Craig H. Tom
ConTel Information Systems, Inc.
Government Systems Division
Littleton, CO 80120
Lee D. Miller
University of Nebraska
Nebraska Remote Sensing Center
Lincoln, NE 68588

An Automated Land-Use Mapping Comparison of the Bayesian Maximum Likelihood and Linear Discriminant Analysis Algorithms

Linear discriminant analysis proved superior to the Bayesian maximum likelihood in accuracy, time, cost of automated land-use mapping, and nonsensitivity of the variance and number of mapping variables.

Abstract: *The Bayesian maximum likelihood, a widely used parametric classifier, was tested against linear discriminant analysis, a data-based formulation, using the* GLIKE *and* CLASSIFY *decision/classification algorithms, respectively, in the Landsat Mapping System at Colorado State University. Identical supervised training sets,* USGS *land-use/land-cover classes, and various combinations of Landsat image and ancillary geodata variables were used to compare thematic mapping accuracies of* GLIKE *and* CLASSIFY *on a single-date summer subscene covering the Denver, Colorado metropolitan area. The "ground-truth" reference was a cellularized* USGS *land-use map of the same time frame.* CLASSIFY*, which accepts* a priori *class probabilities, was a more accurate classifier than* GLIKE*, which assumes equal class occurrences, for all three sets of mapping variables and both levels of detail. Even using the equal class probability assumption of* GLIKE*,* CLASSIFY *was again the more accurate classifier in five of six comparisons with* GLIKE*. These specific results may be generalized to direct accuracy, time, cost, and flexibility advantages of linear discriminant analysis over the Bayesian maximum likelihood, perhaps the most common machine classification technique used today.*

Introduction

As the use of aircraft and satellite remote sensing data has transitioned from basic research to both quasi- and fully operational applications, the emphasis on digital data has steadily increased. Additionally, digital image processing will expand at a much faster rate than in the past with the increased number of operational satellite programs and the increased volume of remote sensing data soon to be available from the following orbital sensors or satellites (Spann, 1980):

- Landsat-D (now Landsat-4);
- Gravsat;
- Magsat;
- Shuttle Imaging Radar;
- Shuttle Large Format Camera;
- Shuttle Multispectral Infrared Radiometer;
- Stereosat;

Photogrammetric Engineering and Remote Sensing,
Vol. 50, No. 2, February 1984, pp. 193-207.

0099-1112/84/5002-0193$02.25/

- Tethered Magnetometer;
- JEOS (Japan); and
- SPOT (France).

Digital image processing techniques are being increasingly applied to Landsat multispectral scanner (MSS) data to generate thematic land-use/land-cover maps. Both supervised and unsupervised classification procedures have evolved to manipulate digital remote sensing data. The supervised method requires a human analyst to designate recognizable "training sets" on an image, while the unsupervised method utilizes numerical clustering algorithms iteratively done by computer.

TABLE 1. "TURNKEY" HARDWARE/SOFTWARE SYSTEMS OVERVIEW. Maximum likelihood classification algorithms were available in five of six "turnkey" image processing/multispectral analysis hardware/software systems commercially available in the United States (Carter *et al.*, 1977).

Turnkey System	Source Organization	Computer Hardware	Required Memory	Supervised Classification	Programming Languages
MDAS*	Bendix Corp.	DEC PDP-11s	131,072 bytes	Bayesian max. likelihood, Gaussian std.	FORTRAN4 80% macro 20%
Vision One/20	Comtal Corp.	DEC PDP-11s/VAX, DG Nova/ Eclipse, H-P 2100/ 3000, SEL	?	parallelpiped	Macro 11
IDIMS	Electromagnetic Systems Labs	H-P 3000 II	131,072 bytes	advanced stored table (Bayesian maximum likelihood), maximum likelihood (array processor), Euclidean min. distance	FORTRAN SPL
IMAGE-100*	General Electric Company	DEC PDP-11s	32,768 words	parallelpiped, table lookup, nonparametric maximum likelihood,** hardware bulk classifier, including table look-up and parametric max. likelihood**	FORTRAN4 90% PAL 10%
System 101	International Imaging Systems	H-P 3000 III	131,072 bytes	maximum likelihood, parallelpiped, table lookup, Euclidean min. distance	FORTRAN SPL
Earth-view	Interpretation Systems Inc.	DEC PDP-11s	32,768 words	Bayesian max. likelihood, parallelpiped, table lookup, nonparametric	FORTH hi level 85% code 15%

* production discontinued
** system option

Parametric decision/classification algorithms, such as Bayesian probabilities and/or maximum likelihood functions, are commonly used in "turnkey" hardware/software systems (Table 1) and "public domain" software systems (Table 2). However, the maximum likelihood decision rule requires a large number of multiplications and logical comparisons for each decision, particularly when many MSS channels and mapping classes are used.

Many alternative classifiers have been tested because of the considerable amount of machine time consumed using maximum likelihood procedures. These studies have used a composite sequential clustering (Su *et al.*, 1972), an elliptical boundary condition model (Richardson *et al.*, 1971), and a lookup table procedure 30 times faster than the Bayesian maximum likelihood and broadly as accurate (Eppler *et al.*, 1971).

Another multivariate classifier is linear discriminant analysis. *A priori* data, in the form of selected samples of known land-use/land-cover types, are extracted from the pictorial scene and used as training sets to structure the discriminant function of the form

$$Y = a_1x_1 + a_2x_2 + \ldots + a_nx_n$$

where $x_1, x_2, \ldots, x_n$ are the image/geodata variables, and $a_1, a_2, \ldots, a_n$ are coefficients computed to determine a single value for Y, the linear compound, that minimizes misclassifications. Therefore, linear discriminant analysis telescopes a multivariate problem down into a linearly ordered situation (Tom and Miller, 1982).

Linear discriminant analysis computes a transform which gives the minimum ratio of the difference between a pair of class means to the multivar-

TABLE 2. "PUBLIC DOMAIN" SOFTWARE SYSTEMS OVERVIEW. Maximum likelihood classification algorithms were featured in four of seven "public domain" image processing/multispectral analysis software systems available through the Univ. of Ga. COSMIC software clearinghouse (Carter *et al.*, 1977; Univ. of Ga. Computer Center, 1981). The Bayesian maximum likelihood algorithm tested in this study was directly derived from the Purdue/LARSYS software system.

