laws passed in the last two decades. Natural resource legislation includes the Multiple Use-Sustained Yield Act of 1960, the National Environmental Policy Act (NEPA) of 1970, and the Resources Planning Act of 1974 (RPA).

Of the three, the RPA has the most significant impact on forestry and remote sensing applications. The legislation specifies that the Forest Service assess all forest and rangeland resources in the U.S. and develop long range resource plans based on these assessments. Assessments are to be made every 10 years beginning in 1979, and the resource plans drawn from the assessment would be responsive to changes anticipated in the coming decade (Glascock, 1976).

The RPA is a landmark piece of legislation, and demands the most intensive information of our natural resources. Essentially, the law demands a full national forest inventory once every 10 years. The inventory process may be made more timely and accurate by using Landsat data to supply some of the necessary information in a cost effective manner (Nichols, 1974; Heller, 1976).

Current forest inventory systems generally use ground information (permanent plots or sample plots) alone or in conjunction with aerial photography to evaluate their resources. Aerial photos are commonly used to stratify a forest. These strata are then sampled on the ground to obtain more detailed information (Husch, Miller, and Beers, 1972).

Such sampling designs (one and two stage) are in use throughout the country. The State of Maine used a combination of techniques, including remeasured plots and an independent two stage (ground and air-photo information) sample to reinventory their lands in 1970 (Ferguson and Kingsley, 1972). Colorado's first statewide inventory used a two

stage design (Miller and Choate, 1964) as do the National Forests located in Colorado (Born, 1977; USFS, 1978; Matteson, 1978). Two stage designs have been used in Minnesota (USFS, 1974) and in Missouri (Spencer and Essex, 1976).

Recently, Landsat data have been incorporated into the forestland sampling techniques, thus forming a three or multistage design. The Landsat data may be used to initially stratify the area of interest. Airphoto and ground samples may then be allocated within the strata of interest in order to adequately characterize the forestlands.

Landsat data was used to stratify areas in the western part of Washington into broad density and cover type classes (Harding and Scott, 1978). A multistage technique was also used on the Plumas National Forest (Nichols, 1974). Both studies concluded that the use of Landsat data saved considerable money that otherwise would've been spent on ground and/or airphoto sampling. Nichols maintained that accuracy was not sacrificed by including the satellite platform data.

The RPA specifies that the U.S. Forest Service must reinventory its lands once every ten years. Landsat data, current forest survey information, and machine processing may help achieve this goal. Analysis methods available to the Landsat data user are briefly reviewed in the following section.

Analysis of Landsat Imagery

The numerical data telemetered to earth stations by Landsat may be used for both qualitative and quantitative analyses. At one extreme the reflectance values may be assembled to produce a reconstructed image of the scene viewed by the satellite. This visual model may then be photointerpreted. At the other extreme is a hypothetical system which inputs the Landsat data, formats it, analyzes it, and outputs desired land use statistics, all without the analyst ever seeing the scene image. Between these extremes lay many documented techniques which facilitate extraction of information from the data.

Visual Interpretation of Landsat Data

Landsat data may be used to form a photographic image of the scene called a color composite (bands 4, 5, and 7), or black and white images can be constructed using information from a single band (usually band

5). These images are used

1. to give the analyst a better feeling for the area with which he is working;
2. to discern cloud cover conditions;
3. to broadly stratify a scene into smaller areas of interest.

Howard (1976) and Jaakola (1976) used color composites and photointerpretation techniques to separate forested and nonforested areas. Both divided their forestlands into broad subclasses. Jaakola concluded:

> "This kind of satellite image analysis, as supported by a fairly small amount of surface observations, gives reliable acreage estimates which can be utilized directly in planning the uses of the resources, or, indirectly, in stratifying the area for more detailed forest inventories."

Heller (1976) studied the different formats available and found that forest managers of large ownerships (greater than 4,000 hectares) could benefit by using Landsat enlargements for planning. He found 1:250,000 color composite imagery most useful. Comparing machine processing and photointerpretation procedures, Heller et al. (1975) stated:

> "Classification can be done most effectively by computer, but photointerpretation produces equally accurate results. Choice would depend on availability of trained people and equipment."

Visual interpretation of the Landsat imagery gives the user a broad overview of the study area, and may be used for type mapping and planning. However, machine processing of the Landsat data has a major advantage in that it is a repeatable sequence. Classifications are consistent since they're based on mathematical precepts, though the method used to develop the training statistics can significantly influence the quality of the classification. Photointerpretation is subjective; given the same interpreter and the same study area on two different days, the type map will be drawn differently. Hence photointerpretation of the Landsat data disregards the data's most important attribute, its numerical characteristics.

Machine Processing - An Overview

Machine processing of Landsat data demands that the computer be 'taught' to recognize informationally important spectral classes. The computer might be programmed to define spectral classes using ground data selected on the basis of informational importance (supervised

technique) or a scene could be clustered into a given number of spectral classes (cluster analysis or unsupervised technique). Those classes are then identified using ancillary information. Each method is overviewed below so that the reader might become more acquainted with the differences between the two. Studies using each of the techniques are reviewed in subsequent sections.

Supervised Development of Training Statistics - An Overview

Much of the multispectral classification work done today utilizes the supervised method to develop the training statistics for the classifier. A supervised approach to developing training statistics is the more understandable, straightforward method of the two (supervised and unsupervised) considered. Essentially the analyst selects areas of interest, tells the computer of their location and identity, then commands that the computer group all resolution elements in a scene into one of the appropriate classes specified by the training areas. Though perhaps more logical, the supervised approach is dependent upon the ability of the analyst to define spectrally separable informational classes. If the classes defined are not spectrally separable, cross-classification will result and accuracies will suffer.

A supervised approach to developing training statistics is a logical sequence of events, given below:

1. Specify the informational classes of interest.
2. Locate homogeneous training fields in each of the informational classes using ancillary data.
3. Formulate the statistics for each informational class.
4. Evaluate the statistics and subdivide those spectral classes which are multimodal.
5. Use these statistics to classify the area of interest.
6. Evaluate the classification.

Supervised classification techniques involve the use of training points or fields (ground information) which ideally include all possible spectral variations for a particular cover type (in a given scene). In order to run an accurate classification using a supervised approach, the following guidelines have been defined (Ellis, 1978):

1. Classes (informational) must be as spectrally distinct from each other as possible.

2. Training fields should be as homogeneous as possible.
3. Bimodal (mixed) spectral classes should be avoided.
4. Training fields should be representative over the entire area classified.
5. Training fields should be 40 acres in size or greater (approximately 36 pixels).

It should be noted that the training field size criterion was suggested in relation to a shrublands classification study undertaken in the central Rocky Mountains and on the Colorado Plateau. Training field size may vary according to the complexity of the study area; the more complex the area, the smaller the size of the training fields.

Ellis found two common problems inherent in the supervised classification procedure:

1. The analyst does not identify and define important spectral classes, thereby confusing and decreasing the effectiveness of the classifier, or he identifies an informational class that is not spectrally differentiable.
2. Within an informational class, training sites selected do not adequately characterize that class.

Despite problems and limitations, the supervised classification technique has been used successfully by individuals classifying forested and nonforested areas.

Unsupervised Development of Training Statistics - An Overview

An alternate approach to developing training statistics reduces analyst involvement in the definition of the spectral classes. The unsupervised approach to developing training statistics involves a clustering processor which groups the data into spectrally homogeneous classes. These spectral classes are then identified using any pertinent information available to the analyst, such as airphotos of the clustered area, vegetation maps, topographic maps, or ground checked points or fields. The analyst must specify the number of spectral classes to be output by the cluster processor. If one specifies too few clusters, the spectral classes will be multimodal, have large variances, and may not satisfy the maximum-likelihood classification processor assumptions. If too many classes are specified, one introduces noise into the system and the analyst may have problems identifying the output clusters.

A sequence of steps may be outlined for an unsupervised development of training statistics:

1. Cluster the study area.
2. Identify each spectral class formed using ancillary data.
3. Use the labelled cluster statistics to classify the area of interest.
4. Evaluate the classification.

If the 'proper' number of spectral classes is output, the clusters formed should be unimodal. If multimodal clusters appear, the area should be reclustered to produce more spectral classes. On the other hand, if too many spectral classes have been formed:

1. identification of the spectral classes may be difficult;
2. needless computer time would've been spent clustering;
3. needless computer time would be spent classifying the area; and
4. those spectral classes that are not spectrally separable should be merged.

The supervised method demands that the ancillary information be used first to define areas of interest. The analyst anticipates that the classes of interest are spectrally separable. The unsupervised approach clusters the data points into spectrally separable groups, then uses the ancillary information to identify them. These methods may be combined and altered to produce different methods of developing training statistics. One of these, the Multicluster Block method, is discussed in future sections.

Results of Studies Using the Supervised and Unsupervised Techniques

Studies Using the Supervised Method of Developing Training Statistics

Investigators have used Landsat (4 channel), Skylab (13 channel), and aircraft (variable number of channels) data to classify forestlands. In addition, multitemporal overlays have been used in some of the studies. These data sets incorporate multidate acquisitions over the same study area. The acquisitions are precision registered and analyzed, the hope being that the additional information will increase the classification accuracy.

Initially, machine processing of multispectral information made use of aircraft data, since satellite platform imagery was not available to the public prior to 1972. Smedes et al. (1969), Rohde and Olsen (1972), and Coggeshall and Hoffer (1973) successfully classified forested areas using aircraft data. Smedes used the best four channels of 17 channel MSS data flown over Yellowstone National Park. He developed training statistics for eight land-cover types and judged the classification accuracy to be 85%. Three years later, using 6 bands of low altitude aircraft data (0.4 to 1.0 μm wavelength region), Rohde and Olsen analyzed pine, spruce, red oak, white oak, black walnut, black locust, and sugar maple plantations. Accuracy of identification was estimated to be 85%. Coggeshall and Hoffer experimented with different combinations of 12 channel multispectral scanner data and concluded that accurate classification of deciduous and coniferous forests (overall accuracy, 80.5%) could be obtained using visible wavelength bands and either near or middle infrared bands.

Investigations using satellite platform imagery began soon after the launch of ERTS-1 (Landsat 1) in July, 1972. A number of studies have been completed outlining the capabilities of single acquisition Landsat data and the supervised technique for developing training statistics. Nichols (1974) used ERTS-1 data as the first stage of a multistage sampling scheme on the Plumas National Forest in California. The predominantly old growth forest was stratified into subclasses related to timber volume using photointerpretation and computer processing of the ERTS data. The multistage approach resulted in equal precision of estimation of timber volume at a 44% cost savings when compared to conventional methods. The timber inventory incorporating the Landsat data permitted estimation of the following parameters:

1. number of trees per acre;
2. square foot basal area;
3. basal area growth over a five year period;
4. cubic foot volume;
5. Scribner board foot volume;

at a cost of $0.072/hectare ($0.029/acre) (Gialdini, 1975).

Other studies incorporating Landsat data and a supervised approach include Kan and Dillman (1975) who reported 70-80 percent accuracies in separating hardwoods, softwoods, and regeneration; and Bryant, Dodge, and Warren (1977) who concluded that their supervised classification of hardwoods, mixed woods, conifers, and boglands showed large discrepancies when compared to a Seven Islands Company inventory. Another negative note concerning the usefulness of Landsat maps generated using supervised techniques was sounded by Mead and Meyer (1977). They ran a number of classifications using supervised training areas in Itasca County, Minnesota. They attempted to break out 11 land use categories - water, lowland conifer, upland conifer, mixed forest, bursh and shrub, grassland and open, agriculture, mixed lands, sedge meadows, urban, and sphagnum/leatherleaf. Their results are best summarized in the final paragraph of their report:

> "Evaluation of the various map solutions by experienced field resource management cooperators resulted in the judgement that classification accuracies were so low as to preclude practical use for their purpose at this time."

Classification results may be improved by increasing the amount of information available to the classifier. This additional information is digitized (if necessary) and precision registered with the existing Landsat data base. The additional information is added as new channels of data which may contain:

1. soils or parent material data,
2. topographic data,
3. multispectral scanner data,
4. planimetric data (land use or ownership information).

Several studies have utilized temporal (multidate) overlays and have found that the supplemental information increases classification accuracies. Williams (1976) ran a supervised classification on 24,300 hectares of commercial forestlands in North Carolina's Southern Pine Region (see also Williams and Haver, 1976). The area runs the gamut of forestland conditions, containing clearcuts, various regrowth stages, and artificial and natural pine regeneration. Using Penn State's ORSER system (Office for Remote Sensing of Earth Resources), principal components analysis, and winter-summer data temporally overlaid, he

attempted to stratify the southern pine forest by crown closure. Results showed a 94% agreement between computer and photointerpretation in breaking out hardwood stands, 96% agreement in pine stands, 54% agreement on clearcuts (major confusion class, pine regeneration), overall 90%.

In northern Italy, Lapietra and Megier (1976) used Landsat MSS data to estimate acreages of poplar plantations. Using four channels (single acquistion), the investigators achieved acreage estimation accuracies of 80%. Using three acquisitions (12 bands) and principal components analysis to reduce 12 bands to 6, accuracies jumped to 95% in the best case.

Kalensky and Wightman (1978) overlaid four Landsat acquisitions; the results of their investigation and a number of earlier studies[1] are shown in Table 2.1. The classifications were done using the supervised approach on the MICA system (Modular Interactive Classification Analyzer - developed at the Canada Centre for Remote Sensing, Ottawa, Ontario). Note that the average map accuracy (which takes into account the spatial displacement of a spectral class) was 13.6% lower than the average classification accuracy. The significant decrease in map accuracy may be due to

1. the effects of a 1.1 acre resolution element upon map accuracy;
2. possible data registration inaccuracies;
3. inaccuracies in the base map used as 'ground truth'.

Kalensky and Wightman (1978) found the most accurate supervised classification was obtained using multidate imagery, bands 5 and 7. The second most accurate classification used three acquisitions - 12 channels - and the third most accurate classification used four acquisitions - 16 channels.

To summarize, the supervised technique has the ability to produce fairly accurate classifications (80-95%) in forested areas. Accuracies improve if multidate overlays are used. Unfortunately, precision

---

1. The studies were done at the Forest Management Institute of the Canadian Forestry Service between 1974 and 1978.

Table 2.1 Classification and mapping accuracies of a digital image processing system (from Kalensky and Wightman, 1978).

CLASSIFIER	INPUT IMAGERY LANDSAT-1/MSS	CLASS		
		Agricultural land	Coniferous forest	Deciduous forest
		Classification Accuracy/Mapping Accuracy (%)		
Max.-likelihood single-date B4, B5, B6, B7	Sept. 05, 1972	87/73	87/76	78/65
	June 03, 1973	64/48	75/64	75/52
	Oct. 06, 1973	70/53	70/64	69/40
	March 18, 1973	94/79	75/66	70/51
Max.-likelihood multidate B4, B5, B6, B7	Sept. 05, 1972 March 18, 1974	88/76	88/74	74/63
	Sept. 05, 1972 June 03, 1973 Oct. 06, 1973	96/77	83/77	84/65
	Sept. 05, 1972 June 03, 1973 March 18, 1974	91/78	87/74	76/63
	Sept. 05, 1972 June 03, 1973 Oct. 06, 1973 March 18, 1974	92/79	85/76	84/67
L-L single-date B5, B7	Sept. 05, 1972	85/74	87/75	78/65
Max.-likelihood multidate B5, B7	Sept. 05, 1972 June 03, 1973 Oct. 06, 1973	97/78	85/78	85/65
Image 100 rectangular parallelpiped B4, B5, B6, B7	Sept. 05, 1972	61/58	82/78	80/64

overlays are expensive.[1] The supervised technique has two major drawbacks. First, the informational classes designated by the analyst may not be spectrally separable, leading to classifier confusion. Second, the front end of the process is analyst intensive. The unsupervised technique minimizes these drawbacks. The results of studies using this technique follow.

Studies Using the Unsupervised Method of Developing Training Statisitics

Studies using an unsupervised approach are not widespread, and most of the research has been conducted at the Laboratory for Applications of Remote Sensing (LARS), Purdue University, in Indiana.

1975 saw the culmination of a two year study of ERTS-1 data capabilities in mountainous terrain in Colorado. Both supervised and unsupervised approaches were tested to discern which method was best for producing training statistics. The study found that the modified clustering technique (later called the Multicluster Blocks technique) yielded consistently higher classification accuracies. By using one tenth of one percent of the total data set for training, a 94% correct Level II classification performance was obtained (Hoffer et al., 1975).[2]

Fleming, Berkebile, and Hoffer (1975) studied four different unsupervised analysis approaches on the Ludwig Mountain quadrangle in southwestern Colorado. Of the four methods studied on the 15,140 hectare study site

1. unsupervised cluster, 10 spectral classes;
2. unsupervised cluster, 20 spectral classes;
3. modified supervised; and
4. modified cluster, now called Multicluster Blocks;

the modified cluster analysis yielded the highest accuracy (76.6, 78.5, 70.0, and 84.6 percent, respecively).

Intensive Study Sites (ISA) in the western sections of the state

---

1. Precision registration may become economically more attractive once NASA begins geometrically correcting the Landsat data prior to sale.

2. Levels of classification detail used in this study (more complete definitions available in Anderson et al., 1976):
   Level I: forest, grassland, barren, water.
   Level II: conifer, deciduous, grassland, barren, and water.

of Washington were selected in order to inventory Washington's public lands (Edwards, 1977; Harding and Scott, 1978). Each ISA was approximately 16 by 20 miles, 512 by 512 pixels. They first tried an unsupervised classification on a few large blocks within each ISA, but due to the spectral complexity of the blocks they were unable to accurately identify the 55 to 65 spectral classes output. Their alternate analysis procedure involved using training blocks to develop the training statistics for pure stands of hardwoods and softwoods (supervised). Mixed stand training statistics were developed using an unsupervised approach since the analysts found it impossible to select 'typical' stands necessary to characterize mixed stands. The combined training statistics yielded 279 separate spectral classes(8 Landsat scenes used) which were ultimately grouped into six broad cover types. No accuracy statement was available for the classification, though blocks of cla s-ification results were scrutinized by a Resource Analyst and accepted if, in his estimation, they were adequate.

Fleming and Hoffer (1977) compared six different methods of developing training statistics, ranging from a purely supervised technique to a purely unsupervised technique and found that a Multicluster Blocks approach "(A) reduced the CPU time, (B) required relatively few man hours of time, (C) utilized the lowest amount of supprot data, and (D) yielded the highest overall classification accuracy." They also concluded that the "supervised approach was the most ineffective method of developing training statistics ...".

Conflicting results were obtained by Schubert (1978). He studied three methods of classifying tracts in central Alberta. Unsupervised accuracy of classification was 77.3% for all classes (10 land use classes), supervised - 78.3%. It cost three times as much to develop the unsupervised statistics, due in part to the novel method used to generate them. Additional classifications were run 40 miles north of the original study area, in a different soil-climatic zone. Here the supervised classification accuracies fell below the unsupervised. It is interesting to note that the most accurate method (83%) was an automatic classification system employing ratios and previously established ratio signatures.

This review of the literature did not provide a clear answer to the question of which technique (supervised or unsupervised) is better. The many studies come up with different conclusions. A supervised approach might be considered appropriate for simplistic situations such as agricultural settings since spectral variability within an informational class is limited by the homogeneous nature of the crops. An unsupervised approach lends itself to continuously changing, complex situations such as mixed forestlands where formation of the spectral classes might best be left to the computer.

A number of unsupervised methods of developing training statistics exist. Work by Hoffer et al. (1975), and Fleming and Hoffer (1977) have shown the Multicluster Blocks approach to be most efficent. However this method was not developed to optimize the use of sampled ground information available on federal, state, and private (commercial) forests. The EODLARSYS software system, specifically those processors used in Procedure 1, was designed to accomodate such grid data. The five processors of interest (DOTDATA, ISOCLS, LABEL, CLASSIFY, and DISPLAY) have not been tested on forest cover types. The five are evaluated in this study, and the three responsible for developing training statistics (DOTDATA, ISOCLS, and LABEL) are compared to the Multicluster Blocks approach.

