Classification of Remotely Sensed Image Data Using Multitype Information

H. M. Kalayeh
D. A. Landgrebe

School of Electrical Engineering
Technical Report TR-EE 82-29

Laboratory for Applications of Remote Sensing
Technical Report 082782

Purdue University
West Lafayette, Indiana 47907

August 1982

CLASSIFICATION OF REMOTELY SENSED IMAGE DATA USING MULTITYPE INFORMATION

H.M. Kalayeh

D.A. Landgrebe

Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, IN 47906-1399, U.S.A.

August 1982

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Classification of multispectral image data based on spectral information has been a common practice in the analysis of remote sensing data. However, the results produced by current classification algorithms necessarily contain residual inaccuracies and class ambiguity. By the use of other available sources of information, such as spatial, temporal, and ancillary information, it is possible to reduce this class ambiguity and in the process improve the accuracy. Therefore, the purpose of this research is to improve the accuracy of the classification by utilizing such multitype information.

To accomplish this objective, three approaches are proposed. The first approach is a stochastic model in the time domain which utilizes spectral and temporal characteristics. The second approach involves the probabilistic and supervised relaxation methods which utilize multitype information. The third approach is a stochastic model in the spatial domain which attempts to extract interpixel

---

This research was supported in part by NASA Grant Nos. NAS9-15466 and NSG-5414.

class-conditional correlation and use this information with spectral characteristics to classify an object.

As a result of adapting the above approaches to the problem, the following five new classifiers are developed.

1. Markov pixel classifier
2. Non-iterative probabilistic relaxation
3. Modified minimum distance object classifier
4. Modified maximum likelihood object classifier
5. Linear minimum distance object classifier

For all the above algorithms, software systems are developed or the existing software programs at the Laboratory for Applications of Remote Sensing (LARS), Purdue University are modified. All these methods are experimentally evaluated.

CHAPTER 1

INTRODUCTION

Looking at the past and seeing what has been accomplished utilizing spectral characteristics and looking at the future and seeing the importance of utilizing temporal and spatial characteristics made us ask, "What else can be accomplished in classification of remotely sensed image data?" Therefore, the main objectives are:

1. To advance the state of the art of pattern recognition by developing algorithms which can utilize combinations of spectral, spatial and temporal information.
2. To improve the accuracy of the classification over the current maximum likelihood pixel classifier.

1.1 Pattern Recognition in Remote Sensing

The field of pattern recognition is concerned with designing machines to recognize patterns. The design procedure typically has two phases:

1. Training or learning phase
2. Decision phase

In the learning phase, it is desired that the machine learn the main characteristics of patterns, then in the decision phase it identifies the class of an unknown pattern. Pattern recognition methods have had large varieties of applications, for example, in information theory, control, image processing and remote sensing. A pattern recognition system in remote sensing consists of four parts; viz., the scene, the sensor system, feature extractor and the classifier (Figure 1.1).

1.2 Remote Sensing of Earth's Resources

In remote sensing, the spectral variations of the electromagnetic energy of the scene have been studied extensively. The spectral response which is a function of wavelength has been modeled as a random process [81,82,84,86]. In practice, the reflected and emitted electromagnetic energy of each pixel in the scene in several important wavelength bands as shown in Fig. 1.2 are measured by an aircraft or spacecraft equipped with a multispectral remote sensor system. The output of the sensor system, as a set of continuous electric voltages, is digitized, calibrated and transmitted to the earth's stations. Then by pattern recognition techniques the data is classified and the useful information is extracted.

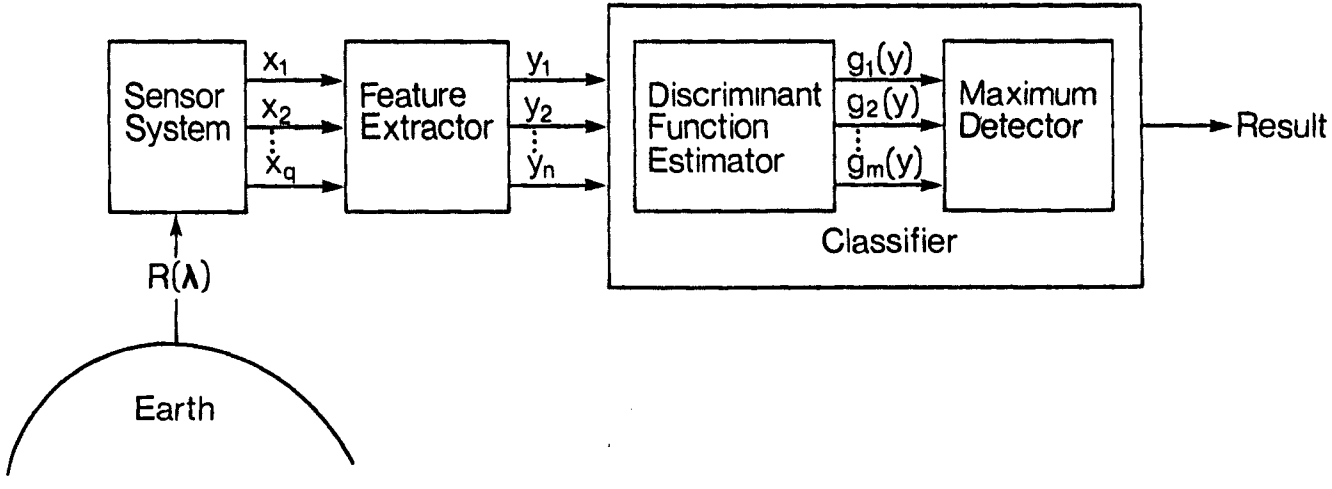

Figure 1.1 A block diagram of pattern recognition system in remote sensing.

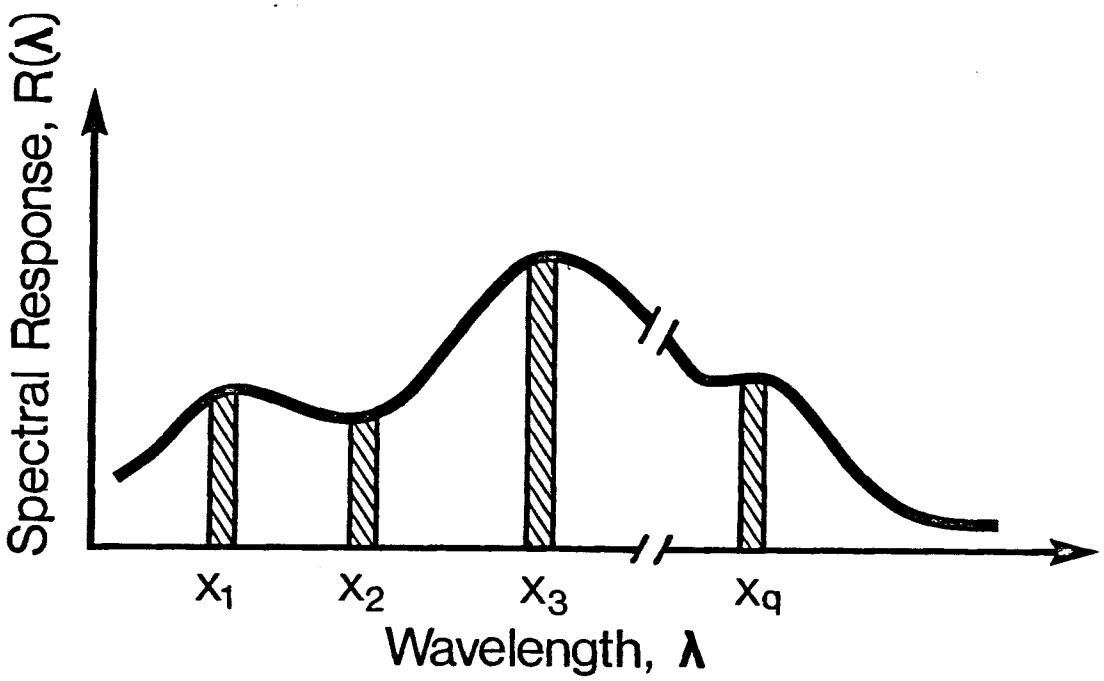

Figure 1.2 The receptor's output provides a q-dimensional vector representation of the spectral response function.

1.3 Review of Literature

1.3.1 Utilizing Spectral Characteristics

There are both parametric and non-parametric approaches which have been studied to characterize spectral information. Let $\mathbb{X}$ be q-dimensional measurement space and $\mathbb{Y}$ be n-dimensional feature space which is the range of measurement space under a linear transformation. The problem of classification is estimating or learning decision boundaries to partition the feature space into non-overlapping regions. In other words, it can be said a pattern classifier is a mapping from n-dimensional feature space into one-dimensional decision space. Let Y be a point in feature space belonging to the ith class and $g_i(Y)$ be a point in decision space. Generally, all algorithms utilizing spectral information can be divided into two main categories: linear and non-linear depending on the functional form of $g_i(Y)$. For more detail, see Swain et al. [1], Fukunaga [2], Duda and Hart [3], Nilson [4], and Mendel and Fu [5]. An extensive bibliography on learning procedures in pattern classifiers is given in [6].

1.3.2 Utilizing Spatial Characteristics

By the term spatial information is meant contextual and textural information. The information provided by the relationship of an object or a pixel to those surrounding

it is referred to as contextual information. The spatial distribution of the reflected and emitted energy of an object is referred to as textural information. In many pattern recognition problems, there exists a spatial context which describes the spatial dependencies among the patterns to be recognized. There are numerous references to the use of context in pattern recognition. See, for example [7-12]. In all these references attempts have been made to utilize the contextual characteristics by discrete one or two dimensional Markov process. An extensive bibliography on the use of context is given in [13]. A different approach which attempts to utilize spectral and spatial context based on compound decision theory is given in [14-16].

Another approach to incorporate the spatial context with spectral information is through the use of probabilistic relaxation methods. Within the last five years, serious efforts have been made to utilize spatial interaction among pixel labels in a local neighborhood by heuristic techniques. The probabilistic relaxation processes have been extensively used in picture processing [17,18] especially for line and curve enhancement [19,20 and 21]. The covergence properties of relaxation have been investigated in [18,24,34]. Because of the heuristic nature of relaxation approaches, several algorithms have been developed [17-35].

Another source of useful information which characterizes an image is the local texture. The textural information can be extracted in one approach based on a "gray tone spatial dependence matrix" [36-38]. In another approach, attempts have been made to model the texture by one or two dimensional autoregression (AR) models [39-44]. In the unilateral AR model [39,40,43,44] the assumption is that the current observation depends only on the past ones and in the bilateral AR model [41,42] the current observation depends on the neighbors on either side. Also, AR processes have been used for modeling of a noisy images and then Kalman filtering approaches are used to reduce the noise [45-52]. An extensive bibliography on statistical and structural approaches to texture is given in [53]. However, most of the references mentioned in this section discuss methods capable of characterizing images in which the measurement on each pixel is a one-dimensional observation. But in a remotely sensed image data a multidimensional observation is available for each pixel. Therefore, care must be taken for modeling. A number of papers on combined use of spectral and textural characteristics for the improvement of multispectral classification of remotely sensed data are given in [38,57-63].

A different approach to utilize spatial characteristics is to partition the scene into statistically

homogeneous objects [54]. Then based on the assumptions that pixels with an object are uncorrelated and normally distributed, the maximum likelihood object classifier is developed [54,55,56]. This scheme provides consistently better performance than maximum likelihood pixel classifier. However, it does not incorporate the context of the objects into the classification process.

1.3.3 Utilizing Temporal Characteristics and Available Ancillary Data

Temporal variations in the scene and available ancillary data such as topographic data, pixel radar response, and soil type maps, are known to be information-bearing. However, because of the complexity which they add to the analysis of spectral and spatial characteristics, they are not being effectively utilized. Thus far, there are three approaches to utilize spectral/temporal and spectral/ancillary data. The earliest approach is simply to increase the dimensionality of feature space by concatenating the available multitemporal measurement vectors, or spectral and ancillary measurement vectors and is called the "stack vector" approach [65,66]. But increasing the dimensionality increases the magnitude of the computation and number of spectral subclasses which must be defined. This scheme requires larger numbers of training samples to characterize the data.

The second approach for joint use of spectral/temporal information has been studied by Swain [67]. The idea has been developed based on a Bayesian strategy (minimum risk) for multitemporal data.

An approach for combined use of spectral/ancillary data has been suggested in [64]. The idea of supervised probabilistic relaxation is used to employ the available ancillary information to improve the accuracy of a predetermined spectral classifier.

The third approach utilizing multitemporal characteristics is to find a mathematical model for spectral development, see for example [68-71]. A more detailed discussion on algorithms utilizing spectral, spatial, temporal and ancillary information is given by Landgrebe [72].

1.4 Summary of Contents

In Chapter 2, multitemporal data is modeled as the output of a stochastic dynamic system, then by the Markov process assumption attempts are made to utilize the temporal characteristics.

In Chapter 3, utilizing spectral, temporal, spatial and ancillary information by probabilistic relaxation techniques is investigated.

In Chapter 4, a two-dimensional autoregressive model is used to extract the texture of an object. Then based on this information and spectral characteristics, the object is classified.

In Chapter 5, a summary and the major contributions of this research are stated. Directions for further study is suggested. Finally, some analytical details, developed computer programs and information about the data set are placed in appendices.

CHAPTER 2

UTILIZING MULTITEMPORAL INFORMATION BY A STOCHASTIC MODEL

2.1 Existing Algorithms Utilizing Multitemporal Information

The spectral variation of energy is a function of time, and at a given time it has been modeled as a random process [81,82,84,86]. However, in practice, at a given time only the spectral variation of energy of selected bands is measured. Therefore, the measurement of energy for each pixel by a remote sensing system can be viewed as performing a statistical experiment whose outcome is vector-valued. The variation of energy as a continuous function of time for q different bands or channels is shown in Figure 2.1. These curves are the spectral/temporal representation of a ground cover type. In practice, only discrete time samples from these continuous functions are measured.

Multispectral image data consists of an observation set $\mathbb{X}$, location set Ω and population set C where

$$\mathbb{X} = X(t),\ s,\ X(t) \in R^q$$

$$\Omega = \{s = (i,j),\ 1 \leq i \leq I,\ 1 \leq j \leq J\}$$

$$C = \{\omega_1, \omega_2, \ldots \omega_m\}$$

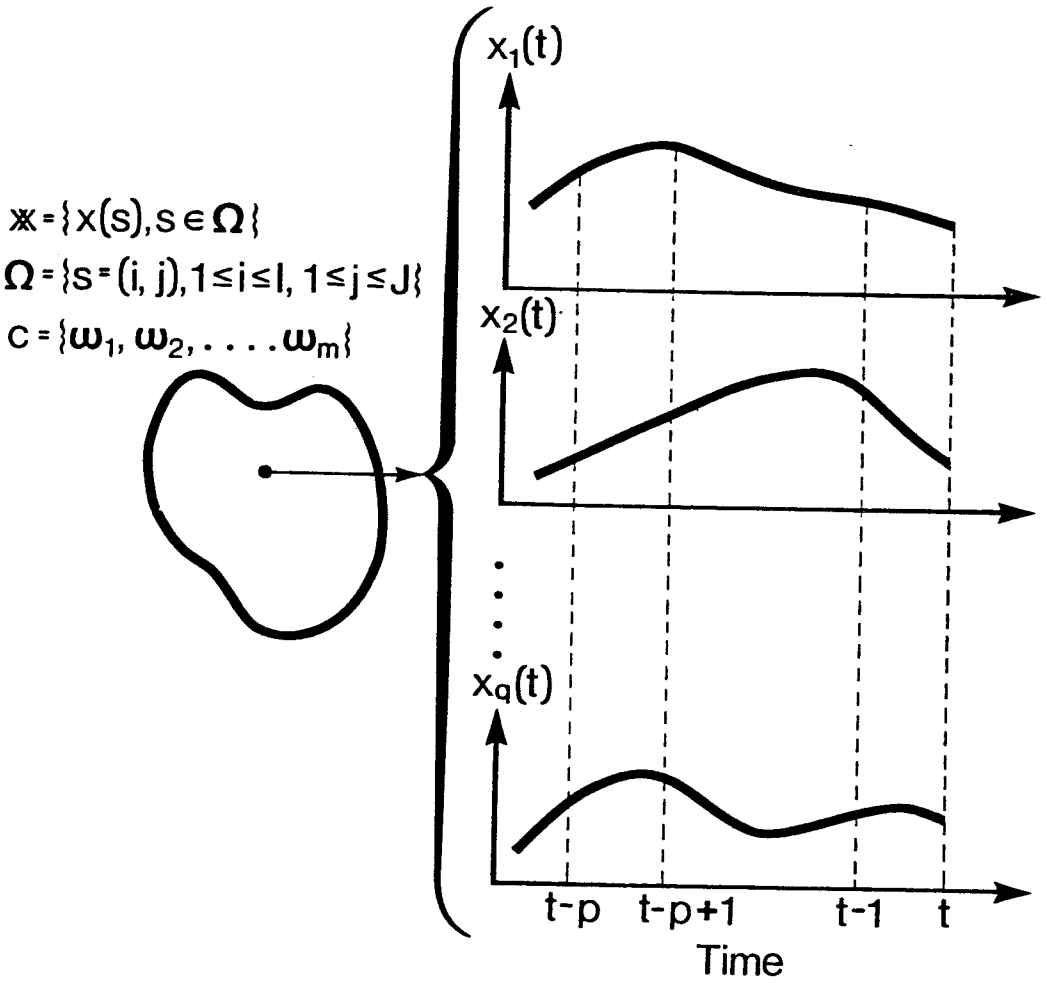

Figure 2.1 Temporal variation of energy as a sample function from a q-dimensional stochastic process.

and m is the number of classes. By sampling the vector continuous signal (P+1) times, we will have (P+1) measurement vectors $X(t)$, $X(t-1),\ldots X(t-P)$ available for each pixel.

It is commonly assumed that $p(X(t)|\omega_i)$ the class conditional density function is multi-variate normal [1]; i.e.,

$$p(X(t)|\omega_i) = N(X(t);\ M_k(t), \Sigma_i(t)) \triangleq$$

$$\frac{1}{(2\pi)^{\frac{q}{2}}|\Sigma_i(t)|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(X(t)-M_i(t))^T \Sigma_i^{-1}(t)(X(t)-M_i(t))\} \tag{2.1}$$

where $M_i(t)$ and $\Sigma_i(t)$ are the mean vector and covariance matrices, respectively. Then assuming a maximum likelihood pixel classification strategy, the decision rule is

$$p(X(t)|\omega_i) = \max_j p(X(t)|\omega_j), \quad j = 1,2,\ldots m, \tag{2.2}$$

$$\rightarrow X(t) \varepsilon \omega_i;\ \text{i.e., } X(t) \text{ is classified into class } \omega_i.$$

The first algorithm which attempts to utilize temporal characteristics is the so-called stack vector approach [65,66]. As mentioned earlier, let $X(t)$, $X(t-1),\ldots,X(t-P)$ be P+1 sampled value vectors of the continuous temporal variation of reflected and emitted electromagnetic energy. The stack vector algorithm is the following:

$$p(X(t),X(t-1),\ldots,X(t-P)|\omega_i) = N(X(t),X(t-1),\ldots,X(t-P);\ M_i,\Sigma_i) \tag{2.3}$$

and decision rule is

$$p(X(t),X(t-1),\ldots,X(t-P)\,|\,\omega_i) =$$

$$\max_j\ p(X(t),X(t-1),\ldots,X(t-P)\,|\,\omega_j),$$

$$j=1,2,\ldots,m \rightarrow [X(t),X(t-1),\ldots,X(t-P)]\varepsilon\omega_i \qquad (2.4)$$

This algorithm sometimes provides increases in accuracy over simple spectral means; however, disadvantages of this method are that it

1. Expands the number of spectral subclasses.
2. Requires large numbers of training samples.
3. Increases the computational complexity.

Another algorithm which attempts to utilize multitemporal information is the cascade pixel classifier [67]. Based on the Bayesian strategy and using some appropriate assumptions, the decision rule for bitemporal information is shown to be:

If $g_i(X(t_2),X(t_1)) =$

$$\max_j\ p(X(t_2)\,|\,\omega_j)\left(\sum_{\ell=1}^{m_1} p(X(t_1)\,|\,V_\ell)P(\omega_j,V_\ell)\right),$$

$$j=1,2,\ldots,m$$

then decide

$$[X(t_2),X(t_1)]\,\varepsilon\omega_i \qquad (2.5)$$

where

$$P(X(t_2)\,|\,\omega_j) = N(X(t_2);\ M_j(t_2),\Sigma_j(t_2))$$

$$P(X(t_1)|V_\ell) = N(X(t_1);\ M_\ell(t_1), \Sigma_\ell(t_1))$$

$X(t_2)$ and $X(t_1)$ are multivariate observations at time t_2 and t_1, respectively. $V_1, V_2, \ldots, V_{m_1}$ and $\omega_1, \omega_2, \ldots, \omega_{m_2}$ denote the set of classes at time t_1 and t_2, respectively. And $P(\omega_j, V_\ell)$ is the joint prior probability of class V_ℓ at time t_1 and class ω_j at time t_2. Problems with this scheme are:

1. It is very sensitive to the joint prior probability; therefore, a good estimate of $P(\omega_j, V_\ell)$ should be available.
2. It is sensitive to missing observation times.

However, the cascade classifier sometimes improves the classification accuracy significantly.

As a third approach there are several algorithms which attempt to utilize temporal characteristics by a regression model. In [71] a procedure statistically modeling only noise and not the signal has been investigated for classifying observations based upon their growth profiles. The model which was proposed in [71] with some modifications is:

$$X_i = AB_i + U_i \qquad (2.6)$$

where

$$X_i = [X_i(1), X_i(2), \ldots, X_i(P)]^T,$$

$X_i(k) = X_i(t_k)$ is an observation at time t_k of one of the available channels from ith class,

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{N-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{N-1} \\ & & & & \\ & & & & \\ 1 & t_P & t_P^2 & \cdots & t_P^{N-1} \end{bmatrix}$$

and

$$B_i = [b_{i0}\ b_{i1}\ \cdots\ b_{iN-1}]^T$$

Assuming U_i is white noise with zero mean and covariance matrix V_i, it can be shown that

$$p(X\ |\omega_i) = N(X_i;\ AB_i, V_i)$$

and the maximum likelihood estimates for B_i and V_i are given by

$$\hat{B}_i = (A^TA)^{-1}A^T[\frac{1}{n_i}\sum_{i=1}^{n_i} X_i] \tag{2.7}$$

$$\hat{V}_i = \frac{1}{n_i}\sum_{i=1}^{n_i}(X_i - A\hat{B}_i)(X_i - A\hat{B}_i)^T \tag{2.8}$$

where n_i is the number of training samples from class i and the decision rule is:

If $P(X|\omega_i) = \max_j P(X|\omega_j)$, $j = 1, 2, \ldots m$,

then classify X into class i. Spectral development of some cover types may be accurately modeled by this algorithm; however, problems with this method are the following:

1. The temporal observation is modeled by a N-1 degree polynomial by which the parameter N must be estimated.
2. The classifier only uses one spectral feature.
3. Increasing P means increasing computational complexity.
4. Small P means lower classification accuracy.

2.2 Proposed Algorithm for Utilizing Temporal Characteristics

Ground cover types are considered as stochastic systems with non-stationary Gaussian processes as input and temporal variations of reflected and emitted electromagnetic energy as output. Then by assumption that the behavior of these stochastic systems is governed by first order Markov processes, multitemporal information may be utilized.

It is logical to assume that the temporal change of the energy of a pixel in all channels as shown in Figure 2.1 could be represented by a continuous time function. A hypothetical distribution of the temporal variations of only one channel for a cover type is shown in Figure 2.2. As mentioned earlier, let X(t) be a q-dimensional random variable (q is the number of spectral features).

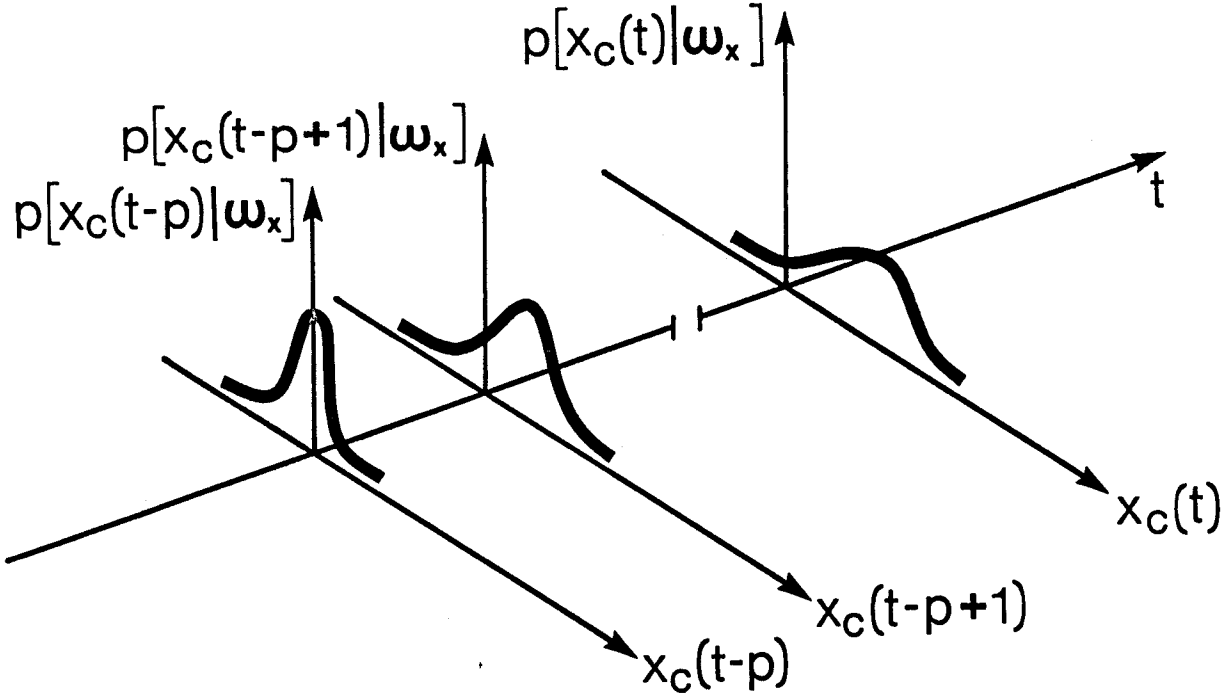

Figure 2.2 A hypothetical distribution of spectral development for a cover type.

Assumptions:

1) $X(t-k)$, $k = 1, 2, \ldots, P$ is a Gauss-Markov random sequence; i.e.

(a) $$p(X(t)|\omega_i) = N(X(t); M_i(t), \Sigma_i(t)) \tag{2.10}$$

(b) $$p(X(t)|X(t-1);\omega_i) =$$

$$N(X(t), M_i(t)+\rho_i(t-1)(X(t-1)-M(t-1)),$$

$$\Sigma_i(t)-\rho_i(t-1)\Sigma_i(t-1)\rho_i{}^T(t-1)) \tag{2.11}$$

where $M_i(t)$ and $\Sigma_i(t)$ are the mean vector and covariance matrix of the ith class at time t, respectively, and $\rho_i(t-1)$ is the temporal correlation matrix of the ith class between time t and $(t-1)$.

2) $X(t)$, $X(t-1), \ldots$ are Markovian sequences, i.e.,

$$p(X(t)|X(t-1), \ldots, X(t-P);\omega_i) =$$

$$p(X(t)|X(t-1);\omega_i) \tag{2.12}$$

It is believed that many natural and man-made dynamic phenomena may be approximated quite accurately by a Gauss-Markov random sequence [83] and a Gauss-Markov random sequence can always be represented by the state vector of a multivariate linear dynamic system forced by a purely random Gaussian sequence in which the initial state vector is Gaussian; i.e.,

$$X_i(t)-M_i(t) = \rho_i(t-1)(X_i(t-1)-M_i(t-1)) + W_i(t) \tag{2.13}$$

where $\rho_i(t-1)$ is the temporal correlation matrix between multivariate observations at time t and t-1 of class i. Let

$$Y_i(t) = X_i(t) - M_i(t)$$

Then from (2.13) we have:

$$Y_i(t) = \rho_i(t-1)Y_i(t-1) + W_i(t) \tag{2.14}$$

where

$$Y(t) \text{ and } W(t) \in R^q$$

$$E[Y(t) \mid \omega_i] = 0 \tag{2.15}$$

$$\text{cov}[Y(t) \mid \omega_i] = \Sigma_i(t) \tag{2.16}$$

$$p(W(t) \mid \omega_i) = N(W(t);\ 0,\ V_i(t)) \tag{2.17}$$

$$E[W(t) \mid \omega_i] = 0 \tag{2.18}$$

$$\text{cov}[W(t) \mid \omega_i] = V_i(t) \tag{2.19}$$

Also, W(t) (error) is orthogonal to Y(t); i.e.,

$$E[Y(t)W^T(t) \mid \omega_i] = 0 \tag{2.20}$$

where E and cov denote the expectation and the covariance matrix, respectively.

2.3 Parameter Estimation for Multitemporal Observation

Suppose one has n_i labeled observations from each class ω_i for i = 1, 2,...m and that each of these observations has been observed at P + 1 distinct times. By the Gauss-Markov and Markovian assumptions of observations we can write

$$p(Y(t), Y(t-1), \ldots, Y(t-P)|\omega_i) =$$

$$\left[\prod_{j=1}^{P} p(Y(t-j+1)|Y(t-j);\omega_i)\right] p(Y(t-P)|\omega_i) \tag{2.21}$$

Assuming the training samples are independent, then we have:

$$p(Y_1(t),Y_1(t-1), \ldots, Y_1(t-P), Y_2(t-1), \ldots, Y_2(t-P),$$

$$\ldots, Y_{n_i}(t),Y_{n_i}(t-1), \ldots, Y_{n_i}(t-P)|\omega_i) =$$

$$\prod_{k=1}^{n_i}\left\{\left[\prod_{j=1}^{P} p(Y_i(t-j+1)|Y_i(t-j;\omega_i)\right] p(Y_i(t-P)|\omega_i)\right\} \tag{2.22}$$

The maximum likelihood estimates $\Sigma_i(t-P)$, $\rho_i(t-j)$, $V_i(t-j+1)$, $j = 1, 2, \ldots P$, $i = 1, 2, \ldots m$ are given by

$$\hat{\Sigma}_i(t-P) = \frac{1}{n_i}\sum_{k=1}^{n_i} Y_k(t-P)Y_k^T(t-P) \tag{2.23}$$

$$\hat{\rho}_i(t-j) = \left[\sum_{k=1}^{n_i} Y_k(t-j+1)Y_k^T(t-j)\right]\left[\sum_{k=1}^{n_i} Y_k(t-j)Y_k^T(t-j)\right]^{-1} \tag{2.24}$$

$$\hat{V}_i(t-j+1) = \frac{1}{n_i}\left[\sum_{k=1}^{n_i} (Y_k(t-j+1) - \hat{\rho}_i(t-j)Y_k(t-j))(Y_k(t-j+1) - \hat{\rho}_i(t-j)Y_k(t-j))^T\right] \tag{2.25}$$

For more detail on estimates of the parameters see Appendix A.