Software System	Source Organization	Computer Hardware	Required Memory	Supervised Classification	Programming Languages	Cosmic Reference
ASTEP	NASA/GSFC	Univac 1108	40,960 words	maximum likelihood	FORTRAN5 assembly	M75-10114
CAMSP	IBM Corp.	IBM 360s	307,200 bytes	maximum likelihood	assembly 100%	MSC-14979
CLASSPAK	NASA/GSFC	DEC PDP-11s	131,072 bytes	maximum likelihood	FORTRAN assembly	GSC-12374
ELLTAB	NASA/JSC	Univac 1108	?	table lookup	FORTRAN5 100%	MSC-14866
LARSYS III.1	Purdue Research Foundation	IBM 360s	524,288 bytes	maximum likelihood, sample (per field) classifier, multiimage layered classifier*	FORTRAN4 90% assembly 10%	MSC-14823
MAXL4X	NASA/ERL	Varian V-70	71,680 words	table lookup	FORTRAN4 assembly	ERL-10007
VICAR/ IBIS	Cal Tech/Jet Propulsion Laboratory	IBM 360s	153,600 bytes	Bayesian*, parallelpiped table lookup*, parallelpiped table lookup with Bayesian secondary classifier*, stored table Bayesian*	FORTRAN4 70% assembly 30%	NPO-14893

* not available in COSMIC program library version

iate variance within the two classes in the simple linear case. Visualizing these two classes as consisting of two swarms of data points in two-spectral space, the one optimum orientation is derived along which the two groups have the greatest separation while simultaneously minimizing the internal spread or inflation of the distribution of each group. An adequate separation between groups A and B cannot be made using either variable x_1 or x_2 (Figure 1). However, it is feasible to find an orientation along which the two groups are separated the most and inflated the least, with the coordinates of this axis of orientation being the linear discrimination function.

The Bayesian maximum likelihood and linear discriminant analysis algorithms have been both implemented in the Landsat Mapping System (LMS), a second-generation multispectral analysis hardware/software package developed at Colorado State University. The LMS package is structured to accepted both digital Landsat imagery and other spatial data for input, preprocessing, supervised feature extraction, decision/classification, and display output (Miller *et al.*, 1977b).

The maximum likelihood algorithm is FORTRAN-coded in LMS as GLIKE, a Bayesian maximum likelihood classifier taken directly from the Purdue LARSYS package (Smith *et al.*, 1972). The linear discriminant analysis algorithm is embodied in LMS as CLASSIFY, and is the modified verion of BMDO7M from the UCLA biomedical statistical package (Dixon, 1967). Both of these LMS classifiers are broadly representative of their respective techniques, and any functional differences in these specific formulations can also be attributed to the generalized procedures. A detailed overview of the LMS GLIKE and CLASSIFY algorithms is presented for greater understanding (Table 3).

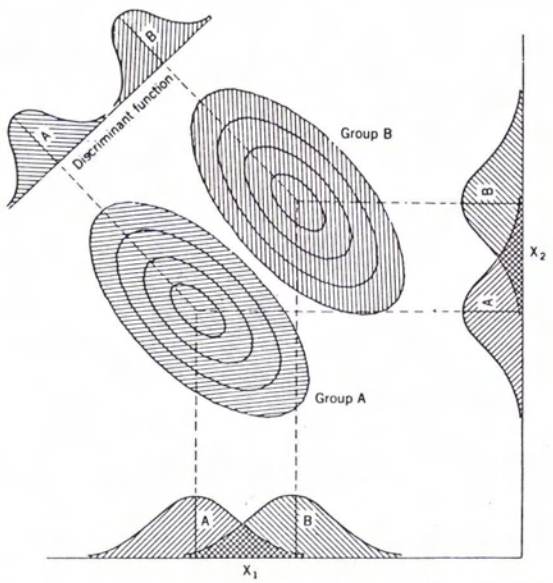

FIG. 1. Simple linear discriminant analysis diagram. This plot of two bivariate distributions shows an overlap between groups A and B along both variables x_1 and x_2. An orientation is computed along which the two groups are separated the most and inflated the least. The coordinates of this orientation are the linear discriminant function, and the clusters become distinguishable by projecting members of the two groups onto the discriminant function line (after Davis, 1973).

The general hypothesis of this research was that linear discriminant analysis would offer speed and cost advantages over the Bayesian maximum likelihood at little or no sacrifice of accuracy in automated land-use/land-cover classification of Landsat MSS data, as demonstrated by controlled tests of the LMS GLIKE and CLASSIFY algorithms on a moderate-sized subscene with independent ground-truth data. A secondary hypothesis was that spatially registered Landsat image and ancillary geodata (excluding land-use data) could materially increase automated land-use/land-cover classification accuracy when used together. The prospect of combining both image and nonimage forms of digital spatial data becomes increasingly attractive with the advent of automated geographic information systems.

Study Site

The Denver, Colorado Metropolitan Area was designated as the study area for this comparative pattern recognition algorithm research. The study used the 15 August 1973 Landsat-1 scene identified as path 36, row 32. A digital landscape model was created to organize and overlay spatial data from satellite imagery, existing maps, and census tables into a computer framework (Tom and Miller, 1980b). This assemblage provided a multivariate, multitemporal mathematical model which represented the landscape much as a three-dimensional model of the physical terrain is represented by a topographic map (Miller *et al.*, 1977a).

Approach and Results

Ground control consisted of a 1:100,000-scale U.S. Geological Survey (USGS) land-use map of Denver encoded in the USGS Circular 671 classification system (Anderson *et al.*, 1972). The original USGS map was manually compiled from 1:121,000-scale, high-altitude NASA U-2 color infrared aerial photos of 1972–1973 (Driscoll, 1975), and was subsequently cellularized as 4-ha (10-acre) squares for machine analysis (Table 4).

The 15 August 1973 Landsat-1 scene was geometrically corrected to yield a north-south-oriented square of 38.6 km (24 statute miles) on a side and centered on the Denver Metropolitan Area. The resultant subscene contained 576 rows and 576 columns of 0.45-ha (1.11-acre) square pixels. A nearest-

TABLE 3. OVERVIEW OF THE "GLIKE" (BAYESIAN MAXIMUM LIKELIHOOD) AND "CLASSIFY" (LINEAR DISCRIMINANT ANALYSIS) ALGORITHMS. These two supervised machine classifiers, part of the LMS multispectral analysis package, were tested for comparative land-use/land-cover mapping accuracy on a single-date LANDSAT-1 summer subscene, using the same training sets, USGS classes, and mapping variables.

