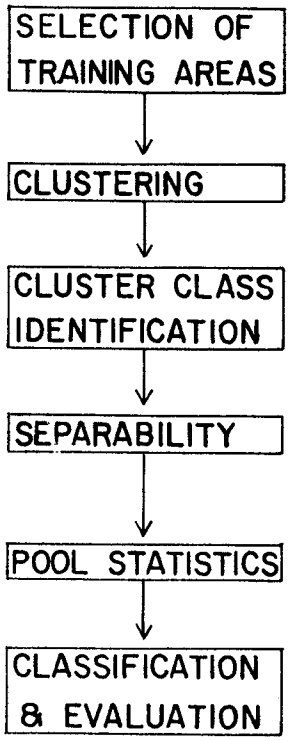
LARS Information Note 072475



Computer-Aided Analysis of LANDSAT-1 MSS Data: A Comparison of Three Approaches, Including a "Modified Clustering" Approach

by
M. D. Fleming
J. S. Berkebile
R. M. Hoffer

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

1975

COMPUTER-AIDED ANALYSIS OF LANDSAT-1 MSS DATA: A COMPARISON OF THREE APPROACHES, INCLUDING A "MODIFIED CLUSTERING" APPROACH*

M. D. Fleming, J. S. Berkebile and R. M. Hoffer

Department of Forestry and Natural Resources
in cooperation with the Laboratory for Applications of Remote Sensing,
Purdue University, West Lafayette, Indiana

I. ABSTRACT

Three approaches to computer-aided analysis of LANDSAT-1 MSS data were evaluated utilizing data from a test site in rugged, mountainous terrain. The approaches compared include non-supervised (clustering), modified supervised, and modified clustering. Test field results indicated classification accuracies of 78.5%, 70.0%, and 84.7%, respectively for the three analysis techniques. The modified clustering approach proved to be the optimal computer-aided analysis technique of those tested because of minimal computer time required, highest classification accuracy, and most effective analyst/data interaction. A detailed description of this analysis technique is included.

II. INTRODUCTION

Over the past decade, tremendous progress has been made in the development of computer-aided analysis techniques (CAAT) involving the application of pattern recognition theory to multispectral scanner data. "Supervised" analysis techniques, involving a training sample approach, and "non-supervised" or clustering techniques have been used with considerable success (Phillips, 1973). However, difficulties are often encountered in relating the cover type categories to the spectral classes present in the data from areas of complex vegetation types and rugged terrain. For example, the supervised approach requires the analyst to select homogeneous training samples which would represent all possible variations in spectral response for each cover type. In the mountainous terrain of the San Juan Mountains of southwestern Colorado, selection of such a training data set proved extremely difficult because of the spectral differences caused by variations in slope and aspect, as well as to the many spectral differences in the cover types themselves.

With the non-supervised approach, the analyst must specify the total number of spectral classes into which the data is to be grouped. The complexity of the study area required such a high number of individual spectral classes that identification of each spectral class proved extremely difficult. It was therefore essential that a more effective procedure be defined to accurately map forest and other cover types when utilizing the LARSYS computer software system and LANDSAT-1 MSS data obtained over a spectrally complex area, such as the Rocky Mountains of Colorado. The objective of this study was to develop a more effective analysis technique and compare the classification accuracy obtained against the more standard supervised and non-supervised approaches.

III. TEST SITE AND DATA DESCRIPTIONS

To compare the three analysis techniques, the Ludwig Mountain study area (15,140 hectares) was selected. The area provides a suitable test for the three techniques because it involves a mountainous area that is spectrally complex due to the variation in cover types (species and crown closure) and the varying

*The research reported in this paper was supported by NASA Contract NAS 9-14016, NASA Contract NAS 5-21880, and NASA Contract NAS 9-13380.

topography (slope, aspect, and elevation).