Two Unsupervised Approaches - Results

The purpose of any classification scheme is to output a product of some use to the analyst or his clients. Products may vary from tables of acreage statistics to vegetation or land use maps at selected scales. In order to output these products, a classifier must be statistically trained to recognize informationally important spectral classes. Two methods of developing these training statistics are evaluated in this study. The actual processors involved in the study are reviewed in the Materials section; studies involving the two approaches are reviewed below.

Multicluster Blocks

The methods involved in developing training statistics using the McB approach are described in the Procedures section. Following are some of the results obtained using this approach.

The Multicluster Blocks procedure has been developed and tested by Mike Fleming in a number of studies located in southwestern and north central Colorado (Hoffer et al., 1975; Fleming and Hoffer, 1977; Hoffer and Fleming, 1978; Krebs and staff, 1976). All of these studies utilized ERTS-1 data acquired in 1973, and many of the studies have been undertaken in cooperation with INSTAAR (Institute of Artic and Alpine Research, University of Colorado).

Hoffer and staff (1975) reported an average level II classification accuracy of 94.5 percent on the Vallecito intensive study area. Classification (test field) accuracies ranged from 85.4 percent (deciduous) to 100 percent (water). Similar results were obtained on other intensive study areas throughout Colorado.

Another study (Krebs and staff, 1976) attempted to classify the Platoro Quadrangle (southwestern Colorado) into community types, specifically

1. dense conifer,
2. sparse conifer,
3. aspen,
4. oak,
5. coniferous-deciduous mixes,
6. various grassland types,
7. bare rock,
8. water.

Using the 'narrow' definition of the cover types,[1] overall classification accuracy was 69%, with bare rock classified most inaccurately, 43.4 percent. A 'moderate' definition of community types yielded an 82.7 percent overall accuracy, with accuracies ranging from 61.2 percent (sparse conifer) to 95.8 percent (coniferous/deciduous mixture). Using the classification products output (narrow definition) the U.S. Forest Service ran their own evaluation. From Kreb's report (pg 74) comes the following:

> "The gut level feelings of those working with the classification were that the data was good, but accuracy figures of 40 to 60 percent doomed any attempts for the U.S. Forest

---

1. Spectral classes obtained during clustering were assigned to a community type based on the type's definition. Use of the narrow definition indicates that, for instance, those spectral classes found only in a coniferous/deciduous mix would be called a coniferous/deciduous spectral class. A moderate definition of the c/d mix (a predominately coniferous mixture) would allow both dense conifer and c/d spectral classes to be assigned to it.

Service to actually use the data in planning efforts."

The study found Landsat data useful in classifying an area to Level II, but discovered community levels could not be delineated adequately.

The Multicluster Blocks technique, at first called the Modified Cluster Technique, has also been used with Skylab data.

> "Classification of Skylab MSS data resulted in a classification performance of 85.0% for the major cover types, which was somewhat less accurate than the Landsat classification, probably due to the poor quality of the Skylab data."
> (Hoffer and Fleming, 1978)

This report goes on to state that attempts to map forest cover types (Level III) using Landsat MSS data indicate

> "... that results at this level of detail would probably not be accurate enough to provide useful information for most users.".

The computer work in the four studies mentioned above were done by the same analyst, Mike Fleming. The current study serves not only as a comparison between the Procedure 1 processors and the Multicluster Blocks technique, but it will also test the reproducibility of the McB technique by a different analyst using data from the same general area.

Procedure 1

The EODLARSYS software package has 13 different processors available to the user. Only five of the processors concern this study, and only three of the five are involved in developing training statistics. In reviewing the literature, the term Procedure 1 is often encountered, and its definition varies. In its narrowest context, Procedure 1 refers specifically to methods of operation utilized in LACIE (Large Area Crop Inventory Experiment) where

1. random points are located in the Landsat scene and identified using ancillary information;
2. a portion of these dots are used to seed the clustering processor ISOCLS;
3. all processors - DOTDATA, ISOCLS, LABEL, CLASSIFY, and DISPLAY have fixed parameter levels.

For the purposes of this report, the following definitions suffice:

Procedure 1: Points 1 and 2 above are valid, howefer processor parameters are not strictly defined.

Self-Start Procedure 1: Same as the definition above, with the

exception that the clustering processor ISOCLS is not seeded with dots, it self-starts.

Most of the work using the P-1 processors has been done in agricultural areas under the auspices of LACIE. No work has been done in a forested area, though Reeves (1978) successfully differentiated rangeland from nonrangeland with probability of correct classification ranging from 93 to 100 percent. LACIE used the P-1 software system to monitor worldwide wheat production. The estimates were produced using repeatable procedures and were not revised using other information sources. The users established a 90/90 percent accuracy criterion; i.e., their regional or countrywide wheat harvest estimates should be within 10% of the actual harvest taken for that region 90% of the time (9 times out of 10 years). The LACIE studies showed that the analysis procedures are capable of meeting the performance criteria (MacDonald, 1976).

Procedure 1 software was developed for LACIE. The system is intended to monitor large agricultural areas where monotypic, regularly shaped crops predominate. The developers (Lockheed, under NASA contract) have made no claims about P-1's ability to accurately classify heterogeneous lands. Extension of P-1 into forested areas may not yield results as accurate as those specified for LACIE.

CHAPTER 3 - MATERIALS

Study Area

The study areas are located in the eastern half of the 844,000 hectare (2.086,484 acre) San Juan National Forest in southwestern Colorado (see Figure 3.1). The two study areas were chosen on the basis of

1. heterogeneity of cover types and topographic variation,
2. presence of locatable features, and
3. availability of reference data.

The San Juan Mountains are a rugged mixture of sedimentary rocks and tertiary volcanics which supported many silver mines near the turn of the century. Most mines have been abandoned and many mountain sides are pockmarked with sterile mine tailings and disintegrating buildings. Elevations in the San Juans range from approximately 2100 meters (about 7,000 feet) to over 4270 meters (14,000 feet) along the Continental Divide. Most of the areas below treeline (approximately 3350 meters) are heavily forested, with species distribution dependent upon elevation and aspect of the locale (see Fleming and Hoffer, 1977). Grasslands are prevalent in valley bottoms, and are often used for pasture. Extensive fields of hardy grasses and wildflowers dominate much of the tundra (above treeline). Mans activities manifest themselves in man-made reserviors, clearcuts, hay farming or pasturelands at the lower elevations, and four wheel drive trails crisscrossing forest and tundra.

The first study site is located 29 km northeast of Durango and encompasses approximately 4678 hectares (11,560 acres), including the northern third of the Vallecito Reservior. Elevations on this 100 by 100 pixel block range from 2332 meters at the reservior to 3293 meters in the northwest corner of the study area. Ponderosa pine is found predominantly on the lower hillsides surrounding the lake. As elevation increases, ponderosa gives way to the Douglas fir/white fir cover type.

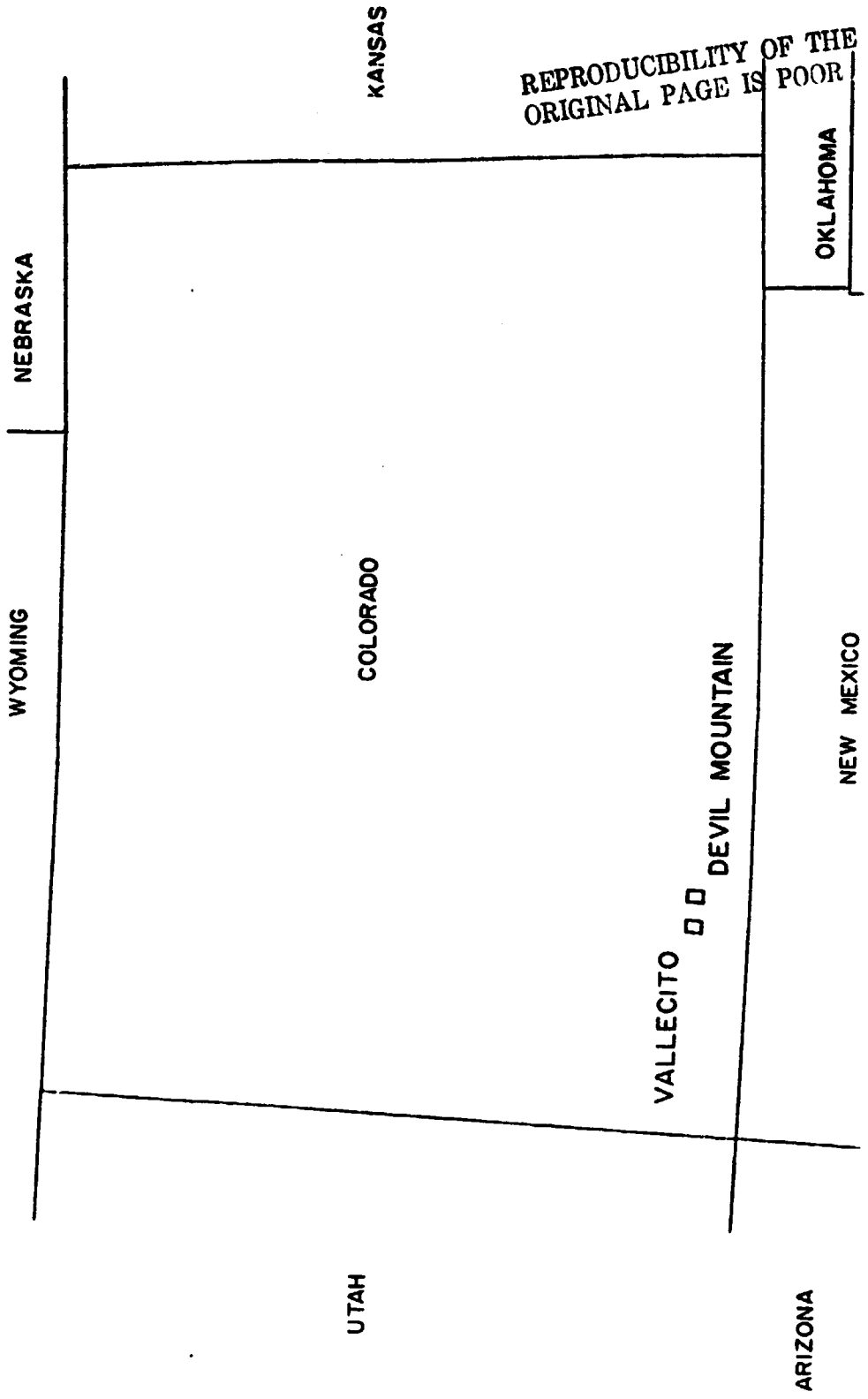


Figure 3.1 Location of the two study sites used to evaluate the Procedure 1 and Multicluster Blocks approaches.

At still higher elevations species composition graduates into the Englemann spruce/subalpine fir cover type at approximately 2740 to 2900 meters (9000 to 9500 feet), depending on the aspect of the site. Quaking aspen is ubiquitous, forming mixed or pure stands wherever cutting has occured or canopies have opened. Gambel oak may also be found in pure, small, shrubby stands at the lower elevations, generally below 2600 meters on the south facing slopes (Miller and Choates, 1964).

The second test site, the Devil Mountain Quadrangle,is approximately 48 km (30 miles) east of Durango and covers 15,156 hectares (37,450 acres). The Piedra River runs north-south roughly bisecting the quadrangle. The lowest elevation of the quadrangle lies in this river channel at the point where the river leaves the quad - 2012 meters (6600 feet). The highest elevation is atop Devil Mountain - 3036 meters (9957 feet). Like the Vallecito quadrangle, the area is extremely heterogeneous, with ponderosa pine flanking the lower elevations of the Piedra River valley. The ponderosa pine may be dense enough to generate a parklike grasslands understory, but in less dense stands is often found in conjunction with Gambel oak. As elevation increases or as aspect changes from south to north, the pine is found in mixture with Douglas fir, white fir concomitant. Aspen is found with increasing frequency above 2200 meters, and forms mixtures with the Douglas fir, white fir, and ponderosa pine. Englemann spruce and subalpine fir are found at the higher elevations in conjunction with Douglas fir, white fir, and aspen. Aspen forms large, pure stands between 2750 - 3050 meters (9,000 to 10,000 feet), perhaps as a result of small lightning fires or harvesting activities. Clearcuts are noticeable, especially in ponderosa pine stands near the Piedra River.

Supporting Information

Selection of the study sites were based not only on the physical attributes of the countryside, but also on the availability of ancillary information. The San Juan National Forest and surrounding federal lands have been the subject of a number of remote sensing research endeavors carried out by LARS and INSTAAR personnel. Hence, much detailed information was available which aided in the analysis procedure.

Type maps drawn by the Institute for Artic and Alpine Research, University of Colorado provided a baseline classification of the two study areas. The species type lines have been drawn atop an acetate reproduction of 1:24,000 U.S.G.S. $7\frac{1}{2}$ minute quadrangle maps. The type lines were the result of photointerpretation efforts at INSTAAR using color infrared photographs. Various areas were field checked to insure accurate species identification. These type maps were laid atop 1:24,000 lineprinter output to aid in the identification of clusters or to check classification results.

Also available on the two study sites were 1:120,000 color infrared photographs taken by NASA in 1973 in conjunction with the Landsat 1 study. Cluster classes were identified using these photos, the type maps, a Zoom Transfer Scope, and 1:24,000 scale cluster output.

The Landsat data used in the analysis was obtained on June 5,1973 at 9:20 am local time. Though some cloud buildup was noted on the CIR photos west of the Vallecito study area and north of Devil Mountain Quadrangle, no clouds were present on either study area on the Landsat data.

The Computing System

The entire analysis was done on LARS' IBM 370/148 medium speed computer operated by the Laboratory for Applications of Remote Sensing, Purdue University. The central processing unit has a number of peripheral devices including tape and disk drives, tape and disk storage, remote site terminal access, lineprinters, cardpunches, card readers, and a matrix printer. The computer is located in the Flexlab II building, Purdue Industrial Research Park, West Lafayette, Indiana. However remote terminal sites include (in addition to LARS terminals in Flexlab I) Goddard Space Flight Center in Greenbelt, Maryland; Johnson Space Center in Houston, Texas; Indiana State University in Terre Haute, Indiana; Wallops Flight Center, Wallops Island, Virginia; St. Regis Paper Company, Jacksonville, Florida; and Alabama A and M University, Huntsville, Alabama.

Software - The Programming Aspect

Two software systems were used in this study. Both are special purpose program sets developed to aid applications-oriented scientists

in the analysis of multispectral data taken by a wide variety of spectral sensors. The software packages are examined below.

LARSYS - Purdue's MSS Data Analysis System

The idea behind the LARSYS software system germinated in the mid 1960's. Since that time hundreds of people have helped to develop and document this software package "specifically designed for remote sensing data analysis" (Spencer and Phillips, 1973). This program set developed in conjunction with technical support from IBM and grants from NASA includes 18 fully documented processors.

Five LARSYS processors were used in this study. The first three, CLUSTER, MERGESTATISTICS, and SEPARABILITY, were used to develop training statistics using a Multicluster Blocks approach. CLASSIFY and PRINTRESULTS were responsible for classifying the area and providing accuracy estimates. The processors are described below; for more rigorous explainations, see Spencer and Phillips, 1973.

CLUSTER: After the analyst has selected the training areas, each must be clustered. Clustering partitions the data into smaller, simpler spectral groups. The purpose of clustering is to form "groups ... of data that have elements similar to one another within the cluster and different from the elements of the other clusters." (Kan, 1972).

The LARSYS CLUSTER processor groups the data vectors by first arbitrarily selecting initial cluster means and then assigning each pixel to that mean closest to it (minimum euclidean distance criterion). Once all pixels are assigned, cluster means are recalculated and the assignment is repeated. The algorithm stops when a prespecified percentage of the pixels (CONVERGENCE) are stabilized (i.e., are reassigned to the same cluster in two consecutive iterations).

The user may adjust five parameters which affect the clustering results. Other control options may be used, but they control input and output material. The following controls are of interest:

CONV XX.X Convergence specifies the percentage of pixels that must maintain cluster allegence between successive iterations in order for processing to cease (default 100.0).

MAXCLAS X This parameter specifies the number of clusters desired (default 5).

INTV X The Interval card specifies what portion of the points in a cluster are used to calculate the cluster mean. An interval of 1 indicates that all pixels in the scene are assigned to the closest cluster mean and are then used to compute the cluster mean after each iteration. Interval 2 means every other pixel is used to compute the mean. Interval 3, every third pixel is used. Once the convergence parameter is satisfied, all pixels are assigned to a cluster and the final cluster means calculated (default 1).
THRESH XX.X The threshold control card specifies a separability quotient value below which a suggestion is made to group the clusters.

The cluster map formed using these and other parameters is compared to photographs and the spectral classes, if acceptable, are identified.

The McB procedure produces as many statistics decks as there are training blocks. The classification processor uses a single statistics deck to classify an area into various informational classes. The following processor allows the analyst to formulate the final statistics deck from the training block decks containing the identified spectral class statistics.

MERGESTATISTICS: The MERGE processor combines statistics from all training blocks to provide the single deck required by the classifier. In addition, similar spectral classes (within one or more blocks) may be combined, useless or confusion classes may be deleted. The resulting statistics deck may be used to classify test areas and, if results prove unsatisfactory, the original block statistics decks may be remerged differently.

SEPARABILITY: In order to evaluate the consequences of any manipulation performed using the MERGE processor, the statistical separability of the spectral classes are calculated. The separability of the classes is determined using a distance criterion called the transformed divergence. As the transformed divergence values increase, the probability of correct classification increases, assuming the spectral classes are correctly labelled.

CLASSIFYPOINTS: This LARSYS processor is a Gaussian maximum-likelihood (ML) classifier which assigns each pixel to an informational class

defined by the analyst in the training statistics deck. Using a pixel's spectral values obtained from the data tape, the processor calculates the probability that the data vector belongs to each of the spectral-informational classes. CLASSIFYPOINTS assigns the pixel to the class with the highest probability. The classifier assumes the spectral classes used in the training statistics deck are unimodal and exhibit a normal distribution. The classification results, written to disk or tape, can be evaluated using PRINTRESULTS.

PRINTRESULTS: The final step in any classification procedure involves output of the final classification products and an evaluation of the results. If desired, an alphanumeric 1:24,000 scale lineprinter map may be output. Of particular value are classification performance tables concerning training and/or test fields delineated by the analyst. Acreages of each informational class may be tabulated using another processor (GLPRINT) if the total acreage of the scene is known.

The experimental methods which utilize the programs outlined above are explained in the Procedures section.

EODLARSYS - Johnson Space Center's Software

In 1970, LARSYS Version 2 was implemented on a UNIVAC 1108 EXEC 2 System as a batch machine at NASA/Johnson Space Flight Center in Houston. Since 1970, personnel at the Earth Observations Division have independently modified the original LARSYS package into a new, nonetheless relate, software system. Similarities are noted between a number of processors in LARSYS and EODLARYSS, though some EODLARSYS processors have been created to perform new functions.

The five EODLARSYS processors of interest to this study were developed in an attempt to more fully automate the classification process. The first three, DOTDATA, ISOCLS, and LABEL are responsible for developing a statistics file similar to the LARSYS statistics deck. The last two, CLASSIFY and DISPLAY, use the training statistics to classify the area and produce performance tables and 1:24,000 scale lineprinter maps.

These five processors are used in Procedure 1, a method designed to classify agricultural areas using information collected on randomly selected pixels (dots) in the scene. Point information is currently

available on some forestlands in the form of Continuous Forest Inventory (CFI) or sample points or plots which are locatable in the Landsat scene. The Procedure 1 processors can make direct use of this information to develop labelled training statiscics which may be used to classify an area.

