N73-16489

LARS Print 030772

CASE FILE COPY

Minimum Distance Classification in Remote Sensing

by
A.G. Wacker and
D.A. Landgrebe

The Laboratory for Applications of Remote Sensing

Purdue University
Lafayette, Indiana

L.A.R.S. Print #030772

MINIMUM DISTANCE CLASSIFICATION IN REMOTE SENSING*

by

A. G. Wacker and D. A. Landgrebe

SUMMARY

The utilization of minimum distance classification methods in remote sensing problems, such as crop species identification, is considered. Minimum distance classifiers belong to a family of classifiers referred to as sample classifiers. In such classifiers the items that are classified are groups of measurement vectors (e.g. all measurement vectors from an agricultural field), rather than individual vectors as in more conventional vector classifiers.

Specifically in minimum distance classification a sample (i.e. group of vectors) is classified into the class whose known or estimated distribution most closely resembles the estimated distribution of the sample to be classified. The measure of resemblance is a distance measure in the space of distribution functions.

The literature concerning both minimum distance classification problems and distance measures is reviewed. Minimum distance classification problems are then categorized on the basis of the assumption made regarding the underlying class distribution.

Experimental results are presented for several examples. The objective of these examples is to: (a) compare the sample classification accuracy (% samples correct) of a minimum distance classifier, with the vector classification accuracy (% vector correct) of a maximum likelihood classifier; (b) compare the sample classification accuracy of a parametric with a nonparametric minimum distance classifier. For (a), the minimum distance classifier performance is typically 5% to 10% better than the performance of the maximum likelihood classifier. For (b), the performance of the nonparametric classifier is only slightly better than the parametric version. The improvement is so slight that the additional complexity and slower speed make the nonparametric classifier unattractive in comparison with the parametric version. In fact disparities between training and test results suggest that training methods are of much greater importance than whether the implementation is parametric or nonparametric.

INTRODUCTION

A fairly common objective of remote sensing in connection with earth resources is to attempt to establish the type of ground cover on the basis of the observed spectral radiance. The examination of systems capable of achieving this objective shows that a

---

*This paper was presented at the First Canadian Symposium for Remote Sensing, February 7-9, 1972, Ottawa, Canada. Dr. Wacker is Associate Professor, Department of Electrical Engineering, University of Saskatchewan, Sakatoon, Saskatchewan. Dr. Landgrebe is Director, Laboratory for Applications of Remote Sensing and Professor, Department of Electrical Engineering, Purdue University, West Lafayette, Indiana. The work described was sponsored by NASA under Grant No. NGL 15-005-112.

certain duality of system types exists. Landgrebe[1] refers to the two types as image-oriented systems and numerically-oriented systems. The duality exists primarily for historical reasons as a consequence of the independent development of photographically oriented and computer oriented technology. The primary distinction between the two system types is that in image oriented systems a visual image is an essential part of the analysis scheme while in numerically oriented systems the visual image plays a secondary role. In Fig. 1 the location of the "Form Image" block in relation to the "Analysis" block characterizes the two system types.

In numerically oriented remote sensing systems it is frequently possible to design the data collection system in such a manner that classification becomes a problem in pattern recognition. This situation prevails if one attempts to study earth resources through the utilization of multispectral data-images. The term multispectral image (i.e. without the modifier data) is used to refer to one or more spectrally different superimposed pictorial images of a scene. The modifier data is added to indicate that images are stored as numerical arrays as opposed to visual images.

To obtain a multispectral data-image of a scene, the scene in question is partitioned on a rectangular grid into small cells (pixels) and the radiance from each pixel for each wavelength band of interest is measured and stored. The set of measurements for a pixel constitutes the measurement vector for that pixel. A multispectral data-image for a scene is simply the complete collection of all measurement vectors for the image. The spatial coordinates (i.e. row and column number) of each pixel are of course also recorded to uniquely identify each measurement vector. Fig. 2 depicts the situation.

The methods used to generate multispectral data images can conveniently be divided into two categories. In the first category, film is used to record the image. The film is subsequently scanned and digitized to produce a data-image. The multispectral property is obtained either by scanning several images photographed through different spectral windows, and overlaying the data; or by utilizing color film and separating the spectral components during the scanning procedure. In the second category the image is generated electrically and stored in an electrically compatible form, usually on magnetic tape as either an analog or digital signal. The electrical signal to be stored can be generated by a number of different systems; the multispectral scanner and return beam vidicon probably qualify as the two most common examples. For the scanner the multispectral property is obtained by filtering of the spectral signal collected through a single aperture prior to recording, or by the superposition of several unispectral images collected through different apertures.

As already stated, pattern recognition techniques can serve as the basis for affecting classification of multispectral data-images. Much of pattern recognition theory is formulated in terms of multidimensional spaces with the dimensionality of the space equal to the dimensionality of the vectors to be classified. This vector dimensionality is, of course, determined by the number of attributes or properties of each pixel to be considered in the classification (e.g. number of spectral bands). Classifying a multispectral data-image by classifying the observation vectors from such an image on a pixel-by-pixel basis falls naturally into this common pattern recognition framework. In contrast to this vector-by-vector approach there are classification schemes which collectively will be referred to as "sample classification schemes". In such schemes all vectors to be classified are first segregated into groups (i.e. samples) such that all the vectors in a group belong to the same

class. The whole group of vectors is then classified simultaneously. The minimum distance method considered is one such classification scheme.

In utilizing sample classification schemes two distinct problems can be identified. The first is concerned with partitioning the measurement vectors into homogeneous groups, while the second is concerned with the classification of these groups. Except for the comments in the next paragraph consideration is restricted to the second problem.

It frequently occurs for multispectral data-images that many of the adjacent measurement cells belong to the same class. For example in an agricultural scene each physical field typically contains many pixels. In fact it is precisely this condition that prompts the investigation of sample classification schemes. In such situations the physical field boundaries serve to define suitable samples for problems like crop species identification, and it is in this context that sample classifiers might also be referred to as per-field classifiers. It is apparent that for the situation just described one method of automatically defining samples is to devise a scheme that automatically locates physical field boundaries in the multispectral data-imagery[2,3]. For the minimum distance classification results presented later, physical field boundaries will actually be used to define the samples, but the field boundaries are located manually rather than automatically. A second and perhaps more promising approach to the problem of defining samples is via observation space clustering. In this approach vectors from an arbitrary area are clustered in the observation space, and all the vectors assigned to the same **cluster constitute a sample irrespective of their location in the arbitrary choosen area. In this case the term <u>fields</u> no longer seems appropriate and consequently the term <u>sample classifier</u> is preferred over the term <u>per-field classifier</u>.**

It is apparent that sample classification schemes cannot be used in all situations where a vector-by-vector approach is possible. A basic requirement is that the data to be classified can either be segregated into homogeneous samples or occur naturally in this form. Where the minimum distance scheme can be applied it intuitively has several potential advantages over a vector-by-vector classifier; in particular it is potentially faster and more accurate.

It seems logical that provided the time required to automatically define the samples is not too great, then sample classifiers should be faster than a vector-by-vector classifier. This is of considerable importance in utilizing a numerically oriented remote sensing system to survey earth resources because a characteristic of such surveys is the tremendous volume of data involved. One would also anticipate that the vector classification accuracy (% vectors correctly classified) for vector-by-vector classifiers would be lower than the sample classification accuracy (% samples correctly classified) for sample classifiers. The reason for this is that in sample classifiers all the information conveyed by a group of vectors is used to establish the classification of each vector, whereas in vector-by-vector classifiers each vector is treated separately without reference to any other vector. In a sense sample classifiers utilize spatial information because vectors are classified as groups, which naturally have some spatial extent. No spatial information is used in vector-by-vector classifiers, consequently, sample classifiers should perform better since spatial information is certainly of some value.

MINIMUM DISTANCE CLASSIFICATION

Problem Formulation

In a certain sense minimum distance classification resembles what is probably the oldest and simplest ap-

proach to pattern recognition, namely "template matching". In template matching a template is stored for each class or pattern to be recognized (e.g. letters of the alphabet) and an unknown pattern (e.g. an unknown letter) is then classified into the pattern class whose template best fits the unknown pattern on the basis of some previously defined similarity measure. In minimum distance classification the templates and unknown patterns are distribution functions and the measure of similarity used is a distance measure between distribution functions. Thus an unknown distribution is classified into the class whose distribution function is nearest to the unknown distribution in terms of some predetermined distance measure. In practice the distribution functions involved are usually not known, nor can they be observed directly. Rather a set of random measurement vectors from each distribution of interest is observed and classification is based on estimated rather than actual distributions.

It is necessary to define more precisely what constitutes a suitable distance for minimum distance classification. Mathematically the terms "distance" and metric are used interchangeably. For our purpose it is convenient to distinguish between the two terms. In essence all that is required for a well-defined minimum distance classification rule is a measure of **similarity between distribution functions which need not necessarily possess all the properties of a metric. The term <u>distance</u> refers to any suitable similarity measure; the term metric is used in the normal mathematical sense. More specifically a metric** on a set S is a real valued function $\delta(.,.)$ defined on S X S (X indicates cartesian product) such that for arbitrary F,G,H in S

(a)	$\delta(F,G) \geq 0$	1
(b)(1)	$\delta(F,F) = 0$	2
(2)	If $\delta(F,G) = 0$ then $F = G$	3
(c)	$\delta(F,G) = \delta(G,F)$	4
(d)	$\delta(F,G) + \delta(G,H) \geq \delta(F,H)$	5

A distance, as used herein, is defined to be a real valued function $d(.,.)$ on S X S such that for arbitrary F,G,H in S at least metric properties a,b(1) and usually b(2) and (c) hold. For theoretical proofs it is in fact often desireable to require that d be a true metric while in practical application such a restriction is usually not necessary.

