Seventh International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Range, Forest and Wetlands Assessment

June 23 - 26, 1981

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

AN EVALUATION OF ISOCLS AND CLASSY CLUSTERING ALGORITHMS FOR FOREST CLASSIFICATION IN NORTHERN IDAHO

LEE F. WERTH

Lockheed/EMSCO
Houston, Texas

I. ABSTRACT

This paper represents one of many studies funded under AgRISTARS to determine how Landsat can contribute to the Forest Service's Renewable Resource Inventory effort as mandated by the National Forest Management Act of 1976. The specific objective of this study was to test clustering algorithms used at Johnson Space Flight Center - ISOCLS and CLASSY.

II. INTRODUCTION

Forest and range managers require accurate and timely vegetation inventory information such as species composition, density, productivity, and condition to make sound management decisions. Data gathered by satellite multispectral scanners and analyzed using various classification schemes may help achieve the inventory goals of the AgRISTARS Program. Earth Observations Division-Laboratory for Applications of Remote Sensing System (EOD-LARSYS) has several classifier options that can be evaluated to find the best one for forest classification.

LARSYS was developed by the Laboratory for Application of Remote Sensing (LARS) to analyze multispectral scanner (MSS) digital tapes. The current version of the software is known as EOD-LARSYS because of the addition of new techniques and modifications of existing techniques. A test of ISOCLS and CLASSY clustering algorithms has been selected as a task of RRI to evaluate the classification schemes for application to forest and rangeland inventories.

A. SCOPE

The scope of this task is to determine how accurately forest and rangeland classes (USGS - Anderson) can be classified using ISOCLS and CLASSY. If one of the algorithms can accurately classify forest classes, it will be a useful classification tool for forest and rangeland inventories.

B. APPROACH

To address this task, the two algorithms will be applied to forest and nonforest classes for one 1:24,000 quadrangle map in northern Idaho. The algorithms to be evaluated include Interactive Self-Organizing Clustering System (ISOCLS) and CLASSY. The classification and mapping accuracies of the classes for ISOCLS and CLASSY were evaluated with 1:30,000 color infrared (CIR) aerial photography. Confusion matrices for the two clustering algorithms were generated and evaluated to determine which one is most applicable to forest and rangeland inventories on future projects.

C. OBJECTIVES

The objectives of this study are: (1) to evaluate ISOCLS and CLASSY for land cover classification, and (2) to determine the classifier to use in other AgRISTARS Program test sites.

III. AREA DESCRIPTION

The study area selected for the ISOCLS-CLASSY comparison is in the Clearwater National Forest (Region 1) of northern Idaho. The Elk River 7-1/2- minute quadrangle (1:24,000) was selected by the Clearwater Forest staff upon the request of AgRISTARS for an area to test various remote sensing systems. The study area is located approximately 83 kilometers northeast of Moscow, Idaho, and represents a complex heterogeneous site of northern Rocky Mountain coniferous forest. The climate, topography (elevation, aspect, and slope), and soils combine to produce coniferous cover types of mixed conifer, western white pine (Pinus monticola), western larch (Larix occidentalis), Douglas-fir (Pseudotsuga menziesii), sub-alpine fir-spruce (Abies lasiocarpa-Picea engelmannii, western red cedar (Thuja plicata), and mountain hemlock (Larix mertensiana). The most abundant one is the mixed conifer type in which a single species is not dominant and the amount of each species in the mixture varies by location. Some of the species that may occur in the mixed type are western larch, Douglas-fir, grand fir (Abies grandis), western red cedar, western white pine, ponderosa pine (Pinus ponderosa, Engelmann spruce (Picea engelmannii), lodgepole pine (Pinus contorta), and sub-alpine fir. Associated species of the larch type may include western red cedar, grand fir, Douglas-fir, ponderosa pine, and western white pine. Grand fir occurs with the western red cedar type and Douglas-fir and mountain hemlock occur with the sub-alpine fir-spruce type. Associated species of the mountain hemlock type include lodgepole pine and whitebark pine (Pinus albicaulis).

