Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

June 27 - 29, 1979

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana 47907 USA

IEEE Catalog No.
79CH1430-8 MPRSD

USING GUIDED CLUSTERING TECHNIQUES TO ANALYZE LANDSAT DATA FOR MAPPING FOREST LAND COVER IN NORTHERN CALIFORNIA

LAWRENCE FOX III, AND KENNETH E. MAYER
Humboldt State University

I. ABSTRACT

Three approaches to computer assisted Landsat multispectral classifications are described. The supervised classification technique enables the analyst to focus on land cover categories of interest. The unsupervised approach uses the statistical properties of the image to identify spectrally pure classes. Guided clustering combines the characteristics of both approaches to develop the maximum number of low variance classes for each land cover category defined.

The application of guided clustering to forest land classification is explained. EDITOR software was used to merge and edit spectral statistics to produce the maximum number of low variance, statistically separable classes. Color-infrared aerial photography was used to assign meaningful forest cover labels to spectral classes of unknown vegetative composition. Classification accuracies were high (91.6% omission, 91.4% commission).

II. INTRODUCTION

Landsat has provided the scientific community the opportunity to acquire digital multispectral data repetitively over extensive regions of the earth's surface. Application of these data to resource management problems has required the development of interpretive methodologies that allow quick, consistent, and accurate extraction of pertinent information.[1] Interpretive methodologies such as guided clustering (also referred to as controlled/modified clustering) have accellerated the use and accuracy of the information gathered from these data.[1-3]

The U.S. Fish and Wildlife Service (USFWS) in cooperation with Humboldt State University and the NASA Ames Research Center (ARC), used Landsat Multispectral Scanner data to inventory forest cover and land condition on the Hoopa Valley Indian Reservation in Northern California. Information gathered from this inventory is being used by the USFWS in their continuing investigation of the declining anadromous fish population within the Klamath and Trinity rivers.

Data analysis was accomplished through the use of the Earth Resources Technology Satellite Data Interpreter and TENEX Operational Recorder (EDITOR) administered through the Institute for Advanced Computation (IAC), an associate of NASA ARC.[4] EDITOR software is a threefold system for interactive image processing. It consists of a series of subroutines which allows the analyst freedom in performing cluster analysis within specified training areas (guided clustering).

The purpose of this paper is to discuss the successful use of guided clustering in defining the maximum number of spectral classes within any one forest cover or land condition category.

III. ANALYSIS TECHNIQUES

Supervised and unsupervised classification techniques are the two commonly recongnized approaches to Landsat multispectral classifications.[5] These techniques have been used successfully in the past with some operational difficulties.[1] The supervised approach allows the analyst the ability to identify training areas on the ground which represent specific land cover/land use categories. Training areas are used to develop sets of multivariate statistics which contain means, variances, and covariances. Statistics generated from these areas are then used to classify areas of unknown vegetative composition. A Gaussian maximum likelihood classifier

is a common algorithm used in this process.

Errors often arise in supervised classifications when variances are high (15-30 digital numbers squared) within a training area. Given a constant Euclidean distance, statistical distances between classes are reduced when variances are high. This results in fewer unique spectral classes being defined and increased spectral confusion. High variances are especially common when supervised classifications are performed on areas of natural vegetation since areas that appear to be single cover types on aerial photographs may actually consist of several spectral classes. The analyst is often forced to accept high variances when using supervised techniques to classify a heterogeneous cover type.

The unsupervised technique uses the statistical properties of the image as the basis for classification. The analyst estimates a reasonable number of spectral classes that will be representative of the study area. Multivariate clustering algorithms are used to assign pixels to the selected spectral classes. The separability or divergence statistics for these classes are evaluated to determine their spectral proximity. If classes are inseparable, clustering will be performed again with fewer classes. This will assure that the maximum number of low variance classes will be defined.