Characteristics	"Glike" (Bayesian Maximum Likelihood)	"Classify" (Linear Discriminant Analysis)
Capabilities	Categorizes digital multispectral image data into predefined land-use/land-cover classes of interest, using supervised pattern recognition processes on *n* training set class means and *n* class variance-covariance matrices.	Categorizes digital multispectral image data into predefined land-use/land-cover classes of interest, using supervised pattern recognition processes on *n* training set class means and a single, within-groups variance-covariance matrix.
Data Requirements	*A priori* identification of training areas for each specific mapping class in order to compute sample field statistics to serve as estimators of population parameters.	*A priori* identification of training areas for each specific mapping class in order to compute sample field statistics to serve as estimators of population parameters.
Analysis Methods	A parametric approach which uses a likelihood ratio decision rule based on the Bayesian formulation and conditional multivariate Gaussian probability densities (Smith *et al.*, 1972).	A data-based formulation which calculates functions which discriminate between mapping classes in an optimal manner. The discriminant functions determine boundaries which produce a set of subspaces, one subspace for each class. The location of the boundaries is such that a minimum of misclassifications (i.e., individual points lying in the incorrect subspace) occur (Jordan *et al.*, 1978).
Limitations	Individual dispersion matrices tend to become singular when numerous mapping classes are used; and it is more time-consuming computationally to use individual dispersion matrices (Hsu, 1978).	The linear discriminant approach reduces a multidimensional problem to a more manageable linear problem. This many-to-one mapping, at least in theory, cannot reduce the minimum achievable error rate. However, some of the theoretically attainable accuracy can be sacrificed for the advantages of working solely with a linear model (Duda and Hart, 1973).
Statistical	The training data for each mapping class adequately represents the actual class; *A priori* class probabilities assumed to be equal (Maxwell, 1976); The data are multivariate normally distributed (Smith *et al.*, 1972); The data are independent and have equal variances (Mendenhall and Scheaffer, 1973).	The training data for each mapping class adequately represents the actual class; *A priori* class probabilities may be specified, or assumed to be equal otherwise (Dixon, 1967); No underlying statistical model is assumed (Duda and Hart, 1973); The data are independent and have equal variances (Mather, 1976).
Possible Consequences Of Statistical Violations	Classification accuracy is not very sensitive to even a moderately severe violation of the normality assumption (Swain and Davis, 1978).	Discriminant analysis is not seriously affected by limited departures from normality or limited inequality of variances (Davis, 1973).
Ramifications	Transformation may be used on original "raw" data to generate normally distributed data sets; or multimodal classes may be separated into more normally distributed subclasses (Smith *et al.*, 1972).	The Bayesian maximum likelihood and linear discriminant analysis procedures yield the same results if the data sets are independent, normally distributed, and have equal variances (G. H. Rosenfield, unpublished data, 1983).

neighbor restitution algorithm with Earth rotation, scanline skew, nonlinear mirror velocity, frame rotation, and pixel resampling without ground control points was used to generate the spectral data set (Tom and Miller, 1980a). This rectification allowed the spatial registration of digital Landsat image data with ancillary land-use, physiographic, transportation access, and socio-economic geodata (Table 5).

A one-ninth subscene was generated by rectilinearly resampling every third row and third column for a total sample of 192 rows and 192 columns from the original 576- by 576-element composite land-

TABLE 4. HIERARCHICAL USGS CIRCULAR 671 LAND-USE/LAND-COVER CLASSIFICATION SYSTEM USED FOR THE DENVER METROPOLITAN AREA. Only the first- and second-order mapping classes are standardized for aircraft and/or satellite image classification/interpretation. The detailed third-order classes are completely user-defined. A 1972-1973 USGS photointerpreted land-use map (Driscoll, 1975) was used as the ground-truth reference against which all the machine classifications were compared for mapping accuracy at a given hierarchical level of detail.

Multilevel Digital Codes			First-Order Land-Use/Land-Cover Type Second-Order Land-Use/Land-Cover Type Third-Order Land-Use/Land-Cover Type
1			Urban and Built-Up Land
	11		Residential
	12		Commercial and Services
		121	Recreational
	13		Industrial
	14		Extractive
	15		Transportation, Communications, and Utilities
		151	Utilities
	16		Institutional
	17*		Strip and Clustered Development
	18*		Mixed Urban
	19		Open and Other Urban
		191	Solid-Waste Dump
		192	Cemetery
2			Agricultural Land
	21		Cropland and Pasture
		211	Nonirrigated Cropland
		212	Irrigated Cropland
		213	Pasture
	22*		Orchards, Groves, and Other Horticultural Areas
	23*		Feeding Operations
	24*		Other Agricultural Land
3			Rangeland
	31		Grass
	32*		Savannas
	33		Chapparal (taken as brushland)
	34*		Desert Shrub
4			Forest Land
	41		Deciduous
		411	Deciduous/Intermittent Crown
	42		Evergreen (Coniferous and Other)
		421	Coniferous/Solid Crown
		422	Coniferous/Intermittent Crown
	43*		Mixed Forest Land
5			Water
	51		Streams and Waterways
	52		Lakes
	53		Reservoirs
	54*		Bays and Estuaries
	55*		Other Water
6*			Nonforested Wetland
	61*		Vegetated
	62*		Bare
7			Barren Land
	71*		Salt Flats
	72*		Beaches
	73*		Sand Other Than Beaches
	74		Bare Exposed Rock
		741	Hillslopes
	75*		Other Barren Land
8*			Tundra
	81*		Tundra
9*			Permanent Snow and Icefields
	91*		Permanent Snow and Icefields

* Land-use/land-cover type not found in the Denver Metropolitan Area.

TABLE 5. LIST OF 48 SPATIAL LANDSCAPE MODELING VARIABLES. The 38.6-km by 38.6-km (24-statute-mile) Denver Metropolitan Area was imaged by ten LANDSAT-1 channels and channel ratios, as well as 38 collateral geodata planes. Each landscape data plane represented a spatially distributed, single-variable image or map in a fully registered stack.

