Reprinted from

Symposium on

Machine Processing of

Remotely Sensed Data

June 21 - 23, 1977

The Laboratory for Applications of
Remote Sensing

Purdue University
West Lafayette
Indiana

IEEE Catalog No.
77CH1218-7 MPRSD

THE USE OF ANALYSIS OF VARIANCE PROCEDURES FOR DEFINING GROUND CONDITIONS OF CATEGORIES GENERATED IN AN AUTOMATIC ANALYSIS OF LANDSAT MSS DIGITAL DATA

STEVEN J. DAUS AND MICHAEL J. COSENTINO
University of California

I. INTRODUCTION

Increasing pressure is being applied to resource management agencies to inventory the resources under their jurisdiction in order to formulate appropriate management plans. Present inventory procedures are inadequate when considering budget levels and the tremendous areas to be inventoried. It is necessary, then, to develop inventory procedures which provide for detailed ground information over large areas with adequate update capabilities. Inventory methods are now being developed and refined which integrate information from multiple levels of remotely sensed data and ground data in appropriate statistical frameworks.

This paper describes the methods and statistical procedures, an analysis of variance routine (ANOVA), for relating ground information to spectral classes resulting from the automatic classification of LANDSAT MSS data. The procedures were a portion of a research project completed for the FIRESCOPE program of the US Forest Service.

II. PROJECT OVERVIEW

A. Research Objectives

The objective of the research carried out by personnel at the Remote Sensing Research Program (RSRP), University of California, Berkeley, was to develop the techniques necessary to provide wildland vegetation information to the fuel modeling, fire management, and fire modeling personnel at the U.S. Forest Service Riverside Fire Laboratory. The study was to produce an acre-by-acre vegetation class map of portions of seven Southern California counties through the use of the most recently developed remote sensing techniques. The vegetation class parameters included vegetation type, vegetation density, broad genus association and maturity.

B. Study Site

The study site was located in Southern California and included the coastal areas from Santa Barbara to San Diego and inland to Bakersfield (approximately eight million acres). The vegetation included mixed (or hard) chaparral, soft chaparral, oak-brush, oak woodlands, savannahs, and grass. The study site boundaries are illustrated in Figure 1.

C. Approach and Input Products

A LANDSAT-based integrated inventory approach was formulated in order to overcome the problems of 1) a large, complex study area, 2) detailed ground information requirements, and 3) time and budgetary constraints. By correlating light reflectance data obtained from the LANDSAT system with data on the vegetation obtained from ground and photo sampling, RSRP personnel were able to provide a relatively inexpensive means to map the 8-million acre study area (approximately 6¢/acre) on an acre-by-acre basis. Each acre was mapped according to percent crown cover and maturity of broad genus associations.

Six LANDSAT frames were required for complete coverage of the study area (Figure 1). The digital data were used to generate spectrally based computer classes and as a basic structure for reporting the acre-by-acre information. In addition to the LANDSAT data, conventional color, 35 mm aerial photography was acquired for sampling purposes. Ten photo plots per 1.5 mile flight line were acquired by RSRP staff. Each plot consisted of a medium scale photo (1:8,000) and a simultaneously acquired large scale stereo pair (1:1000). The ground information interpreted from these photos was then used to define the ground condition of the corresponding spectral classes. For limited portions of the area, small scale photography (1:125,000) was available in the form of color infrared (CIR) transparencies.

III. METHODS AND PROCEDURES

The use of the ANOVA routine required special applications of the data handling methods in order to maintain the validity of the statistical tests. The procedures which were affected included; the training of the maximum likelihood classifier, the allocation and interpretation of the large scale

photography, and the association of the photo plots to the spectrally based computer classes. The following will describe the procedures and their relation to the application of the analysis of variance.