2.4 Decision Rule

To classify an unknown profile Y(t), Y(t-1) Y(t-P), the classification rule is:

$$\text{If} \quad \hat{p}(Y(t),Y(t-1),\ldots,Y(t-P)|\omega_i) = \max_k \hat{p}(Y(t),Y(t-1),\ldots,Y(t-P)|\omega_k),$$

$$k = 1,2,\ldots,m$$

then assign

$$[Y(t),\ Y(t-1),\ldots,\ Y(t-P)] \quad \text{to class } i \tag{2.26}$$

where

$$\hat{p}(Y(t),Y(t-1),\ldots,Y(t-P)|\omega_i) =$$

$$\left[\prod_{j=1}^{P} N(Y(t-j+1);\ \hat{\rho}_i(t-j)Y(t-j),\ \hat{V}_i(t-j))\right]\left[N(Y(t-P);\ 0,\ \hat{\Sigma}_i(t-P))\right]$$

This classifier will be called a Markov classifier and its block diagram is given in Figure 2.3.

In deriving the decision rule, it has been assumed that the multichannel temporal change curve is a first order vector Markov process and also the observations are a Gauss-Markov vector random sequence. However, the temporal change curve can be modeled by a second or higher order

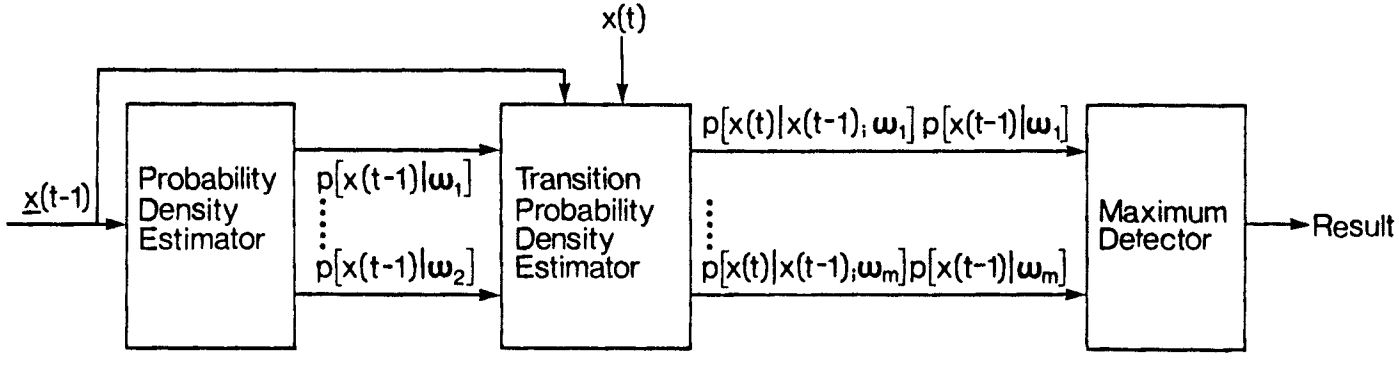

Figure 2.3 The Markov classifier model.

vector Markov process. For example, for a second order Markov process, the stochastic dynamic model is given by:

$$Y_i(t) = \rho_i(t-1)Y_i(t-1)+\rho_i(t-2)Y_i(t-2)+W_i(t) \qquad (2.27)$$

where $Y_i(t) = X_i(t) - M_i(t)$

and the discriminant functions are:

$$\hat{p}(Y(t),Y(t-1),\ldots,Y(t-P)|\omega_i) =$$

$$\left[\prod_{j=2}^{P} N(Y(t-j+2);\ \hat{\rho}_i(t-j+1)Y(t-j+1) + \hat{\rho}_i(t-j)Y(t-j),\ \hat{V}_i(t-j+2)\right] \cdot$$

$$\left[N(Z(t-P);\ 0,\ \hat{\Sigma}_i(t-P)\right] \qquad (2.28)$$

where

$$Z(t-P) = \begin{bmatrix} Y(t-P+1) \\ \\ Y(t-P) \end{bmatrix} \qquad (2.29)$$

$$\Sigma_i(t-P) = \mathrm{cov}[Z(t-P)|\omega_i] \qquad (2.30)$$

Obviously the second order stochastic dynamic Markov model is more complex than first order Markov model.

2.5 Validity of the Assumptions and Advantage of the Proposed Approach

In remote sensing it is commonly assumed that the class-conditional density function is approximately

multivariate normal. This assumption is usually justified, particularly if spectral classes are defined by a suitable mode-seeking method. Let X(t) and X(t-1) be a q-variate random vector observation at time t and t-1, respectively. Based on the above Gaussian assumption, we can write:

$$p(X(t)|\omega_i) = N(X(t);\ M_i(t),\ \Sigma_i(t)) \tag{2.31}$$

$$p(X(t-1)|\omega_i) = N(X(t-1);\ M_i(t-1),\ \Sigma_i(t-1)) \tag{2.32}$$

Let $$W(t) = X(t)-M_i(t)-\rho_i(t-1)(X(t-1)-M_i(t-1)) \tag{2.33}$$

where

$$M_i(t) = E[X(t)|\omega_i] \tag{2.34}$$

$$M_i(t-1) = E[X(t-1)|\omega_i] \tag{2.35}$$

$$\rho_i(t-1) = E[X(t)X_i^T(t-1)]\{E[X(t-1)X_i^T(t-1)]\}^{-1} \tag{2.36}$$

Since the linear combination of normal random vectors are also normal [81]; therefore,

$$p(W(t)|\omega_i) = N(W(t);\ 0,\ V_i(t)) \tag{2.37}$$

where $$V_i(t) = \mathrm{cov}[W(t)|\omega_i] \tag{2.38}$$

Now, suppose the objective is to utilize temporal observations at two stages which correspond to considerable differences in canopy structure. Therefore, let $p(X(t),X(t-1)|\omega_i)$, i=1,2,...m be the discriminant functions for the m classes. Then by the Bayes rule, we can write

$$p(X(t),X(t-1)|\omega_i) = p(X(t)|X(t-1);\omega_i)\ p(X(t-1)|\omega_i) \tag{2.39}$$

For the proposed model, we need only assume that

$$p(X(t)|X(t-1);\omega_i) = N(X(t);D_i(t),V_i(t)) \tag{2.40}$$

where

$$D_i(t) = E[X(t)|X(t-1);\omega_i] \tag{2.41}$$

$$V_i(t) = \text{cov}[X(t)|X(t-1);\omega_i] \tag{2.42}$$

It is shown in [81] that if X(t) and X(t-1) are jointly normal, i.e.,

$$p(X(t),X(t-1)|\omega_i) = N(X(t),X(t-1);M,\Sigma) \tag{2.43}$$

then

$$p(X(t-1)|\omega_i) = N(X(t-1), M(t-1), \Sigma(t-1)) \tag{2.44}$$

and

$$p(X(t)|X(t-1);\omega_i) = N(X(t);D_i(t),V_i(t)) \tag{2.45}$$

where

$$M = \begin{bmatrix} M(t-1) \\ M(t) \end{bmatrix} \tag{2.46}$$

$$\Sigma = \begin{bmatrix} \Sigma(t) & \Sigma(t,t-1) \\ \Sigma(t-1,t) & \Sigma(t-1) \end{bmatrix} \tag{2.47}$$

The converse is not necessarily true if they are not independent. As mentioned earlier, we know that usually X(t) and X(t-1) for a given class are marginally normal and also are not independent. Therefore, if we assume that X(t) and X(t-1) for each class are jointly normal, the normality assumption of the transition probability density function

is true. However, the validity of the joint normal assumption is left to be investigated.

The advantages of the proposed model are the following:

1) Utilization of the temporal correlation between patterns in different stages and incorporation of this information into the classification process is provided to improve the accuracy.
2) Faster computation is provided over the stack vector approach and the cascade classifier.

2.6 Experimental Results

2.6.1 Data Set

Multitemporal spatially registered Landsat multispectral scanner (MSS) data acquired over Henry County, Indiana, in 1978 were selected to evaluate the performance of the Markov pixel classifier. The acquisition dates for this data set are: June 9, July 16, August 20, and September 26. The number of channels available for the Landsat MSS is four. Channels one and two are in the visible range and channels three and four are in the reflective infrared region of the electromagnetic spectrum. The informational classes are corn, soybean and other.

2.6.2 Training Methods

If field boundaries are chosen with care, then typically data from an individual field, regardless of crop type, is usually reasonably unimodel and symmetrical. However, occasionally individual fields do exhibit bimodality, and combined data from different fields of the same crop type frequently exhibits bimodality. Therefore, in order to approximately satisfy the Gaussian assumption, the following two training methods are considered.

Histogramming Method. A large number of fields are histogrammed for each main class and based on these histograms the subclasses which approximately satisfy the normal assumption are defined.

Clustering Method. All training fields for each main class are clustered into various numbers of modes and subclasses are defined on the basis of mode separability.

Experiment 2.1

In order to make a comparison between the cascade pixel classifier and the Markov pixel classifier, bitemporal data were analyzed. In this experiment, we let t = July 16 and t-1= June 9 (27 days apart). Then bitemporal registered data of these two dates were used to evaluate the relative performance of these classifiers. Spectral classes were defined by the histogramming method and the parameters estimated. Finally, test fields were used to

make a comparison among the performance of the maximum likelihood pixel classifier, the cascade pixel classifier and the Markov pixel classifier. In all the experiments that had been performed, the transition probabilities as suggested in [67] were: $P(\omega_j,V_\ell) = 0.8$, $\omega_j = V_\ell$, $P(\omega_j,V_\ell) = 2/(m2-1)$, $\omega_j \neq V_\ell$. The results of this experiment are shown in Figure 2.4 and more details of the results are given in Table 2.1. Information about the software system and the training data are given in Appendix G. The results show that the Markov pixel classifier has substantially better performance than the cascade and either of the uni-temporal maximum likelihood pixel classifiers. The Markov classifier improved the overall performance by about 10 percent.

Experiment 2.2

The same data set used in Experiment 1 is used here, with t = August 20 and t-1 = July 16 (34 days apart). Exactly the same procedures as Experiment 1 for training the classifiers were performed and the results are given in Figure 2.5 and Table 2.2. The results show that the maximum likelihood pixel classifier at time t (August 20) has higher overall performance. However, it is worthwhile to note that the classification accuracy for each class by the Markov pixel classifier is uniform and above 75%. This is

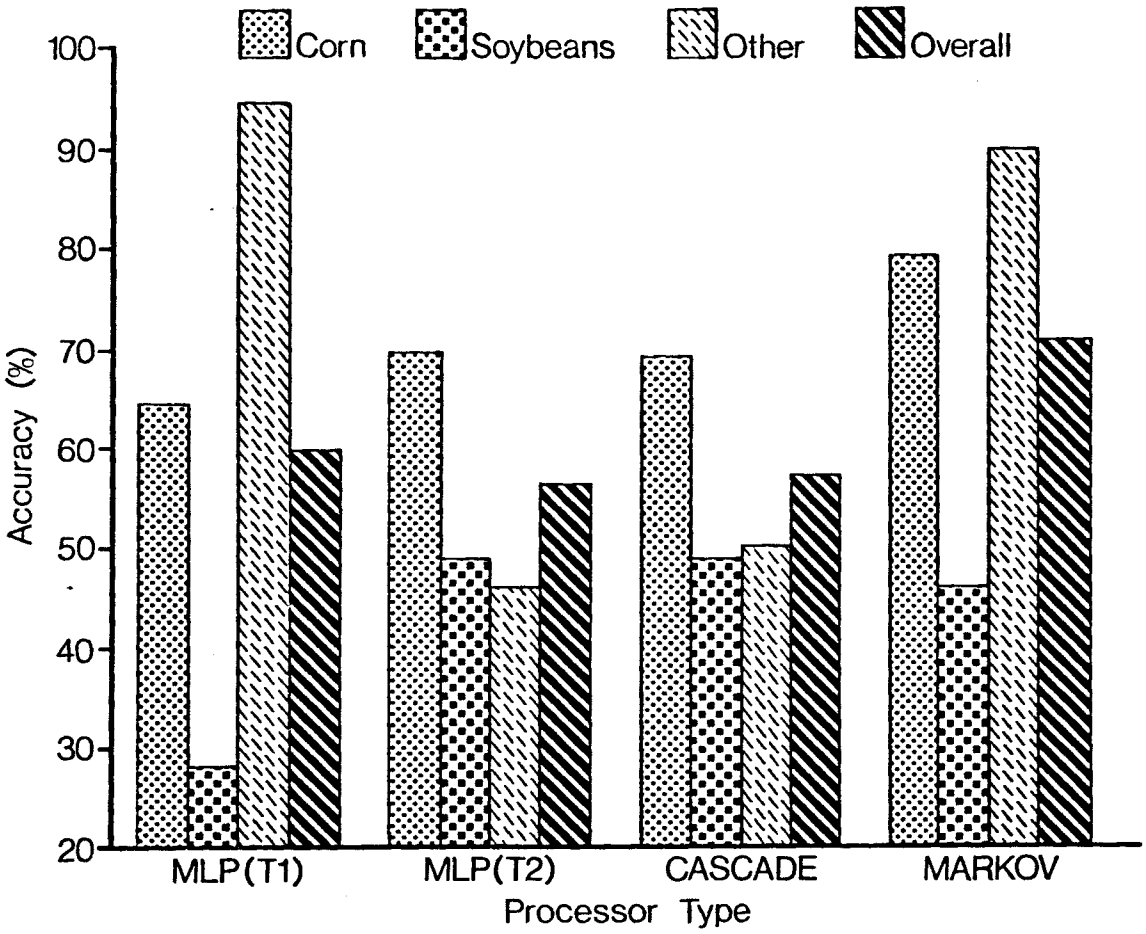

Figure 2.4 Overall classification performance vs. processing scheme (Henry County data; June 9 and July 16).

Table 2.1 Classification performance by class for different classifier (Henry County data; June 9 and July 16).

(a) June 9, 1978

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	64.8	260	114	27
2 SOYB	366	27.9	169	102	95
3 ELSE	285	94.0	9	8	268
TOTAL	1052	59.9	438	224	390

(b) July 16, 1978

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	69.6	279	19	103
2 SOYB	366	48.9	96	179	91
3 ELSE	285	45.6	123	32	130
TOTAL	1052	55.9	498	230	324

(c) Multitemporal Results (Cascade Classifier)

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	69.3	278	19	104
2 SOYB	366	48.4	100	177	89
3 ELSE	285	49.8	112	31	142
TOTAL	1052	56.7	490	227	335

(d) Multitemporal Results (Markov Classifier)

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	78.8	316	38	47
2 SOYB	366	45.4	87	166	113
3 ELSE	285	94.7	10	5	270
TOTAL	1052	71.0	413	209	430

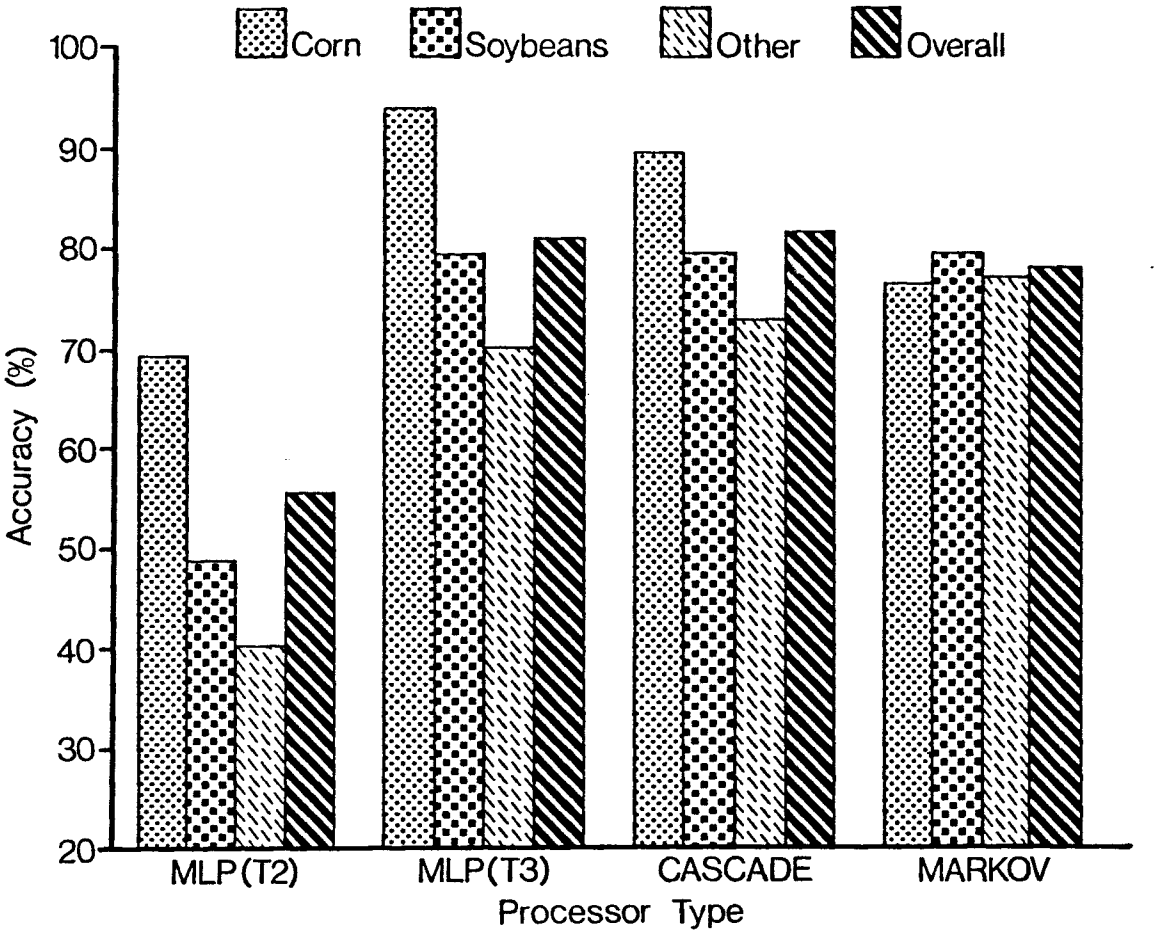

Figure 2.5 Overall classification performance vs. processor scheme (Henry County data; July 16 and August 20).

Table 2.2 Classification performance by class for different classifier (Henry County data; July 16 and August 20).

(a) July 16, 1978

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	69.6	279	19	103
2 SOYB	366	48.9	96	179	91
3 ELSE	285	45.6	123	32	130
TOTAL	1052	55.9	498	230	324

(b) August 20, 1978

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	93.8	376	5	20
2 SOYB	366	79.5	48	291	27
3 ELSE	285	65.6	87	11	187
TOTAL	1052	81.2	511	307	234

(c) Multitemporal Results (Cascade Classifier)

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	90.3	362	10	29
2 SOYB	366	79.8	47	292	27
3 ELSE	285	67.4	82	11	192
TOTAL	1052	80.4	491	313	248

(d) Multitemporal Results (Markov Classifier)

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	76.3	306	1	94
2 SOYB	366	79.2	1	290	75
3 ELSE	285	76.8	64	2	219
TOTAL	1052	77.5	371	293	388

very important in estimation of crop areas and crop production.

Experiment 3.3

The same data set and the same training procedure used in Experiments 1 and 2 are used here, except t = August 20, and t-1 = June 9 (71 days apart). The classification results are given in Figure 2.6 and Table 2.3. Again, the Markov pixel classifier has higher accuracy than the maximum likelihood and cascade pixel classifiers. As shown in Table 2.3 the performance of the maximum likelihood classifier in June is poor and in August is reasonably good because corn and soybeans are separable in August but not in June. However, we see that the Markov classifier has improved the overall performance by about five percent by incorporating the temporal correlation of observations in June and August into the classification process.

2.7 Summary and Conclusions

The temporal variation of energy has been considered as the output of a stochastic dynamic system. Then based on the assumption that the observed temporal data are a Gauss-Markov sequence, a new-classifier, the so-called Markov classifier, is developed. This stochastic model successfully utilizes multitemporal data characteristics. Actually, the spectral development curves of the classes have been

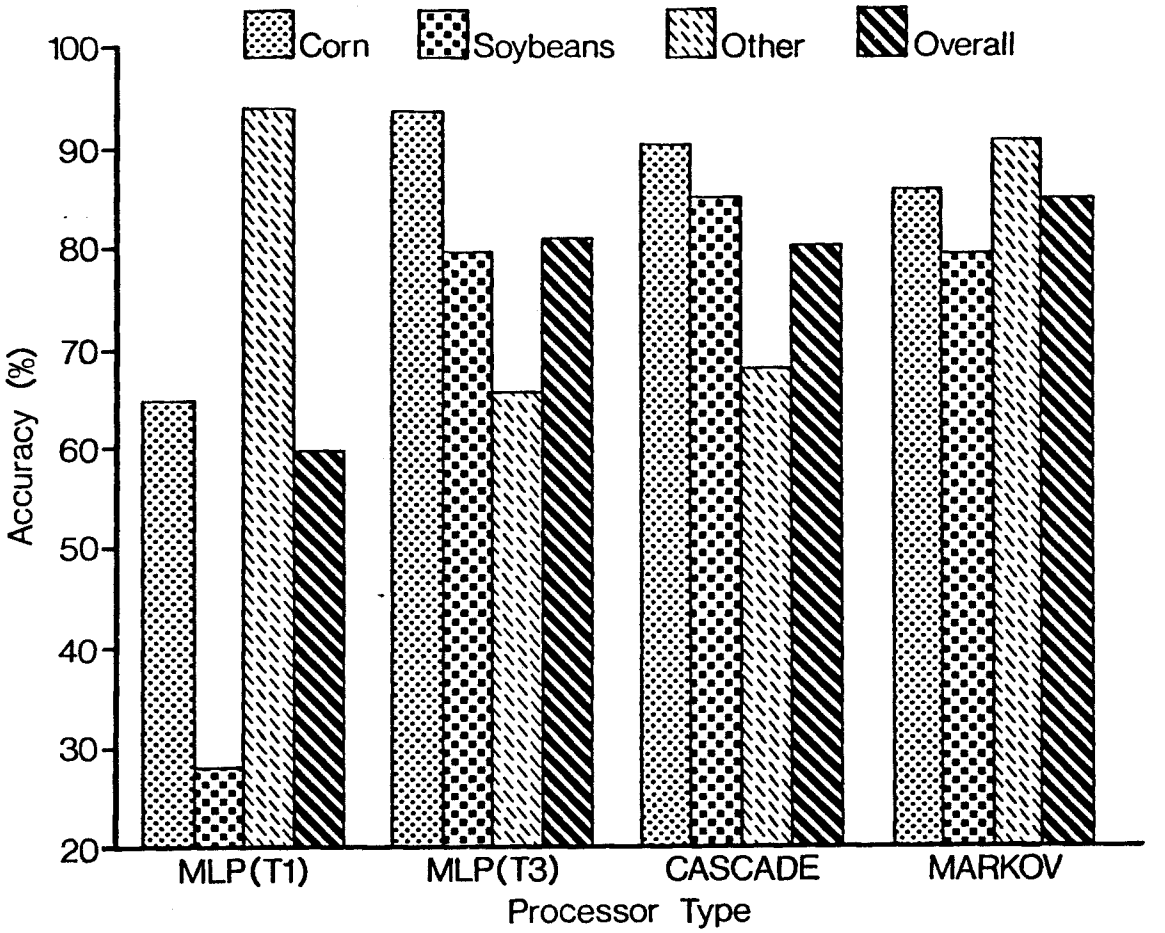

Figure 2.6 Overall classification performance vs. processor scheme (Henry County data; July 9 and August 20).

Table 2.3 Classification performance by class for different classifier (Henry County data; June 9 and August 20).

(a) June 9, 1978 data

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	64.8	260	114	27
2 SOYB	366	27.9	169	102	95
3 ELSE	285	94.0	9	8	268
TOTAL	1052	59.9	438	224	390

(b) August 20, 1978

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	93.8	376	5	20
2 SOYB	366	79.5	48	291	27
3 ELSE	285	65.6	87	11	187
TOTAL	1052	82.1	511	307	234

(c) Multitemporal Results (Cascade Classifier)

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	90.0	361	10	30
2 SOYB	366	79.8	47	292	27
3 ELSE	285	66.3	85	11	189
TOTAL	1052	80.0	493	313	246

(d) Multitemporal Results (Markov Classifier)

Group	No. of Samples	Percent Correct	No. of Samples Classified Into CORN	SOYBEANS	OTHERS
1 CORN	401	85.0	341	28	32
2 SOYB	366	79.5	1	291	74
3 ELSE	285	91.2	11	14	260
TOTAL	1052	84.8	353	333	366

modeled by a stochastic Markov process. The experimental results show that the Markov classifier has significantly better performance than the maximum likelihood and cascade pixel classifiers.

CHAPTER 3

PROBABILISTIC RELAXATION ON MULTITYPE DATA

Classification of multispectral image data based on spectral information has been a common practice in the analysis of remote sensing data. However, the results produced by current classification algorithms necessarily contain residual inaccuracies and class ambiguity. By the use of other available sources of information, such as spatial, temporal, and ancillary information, it is possible to reduce this class ambiguity and in the process improve the accuracy.

In this chapter, probabilistic and supervised relaxation techniques are adapted to the problem. The probabilistic relaxation labeling algorithm (PRL) given in [26], which in remote sensing pixel labeling usually improves performance but deteriorates after the optimum number of iterations, is modified. Experimental results show that the modified relaxation algorithm reduces the labeling error in the first few iterations, then remains constant at the achieved minimum error. Also a noniterative labeling algorithm which has a performance similar to that of the modified PRL is developed. Experimental results from Landsat and Skylab data are included.

3.1 Probabilistic Labeling

Our objective is to develop heuristic algorithms to utilize a combination of spectral, spatial, temporal, and ancillary information. In remote sensing, the spectral variations of electromagnetic energy of the scene have been studied extensively. The spectral response, which is a function of wavelength, has been modeled as a random process [81,82,84,86]. Another source of useful information is the spatial context of a pixel.

The dependencies between pixel labels are referred to as contextual information. In many pattern recognition problems, there exists contextual information which describes the spatial dependencies among the patterns to be recognized [13]. Also, temporal variations in the scene and available ancillary data, such as topographic data, pixel radar response, and classification labeling maps, are known to be information-bearing [72]. Based on these sources of information, the class membership probabilities may be estimated by probabilistic labeling methods.

3.1.1 Probabilistic Labeling by Maximum Likelihood Classifier

Probabilistic labeling is a process of estimating the initial labeling probabilities. Let X be a point in q-dimensional measurement space containing m classes. Also assume that the probability density function associated

with each class is Gaussian. Let $p(X|\omega_k)$ and $P(\omega_k)$ be the class-conditional density function and prior probability of the kth class, respectively. To characterize each class, the class mean vector and covariance matrix are estimated from training samples. Then pixel-label probabilities are estimated by calculating the a posteriori probabilities $P(\omega_k|X)$, as follows:

$$P_i^o(\omega_k) \overset{L}{=} P(\omega_k|X) = \frac{p(X|\omega_k)\ P(\omega_k)}{\sum_{\ell=1}^{m} p(X|\omega_\ell)\ P(\omega_\ell)}$$

$$k = 1, 2, \ldots m \qquad (3.1)$$

where $P_i^o(\omega_k)$ is the initial estimate of probability of the ith pixel's label.

3.1.2 Probabilistic Labeling by Cascade Classifier

To utilize spectral and temporal information jointly, a classifier based on the Bayesian strategy has been proposed [67]. Let $X(t_2)$ and $X(t_1)$ be multivariate observations at time t_2 and time t_1, respectively. And let $\{V_j,\ j = 1,2,3\ldots m_1\}$, and $\{\omega_K,\ K = 1,2,\ldots m_2\}$ be the set of possible classes at time t_1 and time t_2, respectively. It can be shown that [67] the estimate of class membership for bitemporal observations is given by

$$P_i^{\ 0}(\omega_k) = \frac{p(X(t_2)|\omega_k) \sum_{\ell=1}^{m_1} p(X(t_1)|V_\ell)P(\omega_k,V_\ell)}{\sum_{j=1}^{m_2} p(X(t_2)|\omega_j) \sum_{\ell=1}^{m_1} p(X(t_1)|V_\ell)P(\omega_j,V_\ell)} \qquad (3.2)$$

In practice, $p(X(t_2)|\omega_k)$ and $p(X(t_1)|V_\ell)$ are available after estimating their corresponding mean vectors and covariance matrices from training data. As suggested in [67], the prior joint probability $P(\omega_k, V_\ell)$ may be estimated as:

$$P(\omega_k, V_\ell) = P(\omega_k|V_\ell)P(V_\ell) \tag{3.3}$$

Assuming $P(V_\ell) = \frac{1}{m_1}$, then the transition probabilities are given by

$$P(\omega_k|V_\ell) = P^O \text{ for } \omega_k = V_\ell$$

and

$$P(\omega_k|V_\ell) = \frac{1-P^O}{m_2-1} \text{ for } \omega_k \neq V_\ell \tag{3.4}$$

where $0 \leq P^O \leq 1$.