Landscape Submodel	Landscape Variable	Source of Data	Variable Type
Landsat-1 Image	MSS-4 (visible green)	15 Aug 73 Landsat-1 digital tape	Numerical
	MSS-5 (visible red)	"	"
	MSS-6 (solar infrared1)	"	"
	MSS-7 (solar infrared2)	"	"
	MSS-5/MSS-4 ratio	Computed as MSS-5/MSS-4 ratio	"
	MSS-6/MSS-4 ratio	Computed as MSS-6/MSS-4 ratio	"
	MSS-7/MSS-4 ratio	Computed as MSS-7/MSS-4 ratio	"
	MSS-5/MSS-6 ratio	Computed as MSS-5/MSS-6 ratio	"
	MSS-7/MSS-5 ratio	Computed as MSS-7/MSS-5 ratio	"
	MSS-7/MSS-6 ratio	Computed as MSS-7/MSS-6 ratio	"
Land-Use	1963 photo land-use	1:20,000-scale B/W aerial photos	Categorical
	1970 photo land-use	1:24,000-scale B/W orthophotos	"
	1972-1973 USGS photo land-use	1:100,000-scale USGS land-use map	"
	1963-to-1970 land-use changes (from)	Computed from 1963 and 1970 land-use	"
	1963-to-1970 land-use changes (to)	"	"
	1963-to-1970 alphanumeric land-use changes (from)	"	"
	1963-to-1970 alphanumeric land-use changes (to)	"	"
Physiographic	Topographic elevation	1:24,000-scale USGS topographic maps	Numerical
	Topographic slope	Computed from topographic elevations	"
	Topographic aspect	"	"
	Surficial geology	1:62,500-scale USGS geologic map	Categorical
	LANDSAT image insolation	Computed from elevation, slope, aspect	Numerical
	LANDSAT MSS-4/insolation ratio	Computed as MSS-4/insolation ratio	"
	LANDSAT MSS-5/insolation ratio	Computed as MSS-5/insolation ratio	"
	LANDSAT MSS-6/insolation ratio	Computed as MSS-6/insolation ratio	"
	LANDSAT MSS-7/insolation ratio	Computed as MSS-7/insolation ratio	"
Road Transportation Access	Composite minor road minimum distance (MD)	Computed from 1:45,000-scale state highway department map	Numerical
	Composite major road MD	"	"
	Freeway MD	"	"
	Freeway interchange MD	"	"
	Built-up urban area MD	"	"
Socio-Economic	Total population	1970 Census reports and 1:84,500-scale Census tract maps	Numerical
	Total families	"	"
	Total year-round housing units	"	"
	Total vacant housing units	"	"
	Total occupied housing units	"	"
	1969 mean family income	"	"
	Median housing-unit value	"	"
	Median housing-unit rent	"	"
	Total one-car families	"	"
	Total two-car families	"	"
	Total three-/three-plus car families	"	"
	Total census tract acreage	Computed from Census tract maps	"
	Population density per acre	Computed as total population/total census tract acreage	"
	Average number of cars per family	Computed as total one-, two-, three-, and three-plus car families/total families	"
	Average number of families per acre	Computed as total families/total census tract acreage	"
	Average number of year-round units per acre	Computed as total year-round housing units/total census tract acreage	"
	Average number of vacant housing units per acre	Computed as total vacant housing units/ total census tract acreage	"

scape model. Consequently, the classification accuracy of any algorithm, combination of mapping variables, or training set structure could be verified on a point-to-point basis by comparison with the cellularized 1972–1973 USGS land-use map.

Training set statistics for both the Bayesian maximum likelihood and linear discriminant analysis were generated by a new self-verifying, grid-sampling training *point* approach. This point training set was created by again resampling every third row and third column of the one-ninth subscene, yielding a one-ninth times one-ninth or 1/81-sampled image of 4,100 training points of known USGS land-use/land-cover type. This systematic point sampling process represented an efficient distillation of the landscape model, while providing uniform coverage of the study area (Tom and Miller, 1982).

Three combinations of image/non-image mapping variables were applied to classify the one-ninth subscene. These mapping combinations were selected to appraise the classificational utility of increasing amounts of collateral geodata beyond the basic MSS bands. The four-variable combination was the basic four MSS bands. The six-variable set included MSS-4, MSS-7, MSS-7/MSS-5 ratio, MSS-5/MSS-4 ratio, topographic elevation, and urban built-up area minimum distance, a computed spatial distance parameter that expresses the shortest straight-line distance from any cell to the closest urban built-up area (Tom *et al.*, 1978). Lastly, the 22-variable set contained the maximum number of the 41 nonland-use variables (Table 5) which could be used for maximum likelihood mapping with GLIKE. That is, the restricted subset of the 48 spatial landscape modeling variables with nonzero individual dispersion matrices in the 13 second- and 11 third-order USGS land-use classes and, therefore, permitting matrix inversion by GLIKE (Appendix A). These were the four basic MSS bands, six MSS ratios, four MSS/insolation ratios, topographic elevation, topographic aspect, landsat image insolation, and five road transportation access minimum-distance variables.

These three mapping combinations were applied to the one-ninth subscene using both the GLIKE and CLASSIFY algorithms, and checked for general first-order accuracy on a point-to-point basis for all of the 36,864 points against the 1972–1973 USGS land-use reference. The GLIKE average accuracies checked to six first-order USGS classes were 54, 68, and 61 percent, respectively, for the four-, six-, and 22-variable mapping combinations, while the CLASSIFY average accuracies (with *a priori* class probabilities) were 65, 73, and 76 percent, respectively (Table 6).

The GLIKE average accuracies checked to 13 second- and 11 third-order USGS classes were 4, 15, and 2 percent, respectively, for the four-, six-, and 22-variable mapping combinations, while the CLASSIFY average accuracies (with *a priori* class probabilities) were 38, 46, and 49 percent, respectively (Table 7).

The poor performance of GLIKE, especially for the detailed second- and third-order USGS classes, was indeed surprising. A detailed explanation is not currently available. However, the fact that the first-order GLIKE mapping accuracy was comparatively high relative to the second- and third-order mapping accuracy may provide a clue. That is, the capability for specifying *a priori* class probabilities in CLASSIFY materially improved all its classifications (Appendix B). Because no such capability existed in GLIKE, it could achieve comparable results to CLAS-

TABLE 6. COMPARATIVE FIRST-ORDER LAND-USE/LAND-COVER MAPPING ACCURACIES OF "GLIKE" AND "CLASSIFY" (WITH *A PRIORI* [PRIOR] CLASS PROBABILITIES) ALGORITHMS. A 192-row by 192-column subscene was classified with both algorithms using three different combinations of LANDSAT-1 image and/or ancillary geodata variables. The same supervised training areas were used for both algorithms. The accuracy of these classifications was checked cell-by-cell for six first-order classes (Table 5) against the digital 1972-1973 USGS reference. Calculated z values and associated significance levels (P) showed that there were highly significant differences between the two classifiers in providing correct classifications (after Snedecor and Cochran, 1967). Image taken 15 August 1973.

Number of Mapping Variables	Computer Classification Algorithm Used	Verified Accuracy, Pixels	Verified Accuracy, Percent	Computed Z Statistic
Four	Glike	19,751	53.58	z = 32.5*
	Classify (Prior)	24,051	65.24	P = 0.000
Six	Glike	25,093	68.07	z = 14.9*
	Classify (Prior)	26,938	73.07	P = 0.000
Twenty-Two	Glike	22,662	61.47	z = 41.7*
	Classify (Prior)	27,855	75.56	P = 0.000

* Significant at 0.001 probability level.