The study area includes the entire Ludwig Mountain quadrangle, which is located approximately 25 kilometers east of Durango, Colorado. The quadrangle is approximately 11 kilometers by 14 kilometers, covering 15,136 hectares (37,400 acres) and has rugged terrain with elevations ranging from 2134 meters to 3109 meters.

Located at the southern edge of the San Juan Mountain range, the Ludwig Mountain quadrangle is dominated by Ponderosa pine (Pinus ponderosa) forest, but Douglas fir (Pseudotsuga menziesii var. glauca), Engelmann spruce (Picea engelmanii), and subalpine fir (Abies lasiocarpa) are found at the higher elevations and on steep north slopes. At lower elevations the drier, steep, southern slopes are dominated by Gambel oak (Quercus gambelii), and the valley bottoms are agricultural land (predominantly hayfields).

A LANDSAT-1 MSS data set collected Sept. 8, 1972 over the Ludwig study area was free of clouds and snow, and therefore was utilized for the computer-aided analysis. The MSS data (Scene ID 1047-17200) were corrected (Anuta, 1973) to produce a 1:24,000 geometrically correct map when displayed as line printer output. The support data set, or "ground truth", used to aid the analyst included; (1) high-altitude, WB-57F, color infrared photography (1:120,000 scale), (2) 1:24,000 scale forest type map and (3) ground observations by INSTAAR (Institute of Arctic and Alpine Research, University of Colorado) and LARS personnel. Personnel from INSTAAR developed and ground checked the type map. They also utilized this type map and the aerial photos to define the test areas used to quantitatively evaluate the classification results.

IV. BASIC APPROACHES

In utilizing the LARSYS software for analyzing multispectral scanner data, the general procedure normally followed involves:

1. Definition of a group of spectral classes (training classes);

2. Specifying these to a statistical algorithm which calculates defined statistical parameters;

3. Utilizing the calculated statistics to "train" a pattern recognition algorithm;

4. Classifying each data point within the data set of interest (such as an entire ERTS frame) into one of the training classes; and finally,

5. Displaying the classification results in map and/or tabular format, according to the specifications of the analyst.

During the past few years, experience at LARS has shown that there are many possible refinements in the methodology utilized by the analyst for obtaining training classes (step 1 above), while the rest of the procedure varies little from one analysis task to another. The most common techniques for defining training classes involve the "supervised" approach and the "non-supervised" (clustering) approach.

In the "supervised" approach, the analyst selects areas of known cover types and specifies these to the computer as training fields, using a system of X-Y coordinates. The statistics are obtained for each cover type category. The data are then classified, and the results evaluated. Because the analyst has defined specific areas of known cover types for computer training, such classifications are referred to as "supervised".

The second method uses a clustering algorithm which divides the entire training area into a number of spectrally distinct classes. The analyst must specify the number of spectral classes into which the data will be divided. The spectral classes defined by the clustering algorithm are then used to classify the data, but at this point the analyst does not know what cover type is defined by each of the spectral classes. Normally, after the classification is completed, the analyst will identify the cover type represented by each spectral class using available support data, such as cover type maps. Because the analyst need not define particular portions of the data for use as training fields, but must only specify the number of spectral classes into which the data are to be divided, a classification using this procedure is called "non-supervised". Because of the difficulty in knowing how many spectral classes are included in a single species or cover type, previous work (Hoffer, 1974) had indicated that the non-supervised approach was usually more satisfactory

when analyzing MSS data obtained over wildland areas.

Additionally, two variations of these basic methods for defining training classes have been developed. One is to select training areas of known cover type (a supervised approach up to this point), but then utilize the clustering algorithm to refine the data into a number of unimodal spectral classes for each cover type. This method will be referred to as a "modified-supervised" approach. The second variation involves designating small blocks of data (30-60 lines by 40-60 columns) to the clustering algorithm and then identifying each spectral class within these small "cluster training areas". The statistics for the desired informational classes are then formulated by combining spectral classes from the several cluster training areas. This last method is called the "modified non-supervised" or "modified-clustering" approach, and is later described in greater detail.

