DOE/EY/22874-64

EXTENSIONS OF LINEAR DISCRIMINANT ANALYSIS FOR STATISTICAL CLASSIFICATION OF REMOTELY SENSED SATELLITE IMAGERY

PAUL SWITZER

TECHNICAL REPORT NO. 30

NOVEMBER 1979

PREPARED UNDER THE AUSPICES OF

SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

SIMS

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Prepared For
THE U.S. DEPARTMENT OF ENERGY
UNDER CONTRACT NO. EY-76-S-02-2874

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

EXTENSIONS OF LINEAR DISCRIMINANT ANALYSIS FOR STATISTICAL CLASSIFICATION OF REMOTELY SENSED SATELLITE IMAGERY

by

PAUL SWITZER

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

TECHNICAL REPORT NO. 30

NOVEMBER 1979

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH

PREPARED UNDER SUPPORT TO SIMS FROM

DEPARTMENT OF ENERGY (DOE)

ROCKEFELLER FOUNDATION

SLOAN FOUNDATION

ENVIRONMENTAL PROTECTION AGENCY (EPA)

NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

EXTENSIONS OF LINEAR DISCRIMINANT ANALYSIS FOR STATISTICAL CLASSIFICATION OF REMOTELY SENSED SATELLITE IMAGERY[†]

Paul Switzer
Stanford University

ABSTRACT

Linear discriminant analysis is a commonly used statistical tool for classification of surface features using staellite surface reflectance data. Extensions of this basic tool promise substantial improvements. In particular, we examine the added effectiveness of integration of spatial autocorrelation into the discriminant model, resolution of nonhomogeneous pixels, and data based prior probability estimates of class membership, and the use of unclassified pixels as part of the discriminant function "training" set.

[†] Presented to the International Geological Congress, Paris, July 1980.

INTRODUCTION

Digitized imagery from the Landsat earth satellite consists of an integrated energy measurement for each of four wavelength bands. This four-tuple, the data vector, is reported for surface area elements, pixels, which are about an acre in size. The statistics of the data vector are expected to depend on physical properties of the surface so that the data vector might be used to classify pixels according to surface type. This will usually require an ensemble of "training" pixels of known surface type for each of the contemplated categories to which pixels might be assigned. A convenient method for extracting some of the classification information in the training data is multiple linear discriminant analysis, in particular when no single data channel seems able to distinguish the surface categories of interest.

The method of linear discriminant analysis is in common use for this purpose, although it is sometimes recognized that it does not extract all the classification information that may be available. However, the ready availability of packaged software for linear discriminant analysis probably implies its longevity. This paper will examine how this standard analysis can be adapted using slight modifications to enhance its power in certain situations where the geographic alternation of surface categories is slow relative to pixel size, where the alternation is rapid enough that single pixels are frequently mixtures of two surface categories, and where the surface category frequencies are likely to vary substantially from one scene, or portion of a scene, to another.

The refinements discussed in this paper are not likely to yield dramatic improvements in the form of sharply reduced classification error rate. They are certainly secondary in importance to, say, careful preprocessing of data (nonlinear transformations of the data vector such as taking logarithms, forming ratios between channels, etc.), careful definition of surface categories, use of time repetition, careful selection of training pixels, and suitable geographic scale for a single analysis.

Also, this paper does not discuss what I call "many-parameter" classification schemes such as quadratic discrimination or density estimation methods. Sometimes they are much better than standard linear discriminant analyses, but since they are much more finely tuned to the training data, care must be taken not to exaggerate their performance in practice. Indeed, in seeking to modify standard linear discriminant analysis, an important goal has been to keep the dimensionality, i.e. number of adjustable parameters, as small as possible.

EXPLOITING SPATIAL CONTINUITY IN MULTI-PIXEL CLASSIFICATION

It is usual to classify a given pixel at position (x_1,x_2) using only the four-channel data vector $\underset{\sim}{Z}(x_1,x_2)$ for that pixel. However, the alternation of surface types is commonly on a scale larger than that of a single pixel. This spatial continuity of surface types should be exploited in the classification process. One possible approach is to explicitly augment the four-channel data attached to a given pixel by using additional variables corresponding to the four-channel data for neighboring pixels.