TABLE 7. COMPARATIVE SECOND- AND THIRD-ORDER LAND-USE/LAND-COVER MAPPING ACCURACIES OF "GLIKE" AND "CLASSIFY" (WITH *A PRIORI* [PRIOR] CLASS PROBABILITIES) ALGORITHMS. A 192-row by 192-column subscene was classified with both algorithms using three different combinations of LANDSAT-1 image and/or ancillary geodata variables. The same supervised training areas were used for both algorithms. The accuracy of these classifications was checked cell-by-cell for 13 second- and 11 third-order classes (Table 5) against the digital 1972-1973 USGS reference. Calculated z values and associated significant levels (P) showed that there were highly significant differences between the two classifiers in providing correct classifications (after Snedecor and Cochran, 1967). Image taken 15 August 1973.

Number of Mapping Variables	Computer Classification Algorithm Used	Verified Accuracy, Pixels	Verified Accuracy, Percent	Computed Z Statistic
Four	Glike	1,584	4.30	z = 122.7*
	Classify (Prior)	13,972	37.90	P = 0.000
Six	Glike	5,690	15.44	z = 96.1*
	Classify (Prior)	17,056	46.27	P = 0.000
Twenty-Two	Glike	895	2.43	z = 170.3*
	Classify (Prior)	18,002	48.83	P = 0.000

* Significant at 0.001 probability level.

SIFY given only six broad first-order classes, but the statistical similarity of the many second- and third-order classes proved too much for it to handle with only assumed equal class probabilities.

This *a priori* versus equal class probability hypothesis was tested by rerunning the CLASSIFY algorithm for the four-, six-, and 22-variable mapping combinations, and quantifying the accuracy changes using the equal class probability assumption of GLIKE. Consequently, CLASSIFY had reduced average accuracies with equal mapping class probabilities which, checked to six first-order classes, were 51, 71, and 73 percent, respectively, for the four-, six-, and 22-variable mapping combinations (Table 8), while the combined second- and third-order accuracies were 19, 32, and 37 percent, respectively (Table 9).

The use of equal mapping class probabilities presumes equal likelihood of each land-use/land-cover type in the larger, unknown image, while weighted class probabilities presume that the analyst has some *a priori* knowledge of the proportion of predefined mapping classes in the study area. Some knowledge of the relative amount of each class,

TABLE 8. COMPARATIVE FIRST-ORDER LAND-USE/LAND-COVER MAPPING ACCURACIES OF "GLIKE" AND "CLASSIFY" (WITH DEFAULT [EQUAL] CLASS PROBABILITIES) ALGORITHMS. A 192-row by 192-column subscene was classified with both algorithms using three different combinations of LANDSAT-1 image and/or ancillary geodata variables. The same supervised training areas were used for both algorithms. The accuracy of these classifications was checked cell-by-cell for six first-order classes (Table 5) against the digital 1972-1973 USGS reference. Calculated z values and associated significant levels (P) showed that there were highly significant differences between the two classifiers in providing correct classifications (after Snedecor and Cochran, 1967). Image taken 15 August 1973.

Number of Mapping Variables	Computer Classification Algorithm Used	Verified Accuracy, Pixels	Verified Accuracy, Percent	Computed Z Statistic
Four	Glike	19,751	53.58	z = 8.26*
	Classify (Equal)	18,631	50.54	P = 0.000
Six	Glike	25,093	68.07	z = 7.33*
	Classify (Equal)	26,011	70.56	P = 0.000
Twenty-Two	Glike	22,662	61.47	z = 33.8*
	Classify (Equal)	26,937	73.07	P = 0.000

* Significant at 0.001 probability level.

TABLE 9. COMPARATIVE SECOND- AND THIRD-ORDER LAND-USE MAPPING/LAND-COVER MAPPING ACCURACIES OF "GLIKE" AND "CLASSIFY" (WITH DEFAULT [EQUAL] CLASS PROBABILITIES) ALGORITHMS. A 192-row by 192-column subscene was classified with both algorithms using three different combinations of LANDSAT-1 image and/or ancillary geodata variables. The same supervised training areas were used for both algorithms. The accuracy of these classifications was checked cell-by-cell for 13 second- and 11 third-order classes (Table 5) against the digital 1972-1973 USGS reference. Calculated z values and associated significance levels (P) showed that there were highly significant differences between the two classifiers in providing correct classifications (after Snedecor and Cochran, 1967). Image taken 15 August 1973.

Number of Mapping Variables	Computer Classification Algorithm Used	Verified Accuracy, Pixels	Verified Accuracy, Percent	Computed Z Statistic
Four	Glike	1,584	4.30	z = 64.1*
	Classify (Equal)	7,030	19.07	P = 0.000
Six	Glike	5,690	15.44	z = 54.8*
	Classify (Equal)	11,915	32.32	P = 0.000
Twenty-Two	Glike	895	2.43	z = 131.7*
	Classify (Equal)	13,728	37.24	P = 0.000

* Significant at 0.001 probability level.

whether derived from field surveys, image sampling, or ancillary geodata sources, is computationally useful. Previous results from CLASSIFY with default (equal) and *a priori* class probabilities checked to the six first-order USGS classes (Table 10) and the combined 13 second- and 11 third-order USGS classes (Table 11) directly showed the reduced mapping accuracy using assumed equal class occurrences. The drop was even more pronounced for the combined second- and third-order classes (Table 11), and mirrored the same effect noted earlier in GLIKE, although not to the same severity.

A clear machine time and cost differential emerged between the two LMS algorithms. GLIKE took 508, 845, and 2,212 machine seconds on a Control Data Corporation 6400 computer, respectively, for the four-, six-, and 22-variable mapping combinations, while CLASSIFY took only 406, 430, and 612 seconds, respectively (Table 12). Thus, CLASSIFY took only 80, 51, and 28 percent of the machine

TABLE 10. COMPARATIVE FIRST-ORDER LAND-USE/LAND-COVER MAPPING ACCURACIES OF "CLASSIFY" (WITH DEFAULT [EQUAL] AND *A PRIORI* [PRIOR] CLASS PROBABILITIES) ALGORITHM. A 192-row by 192-column subscene was classified with both class probabilities using three different combinations of LANDSAT-1 image and/or ancillary geodata variables. The same supervised training areas were used for the algorithm. The accuracy of these classifications was checked cell-by-cell for six first-order classes (Table 5) against the digital 1972-1973 USGS reference. Calculated z values and associated significance levels (P) showed that there were highly significant differences between the two class probabilities in providing correct classifications (after Snedecor and Cochran, 1967). Image taken 15 August 1973.

