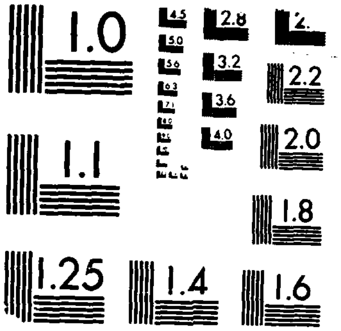
AD-A192 525 SEMI-SUPERVISED TWO STAGE CLASSIFICATION TECHNIQUE(U) 1/3
ARMY MILITARY PERSONNEL CENTER ALEXANDRIA VA
D A TOOMEY 31 JUL 87

UNCLASSIFIED F/G 14/5 NL



MICROCOPY RESOLUTION TEST CHART

UREAU OF STANDARDS-1963-A

2

SEMI-SUPERVISED TWO STAGE CLASSIFICATION TECHNIQUE

1LT DANIEL A. TOOMEY
HQDA, MILPERCEN (DAPC-OPA-E)
200 Stovall Street
Alexandria, VA 22332

Final Report
31 July 1987

Approved for public release; distribution unlimited.

A thesis submitted to University of Wisconsin-Madison, Madison Wisconsin, in partial fulfillment of the requirements for the degree of Master of Science.

664-6967 (work) 461-9584 (Home)

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188
Exp. Date Jun 30, 1986

1a. REPORT SECURITY CLASSIFICATION	1b. RESTRICTIVE MARKINGS AD-A192 525
2a. SECURITY CLASSIFICATION AUTHORITY 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	3. DISTRIBUTION/AVAILABILITY OF REPORT
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Semi-Supervised Two Stage Classification Technique	5. MONITORING ORGANIZATION REPORT NUMBER(S) Final Report 31 July 1987
6a. NAME OF PERFORMING ORGANIZATION — 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code)	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION — 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. — PROJECT NO. — TASK NO. — WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)

12 PERSONAL AUTHOR(S)
31 July 1987

13a. TYPE OF REPORT Master of Science	13b. TIME COVERED FROM ________ TO ______	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT 228

16. SUPPLEMENTARY NOTATION

Approved for public release; distribution unlimited.

17 COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	~~University of Wisconsin-Madison~~ See the attached Terms on the reverse side.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

A Semi-Supervised Two Stage Classification Technique has been developed on the IBM PC-AT computer at the Environmenal Remote Sensing Center, University of Wisconsin-Madison. This technique is used to classify multispectral digital images. It involves two stages. The first is a hybrid clustering technique and the second is a reclassification (post-classification process) of designated spectral classes in a spectrally classified image with ancillary information.

In the first stage, the analyst directs the clustering algorithm by delineating a certain number of training "areas" so that an unsupervised clustering algorithm can identify a user defined number of spectral clusters in each area. These clusters are then implemented as seeds to collect further information from throughout the entire image. (over)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED □ SAME AS RPT □ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) — 22c. OFFICE SYMBOL

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

Block 19 (continued)

In the second stage, ancillary data is employed as a Second Stage of digital information to reclassify certain spectrally classified land cover types to increase the classification accuracy. Two types of reclassification can be applied, a statistical method and threshold level approach. The statistical reclassification uses Second Stage statistical input, while the threshold level approach implements a Second Stage image file that is amenable to thresholding.

hybrid clustering
spectral ~~[illegible]~~ image classification
post-classification
unsupervised clustering
supervised clustering
~~mismerging~~
spectral mismerging,
Ancillary data
Reclassification
(statistical,
threshold)
variance threshold
Transformed divergence

Subject Terms

APPROVED:

Frank L. Scarpace

Frank L. Scarpace
Associate Professor
Environmental Studies and
Civil and Environmental Engineering

Date: 7/31/87

A digitized color infrared aerial photograph, of the Chesapeake Bay region, is classified using both stages of the classification technique to demonstrate the potential of reclassifying only certain spectral classes with ancillary data. A statistical reclassification is done according to texture as the Second Stage; a threshold range reclassification is accomplished with a vegetation index ratio; and a threshold level reclassification is performed using a polygon-masked image.

ABSTRACT

A Semi-Supervised Two Stage Classification Technique has been developed on the IBM PC-AT computer at the Environmental Remote Sensing Center, University of Wisconsin-Madison. This technique is used to classify multispectral digital images. It involves two stages. The first stage is a hybrid clustering technique and the second is a reclassification (post-classification process) of a spectrally classified image with digital ancillary information.

In the first stage, the analyst directs the clustering algorithm by delineating a certain number of training areas so that an unsupervised clustering algorithm can identify a user defined number of spectral clusters in each area. These clusters are then implemented as seeds to collect further spectral information from throughout the entire image. Mismerging of spectral clusters to the seeds is prevented by a user defined variance threshold and a transformed divergence computation.

The resulting clusters, training sets, are implemented into a statistical classifier to segment the scene according to spectral information.

In the second stage, ancillary data is employed as a Second Stage of digital information to reclassify certain spectrally classified land cover types to increase the classification accuracy. Two types of reclassification can be applied, a statistical method and threshold level approach. The statistical reclassification uses Second Stage statistical input, while the threshold level approach implements a Second Stage image file that is amenable to thresholding.

A SPOT satellite sub-scene over the Greater-Madison area in Wisconsin is segmented utilizing the Semi-Supervised clustering approach. The FINDSET algorithm is an unsupervised clustering algorithm that is presently employed at the Environmental Remote Sensing Center. A comparison between the Semi-Supervised approach and the FINDSET algorithm is assessed.