The five EODLARSYS processors used in Procedure 1 are described below.

DOTDATA: The DOTDATA processor is (in this study) the first program run. It allows the analyst to input supervised training points whose identities are known. The computer retrieves the spectral reflectance values for each pixel input to DOTDATA, and compiles a numbered listing of the information. Each data vector (pixel) now has its own unique number which can be used to refer to the pixel in processors following it.

The dots defined to the DOTDATA processor may be divided into Type 1 and Type 2 dots. Type 1 dots may be used as cluster center starting dots in ISOCLS, and are used by the LABEL processor to identify the clusters formed in ISOCLS. Type 2 dots are used for bias correction, a classification accuracy calculation explained in Reeves (1978), or Wills, Gardner, and Aucoin (1977). Only Type 1 dots were used in this study.

ISOCLS: This versatile clustering function accepts as initial cluster centers

1. dots (pixels) from DOTDATA,
2. multispectral starting values input by the analyst, or
3. nothing - it may self-start.

The clustering processor outputs unlabelled (unidentified) spectral classes (cluster means and covariances) for the LABEL processor.

ISOCLS has 9 parameters which affect the number of clusters output:

NMIN	SEP	STDMAX
PMIN	ISTOP	CLUSTER
DLMIN	PERCENT	SEQUEN

Four of these were investigated in the parameter study, and these four are described below.

SEQUEN XXX: ISOCLS manipulates groups of pixels by first assigning all pixels to particular groups, then splitting or combining those groups

to form new groups (or clusters). The analyst controls the order of splitting and combining by using the SEQUEN control card (default SC, Split, Combine).

ISTOP X: The maximum number of iterations performed in the initial split sequence (default 10). For example:

SEQUEN SC
ISTOP 10) would yield 10 splits and one combine iterations.

SEQUEN SSC
ISTOP 10) would yield 11 splits and one combine.

SEQUEN SCSC
ISTOP 8) would yield 8 splits, a combine, a split, and a combine.

Whether or not a particular cluster is split or combined on a S or C iteration depends on the values of the following parameters.

STDMAX X.X: Any cluster with a standard deviation greater than X.X in any channel is split on a split iteration (default 4.5).

DLMIN X.X: Any clusters whose means are closer than X.X units are combined on a combine iteration (default 3.2).

Parameter studies by Moritz, Pore, Yao (1978), and Pore, Moritz, Register, and Yao (1978) using the ISOCLS processor have found that classifications of LACIE segments were more accurate when ISOCLS was not allowed to iterate. Hence, for agricultural land classification, the following parameter levels have been suggested: STDMAX 15, ISTOP 1, DLMIN 0 (LACIE parameters). The clusters obtained from a three pass grouping[1] are very much dependent on the initial cluster centers (input from DOTDATA). In the studies cited above, the initial centers were chosen randomly from a dotfile of a LACIE segment. The approach may prove disasterous in a forest scene because:

1. The means and variances of the spectral classes which characterize the area are dependent on the dots used to seed the processor.
2. The clustering function is not allowed to iterate. Hence adjustments in the cluster means which would ordinarily reduce

---

1. The ISOCLS parameters given above are those used in the LACIE study. ISOCLS is a three pass processor in this instance, with the second and third passes deleting any clusters which are too small (number of pixels). A one pass grouping function could be instituted if ISTOP is set equal to 0.

the sum of squares error (SSE) are not made.

However, the three pass clustering algorithm (LACIE parameters) merits investigation.

LABEL: Unlabelled cluster statistics from ISOCLS are compared to the Type 1 dots compiled in DOTDATA. The analyst may choose one of two procedures to identify the clusters:

1. K-Nearest Neighbor: This alternative finds the given number (K) of type 1 dots closest to the cluster mean, and labels that cluster using the dot identifications. Thus, if the analyst selected 5-Nearest Neighbor procedure, and the majority of the five closest dots were grass, the cluster would be labelled grass. 'Closeness' is determined by geometric (L1 and L2) distances in n space, where n equals the number[1] of channels used for clustering and dot labelling purposes. In the event of a tie, K minus one dots are considered, the dot farthest distant from the cluster mean is dropped from consideration.

2. All-of-a-Kind: Another method used to label clusters involves checking the identification of all Type 1 dots within the cluster. If all of the identified pixels in that cluster belong to the same category, then the cluster is labelled accordingly. If all the Type 1 dots within that cluster do not have the same identification, then the labelling procedure defaults to K-Nearest Neighbor.

Pore et al. (1978) found that results were consistently more accurate when All-of-a-Kind (instead of K-Nearest Neighbor) was used to label the agricultural statistics. Using 4, 8, 12, and 16 channels of multispectral information, accuracy increases ranged from 1.17 to 9.68 percent. In the course fo this study, I attempted to use the All-of-a-Kind labelling procedure, but the processor kept defaulting to K-Nearest Neighbor since clusters formed in a forested scene commonly hold Type 1 dots in more than one category. Therefore, the K-Nearest Neighbor

---

1. In other words, if using L1 distance to determine distance from a dot to a cluster mean - 4 channels of data, where:

   D_i = multispectral value of that pixel in channel i, i = 1 to 4.

   C_i = multispectral value of the cluster mean in channel i, i = 1 to 4.

   L1 distance = $|D_1-C_1| + |D_2-C_2| + |D_3-C_3| + |D_4-C_4|$

   L2 distance = $\sqrt{(D_1-C_1)^2 + (D_2-C_2)^2 + (D_3-C_3)^2 + (D_4-C_4)^2}$

labelling procedure was used throughout the course of this study to insure labelling consistency.

CLASSIFY: One of two procedures may be used to assign a pixel to one of the groups defined by the training statistics deck. The first procedure involves a standard, maximum-likelihood classifier. The probability that a given pixel belongs to a given subclass is calculated for each subclass. The pixel is assigned to the subclass which exhibits the highest probability.

The categories classifier (the second of the two procedures) computes a probability density function for each category defined for the statistics deck. A category is defined as a group of informationally related spectral subclasses. In other words, the Hardwood informational class may be represented by a number of spectral subclasses. That group of spectral subclasses is considered a category. The processor classifies a pixel into the category exhibiting the highest probability. The pixel is then classified into a subclass within the category using maximum-likelihood. The categories classifier is also called the Sum-of-Normal-Densities (SoND) classifier. The SoND classifier is invoked (in this study) using the CATEGORY FILE card. This card tells the processor that the classnames in the labelled statistics deck are to be used as categories. So, if the Hardwood class contains three subclasses (three clusters), the processor sums the density function for the three subclasses to produce the density function for the category Hardwood (see Figure 3.2). These category density functions are used to classify a given pixel into one of the categories. Once a pixel is assigned to a category, the probability that it belongs to each one of the subclasses within that category is computed, and the pixel is assigned to that subclass with the highest probability.

DISPLAY: Procedure 1 accuracy statements may be calculated using Type 1 dot classification accuracies or using test field accuracies. If the former is chosen, the processor evaluates and adjusts classification accuracy using Type 1 and 2 dots.

Test fields or points may be input, in which case Type 1 and 2 dot accuracy calculations are not determined. Throughout the course of the classification procedure, the analyst has the ability to outline

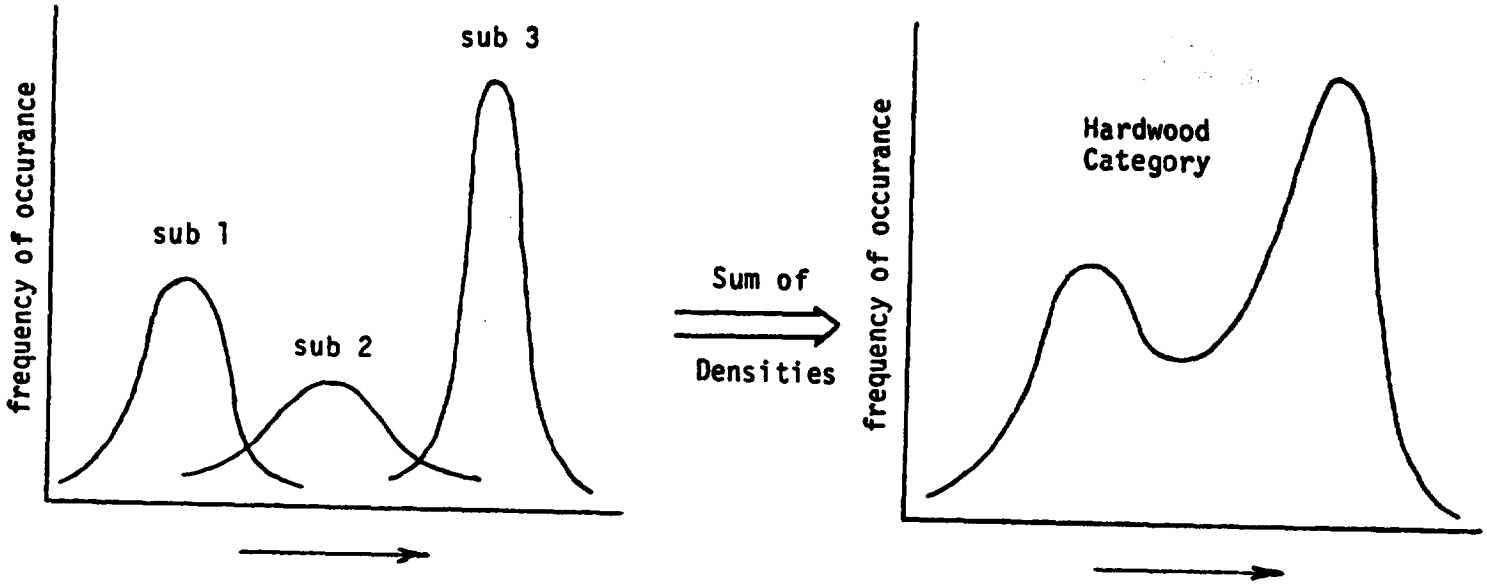


Figure 3.2 The effect of using a Sum-of-Normal-Densities classifier - formation of category density functions from subclass statistics.

areas that are of no interest (Designated Other) or areas that cannot be identified, such as those areas covered by clouds (Designated Unidentifiable). Test fields may only be used if DO (Designated Other) or DU (Designated Unidentifiable) fields are not specified. DO and DU areas were not used in this study thus allowing use of test fields to provide the accuracy considerations.

Of the 13 processors currently online in the EODLARSYS software system, only the five described above were used in this study. The descriptions given for each of these five processors are not complete; many options have not been discussed since they were not directly involved in the study. Stewart and Aucoin (1978) supply the most concise documentation on these processors as implemented on the LARS IBM 370/148. Additional information may be found in user documentation written by Minter, Wills, and Gardner, 1977, and Wills, Gardner, and Aucoin, 1977.

CHAPTER 4 - PROCEDURES

Introduction

In order to meet the major objective and sub-objectives outlined in the Introduction, the study was broken into two phases. The first phase consisted of the ISOCLS (EODLARSYS clustering function) parameter study, since little was known of the effects of the various parameters on the statistics output by ISOCLS. The second phase used information gathered in the first phase to establish processor parameters so that a valid comparison of McB and P-1 could be made. These phases are described in subsequent sections.

Procedures of Interest - the Multicluster Blocks Approach and Procedure 1

Before any experimental procedure is described, the reader should understand the methodologies involved in the two approaches being compared. The processors used in each of the approaches (McB and P-1) are described in the Materials section. The use of these processors to produce a classification is reviewed below.

The Multicluster Blocks Approach

The Multicluster Blocks procedure involves selecting relatively small, heterogeneous blocks (1600 to 3600 pixels) to train the computer. Four primary considerations in selecting the location of the training blocks are defined by Fleming and Hoffer, 1977. The blocks should have

1. "a representative sample of each informational class ..",
2. "three to five cover types ..",
3. "a precisely locatable feature ..", and
4. available photography.

Each block is initially clustered into approximately 14 to 16 spectral classes, though more or less may be desired depending on the spectral heterogeneity of the block. A Zoom Transfer Scope and the appropriate photography are used to identify the spectral classes.

The spectral classes output from each of the blocks must be pooled to combine the training statistics into a single deck. Fleming and Hoffer (1977) found that a three iteration, moderately modified pooling procedure yielded the most accurate results.[1] The first iteration pools those spectral classes with a divergence value of 500 or less.[2] The second and third iterations pool classes with divergence values less than or equal to 1000 and 1500 respectively.

The training statistics formed are used to classify a small portion of the study area, perhaps one of the training blocks. If the results are not satisfactory, and the original clusters are acceptable, then only the pooling need be redone, an inexpensive (though analyst intensive) procedure. Once acceptable training statistics are formulated, the entire area of interest is classified.

The Procedure 1 Approach

Procedure 1 uses the five processors described in the Materials section to classify the area of interest. The analyst has control over the processors' parameter levels and must input the locations of pixels of known identity. If the pixel identification information is available from some existing source, such as forest inventory data, analyst involvement is markedly reduced.

The dot information is compiled and may be used to seed the clustering processor. Regardless of the clustering method, the dots, whose identity and spectral reflectance values are known, are used to identify the spectral classes output by the clustering processor. The

---

1. An analyst may accept the judgement of the SEPARABILITY processor and pool those classes suggested by the processor (unsupervised pooling). Alternatively, the analyst may use his own discretion when pooling, basing his decisions on the identity of the spectral classes, the bispectral plot, and the divergence values between the classes (modified pooling).

2. Transformed divergence measures spectral class separability. A value of 0 means the two classes are identical, a value of 200 (the maximum) means the classes are extremely dissimilar, very separable.

advantage of such a system rests with the fact that the labelled statistics are free from any biases[1] except those inherent in the initial identification of the dots. The disadvantage stems from the fact that, unless the analyst intervenes, each spectral class produced is labelled and used in the final training statistics. Border spectral classes (i.e, hardwood-conifer mixes or grass-hardwood mixes) must be labelled either hardwood, conifer, or grass. Hence the labelling process may in some instances promote misclassification.

The labelled statistics are used to train the classification processor. Accuracy of classification may be assessed by comparing the classification's identity of the input dots to their original identity or by using test fields.

ISOCLS Parameter Study

The objective of this portion of the study was to develop a set of ISOCLS parameters which could be used to cluster a forested area accurately and economically. The 100 x 100 pixel Vallecito Reservoir area was chosen for this portion of the investigation. The criteria used to judge cluster performance were

1. number of clusters produced,
2. CPU time used, and
3. classification performance.

The third criterion unfortunately is influenced by how well the LABEL processor does its job. LABEL's performance in turn is influenced by the ability of the analyst to correctly identify the dots used in the dotfile. Finally, classification performance is influenced by how well test fields were selected and identified. In spite of the subjectivity of this performance criterion, accuracy of classification was used in conjunction with the other two criteria.

Type 1 dots and test fields were delineated on the study site. The dots were used to seed the ISOCLS processor and were also used by LABEL to identify the clusters formed in ISOCLS. These dots were located using a Systematic sample grid system that is used by the USFS Northcentral Forest Experiment Station to inventory their forests in

---

1. Biases may be introduced by analyst identification and manipulation of the training statistics.

Minnesota. The grid covers a township sized area and is composed of 121 dots, arranged in 11 rows of 11 dots. Each row is canted five degrees above the horizontal to avoid having entire rows or columns fall on a N-S or E-W road (see Figure 4.1). A 1:24,000 township sized grid was drawn to locate the Type 1 dots on a lineprinter cluster map of the study area. Sixty dots were located on the Vallecito study area using the grid; these dots simulate the information which might be obtained from Forest Service records. An additional 23 dots were located in cover type categories that had not been sufficiently represented by the systematic sample. These 83 points were identified using the Zoom Transfer Scope, a 1:24,000 lineprinter map of the Vallecito area, and 1:120,000 color infrared photos. Table 4.1 shows the composition of the Type 1 dots used.

Table 4.1 Type 1 dots used in the ISOCLS parameter study (Vallecito study area), by class.

	Located		
Classname	Systematically	by Cover Type	Total
Hardwood	17	0	17
Conifer	33	0	33
Grass	6	8	14
Barren	0	9	9
Water	4	6	10
Total	60	23	83

Manually selected test fields were also delineated on the 100 by 100 pixel area. The accuracy of classification of these test fields was one criterion used to judge how well a particular set of parameters performed in classifying the area.

The same dotfile (the same 83 points) and the same test fields were used throughout the parameter study.

The parameter study was broken into two parts. The first portion of the parameter study investigated the effects of the following ISOCLS parameters: PERCENT, STDMAX, DLMIN, ISTOP. Initially each of the parameters was studied by holding all other cluster parameters at their

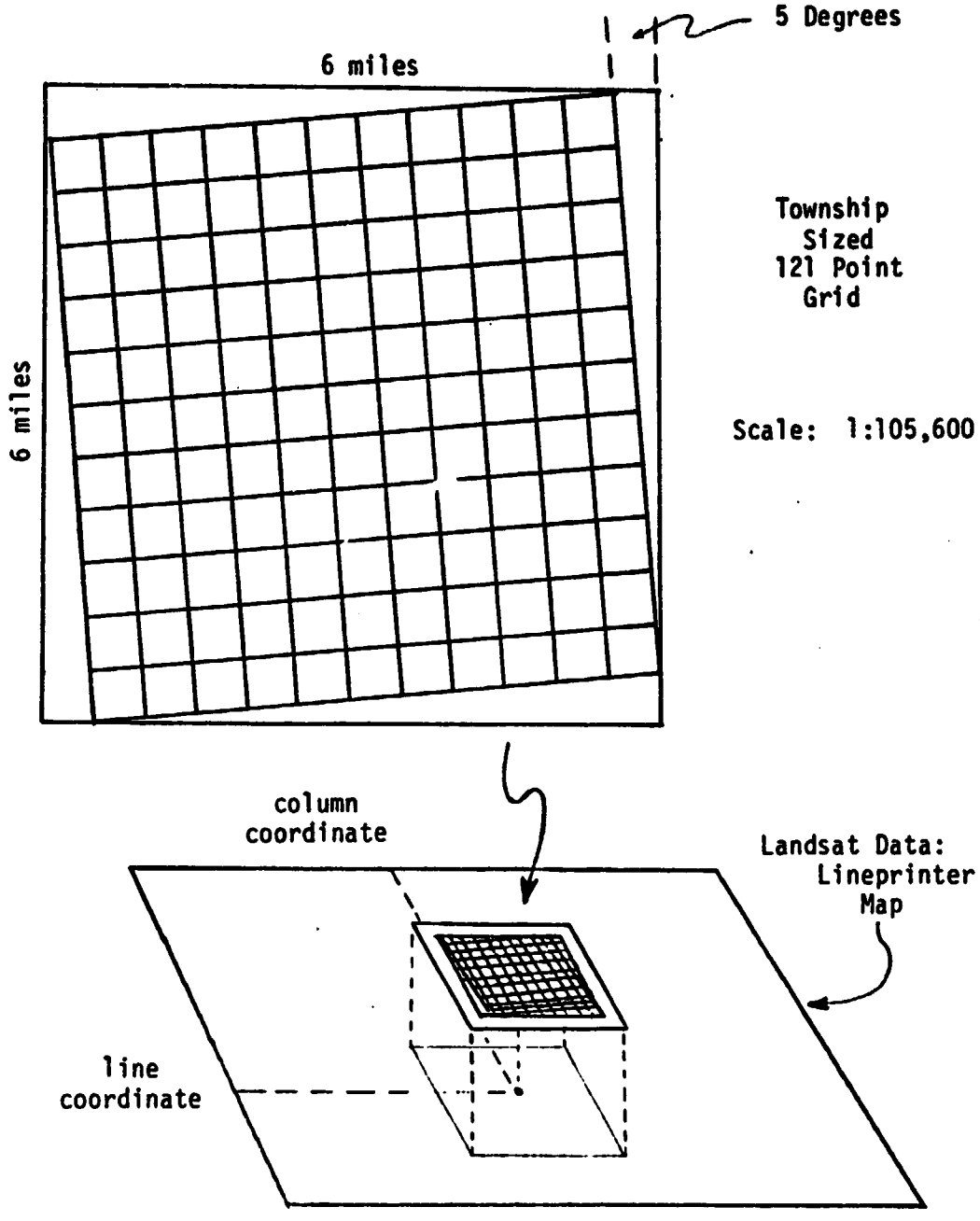


Figure 4.1 Example of 121 point grid used to locate Type 1 dots on the Vallecito and Devil Mountain Quadrangles.

default values and varying the value of the parameter in question. On the basis of the responses of the ISOCLS processor to various parameter manipulations, parameter combinations were choosen for evaluation.