Other land cover/land use classes in the quad area include clearcuts, riparian area, urban area (town of Elk River), meadows, and water (Elk Creek Reservoir). The clearcuts which vary in size and composition occur throughout the quad, but are most common on the west half. Most clearcuts have shrubs as the dominant vegetation, but may also have rock outcrops, large cedar stumps or overtopping regeneration of Douglas-fir, sub-alpine fir, lodgepole pine, and Engelmann spruce. The most dominant shrub in some clearcuts is shinyleaf (Ceanothus velutinus) or redstem ceanothus (Ceanothus sanguineaus). Other shrubs that may be dominant or occur with the ceanothus types are elderberry (Sambucas caerulea), willow (Salix spp), Rocky Mountain maple (Acer glabrum), alder (Alnus sinuata), and ninebark (Physocarpus malvaceus). The riparian area is a cottonwood-willow type (Populus trichocarpa-Salix spp) and the meadows contain grasses, sedges, forbs, and shrubs. Some species occuring in the meadows are Idaho fescue (Festuca idahoensis) bluebunch wheatgrass (Agropyron spicatum), sedges (Carex spp), yarrow (Achillea milletolium), goatweed (Hypericum perforatum), wild strawberry (Fragaria vesca), and snowberry (Symphoricarpos albus). There are wetland species of cattail (Typha latifolia), bulrush (Scripus spp), duckweed/waterlily (Lemmna spp/Nypha sp), and willow where Elk Creek empties into Elk Creek Reservoir and along the fringe of the reservoir.

All five of the management units for the Elk River Planning Unit (Palouse Ranger District) are represented in the study quad. The management units are Elk River foothills, granitic uplands, Elk Creek breaks, intermediate mountain slope lands, and high ridge lands. The elevation in the study area varies from 853 meters in the valleys to 1692 meters on Windy Point. The slope gradient varies from 0 to 10 percent in the valleys to 60 percent plus on the mountain faces.

IV. METHOD

The objective of this study was to compare two unsupervised techniques or clustering algorithms on one 7-1/2-minute quad in northern Idaho. The data source was satellite digital radiometric values, stored on CCTs.

A. DATA SELECTION

Digital tapes of the Landsat scene (ID 293317281) of Idaho were ordered from Goddard Space Flight Center (GSFC) under the AgRISTARS Program. August 12, 1977, Landsat digital tapes were selected because they represent the peak of forest development and are high quality tapes that are cloud free. The color infrared (CIR) optical bar aerial photographs that will be used to evaluate classification accuracy were taken August 29, 1978. An orthophoto quad (made with aerial photographs taken September 24, 1975) of the Elk River test site will also be used to supplement the CIR photographs.

B. PREPROCESSING

The Elk River data set represents a 132 square kilometer area that is recorded on 28,731 data points or samples. Each data point represents approximately 0.45 hectare on the ground. Digital tapes from GSFC were sent to Purdue Laboratory for Application of Remote Sensing (LARS) to reformat into a Universal format and to perform a geometric correction of the Elk River data set. Pixel radiance values were resampled using the nearest neighbor rule.

C. PROCESSING

All digital processing was done on the Laboratory for Application of Remote Sensing (LARS) remote terminals located in Building 17 at JSC. The LARS host computer at Purdue University, West Lafayette, Indiana, is an IBM 3031 (formerly a 370). The software to run the hardware is called LARSYS and the version used at JSC is EOD-LARSYS.

Clustering. The two clustering algorithms selected for the study were Interactive Self-Organizing Clustering System (ISOCLS) and CLASSY. Both algorithms have been used at JSC for the Large Area Crop Inventory Experiment (LACIE). Basically a clustering algorithm searches for the inherent separability or structure of the data without any prior knowledge or training. This is opposed to a supervised technique which does require training to separate the data.