Acknowledgements

I wish to thank all of my educators and mentors throughout the course of my academic experience, especially the invaluable guidance and support of Frank L. Scarpace, my major advisor.

I am very grateful for the complete support from the McAIERS (The Microcomputer Aerial Imagery Enhanced Reconnaissance System) research team during my thesis research. Being a part of McAIERS has offered many valuable experiences in the field of remote sensing.

I would like to extend my appreciation for the guidance and assistance from the professors on my committee: Ralph W. Kiefer, James P. Scherz and Paul R. Wolf.

I would like to thank Professor Thomas M. Lillesand, director of the Environmental Remote Sensing Center, for permission to work with the SPOT satellite data from the SPOT Early Assessment Program (PEPS).

Two years ago I began my graduate years of study in the field of remote sensing, and along the way many have shared their knowledge and wisdom. I would like to extend my appreciation to all of these special people, and especially Kurt Worrell, the McAIERS systems manager and principal computer programmer, and Paul Thies, a friend and mentor.

I would like to thank all my fellow graduate students, friends and family; dear friends who helped me along the way and offered words of encouragement when it seemed to be the best medicine.

And an added note of appreciation to the secretaries, the unsung heroes who assisted with the administrative details involved in graduate school: Marcia M. Verhage, Environmental Remote Sensing Center; and Susan M. Brunsell, Civil and Environmental Engineering.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

To my Parents

Chapter I

Introduction

The following thesis research studies the use of unsupervised clustering algorithms and the implementation of ancillary data in the automated classification process of remotely sensed data. This study is a development and evaluation of a Semi-Supervised Two Stage classification technique. This classifier combines the advantages of an unsupervised classification with the direction (guidance) attributed a supervised classification.

In previous years, unsupervised clustering algorithms that are not considered statistically rigorous have been found to contain biases toward certain aspects of remotely sensed data. FINDSET is such an algorithm and is presently used at the Environmental Remote Sensing Center at the University of Wisconsin-Madison (see section 2.11). It is based on the SEARCH algorithm of the ELAS package (NASA, 1981)(see section 2.10). A Semi-Supervised approach developed in this thesis research reshapes the traditional algorithm structure of FINDSET to potentially reduce these biases.

Currently ancillary data is indiscriminately used in assisting the spectral classification of the entire image of interest, thereby classifying all the land cover in an image with spectral and ancillary data. From the results of previous research, discussed in chapter three, it can be concluded that ancillary data is more appropriate in classifying only certain land cover types in an image. A reclassification approach that implements ancillary data in a discriminant manner, directing the reclassification of only certain cover types, could be more useful to the remote sensing community.

The thesis hypothesis is 1) that a Semi-Supervised Two Stage Classification technique can be developed that would reduce some of the biases possibly inherent in the FINDSET algorithm and 2) that the application of ancillary data as a Second Stage used in a discriminant manner will improve the classification accuracy.

This research involves two objectives :

1) Develop and evaluate a clustering technique that could reduce some of the biases that may be found in FINDSET. The clustering approach developed is termed Semi-Supervised. The Semi-Supervised method stratifies the FINDSET approach to clustering, and includes additional statistical metrics to assist in forming spectral statistics for the land cover classes in the image.

2) To implement a post-classification routine (after the spectral classification) that would implement ancillary data in a discriminant manner; reclassifying only certain land cover classes, spectrally classified, with additional digital information.

These two processes are not necessarily sequential in application. The Semi-Supervised approach does not have to implement a reclassification; and the post classification reclassification does not require a spectral classification that results from the Semi-Supervised process.

The remaining chapters discuss background (literature review), the developed algorithms and methodology of the analysis. Chapter 2 explains the classification techniques and processes of remotely sensed data. Chapter 3 discusses the application of ancillary data in the classification process. Chapter 4 details the Semi-Supervised approach in clustering analysis and the Second Stage reclassification post-classification approach. Chapter 5 details the methodology involved in analyzing the algorithm and describes the study sites selected. In chapter 6 the results of the research are discussed. Following the conclusions in chapter 7 is an appendix of source code of the programs.

Chapter II

Classification Procedures and Techniques of Remotely Sensed Data

2.1 Introduction

The following is a discussion of the current techniques in automated image classification procedures of remotely sensed data. The following is not an exhaustive review of the various classification approaches, but serves to inform the reader of some of the methods available. Clustering analysis is detailed in the latter part of this chapter. Although the thesis research involves a clustering analysis and a reclassification technique, a review of basic classification processes is established as useful background information for the reader.

2.2 General Discussion on Automated Classification

Remotely sensed data often involves multispectral information in the form of a digital image. Remotely sensed images capture the relative spectral reflectance for all the earth resources within the area covered by the scene. A multispectral image is comprised of multiple bands of data, each representing reflectance values for certain ranges of the electromagnetic spectrum. For example, a photographic false color composite is a typical multispectral image composed of three bands of data; typically the green band (.5 - .6 micrometers(μm)), the red band (.6 - .7 μm), and the near-infrared band (.7 - .9 μm). Different earth resources in an image can be described by certain combinations of spectral reflection values from the individual bands. Through automated classification techniques with computers, all the earth resources in an image can be classified into the appropriate land cover class. Automated classification according to multispectral information is called spectral pattern recognition (Lillesand and Kiefer, 1987). This chapter details some of the methods of spectral pattern recognition used in research today.