3.1.3 Probabilistic Labeling by Markov Classifier

The Markov classifier, which utilizes multitemporal information very effectively, is discussed in Chapter 2. For bitemporal obvservations pixel-label probabilities can be estimated as:

$$P_i^{\,0}(\omega_k) \triangleq \frac{p(X(t)|X(t-1),\omega_k)\,p(x(t-1)|\omega_k)}{\sum_{\ell=1}^{m} p(X(t)|X(t-1),\omega_\ell)\,p(X(t-1)|\omega_\ell)} \tag{3.5}$$

The main objective of estimating the initial labeling probabilities by the cascade or Markov classifier is to

incorporate the temporal information in the classification process and, therefore, reduce the ambiguity of the initial labeling. Later, we will show how the initial labeling ambiguity can be reduced even more by utilizing spatial information.

However, if the initial labeling probabilities cannot be statistically estimated, then we may assign probabilities to the predetermined labels, as follows:

$$P_i^o(\omega_k) = W$$

$$P_i^o(\omega_\ell) = \frac{1-W}{m-1}, \quad \ell = 1, 2, \ldots m \qquad (3.6)$$

$$\ell \neq k$$

where it is assumed that the ith pixel's label is ω_k and $\frac{1}{m} < W \leq 1$. This way of assigning the initial labeling probabilities will be referred to as the arbitrary weighting method because we use the initial labeling and then assign arbitrary weights which are between zero and one and agree with the labeling.

This weighting method is faster computationally because the previously mentioned methods for probabilistic labeling are not needed. However, this method also is biased. We know that in order to support the initial labeling, the labeling weight W should be greater than $\frac{1}{m}$ and less than or equal to one. Choosing $\frac{1}{m} < W \leq 1$ is not

justifiable unless we have some prior information about the performance of the algorithm which performs the initial labeling. Then based on that, we may be able to make the range for W narrow.

3.2 Existing Algorithms Utilizing Multitype Information

3.2.1 Spectral-Spatial Classifier Based on Compound Decision Theory

In general, contextual classifiers attempt to incorporate pixel context or information surrounding a pixel into the identification process of that pixel. Multispectral image data consist of a set Ω, set $\{X(s), s \in \Omega\}$ and set $C = \{1,2,...m\}$ where Ω is a two-dimensional array of pixel locations, i.e., $\Omega = \{s = (i,j),\ 1 \leq i \leq I,\ 1 \leq j \leq J\}$, and $X(s)$ is a q-dimensional random observation at point s. Also let N denote the neighbor set. Typically, $N = \{(0,0), (0,-1), (-1.0)\}$ or $N = \{(0,0), (0,-1)\ (-1,0), (0,1)\ (1,0)\}$. Illustrative examples of different neighbor sets are given in Figure 3.1. The discriminant function for a pixel at point $s + (0,0) = s$ based on its neighborhood is given by [14-16]:

$$P(\omega(s) \mid \{X(s + (i,j)),\ (i,j) \in N\}) \propto \sum_{\substack{(i,j) \in N \\ \omega(s+(i,j)) \in C \\ (i,j) \neq (0,0)}} \left(\left[\prod_{(k,\ell) \in N} p(X(s + (k,\ell)) \mid \omega(s + (k,\ell))) \right] \cdot P(\{\omega(s + (i,j)),\ (i,j) \in N\}) \right) \tag{3.7}$$

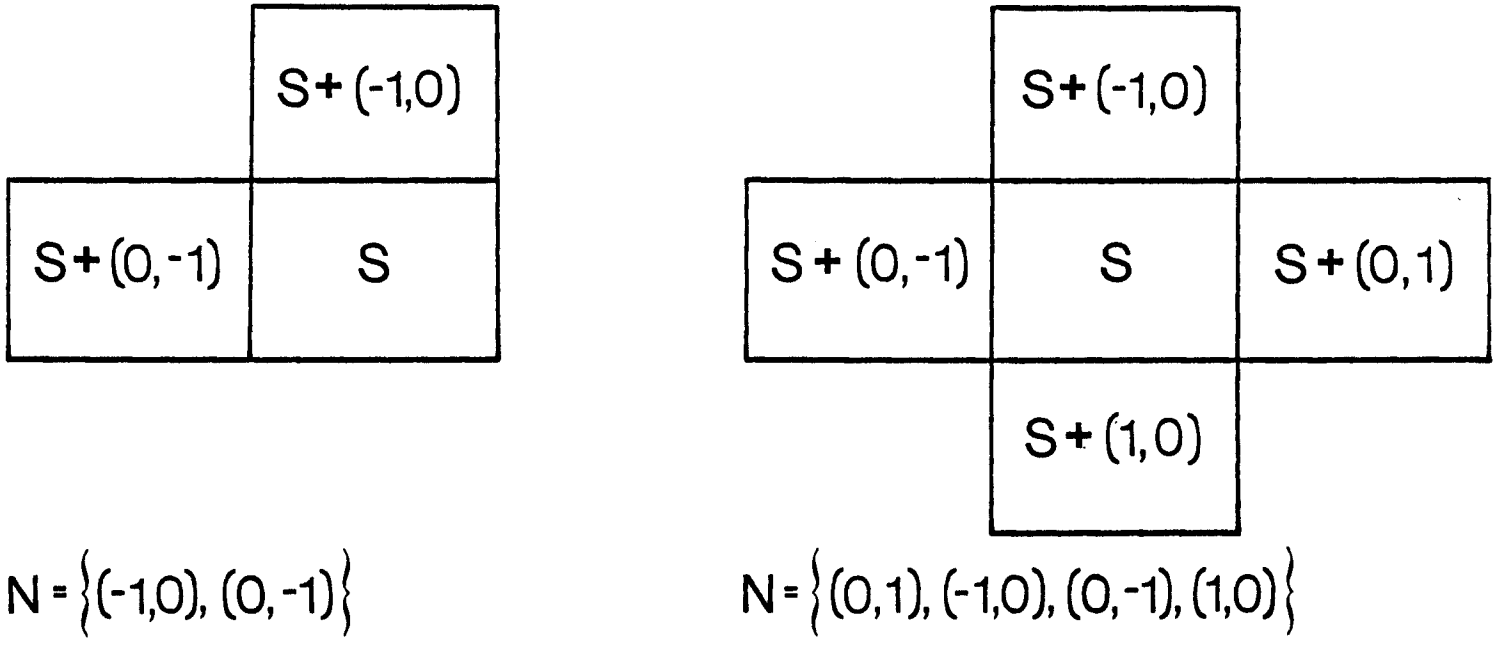

Figure 3.1 Examples of different neighbor sets.

where $X(s)$ q-dimensional random measurement for a pixel at point s;

$\omega(s)$ class of a pixel at point s;

$p(X(s)|\omega(s))$ class-conditional density function of $X(s)$;

$p(\{\omega(s+(i,j)) \quad , (i,j) \varepsilon N\})$

is the joint probability set of possible classes in the neighborhood.

The contextual information is contained in $p(\{\omega(s+(i,j)), (i,j) \varepsilon N\})$. However, in practice, this has to be estimated from a prelabeling process and to have a good estimate of $P(\{\omega(s+(i,j)), (i,j) \varepsilon N\})$ is computationally costly. A simpler algorithm which attempts to utilize contextual information for further study is given in Appendix B.

3.2.2 Utilizing Spectral, Spatial Characteristics by Probablistic Relaxation Algorithm

Relaxation labeling processes use an iterative heuristic approach which attempts to extract contextual

information in a scene to reduce the ambiguity of a predetermined labeling. Relaxation labeling techniques use two sources of information, an initial (ambiguous) labeling and information imbedded in the spatial context of a pixel. A block diagram of a post classifier which utilizes probabilistic and supervised relaxation is given in Figure 3.2.

Let us consider the probabilistic relaxation algorithm which has been suggested by Zucker et al. [26]. Let $P_i^n(\omega_k)$ denote the estimate of the probability that on the nth iteration the label or class of the ith pixel of a scene is ω_k; k = 1, 2,...m. Then define

$$P_i^{n+1}(\omega_k) = \frac{P_i^n(\omega_k)\ Q_i^n(\omega_k)}{\sum_{\ell=1}^{m} P_i^n(\omega_\ell)\ Q_i^n(\omega_\ell)} \tag{3.8}$$

where $Q_i^n(\omega_k)$ is called the neighborhood function and is defined by

$$Q_i^n(\omega_k) = \sum_{j=1}^{J} d_{ij} \sum_{\ell=1}^{m} P_{ij}(\omega_k|\omega_\ell)\ P_j^n(\omega_\ell) \tag{3.9}$$

In this equation $P_{ij}(\omega_k|\omega_\ell)$ is the probability that pixel i is from class ω_k given that pixel j is from class ω_ℓ. The d_{ij} are a set of neighborhood weights which satisfy

$$\sum_{j=1}^{J} d_{ij} = 1 \tag{3.10}$$

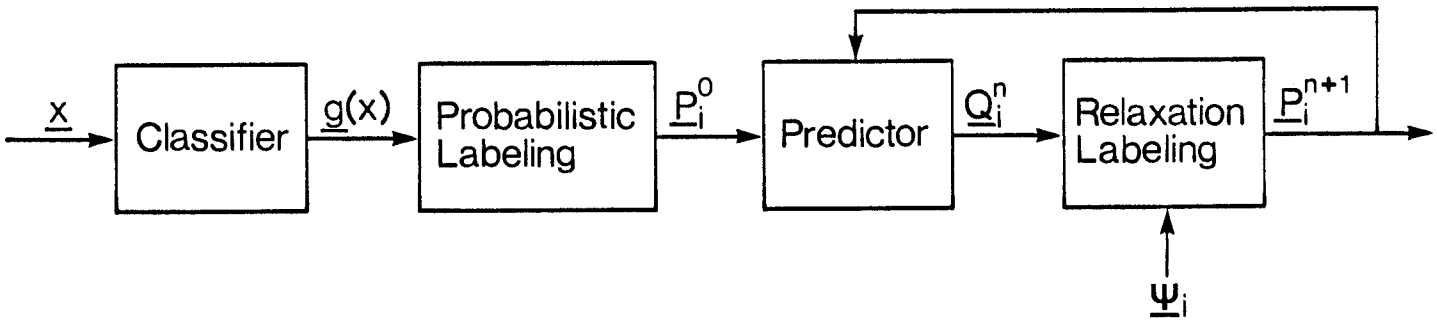

Figure 3.2 Block diagram of a post classifier.

with J as the number of pixels in the neighborhood and m as the number of classes. Examples of J = 5 and J = 9 are given in Figure 3.3. In all our analysis, the J = 5 neighborhood will be used.

3.2.3 Utilizing Spectral, Spatial, Ancillary Information by a Supervised Relaxation Algorithm

Supervised relaxation processes [64] are a more general version of probabilistic relaxation methods which attempt to utilize multitype data characteristics. In supervised relaxation, first an appropriate likelihood for the label of each pixel is estimated based on the statistical information of available data. Then the neighborhood function for the label most favored by ancillary data is increased and others decreased in proportion to their support from the ancillary data source. The relaxation algorithm does not know, of course, which are the correct and which are the incorrect labels. It only "knows" which labels are consistent with their neighbors and with the ancillary data. Consequently, an image with initial labeling errors will be iterated until consistency between spectral, spatial and ancillary information is achieved.

Let us consider the supervised relaxation algorithm which is suggested by Richards et al. in [64].

$$P_i^{n+1}(\omega_k) = \frac{P_i^n(\omega_k)\ R_i^n(\omega_k)}{\sum_{\ell=1}^{m} P_i^n(\omega_\ell)\ R_i^n(\omega_\ell)} \qquad (3.11)$$

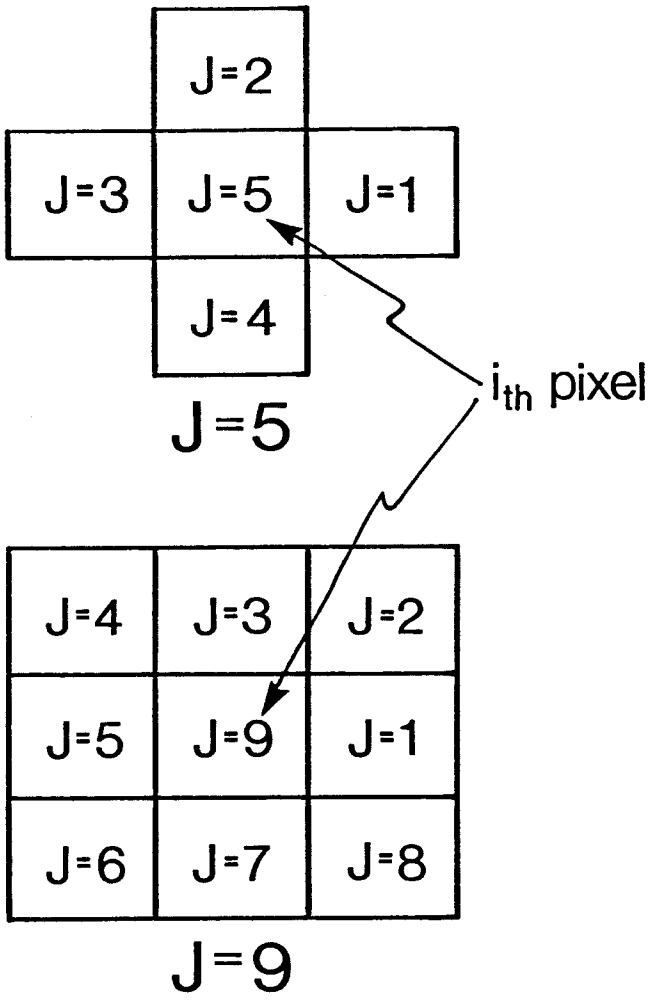

Figure 3.3 Example J-pixel neighborhoods.

where $$R_i^n(\omega_k) = Q_i^n(\omega_k)\Psi_i(\omega_k) \tag{3.12}$$

$$Q_i^n(\omega_k) = \sum_{j=1}^{J} d_{ij} \sum_{\ell=1}^{m} P_{ij}(\omega_k|\omega_\ell)\ P_j^n(\omega_\ell) \tag{3.13}$$

and

$$\Psi_i(\omega_k) = [1+\beta\ (m\phi_i(\omega_k) - 1)] \tag{3.14}$$

In the above equations $P_i^n(\omega_k), Q_i^n(\omega_k)$ are the same as we defined earlier and $\Psi_i(\omega_k)$ is an estimate of the likelihood for the ith pixel's label on basis of ancillary data. In (Eq. 3.14), $\phi_i(\omega_k)$ is the probability that the ith pixel belongs to class ω_k based on ancillary information, and β is a parameter that adjusts the degree of supervision; it is between zero and one. The parameter β is chosen heuristically; however, it should reflect one's confidence in the ancillary data in comparison to the other sources of information. As before, m is the number of possible classes or labels.

3.3 Proposed Algorithms for Utilizing Multitype Information

The spatial context of a pixel or dependency among the labels in a neighborhood is incorporated via $P_{ij}(\omega_k|\omega_\ell)$, the transition probability that pixel i is from class ω_k given that pixel j (one of its neighbors) is from ω_ℓ. In past

practice, $P_{ij}(\omega_k|\omega_\ell)$ has been estimated from the result of probabilistic labeling over the whole data set, which means the transition probabilities are assumed constant over the data set. In fact, in an actual data set, they may be expected to vary from place to place. What we are suggesting is, $P_{ij}(\omega_k|\omega_\ell)$ should slowly vary over the data set and the following procedure is suggested to estimate these transition probabilities.

1. Depending on the number of classes, choose a square window of size L x L centered at the ith pixel. For example, for two classes, we have chosen a window of size 5 x 5 and for the three classes a window of size 6 x 6 may be considered.

2. Estimate the probability of jth pixel's label by

$$P_j(\omega_k) = \frac{1}{L^2} \sum_{r=1}^{L^2} P^o_{jr}(\omega_k) \tag{3.15}$$

where $P_{jr}(\omega_k)$ is the initial estimate of a pixel's label at location jr of the chosen window.

3. Estimate the transition probability by

$$P_{ij}(\omega_k|\omega_\ell) = \frac{P_{ij}(\omega_k,\omega_\ell)}{P_j(\omega_\ell)} \tag{3.16}$$

and the joint probability by

$$P_{ij}(\omega_k,\omega_\ell) = \frac{1}{(L-1)^2} \sum_{r=1}^{(L-1)^2} P_r^o(\omega_k)\left[\tfrac{1}{4}\sum_{j=1}^{4} P_{rj}^o(\omega_\ell)\right] \qquad (3.17)$$

where $P_r^o(\omega_k)$ is the initial estimate of rth pixel surrounding the ith pixel and including ith pixel itself. And $P_{rj}^o(\omega_\ell)$ is the initial estimate of jth pixel surrounding the rth pixel but excluding it.

Now, by using this adaptive procedure, the spatial context of each pixel is estimated and incorporated by the neighborhood function to predict the estimate of the probability of each pixel's label. It is believed this simple algorithm can extract most of the contextual information by only one iteration. The adaptive labeling algorithm is given by:

$$Q_i^n(\omega_k) = \sum_{j=1}^{J} d_{ij} \sum_{\ell=1}^{m} P_{ij}(\omega_k|\omega_\ell) P_j^n(\omega_\ell) \qquad (3.18)$$

Let $d_{ij} = \frac{1-d_i}{J-1}$, then it can be shown that

$$Q_i^n(\omega_k) = q_i^n(\omega_k) + d_i[P_i^n(\omega_k) - q_i^n(\omega_k)] \qquad (3.19)$$

where

$$q_i^n(\omega_k) = \sum_{\ell} P_{ij}(\omega_k|\omega_\ell)\left[\frac{1}{J-1}\sum_{j=1}^{J-1} P_j^n(\omega_\ell)\right] \qquad (3.20)$$

and

$$P_i^{n+1}(\omega_k) = q_i^n(\omega_k) + d_i\left[P_i^n(\omega_k) - q_i^n(\omega_k)\right] \qquad (3.21)$$

The new formulation of the probabilistic relaxation will therefore be

$$p_i^{n+1}(\omega_k) = \frac{d_i\left[p_i^n(\omega_k)\right]^2 + (1-d_i)p_i^n(\omega_k)\ q_i^n(\omega_k)}{\sum_{\ell=1}^{m}\left(d_i\left[p_i^n(\omega_\ell)\right]^2 + (1-d_i)p_i^n(\omega_\ell)q_i^n(\omega_\ell)\right)} \tag{3.22}$$

In Eq. 3.17, if we let $d_i = 1-\gamma_i$, then we can write

$$p_i^{n+1}(\omega_k) = p_i^n(\omega_k) + \gamma_i[q_i^n(\omega_k) - p_i^n(\omega_k)] \tag{3.23}$$

A summary of all the algorithms is given in Table 3.1.

In the above algorithms, if $d_i = 0.0$, then the label of the ith pixel will be decided, based on spatial information (assuming its initial label probability is not zero or one). If $d_i = 1.0$, then we are not using any spatial information for the ith pixel.

As mentioned in Section 3.2.3, the supervised probabilistic relaxation algorithms are heuristic techniques which attempt to reduce the ambiguity of a predetermined labeling by measuring consistency of pixel labels based on multitype data characteristics. Labeling consistency is measured by multiplying appropriate label likelihoods, which can be obtained from spectral, spatial, and ancillary

Table 3.1 Summary of probabilistic and supervised relaxation algorithms.

Algorithm		Probability	
		Initial Labeling	Transition
(1) Probabilistic Relaxation Labeling (PRL)	$P_i^{n+1}(\omega_k) = \frac{P_i^n(\omega_k)Q_i^n(\omega_k)}{\sum_{\ell=1}^{m} P_i^n(\omega_\ell)Q_i^n(\omega_\ell)}$ $Q_i^n(\omega_k) = \sum_{j=1}^{J} d_{ij} \sum_{\ell} P_{ij}(\omega_k \mid \omega_\ell) P_j^n(\omega_\ell)$	Weighting method Probabilistic labeling	Over the region Window
(2) Iterative Adaptive Labeling (IAL)	$P_i^{n+1}(\omega_k) = P_i^n(\omega_k) + \gamma_i [q_i^n(\omega_k) - P_i^n(\omega_k)]$ $q_i^n(\omega_k) = \sum_{\ell} P_{ij}(\omega_k \mid \omega_\ell) [\frac{1}{J-1} \sum_{j=1}^{J-1} P_j^n(\omega_\ell)]$ $0 \leq \gamma_i \leq 1$	Weighting method Probabilistic labeling	Over the region Window
(3) Non-Iterative Adaptive Labeling (NAL)	The same as Algorithm 2 with only one iteration		
Supervised Relaxation	The supervised version of Algorithms 1, 2, and 3 will be referred to as algorithms 4, 5, and 6, respectively.		

information. In our analysis, the following ancillary information was utilized:

1. Probability of a given label based on elevation data. This data represents a quantitative version of the fact that some classes are more likely than others at a given elevation; this is particularly true in regions of high terrain relief. If we constantly remind the relaxation process about these features, then the algorithm performance may be expected to improve.

2. Objects in the scene having narrow shapes, such as roads and rivers ("geometric features") may consist of spectrally separable classes which can be accurately classified by maximum likelihood and minimum distance pixel classifiers. If the labeling results of these classifiers are used to supervise the relaxation, the correct labeling of these geometric features can be preserved.

3. The results of classification based on temporal information, for example at time t-1, can be used to supervise relaxation labeling at time t or vice versa.

3.4 Experimental Results

In order to evaluate the performance of the above heuristic algorithms, two data sets were selected. Data Set 1 was multitemporal spatially registered Landsat MSS data acquired over Henry County, Indiana in 1978. Data Set 2 was multispectral Skylab S-192 data from northeast of the Vallecito Reservoir region in the Colorado Rockies. This data set was classified into a number of tree species using the maximum likelihood classifier. The classification map so produced was rearranged for simplicity into the two categories of spruce/fir and other. For the region, elevation data as well as a probability model for the occurrence of spruce/fir vs. elevation were chosen as an ancillary data variable. Information about software systems and data sets is given in Appendix G.

Experiment 3.1

The objective of this experiment is simply to investigate the probability of error when the initial labeling probabilities are assigned by the weighting method. Therefore, using the maximum likelihood classifier, Data Set 1 (Landsat MSS data), collected on August 20, was classified into corn/soybean and other. A block of 30 x 30 pixels from this data set was selected, and by the weighting method, the initial labeling probabilities were assigned. Then Algorithm 1 (PRL) was applied to the selected block of data, using 40 iterations. The result is shown in Figure 3.4. It suggests the following:

1. If the result of the spectral classifier is weakly supported ($P_i^o(\omega_k) \simeq \frac{1}{m}$) for all pixels), then there is really no useful spatial information available for the relaxation process to utilize to reduce the ambiguity. As seen in Figure 3.4, the results of relaxation applied in such a situation may be even worse than the initial labeling error.

2. Since choosing W arbitrarily is not justifiable, the initial labeling probabilities should be estimated from the probability density function, if they are available.

The effect of relaxation from a spatial standpoint can be seen by comparing Figures 3.5 (ground truth), 3.6 (initial labeling and 3.7 (final labeling). To be clearer, a spatial or geographic standpoint usually means the true labels are spatially clustered and this is especially true for agricultural fields. However, if we look at the ground truth provided by an operator (Figure 3.5), we may observe isolated pixels from a class surrounded by another class. Simply, there may be ambiguity in the ground truth, and usually by examining the initial labeling which is shown in Figure 3.6, more ambiguity can be seen. However, the results of relaxation labeling after 40 iterations (Figure 3.7) may be closer to reality. Therefore, labeling error may be even less than what we have observed.

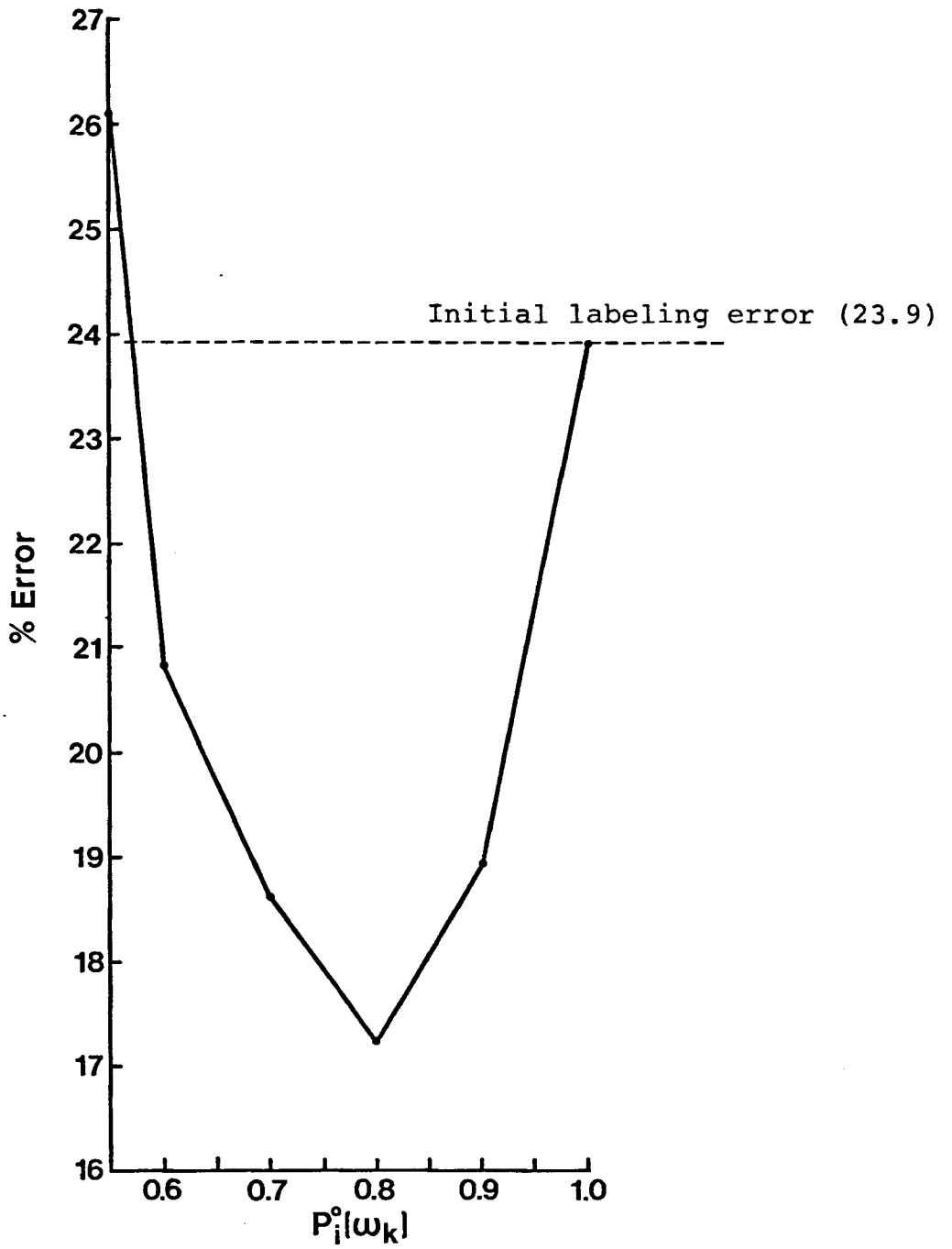

Figure 3.4 Error vs. the initial labeling probability at 40th iteration.

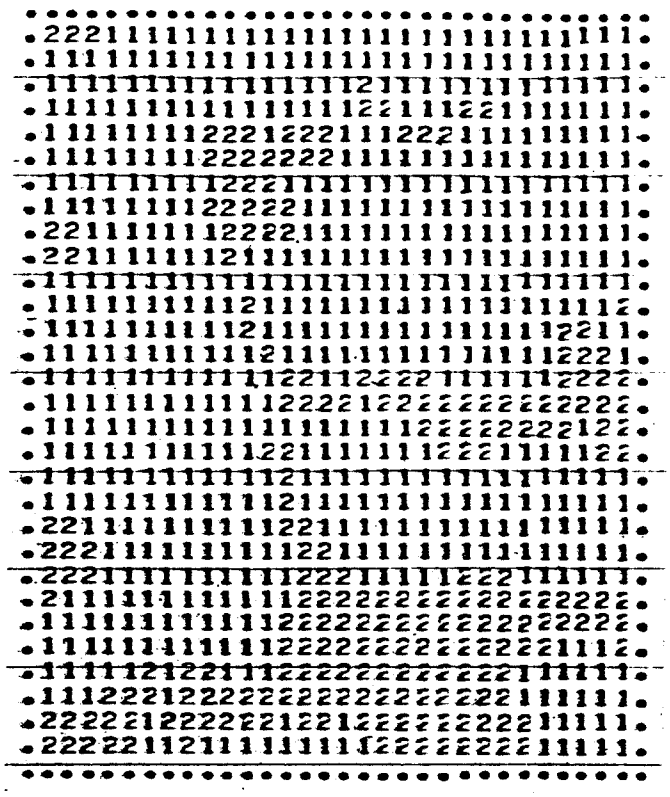

Figure 3.5 Two-category (corn/soybeans-1, other-2) ground truth for the 30 x 30 pixel Landsat MSS data acquired over Henry County, Indiana on August 20, 1978 (considered as "true" labeling).

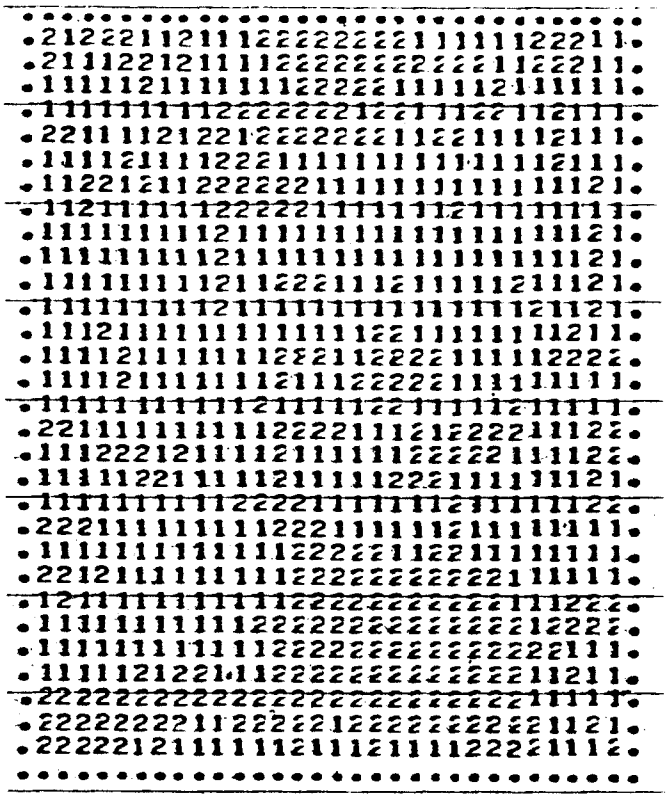

Figure 3.6 Initial labeling for the 30 x 30 pixel image, obtained from maximum likelihood classifier (% labeling error is 23.9).