Number of Mapping Variables	Computer Classification Algorithm Used	Verified Accuracy, Pixels	Verified Accuracy, Percent	Computed Z Statistic
Four	Classify (Equal)	18,631	50.54	z = 40.9*
	Classify (Prior)	24,051	65.24	P = 0.000
Six	Classify (Equal)	26,011	70.56	z = 7.58*
	Classify (Prior)	26,938	73.07	P = 0.000
Twenty-Two	Classify (Equal)	26,937	73.07	z = 7.74*
	Classify (Prior)	27,855	75.56	P = 0.000

* Significant at 0.001 probability level.

TABLE 11. COMPARATIVE SECOND- AND THIRD-ORDER LAND-USE/LAND-COVER MAPPING ACCURACIES OF "CLASSIFY" (WITH DEFAULT [EQUAL] AND *A PRIORI* [PRIOR] CLASS PROBABILITIES) ALGORITHM. A 192-row by 192-column subscene was classified with both class probabilities using three different combinations of LANDSAT-1 image and/or ancillary geodata variables. The same supervised training areas were used for the algorithm. The accuracy of these classifications was checked cell-by-cell for 13 second- and 11 third-order classes (Table 5) against the digital 1972-1973 USGS reference. Calculated z values and associated significance levels (P) showed that there were highly significant differences between the two class probabilities in providing correct classifications (after Snedecor and Cochran, 1967). Image taken 15 August 1973.

Number of Mapping Variables	Computer Classification Algorithm Used	Verified Accuracy, Pixels	Verified Accuracy, Percent	Computed Z Statistic
Four	Classify (Equal)	7,030	19.07	$z = 57.9$*
	Classify (Prior)	13,972	37.90	P = 0.000
Six	Classify (Equal)	11,915	32.32	$z = 39.2$*
	Classify (Prior)	17,056	46.27	P = 0.000
Twenty-Two	Classify (Equal)	13,728	37.24	$z = 32.0$*
	Classify (Prior)	18,002	48.83	P = 0.000

* Significant at 0.001 probability level.

times required by GLIKE, respectively, for the four-, six-, and 22-variable mapping combinations. The GLIKE algorithm thus exhibited machine time and cost sensitivity to the number of mapping variables used for classification.

The CLASSIFY algorithm had a clear cost advantage over GLIKE when the machine costs were weighted by correctly classified pixels for costs per correct pixel. Here, CLASSIFY was 1.5, 2.1, and 4.4 times more cost-effective than GLIKE for the first-order mapping of the four-, six-class, and 22-variable mapping combinations, respectively, while for the combined second- and third-order mapping, CLASSIFY was 11.3, 6.0, and 73.7 times more cost-effective, respectively (Table 12).

Finally, the use of band ratios derived from the

TABLE 12. COMPARATIVE MACHINE CLASSIFICATION TIMES, TOTAL COSTS, AND COSTS PER CORRECT PIXEL OF "GLIKE" AND "CLASSIFY" (WITH DEFAULT [EQUAL] AND *A PRIORI* [PRIOR] CLASS PROBABILITIES) ALGORITHMS. A 192-row by 192-column subscene was classified with both algorithms using three different combinations of LANDSAT-1 image and/or ancillary geodata variables. The same supervised training areas were used for both algorithms. The accuracy of these classifications was checked cell-by-cell for six first-order and 13 second- and 11 third-order classes (Table 5), respectively, against the digital 1972-1973 USGS reference. The machine processing was performed on a Control Data Corporation 6400 computer. Image taken 15 August 1973.

Number of Mapping Variables	Computer Classification Algorithm Used	Machine Time, Seconds	Machine Cost, Dollars	Cost per Correct First-Order Pixel, Cents	Cost per Correct Second-/Third-Order Pixel, Cents
Four	Glike	508.24	40.94	0.21	2.59
	Classify (Equal)	401.21	32.32	0.17	0.46
	Classify (Prior)	406.33	32.73	0.14	0.23
Six	Glike	844.97	68.07	0.27	1.20
	Classify (Equal)	414.90	33.42	0.13	0.28
	Classify (Prior)	429.45	34.59	0.13	0.20
Twenty-Two	Glike	2212.42	178.22	0.79	19.91
	Classify (Equal)	610.28	49.16	0.18	0.36
	Classify (Prior)	612.24	49.32	0.28	0.27

TABLE 14. IMPROVEMENT IN THE COMBINED SECOND- AND THIRD-ORDER LAND-USE/LAND-COVER MAPPING ACCURACIES OF "CLASSIFY" (WITH *A PRIORI* [PRIOR] CLASS PROBABILITIES) ALGORITHM BY THE ADDITION OF MSS BAND RATIOS AND ANCILLARY GEODATA. Calculated z values and associated significance levels (P) showed that there were highly significant differences between the four- and mapping combination and the addition of MSS band ratios and ancillary geodata in providing correct classifications (after Snedecor and Cochran, 1967). Image taken 15 August 1973.

Number of Mapping Variables	Computer Classification Algorithm Used	Verified Accuracy, Pixels	Verified Accuracy, Percent	Computed Z Statistic
Four	Classify (Prior)	13,972	37.90	$z = 23.1$*
Six		17,056	46.27	P = 0.000
Four	Classify (Prior)	13,972	37.90	$z = 30.1$*
Twenty-Two		18,002	48.83	P = 0.000
Six	Classify (Prior)	17,056	46.27	$z = 6.96$*
Twenty-Two		18,002	48.83	P = 0.000

* Significant at the 0.001 probability level.

basic MSS bands, as well as map-derived ancillary geodata, as additional pseudospectral bands increased thematic mapping accuracy over the basic four-band mapping combination. Previous results from CLASSIFY with *a priori* class probabilities checked to the six first-order USGS classes (Table 13) and the combined 13 second- and 11 third-order USGS classes (Table 14) directly showed the increased classificational accuracy using band ratios and, in particular, ancillary geodata.

SUMMARY

Several significant conclusions emerged from these comparative LMS algorithm tests involving GLIKE (Bayesian maximum likelihood) and CLASSIFY (linear discriminant analysis), as follows:

- Linear discriminant analysis was a more accurate classifier than the Bayesian maximum likelihood;
- Bayesian maximum likelihood machine time was greater than that for linear discriminant analysis;
- Bayesian maximum likelihood machine cost per correctly classified pixel was much greater than that for linear discriminant analysis; and
- Linear discriminant analysis machine time was much less sensitive to the number of mapping variables and zero mapping class variance than the Bayesian maximum likelihood.