A multispectral image is comprised of picture elements, pixels, which are quantitative representations of the spectral reflection of an area on the earth's surface. The size of the area depends on the resolution of the sensor. Spectral reflection is a relative measurement of the reflected energy from an earth resource at various wavelengths and is recorded as a digital value by the sensor. An image pixel is a vector involving spectral reflectance values from each spectral band, registered to represent the same area on the ground. An image pixel is a vector of n dimensions. N is defined as the number of spectral bands available to comprise the image. An image pixel can be called a pattern vector, or feature vector, and is plotted in n - dimensional measurement space (Figure 1). Measurement space, or vector space, is described by the spectral bands of data, each band representing a different dimension in vector space. There are two spectral bands illustrated in figure 1. Throughout the text two dimensional figures will be diagramed, such as figure 1, which contain two axis, each symbolically represented by a band number. In this instance band 3 could represent the red band and band 4 the near-infrared band.

Each land cover category within a digital image can be represented by a spectral class having certain spectral characteristics. The spectral characteristics of each class can be depicted by selecting samples representing that category in the image. These samples or feature

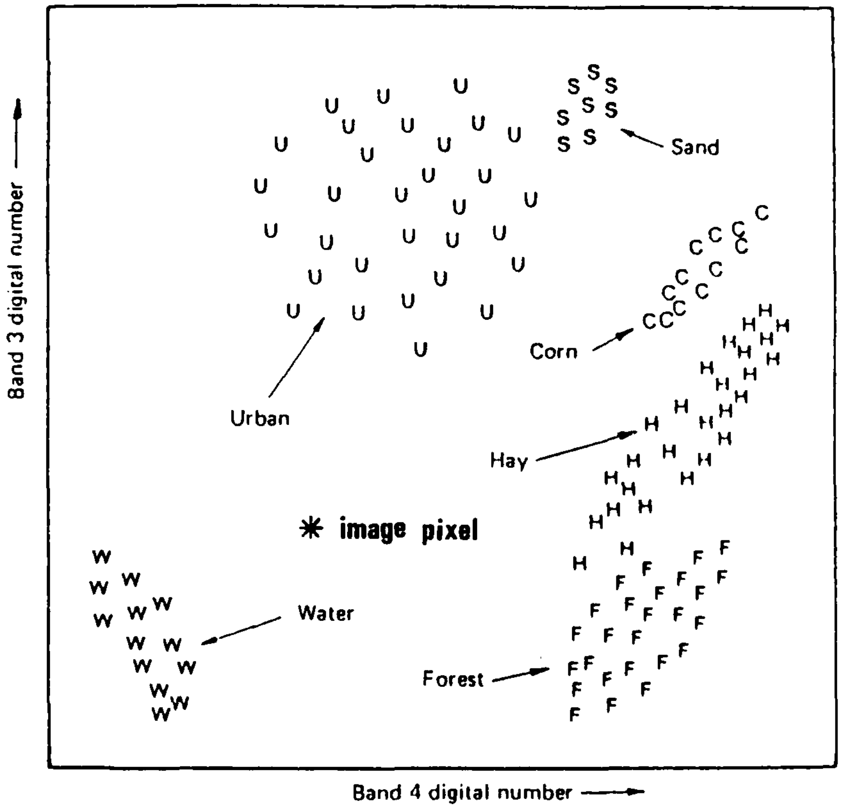


Figure 1. Spectral clusters in 2-dimensional measurement space. (From Lillesand and Kiefer, 1987.)

vectors are the spectral signatures of the category, and as vectors can be plotted in measurement space. Patterns of the same class tend to have similar spectral attributes and cluster in vector space forming clouds of feature vectors (Figure 1). A cluster or several clusters represent each land cover type in an image. These clusters occupy certain areas in multidimensional space, segmenting measurement space into regions that represent the various land cover types. The intent of spectral classifiers is to segment the pixels, vectors, into the appropriate regions of vector space, labelling it as a member of that land cover class.

Before the image can be classified, the clusters of feature vectors for each land cover type must first be identified, so that the spectral regions can be defined for each category. This is accomplished in the training or learning phase by identifying sample patterns for each of the land cover types in an image. The sample patterns can be acquired in a supervised or unsupervised manner. In the supervised process the computer is guided by the analyst to identify the spectral characteristics of each land cover class. In the unsupervised approach, a computer algorithm identifies the different spectral categories in the multispectral data, with little or no input from the analyst.

The remaining sections of this chapter will detail some of the standard procedures of automated image segmentation: the supervised and unsupervised learning phases in classifications; three image classification decision rules: the box-filter, minimum distance to mean, and the maximum likelihood rule; and also review some algorithm developments for unsupervised clustering classifications.

2.3 Supervised and Unsupervised Classifications

The statistical method is one of the standard approaches in pattern classification. It is assumed that the cluster distributions, for each class, in measurement space can be described by statistical parameters: the mean vector, the number of standard deviations from the mean, and the covariance matrix. The mean vector consists of mean values, in each spectral band, that represent the average spectral response for that land cover class in that band. The means from all the spectral bands identify the centroid of the cluster in measurement space. The standard deviations describe the variance of the data around the means in each of the bands. And the covariance matrix describes the dispersion of the pattern vectors around the centroid; it thus details the shape of the cloud.

The multidimensional distribution of the cluster is assumed to be modeled as a multivariate gaussian (normal) distribution. Since the shape of the distribution is assumed to be known, only the parameters of the distribution need to be determined and stored to describe the cluster. These parametric statistical assumptions are implemented in classifications to segment vector space into spectral class regions. These parameters can be determined through supervised or unsupervised techniques.

The supervised approach requires that the analyst delineates training areas for each land cover category, from which the computer obtains training samples for the class. From these feature vectors the distribution parameters are computed to represent the spectral class statistically. To train the computer, polygons are placed over areas of the image, on the computer monitor, that were considered ideal training areas, representing the spectral identities of the land cover. These sites are considered ideal because of their spectrally homogeneous appearance; and through field checks and verifications with other ancillary information. Through this process the analyst is able to describe the spectral characteristics of the clusters that represent the different cover types.