As an example we may introduce the augmented data vector $\underset{\sim}{Z}^*$ with eight components for a pixel at position (x_1,x_2) thus:

$$\underset{\sim}{Z}^*(x_1,x_2) = (\underset{\sim}{Z}(x_1,x_2),\ \underset{\sim}{Z}^e(x_1,x_2))'$$

where

$$\underset{\sim}{Z}^e(x_1,x_2) = 1/4[\underset{\sim}{Z}(x_1+1,\ x_2) + \underset{\sim}{Z}(x_1,\ x_2+1) + \underset{\sim}{Z}(x_1-1,\ x_2) + \underset{\sim}{Z}(x_1,\ x_2-1)] \ .$$

That is, $\underset{\sim}{Z}^e(x_1,x_2)$ is the average four-channel data vector for the four neighboring pixels as shown by the hatched area in Figure 1. In a like manner we could also define $\underset{\sim}{Z}^c(x_1,x_2)$ as the average of the data vectors for the neighboring pixels as shown by the shaded areas of Figure 1 and thereby further augment the length of the available data for classifying the pixel at (x_1,x_2).

One virtue of this approach is that we may apply standard methods of linear discriminant analysis to the higher-dimensional augmented data

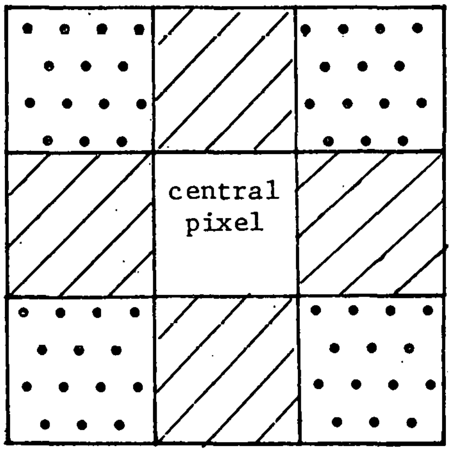

Figure 1. Scheme of neighbor pixels to be used in classification.

vectors $\underset{\sim}{Z}^*$, just as we might have done with the unadorned vectors $\underset{\sim}{Z}$. Furthermore, the augmented vector method is an approximation to an exact posterior analysis in the context of a particular probabilistic model, viz.

$$\underset{\sim}{Z}(x_1,x_2) = \sum_{i=1}^{K} \underset{\sim}{\mu}_i \, \underset{\sim}{\delta}_i(x_1,x_2) + \underset{\sim}{\varepsilon}(x_1,x_2)$$

where $\underset{\sim}{\delta}_i$ is a spatially correlated random indicator function for the surface category i, K = total number of categories into which classification can be made, $\underset{\sim}{\mu}_i$ is the mean four-channel vector for category i and $\underset{\sim}{\varepsilon}$ is a four-dimensional spatially correlated zero-mean random noise function. Now suppose the noise process $\underset{\sim}{\varepsilon}$ is modelled to be Gaussian with a locally spatially isotropic covariance and that the degree of spatial continuity is such that the probability is close to 1 that a pixel and its immediate neighbors all belong to the same category. Then the logarithms of the posterior probabilities that $\delta_i(x_1,x_2) = 1$, i.e. that the pixel at (x_1,x_2) belongs to category i, for $i=1, 2, \ldots, K$, given the augmented data $\underset{\sim}{Z}^*(x_1,x_2)$, is given approximately by the output of a standard linear discriminant analysis applied to the $\underset{\sim}{Z}^*$ data with specified prior probabilities. It is the local isotropy of the noise process which leads one to average the neighboring pixels and thus to reduce the dimension of the analysis immediately from 20 to 8.