Three of the four methods described above were used to obtain training classes for the Ludwig Mountain quadrangle using LANDSAT-1 data. The supervised approach (manual selection of training fields) was not used because of the extreme spectral variation within and between cover types in the Ludwig Mountain quadrangle, as indicated by multimodal classes within each cover type (i.e., deciduous, agricultural, etc.). Such spectral complexity adds to the spectral overlap between cover types, and as mentioned, previous work suggested that the manual approach would not yield satisfactory results for this complex region.

The Ludwig Mountain quadrangle was specifically selected for development of a satisfactory analysis procedure because it is a topographically complicated area which contains a wide variety of cover types. Therefore, if an efficient analysis technique could be defined for this for this area, it seemed reasonable to assume that the same technique would also be suitable for other, less difficult, analysis areas.

To evaluate each method's performance and to prevent possible bias in evaluation, 34 test areas were located by personnel from the Institute of Arctic and Alpine Research (INSTAAR), University of Colorado, prior to initiation of the analysis. These test areas included 659 LANDSAT-1 resolution elements within the quadrangle.

V. COMPARISON OF ANALYSIS TECHNIQUES: CLASSIFICATION PERFORMANCE

NON-SUPERVISED APPROACH

Using the non-supervised approach, training classes for the Ludwig Mountain quadrangle were obtained by means of the clustering algorithm which was instructed to define 10 spectral classes. After the 10 spectral training classes were generated the analyst needed to relate the spectral classes to the cover types. To do this, each spectral class was identified using the vegetation map supplied by INSTAAR and color infrared aerial photography. The classification was then evaluated using the test fields previously defined. For the non-supervised approach, the test fields indicated an overall accuracy of 76.6% (Figure 1).

A comparison between the computer printout "map" of the area and the type map revealed that 10 spectral classes were not sufficient. Some spectral classes represented more than one cover type, and some cover types were represented by more than one spectral class. Most of the misclassification error was caused by single spectral classes that represented more than one cover type. In particular, there were two spectral classes that each represented coniferous forest in one location and deciduous forest in another. It could also be seen that cover types that represented less than 5% of the area (including water, cloud, cloud shadow, and bare rock) were not effectively separated from other classes by the clustering algorithm. For example, water, cloud shadow, and one forest type were included in a single spectral class. To obtain reasonably accurate classification results, one spectral class should not represent more than one cover type. Therefore, in an attempt to alleviate this problem, the number of spectral classes was increased from 10 to 20.

Non-supervised classification using the 20 spectral classes yielded a test field performance of 78.5% (Figure 1). The tabular results showed that there were still several spectral classes that represented more than one cover type. Most of the error was caused by confusion between coniferous forest and deciduous forest, and between coniferous forest and agricultural land. Comparing the classification and the type map showed that the confusion was primarily due to different crown closure densities in the coniferous forest. Because of the relatively large

variance in all the spectral classes, the low density coniferous forest was being identified as either grass (agricultural land) or oak (deciduous forest). This indicated to the analyst that even more spectral classes were needed, but it was already difficult to identify the actual cover type associated with each of the classes. Using additional spectral classes to reduce the variance would have made identification of the many spectral classes even more difficult. Therefore, another approach was required to achieve better spectral representation of the cover types.

MODIFIED SUPERVISED APPROACH

The next technique tested was the "modified supervised" approach for obtaining training statistics. The coordinates for training fields were determined by overlaying a geometrically-corrected, 1:24,000 computer printout of a single channel of LANDSAT data onto a type map of the same scale. To statistically define each cover type, training fields for each type were selected throughout the area. The histograms generated for each cover type showed multimodal distributions. Since such distributions violate the basic assumption of the LARSYS perpoint classifier (a maximum likelihood algorithm, based on Gaussian distribution of the data), the training fields had to be modified before classifying the data. To do this, the clustering algorithm was used.