A. Unsupervised Training

This procedure determined the degree of spectral similarity necessary to define a spectral type or class. For each of the six LANDSAT scenes, five 100 x 100 pixel training areas representing the range of ground conditions were selected, and the data from these areas were input to the clustering program ISOCLAS[1]. In this program, pixel values with similarity deemed statistically significant were considered as one type (cluster). The program operator established (1) the maximum allowable range of values to be considered in one type, (2) the maximum allowable overlap (if any) between clusters, and (3) the minimum number of pixels per type. The clustering specifications were adjusted until the spatial distributions of the clusters approximated physiographic features which appeared on the small scale (1:125,000) CIR transparencies. The ISOCLAS clusters, Gaussian distributions described by means, standard deviation and covariance matrices were used as training classes. Approximately 50 training classes per scene were developed as a result. This training procedure (unsupervised) was used in order to minimize the bias which could be realized in the ANOVA evaluation procedures. If a supervised approach had been used the computer classes resulting from the classification would have been based on the ground variables being tested and would not have been independent of them during the analysis. The use of the unsupervised approach allowed for a more clear interpretation of the information content of the spectrally separable computer classes (as defined during the ISOCLAS operation).

B. Classification

The classification program CALSCAN was "trained" to recognize spectral classes using the training generated by ISOCLAS, and each pixel was classified. The classification was subjectively evaluated by comparing the spatial distribution of the classes with vegetation type maps and 1:125,000 color infrared photography. The evaluation showed that the spatial distribution of the classes reflected corresponding differences in topography, vegetation type and type mix, vegetation density, and apparent vegetation condition.

C. Defining Ground Conditions for Each Spectral Class

To test the hypothesis that a spectral class represented a particular vegetation class, large scale (1:1000) photo plots were manually interpreted in order to describe the vegetation associated with each spectral class. This was accomplished by (1) matching pixel coordinates to photo plot coordinates, (2) describing and defining the ground conditions of the spectral classes, and (3) statistically testing the hypothesis that the spectral classes were different from one another in terms of their associated vegetational components.

Coordinate Transformation. Each large scale photo plot was located on USGS topographic maps and its Universal Transverse Mercator (UTM) coordinate determined. Previously derived transformation equations, linear regression equations developed using a series of control points for which accurate UTM and LANDSAT coordinates were available, were used to identify the specific LANDSAT pixel cooresponding to each large scale photo plot. The specific computer class was determined for each large scale photo plot by finding the predicted location on the computer printouts and reading the computer code of the pixel as shown in Figure 2. The photo plots were then listed by computer class.

Vegetation Class Definition. Two photo plots per spectral class were selected and interpreted for ground information and then this information was verified and refined by ground sampling. A 96-point dot grid was overlayed on the photo plots and percent crown cover determined for the following cover categories:

1. Barren
2. Grass
3. Mixed chaparral (Ceanothus, Quercus, Arctostaphylos, Rhus)
4. Soft Chaparral (non-woody shrubs)
5. Chamise chaparral (Adenostoma fasciculatum)
6. Conifer
7. Hardwood

The selected photos were then interpreted for one of four maturity classes: pioneer, immature, mature, decadent, based upon a subjective interpretation of stand age, vigor, and living vs. nonliving canopy material. In addition, the geographical aspect of the photo plots were determined from topographic maps.

The ground sampling procedure consisted of locating areas on the ground which corresponded to a selected sub-sample of the photo plots. At each ground plot, the interpreted genus composition was verified, the litter was classified as persistent or non-persistent, and the stand live-to-dead ratio was estimated.

IV. The ANOVA Approach

The size of the area precluded obtaining 100% ground or photographic data so it was necessary to estimate the population parameters by statistically valid sampling and analysis procedures. The large scale stereo pairs (scale 1:1,000) were considered adequate to define ground information to the level required by the Forest Service in the regional context, and were considered as the primary ground samples. Consistent with the assumption of the ANOVA, photo plots were systematically located along a randomly located flight

lines.

To define the ground conditions of the computer classes in terms of several vegetation parameters and test the hypothesis that the classes were statistically different from one another, ie. evaluate the ability of the procedure to yield significantly different classes, a one way analysis of variance[2] was used to complete the analysis and make estimates. The computer classes were viewed as a set of different "treatments" and the response variables were the vegetation parameters defined by the Forest Service.

The ANOVA package DANIEL[3] is designed to carry out the statistical calculations involved in fitting a linear regression equation

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p$$

to a set of n data points. Let the points to be fitted be donated by $(X_{1i}, X_{2i}, \ldots, X_{pi}, Y_i)$, i=1, . . ., n, where X_{ji} is the value of the jth independent variable X_j for the ith observation, (in this case the ground variables interpreted for each computer class) and Y_i is the value of the dependent variable for the ith observation (the computer classes). The program fits the equation to the points by the method of least squares, providing estimates of the regression coefficient B_i, the standard errors of these estimates, fitted values are residuals, t - and F-statistics, multiple and partial correlation coefficients, and various other statistics of interest interest. The package was run on a minicomputer with 64K words of storage in core and a link to a movable head disk with 12 million words of storage.

