YRB 73-08

DISCRIMINANT OR CLUSTER ANALYSIS

Harold F. Huddleston

Paul D. Hopkins

Statistical Reporting Service
U.S. Department of Agriculture

DISCRIMINANT OR CLUSTER ANALYSIS

This paper discusses the use of SAS for several types of pattern recognition problems which have been studied by the Research Division of SRS. Some results for several small data sets are presented as a means of indicating the type of data and the type of solution obtained. Specifically, we focus on two types of analysis:

(1) The Maximum Likelihood Discriminant Function where the objective is to classify individual data points;

(2) Techniques involving relationships between groups of individual data points which employ both maximum likelihood discriminant methods and clustering methods in a sequential manner.

Maximum Likelihood Discrimination

The applications, which our agency have, typically involve discriminating between natural populations such as:

(1) Fruit on trees from background noise like: leaves, limbs, bark, sky, ground and shadows using ground (sideview) photography;

(2) Fruit trees in an orchard from background noise like: hedge or border trees, ground, water, using aerial photography;

(3) Identify individual field crops from background noises like: woods, native pasture, wasteland, residential areas, parks, etc., using aerial photography or LANDSAT imagery.

The first step involves obtaining mean vectors and covariance matrices for known samples of the target of interest(s) and the individual categories of the background objects. If it appears advantageous to combine overlapping or multi-mode background categories, this is done prior to determining

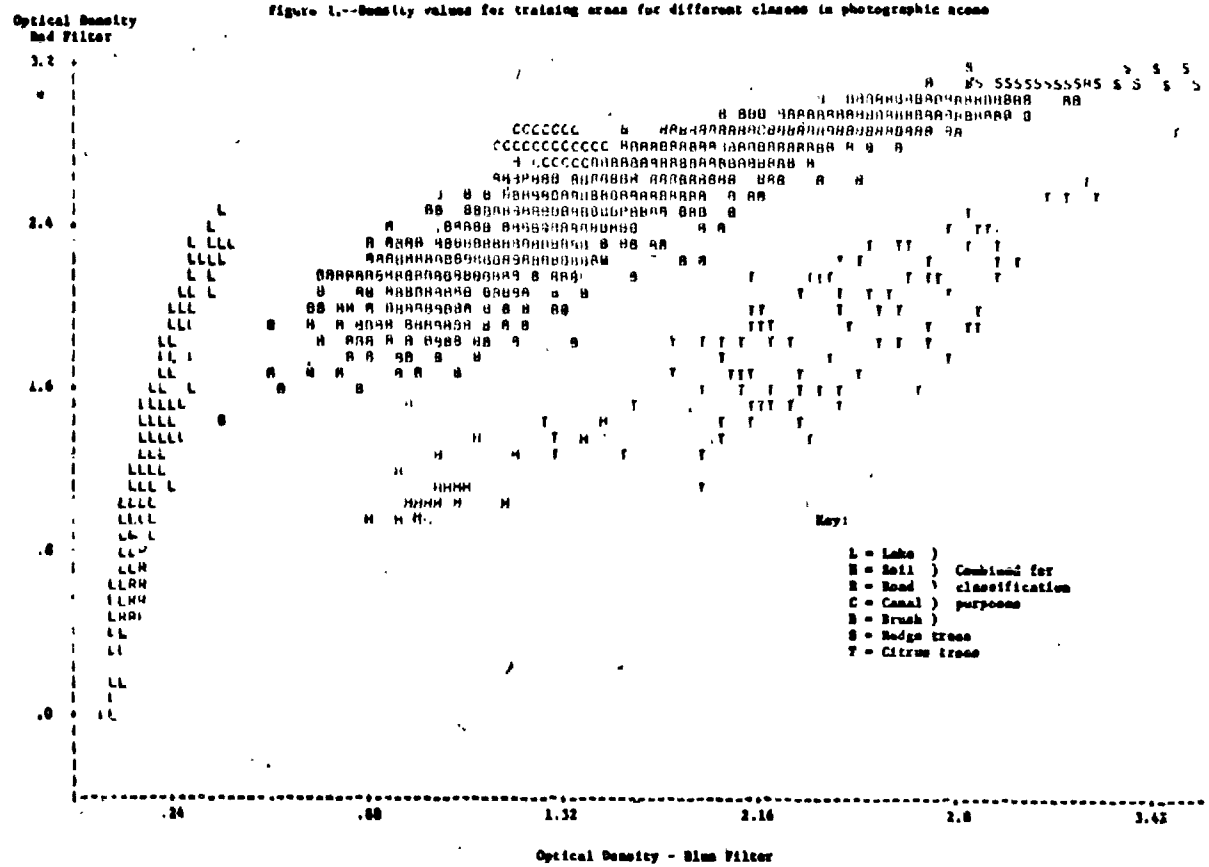
the actual discriminant categories to be used for unknown data points. This decision is based on tests of significance for mean vectors and visual inspection of the marginal distribution to verify that a statistical significance implies "practical" importances. The variance-covariance matrices are also tested for homogeneity to decide whether to employ a linear or quadratic discriminant function. At this point, the known samples are classified and the classification matrix or confusion matrix is obtained to see how good the discrimination is between categories for the known sample of data. The classification of the unknown data elements is made based on the use of prior probabilities supplied from another source in most applications (i.e., where available). Equal prior probabilities are not very realistic when the multivariate distributions for the categories are overlapping.