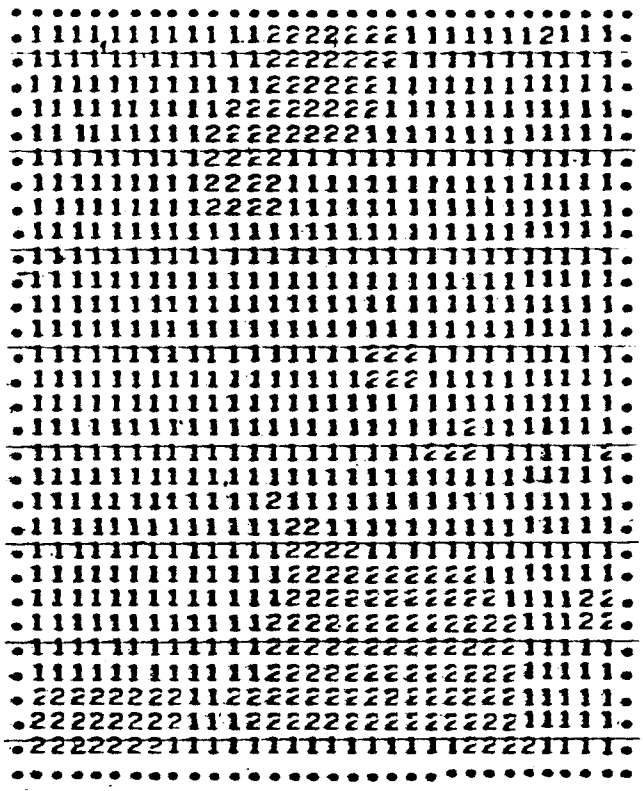

Figure 3.7 Final labeling of the 30 x 30 image after 40 iterations with d_i=0.10, $P_i^o(\omega_k)$=0.8, $P_{ij}(1/1)$=0.769 and $P_{ij}(2/2)$=0.693 (% labeling error is 17.2).

Experiment 3.2

A block of 40x30 pixels different from Experiment 3.1 was chosen from Data Set 1 (Landsat MSS data) collected on August 20. Then Algorithm 1 (PRL) with two different methods of estimating the initial probabilities
the selected block of data and the results are shown in Figure 3.8. The results suggest:

1. By employing the probabilistic relaxation labeling as a post classifier, we can reduce the probability of error.

2. The performance of PRL with probabilistic labeling is better than PRL with the weighting method.

Experiment 3.3

A block 30x30, the same as Experiment 3.1, was chosen from Data Set 1. Using the maximum likelihood classifier, the data were classified into two classes: corn/soybeans and other. Again a comparison of the weighting method and the probabilistic labeling by maximum likelihood is made, as shown in Figure 3.8. The same conclusions as in Experiment 3.2 are indicated from these results; however, the accuracy improvement is larger in this case.

Experiment 3.4

The objective of this experiment was to study the performance of Algorithm 1 (PRL) with two different ways of estimating the transition probability. The difference

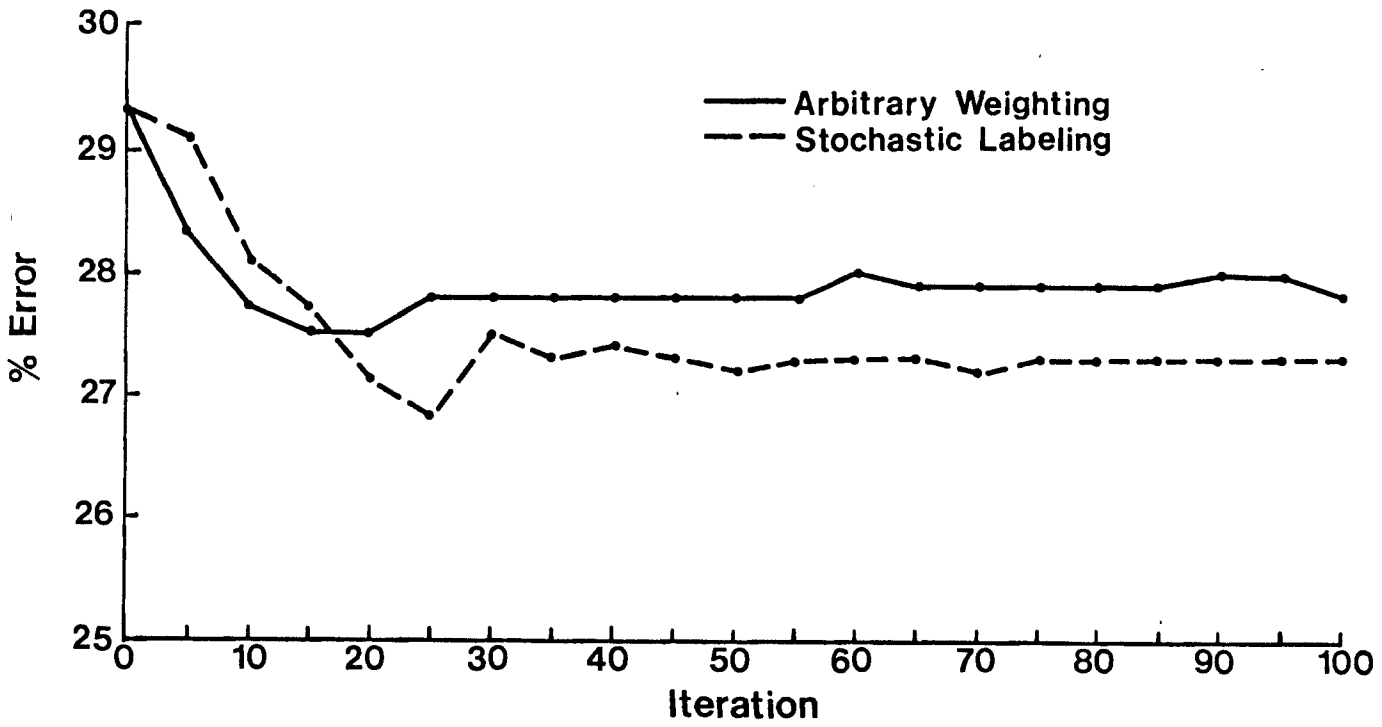

Figure 3.8 Comparison of arbitrary assigning probability to the initial labeling and probability assigned by probabilistic labeling.

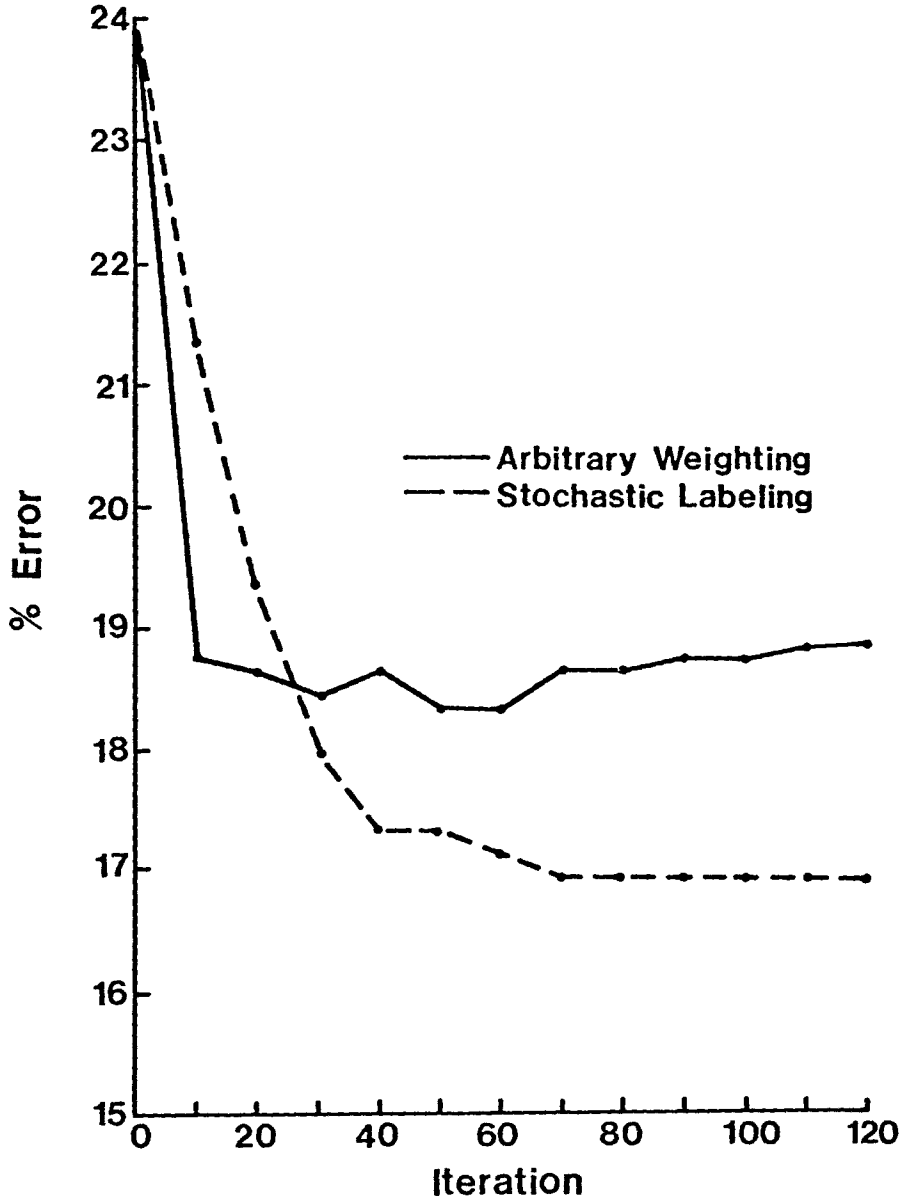

Figure 3.9 Comparison of arbitrary assigning probability to the initial labeling and probability assigned by probabilistic labeling.

between thses two methods is that one (region or block) is independent of the location and the other (window) is dependent on the location of a pixel under consideration, i.e., adaptive. The block data set of Experiment 3.2 was chosen and initial probabilities were estimated by probabilistic labeling based on the maximum likelihood classifier. The results are given in Figure 3.10 and indicate the following:

In both cases there was at least some improvement in accuracy; however, algorithm 1 (PRL), adaptively estimating the transition probability, does not exhibit the deterioration phase. The original algorithm suggested in [26] decreased the labeling error, passed through a turning point, and increased again before settling down to a pessimistic final value.

Experiment 3.5

A block of size 30x30 pixels from Data Set 1, collected on September 26, was chosen. Then the same procedures as in Experiment 3.4 were applied to this block of data. Results are given in Figure 3.11. The same conclusion as for Experiment 3.4 can be drawn from these results.

Experiment 3.6

The objective of this experiment was to study the performance of Algorithm 1 (PRL) and Algorithm 3 (NAL). A block of size 30x30 pixels from Data Set 1, collected on August 20, was chosen. Initial probabilities were

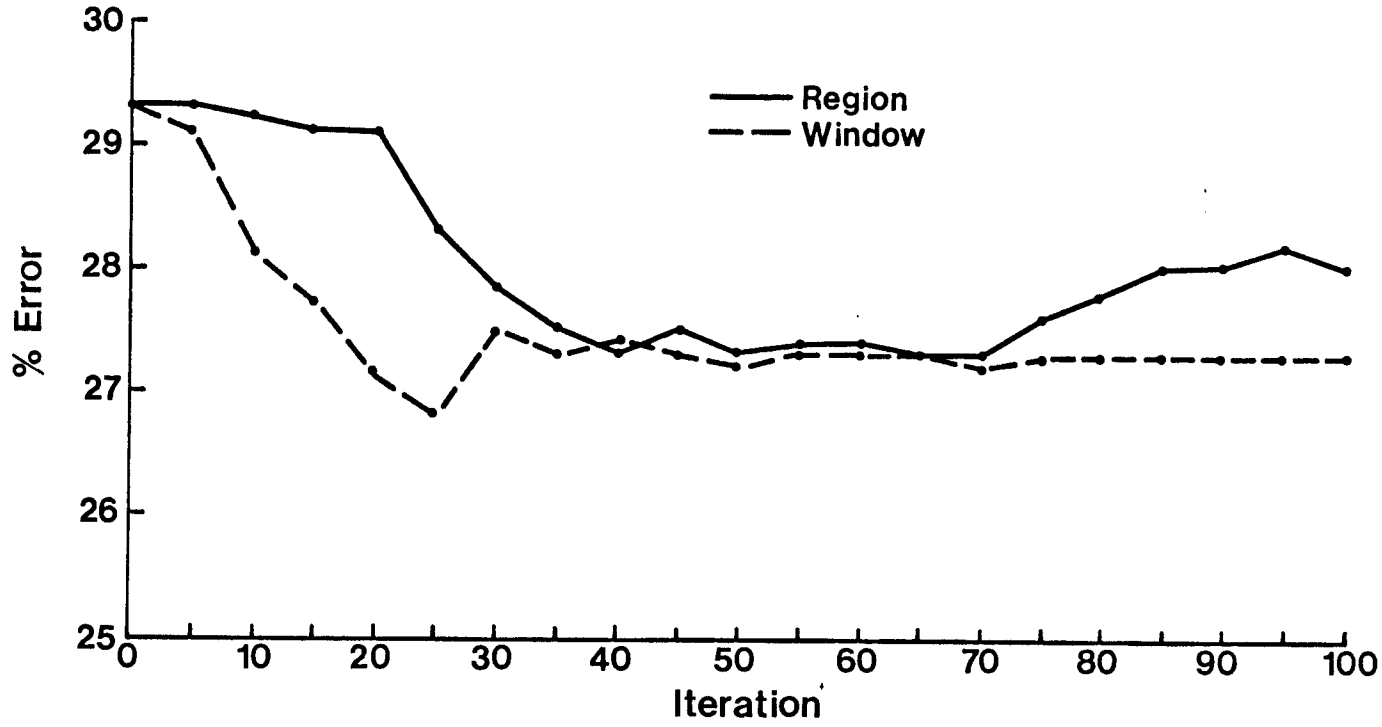

Figure 3.10 Comparison of performance of the Algorithm 1 with estimating the transition probability over a region and over a window.

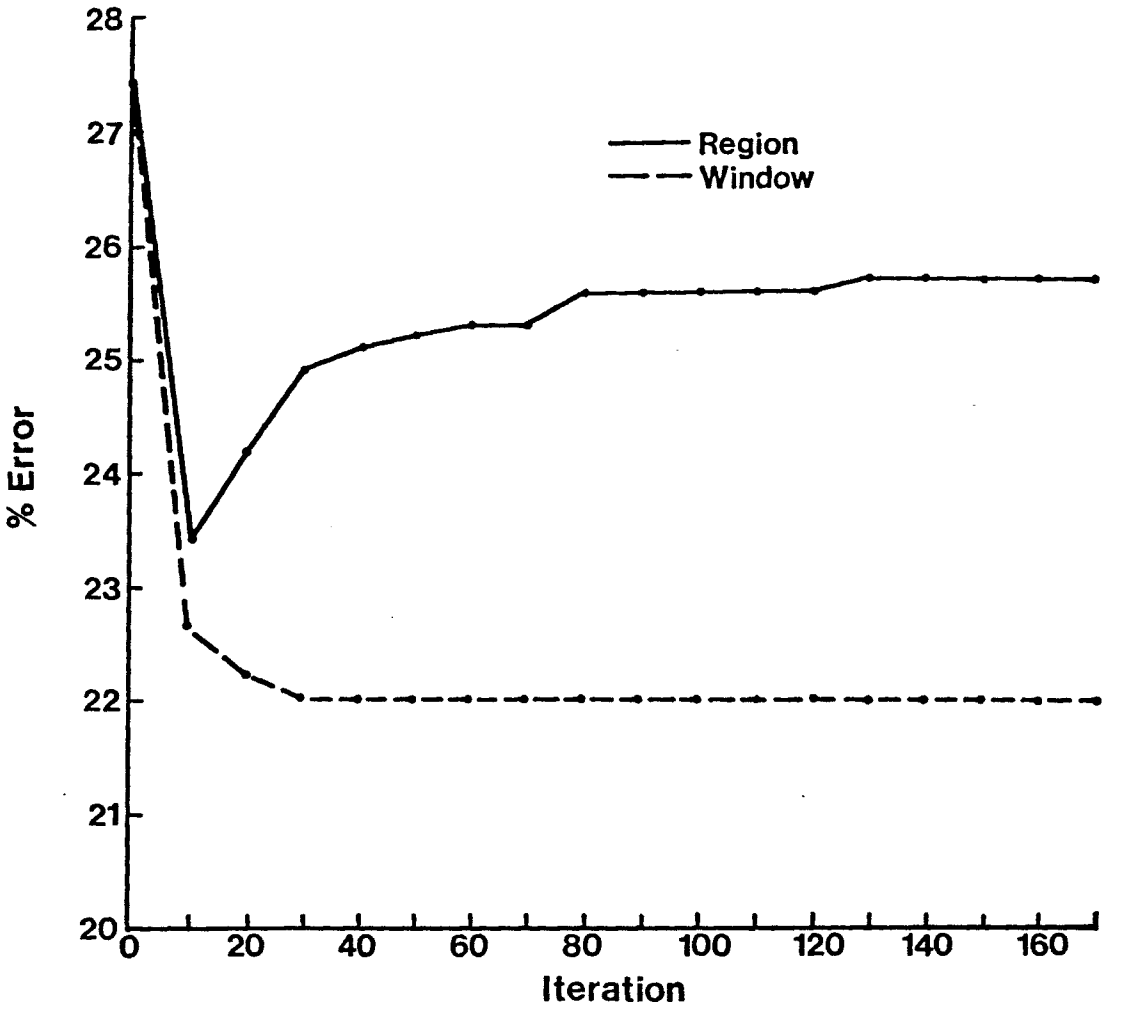

Figure 3.11 Comparison of performance of Algorithm 1 with estimating the transition probability over a region and over a window.

estimated by probabilistic labeling and the transition probabilities were estimated over a window of size 5x5 pixels. The parameter d_i was chosen, as in previous experiments, to be 0.1. Results are given in Figure 3.12. The results suggest that the performance of noniterative adaptive labeling and probabilistic labeling are almost the same.

Experiment 3.7

A block of size 30x30 pixels from Data Set 1, collected on September 26, was chosen. Then the same procedures as in Experiment 3.6 were applied to the block data. Results are given in Figure 3.13. The same conclusion as for Experiment 3.6 can be drawn from these results.

Experiment 3.8

A block of size 30x30 pixels from multitemporal Data Set 1, collected on August 20, 1978 (time t_1) and September 26, 1978 (time t_2), was chosen. The initial labeling probabilities at times t_1 and t_2 were estimated by the maximum likelihood method and the transition probabilities were estimated over a window of size 5x5 pixels. Algorithm 1 (PRL) with $d_i=1-\gamma_i=0.1$ was compared to Algorithm 4 (with $d_i=0.0$ and $\beta=0.5$). Information at time t_2 was used as ancillary information to supervise Algorithm 1. The objective of this experiment was to preserve some geometric

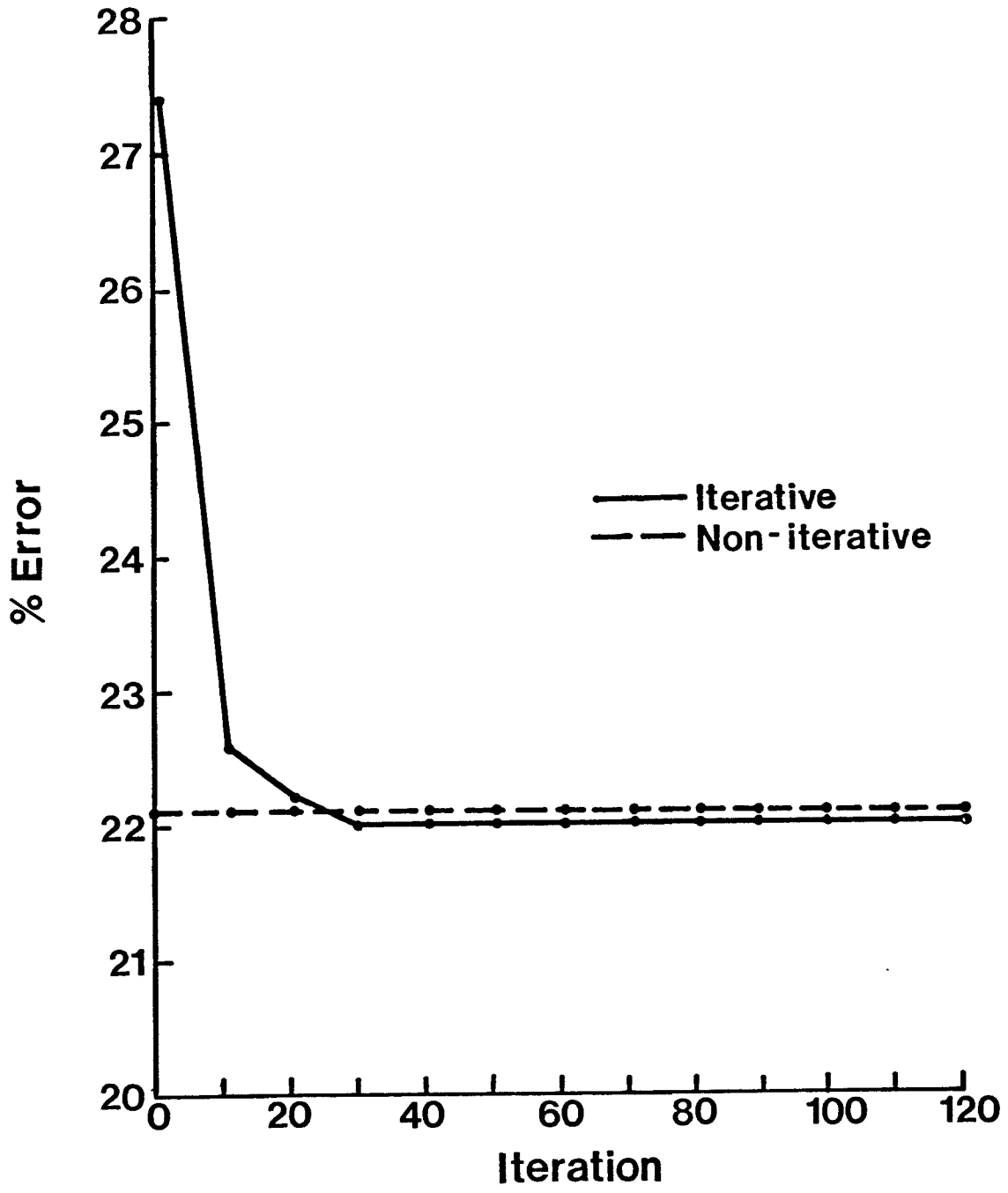

Figure 3.12 Comparison of iterative and non-iterative algorithms (Algorithms 1 and 3).

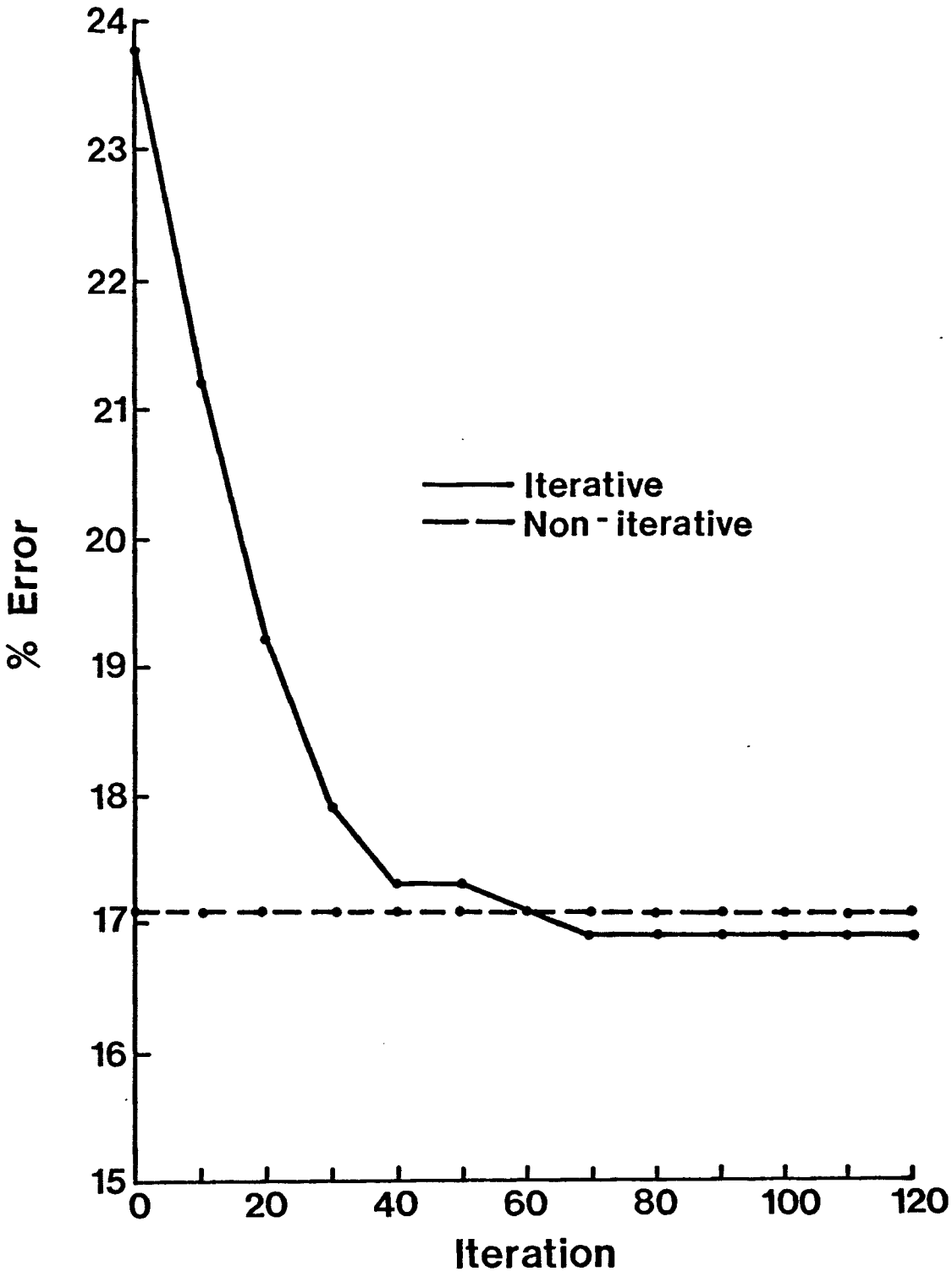

Figure 3.13 Comparison of iterative and non-iterative algorithms (Algorithms 1 and 3).

features (roads, line pixels and isolated pixels) and therefore improve the performance of Algorithm 1. The results are given in Figure 3.14.

Experiment 3.9

The objective of this experiment was to improve the performance of Algorithms 1 and 2 by supervising them by labeling results of a linear classifier. A block of 40 x 30 pixels from Data Set 1 was chosen. Then the performances of Algorithms 1,4, and 5 were evaluated using initial labeling probabilities assigned by the weighting method, estimating the transition probabilities over the chosen block, and setting $d_i=0$ and $\beta=0.25$. The results are given in Figure 3.15. The results show that Algorithm 5 has a better performance than Algorithms 1 and 4. Also Algorithm 5 reaches its fixed point or steady state in few iterations.

Experiment 3.10

A block of size 129 x 91 pixels from Data Set 2 was chosen. The accuracy of the labeling was measured by using 88 pixels whose correct labeling was known. Then the performances of Algorithms 1,3,4, and 6 were compared, estimating initial probabilities by weighting method and estimating the transition probabilities over the whole region. The results are shown in Figure 3.16 and suggest the use of a supervised non-iterative approach for reduction of the labeling ambiguity.

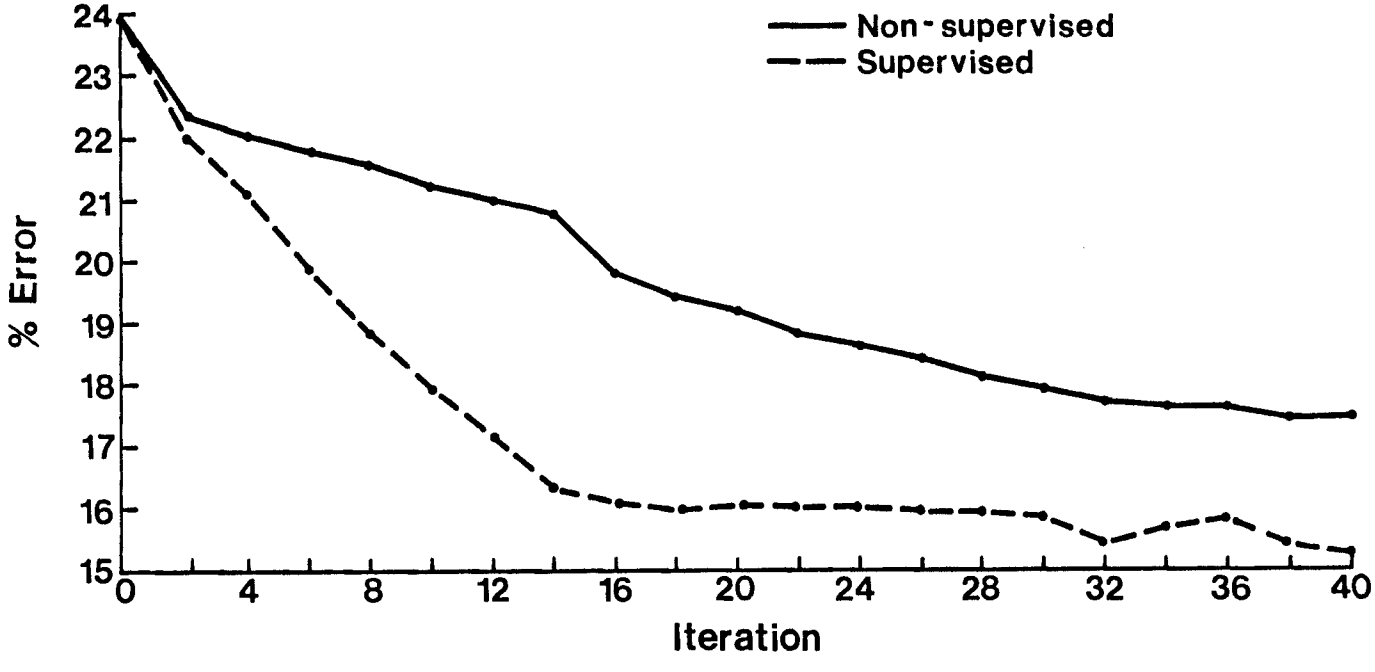

Figure 3.14 Comparison of performance of the supervised and non-supervised relaxation algorithm (Algorithms 1 and 5).