A further reduction is possible when the noise covariances can be approximately factored in the following way (or equivalently the within-category covariances of the data vectors): the covariances between channel pairs for a single pixel are diminished by a constant factor for channel pairs from neighbor pixels. If s_{jk} is the covariance between channels j and k within the same pixel, then the covariance between channels j and k of different pixels would be γs_{jk}, where γ is an attenuation factor depending only on the distance between the two pixels.

In particular, the within-category covariance matrix for the augmented eight-dimensional data vector Z^* would then have the form

$$S^*_{8\times 8} = \begin{pmatrix} S & \beta S \\ \beta S & \alpha S \end{pmatrix}$$

where $S_{4\times 4}$ is the original four-channel within-category covariance matrix and α,β $(\alpha > \beta)$ are scaling constants between zero and one. One implication of this factorization is that the linear discriminant coefficients applicable to the channel data from neighbor pixels are proportional to the coefficients for the central pixel. It follows that a standard linear discriminant analysis may be performed on four-channel data $\underset{\sim}{Z}^{**}$ which is a fixed linear combination of the central pixel data and the neighbor data, viz.

$$\underset{\sim}{Z}^{**} = (\alpha-\beta)\underset{\sim}{Z} + (1-\beta)\underset{\sim}{Z}^{e} .$$

Hence the factored structure reduces the analysis further to a four-dimensional problem. [This reduction ignores the fact that the K category mean vectors for the neighbor pixels will be somewhat closer together than the category mean vectors for the central pixels. This happens because the neighbor pixels will not all belong to the same category as the corresponding central pixel. A more detailed analysis would result in diminished weights for neighbor pixels and would be consequential for spatially patchy classifications.] The virtue of such an approach is that the standard methods may be used on a low-dimensional analysis while exploiting spatial continuity in a moderately rigorous way.

Such factored covariance structure is not uncommon. Kowalik (1979) has calculated S^* for three-category data comprising 322 pixels. The ground-truth map is shown in Figure 2. His S^* matrix showed the factored structure with $\alpha = 0.70$ and $\beta = 0.60$, approximately. The closeness of α and β in this example indicates that the average of the neighboring pixels will get substantially more weight in the linear discriminant analysis than the central pixel.

The standard discriminant analysis based on Z^{**} for the Kowalik example gives an overall error rate of 34%. This should be compared with an error rate of 44% using only the central pixel for classification, i.e. ignoring spatial continuity. Alternate rows of pixels comprised the training data for this example and classification error rates were estimated independently from the intervening rows. The resulting estimated classifications are shown in Figure 3 and should be compared with each other and with ground-truth in Figure 2.

The gain is modest here but, in general, there is potential for substantial improvement in the error rate. In the Gaussian model it is possible to predict the improvement once the factors α,β have been estimated. This is illustrated most easily for the two-category case: if E is the error rate associated with using only the central pixel, then the inclusion of the neighbor pixels should reduce the error to

$$E^* = \Phi(\nu\Phi^{-1}(E)) \quad \text{where} \quad \nu^2 = (1-2\beta+\alpha)/(\alpha-\beta^2) ,$$

and Φ is the standard Gaussian cumulative distribution function. For example, if $\alpha = 0.70$ and $\beta = 0.60$, the above formula would have predicted that an error rate of 30% would be reduced to 26% using neighbor data in a two-category problem.

Key for Figures 2, 3, 4:

Volcanic unit Tertiary Ash Flow Tuff	V V V V V V
Limonitically altered Jurassic quartz monzonite	▩
Jurassic quartz monzonite (not limonitically altered)	□

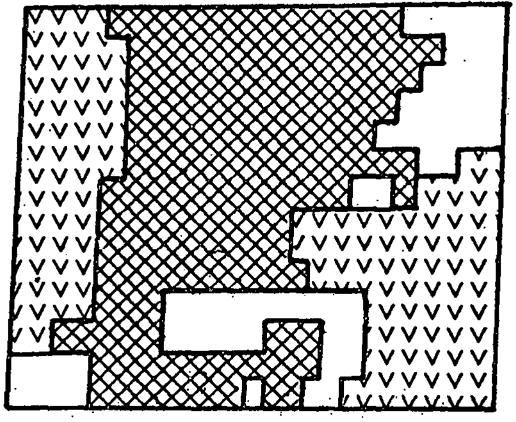

Figure 2. True rock type outcrop map.