Problems with this technique occur because the analyst must "estimate" a resonable number of spectral classes. If too few classes are choosen initially, there will be a loss of spectral integrity within the classification. The classes defined may actually represent two or more spectral classes. It is difficult to determine from the class statistics (means, variances, and separabilities) whether enough spectral classes have been choosen, as variances are often not high enough to cause alarm (3-8 digital numbers squared). Another problem that occurs with the unsupervised technique is that the analyst may have little concept of the land cover categories represented by the spectral classes isolated. Since no training areas are defined, it becomes difficult to assign meaningful land cover labels to individual, or groups of spectral classes. From our investigation, we found that the number of spectral classes defined was more than twice the number of land cover categories required for the inventory. To alleviate the above problems we employed both supervised and unsupervised techniques.

In this study the classification was approached using a supervised strategy and clustering within training areas, referred to as guided clustering. Training fields were defined for each vegetation category within the study area. Histrograms were were constructed from pixel digital counts in each channel for the training areas defined. A visual inspection of these histograms indicated the probable number of spectral classes present. These training areas contained between 3 and 6 spectral classes each. A minumum distance clustering algorithm was used to create spectral classes (spectral statistics). Swain-Fu distance was used as the separability statistic to insure spectral separation.[6] A separability of 0.0 to 0.45 for any two classes required that clustering be repeated with fewer classes. Statistical modeling suggested 0.45 as the minumum separability needed to classify with an approximate 0.95 probability of correct classification.[7] Clustering was also repeated when class variances were high ($<$10 digital numbers squared). This provided the opportunity to split high variance classes in order to reduce variances and increase classification accuracy. This technique was successful, resulting in the identification of the maximum number of separable classes for each land cover category. It is important to note that high light reflectance categories such as snow, always exhibited a high class variance. If a spectral class had a high variance and was significantly different from the other classes, it was saved and included in the statistics file.

Approximately 10 to 15 training areas containing 50-100 pixels, were selected for each land cover category. Clustering within these areas was performed independently creating a series of statistics files, some of which contained similar spectral statistics. These statistics files for the vegetation cover categories were merged and edited to remove spectral confusion. Separability statistics were analyzed at each step, and separable classes ($<$.45) were retained in the merged file. A class was always deleted when it conflicted with two or more other classes. When a class conflicted with only one class, they were pooled together creating a new spectral class containing the combined spectral properties of the pooled classes (Table 1). If the pooled class was separable from the other classes in the statistics file, it was retained. This process continued until all of the vegetation cover categories had been included.

Table 1. Separability Matrix - Swain-Fu Distance.
Class 1 should be pooled with class 2 creating a new class 1. Class 3 should be deleted as it conflicts with class 4 and 5.

CLASS	1	2	3	4	5
1					
2	0.14#				
3	0.58	0.85			
4	1.76	1.82	0.21#		
5	2.54	2.04	0.34#	0.75	

Upon completion of guided clustering, an unsupervised classification was completed on the same Landsat scene. The spectral statistics from the unsupervised classification were merged with the final statistics created from guided clustering. The merged file was edited to remove spectral confusion. This insured the inclusion of any spectral classes not present in the training areas. Using this technique, it was possible to define a maximum number of low variance spectral classes for the entire study area. The class statistics were used to drive a Maximum Likelihood classfiication for all of the training areas and the classification was printed out in an alphanumeric code at approximately 1:24,000 scale. The shape and location of each training area was preserved on this printout.

U-2, 1:32,500, color-infrared photography was interpreted to determine the exact vegetation cover category at various points within each training area. This detailed photo interpretation enabled us to assign meaningful vegetation cover labels to the spectral classes defined within the training areas (Figure 1). However, spectral classes still existed without vegetation cover labels, as some classes did not appear in the training areas. A large window (100,000 pixels) was selected from the Landsat scene that was representative of the study area and classified with the final statistics. The remaining un-named spectral classes were identified and labeled through further detailed photo-interpretation.

# indicates values below 0.45.

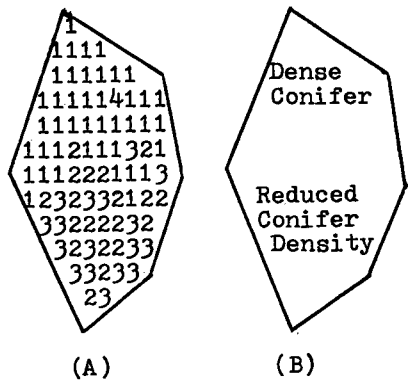

Figure 1. Detailed Photo-Interpretation.
Maximum Likelihood classification results for a typical training area (A). Numbers 1-4 represent Landsat pixels. Vegetation mapping by photo-interpretation for the same training area (B). The analyst would conclude that class "1" represents dense conifer forest, classes "2" and "3" represent a reduction in conifer density, and class "4" is still unknown.