Not only are distances between individual distribution functions of interest but since each class could conceivably be represented by a set of distribution functions the distance between sets of distributions is also **of interest. Definition 1 defines the distance between sets of distributions.**

Definition 1 - Let the distance $d(F,G)$ be defined for all F,G, in A, where A is an arbitrary set of cdf's of interest. If A_1 and A_2 are nonempty subsets of A then the distance $d(A_1, A_2)$ between the sets A_1 and A_2 is defined as

$$d(A_1, A_2) = \inf_{\substack{F \in A_1 \\ G \in A_2}} d(F,G) \qquad 6$$

Note that Definition 1 applies to finite and infinite sets of distribution functions. Of course, if the sets are finite then taking the infimum is equivalent to taking the minimum.

Futhermore, if each set consists only of a single distribution function then the distance between the sets is precisely the distance between the distribution functions. The distance between a distribution function and a set of distribution functions is also included as a special case. It is necessary to make some comments about the usage of the notation $d(F,G)$. Some of the distance measures considered are expressed in terms of probability density functions (pdf's) rather than cumulative distribution functions (cdf's). The convention adopted is that the notation $d(F,G)$ is still used and referred to as the dis-

tance between cdf's, even though the distance is expressed in terms of the densities of F and G (i.e. in terms of f and g).

The minimum distance classification scheme can now be formally defined. It is convenient to use a decision theoretic framework for this purpose. In general to specify a problem in this framework it is necessary to specify:

(a) Z - the sample space of the observed random variable.

(b) Ω - the set of states of nature; that is, the set of possible cdf's of the random variable. If the functional form of the cdf is known, then Ω can be identified with the parameter space.

(c) $\overset{*}{A}$ - the action space; that is the set of actions or decisions available to the statistician.

(d) L (a,F) - loss function defined on $\overset{*}{A}X\Omega$ which measures the loss incurred if $F\varepsilon\Omega$ is the true state of nature and action $a\varepsilon\overset{*}{A}$ is the action taken.

The general formulation of the minimum distance problem in this framework follows:

(a) $Z = E^q$ (q-dimensional Euclidean space)

(b) $\Omega = [\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(k)}]$ where $\Omega^{(i)}$ is the set of possible distribution **functions for the ith class, i = 1, 2, ..., k.**

(c) $\overset{*}{A} = [a_1, a_2, \ldots, a_k]$ where a_i is the decision to decide the random sample to be classified belongs to the ith class, i = 1, 2, ..., k.

(d) $L(a,F) = 0$ if $F\varepsilon\Omega^{(i)}$ and action a_i was taken
$L(a,F) = 1$ otherwise.

A decision rule is a function defined on Z and taking values in $\overset{*}{A}$. The minimum distance decision rule is given by definition 2.

Definition 2 - Let Y be the vector of all sample observations. The minimum distance decision rule $D_{MD}: Z \to \overset{*}{A}$ is $D_{MD}(\underline{Y}) = a_i$ (i.e., decide the random sample to be classified belongs to class i) in case

$$d(\tilde{F}_N, \Lambda^{(i)}) = \min_{j = 1,\ldots,k} d(\tilde{F}_N, \Lambda^{(j)}) \qquad 7$$

Where $\Lambda^{(i)}$ is the set of cdf's selected to represent the ith class and $\tilde{F}_N$ is a sample-based estimate of the cdf of the random sample classified.

Several items in definition 2 require clarification. The vector Y includes not only the random sample to be classified, but also any other observations used in the classification procedure. For example, if training samples are used for each class, these are included in Y. The sets $\Lambda^{(i)}$ also require comment. $\Lambda^{(i)}$ may be the set of all possible distributions for class i (i.e. $\Lambda^{(i)} = \Omega^{(i)}$) or it may be a subset of $\Omega^{(i)}$ or the sample-based estimates of a set cdf's selected to represent class i. Finally the term sample-based estimate is used to refer to any estimate of a cumulative distribution function or its corresponding density which is based on a random sample from the distribution in question. A number of suitable estimators exist[4] and the present formulation does not restrict the type of estimator. Later attention will be focused on distance measures based on densities. In the parametric case the densities will be estimated by estimating the parameters describing the densities (parametrically estimated pdf's). In the nonparametric case density estimates will be based on histograms (density histogram estimation). To obtain a density histogram estimate of a pdf the observation space is partitioned into square bins and the probability density estimate in any bin is the percent of vectors used to estimate the density which fall in the bin.

A number of special cases of the

above formulation are not considered. These special cases are basically a consequence of making different assumptions regarding Ω, and $\Lambda = [\Lambda^{(1)}, \Lambda^{(2)}, \ldots, \Lambda^{(k)}]$. In Type I problems the sets of distribution functions representing the classes are assumed to be known sets. Actually, this problem is not of great interest from a practical point of view, since class distributions are not normally known, but it is interesting from a theoretical point of view because of its relative simplicity.

Type I - The $\Omega^{(i)}$'s are known sets of cdf's

Case (a) The sets $\Omega^{(i)}$ are infinite and $\Lambda^{(i)} = \Omega^{(i)}$

Case (b) The sets $\Omega^{(i)}$ are finite and $\Lambda^{(i)} = \Omega^{(i)}$

Case (c) The sets $\Omega^{(i)} = F^{(i)}$ (single cdf/class) and $\Lambda^{(i)} = F^{(i)}$

Type II problems differ from Type I problems in that the possible distribution functions for each class are known to be q-variate distributions but are otherwise unknown. Consequently, all distributions used in the minimum distance decision rule must be estimated. Since in practice only a finite number of estimated distributions can be utilized this factor must **be considered in formulating the problem. If the sets of states of nature (e.g. the $\Omega^{(i)}$'s) are infinite the infinite sets must somehow be replaced by a representative finite set. A similar attitude must be adopted if it is known <u>a priori</u> that the sets $\Omega^{(i)}$ are finite but it is not known precisely how many distribution functions each $\Omega^{(i)}$ contains (e.g. how many subclasses of wheat are there?); or even if the precise number is known, it may not be** known how to obtain a random sample for each distribution function (i.e. how are samples representing different subclasses of wheat selected?). Finally, in the finite case, even if a random sample for each distribution function of interest can be obtained, their number may be so large that for practical reasons it may be desireable to use a smaller number of representative distributions. Thus, the need **arises for a method to select a representative set of distribution functions from a larger (possibly infinite) set. To do this assign a distribution $H^{*(i)}$ to $\Omega^{(i)}$, i - 1, 2, ..., k. That is the events to which probability mass is assigned by $H^{*(i)}$ are sets of** distributions in $\Omega^{(i)}$. To select a random set of cdf's from $\Omega^{(i)}$ (i.e. to select a random set of training samples for the ith class) is now equivalent to selecting a random sample from $H^{*(i)}$.

The above formulation is rather complicated in that a distribution over a space of functions is involved. This complexity can be avoided by restricting consideration to a parametric family characterized by s real parameters. Making the logical assumption that a one to one correspondence exists between cdf's in $\Omega^{(i)}$ and points in the parameter space $\theta^{(i)}(\equiv E^s)$, it is apparent that assigning a distribution $H^{*(i)}$ to $\Omega^{(i)}$ is equivalent to assigning some other distribution $H^{(i)}$ to the parameter space $\theta^{(i)}$. Consequently, in the parametric case rather than deal with $H^{*(i)}$, which is a cdf on a set of distribution function, only $H^{(i)}$ which is a cdf in E^s need be considered.

It is perhaps worthwhile to restate the above ideas in terms of multispectral data-imagery from an agricultural scene before stating them in a more formal manner. In the interest of simplicity and since it is the case of primary interest assume that the true q-dimensional distribution of the radiance measurements from each field belong to the same parametric family which can be characterized in the parametric space E^s. This family may have **a finite or infinite number of members (i.e. subclasses). Further assume that all the fields in a class (e.g. wheat) can be described by a suitable distribution $H^{(i)}$ over the parameter space.** A set of training fields for each class

is selected at random. Because of our formulation this is equivalent to selecting a random sample from the parameter space according to the assumed distribution over the parameter space for that class (i.e. $H^{(i)}$). For each of the randomly selected training fields the radiance measurements are used to get an estimated cdf for that field. In this way estimated cdf's for a representative set of training fields are obtained for each class. **An unknown field is then assigned to the class that has a training field whose estimated cdf is nearest to the estimated cdf of the unknown field.** Since the problem as stated is parametric, one would normally, though not necessarily, use parametrically estimated cdf's.

Type II problems in which the $\Omega^{(i)}$'s are unknown are now formally described. While prime interest is centered in the case where Ω is a parametric family this restriction is not imposed in stating the problem. The description of Type II problems is complicated by the fact that the description of the sets $\Lambda^{(i)}$ is rather involved.

Type II - The $\Omega^{(i)}$'s are Unknown Sets of cdf's

Case (a) - The sets $\Omega^{(i)}$ are infinite in number and $\Lambda^{(i)} = \Omega_{M_i}{}^{(i)}$. The sets $\Omega_{M_i}{}^{(i)}$ are now described. First a set of population cdf's corresponding to a representative set of M_i training fields for class i, i = 1, 2, ..., k is selected. Let $\Omega_{M_i}{}^{(i)}$ be this set for the ith class. That is $\Omega_{M_i}{}^{(i)}$ is **a random sample of size M_i for $H^{*(i)}$. A sample-based cdf is then obtained for each cdf in $\Omega_{M_i}{}^{(i)}$ for i = 1, 2, ..., k. The resultant set of sample-based estimated cdf's is $\Omega_{M_i}{}^{(i)}$. For** the case where parametrically estimated cdf's are used $\Omega_{M_i}{}^{(i)}$ can also be considered to be a random sample of size M_i in the parameter space according to a distribution $H^{(i)}$.