The unsupervised approach segments measurement space into uniquely defined spectral classes, by algorithms that identify these clusters with little guidance from the user. The clusters are then utilized to classify the image. The user then determines the utility of each of these clusters by studying the areas classified in the image.

There are many approaches to unsupervised clustering. Some methods identify the statistical parameters, for each spectral class, to be implemented in a statistical classification program to segment the image. Statistical classification programs are guided by certain decision rules to classify the image, discussed in the next section. Other approaches simultaneously classify the image as the pattern vectors are aggregated into their clusters. These techniques are discussed in further detail in future sections of this chapter.

2.4 Classification Decision Rules

The basic decision rules employed to statistically segment digital images in the world of remote sensing are the box-filter (parallelepiped), minimum distance to mean (Euclidean Distance) and the maximum likelihood classification scheme. To facilitate clear explanations of the following classification methods, only two dimensions of the data will be depicted in the illustrations. Normally all the spectral bands of the image are implemented in the classification process.

These three decision rules require the statistical parameters for each class as input from the analyst, obtained during the learning process. These parameters describe each of the representative land covers in the image, detailing the clusters that occupy certain regions in measurement space. Sequentially each pixel in the image is plotted in vector space and compared to the training clusters. Each pixel is categorized into the class to which it is most similar according to the following decision rules.

2.4.1 Minimum Distance to Mean Decision Rule

A minimum distance to mean decision rule (Lillesand and Kiefer, 1987; Nelson et al., 1981) bases the classification of each feature vector on the Euclidean distance between the centroids of each cluster to the feature vector. After the distance between the unknown feature vector and every training cluster's centroid are calculated, the unknown vector is classified into the category to which it is closest.

A disadvantage with this algorithm is that it is insensitive to the different degrees of variance in the training data. The image pixel, labelled 2, in figure 2 is classified as sand according to the minimum distance to mean rule, even though the pixel should be classified as urban according to the variance of the data. This could be corrected by using a statistical distance metric instead of a Euclidean distance function. A statistic

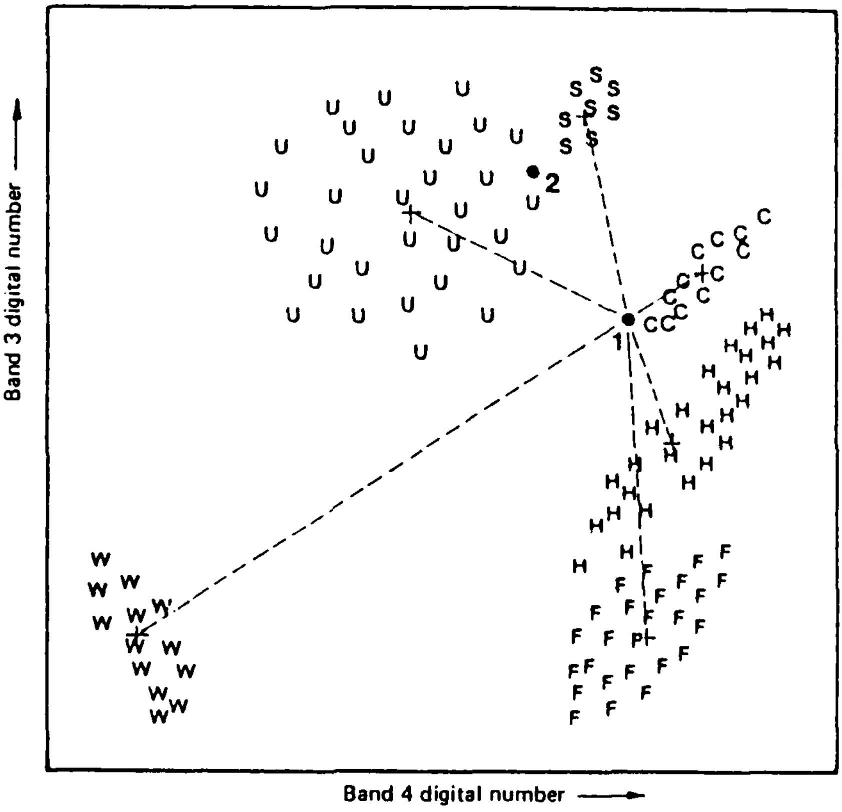


Figure 2. Minimum distance to means decision rule. (From Lillesand and Kiefer, 1987.)

distance function is a measurement that considers the variance of the distributions when calculating the distances between the centers of the clusters. Unfortunately, many statistical separation measures, such as Swain-Fu (Maktav,1985; Armstrong,1977), pairwise divergence (Pearson,1977), Jeffries-Matusita (Maktav,1985; Swain and Davis,1978), and the normalized distance between the mean (Swain and Davis,1978) are criteria to measure the distance between two clusters, not the distance between a point and a cluster as would be required for a minimum distance to mean rule. The Mahalanobis distance (Duda and Hart,1973), however, appears to be a metric that determines the statistical distance between a point and a cluster:

$$r = (x - \mu)^t \ \Sigma^{-1} \ (x - \mu).$$

This distance measurement implicates the dispersion of the cluster, by including the inverse of the covariance matrix. The Mahalanobis distance measure is a complex calculation and defeats one of the advantages of a minimum distance to mean classifier, which is the speed and simplicity of the computations.