ISOCLS is a clustering algorithm that is similar to ISODATA developed by Ball and Hall.[1] If seeded or starting vectors for the desired information classes are not used then the algorithm initializes its own spectral class mean and according to the specified parameters tries to partition the data set into spectral class groups. Unknown samples are compared to see which group they belong to (which spectral class mean vector they are closest to using the city block or L1, distance) or if they require that a new group be formed. After one iteration through the data the mean vectors of the spectral classes are recomputed. The first iteration through the data may terminate with only two or three cluster classes but subsequent iterations will probably produce more cluster classes through a sequence of splitting, combining, and chaining operations. The main difference between ISODATA and ISOCLS is that ISODATA does not have the chaining operation. For a more in-depth description of ISOCLS see Kan.[2]

The number of cluster classes that are produced for a given data set depends, of course, on the data set complexity and selection of parameters. With ISOCLS many parameters can be varied to produce different results. For example, changing STDMAX from 4.5 to 3.0 will result in more cluster classes and setting DLMIN (the distance between cluster centers) from 3.2 to 2.0 will also increase the number of cluster classes produced. Other parameters that can be varied include ISTOP (number of iterations), NMIN (minimum number of samples in a spectral class on the first and next to last iteration), PMIN (minimum number of samples in a spectral class at the last iteration), maximum number of clusters, and percent N (the percentage of stabilized clusters with standard deviations less than the threshold parameter or STDMAX (maximum standard deviation) in the initial split iteration sequence). Since all of the parameters are interrelated, changing any one of them will alter the results; thus, it is difficult to find an optimum set of parameters for a given data set.

CLASSY is a more sophisticated algorithm which alternates maximum likelihood procedures with splitting, joining, and eliminating operations.[3,4,5] CLASSY starts with a model and assumes the data is normally distributed. First the data set is scrambled so that samples may be randomly selected. The algorithm begins with parent clusters and then determines if the distribution of the parent should be broken down into subclusters or be maintained. If the likelihood ratio is higher for the parent then the subclusters, the parent cluster is maintained. The CLASSY algorithm looks at four moments of the mixture density (histogram) to see if the cluster distributions are the normal bell shaped. The moments looked at are mean vector, covariance matrix, skewness (the measure of how symmetrical the tails of the curve are), and kurtosis (the measure of the height of the peak of the curve or the flatness of the curve).

CLASSY is not on EOD-LARSYS, but can be run using the LARS version of IBM Conversational Monitoring System (CMS) 370. The only parameters the analyst can vary with CLASSY are the number of iterations, the smallest cluster size permitted (based on a percent of the data set) and the maximum amount of time the program is allowed to run. CLASSY requires far more subroutines than ISOCLS.

Labeling. After final cluster maps were produced for each algorithm the next step was to use 1:30,000 color infrared (CIR) optical bar camera (OBC) aerial photographs, orthophoto quad and stand maps for assigning information classes to the spectral classes or labeling.

Classification. The spectral class or classes that represented an information class were then assigned an alphanumeric symbol and the resultant map with information classes would then be a classification map.

D. GROUND TRUTH OR REFERENCE BASE

The reference base for evaluating classification accuracy was primarily the 1:30,000 CIR OBC imagery. The optical bar is a panoramic camera that produces exposures on a film strip that covers 3.7

by 59.5 kilometers (at nadir). Since there is so much inherent image distortion in the imagery away from the nadir (optical center of the film), a variety of equal area grids have been developed by the Forest Service Geometronics group in Washington, D.C. for every other exposure of the overlapping stereo pairs. The grid selected for this study contained a 1.01 hectares tick marks.

After a preliminary manual photo-interpretation of the 1:30,000 imagery, any cover types that could not be identified were visited and verified during a field trip.

E. EVALUATION PROCEDURES

The accuracy of the ISOCLS and CLASSY classification maps were assessed with stratified random sampling. The stratum such as grass, cut-over, forest, etc. were delineated on an overlay registered with the orthophoto quad. The number of samples or one pixel test fields in each stratum were determined according to the areal extent (in percent) that each stratum covered on the orthophoto quad. For example, if stratum #1 comprised 15% of the quad then 15% of the samples were taken in stratum #1. The location of the test fields were marked on the classification maps so they could be viewed on a light table after the orthophoto quad (positive transparency) was superimposed on them. The test fields in each stratum were evaluated according to the Landsat and the orthophoto - OBC solution.