All training fields for one cover type were clustered as a group. The exact number of spectral classes into which each cover type was separated depended on the cover's variability (i.e., more variation required, more spectral classes to be defined). Most cover types had to be clustered into four or five spectral classes which appeared to correspond to variations in slope, aspect, and crown closure. After the training statistics had been adequately defined, the entire data set was classified using the standard maximum likelihood algorithm. The test fields used for quantitative evaluation of the results were the same in each of the analysis procedures tested. Using the modified-supervised approach, the test field results indicated a classification accuracy of 70.0% (Figure 1).

The classification had considerable misclassification between deciduous forest, coniferous forest, and agricultural land. This error was primarily due to the confusion between low density coniferous forest, deciduous forest and agricultural land, and was the same type of error that occurred in the non-supervised approach. With this modified-supervised technique, selection of training fields which contained a representative sample of the many spectral classes present was difficult because of the cover type and topographic complexity of the test site. Thus, the effectiveness of the modified-supervised technique was primarily limited by the large spectral variation within the test site, rather than by the difficulty in identifying numerous spectral classes which was the major problem encountered with the non-supervised approach. Since the modified-supervised approach had a lower test field result than the non-supervised technique, it appeared that yet another approach would need to be defined and tested.

MODIFIED CLUSTERING APPROACH

A "modified clustering" method, which is essentially a hybrid of the supervised and non-supervised methods, was the next approach utilized. In this method, several small training areas were designated, each of which contained several cover types. Each area was clustered separately, and the spectral classes for all cluster areas were subsequently combined. In essence, the modified cluster approach entails discovering the natural groupings present in the scanner data, and then correlating the resultant spectral classes with the desired informational classes (cover types, vegetative condition, and so forth).

Again, after the training statistics had been defined, the maximum likelihood algorithm was utilized to classify the entire data set. Qualitative evaluation of the results using this method indicated that the classification map of the Ludwig quadrangle closely resembled the cover type map prepared by INSTAAR. To obtain a quantitative evaluation of the classification, the same test field coordinates used previously were once again utilized. These test field results, indicated an accuracy of 84.7% (Figure 1), which was a substantial increase in accuracy over either of the previously tested approaches.

Detailed analysis and comparison of the classification maps obtained by each of the three training methods tested indicated that the modified clustering procedure was most satisfactory for obtaining the training spectral classes. This detailed evaluation substantiated the quantitative test field results shown in Figure 1. To permit more effective utilization of the LARSYS software system the following discussion describes this particular analysis procedure in enough detail to allow a remote sensing researcher to classify a data set using this analysis technique.

VI. DETAILED DESCRIPTION OF THE MODIFIED CLUSTER TECHNIQUE

Modified cluster is an efficient and effective technique for defining training statistics. It is essentially a hybrid of the supervised and non-supervised training approaches, and overcomes many of the disadvantages inherent in both of these other techniques. Supervised training is limited by the unknown relationship between categories of importance and spectral classes. Non-supervised training is suboptimal since the analyst must estimate and specify the number of spectral classes present in the data. Also, numerous spectral classes are usually required which makes proper interpretation of the results extremely difficult. This hybrid technique, modified cluster, overcomes these obstacles by allowing a more effective analyst/data interaction. Modified cluster requires less computer time to develop training statistics (Table 1) and produces statistics which yield higher classification performance (Figure 1).

Modified cluster is comprised of four basic steps including:

* Step 1 - define training areas dispersed over the entire study site, with three to five cover types present in each training area;

* Step 2 - cluster each training area separately, compare map with support data, and recluster if necessary;

* Step 3 - combine the results of all training areas, using the separability algorithm, and develop a single set of training statistics; and

* Step 4 - classify the training areas as a preliminary test of training statistics, modify statistics deck if necessary, and classify the entire study site.

The following paragraphs will discuss each of these steps in detail.

SELECTION OF TRAINING AREAS

The basic goal when selecting training areas is to obtain a representative sample of all spectral classes present in the study area. To do this, a representative sample of each cover type, including spectral subclasses caused by variations in slope, aspect, and crown density, must be included in at least one but preferably two training areas.