Case (b) - The sets $\Omega^{(i)}$ are finite and $\Lambda^{(i)} = \Omega^{(i)}$ or $\Lambda^{(i)} = \Omega_{M_i}{}^{(i)}\ \Omega^{(i)}$. If the $\Omega^{(i)}$ are finite sets (i.e. finite number of subclasses) then it is desireable to let $\Lambda^{(i)} = \Omega^{(i)}$, where $\Omega^{(i)}$ is the set of sample-based estimated cdf's for the ith class. In cases where the resultant number of subclasses is impractically large and/or only a random set of M_i training fields is available it is necessary to let $\Lambda^{(i)} = \Omega_{M_i}{}^{(i)} \subset \Omega^{(i)}$ and proceed as in case (a).

Case (c) - The set $\Omega^{(i)} = F^{(i)}$ (Single cdf per class) and $\Lambda^{(i)} = F_N{}^{(i)}$.

Distance Measures

The importance in statistics of distances between cdf's has, of course, long been recognized; according to Samuel and Bachi[6] their use appears to fall into two broad categories.

(a) Used for descriptive purposes. For example, as an indicator to quantitatively specify how near a given distribution is to a normal distribution.

(b) Use in hypothesis testing, which is, of course, a special case of decision theory.

There is a tendency for distance functions sufficiently sensitive to detect minor differences in distribution functions (i.e. category (a) use) to be somewhat involved functions of the observations, with the result that their use as test statistics in hypothesis testing has been limited because of the complicated distribution theory. On the other hand, distance functions whose theory is simple enough to be readily used as test statistics often do not distinguish distribution functions sufficiently well. Since in minimum distance classification interest is naturally centered on good discrimination between distribution functions, therefore distance functions that fall into category (b) are nor-

mally used. Since the appropriate distribution theory for hypothesis testing is then in general not known it is impossible to theoretically compute probability of error, but it may be possible to establish reasonably tight upper bounds. The approximate probability of error can of course be determined experimentally.

The literature abounds with references to distance measures and no attempt will be made to give a complete bibliography. A representative sample of distance measures is given in Table 1. This Table includes the most widely used distance measures because of their obvious importance, as well as more obscure distance measures whose application to the present problem appears reasonable. In addition a few miscellaneous distance measures have been included to give an indication of the variety of distances that have been suggested. The distances included in this Table are: Cramer-Von Mises[7,8,9,10], Kolmogorov-Smirnov[11,12,9,10], Divergence[13,14,15], Bhattacharyya[15,16], Jeffreys-Matusita[13,14,17], Kolmogorov Variational[15,18,19], Kullback-Leibler[15,20], Swain-Fu[21], Mahalanobis[22,23], Samuels Bachi[6], and Kiefer-Wolowitz[24]. The references cited are by no means comprehensive. In selecting the references the attempt has been made to cite only the original source in addition to survey papers. The paper by Darling[9], Sahler[10] and to a certain extent Kalaith[15] fall in this latter category.

Most of the references cited are concerned only with the univariate forms of the distance measure. With the exception of the Samuels-Bachi distance, the extention to the multivariate forms is quite natural. Since it is the multivariate forms that are of interest, these, rather than the more common univariate forms, are given in Table 1. For the Samuels-Bachi distance multivariate forms other than the one presented may be possible.

Table 1 also contains information regarding the metric properties of the distance measures when used in conjunction with three families of distribution functions. The families considered are: C, the family of q-variate absolutely continuous distribution functions; MVN, the family of q-variate normal distribution functions; and MVN_Σ, the family of q-variate normal distribution functions with equal covariance matrices. Since MVN and MVN_Σ are subsets of C it is, of course, true that a metric in C is also a metric in MVN and MVN_Σ. A metric in MVN_Σ need not, however, be a metric in MVN or C.

Because of the importance of the multivariate normal distribution, expressions for the distance between two such distributions are given in Table 2 for each of the distances measured in Table 1 in those instances where the expressions are known.

The distances listed in Table 1 are discussed in the references cited and no attempt will be made to discuss them except for some general comments pertaining to their use in minimum distance classification.

Since a large variety of distance measures is available, the problem naturally arises as to which distance measure to use in a given problem. Unfortunately, no complete answer to this question is presently available, but some general comments are possible. The distribution-free properties* that make the Cramer-Von Mises and Kolmogorov-Smirnov distances so popular in the univariate case do not apply in the multivariate case. Since it is the multivariate case that is of interest these distances lose their special appeal. Intuitively a distance like the Kolmogorov-Smirnov distance does not appear to be as good a distance

*** In the univariate case the distribution of the Kolmogorov-Smirnov and the Cramer-Von Mises distances between two estimated distribution functions is independent of the underlying distributions being estimated, provided appropriate estimators are used.**

Table 1

Multivariate Forms of Distance Measures and Their Metric Properties

Name	Form	Metric in C	MVN	MVN$_\Sigma$
Cramer-Von Mises	$W = \{\int_{-\infty}^{\infty}(G(\underline{x}) - F(\underline{x}))^2 d\underline{x}\}^{\frac{1}{2}}$	Yes	Yes	Yes
Kolmogorov-Smirnov	$K = \mathrm{Sup}_{\underline{x}}\lvert G(\underline{x}) - F(\underline{x})\rvert$	Yes	Yes	Yes
Divergence	$J = \int_{-\infty}^{\infty}\mathrm{Ln}(\frac{f(\underline{x})}{g(\underline{x})})(f(\underline{x})-g(\underline{x}))d\underline{x}$	No	No	Yes
Bhattacharyya Distance	$B = -\mathrm{Ln}\int_{-\infty}^{\infty}(f(\underline{x})g(\underline{x}))^{\frac{1}{2}}d\underline{x}$	No	No	Yes
Jeffreys-Matusita Distance	$M = \{\int_{-\infty}^{\infty}(\sqrt{g(\underline{x})} - \sqrt{f(\underline{x})})^2 d\underline{x}\}^{\frac{1}{2}}$	Yes	Yes	Yes
Kolmogorov Variational Distance	$K(p) = \int_{-\infty}^{\infty}\lvert p_g g(\underline{x})-p_f f(\underline{x})\rvert d\underline{x}$	Yes	Yes	Yes
Kullback-Leibler Numbers	$L_{fg} = \int_{-\infty}^{\infty}\mathrm{Ln}(\frac{f(\underline{x})}{g(\underline{x})})\ f(\underline{x})d\underline{x}$	No	No	Yes
Swain-Fu Distance	$T = \frac{\lvert\underline{\mu}_f-\underline{\mu}_g\rvert}{D_f+D_g}$	No	No	Yes
	Where $D = \{\frac{\lvert\underline{\mu}_f-\underline{\mu}_g\rvert^2\ (q+2)}{tr(\Sigma^{-1}(\underline{\mu}_f-\underline{\mu}_g)(\underline{\mu}_f-\underline{\mu}_g)^t)}\}^{\frac{1}{2}}$			
Mahalanobis Distance	$\Delta = \{(\underline{\mu}_g-\underline{\mu}_f)^t\Sigma^{-1}(\underline{\mu}_g-\underline{\mu}_f)\}^{\frac{1}{2}}$			Yes
Samuels-Bachi Distance	$U = \{\int_0^1[F^{-1}(\alpha)-G^{-1}(\alpha)]d\alpha\}^{\frac{1}{2}}$	No	No	No
	where $F^{-1}(\alpha) = \mathrm{Inf}\{c\lvert Q_c\cap Q_\alpha\neq\emptyset\}$			
	and $Q_c = \{\underline{x}\lvert \sum_{i=1}^{q} x_i\leq c\}$, $Q_\alpha = \{\underline{x}\lvert F(\underline{x})\geq\alpha\}$			
Kiefer-Wolfowitz Distance	$V = \int_{-\infty}^{\infty}\lvert F(\underline{x})-G(\underline{x})\rvert e^{-\lvert x\rvert}d\underline{x}$	Yes	Yes	Yes

Notation

(1) F, G are multivariate cdf's with densities f, g; means $\underline{\mu}_f$, $\underline{\mu}_g$; covariances Σ_f, Σ_g; and prior probabilities p_f, p_g.

(2) $\int_{-\infty}^{\infty}()\ d\underline{x}$ designates a multivariate integral.

(3) For Mahalanobis distance F and G are normal with means $\underline{\mu}_f$ and $\underline{\mu}_g$ and have common covariance Σ.

(4) $\lvert\ \rvert$ designates the absolute value or vector norm.

(5) t designates the transpose.