Confusion matrices for each classifier were generated and errors of omission and commission were calculated. Overall and class classification accuracy were also calculated. In addition, overall and class mapping accuracy were calculated according to the method used by Kalensky and Scherk.[6] Mapping accuracy differs from classification accuracy (includes only the omission error) since it includes both the errors of omission and commission.

V. RESULTS AND DISCUSSION

A. PREPROCESSING

The general geometric correction produced by LARS resulted in line printer maps that had a systematic error of 2-3 pixels E-W and 1-2 pixels N-S. The greatest error was in the SW corner of the quad. An attempt was made to improve the positional error with a precision registration but the error was not improved because a sufficient number of well distributed points could not be found. The Elk River data set is in a relatively remote area that does not have well defined roads and stream intersections were not reliable because many of the streams were low at the time of the Landsat overpass (August).

B. PROCESSING

ISOCLS. Many runs of ISOCLS on the Elk River quad were made with different parameters but it became readily apparent that the interdependence of the parameters would prevent the search for an optimum set of parameters. Therefore, it was decided that PMIN and NMIN would be set at -4 and 0, respectively, so that even a one-pixel cluster would be retained if it was unique. The maximum number of clusters allowed was set at 60 and N was set at 80 percent based on LACIE studies.[2] The DLMIN parameter was set at 3.2 after it was observed that values below this produced spectral class pairs that were separated by a DLMIN of less than the amount specified.

The criteria for determining the best value of STDMAX was how well the water pixels were separated from slope shadows and whether the number of spectral classes were too numerous to be identified and labeled. Also, the ease of identifying the meadows and the town of Elk River were other criteria used to judge the algorithm's performance. In essence the analyst evaluated the cluster maps according to the data sets recognizable land cover features. The optimum number of spectral classes was found to be between 21-24 since any lower number resulted in water pixels all over the mountain slopes and cluster maps with more than 24 spectral classes did not have well defined meadows.

An attempt was made to separate out the forest cover types but it became obvious that there was not a unique spectral class or classes for any of the forest types. Also it would have been beneficial if the various cut-over areas could be discriminated because of their variability; but, again this was impossible to do in this study. Most of the cut-over areas were clearcuts but there were also selection and shelterwood cuts. Therefore for the Elk River data set the 24 spectral classes were labeled into the information classes of Coniferous Forest, Cut-over, Grass (meadow) and Water.

The overall classification accuracy using the ISOCLS clustering algorithm was 81% and the overall mapping accuracy was

60% (Table 1). The classification and mapping accuracies for Coniferous Forest were 78% and 78%, respectively. Forest class was confused with Cut-over, Grass and Water. Most of the omission error was with Cut-over. It should be noted, however, that some of the spectral classes that have been labeled Cut-over could indeed be low density Forests. The classification and mapping accuracies of the Cut-over class were 72% and 55%, respectively. The omission error was equally divided between Coniferous Forest and Grass. This is not surprising since the area receives high annual precipitation (1143 mm and greater) and the cut-over areas (especially the clearcuts) are quickly filled in with shrubs and grasses. Also standing trees left on the cut-over areas contribute to the confusion with the Forest class. The classification and mapping accuracies of the Grass class were 100% and 50%, respectively. Both Forest and Cut-over contributed to the commission error. The classification accuracy of the Water class was 100% but since some of the Coniferous Forest was classified as water, the mapping accuracy was only 80%. Since this error occurred along the edge of the reservoir, it could have been caused by pixel positional accuracy or shadows near the waters edge.