The second part of the parameter study investigated the effects of self-starting and seeding ISOCLS with a random sample of Type 1 dots. From previous experience, it is known that:

1. The LACIE study effectively used a two pass clustering algorithm.
2. With all ISOCLS parameters set to LACIE specifications, the number of seed dots input equals the number of clusters output $\pm$ 2 clusters.[1]
3. These clusters often retain the same identity as the starting seed when identified by LABEL.

Hence the parameter study compared the effects of self-starting, LACIE's two pass approach, and a seeded, iterative approach.

The parameter study makes no pretenses at completeness, rather takes the attitude that:

1. a little knowledge might greatly improve classification accuracies and save computer time.
2. other parameter studies might build on the information compiled here.

The results of this work are presented in the Results section of this report.

Classification Study

Establish Training and Testing Blocks, Fields, and Points

The overall objective of this part of the study was to compare the best methods found to date for classifying forested areas using two different training techniques, the Procedure 1 processors (DOTDATA, ISOCLS, and LABEL) and the Multicluster Blocks approach.

Three items were necessary to proceed with the comparisons: Type 1 dots, training blocks, and test fields.

The Type 1 dots were selected using the 121 point township grid laid atop an unsupervised classification (16 classes) of the Devil Mountain Quad. These dots were identified using aerial photography and the Zoom

---

1. The number of clusters produced will always be less than or equal to the number of Type 1 dots input. Generally as the number of seeds rises, variability increases.

Transfer Scope. Additional Type 1 dots were chosen in a cover type only if that type was not sufficently represented by points selected using the grid. In other words, if each of the five cover types did not contain at least ten Type 1 dots,[1] supplimentary dots were arbitrarily located in the inadequately represented cover type(s). Sufficnet dot representation in each cover type is manditory due to the nature of Procedure 1's cluster identification. If, for instance, the K-Nearest Neighbor technique is used to label the clusters, with K set equal to 5, and only two dots have been identified as barren in the dotfile, most likely no cluster will be labelled barren.[2] Since the barren class would not be represented in the labelled training statisitcs, all barren areas in the data would be automatically misclassified. The decrease in classification accuracy would not be a function of ISOCLS or LABEL as much as it would be the fault of the method used to select the Type 1 dots. In order to overcome this problem, each of the five major cover types had at least ten Type 1 dots.

Three training blocks were established on the Devil Mountain Quadrangle. The blocks, ranging in size from 841 pixels to 1804 pixles, were established in heterogeneous areas according to the criteria listed on page 37of this report.

Manually selected test fields were located in homogeneous areas again using the same materials that were used to select the training blocks. The same test fields were used to evaluate each of the classification methods tested (see Table 4.2). The test fields contained no Type 1 dots and were not located in any of the training blocks.

The study site contained 33,123 pixels (183 lines x 181 columns). ISOCLS, using four channels of data data, can accommodate 91,728 pixels. Procedure 1 demands that the entire area be clustered in order to

---

1. The designation of 10 Type 1 dots as the allowable minimum number of dots is arbitrary and peculiar to this study area. This analyst cited ten as a minimum because no more than that could have been reliably located in the barren class.

2. It is actually possible for a cluster to be labelled barren under these circumstances, but the chances of correctly identifying the barren class are small.

develop training statistics. In order to reduce CPU time and remain within the dimensional restrictions of CLUSTER, the quadrangle was clustered on an interval of two.[1]

Table 4.2 Test fields used on the Devil Mountain Quadrangle

Cover Type	Test Fields		Test Field Pixels	
	Number	Percent of Total	Number	Percent of Total
Conifer	39	41.5	932	55.3
Hardwood	33	35.1	457	27.1
Grass	16	17.0	252	14.9
Barren	6	6.4	46	2.7
	94	100.0	1687	100.0

Classification Procedures

A classification processor uses training statistics which enable the classifier to group pixels into the informational classes defined by the analyst. In this study, six different runs were made employing various processor combinations and different parameter settings. Four of the runs used the Procedure 1 approach to developing training statistics (i.e., form a dotfile, cluster the entire area, use the dots to label the clusters); two used the McB approach (i.e., cluster each training block separately, identify and merge the spectral classes). Remember that the P-1 approach does not necessarily mean that only EODLARSYS processors were used; clustering processors were interchanged in some of the runs. Table 4.3 characterizes the six runs made. Each of the six runs were classified using both

a. the EODLARSYS Sum-of-Normal-Densities classifier (CLASSIFY using CATEGORY FILE control card); and

b. the LARSYS standard maximum likelihood classifier (CLASSIFYPOINTS).

Thus, a total of twelve classifications were generated and compared.

---

1. Using 4-channel Landsat data, CLUSTER can handle approximately 10,000 pixels (if N = number of pixels processed, then $N < 40{,}000/n < 25{,}000$, where n = number of channels, see Spencer and Phillips, 1973).

Table 4.3 Development of training statistics using the Multicluster Block and Procedure 1 approaches (clustering processors interchangeable).

Method of Developing Training Statistics	Cluster Processor Used: ISOCLS, Unseeded, 10 Iterations[1]	ISOCLS, Seeded, 3 Iterations	ISOCLS, Seeded, 1 Iteration (LACIE)	CLUSTER[4]
Procedure 1[2]	Run 1	Run 6	Run 5	Run 2
Multicluster Block[3]	Run 3			Run 4

The purpose of making the six runs, and using both classifiers on each run was fourfold:

1. The results of all the runs allow us to draw conclusions on which is the better method of developing training statistics.
2. By comparing Run 1 vs Run 2, and Run 3 vs Run 4, we may draw conclusions on which of the clustering functions - ISOCLS or CLUSTER - is more efficient.
3. The results of Run 1 vs Run 5 vs Run 6 allow us to judge the effects of seeding on CPU time used and accuracy of classification.
4. Within any given run, the Sum-of-Normal-Densities classifier and the standard maximum-likelihood classifier may be compared.

Each run is described in detail below. For a given run, the training statistics processors remain constant; the letter following the number indicates which classifier was used (a. the EODLARSYS Sum-of-Normal-Densities classifier, b. the LARSYS maximum likelihood classifier). Note that the letter 'R' indicates a statistics deck reformatting step.[5]

---

1. The number of iterations refers to the number of split iterations.
2. Processors used to develop training statistics (cluster function variable): DOTDATA-$^{\text{Clustering}}_{\text{function}}$-LABEL.
3. Processors used to develop training statistics (cluster function variable): $^{\text{Clustering}}_{\text{function}}$-MERGESTATISTICS-SEPARABILITY.
4. Cannot be seeded, and is iterative.
5. Two programs written by Carol Jobusch, LARS statistician, were used to make the LARSYS-EODLARSYS statistics deck conversions.

Run 1: Order of Processors:

DOTDATA-ISOCLS-LABEL-{a. CLASSIFY-DISPLAY (SoND)
b. R-CLASSIFYPOINTS-PRINTRESULTS (ML)

The first run was a modified version of Procedure 1. All processors used to develop the training statistics were those used in Procedure 1, however ISOCLS was not seeded. The entire quad was clustered in order to produce the training statistics. Clustering was done on an interval of two, and approximately 20 to 25 spectral classes were desired. Since the analyst rarely knows exactly how many spectral classes will be produced by ISOCLS for a given set of parameters, runs using ISOCLS (Runs 1 and 3) were done first. The number of spectral classes produced by ISOCLS equaled the number of classes requested of CLUSTER in Run 2 (to aid with clustering processor comparisons).

Run 2: Order of Processors:

DOTDATA-CLUSTER-R-LABEL-{a. CLASSIFY-DISPLAY (SoND)
b. R-CLASSIFYPOINTS-PRINTRESULTS (ML)

This run used LABEL to identify the spectral classes formed by the LARSYS clustering function. The unlabelled statistics produced by CLUSTER (MAXCLAS = 20, INTV = 1, CONV = 98.5) were reformatted for use by LABEL, CLASSIFY, and DISPLAY. The labelled statistics output by LABEL were reformatted prior to processing by LARSYS, CLASSIFYPOINTS and PRINTRESULTS (Run 2b).

The same dotfile was used by the LABEL processor in Runs 1, 2, 5, and 6. Runs 1 and 2 used the dotfile only for labelling the statistics output by ISOCLS and CLUSTER respectively. Runs 5 and 6 used the dotfile to seed ISOCLS and to label the clusters produced.

Run 3: Order of Processors:

ISOCLS-R-MERGESTATISTICS-SEPARABILITY-{a. R-CLASSIFY-DISPLAY (SoND)
b. CLASSIFYPOINTS-PRINTRESULTS (ML)

The third run used a Multicluster Block approach to developing training statistics on the Devil Mountain Quad, but used the EODLARSYS clustering function ISOCLS to develop the stat deck. Again the ISOCLS function was used first since the analyst did not have complete control over the number of clusters output.

Each training block selected was clustered individually; 14-16 spectral classes were formed in a block. These spectral classes were

identified by the analyst using a 1:24,000 lineprinter map of each training block, aerial photography, and a Zoom Transfer Scope. All spectral classes formed were manipulated using the LARSYS MERGESTATISTICS and SEPARABILITY processors to produce the final training statistics deck. This stat deck was used to classify the area.

Run 4: Order of Processors:

CLUSTER-MERGESTATISTICS-SEPARABILITY-(a. R-CLASSIFY-DISPLAY (SoND)
b. CLASSIFYPOINTS-PRINTRESULTS (ML)

Run 4 was a standard Multicluster Block approach to developing training statistics, utilizing LARSYS processors throughout. The training blocks used to develop the training statistics in Run 3 were used in this run, and the same number of spectral classes requested for each block as were produced in Run 3. In other words, if ISOCLS clustered the first and second training blocks (Run 3) into 14 and 16 spectral classes respectively, then CLUSTER's MAXCLAS parameter was set at 14 and 16 for these two blocks in Run 4 (CONV = 98.5, INTV = 1).

Runs 1 and 2 (using EODLARSYS software, different clustering processors) had equal numbers of spectral classes, Runs 3 and 4 had equal numbers of spectral classes (or very nearly so). However, the number of spectral classes formed in Runs 1 and 2 did not necessarily equal or even approximate the number formed in Runs 3 and 4. This should not be of any great surprise since one of the advantages of the Multicluster Block method of developing training statistics is its ability to produce large numbers of spectral classes with a relatively small amount of CPU time.

Run 5: Order of Processors: Same as Run 1.

The fifth run used a LACIE classification approach to developing the training statistics for the two classifiers. Unlike the previous runs, ISOCLS was seeded, and was allowed to iterate only twice (one split, one combine). Twenty-three dotfile seeds were used in an attempt to produce 20 (or more) clusters. Various numbers of each type of dot (Hardwood, Conifer, etc.) were input according to

1. approximate photointerpreted areal percentage of cover types, based on a quick scan of the 1:120,000 CIR photos,
2. number of cover type dots in the Type 1 dot data set (i.e., percent of area covered by hardwoods is proportional to the number of hardwood dots/total number of dots), and
3. expected variability within a class (intuitive).

The spectral classes were labelled using L1 distance and 1-Nearest Neighbor.

Run 6: Order of Processors: Same as Run 1.

This run was most similar to Run 5 in that ISOCLS was seeded with Type 1 dots. The only difference between Runs 5 and 6 was that Run 6 allowed ISOCLS to iterate (3 split iterations) and Run 6 used L2 distance and 10-Nearest Neighbor to label the spectral classes (all runs except 5 used this labelling procedure).

The results of these classifications were analyzed using a Newman-Keuls Range Test; the procedure involved steps outlined on page 2.7-11 of Landgrebe (1976). The Maximum-Likelihood and Sum-of-Normal-Densities classifiers were compared using paired-t tests as outlined in Mendenhall (1975). The results of these classifications, and conclusions drawn from this information are given in the subsequent sections.

CHAPTER 5 - RESULTS

Parameter Study

More than 50 separate classifications were done en route to completing this portion of the study, although more work would have to be done in order to fully quantify the effects of the various parameters. The results indicate trends and may provide useful rules of thumb. The reader should realize that replicates were not run. The results shown below are most certainly influenced by characteristics of the data. The shape of the curves may be quite different depending on the dots used to seed ISOCLS, the values of the other clustering parameters, and the spectral characteristics of the study area.

ISOCLS Parameter Study

The Effects of Four Parameters: STDMAX, PERC, ISTOP, DLMIN

The EODLARSYS clustering processor is a complex program which allows the analyst to control the statistical 'shape' of the clusters formed by controlling standard deviations in each channel. Although the program has no single parameter which regulates the number of clusters resulting, the number may be managed by changing the values of certain ISOCLS parameters. The purpose of this study was to quantitatively define the effects of STDMAX, PERCENT, ISTOP, and DLMIN on

1. the number of clusters formed,
2. the computer time used, and
3. in some instances, the accuracy of classification.

The effects of a given parameter on the performance of the clustering processor is often dependent on the values of the other parameters. For instance, changing PERCENT from 80 to 100 may have no effect when all other parameters are set at their default values, but the same change may significantly effect the number of clusters formed (and CPU time used) when the values of STDMAX and DLMIN are reduced. Other parameters are listed only when their values were not the default. Following is a

list of the default values for the parameters mentioned in this study.

STDMAX	4.5	PERCENT	80	ISTOP	10
SEQUEN	SC	DLMIN	3.2	Number of Seed Dots	0
SEP	separates new cluster means by plus-minus one standard deviation in the channel(s) which do not meet the STDMAX criterion.				

In addition to investigating the effects of various parameters, the effect of seeding was considered. In the course of the parameter study, ISOCLS was allowed to self-start, and in other runs was seeded with a subset of the Type 1 dots (from the dotfile). The results of the parameter and seeding study are given below.

STDMAX: ISOCLS splits any cluster which exhibits standard deviation(s) (within a channel) greater than the STDMAX value. Hence as STDMAX decreases, more clusters are split, and CPU time increases.[1] The results of clustering proved to be quite sensitive to changes in the STDMAX parameter. STDMAX had the greatest (and most predictable) effect on the number of clusters formed (see Figure 5.1).

PERCENT: ISOCLS ceases processing

1. when all split and combine sequences are exhausted (as specified by SEQUEN), or
2. when the percentage of stabilized clusters exceeds the PERCENT number.

A stabilized cluster is one whose standard deviations in all channels considered are less than STDMAX. Hence if at the beginning of an iteration 20 clusters have been formed, and either 18, 19, or 20 are stable, then processing will cease if PERCENT is less than or equal to 90.

The PERCENT parameter offers only gross control over CPU time and the number of clusters formed. With all other parameters at default, there was little difference between clusters formed at PERC=80 and PERC=90. Values of 90 and 100 produced identical output.

Using a different set of parameters and seeding ISOCLS with seven Type 1 dots produced the results seen in Figure 5.2. Though at best preliminary, these results indicate PERC=90 may save CPU time without

---

1. CPU times given in this parameter study section refer to the amount of computer time needed to cluster a 100 x 100 pixel forested area (Vallecito Study Area).

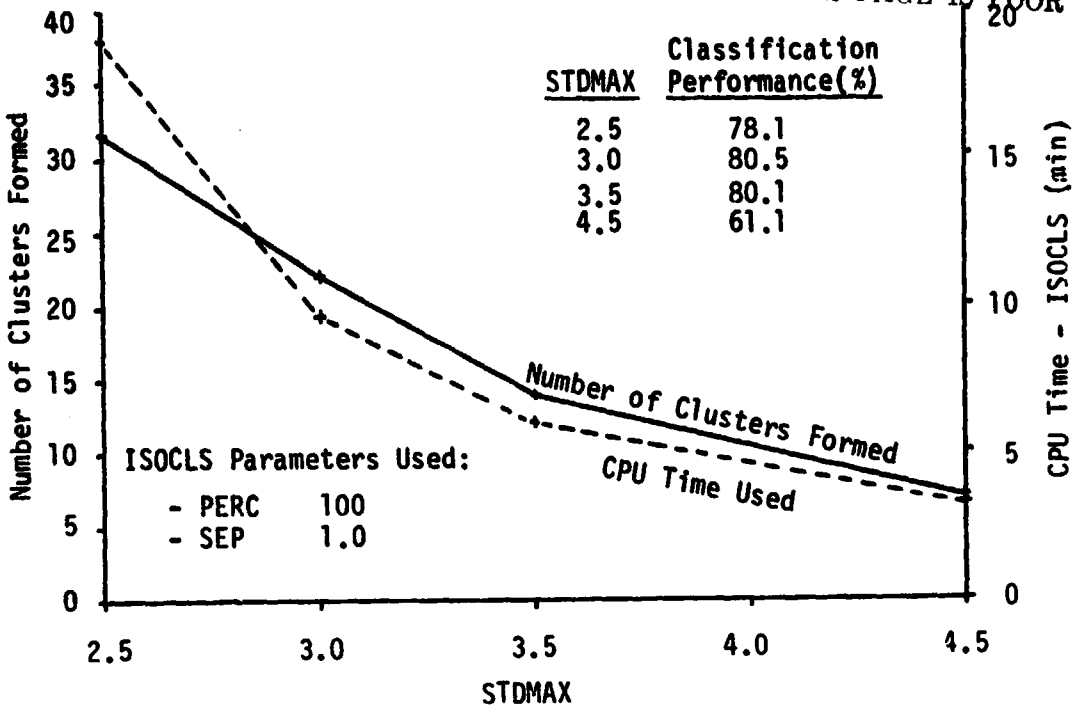


Figure 5.1 Effects of STDMAX on ISOCLS performance. Empirical relationship between STDMAX, number of clusters formed, and CPU time used (Vallecito study area).

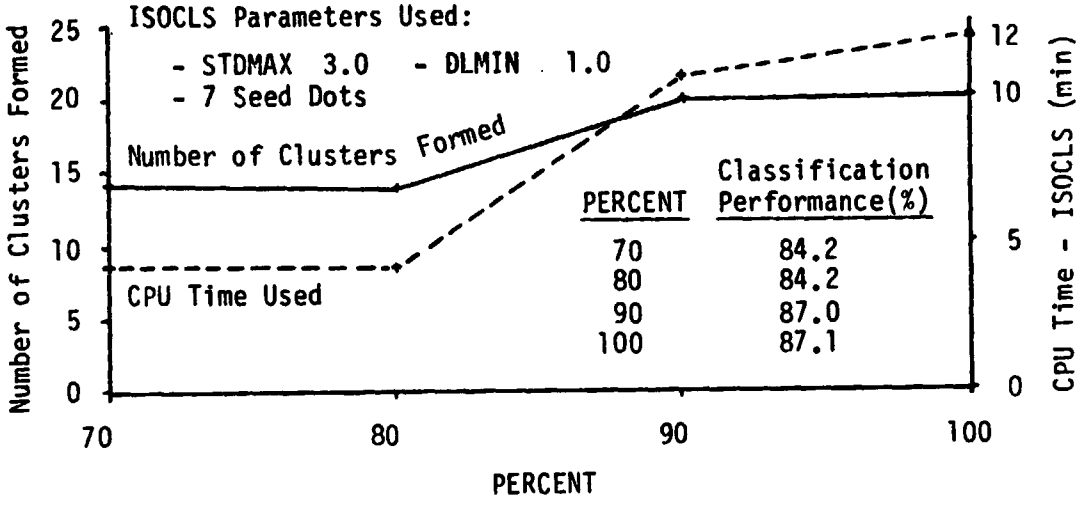


Figure 5.2 Effects of PERCENT on ISOCLS performance. Empirical relationship between PERCENT, number of clusters formed, and CPU time used (Vallecito study area).

a large effect on the number of clusters formed.