Table 2

Distances Between Two Multivariate Normal cdf's

Name	Distance
Divergence	$J = \frac{1}{2}\,\mathrm{tr}[\Sigma_f-\Sigma_g][\Sigma_g^{-1}-\Sigma_f^{-1}] + \frac{1}{2}\,\mathrm{tr}[\Sigma_f^{-1}+\Sigma_g^{-1}][\underline{\mu}_f-\underline{\mu}_g][\underline{\mu}_f-\underline{\mu}_g]^t$
Bhattacharyya Distance	$B = \frac{1}{8}(\underline{\mu}_f-\underline{\mu}_g)^t[\frac{\Sigma_f+\Sigma_g}{2}]^{-1}(\underline{\mu}_f-\underline{\mu}_g) + \frac{1}{2}\,\mathrm{Ln}\frac{\det(\frac{1}{2}[\Sigma_f+\Sigma_g])}{\{\det(\Sigma_f)\det(\Sigma_g)\}^{1/2}}$
Jeffreys-Matusita Distance	$M = [2\{1 - \frac{\{\det(\Sigma_f)\det(\Sigma_g)\}^{1/4}}{\{\det(\frac{1}{2}[\Sigma_f+\Sigma_g])\}^{1/2}}\exp(-\frac{1}{8}(\underline{\mu}_f-\underline{\mu}_g)^t[\frac{\Sigma_f-\Sigma_g}{2}]^{-1}(\underline{\mu}_f-\underline{\mu}_g))\}]^{1/2}$
Kullback-Leibler Numbers	$L_{fg} = \frac{1}{2}\,\mathrm{Ln}\,\frac{\det(\Sigma_f)}{\det(\Sigma_g)} + \frac{1}{2}\,\mathrm{tr}\Sigma_f[\Sigma_g^{-1}-\Sigma_f^{-1}] + \frac{1}{2}\,\mathrm{tr}\Sigma_g^{-1}[\underline{\mu}_f-\underline{\mu}_g][\underline{\mu}_f-\underline{\mu}_g]^t$
Swain-Fu Distance	$T = \frac{\lvert\underline{\mu}_f-\underline{\mu}_g\rvert}{D_f + D_g}$ where $D = \{\frac{\lvert\underline{\mu}_f-\underline{\mu}_g\rvert^2(q+2)^{\frac{1}{2}}}{\mathrm{tr}\,(\Sigma.)^{-1}(\underline{\mu}_f-\underline{\mu}_g)(\underline{\mu}_f-\underline{\mu}_g)^t}\}$
Mahalanobis Distance	$\Delta = \{(\underline{\mu}_g-\underline{\mu}_g)^t\Sigma^{-1}(\underline{\mu}_g-\underline{\mu}_f)\}^{\frac{1}{2}}, (\Sigma=\Sigma_f=\Sigma_g)$

Notation

(1) t means transpose

(2) det means determinant

(3) tr means trace

(4) The normal distributions involved have means $\underline{\mu}_f$ and $\underline{\mu}_g$ and covariance matrices Σ_f and Σ_g.

measure as those involving integration over the whole space. It is also more difficult to compute in parametric situations then some of the integral relations. The Samuels-Bachi distance suffers from a similar computational disadvantage.

The Divergence, Bhattacharyya distance, Jeffreys-Matusita distance, Kolmogorov variational distance and Kullback-Leibler numbers all belong to a class of distance measures which can be written as the expected value of a convex function of the likelihood ratio*. In fact Ali and Silvey[25] have shown that the expected value of any convex function of the likelihood ratio has properties that might reasonably be demanded of a distance measure. In addition Wacker[4] has shown that in feature selection such distance measures have a weak relationship to the probability of error. Kalaith[15] proved the same relationship for Divergence and the Bhattacharyya distance. Since the class of distance measures under discussion is based on pdf's there is probably a tendency for these distances to reflect differences in pdf's rather than cdf's.

Of the distances based on likelihood ratios the Bhattacharyya distance seems to have been gaining in favor. The prime reason for this is apparently the close relation between probability of error and Bhattacharyya distance, as well as the relative ease of computing Bhattacharyya distance in theoretical problems. Other properties of the Bhattacharyya distance which enhance its prestige as a distance measure have been pointed out by Lainiotis[26] and Stein[27]. A property of considerable theoretical utility is the close relation between the Bhattacharyya distance B, the Jeffreys-Matusity distance M and the affinity ρ namely

$$M = 2(1-\rho)^{1/2} = 2(1-e^{-B})^{1/2} \qquad 8$$

Where

$$B = -\mathrm{Ln}\rho \qquad 9$$

* The likelihood ratio of densities $f(\underline{x})$ and $g(\underline{x})$ is $f(\underline{x})/g(\underline{x})$.

and

$$\rho(F,G) = \int_{-\infty}^{\infty}(f(\underline{x})g(\underline{x}))^{1/2}d\underline{x} \qquad 10$$

Because of the above relationships minimum distance classifications made on the basis of the Bhattacharyya distance, Jeffreys-Matusita distance or affinity all yield identical results, and consequently have identical probability of error.

The Jeffreys-Matusita distance is, however, a metric in a much larger class of distributions (see Table 1). This means that theoretical derivations regarding probability of error can be made using the metric properties of the Jeffreys-Matusita distance in this larger class, and the results are applicable if classification is effected using Bhattacharyya distance or affinity as well. This property has been used extensively by Matusita.

While no strong preference for any distance measure can presently be demonstrated the theoretical properties of the Bhattacharyya distance suggests that it might be a reasonable choice and the experimental results presented later are based on this distance measure.

Minimum Distance Classification And Probability of Error

Considerable literature exists on the minimum distance method with Matusita[28-35] and Wolfowitz[36-39] being the chief contributors. Wolfowitz's work is concerned primarily with estimation while much of Matusita's work deals with the decision problem. Contributions have also been made by Gupta[40], Cacoullous[41,42], Sirvastava[43] and Hoeffding and Wolfowitz[44].

In considering minimum distance **decision rules a common requirement is to insist that by using arbitrarily large samples the probability of misclassifying a sample can be made arbitrarily small. This is the notion of consistency and it is a reasonable demand if the pairwise distance between all the sets of distributions**

associated with each class is greater than zero or

$$d(\Omega^{(i)}, \Omega^{(j)}) > 0$$

for all $i, j = 1, 2, \ldots, k; i \neq j$ 11

In parametric problems in which some distribution is assigned to the parameter space the condition specified by 11 is equivalent to requiring that there is no overlap of regions of the parameter space associated with different classes.

It has been shown[40,34,44] that any minimum distance classification problem for which equation 11 holds is consistent (probability of misclassification approaches zero as sample sizes approach infinity) provided the distance and distribution estimator utilized satisfy certain conditions. These conditions are that the distance used must be essentially a metric (metric property b(2) need not hold) and that for the particular distance measure and estimator used, the probability that for the particular distance measure and estimator used, the probability that the distance between the true and estimated distribution **can be made arbitrarily small is one for infinite sample size. Further it is shown that certain distances and estimators satisfy these conditions. In particular in the normal case these conditions are satisfied by using para**metrically estimated densities and the Bhattacharyya distance[35]. Similar consistency results are not known for density histogram estimators. The known properties of consistency are summarized more rigorously and in greater detail by Wacker[4].

It is the property of consistency described in the previous paragraphs which makes the minimum distance decision rule potentially so attractive. In essence consistency says that if the condition specified by 11 is satisfied, and if sufficiently large samples are used then the probability of misclassifying a sample should be very small. Unfortunately in classifying multispectral data-images two problems arise.

(1) The number of distributions associated with any class is very large (perhaps almost infinite) and it is not practical to attempt to store all possible subclass distributions as is essentially assumed in deriving the consistency result described.

(2) It appears that the condition of equation 11 is frequently not satisfied, or at least that distributions from different classes are often so nearly alike that the number of samples required to distinguish them is impractically large.

When the condition specified by equation 11 is violated to the extent that $\Omega^{(i)}$ and $\Omega^{(j)}$ overlap on a set of non zero probability then the minimum distance decision rule can obviously no longer be consistent; in this situation the probability of misclassifying a sample will be finite regardless of sample size. Under these circumstances, except for the simple parametric example treated by Wacker[4], essentially no results are available.

RESULTS

Three different classifiers were used to obtain the experimental results. These classifiers are known as LARSYSAA, PERFIELD and LARSYSDC. **LARSYSAA is a vector-by-vector classifier based on the maximum likelihood decision rule[45], while PERFIELD and LARSYSDC are minimum distance classi**fiers utilizing the Jeffreys-Matusita or equivalent (Bhattacharyya) distance. LARSYSAA and PERFIELD are based on the Gaussian assumption and utilize parametrically estimated pdf's while LARSYSDC utilize density histograms to estimate the pdf's. All three classifiers assume equal subclass probabilities and operate in the supervised mode*.

* Supervised refers to the fact that samples whose classification are known are available to "train" the classifier.

Two examples are discussed. The first example compares the sample classification accuracy (% samples correct) of a parametric with a non-parametric minimum distance classifier. The second example compares the vector classification accuracy (% vectors correct) of the parametric maximum likelihood classifier LARSYSAA with **the parametric minimum distance classifier PERFIELD. The data used in both examples are essentially the same but as subsequently described the training procedures differ considerably.**

The two examples discussed are problems in species identification of agricultural fields. In this context it is usually logical to assume that all the measurement vectors from a given physical field belong to the same class. This assumption was made in defining samples for the minimum distance classifiers and in determining the classification accuracy of the maximum likelihood classifier. In other **words, for the minimum distance classifiers each sample to be classified represents a physical field, while for the maximum likelihood classifier all vectors from a field are assumed to belong to the same class.**

The data for the examples to be discussed has 13 spectral bands and was collected by the University of Michigan Scanner. For ease in referring to different spectral bands the wavelength channel number correspondence of Table 3 is utilized. The data was collected at an altitude of 3000 ft., between 9:45 and 10:45 a.m. E.D.T., on June 30, 1970, from Purdue University flightlines 21, 23 and 24 respectively. The exact location and orientation of these flightlines, which are located in Tippecanoe County, Indiana, is shown in Fig. 3. The flightlines extend the 24 mile length from the north to the south end of the county and are roughly equally spaced in the east-west direction. Since the scanner geometry is such that at an altitude of 3000 feet the field of view is roughly 1 mile, the area covered by the three flightlines, approximately 72 square miles, is about 1/7 of the total area in the county. The scanner resolution and sampling rate are nominally three and six milliradians respectively. This means that at nadir the scanner "sees" a circle about 9 feet in diameter and that the spacing between adjacent pixels is about 18 feet. Since the scanner resolution and sampling rate are independent of look angle the distance between adjacent pixels is approximately 30% larger at the edge of the scanner's field of view with a corresponding change in the shape and area "seen" by the scanner. At the sampling rate indicated there are 220 samples across the width of a flightline and each flightline contains 5000 to 6000 lines. This means each flightline contains somewhat more than 10^6 pixels of which 10% to 20% are typically used for test purposes.