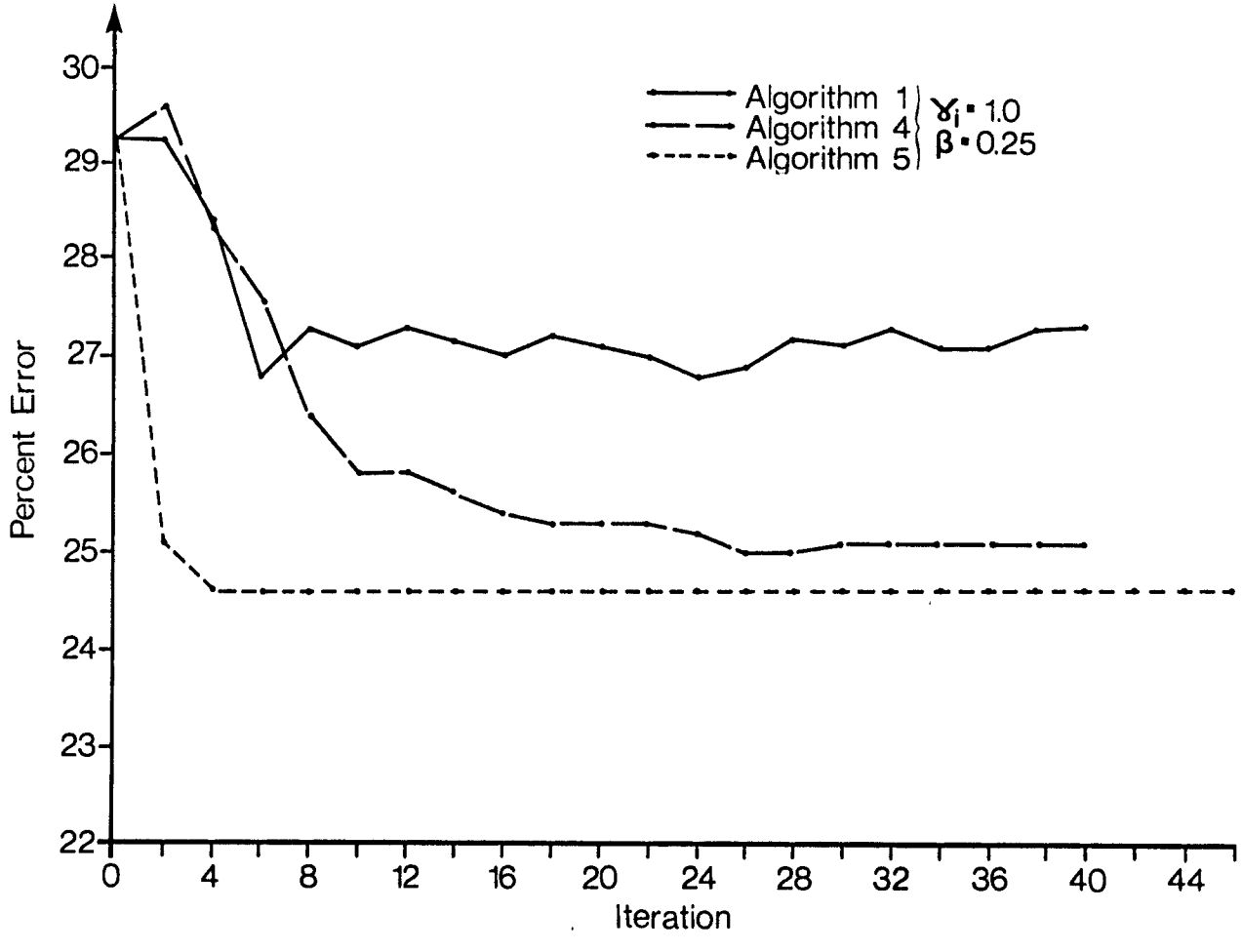

Figure 3.15 Comparison of performance of Algorithms 1, 4 and 5.

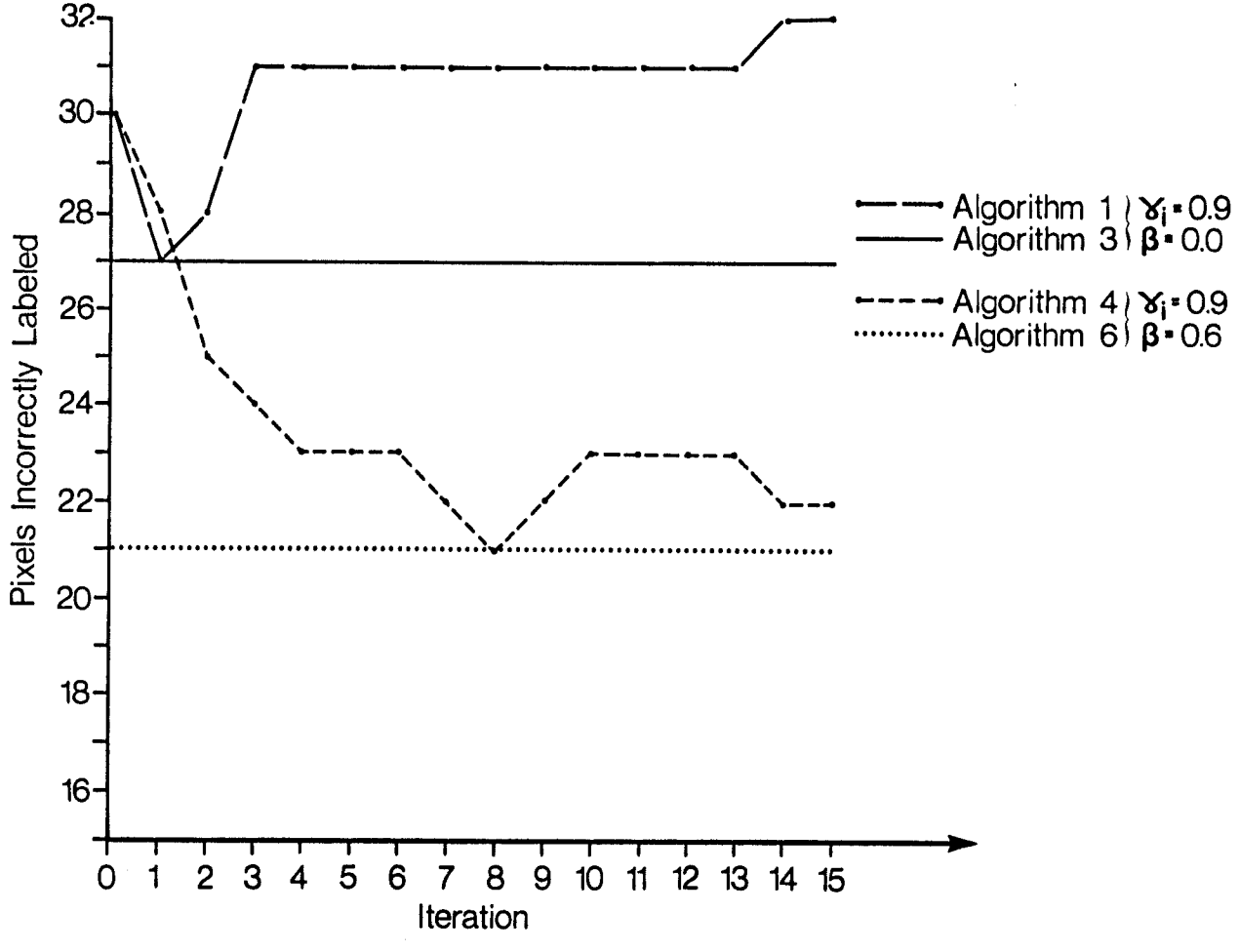

Figure 3.16 Comparison of performance of the Algorithms 1, 3, 4 and 6.

3.5 Conclusion

The probabilistic relaxation technique suggested by Zucker et al. [26] may be applied to the remote sensing data as a post classifier. However, the suggested algorithm usually decreases the labeling error (improving phase), passes through a turning point and increases the labeling error (deterioration phase). We have modified the algorithm by assuming that the transition probabilities are slowly varying over the scene and a method to estimate the transition probabilities has been suggested. The experimental results suggest that the modified algorithm does not exhibit a deterioration phase anymore. Also, a non-iterative adaptive labeling algorithm has been developed which performs as well as the modified probabilistic relaxation algorithm. In addition, in order to be able to preserve the geometric features, i.e., roads, line pixels and isolated pixels, supervised relaxation labeling was developed. By supervising the process by the available ancillary information, we indeed incorporate "memory" into the labeling process to constantly remind the algorithm about some geometric features which are strongly supported by ancillary information. Finally, it has been shown that by utilizing spectral, spatial, and ancillary data, the initial labeling accuracy can be improved.

CHAPTER 4

STOCHASTIC MODEL UTILIZING SPECTRAL AND SPATIAL CHARACTERISTICS

The main objective is to exploit the spatial correlation between the pixels comprising an object by a two-dimensional Markov model and as a result of that develop a new object classifier. First, the minimum distance (MD) and the maximum likelihood (ML) object classifiers are discussed. Then based on a proposed model these two classifiers are modified and a linear object classifier is introduced. Finally, experimental results are presented.

4.1 Object Classifiers

Multispectral image data consist of an observation set $\mathbf{X}$, location set Ω and population set C where:

$$\mathbf{X} = \{\mathbf{x}(s), s \in \Omega\}$$

$$\Omega = \{s = (i,j),\ 1 \leq i \leq I,\ 1 \leq j \leq J\}$$

$$C = \{\omega_1, \omega_2, \ldots \omega_m\}$$

and X(s) is a q-dimensional random observation. Let $\{X(s), s \in \Omega_x\}$ where

$$\Omega_x = \{I_{1x} \leq i \leq I_{2x},\ J_{1x} \leq j \leq J_{2x}\}$$

be the set of observations of an unknown object. Now, the problem is how to classify this object to one of m possible classes.

In remote sensing the set of observations of an object is commonly modeled as:

$$X(s) = M_x + W(s);\ X(s) \varepsilon R^q \tag{4.1}$$

where W(s) is a set of uncorrelated random vectors and coming from a normal population distribution. Let us assume that the object belongs to class ω_x, where $\omega_x \varepsilon C$. Then we can write:

$$E[X(s) | \omega_x] = M_x \tag{4.2}$$

$$E[W(s) | \omega_x] = 0 \tag{4.3}$$

$$E[W(s)\ W^T(t) | \omega_x] = \begin{cases} \Sigma_x & s=t \\ 0 & s \neq t \end{cases} \tag{4.4}$$

Let $p(x(s) | \omega_x)$ denote the class conditional density function for the class ω_x. Based on the above assumption we have

$$p(X(s) | \omega_x) = N(X(s);\ M_x, \Sigma_x)$$

and

$$N(X(s);\ M_x, \Sigma_x) \triangleq \frac{1}{(2\Pi)^{\frac{q}{2}} |\Sigma_x|^{\frac{1}{2}}} \exp\{-\tfrac{1}{2}(X(s)-M_x)^T\ \Sigma_x^{-1}(X(s)-M_x)\} \tag{4.5}$$

In practice, the parameters of classes $\omega_1, \omega_2 \ldots, \omega_m$ the mean vector M_k and covariance matrix Σ_k, (k = 1,2,...m) are estimated from sets of training data supplied for each class.

$$\hat{M}_x = \frac{1}{n_x} \sum_{s\varepsilon\Omega_x} X(s) \tag{4.6}$$

$$\hat{\Sigma}_x = \frac{1}{n_x} \sum_{s\varepsilon\Omega_x} (X(s)-M_x)(X(s)-M_x)^T \tag{4.7}$$

where

$$n_x = (I_{2x}-I_{1x}+1)(J_{2x}-J_{1x}+1) \tag{4.8}$$

The decision rule for a minimum distance object classifier [1] is given by:

$$\text{if} \quad d_{x\ell} = \min_k d_{xk} \rightarrow \{X(s),\ s\varepsilon\Omega_x\} \varepsilon\omega_\ell$$

where

$$d_{xk} \triangleq d[p(X(s)\,|\,\omega_x,\ s\varepsilon\Omega_x),\ p(X(s)\,|\,\omega_k,\ s\varepsilon\Omega_k)] \tag{4.9}$$

where d_{xk} denotes the statistical distance between the probability density functions of class ω_x and class ω_k. Some popular distance measures for two normally density functions are given by:

1. The Bhattacharyya distance

$$B_{xk} = \frac{1}{8}\left(\hat{M}_x - \hat{M}_k\right)^T \left(\frac{\hat{\Sigma}_x + \hat{\Sigma}_k}{2}\right)^{-1} \left(\hat{M}_x - \hat{M}_k\right) + \frac{1}{2}\ell n \left[\frac{|\frac{1}{2}(\hat{\Sigma}_x + \hat{\Sigma}_k)|}{|\hat{\Sigma}_x|^{\frac{1}{2}} |\hat{\Sigma}_k|^{\frac{1}{2}}}\right]$$

2. The Divergence (4.10)

$$D_{xk} = \frac{1}{2}tr\left[(\hat{\Sigma}_x - \hat{\Sigma}_k)(\hat{\Sigma}_x^{-1} - \hat{\Sigma}_k^{-1})\right] + \frac{1}{2}tr\left[(\hat{\Sigma}_x^{-1} + \hat{\Sigma}_k^{-1})(\hat{M}_x - \hat{M}_k)(\hat{M}_x - \hat{M}_k)^T\right] \quad (4.11)$$

3. The Jeffries-Matusita (J-M) distance

$$J_{xk} = \left[2\left(1 - e^{-B_{xk}}\right)\right]^{\frac{1}{2}} \quad (4.12)$$

4. The Transformed Divergence

$$T_{xk} = 2\left[\left(1 - e^{\frac{-D_{xk}}{8}}\right)\right] \quad (4.13)$$

4.2 Maximum Likelihood Object Classifier

The maximum likelihood object classifier also assumes that observations within an object are uncorrelated and normally distributed. Then the decision rule is given by:

$$\text{if} \quad p(\{X(s),\ s\epsilon\Omega_x\}|\omega_\ell) = \max_k p(\{X(s),\ s\epsilon\Omega_x\}|\omega_k), \quad k = 1,2,\ldots,m \quad (4.14)$$

then classify $\{X(s),\ s\varepsilon\Omega_x\}$ into class ω_ℓ. By the above assumption the class conditional density functions can be calculated by

$$p(\{X(s),\ s\varepsilon\Omega_x\}|\omega_\ell) = \prod_{s\varepsilon\Omega_x} p(X(s)|\omega_\ell) \tag{4.15}$$

A block diagram of an object recognition system in remote sensing is given in Figure 4.1.

4.3 Proposed Object Classifiers

It has been observed in [73-76,85] that two pixels in spatial proximity to one another are class unconditionally and class conditionally correlated. The unconditional correlation usually decays slowly with distance but conditional correlation decreases very rapidly. The sources of this spatial correlation can be due to physical properties of the sensor and the target and can also be induced by the atmosphere. Therefore, this spatial correlation introduces redundant data in the object.

Our objective as mentioned earlier is to extract spatial (class conditional) correlation and generate independent observations for each object. Then there will be no redundant information in each object. To do so we are assuming this spatial variation of energy can be modeled by

$$Y(s) = \sum_{(i,j)\varepsilon N} \rho^K_{i,j}\, Y(s+(i,j)) + W(s),\ s\varepsilon\Omega_k \tag{4.16}$$

where

$$Y(s) + X(s) - \hat{M}_x,\ \{X(s),\ s\varepsilon\Omega_k\}\ \varepsilon\omega_k,$$

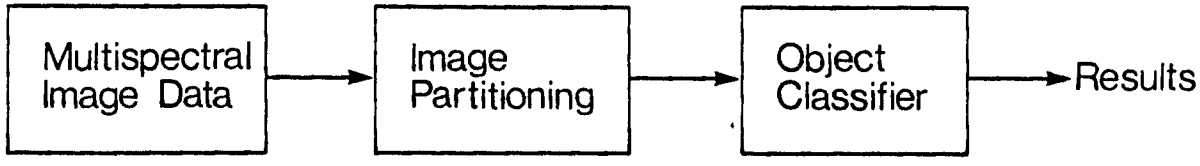

Figure 4.1 Block diagram of an object recognition system.

$$N = \{(0,-1),\ (-1,-1),\ (-1,0)\}$$

$\rho^{K}_{i,j}$ are q x q diagonal matrices and

$$\hat{M}_k = \frac{1}{n_k} \sum_{s \in \Omega_k} X(s)$$

Also, $\{W(s),\ s \in \Omega_k\}$, k = 1,2,...m are Gaussian white noise fields.

$$E[W(s) \mid \omega_k] = 0 \tag{4.17}$$

$$E[W(s)W^{T}(t) \mid \omega_k] = \begin{cases} R_k & s=t \\ 0 & s \neq t \end{cases} \tag{4.18}$$

$$E[Y(s) \mid \{Y(s+(i,j)),\ (i,j) \in N\}; \omega_k] =$$

$$\sum_{(i,j) \in N} \rho^{K}_{(i,j)}\ Y(s+(i,j)) \tag{4.19}$$

$$\mathrm{cov}[Y(s) \mid \{Y(s+(i,j)),\ (i,j) \in N\}; \omega_k] = R_k \tag{4.20}$$

Since W(s)'s are uncorrelated and Gaussian random vectors,

$$p(\{W(s),\ s \in \Omega_k\} \mid \omega_k) = \prod_{s \in \Omega_k} p(W(s) \mid \omega_k) \tag{4.21}$$

By assumption that the observations are Gaussian and come from a first order Markov process the following can be written:

$$p(\{Y(s),\ s\varepsilon\Omega_k\}|\omega_k) =$$

$$\left[\prod_{s\varepsilon\Omega_k} p(Y(s)|\{Y(s+(i,j)),\ (i,j)\varepsilon N\};\omega_k)\right]\cdot$$

$$s\neq(1,j_1),\ (i_1,1);\ j_1=1,2,\ldots J_k;\ i_1=2,3,\ldots I_k$$

$$[p(Y(1,1)\ldots Y(1,J_k),\ Y(2,1),\ldots Y(I_k,1)|\omega_k)] \qquad (4.22)$$

However, practically, it is not possible to estimate the distributions of the pixels on the boundaries but if we do estimate since generally the pixels on the boundaries are mixed, the second term of eq. 4.22 will be almost constant. Hence, in computing the decision rule for classifying an object based on the proposed model, this term may be ignored.

$$p(Y(s)|\{Y(s+(i,j)),\ (i,j)\varepsilon N\};\omega_k) = N(Y(s);\ D_k(s),R_k) \qquad (4.23)$$

where $$D(s) = \sum_{(i,j)\varepsilon N} \rho^k_{i,j} Y(s+(i,j)) \qquad (4.24)$$

It is assumed that ρ_{ij}'s are diagonal matrices. Therefore, the spatial correlation on each channel can be estimated independently of the others. The case when ρ_{ij} is a full matrix is given in Appendix D. When ρ_{ij}'s are diagonal, their estimates are called limited information [78,79] and are given by:

$$\hat{\theta}^K_p = \left[\sum_{s\varepsilon\Omega_k} Z_p(s)\ Z_p^T(s)\right]^{-1} \left[\sum_{s\varepsilon\Omega_k} Y_p(s)\ Z_p(s)\right] \qquad (4.25)$$

where

$$\hat{\theta}^K_p = [\theta^K_p(0,-1),\ \theta^K_p(-1,0),\ \theta^K_p(-1,-1)]^T$$

$$Z_p(s) = [Y_p(s+(0,-1)),\ Y_p(s+(-1,0)),\ Y_p(s+(-1,-1))]^T$$

$$Y(s) = [Y_1(s), Y_2(s), \ldots\ldots Y_q(s)]^T$$

$$p = 1,2,\ldots q \text{ and } k = 1,2,\ldots m$$

The estimate of covariance matrix of W(s) is given by

$$\hat{R}_k = \frac{1}{n_k} \sum_{s \in \Omega_k} (Y(s) - \sum_{(i,j) \in N} \hat{\rho}^K_{i,j} Y(s+(i,j))) \cdot (Y(s) - \sum_{(i,j) \in N} \hat{\rho}^K_{i,j} Y(s+(i,j)))^T \quad (4.26)$$

4.4 Modified Minimum Distance Object Classifier (MMDO)

Let δ^K be a set of parameters for class ω_k. The existing object classifiers characterize each class or object by two parameters, i.e.,

$$\delta^K = \{\hat{M}_k, \hat{\Sigma}_k\}$$

where M_k and Σ_k are the estimate of the mean vector and covariance matrix of class ω_k. But by the proposed model for each class or object we have

$$\delta^K = \{\hat{M}_k, (\hat{\rho}^K_{i,j}, (i,j) \in N), \hat{R}_k\}$$

The proposed model can be thought of as a filter which maps the data into uncorrelated Gaussian random vectors.

The objective here is to modify the MDO classifier. so as to increase its effectiveness. As shown in Figure 4.2a, when classes are not separable, it is very important to improve the performance of the classifier for such cases. The decision rule for the modified minimum distance object classifier is given by:

if $B_{x\ell} = \min_k B_{xk}$

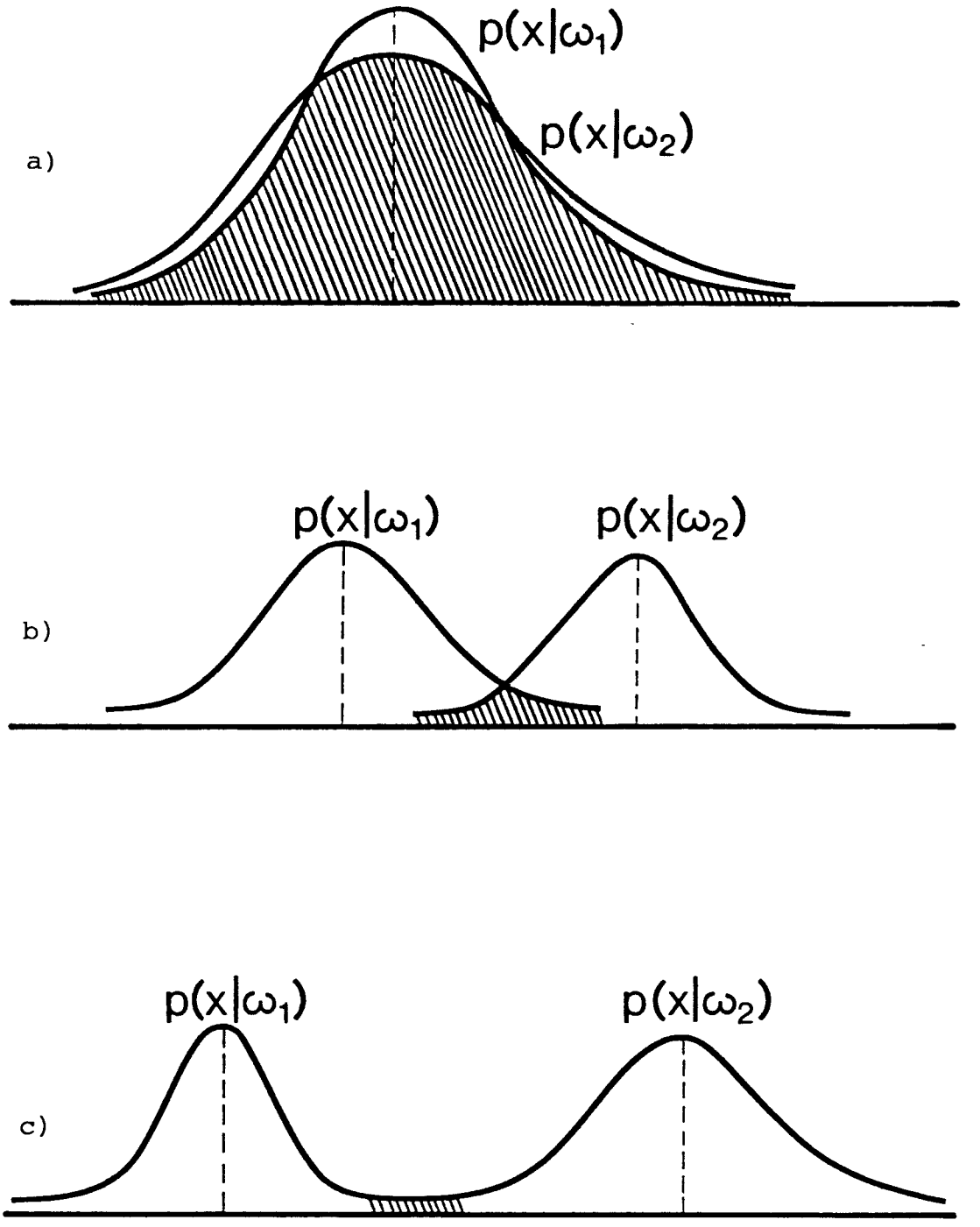

Figure 4.2 Two normal densities with a) low separability, b) medium separability and c) high separability.

$$\rightarrow \{Y(s),\ s\varepsilon\Omega_k\}\varepsilon\omega_\ell \tag{4.27}$$

where

$$B_{x\ell} = \frac{1}{2}\ell n\ \frac{|\frac{1}{2}(\hat{R}_x + \hat{R}_k)|}{|\hat{R}_x|^{\frac{1}{2}} \cdot |\hat{R}_k|^{\frac{1}{2}}}$$

$$\hat{R}_k = \frac{1}{n_k} \sum_{s\varepsilon\Omega_k} W(s)W^T(s)$$

$$W(s) = Y(s) - \sum_{(i,j)\varepsilon N} \hat{\rho}^K_{i,j}\ Y\big(s+(i,j)\big),\ s\varepsilon\Omega_k$$

and

$$Y(s) = X(s) - \hat{M}_k$$

4.5 Modified Maximum Likelihood Object Classifier

As mentioned earlier, the maximum likelihood object classifier assumes that observations from an object are uncorrelated and Gaussian. The decision rule for MLO classifier is given by

$$\text{classify}\quad \{X(s),\ s\varepsilon\Omega_x\}\ \text{in}\ \omega_\ell$$
$$\text{if}\quad \ell n\ p(\{X(s),\ s\varepsilon\Omega_x\}|\omega_\ell) = \max_k\ \ell n\ p(\{X(s),\ s\varepsilon\Omega_x\}|\omega_k) \tag{4.28}$$

where

$$\ell n\ p(\{X(s),\ s\varepsilon\Omega_x\}|\omega_x) = -\frac{n_x}{2}[(q\ \ell n 2\Pi + \ell n|\hat{\Sigma}_k|) + tr(\hat{\Sigma}_k^{-1}\ \hat{Q}_k)] \tag{4.29}$$

$$\hat{Q}_k = \frac{1}{n_x} \sum_{s\varepsilon\Omega_x} (X(s)-\hat{M}_k)(X(s)-\hat{M}_k)^T \tag{4.30}$$

But our assumption is that observations in spatial proximity to one another are class conditionally correlated. Based on the proposed model, the decision rule for the modified maximum likelihood object classifier is given by

$$\text{classify} \quad \{W(s),\ s\varepsilon\Omega_k\} \text{ in } \omega_\ell$$
$$\text{if} \quad \ell n\ p(\{W(s),\ s\varepsilon\Omega_x\}|\omega_\ell) = \max_k \ell n\ p(\{W(s),\ s\varepsilon\Omega_x\}|\omega_k) \tag{4.31}$$

Since we are assuming $\{W(s),\ s\varepsilon\Omega_k\}$ are identically, independently, and normally distributed and from equation 4.16 the Jacobian of the transformation is unity; therefore

$$\ell n\ p(\{W(s),\ s\varepsilon\Omega_x\}|\omega_\ell) = \ell n\ p(\{Y(s),\ s\varepsilon\Omega_x\}|\omega_\ell)$$
$$= \ell n\ p(\{X(s),\ s\varepsilon\Omega_x\}|\omega_\ell) \tag{4.32}$$

$$\ell n\ p\ (W(s),\ s\varepsilon\Omega_x|\omega_k) = -\frac{n_x}{2}[(q\ \ell n 2\Pi + \ell n|\hat{R}_k|) + tr(\hat{R}_k^{-1}\ \hat{Q}_k)] \tag{4.33}$$

$$\hat{Q}_k = \frac{1}{n_x} \sum_{s\varepsilon\Omega_x} W(s)W^T(s) \tag{4.34}$$

$$W(s) = Y(s) - \sum_{(i,j)\varepsilon N} \hat{\rho}^k_{i,j}\ Y(s+(i,j))$$

and $Y(s) = X(s) - \hat{M}_k$

Equation (4.33) clearly shows the dependency of the decision rule on the spatial correlations $\hat{\rho}^k_{i,j}$. If $\hat{\rho}^k_{i,j}$ are different for inseparable classes, then one expects to see the probability of error in MMLO to be less than the MLO classifier.