ISTOP: The analyst may designate the order of split and combine operations. ISTOP informs the processor how many iterations are done in the initial split sequence (indicated by the first 'S' on the SEQUEN control card).

As a preliminary step, ISTOP was changed from 5 to 10, 15, and 20 with all other parameters set at their default values (exception, SEP = 1.0). Evidently only 5 (or less) split iterations were necessary to stabilize 80% (default value for PERCENT) of the clusters formed because there was essentially no change in CPU time used, and no change in the number of clusters produced.

As DLMIN is reduced, one would expect

1. the number of clusters formed to increase, and
2. average intercluster distance to decrease.

The results indicate that DLMIN values below the default do not affect the number of clusters produced and do not significantly affect the CPU time used. Hence the DLMIN default value was used in the subsequent classification study on the Devil Mountain quad. The effect of DLMIN may become more noticeable when the number of initial split iterations is reduced. DLMIN is a divergence value. The average divergence value for all clusters formed in an iteration should increase as the number of iterations increase. Hence as ISTOP is increased, the effects of DLMIN should be reduced (i.e., fewer clusters should be combined for a given DLMIN value).

In summary, as a result of this parameter study, Table 5.1 shows the parameter values defined for use with the ISOCLS processor in the classification study.

The ISOCLS parameters developed on the Vallecito Study area produced less than the desired number of spectral classes on the Devil Mountain quadrangle. Parameters were adjusted (specifically, STDMAX and DLMIN were decreased) to produce 20-25 spectral classes (Run 1) on the entire quadrangle and 14-16 spectral classes on each training block (Run 3). The final parameters used are noted in subsequent sections.

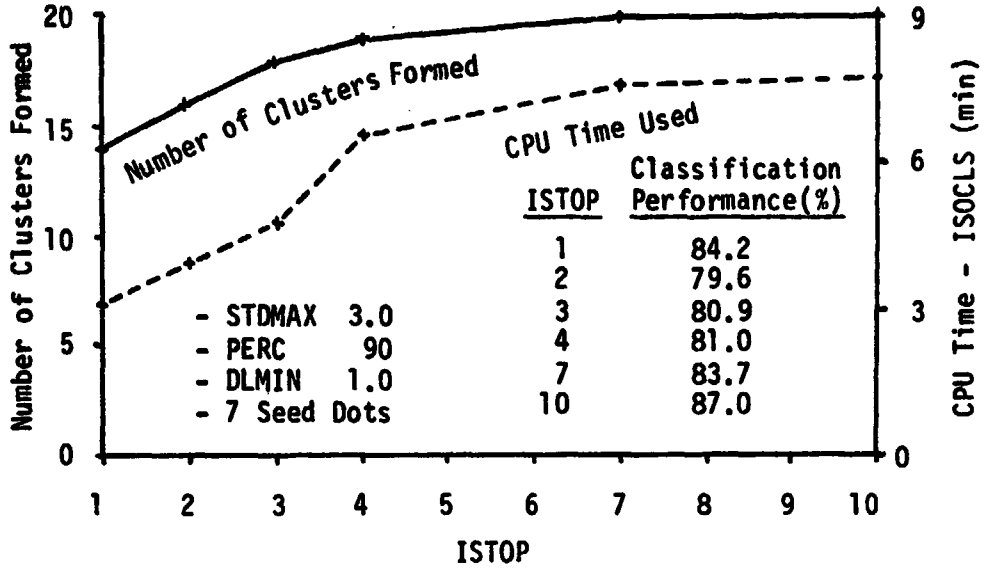


Figure 5.3 Effects of ISTOP on ISOCLS performance. Empirical relationship between ISTOP, number of clusters formed, and CPU time used (Vallecito study area).

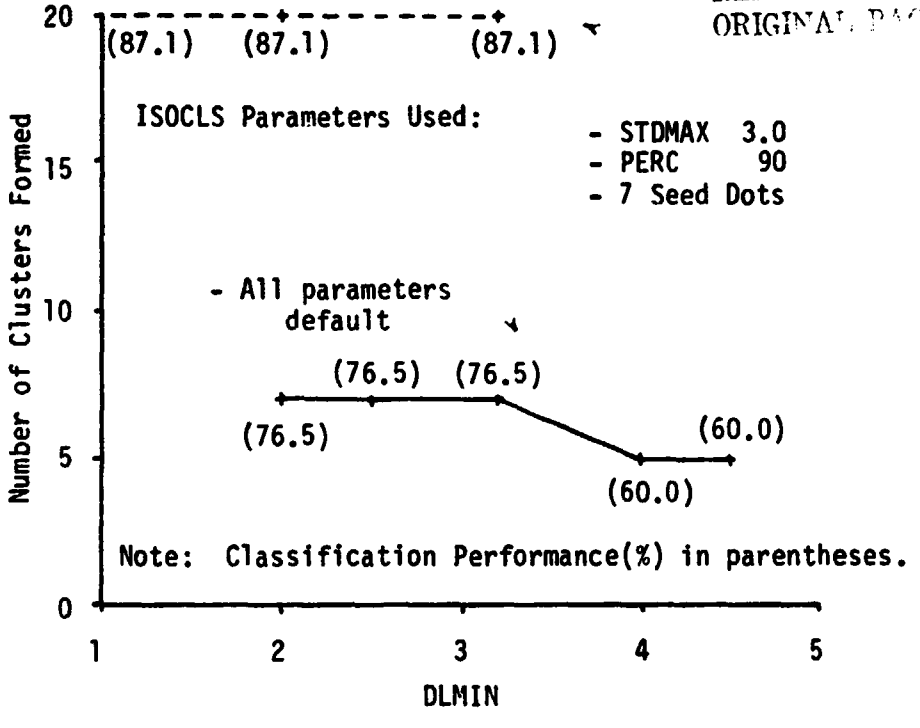


Figure 5.4 Effects of DLMIN on ISOCLS performance. Empirical relationship between DLMIN and number of clusters formed using two ISOCLS parameter sets (Vallecito study area).

Table 5.1 Initial Devil Mountain classification study parameters for ISOCLS, based on the results of the Vallecito ISOCLS parameter study.

Parameter	Run 1	Run 3	Run 5	Run 6
STDMAX	2.6	3.25	15.0	2.8
PERCENT	90	90	80	90
ISTOP	10	10	1	3
DLMIN	3.2	3.2	0	3.2
Dots Used (seeds)	0	0	23	23
SEP	default	default	1.0	default
CLUS	25	16	60	25

The Effects of Seeding ISOCLS

The second phase of the parameter study involved an investigation into the effects of seeding ISOCLS. Three methods were compared

1. Unseeded - 10 split iterations,
2. Seeded - 1 split iteration, and
3. Seeded - 10 split iterations.

Unseeded - 10 Split Iterations: The ISOCLS processor clustered the study area four times. Each time STDMAX was decreased to study the effects of the STDMAX parameter (see STDMAX section, pg 50). The number of clusters output were 7, 14, 22, and 32, accomplished by reducing STDMAX from 4.5 to 2.5. The following ISOCLS parameters were used: PERCENT 100, STDMAX variable, SEP 1.0, all other default.

ISOCLS is little more than a one-pass grouping processor when the parameter levels are set to LACIE specifications. The processor actually goes through the data three times. The first pass assigns each pixel to a seeded cluster mean using a minimum-distance criterion. The last two passes merely delete clusters which are too small (i.e., clusters which contain less than NMIN and PMIN pixels). No splitting or combining is done since the STDMAX parameter is very large (15.0) and the DLMIN parameter is very large (15.0) and the DLMIN parameter is set equal to zero. The other ISOCLS parameters used to implement this procedure are listed under Run 5 in Table 5.1.

Since the number of input dots very nearly equals the number of clusters formed using LACIE parameters, 7, 14, 22, and 32 dots were input to ISOCLS. Table 5.2 presents a listing of the breakdown of the dots used for seeding. These breakdowns were chosen on the basis of expected variability within an informational class. In other words, more variability is expected in the hardwood or conifer stands scattered over the study area than is expected in the water or barren categories, so more seeds were allotted to the forested categories. The actual seeds were chosen randomly; the hardwood seeds were selected at random from all Type 1 hardwood dots. Type 1 pixels were added to each category (hardwood, grass, etc.) in order to increase the number of seeds from 7 to 32.

Table 5.2 Type 1 dots used to seed ISOCLS in each of the five Level II cover types (ISOCLS parameter study-Vallecito study area).

Number of Seeds	Identity of Seeds (Type 1 dots)				
	Hardwood	Conifer	Grass	Barren	Water
7	2	2	1	1	1
14	4	4	3	2	1
22	7	7	4	3	1
32	10	10	6	4	2

Seeded-10 Split Iterations: Realizing that seeding may save CPU time,[1] the same dots (listed above) were input into the ISOCLS processor with the following parameters: PERCENT 100, STDMAX 3.0, all others default. This mode of operation combined the time saving potential of seeding with the benefits of iterative clustering.

These three methods are compared in Figures 5.5 and 5.6.

The results indicate that (on the Vallecito study area):

1. The LACIE parameters require the least CPU time to output a

---

1. If ISOCLS is not seeded, the first split iteration produces only 2 clusters, the second 4, the third 8 (at best), etc. Seeding establishes cluster centers immediately and removes the need for the initial self-start split iterations.

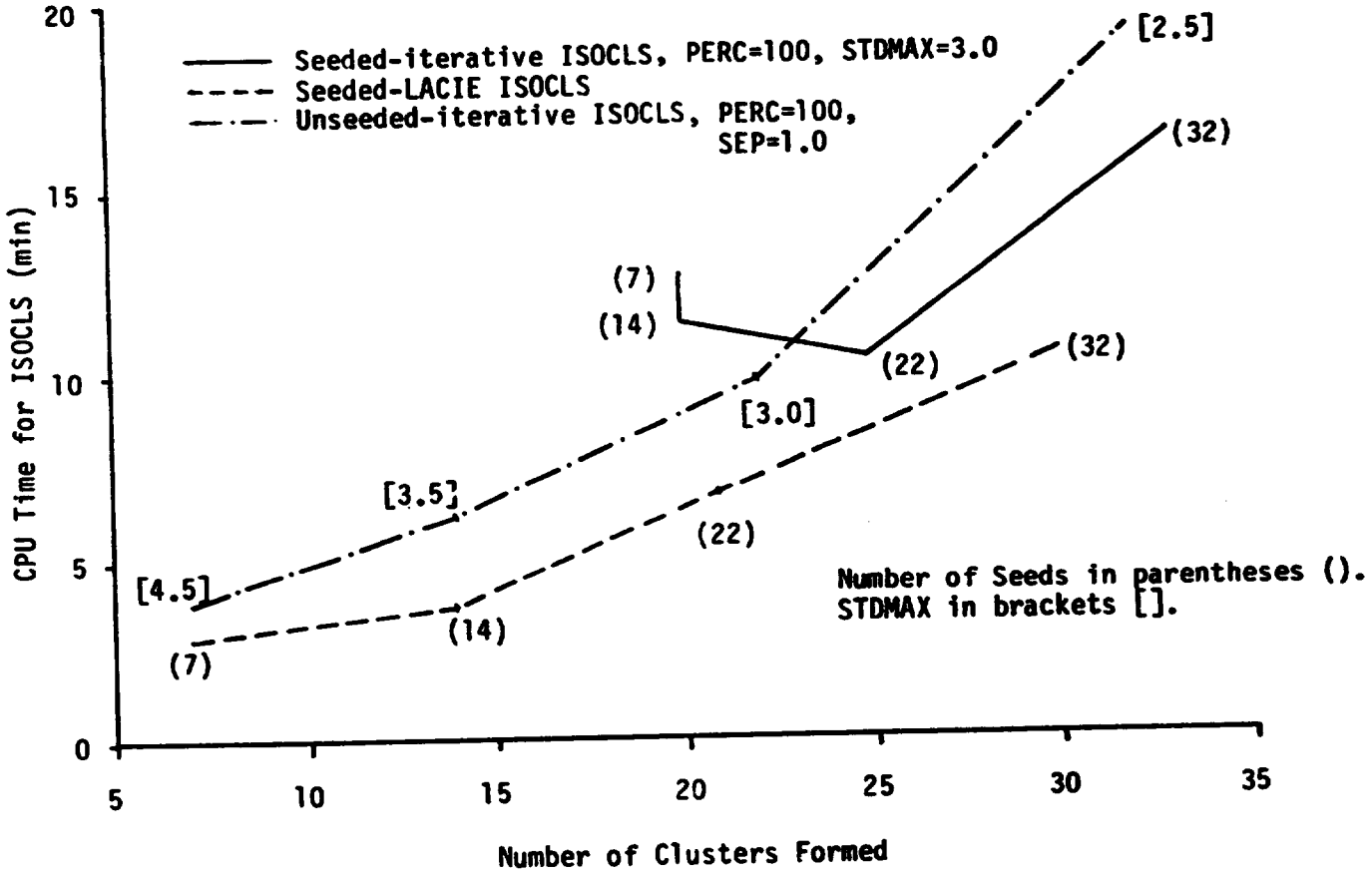


Figure 5.5 Relation between number of clusters formed and CPU time for unseeded-iterative, seeded-LACIE, and seeded-iterative ISOCLS.

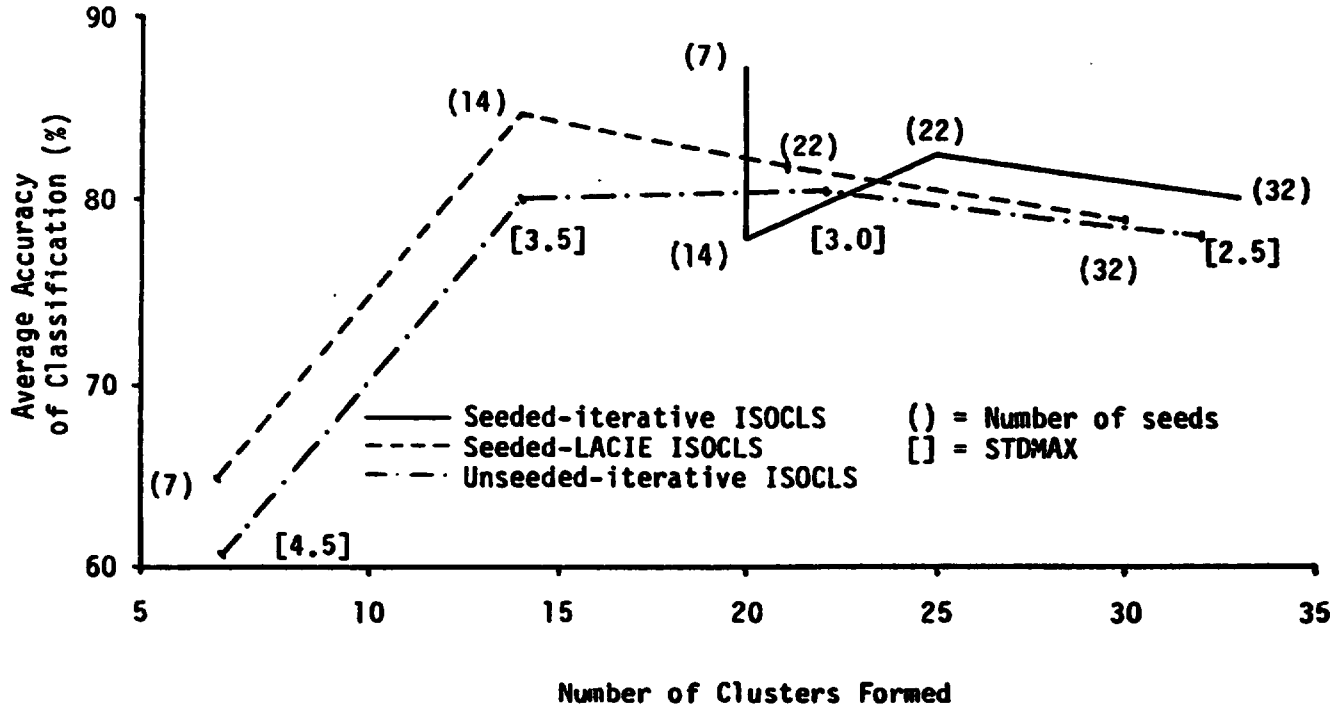


Figure 5.6 Relation between number of clusters formed and average accuracy of classification for unseeded-iterative, seeded-LACIE, and seeded-iterative ISOCLS.

given number of clusters.

2. Seeding does not necessarily save computer time when ISOCLS is allowed to iterate (10 split iterations).
3. Computer time was saved by increasing the number of seeds from 7 to 14 to 22. The savings may have been due to the fact that fewer split iterations were necessary to achieve stable clusters.
4. None of the three methods produced average classification accuracies consistently higher than the other two. The average classification accuracies for all three methods ranged from 78-87% when 14-32 spectral classes were output.
5. The iterative, seeded ISOCLS (7 seeds) formed 20 spectral classes and produced the best results. This particular parameter combination produced the only individual cover type accuracies consistently above 70%.

The accuracy of classification is strongly influenced by how well the statistical classes are labelled. Procedure 1 (of which ISOCLS is part) uses an automatic labelling processor which has parameters subject to analyst control. The effect of this processor on classification accuracy was investigated, and the results are given in the next section.

LABEL - Nearest Neighbor Influences

The LABEL processor identifies spectral classes using one of two methods: 1. K-Nearest Neighbor, 2. All-of-a-Kind (see pg 33for a description). The K-Nearest Neighbor process identifies the spectral class only on the basis of numerical majority of the K-Nearest Neighbor dots. If K=5, the 5 dots closest to the given cluster mean are found. If three dots are hardwood, two are grass, the cluster is labelled hardwood, even if the two grass dots are closer to the cluster mean. No weighting for proximity is done.[1]

---

1. Weighting by proximity could be accomplished by following these steps.
   1. Find the K-Nearest dots.
   2. Calculate the reciprocal of each of the K cluster mean-dot distances.
   3. Sum the reciprocals for each class. The identity of the spectral class is the category with the largest sum.

For example, given that K equals 5 and the five closest dots have the following identities and associated cluster mean-dot distances (euclidean):

Continued, bottom of next page.

The K-Nearest Neighbor LABEL processor was investigated to try and determine the K which yielded the most accurate results. Two different statistics decks were used. The first deck contained 22 spectral classes formed using an unseeded, iterative ISOCLS processor. The second deck contained 21 spectral classes, formed by seeding ISOCLS (1 split iteration) with 22 dots. The results are shown in Figure 5.7.

The same dotfile was used throughout the course of this parameter study. The file consisted of 83 Type 1 dots, 17 hardwood, 33 conifer, 14 grass, 9 barren, and 10 water dots.

Accuracies decreased markedly when 15-Nearest Neighbor was used to label the statistics decks. Ten-Nearest Neighbor yielded the highest (or nearly so) accuracies. The results indicate that K should equal the number of Type 1 dots in the smallest (smallest in terms of number of Type 1 dots) class. In this case, K approximately equal to 9 would produce the best results. The reasoning behind this conclusion is twofold:

1. Given an unknown population mean, the larger the sample size, the higher the probability that the sample mean adequately characterizes the population mean. In terms of this problem, the higher the K, the higher the probability that the spectral classes are correctly labelled.
2. Any K larger than the number of dots in the smallest class tends to 'confuse' LABEL when that class is processed.[2]

---

1. (con't from previous page)

   Normally, this spectral class would be labelled hardwood, since three of the five closest dots are hardwood.