For both examples four principle ground cover categories are considered; wheat, corn, soybeans and other. Although the other class includes a considerable variety of ground cover most of the agricultural fields in this category are either small grains (other than wheat) or forage crops. There are also some bare soils and diverted-acre fields. Some natural categories such as trees and water are also included in this class. For most of the subcategories for the class other ground cover is fairly complete, but the spectral properties of the ground cover are quite variable from field to field within a subcategory. Most of the wheat in the flightline was natyre abd readt for

was mature and ready for harvest. In fact some portion of it had already been harvested. For corn and soybeans the crop canopy at flight time was such that the ground was not covered by vegetation when viewed from above and consequently the radiance is greatly influenced by the soil type. This fact makes it difficult to discriminate corn and soybeans at this time of year and consequently high classification accuracies are not to

be expected, especially since corn and soybeans constitute a considerable fraction of the ground cover.

Table 3

Correspondence Between Channel Numbers and Spectral Bands

Channel Number	Spectral Band (Micrometers)
1	0.40-0.44
2	0.46-0.48
3	0.50-0.52
4	0.52-0.55
5	0.55-0.58
6	0.58-0.62
7	0.62-0.66
8	0.66-0.72
9	0.72-0.80
10	0.80-1.00
11	1.00-1.40
12	1.50-1.80
13	2.00-2.60

While the particular training procedure used in each example is different some general observations are possible. It is evident that some of the variables which affect radiance tend to be constant within a physical field, but vary from field to field. Such variables are usually related to farm management practices and include such factors as variety of species, fertilization rates, crop rotation practices, **etc. Also the variability in soil type can normally be expected to be greater between fields than within fields. Consequently it is not uncommon for all data from one field to be fairly** "uniform" but still be quite different from the data from another field; even though the class (species) is the same in both fields. In terms of probability densities the density from each **individual field might reasonably be approximated by a normal distribution; in that it is typically unimodal and reasonably symmetrical, but the data** from several fields combined frequently exhibit severe multimodality. Under these circumstances, in order that the Gaussian assumption is approximately satisified (for classifiers making this assumption), subclasses are usually defined for each main class, such that the distribution for each subclass is unimodal. Perhaps if data from a sufficient variety of fields could be combined for a given crop species a unimodal distribution would result for each main class and the definition of subclasses would not be necessary, even for a parametric classifier. The class distribution in this case would naturally be broader than the distribution of any "subclass" of which it is composed. It is presently not known in the above situation whether better classification is achieved with parametric (Gaussian) classifiers by using many subclasses whose distribution are relatively narrow, or using fewer subclasses with broader distribution. In practice there appears to be a tendency toward the definition of many subclasses. In nonparametric classifiers it should of course not be necessary to define subclasses as there is no need for densities to be unimodal.

On the basis of the above discussion a fairly general parametric model which at least qualitatively behaves much like the actual multispectral data results when every field associated with each main class is considered as a potential subclass. The variation in distribution parameters from field to field is accounted for by a distribution over the parameter space. This is precisely the problem previously formulated at Type II case (a).

Example 1 - Parametric vs Nonparametric

The classifications performed for this example can be segregated into the four categories shown below.

1) Classifications with the parametric classifier PERFIELD
a) Every training field treated as a subclass.
b) Data from all training fields for each principle class combined (no subclasses).

2) Classifications with the nonparametric classifier LARSYSDC

a) Every training field treated as a subclass.
b) Data from all training fields for each principle class combined (no subclasses).

In the classification procedure each flightline was treated as a separate data set. The training and classification method is described for one flightline with other flightlines receiving similar treatment. Initially test and training data must be defined. Every field of any significant size whose classification had been determined by field observation was included as a possible test or training field. These fields were segregated into the four principle classes. Roughly 10% of the fields in each class were then selected at random to serve as training fields. The remaining fields were used as test fields. Table 4 gives a break down of the number of test and training fields for each flightline. After the training fields had been selected the subclass or class densities were estimated and stored. The test fields were then classified on the basis of their estimated densities by the minimum distance rule. The computations to estimate a density function for PERFIELD are substantially simpler than for LARSYSDC since for PERFIELD only the mean and covariance need be estimated while for LARSYSDC the density histogram must be generated. A bin size of 5 was used for the density histograms in PERFIELD. (The data ranges was 0 to 256). Only 3 of the 13 channels were used in performing the classifications. These were selected in a more or less arbitrary manner, although it was known that the selected set (1,8,11) were among the better subsets of channels.

Table 4

Number of Test and Training Fields

Flight-line	Number of Test(Training) Fields Total	Wheat	Corn	Soy-beans	Other
21	218(22)	23(2)	79(8)	57(6)	59(6)
23	141(15)	18(2)	58(6)	55(6)	10(1)
24	156(18)	19(2)	52(6)	43(5)	42(5)

The results of the classification are shown in Fig. 4. Rather than present the classification results for each flightline individually the performance averaged over the three flightlines is given. The results therefore give some indication of the classification accuracy one might expect on the average for this type of data for the training method used. In view of the random nature of the training procedure it is felt that this is a more meaningful presentation than quoting the results for each flightline individually.

Example 2 - Maximum likelihood vs Minimum Distance Classification

For this example the data from flightlines 21, 22, and 23 was classified using:
a) The parametric maximum likelihood classifier LARSYSAA.
b) The parametric minimum distance classifier PERFIELD.

The training procedure in this case is considerably different than the procedure for Example 1. In this case small areas approximately one acre in size were selected from flightlines 21, 23, and 24 on this basis of a sampling scheme. The sampling scheme simply used every nth acre in the flightline belonging to the class in question as a "training acre". The data from the acres selected in this manner was used to train the classifier. In this manner 59 wheat acres, 44 corn acres, 23 soybean acres and 46 other acres were selected. The sampling rate n was different for the various principle classes. If every training acre were treated as a separate subclass a total of 172 subclasses result. This number exceeds the capabilities of the classification programs. Consequently it was necessary to reduce the number of subclasses to a reasonable number. This was accomplished by means of a clustering program which groups together the acres within each principle class whose estimated pdf's are similar[4]. As a result of this grouping the number of subclasses defined for the principle

classes: Wheat, Corn, Soybeans and Other were 4, 10, 6 and 10 respectively. Density histogram estimates of the resulting 4 wheat subclasses are shown in Fig. 5. Note that even after clustering considerable evidence of multimodality still exists, particularly for the first subclasses. In fact in some channels the contribution of all 4 acres assigned to subclass 1 are clearly evident. It is possible that this data should have been segregated into a greater number of subclasses. After the subclasses had been defined by clustering the statistics (means and covariance) were computed for each subclass. The feature selection capability of LARSYSAA[45] was then used to select the "best" 4 of the 13 channels for classification. This selection is based on the average divergence between all possible subclass pairs, excluding subclass pairs from the same class. On this basis channels 2, 8, 11, and 12 were selected. Using these channels both the training acres as well as the test fields were classified both with LARSYSAA and PERFIELD. The classification results for the training acres are shown in Fig. 6 while the results for the test fields (again averaged over the 3 flightlines) are shown in Fig. 7.

Discussion of Experimental Results

It is suggested that in evaluating a classifier a reasonable index of comparison is the overall average classification accuracy. This performance index has the advantage that it gives an indication of the classification accuracy that might be expected from the classifier for similar data and training procedures. For a relatively small data set, it is usually relatively easy to devise a training procedure or classifier which superficially looks superior but whose apparent superiority disappears when results are averaged over a number of data sets. A disadvantage of the suggested performance index is the necessity to do a reasonable number of classifications.

On the basis of average classification accuracy and the training procedures used there is no evidence that the parametric minimum distance classifier is superior to the nonparametric classifier. Neither is there any evidence that using a relatively large number of subclasses improves classification accuracy on the average. This is contrary to expectations.

Actually when each field is treated as a subclass one would expect **the nonparametric classifier to perform better than the parametric classifier only if the Gaussian assumption was seriously violated for the various training or test fields involved.** Futhermore, for the nonparametric classifier to exhibit any real advantage the nonnormal structure of the data must bear some resemblence from field to field (e.g. modes must appear in same relative positions). Since the nonparametric classifier does not exhibit any superior performance neither of the above factors apparently occur with any consistency.

When the data from all the training fields is grouped one would expect that the data would be multimodal and that the nonparametric classifier would be much superior. The basic fallacy in this reasoning appears to be that although the class distributions are multimodal the samples to be classified are usually unimodal. In other words the distribution of any sample to be classified is not really a random sample from the distribution of any class. Instead it simply tends to account for one of the modes in the class distribution. Futhermore, there is no apparent way of rectifying this situation within the constraint of minimum distance classification.

The fact that the parametric classifier does so well (comparatively) when no subclasses are considered attests to the robustness* of the

* A robust classifier is relatively insensitive to the underlying assumptions about the distributions involved.

Gaussian assumption in minimum distance classification.

It must be recognized that in assessing a classifier factors other than the performance index considered are of importance. One other factor that should be considered is the consistency of the results. That is, how near to the average can one expect to get for any given classification. The variance in the average performance is a measure of this consistency. In this regard, although the number of classifications is small, there is evidence that the nonparametric classifier is better than the parametric version and that for the parametric classifier the variance in average performance is increased by combining the data from many fields. This small advantage hardly warrants the additional complexity of the nonparametric implementation.