4.6 Linear Minimum Distance Object Classifier (LMDO)

When the classes are separable, data is not complex or very limited numbers of training samples are available, we do not need a sophisticated classifier such as MMDO or MMLO. A simple linear minimum distance object classifier (LMDO) can do as well or sometimes even better, because in LMDO classifier, we need only to estimate the mean vector, but in MMDO or MMLO, we must also estimate $\{\rho_{ij},\ (i,j)\in N\}$ and R_k. If a limited number of training samples are available, estimates of R_k and $\{\rho_{ij},\ (i,j)\in N\}$ may be poor. As a result of this, the performance of the MMDO or MMLO may be deteriorated (see Appendices E and F). The decision rule for the LMDO classifier is given by

$$\text{if} \quad \text{tr}\ \hat{Q}_\ell = \min_k \text{tr}\ \hat{Q}_k \tag{4.35}$$

then classify $\{X(s),\ s\in\Omega_x\}$ into class ω_ℓ where

$$\hat{Q}_k = \frac{1}{n_x} \sum_{s\in\Omega_x} (X(s)-\hat{M}_k)(X(s)-\hat{M}_k)^T \tag{4.36}$$

Then

$$\begin{aligned}
\text{tr}\ \hat{Q}_k &= \frac{1}{n_x} \sum_{s\in\Omega_x} (X(s)-\hat{M}_k)^T(X(s)-\hat{M}_k) \\
&= \frac{1}{n_x} \sum_{s\in\Omega_x} [X^T(s)X(s)-X^T(s)\hat{M}_k-\hat{M}_kX(s)-\hat{M}_k^T\hat{M}_k] \\
&= \frac{1}{n_x} \sum_{s\in\Omega_x} X^T(s)X(s) + [-\hat{M}_x^T\hat{M}_k - \hat{M}_k^T\hat{M}_x - \hat{M}_k^T\hat{M}_k]
\end{aligned}$$

where

$$\hat{M}_x = \frac{1}{n_x} \sum_{s\in\Omega_x} X(s)$$

Since

$$\frac{1}{n_x} \sum_{s \in \Omega_x} X^T(s)\ X(s)$$

is common to all classes; therefore, the equivalent discriminant function is

$$-\hat{M}_x^T \hat{M}_k - \hat{M}_k^T \hat{M}_x - \hat{M}_k^T \hat{M}_k \tag{4.37}$$

or equivalently

$$\begin{aligned} L_{xk} &= \hat{M}_x^T \hat{M}_x - \hat{M}_x^T \hat{M}_k - \hat{M}_k^T \hat{M}_x - \hat{M}_k^T \hat{M}_k \\ &= (\hat{M}_x - \hat{M}_k)^T\ (\hat{M}_x - \hat{M}_k) \end{aligned} \tag{4.38}$$

If $L_{x\ell} = \min_k L_{xk}$, then classify the object into class ω_ℓ.

4.7 Experimental Results

Spatially registered multitemporal Landsat multispectral scanner (MSS) data acquired over Henry County, Indiana in 1978 and the aircraft data set of the 1971 Corn Blight Watch flightline 210 were selected to evaluate the performance of the maximum likelihood pixel and object classifier, modified maximum likelihood object classifier, minimum distance object classifier, modified minimum distance object classifier and linear distance object classifer. The acquisition dates for the Landsat MMS data are: June 9, July 16, August 20 and September 26, 1978. The classes corn and soybean were chosen for analysis.

These two data sets were chosen for analysis for the following reasons:

1) Wall-to-wall ground truth is available. This is important both for deriving good quality training samples and for accurate determination of performance.
2) The ground spatial resolution of the aircraft data set is much finer than the Landsat MSS data set. It is important to see how this effects the spatial class conditional correlations.
3) The performance of the proposed classifiers with Landsat MSS (4 channels, low ground resolution and and 6 bit data representation) and aircraft (12 channels, high ground resolution and 8 bit data representation) data sets under different class separabilities could be evaluated.

4.7.1 Training Methods

Histogramming and clustering are two commonly used training methods which could be used to find rectangular shaped objects with approximately Gaussian observations from training fields. We used the histogramming method to define the spectral classes from training fields or objects. The training objects were chosen to be representative of the informational classes. Then based on the proposed two-dimensional Markov process only horizontal and vertical correlations were extracted. Information about the software system and data sets are given in Appendix G.

Experiment 4.1

The test objects of Landsat multispectral scanner data collected on June 9 were classified by six classifiers. The results are given in Figure 4.3.

Usually in early June corn and soybeans are very much like each other. This can be seen from the results by the maximum likelihood pixel and object classifiers or the minimum distance object classifier. However, the two modified object classifiers improved the overall accuracy by about 10%. And this is significant when the separability is not class means dependent.

Experiment 4.2

The objective of this experiment was to show that if the classes are moderately separable, then a linear distance object classifier may do as well as a non-linear object classifier. Therefore, the Landsat MSS data set collected on July 16 in which corn and soybeans are usually separable was analyzed by five different classifiers. The performance of the classifiers by class is given in Figure 4.4. The results show that the overall performance of the MMLO is better than the other classifiers which have about the same performance; however, the linear distance object classifier is much faster than the others.

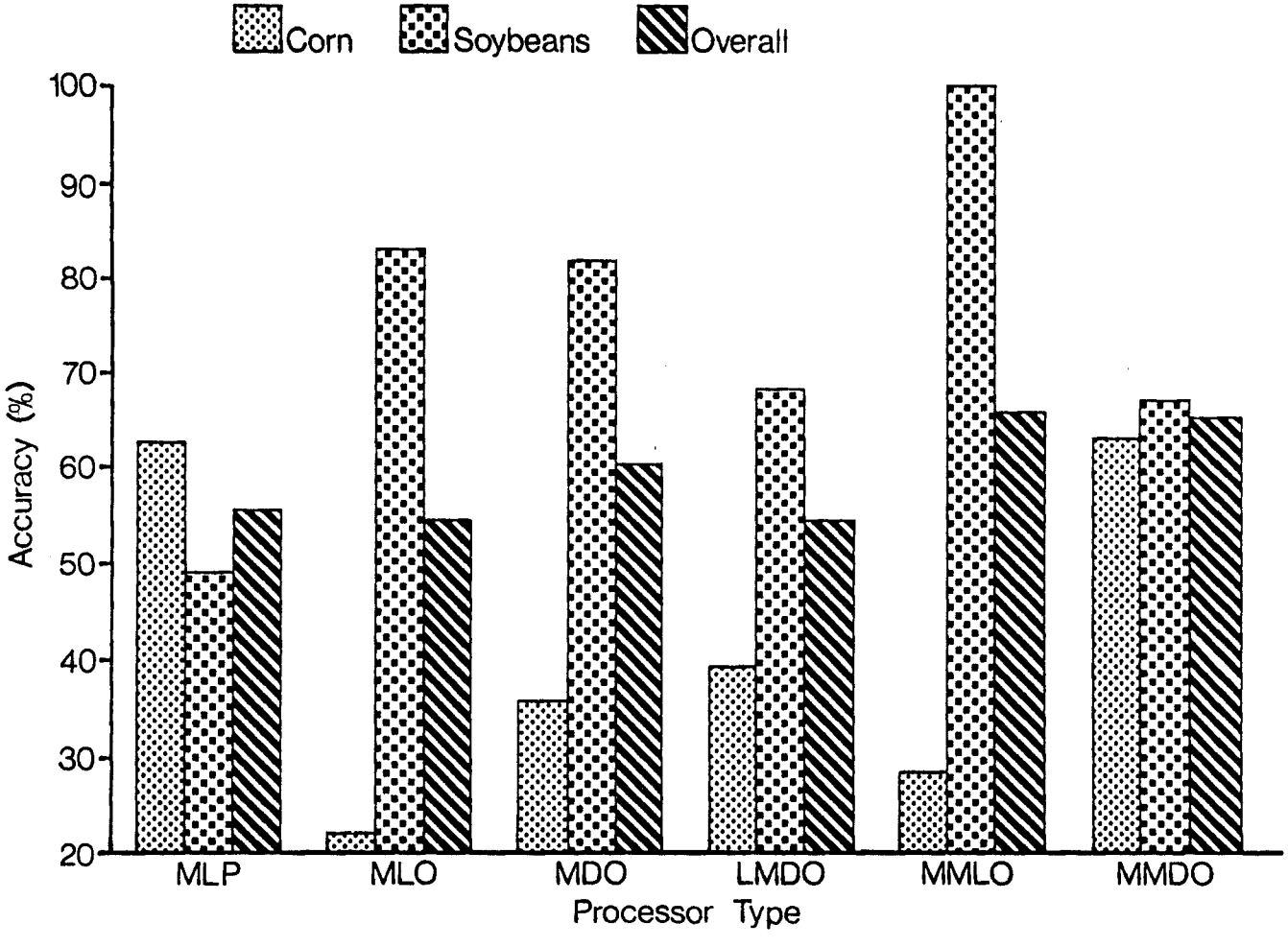

Figure 4.3 Overall classification performance vs. processing scheme (Henry County data; June 9).

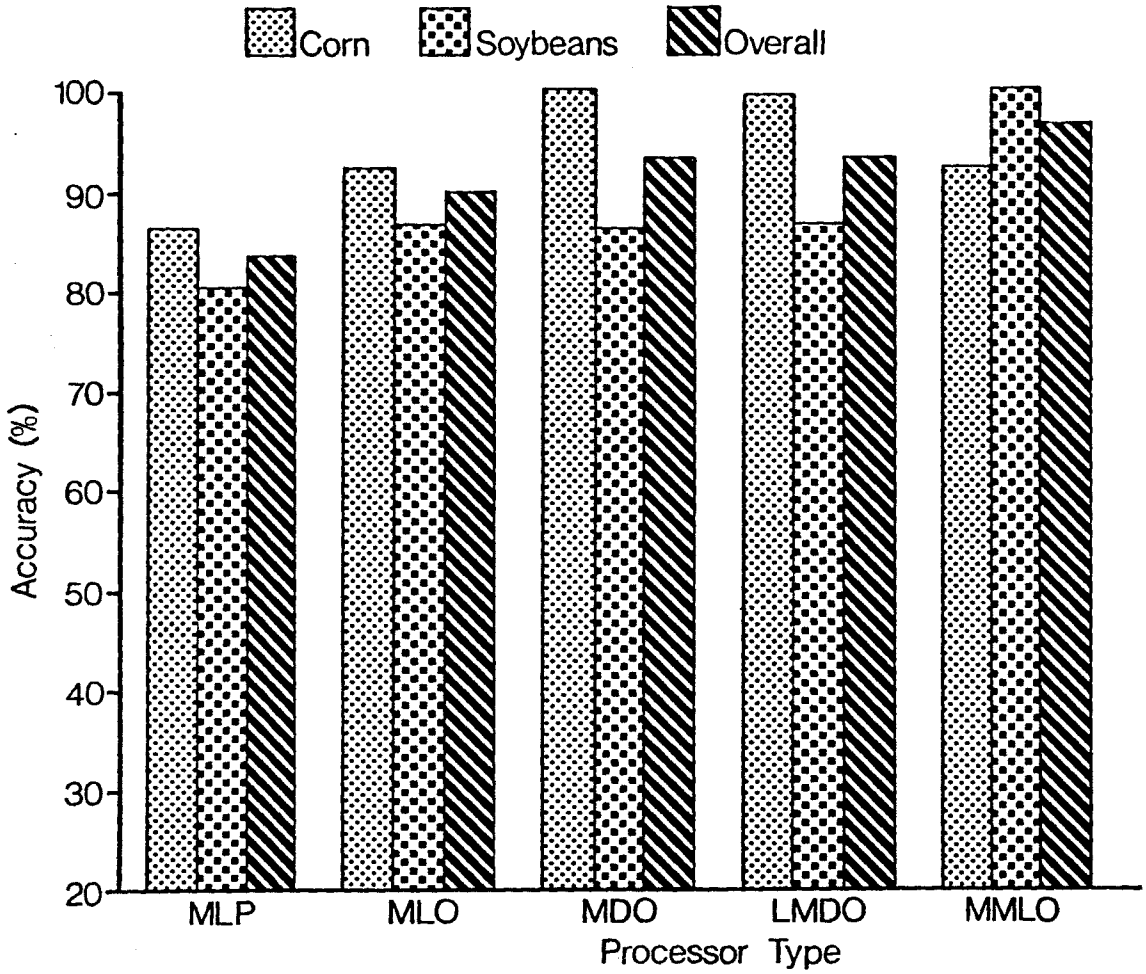

Figure 4.4 Overall classification performance vs. processing scheme (Henry County data; July 16).

Experiment 4.3

Here we just wanted to show another example that if the classes are separable, then a linear distance object classifier may do as well as a non-linear object classifier. The Landsat MSS data collected on August 20 were classified with five classifiers. The results are given in Figure 4.5.

Experiment 4.4

Two classes of wheat and hay were selected from aircraft data of 1971 flightline 210 from the Corn Blight Watch Experiment. The performance of ML pixel classifier, MDO and MMDO classifiers are given in Table 4.1. The results show that when the classes are not very separable, the MMDO classifier which utilizes the textural information together with spectral characteristics has a better performance than the existing object classifiers.

4.8 Conclusion

Based on the assumption that pixels in spatial proximity to one another are conditionally correlated, the two-dimensional stochastic Markov process was proposed to extract this spatial correlation. Then as a result of the model, the maximum likelihood and minimum distance object classifiers were modified and also a linear distance object classifier was introduced. Spatially registered

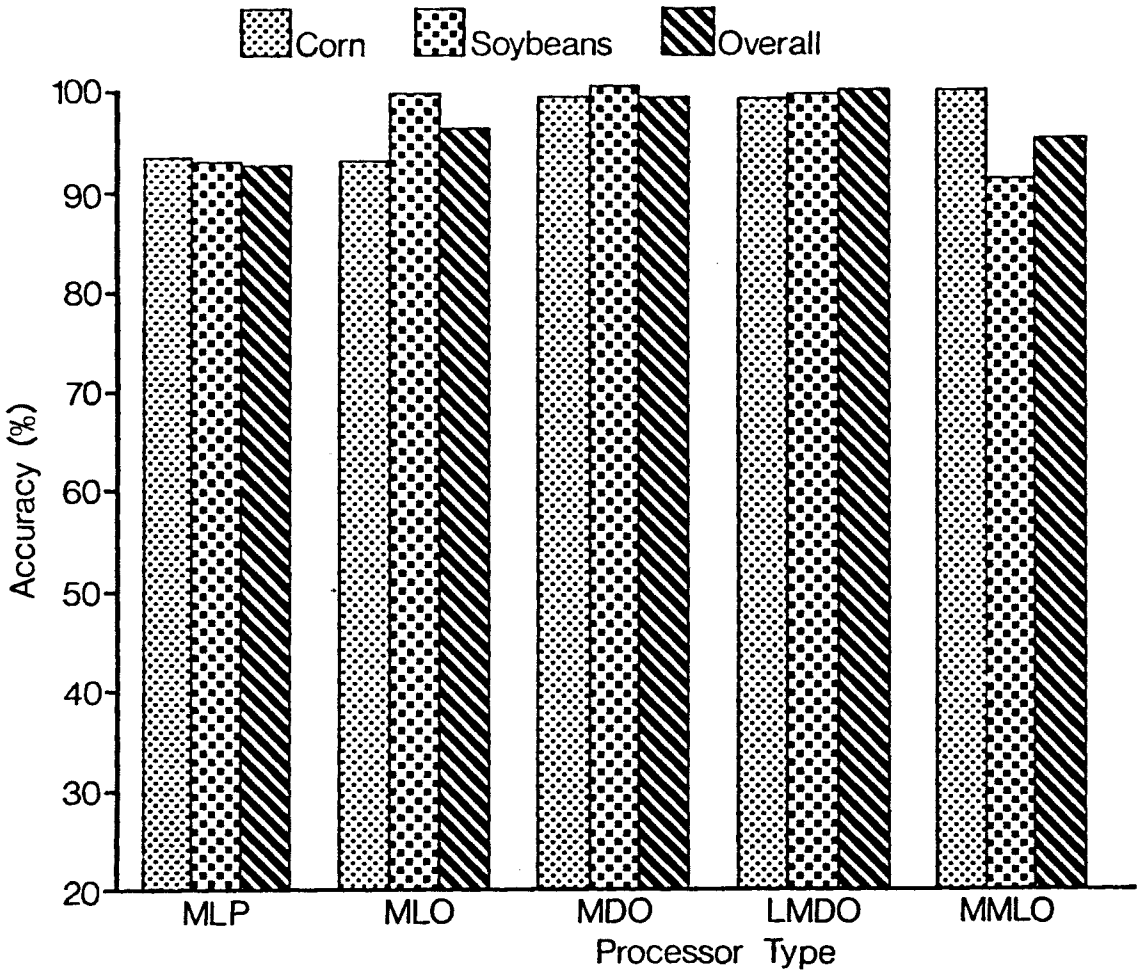

Figure 4.5 Overall classification performance vs. processing scheme (Henry County data; August 20).

Table 4.1 Classification performance by class for different classifier (aircraft data).

Maximum Likelihood Pixel Classifier

Group	No. of Samples	Percent Correct	No. of Samples Classified Into WHEAT	HAY
1 WHEAT	734	98.5	723	11
2 HAY	862	44.0	483	379
TOTAL	1596	69.0	1206	390

Minimum Distance Object Classifier

Group	No. of Fields	% Field Correct	No. of Samp.	% Sam. Correct	No. of Samples Classified Into WHEAT	HAY
1 WHEAT	8	100.0	734	100.0	8	0
2 HAY	7	57.1	862	46.4	3	4
TOTAL	15	80.0	1596	71.1	11	4

Modified Minimum Distance Object Classifier

Group	No. of Fields	% Field Correct	No. of Samp.	% Sam. Correct	No. of Samples Classified Into WHEAT	HAY
1 WHEAT	8	87.5	734	94.6	7	1
2 HAY	7	71.4	862	73.2	2	5
TOTAL	15	80.0	1596	83.0	9	6

multitemporal Landsat MSS (low complexity data) and aircraft (higher complexity data) data sets for classes with different separabilities were analyzed. The results suggest that when classes are not very separable, the modified minimum distance object classifier has significantly better performance than existing object classifiers. Also, when the classes are moderately separable, the linear distance object classifier does as well as the others.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 Summary

The purpose of this research was to develop analytical techniques for incorporating spectral, spatial, temporal, and ancillary data characteristics into the classification process. In Chapter 2, based on the assumption that temporal observations are from a Gauss-Markov process, a new processor, called the Markov pixel classifier, was developed. The results of experiments show that this classifier has better performance than the maximum likelihood and cascade classifiers.

In Chapter 3, probabilistic and supervised relaxation labeling (PRL) techniques were adapted for utilizing multitype data. The PRL algorithm suggested by Zucker et al. [26] was modified and an algorithm called non-iterative adaptive labeling was developed. Also, in order to preserve the isolated, line, and corner pixels and narrow geometric features such as road of the scene and to incorporate "memory" into the probabilistic relaxation process, we supervised the relaxation process with the classification results of the scene at different times by a linear classifier or by using ancillary data. The experiment results

suggest that the performance of the non-iterative adaptive labeling (NAL) is very close or even sometimes better than the iterative PRL algorithm. Also, the results suggest that by supervising the relaxation process or the NAL algorithm, the accuracy of classification can significantly be improved.

In Chapter 4, based on the assumption that pixels in spatial proximity to one another are conditionally correlated, a two-dimensional stochastic Markov process was developed to utilize this correlation and to generate another two-dimensional uncorrelated Gaussian process. As a result of adapting this model to the problem, the minimum distance and maximum likelihood object classifiers were modified. Also, a linear distance object classifier was developed. The experiment results suggest that if the separability between classes is dependent only on covariance matrices, the modified minimum distance object classifier significantly improves classification accuracy over the MDO and MLO classifiers. The results also suggest if there is moderate separability between classes, the linear distance object classifier does as well as the others.

5.2 Recommendations for Further Work

The objective was to incorporate the temporal class conditional correlation, the labeling correlation and spatial class conditional correlation into the classification process. However, to be able to say whether these sources

of information are consistently useful or not, more investigation needs to be performed. For example, one would expect that the temporal correlation is very useful in the accurate estimation of the development stage of a given crop, based on the spectral/temporal observations. In Chapter 2, the temporal variations of energy of the crop was modeled by a stochastic Markov process. It should be possible to adapt this model for predicting the development stage.

In Chapter 4, we developed a stochastic model to utilize class conditional interpixel correlation. It is important to find out for what classes it brings useful information into the classification process and also how the proposed model can be adapted into the image partitioning process. Also in Chapter 4, based on separability between classes, we suggested the type of object classifier that should be used. But it is important to generalize this and, based on some criteria, predict the type of processor that should be used; for example, in a tree classifier at each node. In addition, an iterative contextual classifier for further study is given in Appendix B.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. P.H. Swain and S.M. Davis, eds. Remote Sensing: The Quantitative Approach. McGraw-Hill International Book Co., New York, 1978.

2. K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, New York, 1972.

3. R.O. Duda and P.E. Hart. Pattern Classification and Scene Analysis. Wiley, New York, 1973.

4. N.J. Nilsson. Learning Machines. McGraw-Hill Book Co., Inc., 1965.

5. J.M. Mendel and K.S. Fu. Adaptive, Learning and Pattern Recognition Systems. Academic Press, New York and London, 1970.

6. M. Shimura. "Learning Procedures in Pattern Classifier," Pattern Recognition. pp. 125-138, 1978.

7. J. Raviv. "Decision Making in Markov Chains Applied to the Problem of Pattern Recognition, IEEE Transactions on Information Theory, Vol. IT-3, No. 4, October, 1976.

8. G. David Forney. "The Viterbi Algorithm." Proceedings of the IEEE, Vol. 61, No. 3, March, 1978.

9. K. Abend, T.J. Harley and L.N. Kanal. "Classification of Binary Random Patterns." IEEE Transactions on Information Theory," Vol. IT-11, No. 4, October, 1965.

10. L.N. Kanal, ed. Proceedings of the IEEE Workshop on Pattern Recognition. Thompson Book Company, Washington, D.C., 1968.

11. J.R. Welch and K.G. Salter. "A Context Algorithm for Pattern Recognition and Image Interpretation," IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-1, pp. 24-30, January, 1971.

12. T.S. Yu and K.S. Fu. "Statistical Pattern Recognition Using Contextual Information," Technical Report TR-EE 78-17, School of Electrical Engineering, Purdue University, West Lafayette, IN 47907, March, 1978.

13. G.T. Toussaint. "The Use of Context in Pattern Recognition," Pattern Recognition. Vol. 10, pp. 189-204. Pergamon Press Ltd., 1978.

14. E.F. Kit and P.H. Swain. "An Approach to the Use of Statistical Context in Remote Sensing Data Analysis," 5th Canadian Symposium on Remote Sensing, Victoria, August, 1978.

15. P.H. Swain, H.J. Siegel and B.W. Smith. "Contextual Classification of Multispectral Remote Sensing Data Using a Multiprocessor System," IEEE Trans. Geosci. Remote Sensing, Vol. GE-18, April, 1980.

16. P.H. Swain, S.B. Vardeman, and J.C. Tilton. "Contextual Classification of Multispectral Image Data," Pattern Recognition, Vol. 6, March 1981.

17. A. Rosenfeld, R. Hummel and S. Zucker. "Scene Labeling by Relaxation Algorithms," IEEE Trans. Sys. Man, Cyber. Vol. SMC-6, pp. 420-433, 1976.

18. S. Zucker, E. Krishnamurthy and R. Haar. "Relaxation Processes for Scene Labeling: Convergence, Speed and Stability," IEEE Trans. Sys. Man, Cyber. Vol. SMC-8, No. 1, pp. 41-48, 1978.

19. S. Zucker, E. Krishnamurthy and A. Rosenfeld. "Application of Relaxation Labeling to Line and Curve Enhancement," IEEE Trans. Comp. Vol. C-26. pp. 394-403, plus correction, pp. 900-929, 1977.

20. B. Schachter, A. Lev, S. Zucker and A. Rosenfeld. "An Application of Relaxation to Edge Reinforcement," IEEE Trans. Sys. Man, Cyber. Vol. SMC-7, No. 11, pp. 813-816, 1977.

21. S. Peleg and A. Rosenfeld. "Determining Compatability Coefficients for Curve Enhancement Relaxation Processes," IEEE Trans. Sys. Man, Cyber. Vol. SMC-8, No. 7, pp. 548-555, 1978.

22. A. Lev, S. Zucker and A. Rosenfeld. "Iterative Enhancement of Noisy Images," IEEE Trans. Sys. Man, Cyber. Vol. SMC-7, No. 6, pp. 435-442, 1977.

23. J. Eklundh, H. Yamamoto and A. Rosenfeld. "Relaxation Methods in Multispectral Pixel Classification," Technical Report 662. Computer Science Center, University of Maryland, College Park, MD, July 1978.

24. J. Eklundh and A. Rosenfeld. "Convergence Properties of Relaxation," Technical Report 701. Computer Science Center, University of Maryland, College Park, MD, Oct. 1978.

25. S. Peleg. "Monitoring Relaxation Algorithms Using Labeling Evaluation," Technical Report 842. Computer Science Center, University of Maryland, College Park, MD, 1979.

26. S. Zucker and J. Mohammed. "Analysis of Probabilistic Relaxation Labeling Processes," Proc. IEEE Conf. Pattern Recognition and Image Processing, Chicago, IL. pp. 307-312, 1978.

27. S. Peleg, "A New Probabilisitic Relaxation Scheme," Proc. IEEE Conf. Pattern Recognition and Image Processing, Chicago, IL, pp. 337-343, 1979.

28. S. Peleg, " A New Probabilistic Relaxation Scheme," IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. PAMI-2, No. 4, 1980.

29. R. Hummel and S. Zucker. "On the Foundation of Relaxation Labeling Processes," Proc. IEEE Conf. Pattern Recognition and Image Processing, Vol. 1, pp. 50-53, December, 1980.

30. S. Peleg. "Monitoring Relaxation Algorithms Using Labeling Evaluations," Proc. IEEE Conf. Pattern Recognition and Image Processing, Vol. 1, pp. 54-57, December 1980.

31. C.R. Mclean and C.R. Dyer. "Analog Relaxation Processor," Proc. IEEE Conf. Pattern Recognition and Image Processing, Vol. 1, pp. 58-60, December 1-4, 1980.

32. O. Faugeras and M. Berthod. "Scene Labeling: An Optimization Approach," Pattern Recognition. Vol. 12, pp. 339-347. Pergamon Press Ltd. Printed in Great Britain, 1979.

33. S. Zucker, Y. Leclerc and J.L. Mohammed. "Continuous Relaxation Local Maximum Selection: Conditions for Equivalence," IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. PAMI-3, pp. 117-127, March, 1981.

34. J.A. Richards, D.A. Landgrebe and P.H. Swain. "On the Accuracy of Pixel Relaxation Labeling," IEEE Trans. Sys. Man and Cyber. SMC-11(4). Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana. Technical Report 030180.

35. O.D. Faugeras and K.E. Price, "Semantic Description of Aerial Images Using Stochastic Labeling," Proc. IEEE Conf. Pattern Recognition and Image Processing, Vol. 1, pp. 352-357, December 1980.

36. R.M. Haralick, K. Shanmugam, and Z. Dinstein. "Textural Features Measures for Terrain Classification," IEEE Trans. Sys., Man, Cyber., Vol. SMC-3, November 1973.

37. J. Weska, C. Dyer and A. Rosenfeld. "A Comparative Study of Texture Measures for Terrain Classification," IEEE Trans. Sys., Man, Cyber., Vol. SMC-6, April 1976.

38. D.J. Wiersma and D.A. Landgrebe. "The Use of Spatial Characteristics for the Improvement of Multispectral Classification of Remotely Sensed Data," in Proc. 1976 Symp. Machine Processing of Remotely Sensed Data (West Lafayette, IN). June 29-July 1, 1976.

39. B.H. McCormick and S.N. Jayamamurthy. "Time Series Model for Texture Synthesis," Intl. J. of Computer and Inf. Sciences, Vol. 3, pp. 329-343, 1947.

40. E.J. Delp et al. "Image with a Seasonal Autoregressive Time Series with Application to Data Compression," Proc. Conf. on Pattern Recognition and Image Processing, 1978.

41. P. Whittle. "On Stationary Processes in the Palne," Biometrika, Vol. 41, pp. 434-449, 1954.

42. R. Chellappa, "Spatial Autoregressions in Digital Image Restoration: Simulataneous," TR-984, AFOSR-77-3271 University of Maryland, College Park, MD, December, 1980.

43. R.L. Kashyap, R. Chellappa and N. Ahuga, "Decision Rules for Choice of Neighbors in Random Field Model of Images," Computer Graphics and Image Processing, Vol. 15, pp. 301-318, 1981.

44. S.L. Sclove. "Pattern Recognition in Image Processing Using Interpixel Correlation," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-3, March, 1981.

45. T. Katayama and H. Tsugi, "Restoration of Noisy Images by Using a Two-Dimensional Linear Model," Proc. IEEE, 4th Intl. Joint Conf. on Pattern Recognition, 1978.

46. N.E. Nahi. "Role of Recursive Estimation in Statistical Image Enhancement," Proc. IEEE, Vol. 60, pp. 872-877, July 1972.

47. N.E. Nahi and T. Assefi. "Bayesian Recursive Image Estimation," IEEE Trans. Comput., Vol. C-21, pp. 734-738, July 1972.

48. S.R. Powell and L.M. Silverman. "Modeling of Two-Dimensional Covariance Function with Applications to Image Restoration," IEEE Trans. Automat. Contr., Vol. AC-19, pp. 8-13, February 1974.

49. F.C. Shoute, M.F. Ter Horst and J.C. Williams. "Hierarchic Recursive Image Enhancement," IEEE Trans. Circuits Syst., Vol. CAS-24, pp. 67-78, February 1977.

50. J.W. Woods and CH. Radewan. "Kalman Filetering in Two Dimensions," IEEE Trans. Inform. Theory, Vol. IT-23, pp. 473-482, July 1977.

51. A. Habibi. "Two dimensional Bayesian Estimate of Images," Proc. IEEE, Vol. 60, pp. 878-883, July 1972.

52. C.H. Chen. "Adaptive Image Filtering." Proc. 1979 IEEE Conf. Pattern Recognition and Image Processing, pp. 32-37, August 6-8, 1979.

53. R.M. Haralick. "Statistical and Structural Approaches to Texture," Proc. IEEE, Vol. 67, May 1979.

54. R.L. Kettig and D.A. Landgrebe. "Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects," IEEE Trans. Geosci. Electron., Vol. GE-4, January 1976.

55. D.A. Landgrebe. "The Development of a Spectral/Spatial Classifier for Earth Observational Data," Proc. IEEE Computer Conf. Pattern Recognition and Image Processing (Chicago, IL), May 31-June 2 (1978).

56. D.A. Landgrebe. "The Development of a Spectral/Spatial Classifier for Earth Observational Data," J. Pattern Recogn. Soc., Vol. 12, pp. 16.-175, 1980.