Selection of training areas throughout the entire study area provides a better sample of each cover type and lessens the problems encountered in extrapolating the training statistics to the entire data set. Since each cluster class must be accurately identified, informational support data of good quality (e.g., maps and aerial photography) must be available for all selected training areas. Classification accuracy is heavily dependent upon the precision with which the cluster classes were identified and described. Thus, the more accurate the identification of the spectral cluster training classes, the more accurate the final classification. Selecting training areas that have a precisely locatible feature such as a lake, rock outcropping, etc., allows easier and more accurate correlation between the support data and cluster classes.

Experimentation with different LANDSAT-1 data sets has indicated that the optimum size for training area is approximately 40 lines by 40 columns (1600 pixels or LANDSAT resolution elements). This size area was large enough to yield approximately 100 pixels per spectral class, yet was small enough to be clustered relatively quickly.

Experimentation also indicated that selecting and clustering a training area with three to five spectrally similar cover types optimized the spectral separability between these cover types. Additionally, this procedure indicated whether the various cover types of interest could be defined on the basis of their spectral reflectance. In other words, if a single spectral class was identified as representing several different cover types, a clear relationship did not exist between the spectral classes present and the cover types of interest.

CLUSTERING

The MSS data for each training area are clustered into a number of spectral classes, independent of all other training areas. In this manner, a greater number of spectral classes are obtained, and the amount of computer time required is greatly reduced (as compared to clustering all training areas together). Table 1 shows the comparison between clustering seven training areas separately and clustering all of them together. Through separate clustering, the computer time is reduced by nearly 86%, and the number of spectral classes is increased from 30 to 76. Although there may be

some duplication of spectral classes when clustering independently, these can be easily identified and grouped. More importantly, any classes that represent mixtures of several cover types or pixels that are on the edge between cover types can be identified and deleted without significantly reducing the number of spectral classes.

The number of cluster classes into which each area is divided varies as a function of the data variability. A comparison of several parameters which may be used to help choose the proper number of clusters indicated that the parameters were closely related. These parameters included; average transformed divergence, highest minimum transformed divergence, total variability of all cluster classes, and a transformed scatter ratio (Sinding-Larsen, 1974). The transformed scatter ratio, which estimates how well the data are divided, was used in this study to select the "optimum" number of cluster classes for a training area. Each training area is clustered into 12 through 16 classes, and the transformed scatter ratio is calculated for each number of classes. The optimal class number is selected by minimizing the transformed scatter ratio. If the miximum number is 12 or 16, the transformed scatter ratio is then calculated for the next cluster class number (e.g. 11 or 17, respectively). This process continues until a minimum scatter ratio is found.

After the "optimum" number of cluster classes is found for a training area, each cluster class must be identified as to the actual cover type it represents, by overlaying the cluster map with the support data. Figure 2 is an example of a training area cluster map that has been overlayed with a cover type map. In this case, the cover type map was obtained by interpretation of color infrared aerial photography. The aerial photography could be used directly by projecting the photography onto the cluster map using an overhead projector, zoom transfer scope or vertical sketchmaster. By using the aerial photography directly, precise and detailed information could be obtained for each cluster class than by simply using cover type maps.

POOLING STATISTICS

Because several statistics decks are produced by clustering the data from each training area separately, the separability algorithm is used to combine the cluster classes into the informational-spectral classes of the final statistics deck. The saturating, transformed divergence value (obtained from the separability algorithm) is a measure of the distance between classes in multidimensional space. This measure, which ranges in value from 0 to 2000, is referred to as the "divergence value." Higher divergence values indicate class pairs which are more separable. Past experience of LARS researchers suggests that class pairs with divergence of 1700 or greater will generally yield a bimodal distribution when grouped (which violates the basic assumption of the maximum-likelihood, Gaussian classifier).