The first illustrated example indicates some results for a small data set obtained in Texas. Each data point represents a small square with sides of 240 microns from an aerial transparency.

The transparency was 9 inches by 9 inches taken from an altitude of 600 feet. The scale was approximately 1:1800 or the area covered was about 1,350 feet by 1,350 feet. The data analyzed corresponded to an area 140 feet by 420 feet. The film was scanned with a microdensitometer using four filters: red, blue, green, and clear, and these values plus location, in terms of X and Y coordinates, were recorded on tape. Only the red and blue density readings (variables) were used in the final classification.

The spectral behavior of each group (based on a sampling of the scene) is shown in figure 1. The group of interest is the fruit tree pixels which correspond to the capital T's. This plot of the "training data" indicates

Figure 1.--Density values for training areas for different classes in photographic scene



the lake and road pixels are clearly separated from the tree pixels, consequently pixels with density-blue reading less than 23 are deleted from all further analysis. The groups were then reduced to three: (1) fruit trees, (2) hedges, and (3) other. Tests of homogeneity of the within group covariance matrices indicated that quadratic rather than linear discriminant functions should be used. The training data was then reclassified using unequal prior probabilities based on the estimated fraction of area covered by fruit trees, hedge and other. The chosen discrimination procedure was used to classify all data. The classification matrix obtained for the training data is shown in table 1 below with the column totals indicating the pixels classified in each class. The off diagonal entries indicate misclassified data points. The classification for all other pixels in the scene is shown in table 1a.

Table 1--Classification of Training Pixels

Data points selected from	Classification			Total
	Trees	Hedge	Other	
Trees	87	3	0	90
Hedge	2	19	0	21
Other	0	0	1,650	1,650
Total	89	22	1,650	1,761

Table 1a--Classification of Unknown Pixels

Trees	Hedge	Other	Total
2,279	797	17,932	21,008

This result is not very useful in its present form. Consequently, we express it differently; that is, about 10.4% (2,368/22,769) of the data points and, hence, area is occupied by fruit trees. We would prefer to have a count of the number of trees. However, this problem will be discussed in a later section.

A second illustration indicates some results for a small data obtained in Missouri in August 1972. Each data point represents a square with sides of 240 microns. The film was color infrared (2443) and taken from an altitude of 8,000 feet. The scale was approximately 1:40,000 and the data was obtained from a 70mm positive transparency using a microdensitometer. Each data point represented about one acre.

In this example, the interest centers on estimating crop acreage or the fraction of land occupied by each crop. We had 4 crops and 8 other agricultural land uses which were of interest. For two of the spring planted crops, corn and soybeans, each was visibly in two different stages of development. That is, one group represented relatively large (early) plants and the second group small (late) plants. Thus, 14 groups were used. Small data sets from each of these categories were selected for obtaining mean vectors and variance-covariance matrices. Based on these data, we derived four different discriminant functions. The results of

classifying the training data are given in table 2 below. These results are fairly typical of what we have found over a number of states: namely, quadratic discriminant functions with unequal probability give the "best" results. In addition, the estimated fraction of land in each use has less bias. This results from the covariance matrices not being homogeneous and the multivariate distributions overlapping.

Table 2--Percent of Training Data Set Classified Correctly for Discriminant Procedures

Crop type	Linear functions		Quadratic functions		Number data points
	Equal priors	Proportional priors	Equal priors	Proportional priors	
Corn	34.3	83.8	32.3	80.8	99
Soybeans	41.6	92.5	38.9	84.5	226
Hays	54.3	11.4	82.9	60.0	35
Grain sorghum	60	0	90	40.0	10
Other uses	49.5	2.0	78.6	34.1	182
Overall Accuracy 1/	44.0	54.3	54.5	64.9	552

1/ Weighted by data points in each group.