The accuracy of the final classification was evaluated using the previously mentioned U-2 photography. A black line grid produce on clear mylar that represented Landsat pixels, was locally fit to the photograph.[1] Sampling clusters were choosen at random and sampled without replacement. Binominal approximation theory was used to develop error statments.[8] Overall accuracy was 91.6% considering omission errors and 91.4% relative to errors of commission.

IV. CONCLUSION

Guided clustering has provided the means to produce a classification which contained a maximum number of low variance spectral classes. This meant that each spectral class normally represented one or at most very few similar types of vegetative cover. Usually a single category of cover was represented by several spectral classes. Since variances were low and classes were relatively pure, very little spectral confusion was present in the final classification. Guided clustering seemed especially beneficial when classifying complex ecological communities of heterogenous composition.

V. ACKNOWLEDGEMENTS

Funding for this research was provided by NASA Grants 2244 and 2341. NASA Computer systems were made available by the Ames

Research Center through a remote terminal. The personnel of the U.S. Geological Survey, Geography Program at NASA Ames contributed substantially to the computer analysis. We especially thank Willard Newland and Leonard Gaydos for extensive training and consulting concerning EDITOR software. We wish to thank the many people from the NASA Ames Research Center. Dale Lumb, Susan Norman, and David Peterson were administrative and technical coordinators. A special thanks to Buzz Slye and Don Card for their technical and statistical support.

We gratefully acknowledge the creative efforts of Donna Hankins who provided the impetus for this project and the editorial comments and administrative work of Joseph Webster.

VI. REFERENCES

1. Rohde, W.G. 1978. Digital image analysis techniques required for natural resource inventories. Asso. of Federated Information Processing Societies Conference Processings of the National Computer Conference, June 5-8, 1978. Anaheim, CA. Vol. 47. AFIPS Press. Montvale, N.J. p. 93-106.

2. Gaydos, L. and W.L. Newland. 1978. Inventory of land use and land cover of the Puget Sound Region using Landsat digital data. U.S. Geological Survey, Journal of Research. Vol. 6, No. 6, p. 807-814.

3. Fleming, M.D., S.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. Purdue University, Laboratory for Applications of Remote Sensing, LARS Information Note 072475.

4. Anonymous. 1978. EDITOR Handbook. Institute for Advanced Computation Tech. Memo No. 5662. NASA Ames Research Center, Moffett Field, CA.

5. Sabins, F.F. 1978. Remote Sensing Principles and Interpretation. W.H. Freeman and Company. San Fransisco, CA. p. 263-7.

6. Swain, P.H. 1972. Pattern recognition - A basis for remote sensing data analysis. West Lafayette, Ind., Purdue University, Laboratory for Applications of Remote Sensing. Information note 111572. 40 p.

7. Card, D. 1979. (NASA Ames Research Center) Personal communication.

8. Cochran, W.G. 1977 Sampling Techniques, 3rd Edition. John Wiley and Son Inc., N.Y. p. 66.

Dr. Lawrence Fox III, born in Salt Lake City, Utah, recieved his graduate education at The University of Michigan in Remote Sensing and Natural Resources Management. He began his academic career at Humboldt State University where he is currently teaching remote sensing and air-photo interpretation in the Forestry Dept. Through a grant with the NASA Ames Research Center, he is currently applying Landsat digital analysis techniques to forest inventory needs in Northern California.

Kenneth E. Mayer has been working as a Landsat digital analyst on a NASA grant at Humboldt State University. Ken has both Bachelors and Masters degrees in Natural Resources from Humboldt. Since 1973 Ken has been gaining valuable experience working as a biologist with various government agencies. His current job envolves Landsat application in resource management. He is the field biologist responsible for the Hoopa Valley Indian Reservation watershed inventory using Landsat data.