These results demonstrated the greater utility of linear discriminant analysis as a data-based machine algorithm relative to the Bayesian maximum likelihood, perhaps the most common computer classification technique in use today. Advantages of linear discriminant analysis were indicated in terms of accuracy, time, cost, and nonsensitivity to the statistical variance and number of mapping variables. These economies of the linear discriminant algorithm will be important in making the computer more available and acceptable for machine processing of digital MSS data in the future.

The USGS Circular 671 scheme, developed for aircraft and satellite data inputs, worked reasonable well as a hierarchical classification system at both the first- and second-order levels in this supervised machine-interpretation effort. However, it appeared that much more detailed ancillary geodata will be needed for accurately mapping the third-order USGS urban classes. The inherent remaining problem in urban settings is the repetitive use of man-made building matrials in a diversity of urban land uses, and is worthy of additional research. Fortunately, the USGS land-use classification system is predominately land-cover oriented and is therefore useful for remote sensing discrimination. The ancillary geodata compensated, in large part, for the land-use/land-cover confusion factor in the seven second-order urban classes but did not do well for the four third-order urban classes. Average classification accuracy was 60 percent for the second-order urban classes, but only 13 percent for the third-order classes, using linear discriminant analysis. The generally poor performance of the Bayesian maximum likelihood, especially for the second- and third-order USGS classes, should also receive additional attention by researchers.

Lastly, the increased mapping accuracy of joint Landsat image and ancillary geodata sets strongly suggests that additional automated geographic information systems development could provide more complete, consistent, and objective information and analyses than might be possible using only digital remote sensing data.

Acknowledgments

The authors gratefully acknowledge the indispensable support provided by the Earth Resources Branch (Code 923), NASA/Goddard Space Flight Center, and the Colorado State University Computer Center, as well as the many helpful comments and suggestions received during the preparation of this paper, particularly those of George H. Rosenfield, USGS/Reston, Virginia. The work described in this report was performed for the National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, MD 20771.

References

Anderson, J. R., E. E. Hardy, and J. T. Roach, 1972. *A Land-Use Classification System for Use with Remote Sensor Data.* U.S. Geological Survey Circular 671, U.S. Gov't Print. Off., Wash., D.C., 16 p.

Carter, V. P., F. Billingsley, and J. Lamar, 1977. *Summary Tables for Selected Digital Image Processing Systems.* U.S. Geological Survey Open-File Report 77-414, Reston, Va., 45 p.

Davis, John C., 1973. *Statistics and Data Analysis in Geology.* John Wiley and Sons, New York, 550 p., illus.

Dixon, W. J., 1967. *BMD Biomedical Programs.* Univ. of Calif. Press, Berkeley, pp. 214a-214s.

Driscoll, Linda B., 1975. *Land-Use Classification Map of the Greater Denver Area, Front Range Urban Corridor.* U.S. Geological Survey Misc. Inv. Map-I-856-E.

Duda, R. O., and P. E. Hart, 1973. *Pattern Classification and Scene Analysis.* John Wiley and Sons, New York, 482 p., illus.

Eppler, W. G., C. A. Helmke, and R. H. Evans. 1971. Table Look-Up Approach to Pattern Recognition. In *Proceedings of the Seventh International Symposium on Remote Sensing of Environment,* The Univ. of Michigan, Ann Arbor, pp. 1415-1425.

Hsu, Shin-yi, 1978. Texture-Tone Analysis for Automated Land-Use Mapping. *Photogram. Eng. and Remote Sensing* 44(11):1393-1404.

Jordan, D. C., D. H. Graves, and M. C. Hammetter, 1978. Use of Manual Densitometry in Land Cover Classification. *Photogram. Eng. and Remote Sensing* 44(8):1053-1059.

Mather, P. M., 1976. *Computational Methods of Multivariate Analysis in Physical Geography.* John Wiley and Sons, London, 532 p., illus.

Maxwell, Eugene L. 1976. Multivariate Systems Analysis of Multispectral Imagery. *Photogram. Eng. and Remote Sensing* 42(9):1173-1186.

Mendenhall, W., and R. L. Scheaffer, 1973. *Mathematical Statistics with Applications.* Duxbury Press, North Scituate, Mass., 55 pp., plus appendices.

Miller, L. D., C. H. Tom, and K. Nualchawee, 1977a. *Remote Sensing Inputs to Landscape Models Which Predict Future Spatial Land Use Patterns for Hydrologic Models.* NASA preprint X-923-77-115, Goddard Space Flight Center, Greenbelt, Md., 41 p., illus.

Miller, L. D., E. L. Maxwell, and R. L. Riggs, 1977b. *User's Manual for LMS (LANDSAT Mapping System).* Dep't of Earth Resources, Colorado State Univ., Ft. Collins, misc. paging.

Richardson, A. J., R. J. Torline, and W. A. Allen, 1971. Computer Identification of Ground Pattern from Aerial Photographs. In *Proceedings of the Seventh International Symposium on Remote Sensing of Environment,* The Univ. of Michigan, Ann Arbor, pp. 1357–1376.

Smith, J. A., L. D. Miller, and T. D. Ells, 1972. *Pattern Recognition Routines for Graduate Training in the Automatic Analysis of Remote Sensing Imagery—RECOG.* Science Series 3A, Dep't of Watershed Sciences, Colorado State Univ., Ft. Collins, 86 p., illus.

Snedecor, G. W., and W. G. Cochran, 1967. *Statistical Methods, Sixth Ed.* Iowa State Univ. Press, Ames, 593 p., illus.

Spann, G. William, 1980. Satellite Remote Sensing Markets in the 1980's. *Photogram. Eng. and Remote Sensing* 46(1):65-69.

Su, M. Y., R. R. Jayroe, and R. E. Cummings, 1972. Unsupervised Classification of Earth Resources Data. In *Remote Sensing of Earth Resources,* F. Shahrokhi, ed., Univ. of Tennessee, Tullahoma, pp. 673–694.

Swain, P. H., and S. M. Davis, 1978. *Remote Sensing: The Quantitative Approach.* McGraw-Hill Book Co., New York, 396 p., illus.

Tom, C. H., and L. D. Miller, 1980a. Spatial Land-Use Inventory/Denver Metropolitan Area, with Inputs from Existing Maps, Air Photos, and LANDSAT Imagery. In *Proceedings of the Fourteenth International Symposium on Remote Sensing of Environment,* The Univ. of Michigan, Ann Arbor, pp. 603-612.

———, 1980b. Forest Site Index Mapping and Modeling. *Photogram. Eng. and Remote Sensing* 46(12):1585-1596.

———, 1982. A Comparison of LANDSAT Point and Rectangular Field Training Sets for Land-Use Classification. *International Jour. of Remote Sensing* review draft, 28 p.