CLASSY. CLASSY was developed to be used for clustering agriculture crop areas and not wildlands. The largest data set it could previously cluster was LACIE segments (9.3 by 11.1 kilometers) so software changes had to be made to do a quad size data set. Since this was the first time CLASSY had been used in forest classification, the experience gained in agricultural work was used to preliminarily set the algorithm parameters. It was recommended that the number of iterations be between 3 and 7 and the total run time be set at 150 minutes.[7] Three iterations were usually used for LACIE work but since the data set was very heterogeneous, initial trials were run with 4 and 5 iterations. When 6 iterations were used it was found that the algorithm converged better so most of the runs were made with 6 iterations. Compared to ISOCLS the smallest size cluster class that can be maintained is 1 or 2 percent of the data set, i.e., for 28,731 samples, the smalles cluster class at the 1 percent level would be 287 pixels or samples. Theoretically, any cluster class smaller than this would not be retained at the end of the final iteration. Because of LACIE problems of running CLASSY with 1 percent of the scene it was decided to use 2 percent of the scene as the smallest cluster class. Later on the 2 percent cluster size was changed to 1 percent.

At the 2 percent threshold only 6 cluster classes were produced. The Elk Creek Reservoir is between 50 and 100 pixels in area so it was too small to show on the cluster maps. With the 1 percent threshold the number of spectral classes decreased from 6 to 5 and the water was still too small to be separated. The resulting 5 spectral classes were labeled into the information classes of Coniferous Forest, Cut-over and Grass (meadow).

The overall classification and mapping accuracies using the CLASSY clustering algorithm were 77% and 67%, respectively (Table 2). The Coniferous Forest classification accuracy was 12% higher than ISOCLS but the mapping accuracy was 1% lower. Both Cut-over and Grass were confused with Forest but there was more omission error with Grass than was the case with ISOCLS. Grass also contributed to a commission error for the Forest class. The classification accuracy of the Cut-over class was 45% lower than ISOCLS and the mapping accuracy was 31% lower. Most of the Cut-over samples were actually Forest. Pixel positional accuracy and some of the problems already mentioned with ISOCLS contributed to the errors. The classification accuracy of Grass was 7% lower than ISOCLS but a higher commission error resulted in a mapping accuracy that was 19% lower. Spectral response overlap between grass, shrub and trees is believed to result in the commission error for both ISOCLS and CLASSY. The Water class was too small to be discriminated by CLASSY so no water comparison can be made between the ISOCLS results.

Both ISOCLS and CLASSY indicated that there was a grass meadow in the NW portion of the quad but the aerial photos showed that the area was a cut-over area which had grass between the remaining standing trees. The area in question is not like the grass meadows in the southern part of the quad that are frost pocket areas (created by deforestation) that can no longer support tree seedlings.

The Cut-over class is very important to forest managers and any accuracy lower than 70% probably would not be acceptable to them. Recent Cut-over ares have a better change of being detected; especially, clean clearcuts. However, as cut-over areas are filled in with grass, shrub or seedlings, the chances for discrimination are decreased. More work needs to be done in this area since change detection and updating of geographic

TABLE 1. Accuracy Evaluation of Landsat Forest Classification using ISOCLS Clustering Algorithm.

			Landsat				Omissions		Mapping
	Class	A	B	C	D	Total	No.	%	Accuracy
Aerial Photos	Coniferous Forest (A)	100	13	9	6	128	22	22	78
	Cut-over (B)	6	31	6		43	12	28	55
	Grass (C)			15		15	0	0	50
	Water (D)				24	24	0	0	80
	Total Indicated	106	44	30	30	210			
	Total Committed	6	13	15	6				
	Percent Commission	6	30	50	20				
	Overall Classification Accuracy	81%							
	Overall Mapping Accuracy	69%							

TABLE 2. Accuracy Evaluation of Landsat Forest Classification using CLASSY Clustering Algorithm

			Landsat			Omissions		Mapping
	Class	A	B	C	Total	No.	%	Accuracy
Aerial Photos	Coniferous Forest (A)	135	5	10	150	15	10	77
	Cut-over (B)	24	12	9	45	33	73	24
	Grass (C)	1		14	15	1	7	41
	Total Indicated	160	17	33	210			
	Total Committed	25	5	19				
	Percent Commission	16	29	58				
	Overall Classification Accuracy	77%						
	Overall Mapping Accuracy	67%						

information systems will greatly assist land management planning, especially in the future.