The results comparing the minimum distance and maximum likelihood classifiers **show fairly conclusively that in general the sample classification accuracy of minimum distance classifiers is higher than the vector classification accuracy of maximum likelihood** classifier of the same data. This is true for both the test and training data. It is recognized of course that the quantities being compared are by nature somewhat different but nevertheless they represent the natural method of expressing the classification accuracy of each classifier individually and do afford some measure of comparison. This result agrees with expectations although a greater improvement might have been anticipated.

It is convenient to define the difference between the sample classification accuracy and the vector classification accuracy as the improvement factor. The exact value of the improvement factor depends on the particular data but qualitatively it is obvious that for Type II case (a) problems the improvement will be very small or non existent both when the separation of the parameter space densities for all classes is large (one can't improve a high vector classification accuracy much) as well as when no separation exists (subclasses of different main classes can then not be distinguished by either classifier). The experimental evidence suggest that for moderate overlap of the parameter space densities the improvement factor will be of the order of 5% to 10%.

In concluding it should be mentioned that no comparative computation times have been given. The fact that the experiments involved a number of different programs, two computer systems (one in a time sharing mode) and the inherent dependence of processing time on the Classification Parameters and on the manner in which the data is stored (data retrieval time is by no means negligible) makes it virtually impossible to give meaningful comparative times. Suffice it to say that to classify a typical flightline time would be measured in fractions of an hour to hours on an IBM 360 System Model 44, and that PERFIELD is the fastest classifier, followed by LARSYSDC and LARSYSAA in that order.

CLOSURE

Although only two examples have been presented numerous other classifications have been performed on similar data and the results generally support the results presented. Even considering only the classification discussed the volume of data involved is quite substantial and is certainly adequate for a reasonable test.

For the type of data considered two basic conclusions appear reasonable.

(1) The classification accuracy of a nonparametric minimum distance classifiers, utilizing density histograms for estimating pdf's, is on the average not any larger than the classification accuracy of the parametric (Gaussian) classifier based on parametrically estimated pdf's. The variability in

performance of the nonparametric classifier appears somewhat smaller. Since the parametric classifier requires less storage and is faster than the nonparametric classifier the latter classifier is not an attractive alternative.

(2) The average sample classification accuracy of a parametric (Gaussian) minimum distance classifier is larger than the average vector classification accuracy of a miximum likelihood vector classifier. Ignoring the problem of sample definition the minimum distance classifier is faster and is an attractive alternative to the maximum likelihood classifier in situations where it can be utilized.

The disparity between test and training results for both minimum distance and maximum likelihood classifiers is much greater than the difference due to classifier type or the specific implementation. This suggests that given the present state of the art greater improvement in classification accuracies will probably result from investigations intended to improve the training procedure than from investigation of classifier types.

REFERENCES

1. D.A. Landgrebe, "Systems Approach to the Use of Remote Sensing", LARS Information Note 041571, Purdue University, Lafayette, Indiana, April, 1971.

2. A.G. Wacker and D.A. Landgrebe, "Boundaries in Multispectral Imagery by Clustering," 1970 IEEE Symposium on Adaptive Processes (9th) Decision and Control, pp. X14.1-X14.8, December, 1970.

3. R.L. Kuehn, E.R. Omberg and G.D. Forry, "Processing of Images Transmitted from Observation Satellites, "Information Display, Vol. 8, No. 5, pp. 13-17, September/October, 1971.

4. A.G. Wacker, "Minimum Distance Approach to Classification," Ph.D. Thesis, Purdue University, Lafayette, Indiana, January, 1972. Also available as LARS Information Note 100771, Purdue University, Lafayette, Indiana, October, 1971.

5. Z.W. Birbaum, "Distribution Free Tests of Fit for Continuous Distribution Functions," Ann. Math. Stat., Vol. 24, pp. 1-8, 1953.

6. E. Samuel and R. Bachi, "Measures of Distances of Distribution Functions and Some Applications," Metron, Vol. 23, pp. 83-122, December, 1964.

7. H. Cramer, "On the Composition of Elementary Errors," Skand. Aktuarietids, Vol. 11, pp. 13-74 and 141-180, 1928.

8. R. Von Mises, "Wahrscheinlichkeitsrechnung," Leipzig-Wein, 1931.

9. D.A. Darling, "The Kolmogorov-Smirnov, Cramer-Von Mises Tests," Ann. Math. Stat., Vol. 28, pp. 823-838, December, 1957.

10. W. Sahler, "A Survey on Distribution-Free Statistics Based on Distances Between Distribution Functions," Metrika, Vol. 13, pp. 149-169, 1968.

11. A.N. Kolmogorov, "Sulla Determinazione Empirica Di Une Legge Di Distribuzione," Giorn, dell'Insi

Distribuzione," Giorn, dell'Insit. degli att., Vol. 4, pp. 83-91, 1933.

12. N.V. Smirnov, "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent

Samples," Bull. Math. Univ. Moscow, Vol. 2, pp. 3-14, 1939.

13. H. Jeffreys, "An Invariant for the Prior Probability in Estimation Problems," Proc. Roy. Soc. A., Vol. 186, pp. 454-461, 1946.

14. H. Jeffreys, "Theory of Probability," Oxford University Press, 1948.

15. T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection," IEEE Trans. on Comm. Tech., Vol. COM-15, pp. 52-60, February, 1967.

16. A. Bhattacharyya, "On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions," Bull. Calcutta Math. Soc., Vol. 35, pp.99-109, 1943.

17. K. Matusita, "On the Theory of Statistical Decision Functions," Ann Instit. Stat. Math. (Tokyo), Vol. 3, pp. 17-35, 1951.

18. B.P. Adhikari and D.D. Joshi, "Distance Discrimination et Resume Exhaustif," Pbls. Inst. Stat., Vol. 5, pp. 57-74, 1956.

19. C.H. Kraft, "Some Conditions for Consistency and Uniform Consistency of Statistical Procedures," University of California Publications in Statistics, 1955.

20. S. Kullback and R.A. Leiber, "On Information and Sufficiency," Ann. Math. Stat., Vol. 22, pp. 79-86, 1951.

21. P.H. Swain and K.S. Fu, "Nonparametric and Linguistic Approaches to Pattern Recognition," LARS Information Note 051970, Purdue University, Lafayette, Indiana, June, 1970.

21. P.H. Swain and K.S. Fu, "Nonparametric and Linguistic Approaches to Pattern Recognition," LARS Information Note 051970, Purdue University, Lafayette, Indiana, June, 1970.

22. P.C. Mahalanobis, "Analysis of Race Mixture in Bengal," J. Asiat. Soc. (India), Vol. 23, pp. 301-310, 1925.

23. P.C. Mahalanobis, "On the Generalized Distance in Statistics," Proc. Nat'l. Inst. Sci. (India), Vol. 12, pp. 49-55, 1936.

24. J. Keifer and J. Wolfowitz, "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," Ann. Math. Stat., Vol. 27, pp. 887-906, 1956.

25. S.M. Ali and S.D. Silvey, "A General Class of Coefficients of Divergence of one Distribution From Another," J. Roy. Stat. Soc., Ser. B, Vol. 28, pp. 131-142, 1966.

26. D.G. Lainiotis, "On a General Relationship Between Estimation, Detection, and the Bhattacharyya Coefficient," IEEE Trans. on Information Theory, Vol. IT-15, pp. 504-505, July, 1969.

27. C. Stein, "Approximations of Improper Prior Probability Measures," Dept. of Statistics, Stanford University, Stanford, California, Tech. Report 12, 1964.

28. K. Matusita, "On Theory of Statistical Decision Functions," Ann. Inst. Math. (Tokyo), Vol. 3, pp. 17-35, 1951.

29. K. Matusita, "On Estimation by the Minimum Distance Method," Ann. Inst. Stat. Math. (Tokyo), Vol. 5, pp. 59-65, 1954.

30. K. Matusita, Y. Suzuki, and H. Hudimoto, "On Testing Statistical Hypothesis," Ann. Inst. Stat. Math. (Tokyo), Vol. 6, pp. 133-141, 1954.

31. K. Matusita and H. Akaike, "Decision Rules Based on the Distance for the Problems of Independence Invariance and Two Samples," Ann. Inst. Stat. Math., Vol. 7, pp. 67-80, 1956.

32. K. Matusita and M. Motoo, "On the Fundamental Theorem for the Decision Rule Based on Distance || ||," Ann. Inst. Stat. Math., Vol. 7, pp. 137-142, 1956.

33. K. Matusita, "Decision Rule Based on the Distance for the Classification Problem," Ann. Inst. Stat. Math. (Tokyo), Vol. 8, pp. 67-70, 1956.

34. K. Matusita, "Distance and Decision Rules," Ann. Inst. Stat. Math. (Tokyo), Vol. 5, pp. 59-65, 1954.

35. K. Matusita, "Classification Based on Distance in Multivariate Gaussian Case," Proc. 5th Berkeley Symposium on Math. Stat. and Prob., Vol. 1, pp. 299-304, 1967.

36. J. Wolfowitz, "Consistent Estimations of the Parameters in a Linear Structural Relationship," Skand. Aktuarietids, pp. 132-151, 1952.

37. J. Wolfowitz, "Estimation by the Minimum Distance Method," Ann. Inst. Stat. Math. (Tokyo), Vol. 5, pp. 9-23, 1953.

38. J. Wolfowitz, "Estimation by the Minimum Distance Method in Nonparametric Difference Equations," Ann. Math. Stat., Vol. 25, pp. 203-217, 1954.