The photo interpretation results were tallied and punched into the DANIEL format and the ANOVA performed for each scene. Representative outputs for scene B are shown in Figure 3 and 4. The predicted levels of the information types for each group; the color code of each group on the inplace map; and the original computer classes (of the original fifty) which comprise each group are shown in Figure 3. The parameters of the ANOVA model are shown in Figure 4. Note that a separate model was used for each vegetation response variable. An example of the inplace map is not shown in this paper due to the dependence on color for adequate interpretation of its information content.

V. DISCUSSION

The results of this study indicated that 1.) ANOVA approach is a viable method for relating ground conditions to classes generated in a computer classification and 2.) if the approach is to be of maximum benefit changes must be made in the methodology. The most basic problem which affected the validity of the results was the low number of photo observations available for each plot. The low numbers resulted from a complex interaction between factors which included 1) pre-classification allocation and acquistion of the large scale aerial photography, 2) non-stratification of the classification and analysis procedure, and 3) improper understanding and control of the training procedures.

The majority of the large scale aerial photos were spatially allocated at random over the entire area prior to the computer classification. This approach ignored the potential relative sizes and spatial distributions of the computer classes generated during the classification procedure and created a situation where, 1) there was not an appropriate number of samples to adequately describe the classes and/or 2) the spatial arrangement over-sampled some classes and provided no coverage for others.

By not allocating sample plots, training areas or classifying on a stratified basis it was not possible to refine the ANOVA approach. If stratified procedures had been used it would have been possible to conduct the analysis of variance either in separate runs or as a blocked design with the number of photo plots in each computer classes as a definitive criteria.

The use of the clustering approach for training was theoretically valid and necessary due to the previously discussed bias problem. However, the practical results showed that the technique produced too many initial cluster and that it was difficult (if possible at all) to relate training classes developed on one scene to other scenes. It was not possible to control the clustering algorithm adequately to provide detailed separation without producing a high number of clusters. In addition it was not possible to produce a set of representative classes for the entire area, ie. pool all of the photo plots in classes which were the same on the six scenes, because of the differences in acquisition spectral characteristics. The classes had to be aggregated on a scene-by-scene basis and a separate analysis of variance run for each scene. The method used for aggregating the classes was not optimal because it merely worked backward through the clustering procedure generating classes which were spectrally more general, a procedure as equally uncontrollable as the initial breakdown. This particular procedure was used because the alternative procedures would again have biased the analysis of variance by creating aggregation that were defined by the response variables used in the ANOVA.

The accuracy of the equations used to pair the photos plots with their corresponding LANDSAT pixel could also have affected the validity of the ANOVA procedure. If the transformation equations did not consistently locate the photo plots in the proper pixel the homogeneous structure (a basic assumption of ANOVA) of the classes would be suspect. It would not be valid to compare the information from two photo plots if one had been placed in the class due to mislocation. The transformation equations were

refined to produce an average prediction error of one pixel and whether this accuracy is acceptable when applying the ANOVA procedure in this manner requires further study.

VI SUMMARY

The results of this study have shown that analysis of variance procedures (ANOVA) can be used to relate ground or photo plot data to spectrally based computer classes in order to define the ground conditions of the classes. However, careful structuring of the procedures including; training, allocation of samples, classification and locational transformations, is necessary to maintain the validity of the statistical test.

REFERENCE

1. Kan, E. "The JSC Clustering Program ISOCLS and Its Applications", NASA-JSC, LEC-0483. Houston, 1973.

2. Scheffe, Henry. "The Analysis of Variance", John Wiley & Sons, Inc. New York. 1959.

3. Cuthert, Daniel and F. S. Wood "Fitting Equations To Data" Wiley-Interscience, 1971.

4. Steele, R. S., et al. "A 4-Channel Optical Film Annotator for Production of Planimetrically Correct Images from Digital Data" In 9th International Symposium on Remote Sensing of Environment, Ann Arbor Michigan, Oct. 1974.