Some of the individual forest cover types could have been separated if terrain data had been used. This would be particularly true for the mountain hemlock and the Douglas-fir type.

The classification and mapping accuracy of ISOCLS and CLASSY are not that different except for some of the classes. CLASSY requires more CPU time than ISOCLS per run but more trial and error runs are required by ISOCLS because it has so many parameters that can be varied. ISOCLS in the unseeded mode requires 10 to 20 runs to separate land cover/use classes compared to only 2-3 runs with CLASSY for a given data set. If ISOCLS is used it is recommended that starting vectors (seeded) be used.

VI. CONCLUSION

ISOCLS in a pure unsupervised mode is an ad hoc algorithm that requires many trial and error runs to find the proper parameters such as STDMAX to separate desired information class. On the other hand, CLASSY is a more refined algorithm that tells the analyst more in a single run concerning the classes that can be separated. The major drawbacks to CLASSY are that important forest and range classes that are smaller than a minimum cluster size will be combined with other classes and the algorithm requires so much computer storage that only data sets as small as a quad can be done at one time. However, CLASSY appears to show more promise for forest stratification than ISOCLS and shows more promise for consistency. This study is not conclusive and more research needs to be done comparing the two algorithms in different areas and using any new improvements to either ISOCLS or CLASSY.

VII. ACKNOWLEDGMENTS

The author would like to thank the following Lockheed personnel: Kent Lennington for assistance with CLASSY, and Susanne O'Brien and Pat Aucoin for software development and testing. Thanks also go to Robert Spencer and Irv Johnson of the Clearwater National Forest for field assistance and reference data. Support and comments from NASA technical monitor, Ridgon Joosten, and NFA Program leader, Phil Weber, are greatly appreciated.

VIII. REFERENCES

1. Ball, G. H. and Hall, D. J.: ISODATA, an Iterative Method of Multivariate Analysis of Pattern Classification. Proceedings of the International Communication Conference, Philadelphia, Pennsylvania, June 1966.

2. Kan, E. P. F.: The JSC Clustering Program ISOCLS and Its Applications. Lockheed Electronics Company, Inc., Houston, Texas, Technical Report LEC-0483, July 1973.

3. Lennington, R. K. and Malek, H.: The CLASSY Clustering Algorithm Description, Evaluation, and Comparison with the Iterative Self-Organizing Clustering System (ISOCLS). Lockheed Electronics Company, Inc., Houston, Texas, Technical Memorandum, LEC-11289, March 1978.

4. Lennington, R. K. and Rassbach, M. E.: CLASSY - An Adaptive Maximum Likelihood Clustering Algorithm. Proceedings of the Ninth Annual Meeting of the Classification Society (North American Branch), Clemson, South Carolina, May 1978.

5. Lennington, R. K. and Rassbach, M. E.: Mathematical Description and Program Documentation for CLASSY, An Adaptive Maximum Likelihood Clustering Method. Lockheed Electronics Company, Inc., Houston, Texas, Technical Memorandum, LEC-12177, April 1979.

6. Kalensky, Z. and Scherk, L. R.: Accuracy of Forest Mapping from Landsat Computer Compatible Tapes. Proceedings of the Tenth International Symposium on Remote Sensing of Environment Ann Arbor, Michigan, October 1975.

7. Lennington, K. Personnal communication. June 26, 1979. Lockheed Engineering and Management Services Company, Houston, Texas.

Lee F. Werth received his Ph.D. degree in Forestry from the University of Minnesota. His specialties include the remote sensing of vegetation and wildlife resources using qualitative and quantitative methods. He has done manual interpretation studies of small and large format color infrared aerial photography and digital processing of Landsat data. He is currently with Lockheed's Forestry Projects Office doing digital processing of Landsat data for forest classification and the development of the Land Management Planning Support System for Renewable Resources Inventory under AgRISTARS.