39. J. Wolfowitz, "The Minimum Distance Method," Ann. Math. Stat. Vol. 28, pp. 75-88, 1957.

40. S. Das-Gupta, "Nonparametric Classification Rules," Sankhya, Indiana Jour. of Stat., Series

Indian Jour. of Stat., Series A, Vol. 26, pp.4-30, 1964.

41. T. Cacoullos, "Comparing Mahalanobis Distance I: Comparing Distances between Populations and Another Unknown," Sankhya, Indian Jour. Stat., Series A, Vol. 27, pp. 1-22, March, 1965.

42. T. Cacoullos, "Comparing Mahalanobis DistancesII: Bayes Procedures When the Mean Vector are Unknown," Sankhya, Indian Jour. Stat., Series A, Vol. 27, pp. 23-32, March, 1965.

43. M.S. Srivastava, "Comparing Distances Between Multivariate Populations - The Problem of Minimum Distance," Ann. Math. Stat., Vol. 38, pp. 550-556, April, 1967.

44. W. Hoeffding and J. Wolfowitz, "Distinguisability of Sets of Distributions," Ann. Math. Stat., Vol. 29, pp. 700-718, September, 1958.

45. K.S. Fu, D.A. Landgrebe, and T.L. Phillips, "Information Processing of Remotely Sensed Agricultural Data," Proc. IEEE, Vol. 57, pp. 639-654, April, 1969.