   If a weighted approach to labelling is applied:

Category	Distance
Grass	1.0
Grass	1.5
Hard	2.0
Hard	2.5
Hard	3.0

$$\text{Grass(weighted)} = \frac{1}{1.0} + \frac{1}{1.5} = 1.667$$

$$\text{Hard(weighted)} = \frac{1}{2.0} + \frac{1}{2.5} + \frac{1}{3.0} = 1.233,$$ then the spectral class takes on the identity of the category with the largest sum, in this case, Grass.

2. If there are m points in the smallest class, and K dots are used to label each spectral class, where m is less than K, then at least K-m dots cannot be in that spectral class. Hence LABEL is biased when m is less than K.

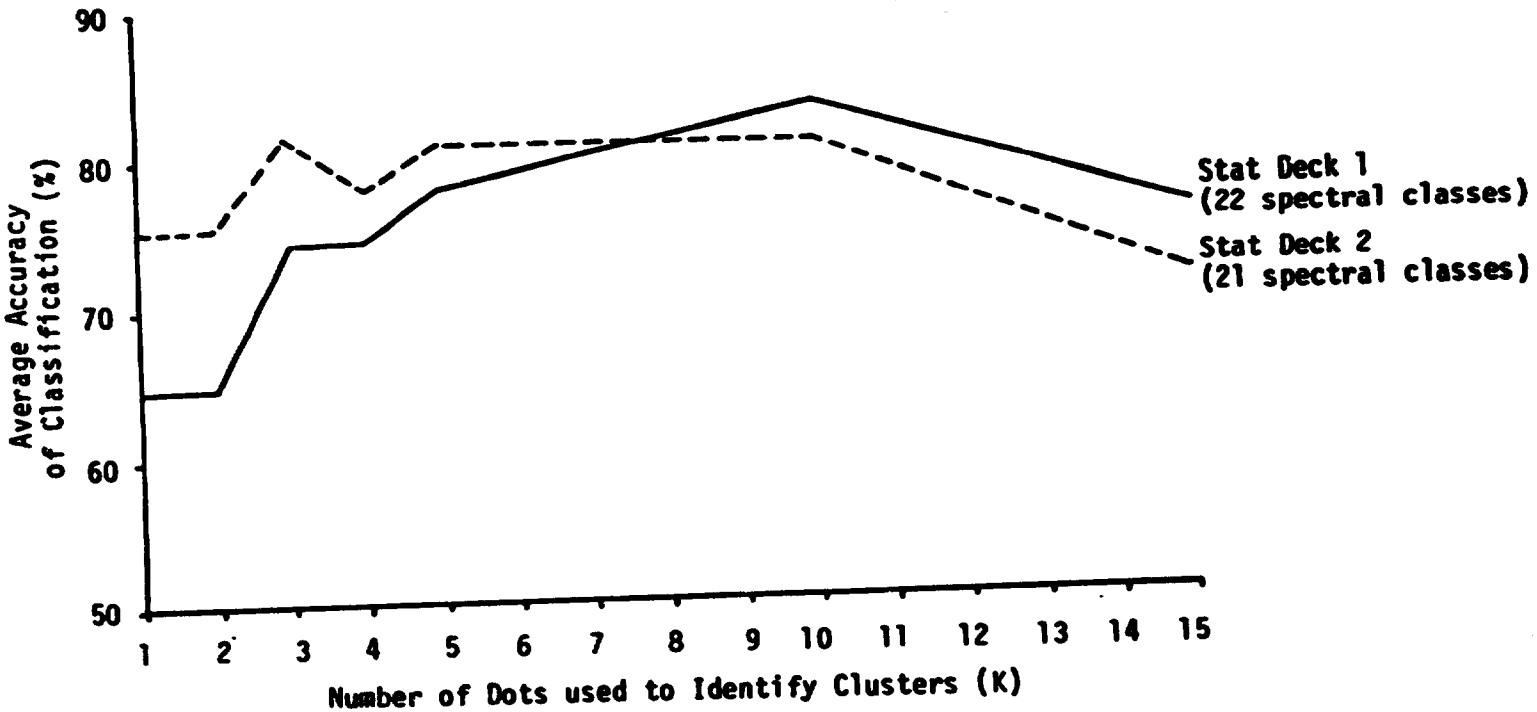


Figure 5.7 Empirical relationship between number of dots used to label the clusters output by ISOCLS (two statistics decks tested) and average accuracy of classification (%).

Based upon these results, it was decided that the second part of this study would use K-Nearest Neighbor labelling, with K equal to the number of Type 1 dots in the class with the smallest number of Type 1 dots.

Comparison of the Multicluster Blocks and Procedure 1 Approaches - Results

This phase of the research would have been greatly simplified if each of the runs had been made with the appropriate parameter levels. The ISOCLS parameter study done on the Vallecito study area indicated clustering trends, but it was found that this parameter study offered little predictive value when Devil Mountain quadrangle MSS data was processed. In Runs 1 and 3 (two of the four runs using ISOCLS), the study area or training blocks had to be clustered three times to obtain the desired number of clusters. Parameter levels which were expected to yield about 15 spectral classes in a training block produced only seven. Parameters which would have output 20 to 25 spectral classes on the Vallecito study area produced only 13 on the Devil Mountain quadrangle. The reduction in number of clusters formed on two different study sites (given the same ISOCLS parameters) may be due to the spectral complexity of the sites. The Vallecito area is relatively complex when compared with the Devil Mountain site which is much more uniformly forested. Hence, an iterative sequence was begun and parameters were manipulated until the desired number of spectral classes were obtained. Repetetive sequences also occured in the two Multicluster Block runs. The training block statistics were merged, often deemed unsatisfactory, and remerged.

The analyst time spent and computer time used in each of the six runs (see Table 5.3) include only that time spent for the final run. In other words, if an area was clustered three times (by ISOCLS) to produce 15 spectral classes, only that computer time used in the third try is added to the total CPU figure. Likewise, if two merge sequences were necessary to produce an adequate final statistics deck using a Multicluster Blocks approach, only the analyst time spent formulating the second deck is summed into analyst time used. Hence the CPU times and analyst's time noted apply to that hypothetical analyst familiar

Table 5.3 Results of the six runs using the Sum-of-Normal-Densities classifier (SoND) and the Maximum-Likelihood classifier.

Run No.	Approach	Analyst Time (hrs)	CPU Used[1] Training Stats (min)	No. of Spectral Classes Developed	Classifier Used	Time (min)[2] for Classifier	Classification Accuracy (%)[3] Con	Hard	Grass	Barr	Avg	Over
1	P-1 ISOCLS unseed,iter	5	7.403	20	SoND	1.168	98.0	83.6	63.5	69.6	78.7	88.1
					ML	1.229	98.3	83.4	64.3	65.2	77.8	88.3
2	P-1 CLUSTER	5	13.502	20	SoND	1.125	95.9	79.9	77.0	45.7	74.6	87.4
					ML	1.245	96.8	80.7	76.2	41.3	73.8	87.8
3	McB ISOCLS unseed,iter	4.5	4.119	15	SoND	1.043	95.1	80.5	50.0	63.0	72.2	83.5
					ML	1.157	95.0	83.6	46.8	60.9	72.6	83.8
4	McB CLUSTER	2.6	7.330	18	SoND	1.119	97.2	74.4	73.8	58.7	76.0	86.5
					ML	1.248	97.7	75.1	69.8	58.7	75.3	86.4
5	P-1 ISOCLS seed,(LACIE)	5	4.217	22	SoND	1.717	99.4	70.0	30.2	45.7	61.3	79.6
					ML	1.321	99.4	69.1	33.7	45.7	62.0	79.9
6	P-1 ISOCLS seed,iter	5	4.693	20	SoND	1.120	98.9	81.4	65.1	39.1	71.1	87.5
					ML	1.239	99.0	81.2	64.7	39.1	71.0	87.4

1. The time used to develop the training statistics include the computer time necessary to run DOTDATA, the clustering processor, and LABEL for the P-1 approaches, and the computer time necessary to run the clustering processor, MERGE, and SEPARABILITY for the Multicluster Block approaches.

2. Number of pixels classified = 1687.

3. Test pixels used: Conifer 932, Hardwood 457, Grass 252, Barren 46, Total 1687 (see Table 4.2).

with the McB approach and merging techniques and/or knowledgeable about the intricacies of the ISOCLS processor and the data set.

The main thrust of the research involved comparing the Procedure 1 method of classifying forestland Landsat MSS data to the McB approach to developing training statistics and classification using those statistics. Below the results obtained for each of the six runs are discussed. Four use a Procedure 1 approach to obtain the classification - Runs 1, 2, 5, and 6. Runs 3 and 4 use a Multicluster Block approach (see Tables 5.3 and 5.4).

Table 5.4 Number of spectral classes in the statistics decks used by the classifiers, by cover type, for each run.

Run No.	Approach	Spectral Classes Used				
		Con	Hard	Grass	Barr	Total
1	P-1 ISOCLS unseed,iter	8	6	4	2	20
2	P-1 CLUSTER	11	5	3	1	20
3	McB ISOCLS unseed,iter	5	5	3	2	15
4	McB CLUSTER	8	4	3	3	18
5	P-1 ISOCLS seed,(LACIE)	13	4	3	2	22
6	P-1 ISOCLS seed,iter	8	6	5	1	20

Discussion of Results of Six Runs

The results of each run are discussed below. The parameter levels used for individual processors (where important) are given and any problems encountered are noted. For a description of each run, refer to the Procedures section (pgs 45-48).

Run 1

This Procedure 1 approach used the ISOCLS clustering processor in an iterative mode (10 splits and one combine). Type 1 dots were not used to seed the processor.

Twenty to twenty five spectral classes were desired on the Devil Mountain test site.[1] The area was clustered three times, the first two produced 13 and 18 spectral classes (STDMAX 2.6 and 2.2 respectively). The 13 spectral class statistics deck produced average and overall classification performances of 72.0 and 86.5% respectively. These accuracies were 6.7 and 1.6% lower than the 20 spectral class statistics deck generated when STDMAX was set to 2.0.[2] Approximately nine CPU minutes were used in the first two attempts.

The LABEL processor identified 8 of the spectral classes as coniferous, 6 hardwood, 4 grass, and 2 barren. These labelled statistics were used by the classifiers; the results are given in Table 5.5.

Compilation of the results from each of the six runs (see Table 5.3) show that these statistics produced the best overall and average classification accuracies.

Run 2

Run 2 is essentially the same as Run 1 except that the LARSYS CLUSTER processor was used (instead of ISOCLS) with the EODLARSYS processors DOTDATA and LABEL to produce the training statistics. Twenty spectral classes were requested, and the entire quad clustered on an interval of two. CLUSTER (CONV 98.5) took twice as much CPU time to produce the same number of clusters though only one clustering attempt was necessary (see Table 5.14).

The labelled training statistics produced classification results lower than the first run's, though only the average classification accuracy was significantly lower (according to the Newman-Keuls Range

---

1. Twenty to twentyfive spectral classes were deemed desirable based on results of the parameter study and on previous classification experience.

2. Other parameter levels used to cluster the Devil Mountain quadrangle (all Run 1 attempts): PERC 90, ISTOP 10, DLMIN 3.2, CLUS 25, Number of Seed Dots 0.

Table 5.5 Classification results using Run 1 training statistics and
a. the LARSYS Maximum Likelihood classifier,
b. the EODLARSYS Sum-of-Normal-Densities classifier
on the Devil Mountain quadrangle.

a. Maximum Likelihood classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	98.3	916	15	1	0
Hard	457	83.4	45	381	30	1
Grass	252	64.3	14	50	162	26
Barr	46	65.2	7	0	9	30

Average Performance: (311.2/4) = 77.8

Overall Performance: (1489/1687) = 88.3

b. Sum-of-Normal-Densities classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	98.0	913	18	1	0
Hard	457	83.6	43	382	31	1
Grass	252	63.5	13	47	160	32
Barr	46	69.6	6	0	8	32

Average Performance: (314.7/4) = 78.7

Overall Performance: (1487/1687) = 88.1

Test, Table 5.12). LABEL found 11 conifer spectral classes, 5 hardwood, 3 grass, and one barren class. The confusion tables produced by the classifiers using these training statistics are reproduced in Table 5.6.

Unlike Run 1, the barren class was identified with a very low accuracy. Many barren test field points were classified as grass. Evidently the CLUSTER processor formed a mixed spectral class which was labelled as grass, a class which might have more accurately been labelled barren. Procedure 1, without analyst participation, uses all spectral classes produced by the clustering processor. Mixed spectral classes (grass-barren mixtures) must be labelled and used. These classes would more than likely be deleted if encountered in a McB approach. In defense of the P-1 approach, the LABEL processor (10-Nearest Neighbor) did better than anticipated, and in general produced statistics of higher quality (as judged by classification accuracy) than those produced by this analyst.

Run 3

Run 3 was the first of two Multicluster Blocks approaches investigated in this study. This run used the ISOCLS clustering processor in an unseeded, iterative mode to cluster three training blocks established on the Devil Mountain quadrangle. Between 14 and 16 spectral classes were desired on each block.

Again the problem of not being able to control the number of spectral classes output by ISOCLS plaqued this run. The number of spectral classes produced by ISOCLS is very difficult to estimate if the processor is not seeded and is allowed to iterate. Each block was clustered three times; the parameters used and the results are given below. Those parameters not mentioned were set to their default values. No seeding was done (no Type 1 dots input).

The data is given first to provide more insight into the ISOCLS processor, second to document two oddities. As explained in the parameter study section, one expects the number of spectral classes produced by ISOCLS to increase as STDMAX and/or DLMIN is reduced. The results from training block 1 (cluster attempts 1 and 2) show this is not always the case. Much mathematical handwaving concerning data

Table 5.6 Classification results using Run 2 training statistics and
a. the LARSYS Maximum Likelihood classifier,
b. the EODLARSYS Sum-of-Normal-Densities classifier
on the Devil Mountain quadrangle.

a. Maximum Likelihood classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	96.8	902	30	0	0
Hard	457	80.7	40	369	47	1
Grass	252	76.2	16	36	192	8
Barr	46	41.3	3	0	24	19

Average Performance: (295.0/4) = 73.8
Overall Performance: (1482/1687) = 87.8

b. Sum-of-Normal-Densities classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	95.9	894	38	0	0
Hard	457	79.9	37	365	54	1
Grass	252	77.0	12	26	194	20
Barr	46	45.7	3	0	22	21

Average Performance: (298.5/4) = 74.6
Overall Performance: (1474/1687) = 87.4

Table 5.7 Parameter values used to cluster training blocks using ISOCLS, Run 3.

Tr Block	Cluster Attempt	Parameter Levels PERC	DLMIN	ISTOP	STDMAX	CLUS	Num Pixels	CPU Time Used(min)	Num Sp Cl Formed
1	1	90	3.2	10	2.8	16	841	0.919	7
	2	90	2.4	10	2.0	16	841	0.598	6
	3+	100	2.0	10	1.7	16	841	0.909	15
2	1	90	3.2	10	2.8	16	1804	1.507	9
	2	90	2.6	10	2.1	16	1804	1.283	13
	3	90	2.5	10	2.0	16	1804	1.219	15
3	1	90	3.2	10	2.8	16	1378	1.452	10
	2	90	2.8	10	2.1	16	1378	1.283	13
	3	90	2.7	10	2.1	16	1378	1.021	15

+ PMIN dropped from 20 to 5.

distribution in the training block and the ISOCLS parameter values could explain the phenomenon noted (scientific hindsight can be very accommodating). Suffice it to say that reducing STDMAX or DLMIN does not necessarily increase the number of spectral classes produced. The second feature of interest is the CPU time used. Due to the characteristics of the IBM computing system, the CPU time required for a processor does not always indicate the actual computing time used for clustering. If the LARS computer is busy, it begins spending an inordinate amount of time spooling programs in and out of memory core. Processing time can be increased by more than 50% for exactly the same job; therefore the CPU time used depends (in part) on system demands. Hence the decrease in CPU time noted for all three training blocks as more spectral classes were produced may be an artifact of the computing system and the demands on that system at the time the jobs were run. Attempts were made throughout the study to minimize the effects of these computing inefficencies by running all the programs involved in this study interactively between 6 pm and 1 am (when system demands are low) or by running jobs night batch.

The fifteen spectral classes produced in each training block were identified using color infrared photography, the cluster map output from ISOCLS, and a Zoom Transfer Scope. The cluster classes output on each training block were very difficult to photointerpret. This

difficulty is reflected in the fact that it took over twice as long to photointerpret the training blocks; it took 2.5 hours to identify the spectral classes produced by ISOCLS, and about 1.1 hours to identify the CLUSTER (Run 4) spectral groupings. The ISOCLS spectral classes seemed confusing, more randomly distributed across cover types. Thirteen spectral classes (out of 45) were identified as mixed classes and were deleted prior to the Merge process. Only 4 (out of 45) were deleted from the statistics produced by CLUSTER in Run 4.

The modified merge procedure described in the Procedures section was done twice. The first sequence produced a statistics deck deemed unsatisfactory because the results of the first merge sequence indicated that the analyst slighted the grass category in favor of hardwood and barren. The same statistics were merged a second time and used by the classifiers; the confusion tables are presented in Table 5.8. The final statistics deck used by the classifier contained 15 spectral classes, 5 conifer, 5 hardwood, 3 grass, and 2 barren classes.

Throughout this study, the quality of the training statistics was judged soley on the classification accuracies obtained. A Newman-Keuls Range Test placed the Run 3 training statistics in the lower half of the six studied. The lower performance of this Multicluster Block attempt was expected due (in part) to the inferior quality of the unlabelled training statistics output by ISOCLS using the parameters listed in Table 5.7.

Run 4

The training statistics output by CLUSTER proved to be much easier to photointerpret compared to those output by ISOCLS in Run 3. The resulting classification accuracies improved significantly.

The three training blocks were clustered individually into 15 spectral classes, taking 5.777 minutes of computer time. In general, ISOCLS could generate the same number of spectral classes as CLUSTER using much less CPU time. However, additional computer time was necessary in order to find the correct parameter levels so that ISOCLS could produce an acceptable number of clusters. The final result was that CLUSTER turned out to be the less expensive and less aggravating of the two clustering processors.

Table 5.8 Classification results using Run 3 training statistics and
a. the LARSYS Maximum Likelihood classifier,
b. the EODLARSYS Sum-of-Normal-Densities classifier
on the Devil Mountain quadrangle.

a. Maximum Likelihood classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	95.0	885	47	0	0
Hard	457	83.6	27	382	47	1
Grass	252	46.8	0	131	118	3
Barr	46	60.9	0	4	14	28

Average Performance: (286.2/4) = 71.6
Overall Performance: (1413/1687) = 83.8

b. Sum-of-Normal-Densities classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	95.1	886	46	0	0
Hard	457	80.5	30	368	58	1
Grass	252	50.0	0	122	126	4
Barr	46	63.0	0	3	14	29

Average Performance: (288.6/4) = 72.2
Overall Performance: (1409/1687) = 83.5

Again the merging process was done twice since results of the first merge were judged unsatisfactory (based on initial classification accuracies obtained for certain cover types, specifically hardwood or grass). The second attempt used the same training statistics; photo-interpretation was checked and the 45 spectral classes recombined. The final statistics deck contained 18 spectral classes, 8 conifer, 4 hardwood, 3 grass, and 3 barren classes. The results are noted in Table 5.9.

Run 4 was the only run which gave statistically similar results as Run 1 in both accuracy categories (0.5 alpha level) (see Table 5.12).

Run 5

This run investigated the effects of the LACIE approach to the classification of forestlands. As such, ISOCLS was seeded with 23 Type 1 dots (10 conifer, 7 hardwood, 4 grass, and 2 barren dots). The parameter levels used for clustering and labelling were those set by LACIE (and listed in Table 5.1, Run 5). LABEL identified the 22 spectral classes (13 conifer, 4 hardwood, 3 grass, and 2 barren classes) and these labelled statistics were used by the classifiers with the results given in Table 5.10.