5. Colwell, R. N., M. J. Cosentino, S. J. Daus, and S. J. Titus. "Southern California Fuels-Oriented Vegetation Mapping Using Multistage Techniques" Final report. USDA, Forest Service Con. # 21-348, Jan 1977.

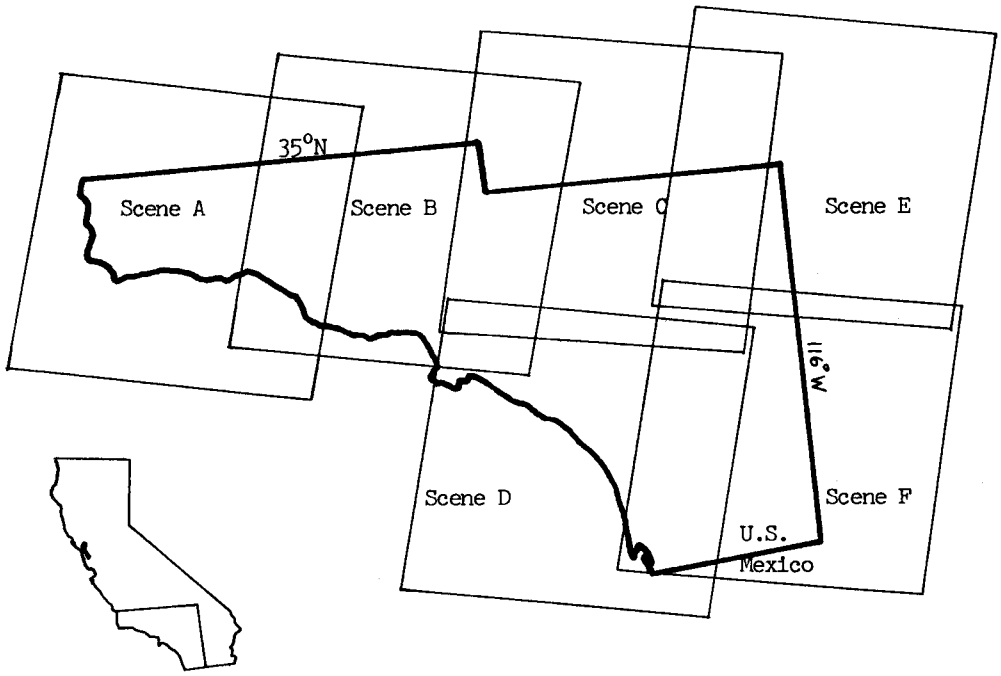

Figure 1. The southern California study site and the arrangement of the LANDSAT frames covering the area. The study site was approximately 8 million acres in size and covered an area in which fire and fuel management is a major consideration.

Figure 2. In order to determine the computer class of each large scale photo plot the LANDSAT coordinate positions were predicted for each plot based on its UTM coordinate. The location of each plot was marked on restricted areas print outs of the classification results and the specific class of the pixel corresponding to each photo plot was recorded.

Table 1. A summary of the remotely sensed data used in this study

IMAGERY	Identification	Data Acquired	Scale
1. LANDSAT MSS DIGITAL			
Scene A	1667-18004	21 May 74	1:1,000,000
B	1702-17535	25 June 74	1:1,000,000
C	1701-17480	24 June 74	1:1,000,000
D	1701-17483	24 June 74	1:1,000,000
E	1700-17422	23 June 74	1:1,000,000
F	1700-17425	23 June 74	1:1,000,000
2. High altitude	CIR trans, 9 x 9	June 1972	1:125,000
3. Large scale	complete strip coverage	June75/March 76	1:8,000
	strip pairs	June75/March 76	1:1,000