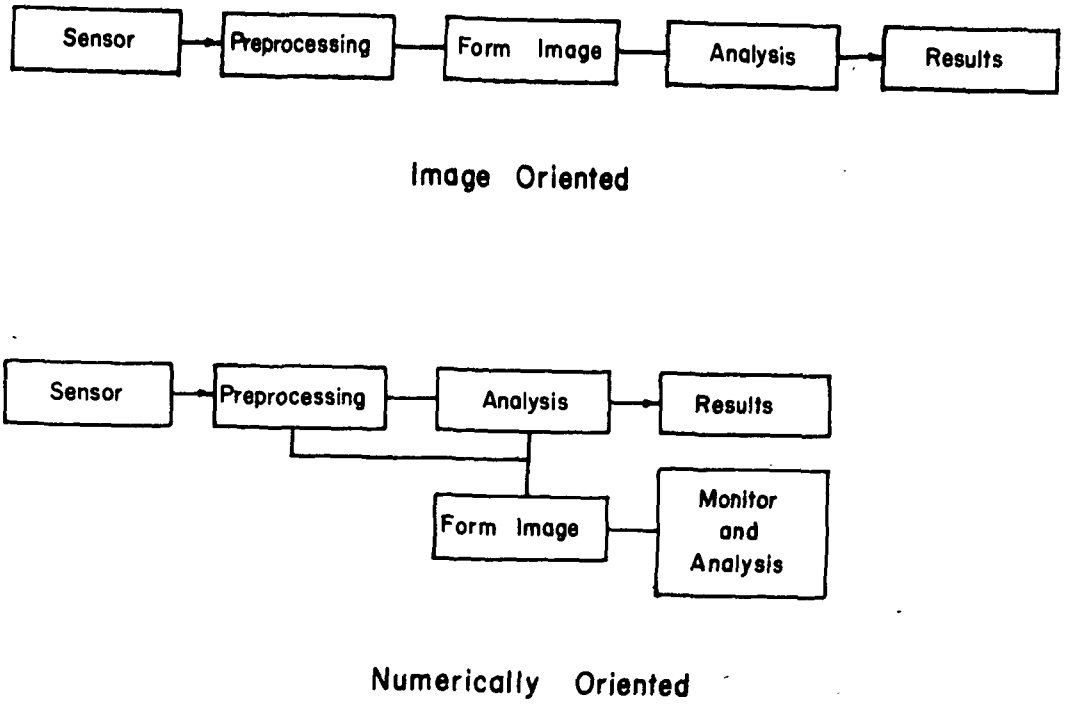

Fig. 1 Organization of Image and Numerically Oriented Systems

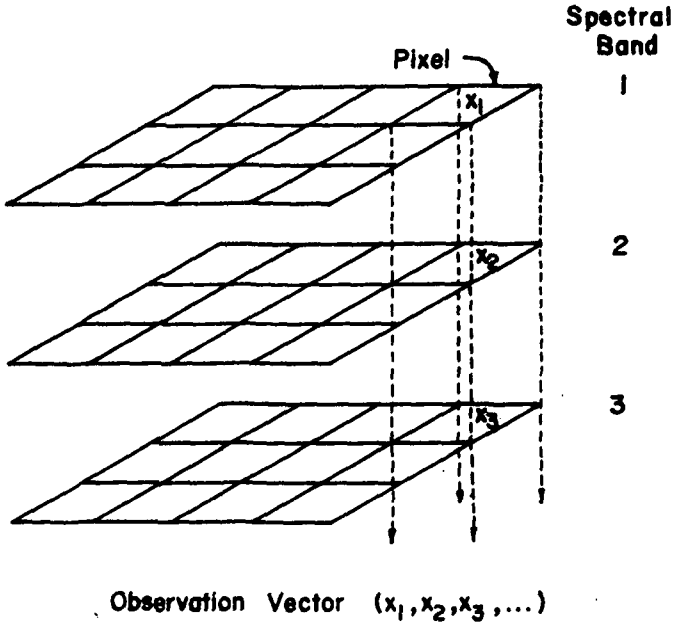

Fig. 2 Formation of Multispectral Data-Image

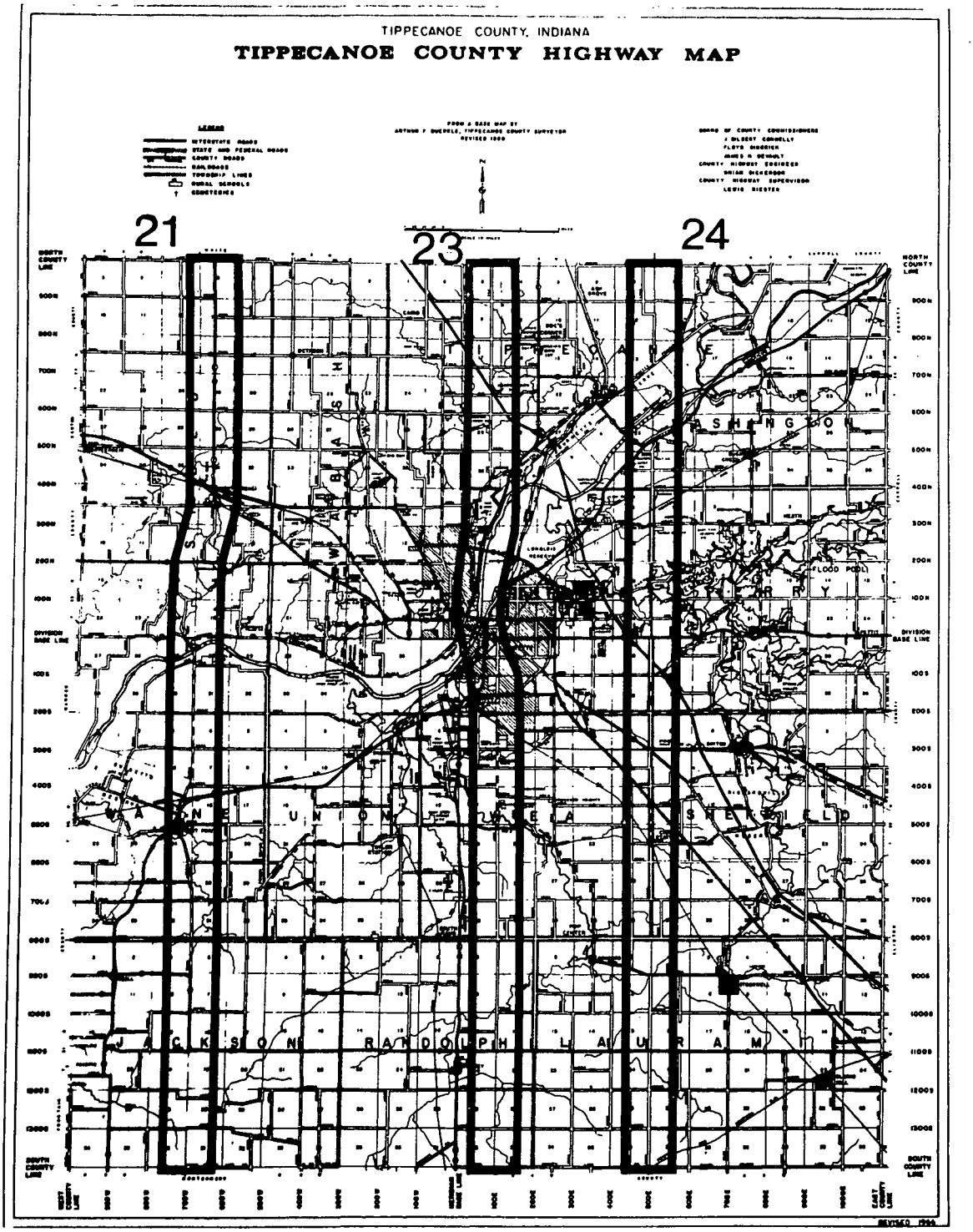

Fig. 3 Location of Tippecanoe County Flightlines 21, 23 and 24

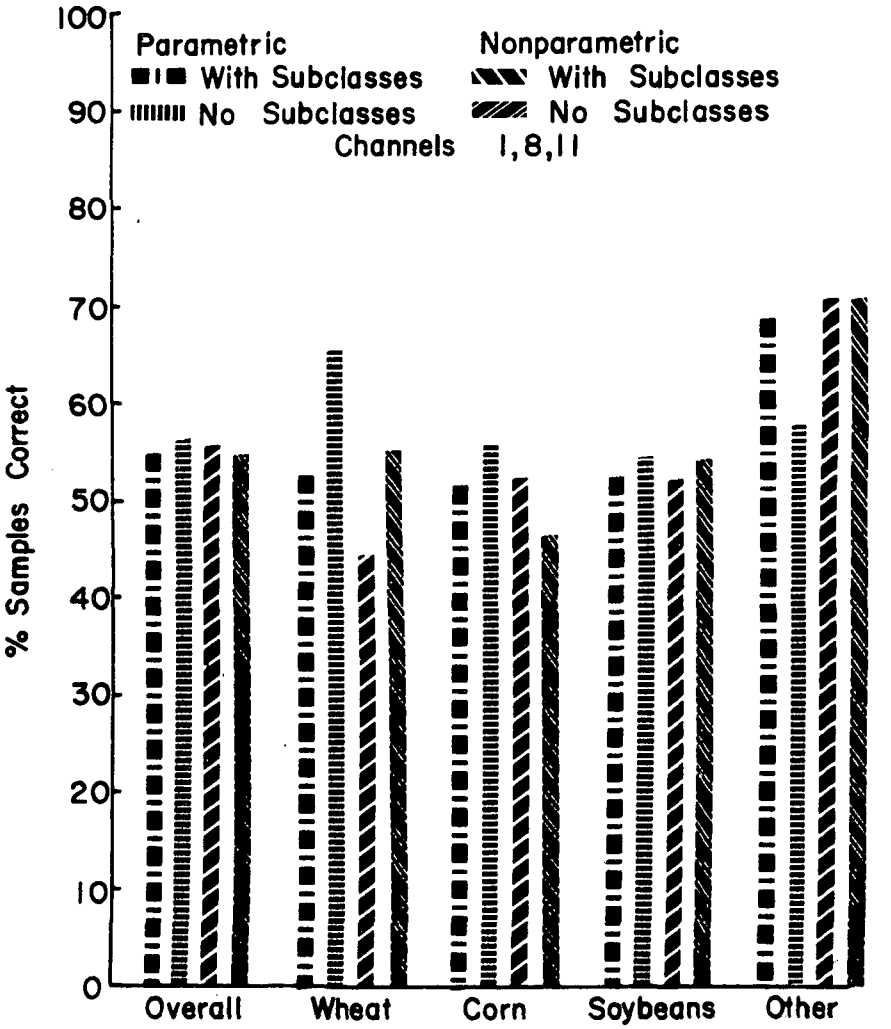

Fig. 4 Comparison of Average Test Performance for Parametric and Nonparametric Minimum Distance Classification Using Bhattacharyya Distance and Random Training

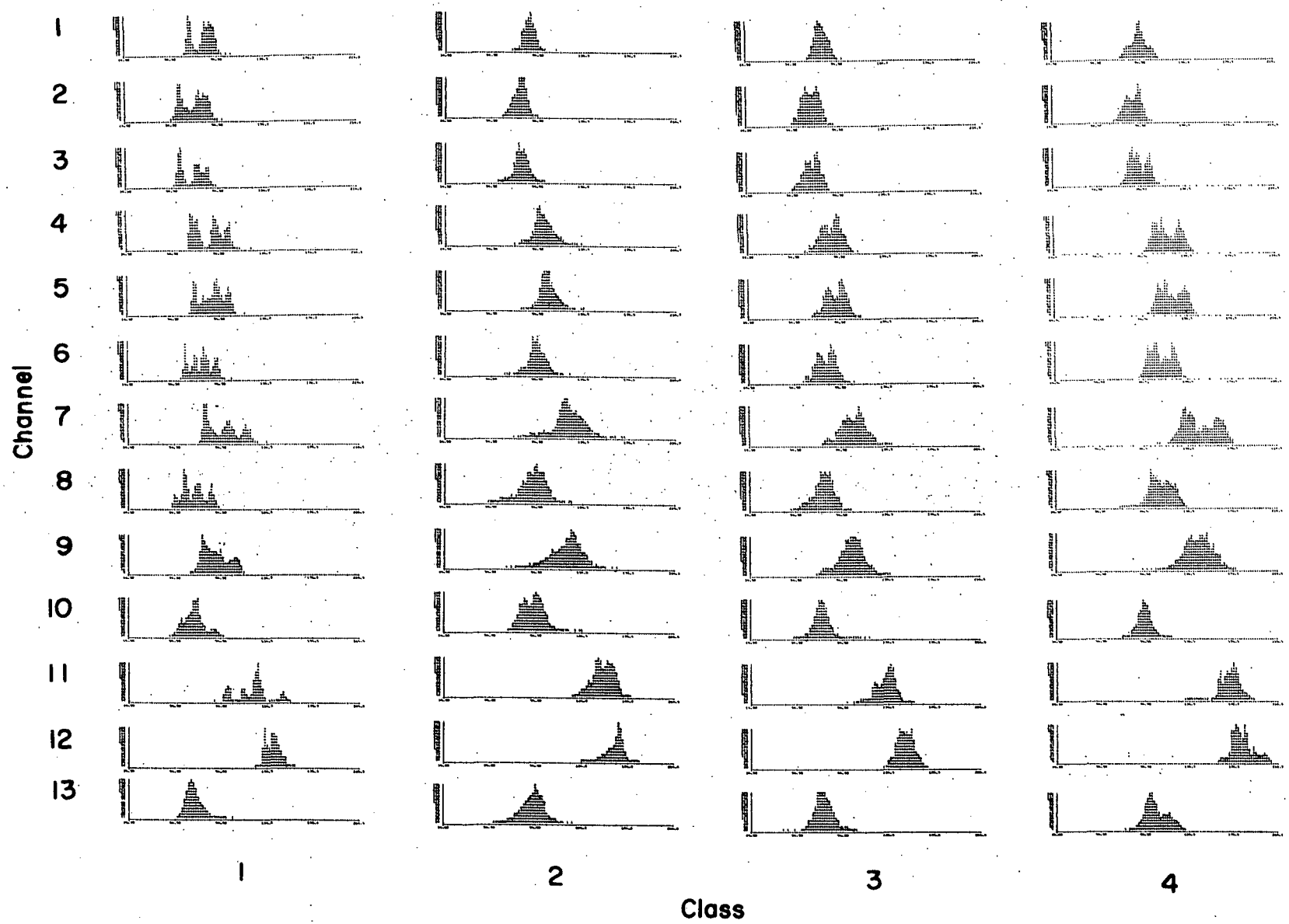

Fig. 5 Histograms for Wheat Subclasses Obtained as Result of Clustering Wheat Acres

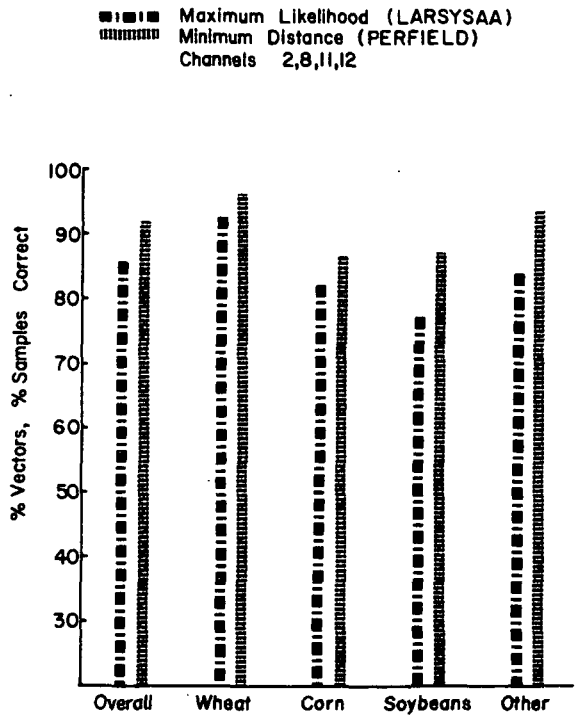

Fig. 6 Comparison of the Training Performance for Minimum Distance and Maximum Likelihood Classification

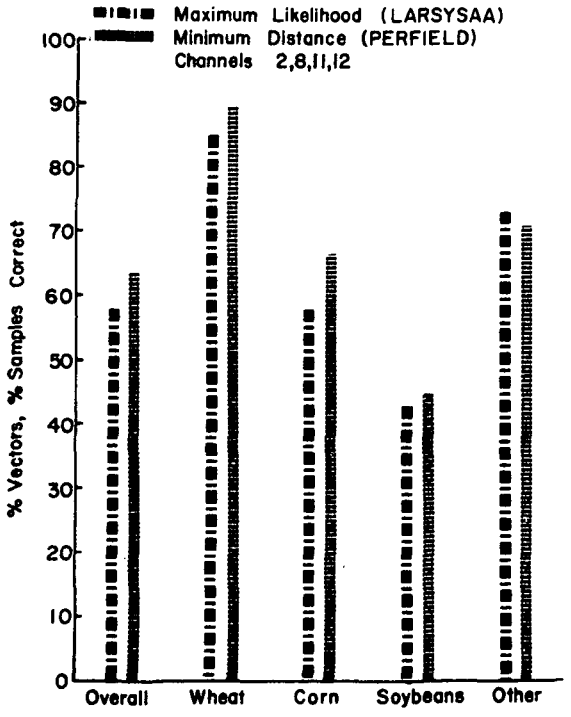

Fig. 7 Comparison of Average Test Performance of Minimum Distance and Maximum Likelihood Classification