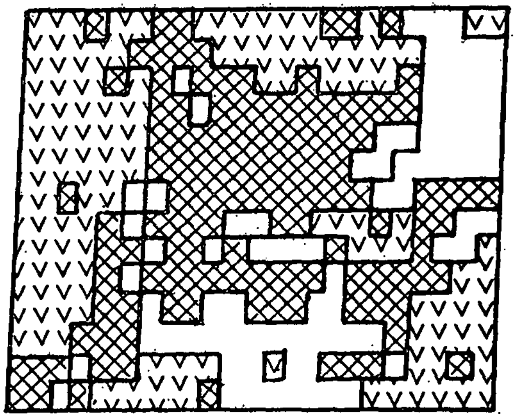

Figure 3. Rock type map estimated using the four Landsat channels, central pixels only.

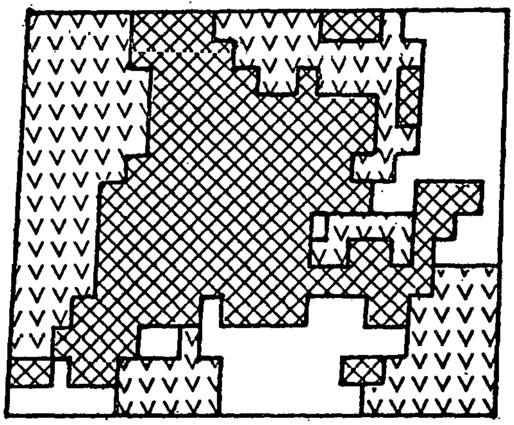

Figure 4. Rock type map estimated using the four Landsat channels, central pixels are adjacent neighbor pixels.

SPATIAL DISCONTINUITY: HETEROGENEOUS PIXELS

The framework of linear discriminant classification presupposes that each pixel has a unique category identification corresponding to one of the training categories. This is a reasonable framework when the scale of geographic continuity of category types is large relative to the pixel size. However, when a substantial number of pixels are likely to be mixtures of more than one category type another approach will be needed.

A number of ad hoc methods for the resolution of single-pixel mixtures have been tried. There is one method which corresponds to a maximum likelihood approach in the context of a multivariate Gaussian model. For example, consider the resolution of a pixel into two categories having multi-channel mean vectors $\underset{\sim}{\mu}_A$ and $\underset{\sim}{\mu}_B$, respectively. Suppose π_A and π_B $(\pi_A+\pi_B = 1)$ are the corresponding unknown areal proportions of the two categories for the given pixel. Then the multi-channel mean vector would be $\pi_A\underset{\sim}{\mu}_A + \pi_B\underset{\sim}{\mu}_B$.

It is difficult, however, to model exactly the appropriate multichannel covariance matrix for mixed pixels and in principle this covariance could depend on π_A, π_B. A great simplification results if the covariance $\underset{\sim}{S}$ is assumed to be approximately the same for any mixture proportions. This covariance simplification may be justified under certain circumstances but its main virtue is that it leads to tractable maximum likelihood estimates of the unknown single-pixel proportions for the two categories. In particular, for an observed pixel data vector $\underset{\sim}{Z}$, let $D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_A)$ and $D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_B)$ be the squared Mahalanobis distances between $\underset{\sim}{Z}$ and each of the category mean vectors and let $D^2(\underset{\sim}{\mu}_A, \underset{\sim}{\mu}_B)$ be the squared Mahalanobis distance between the two mean vectors, i.e.,

$$D^2(\underset{\sim}{Z},\underset{\sim}{\mu}) = (\underset{\sim}{Z}-\underset{\sim}{\mu})' \, \underset{\sim}{S}^{-1} \, (\underset{\sim}{Z}-\underset{\sim}{\mu}) \text{ , etc.}$$