Sequential Discrimination and Clustering

One phase of our research has been to develop a system for identifying and counting objects on photographs. The system has also been used to identify and count objects of interest acquired with a high density of data points so physical shapes could be detected.

The application of this system required an addition to SAS-72 for counting of fruit on trees and counting fruit trees in an orchard from digitized films using a scanning microdensitometer. The first step is to use a discriminant function to classify the data points as part of a data reduction step since clustering techniques prefer small data sets. A clustering technique based on the minimal spanning tree concept (Zahn 1971) has been added to SAS-72 for this purpose. Refer to the first illustration for the purpose of counting the fruit trees indicated by the 2279 data points in table 1a. The application and flow of the procedures are shown in chart 1 for two data sets. In both of these applications we use the spatial properties (size and compactness) of the "target" of interest to develop homogeneous groups. These two tools are utilized sequentially to improve the accuracy of the system.

The classes of objects found in the scene on two of the respective film transparencies were as follows for these two applications:

(a) Counting Oranges

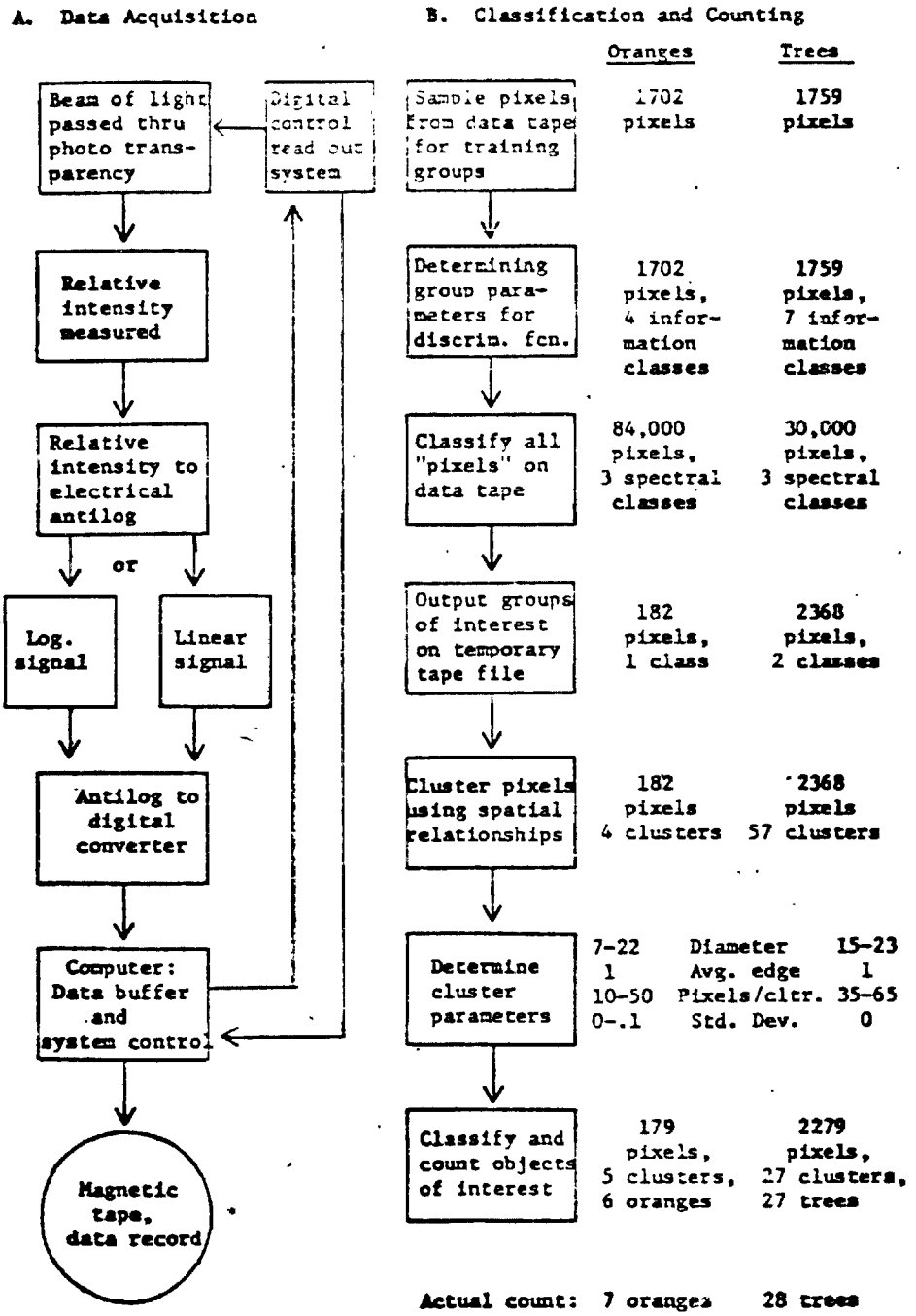
(1) Sky
(2) Ground
(3) Foliage
(4) Oranges