2.4.2 Box Classification Decision Rule

A box classification filter (Lillesand and Kiefer, 1987; Story et al., 1984) bases the classification decision on the ranges of the multispectral dispersion of the training clusters in each of the spectral bands. The unclassified feature vector must fall within the limits of the maximum and minimum training values, in each spectral band, to be a member of that spectral category. Unlike the minimum distance to mean approach, the box filter is sensitive to the variance of the data by considering the dispersion of the spectral values in each band. The high and low values in each band create a box around the center of the cluster (Figure 3). The unclassified pixel must fall within this box to be classified as a member of that class. The box becomes a parallelepiped when more than two dimensions of data are involved in describing the training set statistics. Therefore, this classification scheme is also titled the parallelepiped classifier.

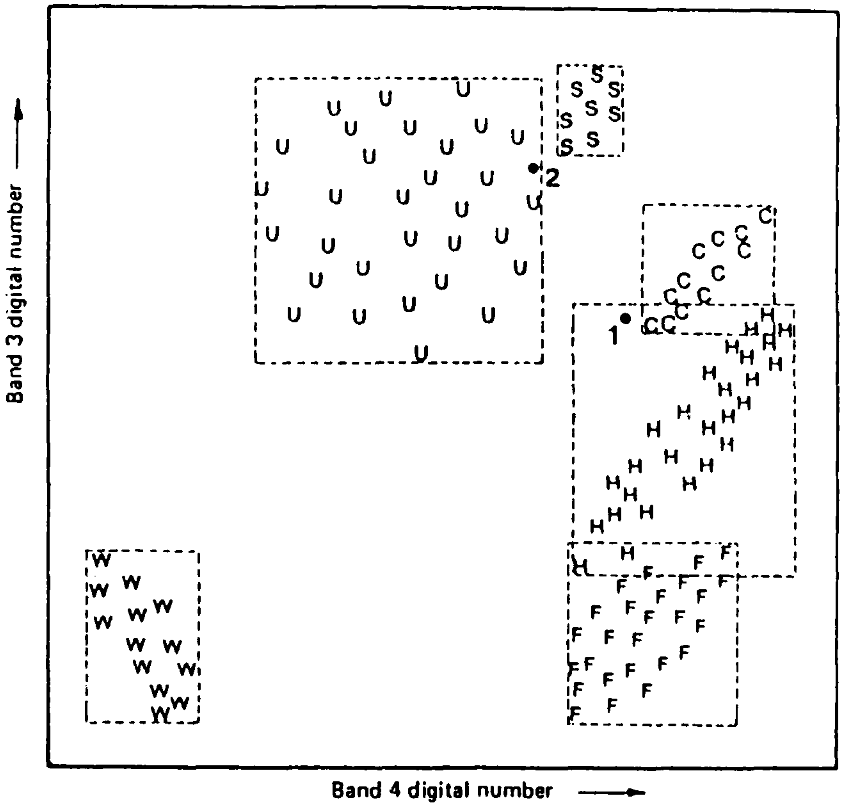


Figure 3. Box-filter decision rule. (From Lillesand and Kiefer, 1987.)

If the unknown pixel falls outside the limits it is considered unclassified. This often results in a large percentage of the image remaining unclassified. In Story et al. (1984), a box-filter classification resulted in 79.9% of the image remaining unclassified. There was, however, a high degree of accuracy within the areas classified, 86.44%. This indicates that the box regions describe the spectral characteristics of the spectral classes particularly well but more training set data is required to classify the remaining areas of the scene. A box-classifier works well except for overlapping classes, where two different regions overlap (Figure 3). Misclassification often results when an unknown vector lies within an overlapping region because the classifier assigns the pixel into the first region it fits. Therefore, the order in which the box filters are inspected biases the results. The overlap is caused by sample distributions that are similar in spectral characteristics and also highly correlated. Distributions having a high covariance, are poorly charact^erized by the parallelepiped decision regions (Lillesand and Kiefer, 1987). Covariance is the tendency of a pattern vector to vary similarly in two or more spectral bands; thus, cluster distributions can appear elongated and often slanted. (the H' vectors in figure 3). Positive correlation involves a dependence in the training data in which high digital values in one band are associated with high digital values in another. Negative correlation takes place when the inverse occurs, high values in one band tend to be associated with low values in the other. Negative correlation is illustrated by the 'W' class in figure 3.

Classification rules based on decision regions described by parallelepiped volumes are not sensitive to spectral data that is highly correlated. Unfortunately, remotely sensed data often exhibits covariance. The parallelepiped is able to describe such a situation much better if the multidimensional rectangle was modified into a series of stepped rectangles (Figure 4). Although this may alleviate some of the problems, another type of decision rule may be more adept at describing spectral patterns of this nature. Such a rule is a maximum likelihood classifier.