Tom, C. H., L. D. Miller, and J. R. Christenson, 1978. *Spatial Land-Use Inventory, Modeling, and Projection/Denver Metropolitan Area, with Inputs from Existing Maps, Airphotos, and LANDSAT Imagery.* NASA Technical Memorandum 79710, Goddard Space Flight Center, Greenbelt, Md., 225 p., illus.

Univ. of Ga. Computer Center, 1981. *Image Processing.* Computer Software Management and Information Center (COSMIC), Univ. of Ga. Computer Center, Athens, Ga., misc. paging.

(Received 26 May 1982; revised and accepted 1 October 1983)

Appendix A
Bayesian Maximum Likelihood

The Bayesian maximum likelihood classifier, such as the GLIKE classification algorithm in the LMS package, assumes that the data are multivariate normally distributed. It utilizes a Bayesian decision rule of the form

$$p(\mathbf{X}|i) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}_i|^{1/2}} e^{-(1/2)(\mathbf{X}-\mathbf{A}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{X}-\mathbf{A}_i)}$$

where

- $p(\mathbf{X}|i)$ is the likelihood of occurence of the feature vector
- $\mathbf{X}$ given that it belongs to the ith class,
- $\mathbf{\Sigma}_i$ is the variance-covariance matrix for the ith class,
- $|\mathbf{\Sigma}_i|$ is the determinant of $\mathbf{\Sigma}_i$,
- $\mathbf{\Sigma}^{-1}$ is the inverse of $\mathbf{\Sigma}_i$, and
- $\mathbf{A}_i$ is the mean feature vector for the ith class.

These conditional probabilities for mapping classes are taken two at a time, ratioed, and evaluated to assign each pixel to the class for which the likelihood, $p(\mathbf{X}/i)$, or the unknown vector is the highest, or

$$g_i(\mathbf{X}) = p(\mathbf{X}/i).$$

The class probabilities, $p(i)$, are assumed equal in the Bayesian maximum likelihood (Maxwell, 1976). However, the modified Bayesian maximum likelihood can exploit *a priori* mapping class probabilities for improved machine classification accuracy by taking

$$g_i(\mathbf{X}) = p(i)p(\mathbf{X}/i).$$

Appendix B
Linear Discriminant Analysis

ANALYSIS

Linear discriminant analysis derives linear combinations of the mapping variables which provide the greatest separation of the group multivariate means, and which provide the least inflation of the within-groups multivariate variance. The statistical algorithm embodied in CLASSIFY computes a discriminant function for each of the mapping classes by selecting the independent variables, the 48 Landsat image and ancillary geodata variables, in a stepwise fashion. The new variable entered at each step is selected on the basis of largest F-value to enter. An original set of observations on a landscape modeling element, for example, is transformed into a single discriminant score by the discriminant function. The score represents the cell's position along the line defined by the linear discriminant function. As stated earlier, the discriminant function collapses a multivariate problem down into a univariate situation.

The discriminant function is found by solving an equation of the form

$$[s_p^2] \cdot [\lambda] = [D]$$

where $[s_p^2]$ is an $m \times m$ matrix of pooled variances and covariances of the m variables. The coefficients of the discriminant function are represented by a column vector of the unknown lambdas. Lowercase Greek lambdas (λ) are used by convention to represent the coefficients of the discriminant function. These are exactly the same as the betas (β) also used by convention in regression equations. These should not be confused with the lambdas used to represent eigenvalues in principal components or factor analyses, nor the lambdas used to represent wave-length in a remote sensing sense.

The right-hand side of the equation consists of the column vector of m differences between the means of the two groups A and B in the simple linear discriminant analysis case. The equations can be solved by inversion and multiplication, such as

$$[\lambda] = [s_p^2]^{-1} \cdot [D]$$

or by the use of a simultaneous equation solution.

ASSUMPTIONS

The significance of the separation between the two groups can be tested, provided that certain assumptions are made regarding the nature of the data used in the discriminant function. These five basic test assumptions about the test data are as follows:

(1) The observations in each group are randomly chosen;
(2) The probability of an unknown observation belonging to either group is equal;
(3) The variables are normally distributed within each group;
(4) The variance-covariance matrices of the groups are equal in size; and
(5) None of the observations used to calculate the function were misclassified.

The most difficult assumptions to justify are (2), (3), and (4). However, the function is not seriously affected by limited departures from normality or by limited inequality of variances. The justification of (2) depends upon an *a priori* assessment of the relative abundance of the groups under examination (Davis, 1973).

TESTS OF SIGNIFICANCE

A test for the significance of the discriminant function is developed from the t-statistic mentioned earlier. A "distance" measure between the two multivariate means can be calculated by simply subtracting R_A from R_B. This is equivalent to substituting the vector of differences between the two group means into the discriminant equation, or setting the individual values of ψ_j equal to D_j. This distance measure is called *Mahalanobis' distance*, or the generalized distance, D^2. It is a measure of the separation between the two multivariate means expressed in units of the pooled variance. Hotelling's T-statistic of this distance has the form

$$T^2 = \frac{n_a n_b}{n_a + n_b} D^2.$$

The T-test can be transformed into a F-test, becoming

$$F = \frac{n_a + n_b - m - 1}{m(n_a + n_b - 2)} \frac{n_a n_b}{n_a + n_b} D^2$$

with m and $(n_a + n_b - m - 1)$ degrees of freedom. The null hypothesis tested by this statistic is that the two multivariate means are equal, or that the distance between them is zero; that is,

$$H_0: [D_j] = 0$$

against

$$H_1: [D_j] > 0.$$

The utility of this as a test of a discriminant function should be clear. If the means of the two groups are very close together, it will be difficult to separate them, especially if both groups have large variances. On the other hand, if the two means are well-separated and scatter around the means is small, discrimination is relatively easy.

REDUCTION OF DIMENSIONALITY

Not all of the variables included in the discriminant function are equally useful in distinguishing one group from another. Those variables that are not particularly useful can be isolated and eliminated from future analyses. Because discriminant analysis is so closely related to multiple regression, most of the procedures for selecting the most effective set of predictors can also be used to find the most effective set of discriminators. For example, the relative contribution of variable j to the distance between the two group means may be measured by a quantity E_j; i.e.,

$$E_j = \lambda_j D_j / D^2$$

where D_j is the difference between the jth means of the two groups. This is only *one measure* of the direct contribution of the variable, j, and does not consider interactions between variables. If two or more of the variables in the discriminant function are not independent, their interactions may contribute to D^2 to a greater extent than the value of E_j suggests. This measure serves roughly the same purpose as standardized partial regression coefficients in multiple regression. Values of E_j may be simply converted to percentages by multiplying by 100.