The Newman-Keuls Range Test showed that the Procedure 1 approach using LACIE parameters was significantly lower in both average and overall accuracy than any other run at the $\alpha = 0.05$ level (see Table 5.12). Of the four Procedure 1 approaches attempted, this one used the least CPU time, since LACIE parameters require the clustering processor to pass through the data only three times. The method used to label the statistics (K-Nearest-Neighbor, K=1) makes this procedure sensitive to mislabelled dots, possibly reducing chances of correct spectral class identification.

The LACIE parameters may work well in agricultural settings where a dot identified as wheat may be counted on to exhibit relatively uniform reflectance values, no matter where it is located in the segment. Extensive heterogeneity is the rule in the forests of the San Juan Mountains; this characteristic is perhaps the prime reason behind the poor showing of the P-1 (LACIE) approach for classifying some covertypes in this forested scene.

Table 5.9 Classification results using Run 4 training statistics and
a. the LARSYS Maximum Likelihood classifier,
b. the EODLARSYS Sum-of-Normal-Densities classifier
on the Devil Mountain quadrangle.

a. Maximum Likelihood classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	97.7	911	20	1	0
Hard	457	75.1	40	343	72	2
Grass	252	69.8	0	62	176	14
Barr	46	58.7	0	0	19	27

Average Performance: (301.3/4) = 75.3
Overall Performance: (1457/1687) = 86.4

b. Sum-of-Normal-Densities classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	97.2	906	23	3	0
Hard	457	74.4	37	340	78	2
Grass	252	73.8	0	53	186	13
Barr	46	58.7	0	0	19	27

Average Performance: (304.1/4) = 76.0
Overall Performance: (1459/1687) = 86.5

Table 5.10 Classification results using Run 5 training statistics and
a. the LARSYS Maximum Likelihood classifier,
b. the EODLARSYS Sum-of-Normal-Densities classifier
on the Devil Mountain quadrangle.

a. Maximum Likelihood classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	99.4	926	6	0	0
Hard	457	69.1	69	316	68	4
Grass	252	33.7	16	52	85	99
Barr	46	45.7	6	0	19	21

Average Performance: (247.9/4) = 62.0
Overall Performance: (1348/1687) = 79.9

b. Sum-of-Normal-Densities classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	99.4	926	6	0	0
Hard	457	70.0	61	320	71	5
Grass	252	30.2	15	45	76	116
Barr	46	45.7	6	0	19	21

Average Performance: (245.3/4) = 61.3
Overall Performance: (1343/1687) = 79.6

Run 6

The final run attempted to combine the "best" features of Runs 1 and 5, i.e., the capability of ISOCLS to accept initial cluster centers and its controlled iteration capabilities. The processor was seeded with the same dots used in Run 5, and ISTOP was increased to three (four passes through the data). Labelling was done using the 10-Nearest-Neighbor approach. The extra iterations minimally increased CPU time. The iterations combined with the increased number of dots used to label the spectral classes resulted in an eight to ten percent increase in classification accuracy over the LACIE approach (Run 5).

Twenty spectral classes were output by ISOCLS, 8 were labelled conifer, 6 hardwood, 5 grass, and 1 barren. See Table 5.10 for the cross-classification results.

The barren cateogry was classified very poorly, the lowest of any of the six runs in part because the category was characterized by only one spectral class. In spite of this, Run 6 results ranked in the upper half of the six runs in overall accuracy, and was not statistically different from Run 1 (see Table 5.12). However, it's capabilities are more clearly shown by evaluating the average classification accuracy, where it ranked in the lower third.

Analysis of the Results of the Six Runs

Performance and Test Fields

The test fields used to evaluate the runs were conscientiously selected to represent the entire range of stand conditions. In other words, a field was considered forested if more than 30% of the area was tree crown. A forested test field was judged to be hardwood if more than 50% of the crowns present were hardwood. Forested test fields were selected in pure stands and in mixed, hence hardwood-conifer cross-classification was expected. Coniferous test fields were rarely misclassified, but hardwood fields were often misclassified into conifers or grass. The conifer-hardwood mixup may be due to the attributes of the hardwood test fields (i.e., many hardwood test fields had high percentages of conifer in them).

The hardwood-grass cross-classification was most likely due to the spectral similarity of the two classes. The difference between the two

Table 5.11 Classification results using Run 6 training statistics and
a. the LARSYS Maximum Likelihood classifier,
b. the EODLARSYS Sum-of-Normal-Densities classifier
on the Devil Mountain quadrangle.

a. Maximum Likelihood classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	99.0	923	8	1	0
Hard	457	81.2	51	371	34	1
Grass	252	64.7	8	79	163	1
Barr	46	39.1	7	0	21	18

Average Performance: (284.0/4) = 71.0
Overall Performance: (1475/1687) = 87.4

b. Sum-of-Normal-Densities classification results:

Group	No of Samps	Perc Corct	Number of Samples Classified into Con	Hard	Grass	Barr
Con	932	98.9	922	9	1	0
Hard	457	81.4	50	372	34	1
Grass	252	65.1	8	79	164	1
Barr	46	39.1	7	0	21	18

Average Performance: (284.5/4) = 71.1
Overall Performance: (1476/1687) = 87.5

might be more pronounced at a later date, but in early June in the southern Rockies, the new hardwood flush has a high infrared reflectance, much like that of the grasslands. The statistics for Runs 3 and 4 were remerged due to this hardwood-grass mixup. Hardwood accuracies were increased only at the expense of grass, and vice-versa. This cross-classification plagued all six runs, but was most noticeable in Runs 3 and 4, the McB approaches.

The barren class had, of the four considered, the worst classification accuracy over the six runs. For the most part, it was generally confused with grassland, though cross-classification with conifer occurred regularly, but less often. The barren cover type makes up only a small percentage of the Devil Mountain quad, and occurs in irregular patches. It occurs most often at the tops of peaks (and grades into grassland-tundra) and comprises a small section of the Piedra River valley, in union with sparse stands of Ponderosa Pine (less than 30% crown closure). Hence the barren-grassland cross-classification may be the result of the test fields' edge effects. Only the four P-1 approaches classified barren test field pixels into the conifer class. Evidently, LABEL was not able to distinguish the unlabelled sparse conifer-barren spectral classes as barren classes since there were not enough Type 1 dots exhibiting similar attributes to constitute a majority. These classes were successfully identified by the analyst using the McB appraoch.

A Comparison of the Methods of Developing Training Statistics

Newman-Keuls Range tests were run on the overall and average accuracies obtained for each of the six runs. The procedure involved steps outlines on page 2.7-11 of Landgrebe (1976).

Procedure 1 approaches are responsible for the highest and lowest classification performances, indicating that the P-1 approach carries much potential if used properly. The P-1 approach using LACIE parameters are definitely not the proper way to classify a forested scene.

Runs 1 and 4 are not statistically different at the 0.05 alpha level, and are the two best approaches to developing training statistics for forestland classification. The use of either of these two methods

would depend upon the size to the area being classified and the ancillary information available to the analyst.

Table 5.12 Newman-Keuls Range Tests of average and overall classification accuracies for the six runs (Devil Mountain study site).

Note: Runs sharing a common line are not significantly different at the 0.05 alpha level.

The maximum likelihood classification accuracies were used in this analysis.

a. Average Classification Accuracy:

Run No.	1 P-1 ISOCLS unseed,iter	4 McB CLUSTER	2 P-1 CLUSTER	3 McB ISOCLS unseed,iter	6 P-1 ISOCLS seed,iter	5 P-1 ISOCLS seed,LACIE
Accuracy:	77.8	75.3	73.8	71.6	71.0	62.0

b. Overall Classification Accuracy:

Run No.	1 P-1 ISOCLS unseed,iter	2 P-1 CLUSTER	6 P-1 ISOCLS seed,iter	4 McB CLUSTER	3 McB ISOCLS unseed,iter	5 P-1 ISOCLS seed,LACIE
	88.3	87.8	87.4	86.4	83.8	79.9

The Procedure 1 approach using ISOCLS as an unseeded, iterative processor would be best where forest inventory information is available and can be located in the Landsat grid. For instance, this approach would be ideal in Minnesota where the USFS has established a grid system of photointerpreted and ground checked points. The field points have been located on a Universal Transverse Mercator rectangular grid coordinate system. The UTM coordinates are easily converted to Landsat coordinates using a simple Fortran program. These points of known identity have been established in forest, agricultural, and urban areas, and on water. In some instances however, certain cover types will not be adequately represented by the inventory information supplied. In such cases, the analyst must locate and identify additional pixels in covertypes lacking adequate representation. Once this inventory information

is input (requiring little analyst interaction), the classification process is completely machine-oriented. The disadvantage to the P-1 procedure is that the entire area of interest must be clustered in order to develop the unlabelled training statistics. Costs may be kept reasonable by clustering large areas on an interval, i.e., cluster a representative sample of the data. The P-1 approach has potential for an automated, operational forestlands classification system, however, the software would have to be streamlined to be cost effective.

The Multicluster Block approach (using CLUSTER) to developing the training statistics becomes more attractive as the size of the study area increases. Savings in computer time, though noticeable, were not great in this study since the entire Devil Mountain quad was not large when compared to the training blocks selected.[1] The McB approach is also suited to areas where ancillary information is scarce. Forest inventory information may be an aid in identifying the spectral classes in each block, but quality airphotos or a type map are more useful.

Following is a cost evaluation of the time used to develop the training statistics for the Devil Mountain quad for each of the six runs.[2] The evaluation is based on a $5.00 charge for a computing minute ($300.00/hour) and $10.00 per hour for analyst cost. The first cost column assumes the analyst had to locate and identify his own Type 1 dots for the P-1 approaches (as was the case in this study). The second cost column assumes forest inventory information is available and the analyst spends one hour reformatting this information to make it available to the DOTDATA processor.

The cost evaluation is really relevant only to those who might wish to institute an operational system using the current software, at best highly unlikely. However, the point can be made that the P-1 approaches are cheaper than the McB approaches when forest inventory

---

1. The McB approach used 12.1% of the available data on the Devil Mountain quadrangle. Procedure 1, clustering on an interval of 2, used 25.2% of the available data.

2. In the case of Runs 1, 2, 5, and 6, the P-1 runs, the processors involved are DOTDATA, ISOCLS or CLUSTER, and LABEL. In the McB approaches, Runs 3 and 4, ISOCLS or CLUSTER, MERGE, and SEPARABILITY CPU times are summed.

Table 5.13 Cost of developing training statistics for the Devil Mountain quadrangle

Run	Approach	Cluster Function	Analyst Time (hours)	Computer Time (min)	Total Cost ($) Estab Dots	Inven Avail
1	P-1	ISOCLS unseed,iter	5	7.403	87.02	47.02
2	P-1	CLUSTER	5	13.502	117.51	77.51
3	McB	ISOCLS unseed,iter	4.5	4.119	65.60	65.60
4	McB	CLUSTER	2.6	7.330	62.65	62.65
5	P-1	ISOCLS seed,LACIE	5	4.217	71.09	31.09
6	P-1	ISOCLS seed,iter	5	4.693	73.47	33.47

information is available. Inventory information reduces analyst involvement in the P-1 process markedly, whereas the inventory information really has no effect at all on the amount of time spent by the analyst involved in a Multicluster Block approach. This point is important when considering future classification systems. As computer technology grows and software is refined, the amount of CPU time used in a given process will be negligible when compared with the analyst expenses. The McB approach is analyst intensive, an analyst must be involved in developing the training statistics. P-1 makes no such demands (or minimal demands) if inventory information is available. As such, P-1 may be the most cost-effective approach to developing training statistics. Table 5.13 also suggests that seeding the ISOCLS processor may be the cheapest method of developing the statistics. This study only scratched the surface of the effects of seeding and number of clustering iterations on the accuracy of classification. The seeded approach may yet prove to be the most accurate and cost effective method of classifying forested lands using available forest inventory data.

The Classifiers

The training statistics developed in each of the six runs were used by two different classification procedures, the EODLARSYS Sum-of-Normal-Densities classifier (CLASSIFY) and the LARSYS maximum likelihood

classifier (CLASSIFYPOINTS). Both are perpoint processors; both are more thoroughly described in the Materials section.

Paired T-tests were run on the transformed accuracies (arcsin $\sqrt{p}$, p = average or overall accuracy) to determine statistically significant differences between the classifiers. No significant difference was found, even at the 0.10 alpha level.

Table 5.14 Differences in average and overall accuracies between the two classifiers for the six runs, Devil Mountain quadrangle.

Note: The letters in parentheses indicate the classifier which produced the higher accuracy. SoND - Sum-of-Normal-Densities
ML - Maximum-Likelihood

Note: Paired-T statistical analysis used transformed data; the numbers below have not been transformed.

	Average Accuracy			Overall Accuracy		
Run	SoND	ML	Difference	SoND	ML	Difference
1	78.7	77.8	0.9 (SoND)	88.1	88.3	0.2 (ML)
2	74.6	73.8	0.8 (SoND)	87.4	87.8	0.4 (ML)
3	72.2	71.6	0.6 (SoND)	83.5	83.8	0.3 (ML)
4	76.0	75.3	0.7 (SoND)	86.5	86.4	0.1 (SoND)
5	61.3	62.0	0.7 (ML)	79.6	79.9	0.3 (ML)
6	71.1	71.0	0.1 (SoND)	87.5	87.4	0.1 (SoND)

In general, the maximum-likelihood classifier produced higher overall accuracies, and the SoND classifier produced better average accuracies (again, statistically indistinguishable).

A study by Scholz, Fuhs, and Hixson, 1979, compared five classification processors, among them the two used in this study. They found that the Sum-of-Normal-Densities classifier took up to five times as much computer time as the maximum-likelihood classifier. In each of the six runs in this study, the classifiers categorized 1687 pixels using all four channels of single acquisition Landsat data. A paired-T test was run to determine if there were significant differences in CPU time used between the classifiers. No differences were found ($t_{calculated} = 0.29$). In fact, in five of the six runs, the

maximum-likelihood classifier took longer, though CPU time differences only involved magnitudes of tenths of a minute (see Table 5.3).

Scholz, Fuhs, and Hixson, 1979, concluded that:

> "the key to accurate classifications is in the development of the training statistics for the classification rather than in the classifier itself."

This research confirms their final conclusions, but disagrees with the classification time element. Perhaps differences would be more noticeable if more pixels were classified (such as the 5 x 6 mile segment - 22,932 pixels - used in the study mentioned).

The Clustering Processors

The processors used to develop the training statistics in Runs 1 and 2 (P-1 approach) and in Runs 3 and 4 (McB approach) were identical except for the clustering processors. Differences in computer time and in classification accuracy then are a direct result of differences in ISOCLS and CLUSTER. The parameters used in each of the runs are detailed in the description of the Run results.

Table 5.15 indicates that it takes CLUSTER approximately twice as long as ISOCLS to output a given number of spectral classes.

Table 5.15 Comparison of CPU time used to develop unlabelled training statistics using the ISOCLS and CLUSTER processors.

Approach	Area Clustered	No. Pixels Clustered	Time (min) used ISOCLS	Time (min) used CLUSTER	No. spectral classes
P-1	Dev Mt Qd	8372	6.04	12.09	20
	Tr Bl 1	841	0.91	0.90	15
McB	Tr Bl 2	1804	1.22	2.86	15
	Tr Bl 3	1378	1.02	2.02	15
McB(total)		4023	3.15	5.78	45

More important however is an analysis of the accuracies (see Runs 1, 2, 3, and 4, average and overall accuracies, Table 5.3). Average accuracies noticeably declined (approximately 4%) when CLUSTER was used instead of ISOCLS in the P-1 approach. The overall accuracy was minimally affected. On the other hand, McB accuracies using statistics generated by CLUSTER (Run 4) were noticeablely higher (4% average, 3% overall) than

the ISOCLS (Run 3) results. The fact that the Run 3 accuracies were lower than Run 4's was no suprise since identification of the ISOCLS training block spectral classes was very difficult.

The results of the cluster processors comparison is best summed up by listing the advantages and disadvantages of each.

ISOCLS

Disadvantages:

1. Cannot specify number of spectral classes output.
2. Intricate parameter controls require much analyst experience if ISOCLS is to be used properly.
3. The spectral classes output were difficult to identify using photointerpretive methods; the 'quality' of the spectral classes seemed low.

Advantages:

1. Can accept beginning cluster means which reduces CPU time.
2. In general, requires less CPU time than CLUSTER to output a given number of spectral classes even when not seeded.

CLUSTER

Disadvantages:

1. Cannot accept initial cluster means (a characteristic of the LARS program only).
2. Requires more CPU time.

Advantages:

1. Much easier to use, parameters simple.
2. Number of spectral classes requested equals the number output.

In summary, do not mix software clustering processors. ISOCLS seems best suited to a P-1 approach, CLUSTER to a McB approach.

CHAPTER 6 - CONCLUSIONS AND RECOMMENDATIONS

Six different methods involving two general approaches were tested to determine the best method of developing training statistics. Experimental design allowed comparison of the two classification processors and the two clustering processors involved in the study.

Conclusions

The Procedure 1 approach to developing training statistics involves the use of the DOTDATA, ISOCLS and LABEL processors. The statistics output by these EODLARSYS programs showed great promise for accurately classifying forested areas. The Procedure 1 approach to developing training statistics is extremely versatile. Only four different parameter sets were tested. Conclusions based on this research may be drawn:

1. The P-1 approach using ISOCLS in an unseeded, iterative mode and the Multicluster Blocks approach using CLUSTER were the two best approaches to developing training statistics. The use of either method would basically depend upon the type and amount of ancillary information available to the analyst.
2. The P-1 approach makes maximum use of current forest inventory information. Availability of this information minimizes the analyst involvement in the classification procedure. The reduction in analyst involvement is critical because a. in an operational mode, analyst time would most likely be the biggest single cost factor and b. it reduces the largest source of bias.

P-1 has it's restrictions. First the dotfile must contain a sufficent number of pixels in each cover type of interest and should cover the range of variability within a cover type. Forest inventory data which provides only information on forested plots may be used, but additional dots must be identified and located in all of the nonforest cover types (urban, grassland, barren, and water). The accurate location and identification of all dots is critical, for mislabelled or poorly located dots may lead to spectral classes incorrectly identified.

3. LACIE parameters should not be used to classify forested areas. The essentially noniterative ISOCLS clustering processor and K-Nearest Neighbor (K=1) method of labelling the ISOCLS statistics do not produce consistent results in a heterogeneous area.
4. Seeding ISOCLS with Type 1 dots reduces CPU time, and shows promise for classifying forested areas. Preliminary research indicates that each cover type should have at least two seed dots to better insure that the smaller cover types are spectrally represented.
5. When K-Nearest Neighbor is used to label the spectral classes, K should equal that number of dots in the covertype with the smallest dot representation in the dotfile.

The Multicluster Blocks approach to developing training statistics produced classification results on par (statistically indistinquishable) with the best P-1 results. The McB approach requires a great deal of analyst interaction which, under certain circumstances, is desireable. Experience has shown that P-1's automated labelling processor often misses critical cover types that make up only a small percentage of the study area. Also, naturally, LABEL cannot abide by arbitrary definitions unless those definitions are exemplified in the dotfile. Crown density differences (for instance, between barren and conifer, where crown densities less than 30% are considered nonforested) may lead to mislabelling. These problems are overcome using the McB approach since spectral class labelling is done by the analyst. The McB approach is also efficent when developing training statistics for large study areas. P-1 requires the entire study area be clustered, and this can become expensive when the area involved reaches into the hundreds of thousands of acres, unless the area is clustered on an interval of 2, 3, or 4 (i.e., sample one fourth, one ninth, or one sixteenth of the pixels). The clustering of representative training blocks (Multicluster Blocks approach) may be more realistic.