Since a large number of cluster classes are usually obtained by clustering each area independently, simultaneous comparison of all class pairs with divergence values less than 1700 is difficult. For this reason, the combining of similar cluster classes is performed in a series of steps. The first step is to calculate the divergence value for each pair of cluster classes. Because cover types are included more than once in the many training areas, there should be several similar, spectral classes for each cover type. We found that combining all pairs with a divergence value of 1000 or less reduced the number of cluster classes by nearly one-half. The low divergence value of 1000 indicated that the spectral classes for that pair were very similar. To distinguish these combined classes from the original cluster classes, the combined classes will be referred to as "spectral classes."

The second step in combining the classes is to calculate the divergence value for each pair of spectral classes. In this step, all spectral class pairs with a divergence value of 1500 or less are combined. The value of 1500 was selected because there are usually still too many pairs with a divergence value less than 1700 to allow easy grouping of the spectral classes (and not many below 1200). When combining the spectral classes, the cover type is checked for

each cluster class included in the spectral class grouping. Any spectral class with more than one cover type present (mixed cover types) is deleted unless the mixed class is a desired informational class. The combined spectral classes are then identified and named, and consequently are called spectral-informational classes.

The process of calculating divergence values and combining classes is repeated several times until the desired separability is achieved between the spectral-informational classes. If more detail is desired for one or more cover types, it may be desirable not to combine some spectral-informational classes and therefore accept misclassification between these classes. This is where the objectives of the analysis become important in deciding the disposition of these classes.

TEST TRAINING STATISTICS

As a final check before classifying the entire study area (and to test of the training statistics), the training areas should be classified. The classification results can then be compared with the support data to make sure no errors were made in labeling classes or that any desirable classes were deleted. If no errors were made, the entire study area can now be classified.

VII. SUMMARY AND CONCLUSIONS

The non-supervised (clustering) analysis procedure was tested, using first 10 and then 20 spectral classes for classification. These classifications yielded test field accuracies of 76.6% and 78.5% respectively. Observation of the tabular results suggested that an insufficient number of spectral classes were utilized in the classifications since many of the spectral classes represented more than one cover type. This was true even when 20 spectral classes had been specified. Increasing the number of spectral classes during clustering would have made interpretation of these into spectral-informational classes an extremely difficult and time consuming task. Therefore, clustering with greater than 20 spectral classes was not attempted.

The modified-supervised approach provided a classification accuracy of 70.0%, a considerably lower performance when compared to the two other approaches investigated. Errors were caused primarily by inadequate representation of the desired cover types by the spectral classes. This occurred because the modified-supervised approach did not enable the analyst to obtain a representative sample of the spectral subclasses within each cover type, particularly for the complex mountainous area involved in this investigation.

The modified-cluster method proved to be the optimal analysis procedure among the various techniques tested in this study because it resulted in considerable improvement in several phases of this analysis, including personnel time, computer time, and classification accuracy. Not only were the test field results considerably higher (84.7%), but a detailed comparison between the computer classifications and the type map indicated even more conclusively that the modified-cluster approach yielded the best classification results. Further testing on additional data sets has further proven that this modified-clustering technique is an effective and valuable tool for computer-aided analysis of LANDSAT-1 data, particularly for geographical areas that are spectrally complex due to the presence of a large variety of cover types and terrain features.

VIII. REFERENCES

Anuta, P. 1973. "Geometric Correction of ERTS-1 Digital Multispectral Scanner Data". LARS Information Note 103073, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

Hoffer, R. M., M. D. Fleming and P. V. Krebs. 1974. "Use of Computer-aided Analysis Techniques for Cover Type Mapping in Areas of Mountainous Terrain". LARS Information Note 091774, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

Phillips, T. L. (ed.) 1973. "LARSYS User's Manual." Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

Sinding-Larsen, R. 1974. "A Computer Method for Dividing a Regional Geochemical Survey Area into Homogeneous Subareas Prior to Statistical Interpretation". Geological Survey of Norway, Trondheim, Norway.