Then the maximum likelihood estimates of the proportions are given by

$$\hat{\pi}_A = 0.5 + 0.5 \, \frac{[D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_B) - D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_A)]}{D^2(\underset{\sim}{\mu}_A, \underset{\sim}{\mu}_B)}$$

$$\hat{\pi}_B = 1 - \hat{\pi}_A \, .$$

If either $\hat{\pi}_A$ or $\hat{\pi}_B$ turns out negative, it is truncated to zero.

The implementation of this estimation procedure for the resolution of mixtures requires estimates of the separate category mean vectors $\underset{\sim}{\mu}_A$, $\underset{\sim}{\mu}_B$ and the within-category covariance matrix $\underset{\sim}{S}$. Such estimates would be available in situations where the training data included homogeneous pixels. After training on the homogeneous data one would then perform an ordinary two-category linear discriminant analysis on the mixed pixels in which the D^2 values for each pixel are part of the usual output. These D^2 values from the standard analysis would then be used in the simple calculation indicated above to get estimates of mixture proportions for each pixel.

As an example, Marsh et al. (1980) considered the classification of 17 pixels in Garfield Flat, Nevada, which were mixtures (in varying proportions) of a clay-silt playa interspersed with phreatophyte vegetation mounds. The training data consisted of 40 homogeneous pixels of unvegetated playa and a similar number of homogeneous phreatophyte pixels. The results using the estimation procedure of this section on the 17 pixels is summarized in Table 1, demonstrating another use of linear discriminant analysis for nonstandard problems.

TABLE 1

GARFIELD FLAT
%vegetation

ACTUAL	ESTIMATED	ACTUAL	ESTIMATED
36.0	32.0	30.5	31.0
23.0	25.5	41.0	36.0
38.5	38.0	25.5	26.5
36.0	36.0	15.5	23.5
31.0	29.0	28.0	32.5
28.5	36.5	41.0	40.5
33.5	23.0	54.0	64.0
25.5	16.0	43.5	48.0
36.0	29.0		

SIMPLE MODIFICATIONS FOR HETEROGENEOUS VARIABILITY

There is another way in which a modification of the usual probabilistic model leads to an easily performed modification of standard linear discriminant analysis. Suppose that some classes are much more dispersed about their mean vectors than other classes. The simplest formalization of this concept says that the channel-to-channel covariance matrix for surface category i has the form

$$\underset{\sim}{S_i} = \underset{\sim}{S}/C_i$$

where the C_i are positive constants determined up to a proportionality factor.

In the usual model the log-probability that a pixel with data vector $\underset{\sim}{Z}$ belongs to category i is a fixed linear function of the Mahalanobis distance $D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_i)$ between $\underset{\sim}{Z}_i$ and the category i mean vector $\underset{\sim}{\mu}_i$. With the above modification of the model this log-probability becomes a fixed linear function of

$$C_i \, D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_i) - \log C_i \; .$$

So the usual output of Mahalanobis distances from a standard linear discriminant analysis should be modified in cases where the C_i's are grossly unequal. For example, if one notices that the four-channel variances are approximately in the same proportion for all categories, then ratios of these variances will provide estimates of the C_i's.

One reason why some categories could be more dispersed about their mean vectors than other categories is that some categories are more generic in nature than others. An argument can be made, in such contexts, for treating a group of similar categories as a single category in the initial linear discriminant analysis. If a pixel is assigned to this supercategory then another linear discriminant analysis is needed to further refine the assignment. This hierarchical use of linear discriminant analysis has the advantage that each stage involves discrimination among only a few classes, but great care is needed in estimating error rates for hierarchical procedures.