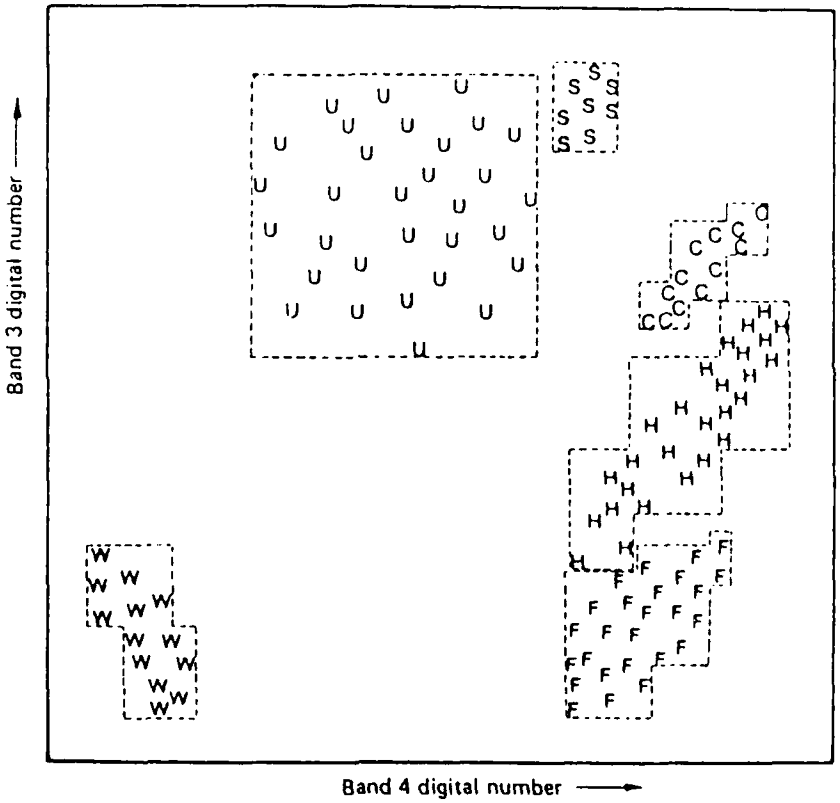


Figure 4. Modified box-filter. (From Lillesand and Kiefer, 1987.)

2.4.3 Maximum Likelihood Decision Rule

The maximum likelihood classification (Lillesand and Kiefer,1987; Story et al., 1984) is based on statistical parameters such as the mean vector and covariance matrix, to estimate the training distributions for each land cover class. This classifier can quantitatively evaluate both the variance and covariance of the spectral data in vector space because of the inclusion of the covariance matrix in the decision rule.

The distribution of the pattern vectors in the training clusters, for each class are assumed to be gaussian in shape (smooth curve normal distribution). With this assumption, the mean vector and covariance matrix can adequately describe the cluster, as was previously explained in section 2.2. With these parameters, decision regions, for each class, are quantitatively described by a discriminant function. Discriminant functions, unlike the parallelepiped are probability density functions that are very good statistical approximations of the shape of the sample distribution in multidimensional space.

With the probability density function of any cover class, we may compute the statistical probability of an unknown feature vector, plotted in measurement space, of being a member of that spectral category. In figure 5 the discriminant functions for a few land cover types are illustrated in two dimensional measurement space by a 3 dimensional surface. The vertical axis indicates the probability of a pixel being a member of a class. The closer the feature vector plots to the center of the distribution, the more probable that it is member of that class. To classify an unknown pixel according to the maximum likelihood rule, the algorithm calculates the probability of that pixel being in all of the land cover classes represented by training statistics. The feature vector would then be assigned to the most probable class, the one with highest probability. This procedure classifies all the pixels in an image, unless a threshold is established for the minimum probability that must be satisfied before a feature vector is classified (for further information on thresholds see section 2.4.4).

Each discriminant function defines lines of equal probability around the cluster center. In multidimensional

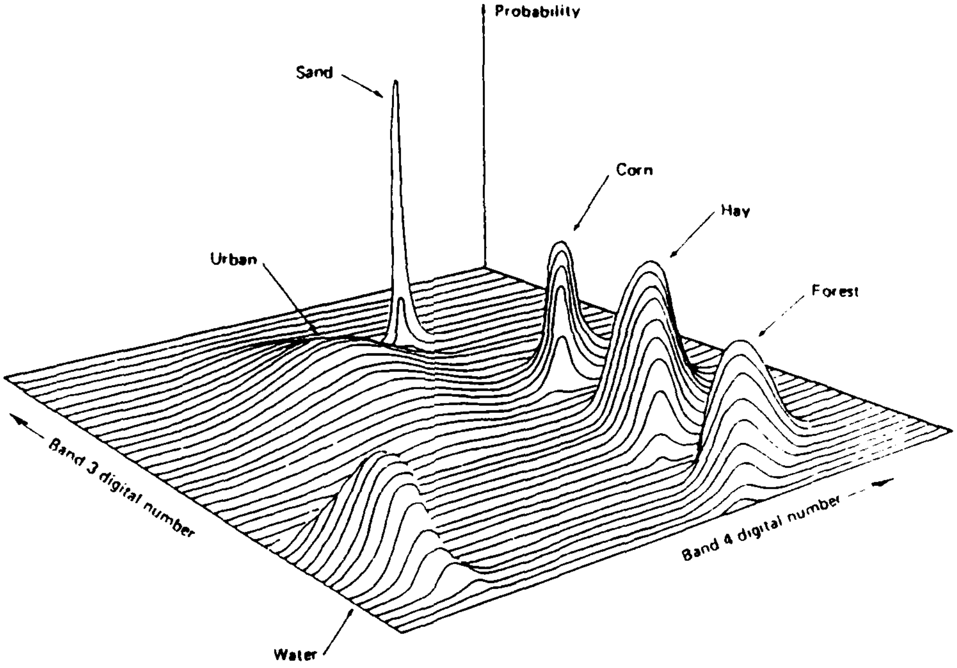


Figure 5. Spectral class discriminant functions. (From Lillesand and Kiefer, 1987.)