57. R.M. Haralick and K.S. Shanmugam, "Combined Spectral and Spatial Processing of ERTS Imagery Data," Remote Sensing of Environment, Vol. 3, pp. 3-13, 1974.

58. H. Maurer, "Texture Analysis with Fourier Series." Proc. 9th Intl. Symp. on Remote Sensing of Environment, Vol. II, pp. 1411-1420, 1974.

59. J.H. Herzoy and B. Strum, "Preprocessing Algorithms for the Use of Radiometric Corrections and Texture/Spatial Features in Automatic Land Use Classification," Proc. 10th Intl. Symp. on Remote Sensing of Environment, Vol. II, pp. 705-724, 1975.

60. U. Wieczovek, "Textural Analysis by Statistical Parameters and Application to the Mapping of Flour-Structures in Wetlands," Proc. 11th Intl. Symp. on Remote Sensing of Environment, Vol. II, pp. 1035-1043, 1977.

61. Iisaka, E. Nakata, Y. Ishu, M. Imanaka and Y. Myazaki, "Application of Texture Analysis Image Enhancement Techniques for Remote Sensing," Proc. 12th Intl. Symp. on Remote Sensing of Environment, Vol. III, pp. 1957-1971, 1978.

62. Lt. Col. R.K. Aggarmala, "Role of Texture in Computer Aided Signature Analysis a Case Study," Proc. 13th Intl. Symp. on Remote Sensing of Environment, Vol. III, pp. 1507-1519, 1979.

63. Shin-Yi Hsu, "Texture-Tone Analysis for Automated Land-Use Mapping," Photogrammetric Engineering and Remote Sensing, Vol. 44, No. 11, pp. 1393-1404, November 1978.

64. J.A. Richards, D.A. Landgrebe and P.H. Swain. "Pixel Labeling by Supervised Probabilistic Relaxation," IEEE Trans. Pattern Analysis and Machine Intelligence. Vol. PAMI-3, No. 2. March 1981. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana. LARS Technical Report 022580.

65. D.A. Landgrebe. "Advancements in Large-Scale Data Processing Systems for Remote Sensing," in Proc. 4th Annual Earth Resources Program Rev. (NASA Manned Spacecraft Center, Houston, TX). January 1972.

66. M.D. Fleming and R.M. Hoffer. "Computer-Aided Analysis Techniques for an Operational System to Map Forest Lands Utilizing Landsat MSS Data," M.S. Thesis, Purdue University, West Lafayette, IN. December 1977.

67. P.H. Swain. "Bayesian Classification in a Time-Varying Environment," IEEE Trans. Syst., Man, Cybern., Vol. SMC-8, December 1978.

68. R.A. Abotteen. "Principal Component Greenness Transformation in Multitemporal Agricultural Landsat Data," Proc. 12th Intl. Symp. on Remote Sensing of Environment, Vol. II, pp. 765-774, 1978.

69. E.P. Crist and W.A. Malila. "A Temporal-Spectral Analysis Technique for Vegetation Application of Landsat." Proc. 14th Intl. Symp. on Remote Sensing of Environment, Vol. II, pp. 1031-1081, 1980.

70. G.F. Byrne and P.F. Cropper. "Land Cover Change Detection by Principal Component Analysis of Multitemporal MSS Data: The Presence of Clouds." Proc. 14th Intl. Symp. on Remote Sensing of Environment, Vol. III, pp. 1375-1755, 1980.

71. J.D. Tubbs, "Classification of Landsat Multitemporal Training Data Based Upon Growth Curve Analysis." Proc. 14th Intl. Symp. on Remote Sensing of Environment. Vol. II, pp. 1123-1127, 1980.

72. D.A. Landgrebe. "Analysis Technology for Land Remote Sensing," Proc. IEEE, Vol. 69, No. 5, May (1981).

73. R.G. Craig and M.L. Labovitz, "Sources of Variation in Landsat Autocorrelation," Proc. 14th Intl. Symp. on Remote Sensing of Environment, Vol. III, pp. 1755-1767, 1980.

74. J.D. Tubbs and W.A. Coberly. "Spacial Correlation and Its Effect upon Classification." Proc. 12th Intl. Symp. on Remote Sensing of Environment, Vol. II. pp. 775-781, 1978.

75. J.D. Tubbs, "Classification Results Using Spacially Correlated Landsat Data," Proc. 13th Intl. Symp. on Remote Sensing of Environment, Vol. III, pp. 1499-1505, 1979.

76. R.G. Craig, "Autocorrelation in Landsat Data," Proc. 13th Intl. Symp. on Remote Sensing of Environment, Vol. III, pp. 1517-1524, 1979.

77. G.E.P. Box and G.M. Jenkins, Time Series Analysis Forecasting and Control, San Francisco, California, Holden-Bay, 1970.

78. R.L. Kashyap and A.R. Rao, Dynamic Stochastic Models from Empirical Data, Academic Press, New York, 1976.

79. R.L. Kashyap and R.E. Nasburg, "Parameter Estimation in Multivariate Stochastic Difference Equations," IEEE Trans. Automatic Control, Vol. AC-19, No. 6, December 1974.

80. H. Akashi, H. Imai and K.A.F. Moustafa, "Parameter Identification Techniques for Multivariate Stochastic Systems," Automatica, Vol. 15, pp. 217-221, 1979.

81. A. Papoulis. Probability Random Variables and Stochastic Processes, New York, MaGraw-Hill, 1965.

82. H.L. Van Trees. Detection Estimation and Modulation Theory, Part I, New York, Wiley & Sons, 1968.

83. E. Bryson, Jr. and Yu-chi Ho. Applied Optimal Control, Optimization, Estimation, and Control, New York, John Wiley & Sons, 1975.

84. D.J. Wiersma and D.A. Landgrebe. "The Analytical Design of Spectral/Measurements for Multispectral Remote Sensor Systems," LARS Technical Report 122678, Laboratory for Applications of Remote Sensing, Purdue University, 1979.

85. R.L. Kettig and D.A. Landgrebe. "Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects," LARS Information Note 050975, Laboratory for Applications of Remote Sensing, Purdue University, 1975.

86. D.J. Wiersma and D.A. Landgrebe. "Analytical Design of Multispectral Sensors," IEEE Transactions on Geoscience and Remote Sensing, Vol GE-18, pp. 180-189, April 1980.

APPENDICES

APPENDIX A

PARAMETERS ESTIMATION FOR THE MARKOV CLASSIFIER

The purpose of this appendix is to find the maximum likelihood estimates of the parameters of the Markov classifier which was described in Chapter 2.

Let $X_{i1}(t)$, $X_{i2}(t)$, ... $X_{in_i}(t)$ be the q-dimensional available training samples of the ith class at time t, and assume that these samples are uncorrelated and from a normal distribution with parameters $M_i(t)$ and $\Sigma_i(t)$. In practice $M_i(t)$ and $\Sigma_i(t)$ are estimated from training samples. It has been shown [2] that the maximum likelihood estimates of a mean vector and a covariance matrix may be estimated by

$$\hat{M}_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}(t) \tag{A-1}$$

$$\hat{\Sigma}_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij}(t)-\hat{M}_i(t))(X_{ij}(t)-\hat{M}_i(t))^T \tag{A-2}$$

In Chapter 2 the temporal variation of energy was modeled by

$$Y_i(t) = \rho_i(t-1)\ Y_i(t-1) + W_i(t) \tag{A-3}$$

where

$$Y_i(t) = X_i(t) - \hat{M}_i(t) \tag{A-4}$$

Note that the model is applied to the prelabeled training samples. The objective here is to estimate $\rho_i(t-1)$, the temporal correlation between observations at time t and t-1, from training samples. The simplest way to estimate $\rho_i(t-1)$ is by the projection principle [81].

$$E[W_i(t)\ Y_i^T(t-1)] = 0 \tag{A-5}$$

From (A-3)

$$W_i(t) = Y_i(t) - \rho_i(t-1)\ Y_i(t-1) \tag{A-6}$$

Then from A-5, we can write

$$E[(Y_i(t) - \rho_i(t-1)\ Y_i(t-1))\ Y_i^T(t-1)] = 0 \tag{A-7}$$

$$E[Y_i(t)\ Y_i^T(t-1)] = \rho_i(t-1)\ E[Y_i(t-1)\ Y_i^T(t-1)] \tag{A-8}$$

Let

$$\Sigma_i(t,t-1) = E[Y_i(t)\ Y_i^T(t-1)] \tag{A-9}$$

and

$$\Sigma_i(t-1) = E[Y_i(t-1)\ Y_i^T(t-1)] \tag{A-10}$$

Then from (A-8)

$$\rho_i(t-1) = \Sigma_i(t,t-1)\ \Sigma_i^{-1}(t-1) \tag{A-11}$$

an estimate of $\rho_i(t-1)$ is given by

$$\hat{\rho}_i(t-1) = \hat{\Sigma}_i(t,t-1)\ \hat{\Sigma}_i^{-1}(t-1) \tag{A-12}$$

where

$$\hat{\Sigma}_i(t,t-1) = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}^T(t-1) \tag{A-13}$$

and

$$\hat{\Sigma}_i(t-1) = \frac{1}{n_i} \sum_{ij=1}^{n_i} Y_{ij}(t-1)\ Y_{ij}^T(t-1) \tag{A-14}$$

APPENDIX B

ITERATIVE CONTEXTUAL CLASSIFIER

The main objective here is to propose for further studies an iterative contextual classifier, which attempts to incorporate the spatial labeling dependencies to reduce the initial labeling error.

A multispectral image data set consists of a location set Ω, an observation set $\{X(s),\ s\varepsilon\Omega\}$, and a class set $C = \{\omega_1,\omega_2,\ldots,\omega_m\}$ where Ω is a two dimensional array of pixel locations, i.e.,

$$\Omega = \{s = (i,j)\ \ 1 \leq i \leq I,\ 1 \leq j \leq J\},$$

X(s) is a q-dimensional random observation at point s and C is the set of all possible classes.

Let $p(\omega_k(s)\,|\,X(s),X(s + (i,j))$, where $(i,j)\ \varepsilon N = \{(0,-1),\ (-1,0),\ (0,-1),\ (1,0)\}$, be the a posteriori probability that, given the observation X(s) at point s and X(s + (i,j)) at point s + (i,j) belonging to one of the neighbors of the pixel at point s. From Bayes rule we obtain

$$p(\omega_k(s),\omega_\ell(s+(i,j)\,|\,X(s),X(s+(i,j))) =$$

$$\frac{p(X(s),X(s+(i,j))\,|\,\omega_k(s),\omega_\ell(s+(i,j))P(\omega_k(s),\omega_\ell(s+(i,j)))}{p(X(s),\ X(s + (i,j)))} \qquad (B\text{-}1)$$

By assuming class conditional independencies, we can write

$$p(X(s),X(s+(i,j))\,|\,\omega_k(s),\omega_\ell(s+(i,j))) =$$

$$p(X(s)\,|\,\omega_k(s))\;\; p(X(s+(i,j))\,|\,\omega_\ell(s+(i,j))) \qquad \text{(B-2)}$$

Also we can write

$$\sum_{\ell=1}^{m} P(\omega_k(s),\ \omega_\ell(s+(i,j))\,|\,X(s),\ X(s+(i,j))) =$$

$$P(\omega_k(s)\,|\,X(s),\ X(s+(i,j))) \qquad \text{(B-3)}$$

and

$$\sum_{k=1}^{m}\sum_{\ell=1}^{m} P(\omega_k(s),\omega_\ell(s+(i,j))\,|\,X(s),X(s+(i,j))) = 1 \qquad \text{(B-4)}$$

thus

$$P(\omega_k(s)\,|\,X(s),\ X(s+(i,j))) =$$

$$\frac{\sum_{\ell=1}^{m} P(\omega_k(s),\omega_\ell(s+(i,j))\,|\,X(s),X(s+(i,j)))}{\sum_{k=1}^{m}\sum_{\ell=1}^{m} P(\omega_k(s),\omega_\ell(s+(i,j))\ X(s),\ X(s+(i,j)))} \qquad \text{(B-5)}$$

By substituting (B-2) in (B-5) we can write

$$P(\omega_k(s)\,|\,X(s),\ X(s+(i,j))) =$$

$$\frac{P(\omega_k(s)\,|\,X(s))\sum_{\ell=1}^{m} P(\omega_\ell(s+(i,j))\,|\,X(s+(i,j)))\ r(\omega_k(s),\omega_\ell(s+(i,j)}{\sum_{k=1}^{m} P(\omega_k(s)\,|\,X(s))\sum_{\ell=1}^{m} P(\omega_\ell(s+(i,j)\,|\,X(s+(i,j)))\ r(\omega_k(s),\omega_\ell(s+(i,j)))} \qquad \text{(3-6)}$$

where

$$r(\omega_k(s),\ \omega_\ell(s+(i,j))) = \frac{P(\omega_k(s),\ \omega_\ell(s+(i,j)))}{P(\omega_k(s))\ P(\omega_\ell(s+(i,j)))} \qquad \text{(B-7)}$$

Let $q^n(\omega_k(s)|X(s))$ be a predicted or an approximate estimate of $P(\omega_k(s)|X(s))$ based on spatial dependency of labeling in a neighborhood at nth iteration. Also let the prediction error be

$$e^n(\omega_k(s)|X(s)) = q^n(\omega_k(s)|X(s)) - p^n(\omega_k(s)|X(s) \qquad \text{(B-8)}$$

where

$$q^n(\omega_k(s)|X(s)) = \tfrac{1}{4}\sum_{(i,j)\in N} P(\omega_k(s)|X(s),X(s+(i,j))) \qquad \text{(B-9)}$$

Since the objective is to incorporate local contextual information, therefore, we expected the prediction error to approach zero in few iterations. The prediction error should strongly affect our decision in the first few iterations because it brings useful information. However, if the prediction error does not approach zero after few iterations, it may be taken to imply that there is insufficient labeling dependency among the pixels in the neighborhood of the pixel under consideration.

Based on the above discussion and from the adaptive labeling algorithm which was developed in Chapter 3, the following algorithm is proposed for further study.

$$p^{n+1}(\omega_k(s)|X(s)) = p^n(\omega_k(s)|X(s)) + \frac{1}{n+1}\ e^n(\omega_k(s)|X(s)) \qquad \text{(B-10)}$$

APPENDIX C

PROGRAMMING CONSIDERATION FOR THE PROBABILISTIC LABELING

To estimate the initial labeling probabilities, we need to compute the term $e^{\ell n\ p(X|\omega_k)}$ which for some pixels ranges over a large negative exponential that may cause underflow or overflow. As discussed in Chapter 3 the probabilistic labeling by maximum likelihood is given by

$$p_i(\omega_k) = \frac{e^{\ell n\ p(X|\omega_k)}}{\sum_k e^{\ell n\ p(X|\omega_k)}} \qquad \text{(C-1)}$$

If the denominator in equation (C-1) is very small, i.e., close to zero, then underflow may occur. To overcome this computational problem

let $M(X) = \max_k \ell n\ p(X|\omega_k), \quad k=1,2,\ldots,m$

Now, let us rewrite equation (C-1) by

$$p_i(\omega_k) = \frac{e^{\{\ell n\ p(X|\omega_k) - M(X)\}}}{\sum_k e^{\{\ell n\ p(X|\omega_k) - M(X)\}}} \qquad \text{(C-2)}$$

In equation (C-2) the underflow problem does not occur because the denominator is always greater than one.

The same method can be used for the probabilistic labeling by the cascade and markov pixel classifiers to avoid the underflow problem. The modified equation for cascade is given by

$$g_k(X_1) = \ell n\ p(X_2|\omega_k) \sum_{\ell=1}^{m_1} p(X_1|V_\ell)P(\omega_k,\omega_\ell)$$

$$= g_k(X_2) + \ell n \sum_{\ell=1}^{m_1} e^{g_\ell(X) + \ell n\ P(\omega_k,V_\ell)} \quad \text{(C-3)}$$

where $g_k(X_2) = \ell n\ p(X_2|\omega_k)$

and $g_\ell(X_1) = \ell n\ p(X_1|V_\ell)$

let $M(X_1) = \max_\ell g_\ell(X_1) \qquad \ell=1,2,\ldots,m \quad \text{(C-4)}$

Then the initial probability can be calculated by

$$p_i(\omega_k) = \frac{e^{\left\{g_k(X_2) + M(X_1) + \ell n\left[\sum_{\ell=1}^{m_1} e^{g_\ell(X_1) + \ell n\ p(\omega_k,V_\ell) - M(X_1)}\right]\right\}}}{\sum_k e^{\left\{g_k(X_2) + M(X_1) + \ell n\left[\sum_{\ell=1}^{m_1} e^{g_\ell(X_1) + \ell n\ p(\omega_k,V_\ell) - M(X_1)}\right]\right\}}} \quad \text{(C-5)}$$

Finally, for the markov pixel classifier the modified equation is given by

$$p_i(\omega_k) = \frac{e^{\{\ell n[p(X(t)|X(t-1);\omega_k)\ p(X(t-1)|\omega_k)] - M(X)\}}}{\sum_k e^{\{\ell n[p(X(t)|X(t-1);\omega_k)\ p(X(t-1)|\omega_k)] - M(X)\}}} \quad \text{(C-6)}$$

where $M(X) = \max_k \ell n[p(X(t)|X(t-1);\omega_k)\ P(X(t-1)|\omega_k)],$

$$k=1,2,\ldots,m \quad \text{(C-7)}$$

APPENDIX D

PARAMETERS ESTIMATION OF THE 2-DIMENSIONAL STOCHASTIC MARKOV MODEL

The purpose of this appendix is to find the maximum likelihood estimate of the parameters of the 2-dimensional stochastic markov model of Chapter 4.

In Chapter 4 the spatial variation of energy for each class has been modeled by

$$y(i,j) = a_1^k y(i-1,j) + a_2^k y(i,j-1) + a_3^k y(i-1,j-1) + w(i,j), \quad (i,j)\varepsilon\Omega_k,\ y(i,j)\varepsilon R \tag{D-1}$$

where

$$y(i,j) = x(i,j) - \hat{m}_k,\ \Omega_k = \{I_{1k} \le i \le I_{2k},\ J_{1k} \le j \le J_{2k}\} \tag{D-2}$$

and m is the estimate of mean of a channel and class under consideration. By using the projection principle [81] which says

$$E[z(i,j)\ w(i,j)] = 0 \tag{D-3}$$

where

$$z(i,j) = [y(i-1,j),\ y(i,j-1),\ y(i-1,j-1)]^T$$

Equation (D-1) can be rewritten by

$$y(i,j) = z^T(i,j)\theta^k + w(i,j) \tag{D-4}$$

and from (D-4)

$$w(i,j) = y(i,j) - z^T(i,j)\,\theta^k \tag{D-5}$$

where

$$\theta^k = [a_1^k, a_2^k, a_3^k]^T$$

By substituting (D-5) in (D-2) we can write

$$E\{z(i,j)[y(i,j) - z^T(i,j)\theta^k]\} = 0 \tag{D-6}$$

From (D-6) we obtain

$$\theta^k = \{E[z(i,j)z^T(i,j)]\}^{-1}\, E[z(i,j)y(i,j)],\ (i,j)\,\varepsilon\Omega_k \tag{D-7}$$

Since it has been assumed that the observations come from a normal distribution, we can estimate θ^k by

$$\hat{\theta}^k = \left[\frac{1}{n_k} \sum_{(i,j)\varepsilon\Omega_k} z(i,j)z^T(i,j)\right]^{-1} \left[\frac{1}{n_k} \sum_{(i,j)\varepsilon\Omega_k} z(i,j)y(i,j)\right] \tag{D-8}$$

The parameter n_k, the number of pixels in class k, can be canceled from equation (D-8) and, therefore,

$$\hat{\theta}^k = \left[\sum_{(i,j)\varepsilon\Omega_k} z(i,j)z^T(i,j)\right]^{-1} \left[\sum_{(i,j)\varepsilon\Omega_k} z(i,j)y(i,j)\right] \tag{D-9}$$

After estimating θ^k, the error signal is predicted for each channel. So, from (D-5) we have

$$\hat{w}(i,j) = y(i,j) - z^T(i,j)\,\theta^k,\ (i,j)\,\varepsilon\Omega_k \tag{D-10}$$

As mentioned in Chapter 4 w(i,j)'s are spatially uncorrelated. However, spectrally they are correlated and the estimate of the covariance matrix of the q-dimension error signal is given by

$$\hat{R}_k = \frac{1}{n_k} \sum_{(i,j)\in\Omega_k} \hat{W}(i,j)\hat{W}^T(i,j) \qquad \text{(D-11)}$$

where

$$\hat{W}(i,j) = [\hat{w}_1(i,j),\ \hat{w}_2(i,j),\ldots\hat{w}_q(i,j)]^T$$

If it is desired to utilize spatial and spectral correlation simultaneously, the model is given by

$$Y(i,j) = A_1^k Y(i-1,j) + A_2^k Y(i,j-1) + A_3^k Y(i-1,j-1) + W(i,j); \quad (i,j)\in\Omega_k, Y(i,j)\in R^q \qquad \text{(D-12)}$$

where A_1, A_2 and A_3 are q x q matrices.

Let $A_\ell^k = \begin{bmatrix} \underline{a}_{\ell_1}^k \\ \underline{a}_{\ell_2}^k \\ \vdots \\ \underline{a}_{\ell_q}^k \end{bmatrix}$ and $\underline{a}_{\ell_1}^k = [a_{\ell_{11}}^k \quad a_{\ell_{12}}^k \ \ldots \ a_{\ell_{1q}}^k]$,

$\ell = 1,2,3.$

Also let $\theta_P^k = [a^k_{1P}, a^k_{2P}, a^k_{3P}]^T$, n, P=1,2,...q. Then from equation (D-12), we obtain

$$y_P(i,j) = Z^T(i,j)\theta_P^k + w_P(i,j) \qquad \text{(D-13)}$$

where

$$Z(i,j) = [Y(i-1,j), Y(i,j-1), Y(i-1,j-1)]^T \qquad \text{(D-14)}$$

is a 3q x 1 vector. An estimate of θ_p^k as considered earlier can be obtained by

$$\hat{\theta}_p^k = \left[\sum_{(i,j)\in\Omega_k} Z(i,j) Z^T(i,j) \right] \left[\sum_{(i,j)\in\Omega_k} Z(i,j) y_p(i,j) \right] \qquad \text{(D-15)}$$

$$p=1,2,\ldots q$$

Now, $\hat{w}_p(i,j) = y_p(i,j) - Z(i,j)\,\theta_p^k$ are spatially and spectrally uncorrelated. Therefore,

$$\hat{R}_k = \frac{1}{n_k} \sum_{(i,j)\in\Omega_k} \hat{W}(i,j)\hat{W}^T(i,j) \qquad \text{(D-16)}$$

should be a diagonal matrix. The estimate of variance of error signal for each channel is given by

$$\hat{r}_{pp}^{\,k} = \frac{1}{n_k} \sum_{(i,j)\in\Omega_k} \hat{w}_p^{\,2}(i,j) \qquad \text{(D-17)}$$

A more detailed discussion on parameter estimation in multivariate stochastic difference equations is given in [77-80].

APPENDIX E

PREDICTING THE REQUIRED NUMBER OF TRAINING SAMPLES

In this appendix a criterion which measures the quality of the estimate of the covariance matrix of a multivariate normal distribution is developed. Based on this criterion, the necessary number of training samples is predicted. Experimental results which are used as a guide for determining the number of training samples are included.

In practice, the number of training samples is frequently limited because it is expensive to collect many training samples. A typical application in which this is the case is the field of remote sensing, and we will use this application to illustrate the technique.

In remote sensing, the reflected and emitted electromagnetic energy of each pixel of a scene in several important wavelength bands is measured by a multispectral remote sensor system mounted on board an aircraft or spacecraft. The output of the sensor system is used to form a point in a q-dimensional space [E-6] A commonly used pattern classification algorithm in this application is the maximum likelihood Gaussian scheme. In this instance, the classes are

each characterized as a Gaussian distribution in q-space and these distributions in turn are specified by estimates of the means and covariances of each. However, we know that the performance of the estimators is dependent on the number of training samples. In the case of limited training samples, the estimates of the first and second order statistics cannot accurately depict all the information which is contained in the data. In particular, the estimate of the covariance matrix may be poor. As a result of this poor estimation, later analysis of the data (for example, classification accuracy and statistical distance measures) will be degraded. See [E-1] for more details. Therefore, it is important to predict how many samples will be needed in order that the performance of the estimators be statistically reasonable. In the following, a criterion is developed to measure the performance of the estimate of the covariance matrix; then the number of required samples is predicted.

E.1 Prediction Criterion

Let $X_1, X_2, \ldots X_n$ be q-dimensional random sample vectors which are drawn from a normally distributed population with parameters $\theta = (M, \Sigma)$, where M is the true mean vector and Σ the true covariance matrix. In practice, M and Σ are not available, so they must be estimated from the observed data. The maximum likelihood estimates of M and Σ are:

$$\hat{M} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{E-1}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (X_i - \hat{M})(X_i - \hat{M}) \tag{E-2}$$

For more detail, see [E-2]

The performance of an estimator is measured by properties, such as whether it provides (a) an unbiased estimate, (b) a consistent estimate, (c) an efficient estimate, and (d) a sufficient estimate. Now, let us study the properties of maximum likelihood estimates of M and Σ. From [2] we have:

$$E[\hat{M}] = M \tag{E-3}$$

$$Cov[\hat{M}] = \frac{1}{N}\Sigma \tag{E-4}$$

$$E[\hat{\Sigma}] = \frac{N-1}{N}\Sigma \tag{E-5}$$

Thus, by definition, $\hat{M}$ is an unbiased estimate of M, but $\hat{\Sigma}$ is not an unbiased estimate of Σ. However, if

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N} (X_i - \hat{M})(X_i - \hat{M})^T \tag{E-6}$$

then $E[\hat{\Sigma}] = \Sigma$ which is unbiased. The density function of $\hat{M}$ and $\hat{\Sigma}$ are:

$$p(\hat{M}) = \frac{1}{(2\pi)^{\frac{q}{2}}\left|\frac{1}{N}\Sigma\right|^{\frac{1}{2}}} \exp\left\{-\tfrac{1}{2}(\hat{M}-M)^T N\Sigma^{-1}(\hat{M}-M)\right\} \tag{E-7}$$

$$p(\hat{\Sigma}) = \frac{(N-1)^{q}|\Sigma|^{(N-q-2)/2}\exp\{-\frac{1}{2}(N-1)\mathrm{tr}\Sigma^{-1}\hat{\Sigma}\}}{2^{(N-1)q/2}\pi^{q(q-1)/4}|\Sigma|^{(N-1)/2}\prod_{i=1}^{q}\Gamma[\frac{1}{2}(N-i)]} \quad \text{(E-8)}$$

That is, $\hat{M} \sim N(M, \frac{1}{N}\Sigma)$, a normal distribution and $\hat{\Sigma} \sim W(\Sigma, N)$, a Wishart distribution. For more details of other properties of these estimators, see [E-2,-3] and for various properties of the Wishart distribution see [E-4].

Though the distribution of $\hat{\Sigma}$ is complex, the performance of the estimates of the covariance matrix which are of interest can be measured by the variance of the diagonal components of Σ, as follows:

$$\hat{\sigma}_{kk} = \frac{1}{N-1}\sum_{i=1}^{N}(X_{i_k} - \hat{m}_k)^2 \qquad k=1,2,\ldots,q \qquad \text{(E-9)}$$

In [3] it is shown that $(N-1)\frac{\hat{\sigma}_{kk}}{\sigma_{kk}}$ has a chi-square distribution with (N-1) degrees of freedom. And

$$E[\hat{\sigma}_{kk}] = \sigma_{kk} \qquad \text{(E-10)}$$

$$E\left[\frac{\hat{\sigma}_{kk}}{\sigma_{kk}}\right] = 1 \qquad \text{(E-11)}$$

$$\mathrm{var}[\hat{\sigma}_{kk}] = \frac{2\sigma_{kk}^2}{N-1} \qquad \text{(E-12)}$$

$$\text{var}\left[\frac{\hat{\sigma}_{kk}}{\sigma_{kk}}\right] = \frac{2}{N-1} \qquad \text{(E-13)}$$

In a similar manner, and in order to facilitate the evaluation of the covariance matrix one can work in a new space via the following transformation:

$$Y = \Lambda^{-\frac{1}{2}}\ \Phi^{T}\ (X - M)$$

where Φ and Λ are respectively the eigenvector matrix and the eigenvalue matrix of Σ.

This transformation leads to:

a) choose the mean M as origin.

b) transform the covariance matrix into the unity matrix.

In effect, we have:

$$YY^{T} = \Lambda^{-\frac{1}{2}}\ \Phi^{T}(X-M)\ \ (X-M)^{T}\Phi\Lambda^{-\frac{1}{2}}$$

and $\text{cov}(Y) = \Lambda^{-\frac{1}{2}}\ \Phi^{T}\ \Sigma\Phi\Lambda^{-\frac{1}{2}}$

So $\Phi^{T}\Sigma\Phi = \Lambda$ because the orthonormal matrix Φ.

Thus $\text{cov}(Y) = I$

In practice Φ and Λ are the eigenvector matrix and the eigenvalue of $\hat{\Sigma}$.

Hence $Y = \hat{\Lambda}^{-\frac{1}{2}}\hat{\Phi}^{T}(X-\hat{M})$

and $\text{cov}(Y) = \hat{I}$ where the diagonal elements are noted as $\hat{\gamma}_{kk}$. Because of the orthonormal transformation, the

features in the new space are independent; therefore, $(N-1)\hat{\gamma}_{kk}$ has chi-square distribution with (N-1) degrees of freedom. For brevity, let:

$$(N-1)\hat{\gamma}_{kk} \sim \chi^2(N-1) \tag{E-14}$$

$$\text{and} \quad \hat{Q} = [\hat{\gamma}_{11} + \ldots + \hat{\gamma}_{qq}] \tag{E-15}$$

$$\text{then } (N-1)\hat{Q} \sim \chi^2(q(N-1)) \tag{E-16}$$

$$E[(N-1)\hat{Q}] = q(N-1)) \tag{E-17}$$

$$E[\hat{Q}] = q \tag{E-18}$$

$$\text{var}[(N-1)\hat{Q}] = 2q(N-1) \tag{E-19}$$

$$\text{var}(\hat{Q}) = \frac{2q}{N-1} \tag{E-20}$$

A logical choice for our prediction criterion is $\text{var}(\hat{Q})$ because it measures the dispersion of the estimate of the covariance matrix.