ESTIMATION OF PRIOR PROBABILITIES WITH LINEAR DISCRIMINANT ANALYSIS

Standard linear discriminant analysis approximates a Bayes optimal classification rule in the context of the usual Gaussian model when the categories are equally likely a priori and where the objective is to maximize the total number of correct classifications. It is equivalent to the rule which assigns a pixel with data vector $\underset{\sim}{Z}$ to category i if the Mahalanobis distance $D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_i)$ between $\underset{\sim}{Z}$ and the category i mean vector μ_i is smaller than it is for any other category $j \neq i$.

Now suppose the classes are not equally likely but are represented in proportions $p_1, p_2, \ldots, p_K$ for the group of pixels to be classified. The Bayes rule is modified in this case as follows:

$$\text{Assign } \underset{\sim}{Z} \text{ to category } i \text{ if } D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_i) + 2|\ell n\, p_i| <$$

$$D^2(\underset{\sim}{Z}, \underset{\sim}{\mu}_j) + 2|\ell n\, p_j| \quad \text{for all } j \neq i .$$

Of course the category frequencies $p_1, p_2, \ldots, p_K$ are not known. Furthermore, they will be different from one scene to another and, except in special cases, it would not be appropriate to use the category frequencies of the training set of pixels.

However, it is possible to obtain estimates of the p_j, for a given scene, from a first pass of a standard linear discriminant analysis which assumes (tentatively) that the p_j are equal. Suppose this first pass assigns the proportion p'_k of the current scene to category k. From the training data we already would have estimates of the proportion of category j pixels which would (erroneously) be assigned to category k, say p_{jk}. Then it follows from the law of total probability that p'_k is an estimate of $\Sigma_j\, p_j\, p_{jk}$. Hence to obtain estimates of the unknown p_j, $j=1, 2, \ldots, K$, one solves the linear system

$$p'_k = \sum_{j=1}^{K} p_j\, p_{jk}, \quad k=1, 2, \ldots, K .$$

Having obtained estimates of the category frequencies $p_1, p_2, \ldots, p_K$, for a given scene, the approximate Bayes rule may now be used for that scene: the Mahalanobis distances obtained from the usual linear discriminant analysis are modified by the addition of $2|\ell n\ p_j|$. Some standard packaged programs, such as BMDP7M, allow "prior probabilities" to be specified, in which case the modification is automatically made by a second pass through the program using the estimated category frequencies as prior probabilities.

An important issue in the use of category frequency estimates is the size of the scene or pixel ensemble for which separate frequency estimates should be obtained. By successively applying the above method to small contiguous ensembles of about 25 pixels, say, one has an alternative approach to capturing the information contained in spatial continuity - although category frequency estimates may then become statistically unstable. In any event, such estimation of frequencies is heavily dependent on having the class mean vectors remain unchanged and may, therefore, not be robust. It could also happen that the solution of the above linear system may give estimates of the p_j which are out of the range of possible values, hence requiring special fix-ups.

Of course the Bayes rule which tries to maximize the total number of correct category assignments may not be what is really wanted where certain misassignments are more serious than others. In particular, when one tries to discern infrequent categories against a pervasive background then such Bayes rules will commonly assign all pixels to background. The problem is not with the rule but rather with a poor specification of the objective. It is usually worthwhile to try to specify an actual "loss" table where

Key:

Mahalanobis Distance	
0 - 4.0	▓
4.1 - 8.0	▒
8.1 - 12.0	░
12.1 - 16.0	⁘
16.1 → +	□

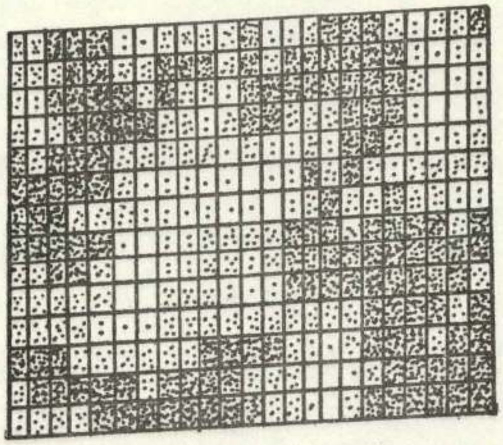

Figure 4a. Mahalanobis distance from the mean data vector for the volcanic rock unit.