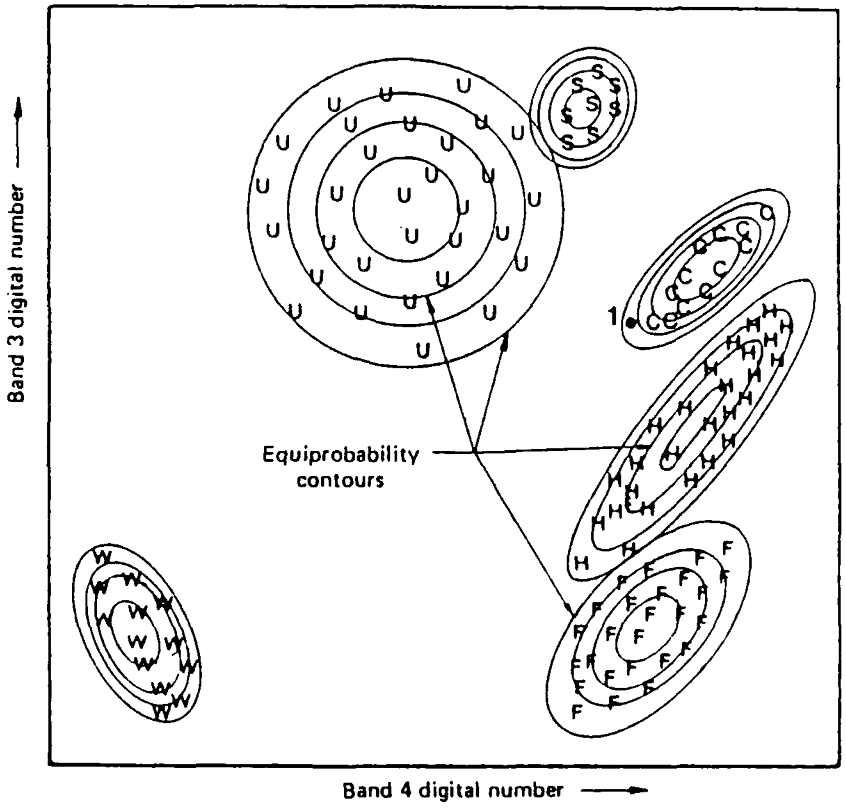


Figure 6. Equiprobability contours. (From Lillesand and Kiefer, 1987.)

space these ellipsoidal contours are hypersurfaces (hyperellipsoids) that are located at equal statistical distances around the center. A two dimensional diagram illustrates these 'equiprobability contours' (Lillesand and Kiefer, 1987) (Figure 6). These contours demonstrate the sensitivity of the maximum likelihood approach in representing correlated data as compared to the rectangular regions depicted by the box-filter.

2.4.4 Maximum Likelihood Classifier - Quantitative

The following section details the multivariate statistical analysis of the maximum likelihood method in a more quantitative sense, according to Swain and Davis (1978). Its intent is to inform the reader of the computational analysis that is involved in the classification of a given pixel in an image.

Discriminant functions are probability density functions for each land cover class, that are employed in a maximum likelihood decision rule to classify an image. If functions $g_i(X)$, $j = 1, 2, .. m$, are a set of m discriminant functions, one for each decision region, then the following decision rule applies:

Classification rule: Let ω_i denote the ith class. Decide that $X \epsilon \omega_i$ if and only if $g_i(X) \geq g_j(X)$ for all $j = 1, 2, .. m$.
(Swain,1977).

It reads: X belongs to class i only if the probability of X being in i, estimated by the discriminant function, is greater than the probability of it being in one of the other classes.

In a univariate case discriminant functions for class i are given by:

$$p(x|\omega_i) = \frac{1}{(2\pi)^{1/2}\,\sigma_i} \exp\left[-1/2 \frac{(x - \mu_i)^2}{\sigma_i^2}\right]$$

where exp[] = e raised to the power indicated.
μ_i = is the mean value of measurements in class i.
σ_i^2 = is the variance of the measurements in class i.

μ_i and σ_i^2 are the parameters to be stored that will define the cluster for each class. Unbiased estimators for these terms are:

$$\hat{\mu}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} x_j$$

$$\hat{\sigma}_i^2 = \frac{1}{q_i - 1} \sum_{j=1}^{q_i} (x_j - \hat{\mu}_i)^2$$

(Fruend,1971)

where q_i is the number of samples in class i.
x_j is the jth sample.

Therefore, the estimated probability function for class i is then;

$$p(x|\omega_i) = \frac{1}{(2\pi)^{1/2}\hat{\sigma}_i} \exp\left[-1/2 \frac{(x - \hat{\mu}_i)^2}{\hat{\sigma}_i^2}\right]$$

The univariate (one dimensional) case can be expanded into a multivariate probability density function. But before this is demonstrated, bivariate terms should be discussed to indicate some of the multidimensional terminology.

The two dimensional bivariate normal density function is given by a cumbersome equation (Swain and Davis, 1978), in which the parameters μ_{ij} and σ_{ijk} are calculated for each set of two dimensional training set statistics to describe the distribution.

where μ_{ij} = is the mean value of the data in channel j, for class i.
σ_{ijk} = is the covariance between channels j and k, for class i.

The two dimensional case can be expanded into n-dimensions and simplified in expression by vector/matrix notation. N is equal to the number of spectral bands.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix}$$

represents the data vector. Each x represents a registered digital value from each spectral band.

$$U_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ . \\ . \\ \mu_{in} \end{bmatrix}$$

represents the mean measurement vector for class i. Each μ represents the mean value for the dispersion of values in each band for class i.

$$\Sigma_i = \begin{bmatrix} \sigma_{i11} & \sigma_{i12} & \ldots\ldots & \sigma_{i1n} \\ \sigma_{i21} & \sigma_{i22} & \ldots\ldots & \sigma_{i2n} \\ . & . & & . \\ . & . & & . \\ \sigma_{in1} & \sigma_{in2} & \ldots\ldots & \sigma_{inn} \end{bmatrix}$$

represents the covariance matrix for class i. The diagonal elements define the variance of the samples in each band, while each off diagonal element details the covariance between two of the spectral bands for class i; together they describe the shape of the cloud in multispectral measurement space.