CLUSTERING OF TRAINING AREAS	TOGETHER	SEPARATELY
NUMBER OF PIXELS	7844	7844
NUMBER OF TRAINING AREAS	7	7
NUMBER OF SPECTRAL CLASSES	30	76
COMPUTER TIME (MINUTES)	68.1	9.7

Table 1. Comparison between the non-supervised and modified cluster methods for defining training statistics.

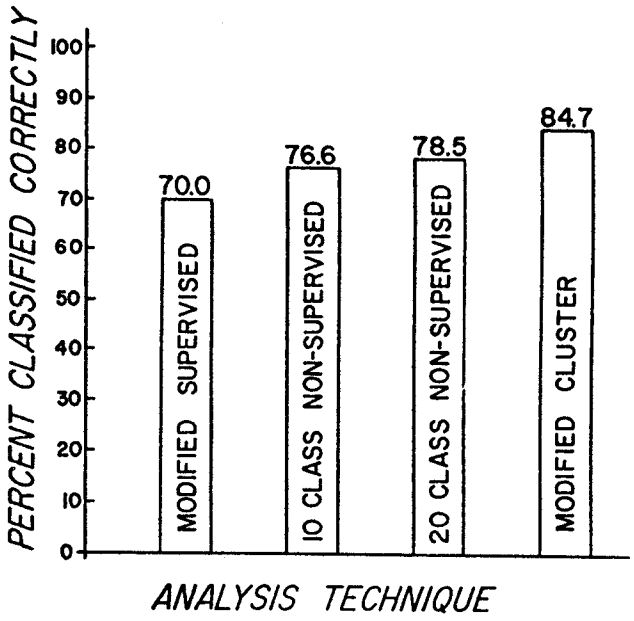
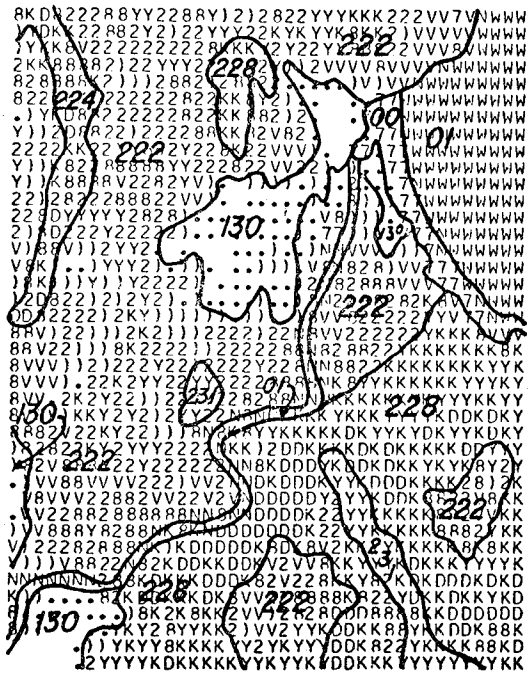


Figure 1. Classification performances of the same LANDSAT-1 data set for four analyses using three different analysis techniques. The values denote the percentage of the data points correctly classified for four cover types including agriculture, water, and deciduous and coniferous forest.



LEGEND

PRINTOUT SYMBOL	TYPE MAP SYMBOL	IDENTIFICATION
7	00	BAREROCK
N	01	RIVER WATER
W	01	RESERVOIR WATER
)	130	GRASSLAND
.	222	PONDEROSA PINE
8	222	PONDEROSA PINE
)	222	PONDEROSA PINE
2	222	PONDEROSA PINE
V	222	PONDEROSA PINE
Y	231	DOUGLAS/WHITE FIR-ASPEN MIX
K	224	DOUGLAS/WHITE FIR
D	228	DOUGLAS/WHITE FIR

Figure 2. Type map from photo-interpretation of support photography overlayed with cluster "map" of LANDSAT-1 data. The analyst utilizes this overlay to determine what informational classes are represented by each spectral class (one spectral class per computer symbol). Spectral classes which denote more than 1 cover type are deleted. This process is duplicated for each training area.