To see how to apply the criterion, suppose it is desired that $\text{var}(\hat{Q}) \leq \alpha$. Therefore, from (E-20)

$$N \geq 1 + \frac{2q}{\alpha} \tag{E-21}$$

Note that the minimum value of N is q + 1, because if N is less than q + 1, then the covariance matrix will be singular. So,

$$\text{var}(\hat{Q})_{\max} = \frac{2q}{N_{\min}-1} = 2 \tag{E-22}$$

A plot of the $\text{var}(\hat{Q})$ as a function of N with q as a parameter is shown in Figure E.1. Now, if for example $\alpha = 0.2$, then $N \geq 1 + 10q$.

The next question to be addressed is how does one choose a reasonable value for α. To answer this question, let us consider the following. As shown in Figure E.1, if $N > 1 + 10q$, then $\text{var}(\hat{Q})$ is decreasing very slowly and its slope $(-\frac{\text{var}(\hat{Q})}{N})$ is small, less than $-.02/q$ because from equation E-20, if $N > 1 + mq$ then slope will be less than $-\frac{2}{m^{2q}}$. This suggests that if $N = 1 + 10q$, then the statistical distance between the true probability density and the estimated one may be close to zero because the estimates of the mean vector and covariance matrix are very close to the true ones ($\text{var}(\hat{Q}) = 0.2$). The transformed divergence [E-5,-6] is a useful statistical distance measure and is given by

$$D_T = 2000[1 - \exp(-D/8)], \tag{E-23}$$

where

$$D = \tfrac{1}{2}\text{tr}(\Sigma - \hat{\Sigma})(\hat{\Sigma}^{-1} - \Sigma^{-1}) + \tfrac{1}{2}\text{tr}(\Sigma^{-1} + \hat{\Sigma}^{-1})(M - \hat{M})(M - \hat{M})^T \tag{E-24}$$

We will use it to experimentally measure the quality of the estimates of the parameters and also as a guide to choosing α or N. The following procedure provides a practical means for doing so:

1. Assume that the true probability density of the data is normal with mean vector M and covariance

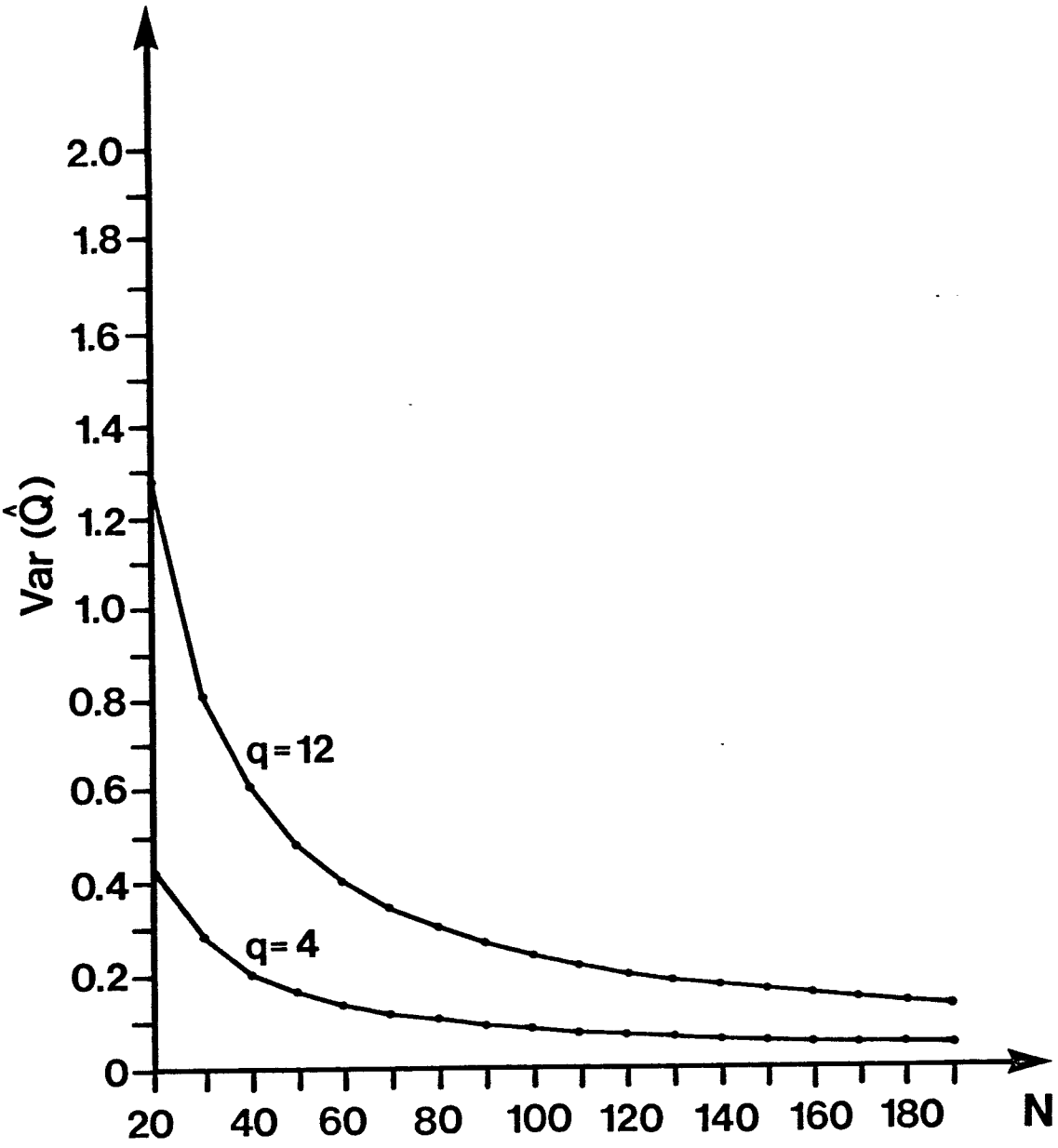

Figure E.1 Variance of $\hat{Q}$ as a function of number of training samples N.

matrix Σ (M and Σ are chosen to be 12 x 1 and 12 x 12 matrices, respectively).

2. Based on the true parameters of the distribution, N_i data points are randomly generated.
3. The parameters of the distribution are estimated based on the N_i randomly generated samples and then, using transformed divergence, the statistical distance between the true probability density and the estimated one is computed.
4. Step 3 is repeated five times and the average transformed divergence is calculated.
5. The average transformed divergence for different values of $\text{var}(\hat{Q})$ is computed and shown in Figure E.2.

The result in Figure E.2 shows almost a linear relationship between D_T and $\text{var}(\hat{Q})$. This implies that when $\text{var}(\hat{Q}) = \text{var}(\hat{Q})_{max} = 2$, then $D_T = (D_T)_{max} = 2000$. This indicates that the quality of the estimates of the parameters (mean vector and covariance matrix) is very poor. However, if $\text{var}(\hat{Q}) = 0.2$, then $D_T = 175$, which suggests that the estimated probability density is very close to the true one. In practice, however, the true parameters of the distribution are not available and neither is the transformed divergence. As mentioned earlier, a logical choice for our prediction criterion is $\text{var}(\hat{Q})$ because it measures the dispersion of the estimate.

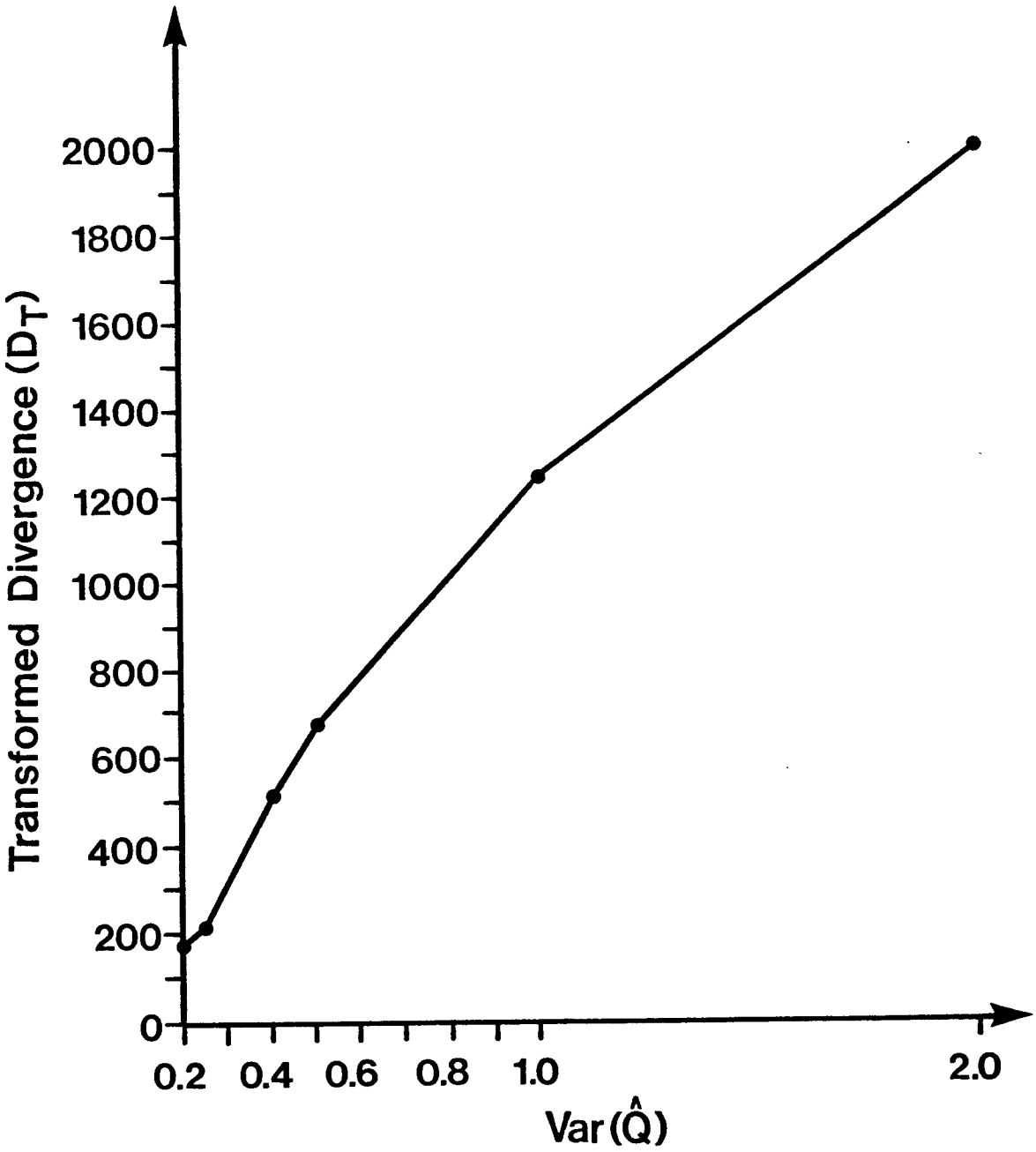

Figure E.2 The average transformed divergence as a function of variance of $\hat{Q}$.

We have found that $D_T = 500$, or equivalently, $\alpha = 0.4$ is a logical threshold to decide whether the estimates of the parameters are good or not. This choice implies that the number of training samples should not be less than 1 + 5q. However, we believe by using information given in Table E.1, one should be able to establish an upperbound on $\text{var}(\hat{Q})$ and consequently estimate the required number of training samples.

Table E.1 Distance between the true distribution and estimated one as a function of $\text{var}(\hat{Q})$ or number of training samples.

$\text{var}(\hat{Q})$	D_T	D	N
1.00	1250	7.85	1 + 2q
0.50	675	3.40	1 + 4q
0.40	500	2.30	1 + 5q
0.25	210	0.80	1 + 8q
0.20	175	0.70	1 + 10q

References

E-1. M.A. Muasher and D.A. Landgrebe. Multistage classification of multispectral earth observational data: The design approach. School of Electrical Engineering, Technical Report TR-EE 81-41, and Laboratory for Applications of Remote Sensing, LARS Technical Report 101381, Purdue University, West Lafayette, IN 47907-0501. Dec. 1981.

E-2. K. Fukunaga. Introduction to statistical pattern recognition. Academic Press, New York, 1972.

E-3. J.P. Bickel and A.K. Docksum. Mathematical statistics: basic ideas and selected topics. Holden-Day, Inc., San Francisco, 1977.

E-4. T.W. Anderson. Introduction to multivariate statistical analysis. Wiley, New York, 1958.

E-5. P.H. Swain and R.C. King. Two Effective Feature Selection Criteria for Multispectral Remote Sensing. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN 47906-1399. LARS Technical Report 042673, Apr. 1973.

E-6. P.H. Swain and S.M. Davis, eds. Remote Sensing: The Quantitative Approach. McGraw Hill, Inc., New York, 1978.

APPENDIX F

FEATURE SELECTION WITH LIMITED TRAINING SAMPLES

A criterion is developed which measures the quality of the estimates of the parameters of multivariate normal distributions for two class problems when limited number of samples are available. This criterion predicts if the Hughes phenomenon occurs. The maximum number of features which does not degrade the accuracy of the classifier is then predicted.

In pattern recognition, it is frequently possible to find a subset of features which gives almost the same or perhaps even better probability of correct classification than if all features are used. In the case of parametric classifiers, if accurate estimates of the parameters are available, then feature selection is done simply to reduce the computational complexity. But if the number of training samples is small, estimates of the parameters may be poor. In this case, feature selection becomes more important; if all features are used, the probability of error will be greater than when only a smaller number of features are used (see

An example of where the proper choice of feature subsets is especially important is that of the decision tree classifier. Let us assume that we are dealing with a binary tree classifier such that a limited number of training samples at each node is given and the problem is to find the maximum number of features that must be used at each node without increasing the probability of error. (For more detail on a binary tree classifier, see [F-2].

F.1 Prediction Criterion for Determining the Maximum Number of Features

If the number of training samples is quite limited, one might suppose that there will be more difficulty estimating covariances than means. With this in mind, let us consider two class problems. Let $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ be estimates of the covariance matrices based on n_1 and n_2 samples of class ω_1 and class ω_2, respectively. Then let $\hat{I}$, $\hat{\Lambda}$ be the estimates of covariance matrices of ω_1 and ω_2 after applying simultaneous diagonalization transformations where $\hat{I}$ is an estimate of the identity matrix and $\hat{\Lambda}$ is a diagonal matrix (see [F-2]). Let $\hat{\alpha}_{ii}$ and $\hat{\lambda}_{ii}$ be the diagonal elements of $\hat{I}$ and $\hat{\Lambda}$, respectively. Then let

$$\hat{Q}_1 = \sum_{i=1}^{q} \alpha_{ii} \qquad \text{(F-1)}$$

$$\hat{Q}_2 = \sum_{i=1}^{q} \frac{\hat{\lambda}_{ii}}{\lambda_{ii}} \qquad \text{(F-2)}$$

$$\hat{Q} = \hat{Q}_1 + \hat{Q}_2 \qquad \text{(F-3)}$$

where α_{ii} and λ_{ii} are the true variances of the ith feature of class ω_1 and ω_2 in the new space and q is the number of features used. In the new space, the features are independent and so are their variances in class ω_1 and class ω_2. Furthermore, we are assuming the elements of the covariance matrices of two classes in the new space are independent. Consequently, it can be said that $\hat{Q}_1$ and $\hat{Q}_2$ are two independent random variables; then we can write

$$\text{var}(\hat{Q}) = \text{var}(\hat{Q}_1+\hat{Q}_2) = \text{var}(\hat{Q}_1)+\text{var}(\hat{Q}_2) \qquad \text{(F-4)}$$

In Appendix E, it is shown that

$$\text{var}(\hat{Q}_i) = 2\ q/(n_i-1) \qquad \text{(F-5)}$$

We will choose $\text{var}(\hat{Q}_1+\hat{Q}_2)$ as our prediction criterion to determine the maximum number of features for which there is no degradation in accuracy. Then we have

$$\text{var}(\hat{Q}) = 2\ q/(n_1-1) + 2\ q/(n_2-1) \qquad \text{(F-6)}$$

The proposed criterion measures the quality of the estimators and for a given n_1 and n_2 suggests the maximum number of features that should be used. For simplicity, suppose q = 1. The probability of error is given by the shaded area of Figure F.1a. Let ε_0 be this probability of error if the parameters of the class distributions are known or if accurate estimates of these parameters are available. To have an accurate estimate of the parameters, a large number of

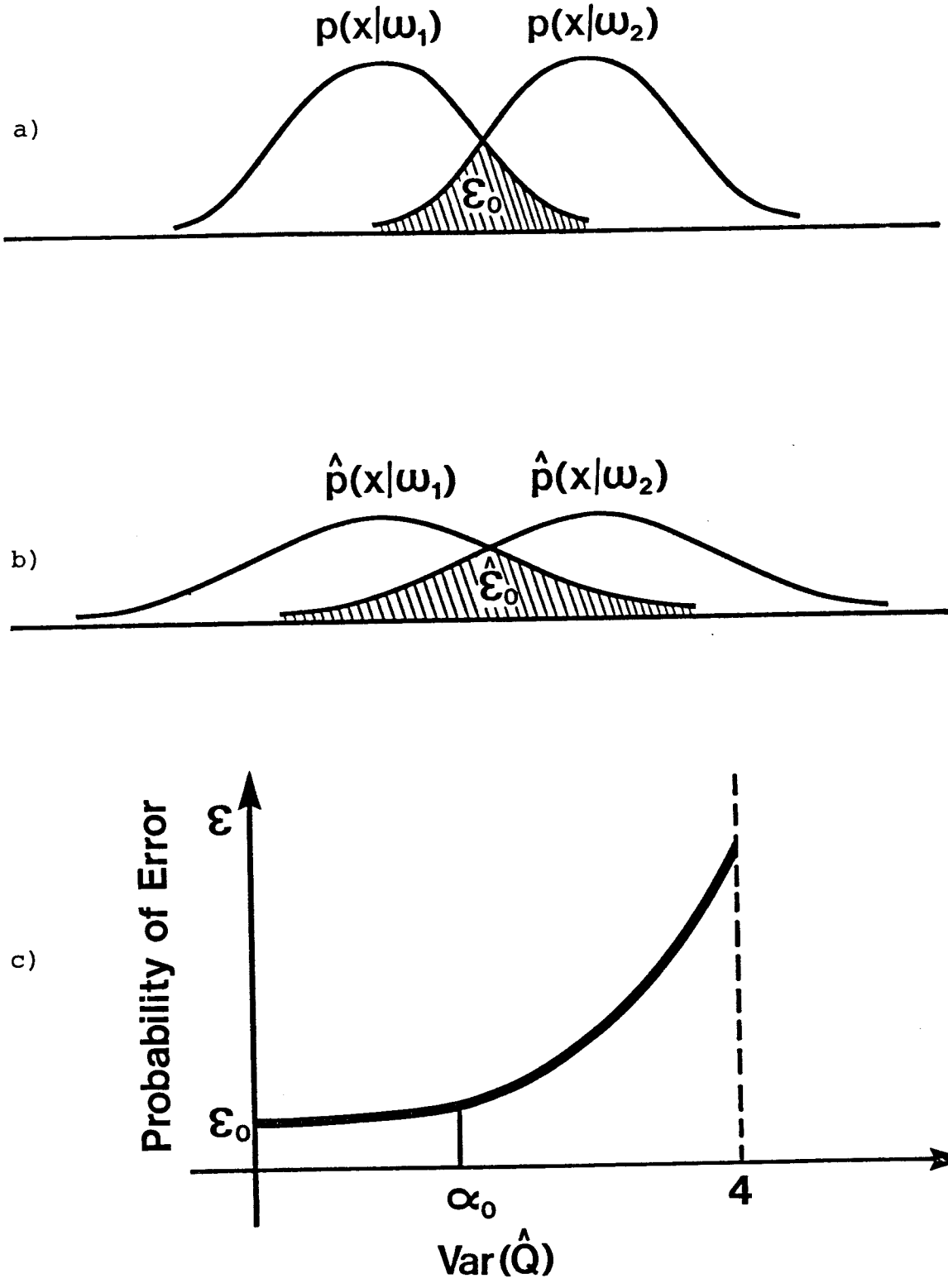

Figure F.1 Degradation in accuracy as explained by class probability densities with a) known and b) estimated parameters and c) a hypothetical curve of the probability of error as a function of $\mathrm{Var}(\hat{Q})$.

training samples is needed. Let $\hat{\varepsilon}_0$ be the probability of error if the parameters of the distributions are estimated from a limited set of training samples. Equation (F-6) indicates when n_i is very small, then $\text{var}(\hat{Q})$ is large. With the presence of a fixed, limited training sample size, any increase in dimensionality necessarily results on the average in a degradation in the accuracy of statistics estimation of the class distributions. Because of variance of the estimated parameters (particularly covariance matrices), one should expect $\hat{\varepsilon}_0$ to be greater than ε_0 (Figure F.1b). of error as a function of var(Q). We expect the probability of error to be almost constant at a low variance, then as the estimate of the covariance matrices becomes poorer and poorer, it begins to increase. For the worst case, when $n = q + 1$ then from equation (F-6) $\text{var}(\hat{Q}) = \text{var}(\hat{Q}) = 4$, the accuracy of the classification will be degraded the most.

Our objective is to find the maximum number of features (corresponding to some threshold value α_0) for a given number of training samples for which there is no degradation in accuracy. It must be recognized, however, that estimated performance is a random variable since it is based upon estimated class statistics. We will therefore determine the maximum number of features based upon an ensemble average performance since we cannot insure what will take place precisely on any one trial. α_0 will be

experimentally determined. The maximum number of features for a given number of training samples which does not degrade the performance of a binary tree classifier at each node can be calculated from equation (F-6) by

$$q = \frac{(n_1-1)(n_2-1)}{2(n_1+n_2-2)} \alpha_0 \qquad \text{(F-7)}$$

In deriving equation (F-7), it has been assumed that for $\text{var}(\hat{Q}) \simeq \alpha_0$ Hughes Phenomenon begins to occur. Figure F.2 shows the $\text{Var}(\hat{Q})$ plotted against the number of features for different numbers of training samples, using equation (F-6), with $\alpha_0 = 1.0$ shown on the curve. It shows that when $n_1 = n_2 = 13$, the number of features to be used is close to 3, and is close to 5 when $n_1 = n_2 = 20$.

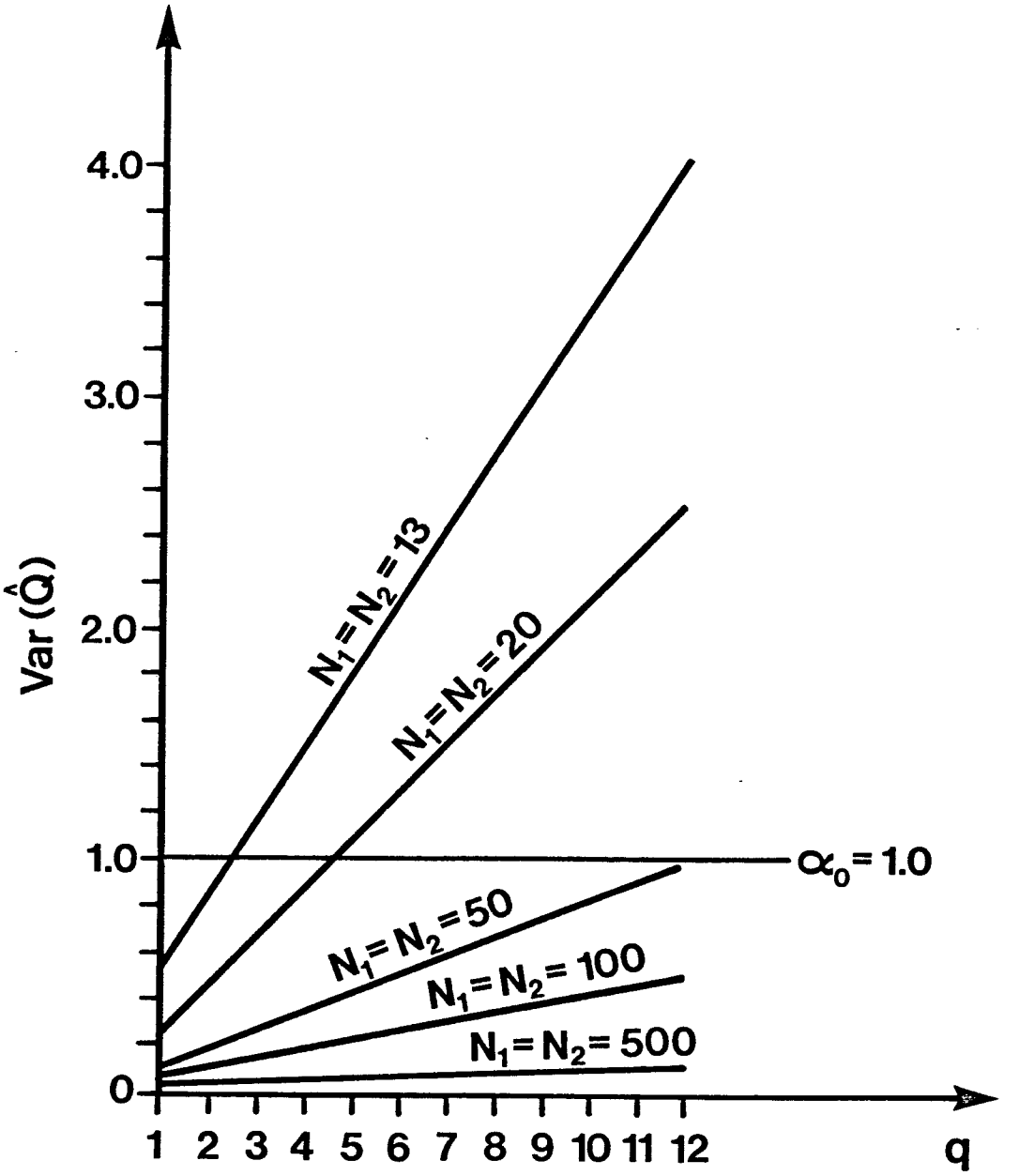

Figure F.2 Variance of $\hat{Q}$ as a function of the number of features for different number of training samples.

References

F-1. M.J. Muasher and D.A. Landgrebe. Multistage Classification of Multispectral Earth Observational Data: The Design Approach. School of Electrical Engineering, Technical Report TR-EE 81-41, and Laboratory for Applications of Remote Sensing, Technical Report 103181, Purdue University, West Lafayette, IN 47907-0501. Dec. 1981.

F-2. K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, New York, 1972.

APPENDIX G

INFORMATION ABOUT THE SOFTWARE SYSTEM AND DATA SETS

Table G.1 Information about the (modified or developed) programs for Markov classifier.

Location / Programs	Tape	File	File name	File type
Programs for writing the likelihood values	858	33	CLASS CLSFY2 CONTEX	ASSEMBLE FORTRAN FORTRAN
Programs for generating statistics	858	35	STASUP STAINT PCHSTA LEARN	FORTRAN FORTRAN FORTRAN FORTRAN
Programs for classification	858	36	CLAINT CLASUP CLSFY1 CLSFY2 CONTEX PRINT REDSAV REDSTA STATS	FORTRAN FORTRAN FORTRAN FORTRAN FORTRAN FORTRAN FORTRAN FORTRAN FORTRAN

Table G.2 Information about the data set and statistics for Markov classifier.

Data/Statistics \ Location	Tape	File	File name	File type
Training data	501	9	STATF	M843
Test data	501	9	CLASS	843F
Statistics	501	6	843F1	STATDECIC
		6	843F2	STATDECIC
		6	843F3	STATDECIC
		6	843T12	STATDECIC
		6	843T23	STATDECIC
		1	843P13B	STATDECIC
Likelihood values ($\ell n\ p(X(t)\,|\,\omega_i)$	500	1 (t=June 9)		
		2 (t=July 16)		

Table G.3 Information about the (modified or developed) programs for the probabilistic relaxation algorithms.

Location / Programs	Tape	File	File name	File type
Programs for writing the likelihood values	858	33	CLASS CLSFY2 CONTEX	ASSEMBLE FORTRAN FORTRAN
Programs for classification	501	8	RELAX RELAX RELSUB1 or RELSUB3	EXEC FORTRAN FORTRAN FORTRAN
			RELAX3 RELAX3 RELSUB1 or RELSUB3	EXEC FORTRAN FORTRAN FORTRAN

Table G.4 Information about data set and the likelihood values for probabilistic relaxation algorithms.

Location / Data/ Likelihood	Tape	File	File name	File type
Test data	501	9	843G12	DATA
			843G22	DATA
			843G13	DATA
			843G23	DATA
			843MD12	DATA
			843MD13	DATA
			843ML12	DATA
			843ML13	DATA
			843ML22	DATA
Likelihood values	4280	1	(843G12 DATA; August)	
	4280	2	(843G22 DATA; August)	
	4280	3	(843G22 DATA; September)	
	4279	1	(SKYLAB DATA)	
	4281	1	(843G13 DATA; September)	

Table G.5 Information about the (modified or developed) programs for MMLO and MMDO.

Location / Programs	Tape	File	File name	File type
Programs for generating statistics	501	9	GSTASUD LEARN SPCOR	EXEC FORTRAN FORTRAN
Programs for classification (MMLO)	501	9	GCLASUP CLAINT CLASUP CLSFY1 CLSFY2 CONTEX REDSAV REDSTA	EXEC FORTRAN FORTRAN FORTRAN FORTRAN FORTRAN FORTRAN FORTRAN
Programs for classification (MMDO)	501	9	GSAMSUP SMCLS2 SPCOR	EXEC FORTRAN FORTRAN

Table G.6 Information about data set and statistics for MMLO and MMDO.

Location / Data/ Statistics	Tape	File	File name	File type
Training data	501	9	STAFF	M843
			STATWH	539
Test data	501	9	CLASS	843F
			SAMP2WH	539CL3
			CLASS	539WH
Statistics	501	6	843F1	STATDECK
		6	843F2	STATDECK
		6	843F3	STATDECK
		5	843I11	STATDECK
		5	843I22	STATDECK
		5	843I33	STATDECK
		5	123843B	STATDECK
		9	WH539MD	STATDECK