The method used to develop the training statistics most certainly influenced the classification more than the classifiers used. This study found that:

1. The differences in classification accuracies between the EODLARSYS Sum-of-Normal-Densities classifier and the LARSYS maximum-likelihood classifier were minimal. Statistical evaluation showed no significant differences between the two classifiers.

2. The CPU times used by the classifiers to classify 1687 pixels were not significantly different. On the average, the maximum-likelihood classifier took more CPU time than the SoND classifier.

The clustering processor used to develop the training statistics significantly affected the classification accuracies. Sweeping conclusions cannot be made due to the versatility of the ISOCLS processor. The six runs indicated:

1. That P-1 approaches should use the ISOCLS processor.
2. A Multicluster Blocks approach yields better results when the CLUSTER processor is used.
3. CLUSTER takes approximately twice the CPU time to output a given number of spectral classes.

These conclusions should be viewed in the light that future ISOCLS seeding and iteration studies may further refine the ISOCLS processor capabilities. CLUSTER grouped the training block data into easily identifiable spectral classes, the ISOCLS classes were very difficult to identify. Seeding and/or parameter changes would affect, and may improve, the quality of the clusters output by ISOCLS, making it useful to the McB approach.

The performance of the ISOCLS processor (number of clusters output, variability within clusters, and CPU time used) is affected by:

1. the parameter levels used;
2. the spectral variability of the study area;
3. the size of the study area.

Following are lists of parameters that might be used on a forested area by an analyst new to the rigors of Procedure 1 and the ISOCLS processor. The parameters provide a handhold, a starting point for the novice P-1 analyst. The parameters given presume that approximately 8500 pixels (4 channel) are processed and the study area is spectrally heterogeneous. Between 20 and 25 spectral classes are desired.

Unseeded: No Type 1 dots input as initial cluster means.

STDMAX	1.8	SEQUEN	SC
PERC	90	DLMIN	3.0
ISTOP	10	CLUS	25

all other parameters default.

If these parameters yield less than the desired number of spectral classes, reduce STDMAX and DLMIN (in 0.2 increments) and/or increase PERC to 100.

If seeds are input, the following parameters might be used.

Seeded: Use 20-25 Type 1 dots if 20-25 spectral classes are desired. The percentage of dots in each cover type should be roughly proportional to the percent composition of the study area (i.e., if 30% of the study area is hardwood forest, then 30% of the seeds should be hardwood dots). Each cover type should have at least two dots input to the ISOCLS processor.

STDMAX	2.8	SEQUEN	SC
PERC	90	DLMIN	3.0
ISTOP	3	CLUS	25

all other parameters default.

If these parameters yield less than the desired number of spectral classes, follow the instructions above and/or increase ISTOP in increments of 2.

Again, the parameters given above are starting points and may not produce the results expected. ISOCLS parameters must be manipulated in order to produce an adequate set of training statistics (a sure thing, right up there with death and taxes).

As Schubert (1978) stated, the ultimate foreseeable goal in Landsat data processing is the development of a fully automatic procedure which would allow a reproducible classification of a given scene. Procedure 1 is a viable method of classifying forested areas. It is ideally suited to utilizing forest inventory data directly, as such it can conceivably be modified into an automated classification system. Many questions remain, some of which are listed in the next section, but an automated forest inventory system is feasible and may attract interest from the private sector.

Research Recommendations

Researchers rarely find themselves at a dead end. The answer to one question gives rise to others. This research has raised points that might be of interest to others, especially those concerned with classification procedures development.

The LABEL processor is very important in the development of P-1 training statistics and may be the weakest link in the classification process. The Nearest-Neighbor concept is just one method of labelling the spectral classes. Pore et al. (1978) reported that All-of-a-Kind labelling (explained, pg 33) improved classification accuracy in agricultural areas. This labelling procedure was used on the Vallecito study area. Every cluster evidently had more than one kind of Type 1 dot included, for each defaulted to K-Nearest Neighbor. It is recommended that future research address the following questions.

1. Due to the heterogeneity of the forested tracts, would majority rule (instead of unanimous rule) better suit forestland statistics labelling (i.e., check all the Type 1 dots in a cluster and label that cluster the identity of the majority of the dots[1])?
2. If K-Nearest Neighbor labelling is used, does weighting for proximity to the mean (explained, pg 58, footnote) have any merit?

A number of questions remain to be answered about ISOCLS. The ISOCLS clustering processor may be seeded with Type 1 dots or with spectral values input by the analyst. The effects of seeding will vary from study area to study area, however the general effects of the proportions (number of Type 1 dots input in each cover type) and numbers of dots input to ISOCLS remain unresolved. It is recommended that the following questions be researched:

1. What are the effects of number of split iterations, number of seeds used, and proportions of Type 1 dots used on CPU time requirements and classification accuracies?
2. Does a proposed technique which might be called dotfile-cluster mean seeding hold any promise? The proposed technique is reviewed below.

The dots used to seed ISOCLS should be representative of the area being clustered so that the 'natural' data grouping in the study area can be readily discerned. Inputting Type 1 dots randomly by cover type may not truly represent the spectral groupings present in the data,

---

1. Ties could be handled in one of two ways:
   a. Default to K-Nearest Neighbor.
   b. Discard dots furthest from the cluster mean until the tie is broken.

especially if Type 1 dot identification errors exist. One possible method of obtaining more truly representative initial cluster centers may be to

a. compile the dotfile information (i.e., locate and identify the Type 1 dots on the study area);

b. cluster the dotfile into an adequate number of spectral classes (adequate meaning perhaps twice the number of major cover type groupings on the study area);

c. label the spectral classes using the original dotfile and the LABEL processor;

d. use the spectral class means of the dotfile to seed ISOCLS.

If the dotfile is large enough (so that most clusters contain 30 or more points) and truly characterizes the study area, ISOCLS might best be used as a grouping processor (LACIE parameters). The effects of iterations on the quality of the training statistics output using this method should be documented.

One of the disadvantages of the Procedure 1 approach is that the entire area, or a systematic subsample of the area, must be clustered in order to develop the unlabelled training statistics. A Monocluster Blocks approach might better utilize computer time while minimizing analyst interaction. Essentially, instead of developing statistics over the entire area, the analyst would select a few training blocks (1600 to 3600 pixels) in heterogeneous areas characteristic of the entire study area. All of the blocks would be clustered simultaneously, the statistics labelled (using LABEL) and the area classified. This approach may combine the CPU time-saving capability of Monocluster Blocks with the automated processing advantages of Procedure 1 while demanding only minimal, additional analyst input. It is therefore recommended that research be conducted to explore the capabilities of a Monocluster Blocks-Procedure 1 approach.

The Multicluster Block approach has been tested by two analysts using the same data set and results are comparable. The McB approach is relatively simple procedure to learn, but does require some analyst experience in order to use it effectively. No future research is necessary to further document this approach. It is suggested that this approach might be used in future research as a standard against which alternate procedures are compared.

Final Statement

The accuracies of classification of the nonforest cover types obtained during this research might be unacceptable in an operational system. Studies have shown that multidate overlays can substantially improve performance. Overlaying data sets may become cost effective, in fact, a standard procedure, when NASA implements software which geometrically corrects Landsat data sets prior to sale.

Improved spatial and spectral resolutions and an increased number of spectral bands (in the visible, reflective, and emissive infrared) will undoubtedly improve classification performances. This satellite platform imagery may be used by automated classification systems which mesh the multispectral information with forest inventory information. The Procedure 1 approach is just one method by which this might be accomplished.

The world's population is currently growing at the rate of 80 million people per year. It has been predicted that by the year 2000 the world's population will have doubled to over 7 billion people. Availability and location of resources take on a new importance. Satellite platform imagery provides the resource analyst synoptic information in a quantitative format directly suited to computer analysis. The marriage of satellite and inventory information (be it forest or agricultural information) in conjunction with automated-computer-aided-analysis techniques offer the possibility of monitoring resources on time schedules on the order of days. Such accurate and timely information may someday be necessary as the world's agricultural areas and forestlands are pushed to their productive limits under environmental conditions less than ideal.

> "Remote sensing from resource satellites provides mankind for the first time with a potential capability for worldwide resource mapping and environmental monitoring in near-real time. The importance of such readily available and objective information about the distribution, quality and exploitation of natural resources can hardly be overemphasized. Man has become accustomed to the uncontrolled use of the environment and has taken for granted the self-renewing and self-correcting ability of nature and the apparent abundance of natural resources. This misconception is coming to an end and we are now beginning to experience the first adverse effects of a deteriorating environment ..."
>
> Kalensky and Wightman (1976).

LIST OF REFERENCES

LIST OF REFERENCES

1. Anderson, J.R., E.E. Hardy, J.T. Roach, and R.E. Witner. 1976. A Land Use and Land Cover Classification System for Use With Remote-Sensor Data. Geological Survey Professional Paper 964, Washington, D.C. 28 pp.

2. Born, J.D. 1977. Renewable Resource Evaluation Data Base and Sampling Design Procedures Handbook for the Rocky Mountain Resource Supply Region. Intermountain Forest and Range Experiment Station, Ogden, Utah.

3. Bryant, E.S., A.G. Dodge, and S.D. Warren. 1978. Satellites for Practical Natural Resource Mapping? A Forestry Test Case. Proceedings, National Workshop on Integrated Inventories of Renewable Natural Resources, Tucson, Arizona. pp. 219-226.

4. Coggeshall, M.E., and R.M. Hoffer. 1973. Basic Forest Cover Type Mapping using Digitized Remote Sensor Data and ADP Techniques. LARS Information Note 030573, Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, IN. 131 pp.

5. Edwards, J.R. 1977. Computer Training Procedures for the Western Washington Forest Productivity Study Utilizing Landsat Data. Proceedings, Symposium on Machine Processing of Remotely Sensed Data, Purdue University, W. Lafayette, IN. pp. 264-269.

6. Ellis, S.L. 1978. Shrubland Classification in Central Rocky Mountains and the Colorado Plateau. Proceedings, National Workshop on Integrated Inventories of Renewable Natural Resources, Tucson, Arizona.

7. Ferguson, R.H., and N.P. Kingsley. 1972. The Timber Resources of Maine. USDA Forest Service Bulletin NE-26, Northeastern Forest Experiment Station, Upper Darby, PA 129 pp.

8. Fleming, M.D., and R.M. Hoffer. 1977. Computer-Aided Analysis Techniques for an Operational System to Map Forest Lands Utilizing Landsat MSS Data. LARS Information Note 112277, Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, IN. 235 pp.

9. Fleming, M.D., J.S. Berkebile, and R.M. Hoffer. 1975. Computer-Aided Analysis of Landsat 1 Mss Data: A Comparison of Three Approaches, Including a 'Modified Clustering' Approach. LARS Information Note 072475, Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, IN. 9 pp.

10. Gialdini, M.J., S. Titus, J.D. Nichols, and R.W. Thomal. 1975. The Integration of Manual and Automatic Image Analysis Techniques with Supporting Ground Data in Multistage Sampling Framework for Timber Resource Inventories, Three Examples. Proceedings, NASA Earth Resources Survey Symposium, Houston, Texas. pp. 1377-1388.

11. Glascock, H.R., Ed. 1976. About the RPA. Journal of Forestry 74(5): pp. 274.

12. Harding, R.A., and R.B. Scot. 1978. Forest Inventory with Landsat, Phase II, Washington Forest Productivity Study. Washington Department fo Natural Resources, Olympia, Washington. 221 pp.

13. Heller, R.C., R.C. Aldrich, R.W. Dunn. 1975. Evaluation of ERTS-1 Data for Forest and Rangeland Surveys. USDA Forest Service Research Paper PSW-112, Pacific Southwest Forest and Range Experiment Station, Berkley, California. 57 pp.

14. Heller, R.C. 1976. Remote Sensors for Airborne and Spaceborne Imagery. Proceedings, Symposium IUFRO Subj. Gr. S.6.05, Oslo, Norway. pp. 37-52.

15. Hoffer, R.M., and M. Fleming. 1978. Mapping Vegetative Cover by Computer-Aided Analysis of Satellite Data. LARS Technical Report 011178, Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, IN. 8 pp.

16. Hoffer, R.M., and Staff. 1975. Natural Resource Mapping in Mountainous Terrain by Computer Analysis of ERTS-1 Satellite Data. Research Bulletin 919, Agricultural Experiment Station, Purdue University, W. Lafayette, IN. 124 pp.

17. Howard, J.A. 1976. Remote Sensing of Tropical Forests with Special Reference to Satellite Imagery. Proceedings, Symposium IUFRO Subj. Gr. S.6.05, Oslo, Norway. pp. 211-226.

18. Husch, B., C.I. Miller, and T.W. Beers. 1972. Forest Mensuration, 2nd edition, Ronald Press Co., New York. 410 pp.

19. Jaakkola, S. 1976. An Automated Approach to Remote Sensing Oriented Forest Resource Surveys. Proceedings, Symposium IUFRO Subj. Gr. S.6.05, Oslo, Norway. pp. 147-156.

20. Kalensky, Z., and J.M. Wightman. 1976. Automatic Forest Mapping Using Remotely Sensed Data. Proceedings, Symposium IUFRO Subj. Gr. S.6.05, Oslo, Norway. pp. 115-137.

21. Kan, E.P. 1972. Data Clustering: An Overview. Technical Report LEC 640-TR-080, Lockheed Electronics Co., Inc., HASD, Houston, Texas.

22. Kan, E.P., and R.D. Dillman. 1975. Timber Type Seqarability in the Southeastern United States on Landsat-1 MSS Data. Proceedings, NASA Earth Resources Survey Symposium, Vol. 1-A: Technical Session Presentations, Houston, Texas. pp. 135-157.

23. Krebs, P.V. and Staff. 1976. Multiple Resource Evaluation of Region 2 U.S. Forest Service Lands Utilizing Landsat MSS Data. Institute of Artic and Alpine Research, University of Colorado, Boulder, Colorado.

24. Lapeitra, G., and J. Megier. 1976. Acreage Estimation of Poplar Planted Areas from Landsat Satellite Data in Northern Italy. Proceedings, Symposium IUFRO Subj. Gr. S.6.05, Oslo, Norway. pp. 157-170.

25. Landgrebe, D.A. 1976. Final Report, NASA Contract NAS9-14015, June 1, 1975-May 31, 1976. Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, Indiana.

26. MacDonald, R.B. 1976. Large Area Crop Inventory Experiment. 2nd Annual William T. Pecora Memorial Symposium, Sioux Falls, South Dakota.

27. Matteson, R. 1978. Personal corrospondence. USDA, Forest Service, San Juan National Forest, Durango, Colorado.

28. Mead, R., and M. Meyer. 1977. Landsat Digital Data Application to Forest Vegetation and Land Use Classification in Minnesota. Proceedings, Machine Processing of Remotely Sensed Data, Purdue University, W. Lafayette, Indiana. pp. 270-279.

29. Mendenhall, W. 1975. Introduction to Probability and Statistics 4th Edition, Duxbury Press, Belmont, California. 460 pp.

30. Miller, R.L., and G.A. Choate. 1964. The Forest Resource of Colorado. USDA, Forest Service, Rocky Mountain Forest and Range Expt. Station, Fort Collins, Colorado, and Intermountain Forest and Range Expt. Station, Ogden, Utah. 55 pp.

31. Minter, R.T., B.E. Wills, and C.T. Gardner. 1977. User Documentation EODLARSYS Earth Observations Division Version of the Laboratory for Applications of Remote Sensing System. LEC-3984, Revision 4, Lockheed Electronics Co., Inc., Houston, Texas.

32. Moritz, T.E., M.D. Pore, and S.S. Yao. 1978. Cluster Parameter Study. Lockheed Electronics Col, Inc., Aerospace Systems Div., Houston, Texas.

33. Nichols, J.D. et al. 1974. ERTS-1 Data as an Aid to Wildland Resource Management in Northern California. Final Report to NASA by Remote Sensing Research Program, University of California, Berkley, California.

34. Pore, M.D., T.E. Moritz, D.T. Register, S.S. Yao. 1978. On Evaluating Clustering Procedures for Use in Classification. Report #LEC-12171, Lockheed Electronics Col, Inc., Houston, Texas.

35. Reeves, C.A. 1978. Procedure-1: Applicability to Rangeland Classification, Final Report. Report #AD-631737-5335-02, Lockheed Electronics Col, Incl, Houston, Texas.

36. Reeves, R. (ed). 1975. Manual of Remote Sensing, Vol. I and II. American Society of Photogrammetry, Falls Church, Virginia.

37. Rohde, W.G. 1978. Potential Applications of Satellite Imagery in Some Types of Natural Resource Inventories. Proceedings, National Workshop on Integrated Inventories of Renewable Natural Resources, Tucson, Arizona.

38. Rohde, W.G., and C.E. Olsen, Jr. 1972. Multispectral Sensing of Forest Tree Species. Photogrammetric Engineering 38(12): 1209-1215.

39. Scholz, D., N. Fuhs, and M. Hixon. 1979. An Evaluation of Several Different Classification Schemes: Their Parameters and Performance. Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, Indiana.

40. Schubert, J.S. 1978. Computer Processing of Landsat Data for Canada Land Inventory Land Use Mapping. Report #13, Gregory Geosciences Ltd., Ottawa, Ontario. 70 pp.

41. Smedes, H.W., K.L. Pierce, M.G. Tanguay, and R.M. Hoffer. 1969. Digital Computer Terrain Mapping from Multispectral Data. Journal of Spacecraft and Rockets 7(9): 1025-1031.

42. Spencer, J.S., and B.L. Essex. 1976. Timber in Missouri, 1972. USDA Forest Service Resource Bulletin NC-30, North Central Forest Experiment Station, St. Paul, Minnesota. 108 pp.

43. Spencer, P.W., and T. Philips (ed). 1973. LARSYS Users Manual, Vol. I-III. Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, Indiana.

44. Stewart, J., and P.J. Aucoin. 1978. Earth Observations Division Version of the Laboratory for Applications of Remote Sensing (EOD-LARSYS) Users Guide for IBM 370/148, Vol. I - System Overview. Report # LEC-12563, Lockheed Electronics Co., Inc., Houston, Texas.

45. U.S. Forest Service. 1974. Photo Sampling Instructions for the Fourth Minnesota Forest Survey. USDA, Forest Service, North Central Forest Expt. Station, St. Paul, Minnesota. 21 pp.

46. U.S. Forest Service. 1978. Reinventory Plan, Pike/San Isabel National Forests, Chapters 5,6, Appendices A, B, and C. USDA, Forest Service, San Juan National Forest, Durango, Colorado. 23 pp.

47. Williams, D.L. 1976. A Canopy Related Stratification of Sourthern Pine Forest Using Landsat Digital Data. Report # X-923-76-188, Goddard Space Flight Center, Greenbelt, Maryland. 10 pp.

48. Williams, D.L., and G. F. Haver. 1976. Forest Land Management by Satellite: Landsat-Derived Information as Input to a Forest Inventory System. Intralab Project # 75-1, NASA, Goddard Space Flight Center, Greenbelt, Maryland. 36 pp.

49. Wills, B.E., C.T. Gardner, and P.J. Aucoin. 1977. 'As Built' Design Specifications for EODLARSYS Procedure 1. Report # LEC-11293, Lockheed Electronics Co., Inc., Houston, Texas.

GENERAL REFERENCES

1. Bishop, B.C. 1976. Landsat Looks at Hometown Earth. National Geographic 150(1): pp. 140-147.

2. Fowells, H.A. (ed). 1965. Silvics of Forest Trees of the United States. Agricultural Handbook # 271, U.S. Department of Agriculture, Washington, D.C. 762 pp.

3. McCormack, D., and D. Witten. 1978. NASA News Press Kit, Landsat C. Release No. 78-22, National Aeronautics and Space Administration, Washington, D.C. 40 pp.

4. Thornbury, W.D. 1969. Principles of Geomorphology 2nd Edition, John Wiley and Sons, Inc., New York, N.Y. 594 pp.