(b) Counting Trees

(1) Lake)
(2) Soil) Combined for classification purpose
(3) Canal)
(4) Road)
(5) Bushes)
(6) Hedge (Trees)
(7) Citrus (Orange) trees

Figure 1 shows the spectral data and separation of the various classes of targets using the red and blue filtered readings obtained from the scanning microdensitometer. The use of different data modes (transmission units versus density units) is also possible since the output analog signal may be either logarithmic or linear with the microdensitometer used. For these two examples, two-dimensional feature selection indicated obtaining the digital data in transmission values for oranges and density values for trees. In the first case, the object of interest (oranges) is "relatively light" and is better separated from other objects by a linear scale. In the second case, the object of interest (trees) is a "dark" object that is better separated from other background objects on a logarithmic scale. The relationship between these two units of measuring light intensity is: Density = Log_{10} (1/transmission). The aperture size (i.e., pixel size) for the 35mm slide was a 100 microns by 100 microns which represent about 1/84,000 of the total area of the slide. For the aerial photo transparency (9 inches by 9 inches), the aperture size was 240 microns by 240 microns which represented about 1/1,000,000 of the total area of the transparency. Chart 1 shows the data reduction and final results for a limited amount of data. This same sequential system can be used with low resolution sensors where the objects to be detected are relatively large and possess distinct spatial or spectral characteristics.

Chart 1.—Schematic diagrams showing data acquisition and classification of picture elements (i.e., pixels)



Appendix. 1972 SAS Procedure Used For Examples Cited

Example 1: Page 4

(a) PROC SORT ; BY GROUP ;

Comment: Sort Groups;

(b) PROC PLOT ROWS = 45 ;
VAR DRED DBLUE ; ID GROUP ;

Comment: Two Dimensional Scatter Plots of Discrimination Variables;

(c) PROC DISCRIM S POOL = TEST PROP ;
CLASS GROUPS ; VAR DRED DBLUE ;

Comment: Initial Classification Into Groups;

(d) PROC DISCOUT POOL = NO PROP ;
VAR DRED DBLUE ; CLASS GROUPS ;

Comment: Same PROC as PROC DISCRIM, with added feature that classification results are saved on a file for later processing.

Example 2: Page 5

(a) PROC MEANS; BY GROUPS ;

Comment: Look at Simple Statistics of Groups;

(b) PROC DISCRIM POOL = YES LIST PROP ;
CLASS GROUP ;
VARIABLES BLUE GREEN RED ;

Comment: Linear Discriminant Functions;
Proportional Group Priors

(c) PROC DISCRIM POOL = YES LIST ;
CLASS GROUP ;
VARIABLES BLUE GREEN RED ;

Comment: Linear Discriminate Function;
Equal Group Priors;

(d) PROC DISCRIM POOL = NO LIST PROP ;
CLASS GROUP ;
VARIABLES BLUE GREEN RED ;

Comment: Quadratic Discriminant Functions; Proportional Group Priors;

(e) PROC DISCRIM POOL = NO LIST ;

CLASS GROUP ;

VARIABLES BLUE GREEN RED ;

Comment: Quadratic Discriminant Functions; Equal Group Priors;

Example 3: Page 7 (Example No. 1 extended for 2279 data points)

(f) PROC PLOT ROWS = 45 ; VAR YORD XORD ;

Comment: X-Y Plot of Pixels Classified as Trees;

(g) PROC MSTCLUS ; VAR YORD XORD ;

BY BLOCK ;

Comment: Cluster Pixels Classified as Trees Using Spatial Variables;

NOTE: Minimal spanning tree cluster analysis added to SAS=72 version executed by subareas (Block) of the photograph. Data points were blocked into subareas to reduce computer costs;

(h) PROC PLOT ROWS = 45 ; VAR SIZE DIAM ;

ID CLUSTER ;

Comment: Done for every combination of cluster variables: size, diameter, average edge length, standard deviation of edge length within cluster;

(i) PROC DISCRIM S POOL = TEST PROP ;

CLASS CLUS-CLS ; VAR SIZE DIAM EDGE ;

Comment: Classify clusters into tree or nontree groups based upon cluster characteristics.