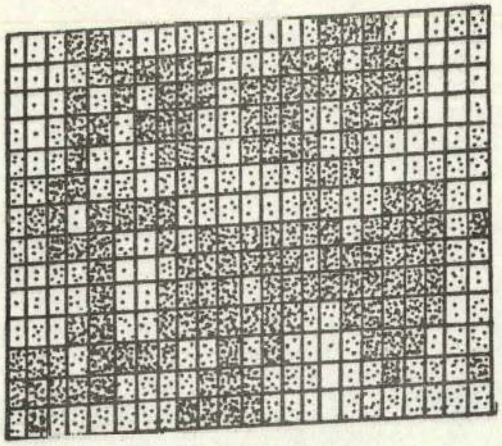

Figure 4b. Mahalanobis distance from the mean data vector for limonitically altered quartz monzonite.

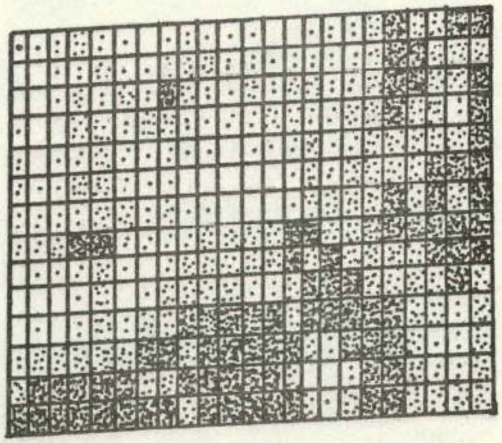

Figure 4c. Mahalanobis distance from the mean data for the quartz monzonite not limonitically altered.

L_{kj} is the loss associated with assigning a k-type pixel to the category j. Presumably L_{jj} would be a negative number, not necessarily the same for all j. The Bayes rule which minimizes the expected loss per pixel is

$$\text{Assign } \underset{\sim}{Z} \text{ to category } i \text{ if } \sum_{k=1}^{K} [L_{kj} - L_{ki}] \cdot P_k > 0$$

$$\text{for all categories } j \neq i$$

where P_k is the posterior probability that a pixel with data vector $\underset{\sim}{Z}$ belongs to the category k. This rule will not necessarily make assignments to the categories which are most probable.

GEOGRAPHICAL DISPLAYS OF PIXEL ASSIGNMENTS

One usually would wish to display the results of discriminant analyses in the form of geographic maps, especially for surfaces possessing a degree of spatial continuity. The most straightforward display assigns a color to each of the possible surface categories and colors each pixel according to its linear discriminant assignment. Figure 3 is an example of such maps. However, such maps suppress a great part of the information contained in the linear discriminant analysis since they convey no information about the certainty of category assignments. This shortcoming can be corrected by displaying a separate shade print map for each category i where the grey-level of a pixel is proportional, say, to the Mahalanobis distance between the pixel data vector and the category i mean vector. Figure 4 shows an example of such shade prints for the Kowalik data.

The foregoing is not to be confused with the usual graphical output of a linear discriminant analysis in which each pixel is plotted as a point in the plane of the first two canonical variables. While these plots show the scatter and overlap of the data for the various categories, the geographic relationships among the pixels is completely ignored in these plots. Canonical variable plots are, therefore, not so useful for geographic ensembles.

REFERENCES

Kowalik, W. S. (1979). Personal communication.

Marsh, S. E., Switzer, P., Kowalik, W. S., and Lyon, R. J. P. (1980). A Method for Resolving the Percentage of Component Terrains Within Single Resolution Elements. To appear in Photogrammatic Engineering and Remote Sensing.