Vegetation Class Color & Code		Spectral Classes	AVERAGE (N-2) PERCENT CROWN CLOSURE											
			Barren	Grass	Mixed Chaparral	Soft Chaparral	Chamise	Total Brush	Conifer	Decid. Hrbel	Non-Decid Hrbel	Total Veg.	Maturity*	Aspect
Red	4	1, 16, 31, 42, 49	84.0	5.0	----	11.0	----	11.0	----	----	----	16.0	1.0	Flat
Blue Green	30	2, 17, 32, 50	----	78.0	----	----	----	----	----	----	22.0	100.0	3.0	----
Yellow Brown	15	3, 18	34.0	52.0	2.5	11.5	----	14.0	----	----	----	66.0	1.5	----
Yellow Green	17	33	41.5	9.0	6.5	43.0	----	49.5	----	----	----	58.5	1.0	----
Orange	19	43	44.5	----	37.5	----	18.0	55.5	----	----	----	55.5	3.0	South
Beige	21	4, 19	6.0	6.0	88.0	----	----	88.0	----	----	----	94.0	3.0	North
White	31	5, 20, 34	9.0	84.0	----	7.0	----	7.0	----	----	----	91.0	1.0	----
Pink	40	6, 21, 35	11.5	----	14.0	----	----	14.0	----	----	----	74.5	4.0	North
Brown	2	22, 36, 44	56.0	10.0	----	----	----	----	----	24.0	10.0	44.0	2.0	Flat
Green	7	7, 23, 37	4.0	----	13.5	----	----	13.5	----	----	82.5	96.0	4.0	----
Dark Violet	33	8, 24, 38, 45	28.0	35.0	----	37.0	----	37.0	----	----	----	72.0	3.0	Flat
Blue Green	30	9, 25, 39	----	75.0	----	----	----	----	----	----	25.0	100.0	3.0	Flat
Purple	37	10, 26, 40	24.5	----	40.5	5.0	30.0	75.5	----	----	----	75.5	3.5	East
Yellow	20	11, 27, 46	6.0	----	61.5	----	----	61.5	----	----	32.5	94.0	3.5	----
Brown	2	12, 28, 41, 47	40.0	44.0	----	----	----	----	----	16.0	----	60.0	2.5	----
Black	1	13, 29	4.0	----	61.5	----	----	61.5	----	----	34.5	96.0	4.0	North
Dark Green	5	14	23.0	----	50.5	----	13.5	64.0	13.0	----	----	77.0	4.0	East
Black	1	15, 30, 48	7.5	----	51.5	----	----	51.5	----	----	41.0	92.5	4.0	North
Blue	10	51, 52		WATER										

*Maturity Codes:
1 = Pioneer
2 = Immature
3 = Mature
4 = Decadent

Figure 3. An example of the computer class specific ground conditions generated in the ANOVA procedure. The body of the table contains average predicted crown closure percentages of each type and maturity and aspect ratings. Also shown are the color codes associated with the hard copy inplace map and the original computer class composition of the aggregated groups.

Scene B--ANOVA Results

Vegetation Response Variable	r^2	Degrees of Freedom			Error MSE	Calculated F
		Treatment	Residual	Total		
barren	.8852	17	18	35	11.4855	8.1657
grass	.8485	17	18	35	17.8209	5.9302
mixed chaparral	.9419	17	18	35	9.6451	17.4915
soft chaparral	.8715	17	18	35	6.7926	7.1827
chamise	.6989	17	18	35	7.5664	2.4580
total brush	.9609	17	18	35	8.2782	26.0379
conifer	.4857	17	18	35	1.8333	1.0000
deciduous	.6234	17	18	35	7.2667	1.7524
non-deciduous	.9664	17	18	35	6.6687	30.4686
total vegetation	.8858	17	18	35	11.6774	8.2163
plot maturity	.9535	17	18	35	.3333	21.7059
aspect	.7059	17	18	35	.9129	2.5412

Critcal F Values for Testing the Hypothesis of No Differences Between Vegetation Classes

$F_{17,18,.05} = 2.238$ $F_{17,18,.025} = 2.626$ $F_{17,18,.01} = 3.17$

Figure 4. Descriptive parameters for the ANOVA models run for each vegetation response variable. The amount of linear variability accorded for in the model, r^2,,is generally high however the error mean square is generally high as well. The calculated F indicates whether there is significant differences among the treatment (computer classes) with respect to the vegetation response variable.

Mr. Daus completed a BS degree in Forestry (1969) and an MS in Wildland Resource Science (1971) at the University of California Berkeley. He is presently completing a Ph.D. in Ecology at the University of California Davis. He is also head of the Image Analysis Unit, Remote Sensing Research Program conducting research in the application of human/computer information analysis systems to natural resources inventories.

Mr. Cosentino completed a BS degree in Forestry (1975) at the University of California, Berkeley He is presently the manager of a fuels related vegetation mapping project in Mendocino County, California.