The multivariate density function is:

$$p(X|\omega_i) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp[- 1/2 \ (X-U_i)^T \ \Sigma_i^{-1} \ (X - U_i)]$$

where $|\Sigma_i|$ is the determinant of the covariance matrix Σ_i , Σ_i^{-1} is the inverse of Σ_i , and $(X-U_i)^T$ is the transpose of the vector $(X-U_i)$.

U_i and Σ_i are calculated by unbiased estimators.

$$\hat{\mu}_{ij} = \frac{1}{q_i} \sum_{l=1}^{q_i} x_{il} \qquad j = 1, 2, .. n$$

$$\hat{\sigma}_{ijk} = \frac{1}{q_i - 1} \sum_{l=1}^{q_i} (x_{jl} - \hat{\mu}_{ij})(x_{kl} - \hat{\mu}_{ik})$$

$$j = 1, 2, .. n$$
$$k = 1, 2, .. n$$

where q_i is the number of training samples in class i.

Now that the multivariate probability density function has been detailed, one term in the maximum likelihood decision rule is left to be discussed, the a priori probability. The a priori probability "is the anticipated likelihood of occurrence" for a certain class within the scene (Lillesand and Kiefer, 1987); regarding some factor, such as the percentage of area that is covered by each class, for example. The a priori probability is a weighting factor that permits land cover rarely evident in the image to be weighted less during the classification process than a cover type which is more prevalent.

> The maximum likelihood decision rule: Decide $X \in \omega_i$ if and only if $p(X|\omega_i)p(\omega_i) > p(X|\omega_i)p(\omega_i)$ for all j = 1, 2, .. m.
>
> (Swain,1978)

where $p(\omega_i)$ is the a priori probability for class i. The discriminant function $p(X|\omega_i)p(\omega_i)$ also incorporates a strategy to minimize the average loss over the entire set of classifications (The Bayes optimal strategy) (Swain and Davis, 1978).

The discriminant function that is expressed as a multivariate density function, with the a priori probability, now becomes:

$$p(X|\omega_i)p(\omega_i) = g_i(X) = \frac{p(\omega_i)}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp[-1/2\ (X-U_i)^T\ \Sigma_i^{-1}\ (X-U_i)\ ].$$

A transformation of this equation into a simpler form is

$$g_i(X) = \log_e p(\omega_i) - \log_e |\Sigma_i| - 1/2\ (X-U_i)^T\ \Sigma_i^{-1}\ (X-U_i).$$

Only the quadratic term must be recalculated for each class with every classification.

One difficulty is that this decision rule classifies all objects in an image. In remote sensing some spectral patterns do not belong to any of the classes described, because of an analyst oversight in creating training sets or because of insignificant training data. The classifier can be designed to reject a low probability of classification by a technique called thresholding. A user specifies a threshold that will guide the classifier in rejecting probabilities below the designated value.

This maximum likelihood decision rule is widely used in remote sensing applications. It can be used to classify an image according to discriminant functions derived from supervised training set statistics or training sets created by an unsupervised clustering technique.

2.4.5 Box Preprocessing Decision Rule

Of all the classification approaches we have discussed above, the box classifier was the most accurate in segmenting the image, often not classifying more than 50% of the image. The maximum likelihood approach, however, classifies all the pixels in an image, with reduced accuracy, if no threshold value is stipulated. Maximum likelihood approaches implemented without a threshold value have the ability to misclassify areas that have not been spectrally described with training set data. In Story et al., (1984) a parallelepiped Bayesian classifier was found to be the most accurate in classifying the whole image when compared to the previous approaches mentioned. In this study the box-filter was found to be 86.44% accurate in the areas that were classified, the maximum likelihood approach (bayesian) was 79.03% accurate in classifying the whole image. The parallelepiped-box classifier was 83.64% accurate in classifying all the pixels in the image.

The parallelepiped-box classifier (Story et al., 1984) performs a box-preprocessing operation on the data, and those pixels that either fall outside the parallelepiped regions or in areas where the boxes overlap are labelled 'undefined' and 'mixed' respectively. The box preprocessor allows the analyst to specify a confidence interval to adjust the ranges of spectral values to be used as limits within each band of data. The confidence interval involves the designation of a certain number of standard deviations from the mean. The user then has the option of reclassifying all of the 'undefined' or 'mixed' categories of pixels using a maximum likelihood classifier.

The box-preprocessing classification program available at ERSC performs a maximum likelihood operation on pixels considered 'mixed' and permits the user to designate how the 'undefined' pixels should be treated. All the classifications discussed in chapters 5 and 6, unless otherwise stated, are segmented using this program and all 'undefined' pixels are labeled unclassified.

2.5 Unsupervised Clustering Methods

Unsupervised classification involves computer algorithms which automatically analyze the spectral data and identify the various classes that are present. In the supervised process, clustered spectral data, representing information classes, were identified by the analyst in a supervised training selection process. Unsupervised clustering algorithms inspect all the spectral data of the entire image and aggregate similar feature vectors together to describe the land cover categories of an image. The meaning of the clusters is determined by the analyst after the image is classified.

Duda and Hart (1973) offered some basic reasons for using unsupervised classification algorithms: 1) the collection and labelling of large sets of sample data, from the image, in the supervised training process can be surprisingly costly and time consuming and 2) in the early stages of the investigation it may be valuable to gain some insights into the number and type of land cover categories present, to assist in selecting sites for training the computer.

There are numerous clustering schemes that have been developed to identify separable spectral classes in remotely sensed data. The research discussed in the following sections is not intended to be an exhaustive review of all algorithm developments, but it is a semi-comprehensive outline of clustering methods.

Clustering is a method of aggregating spectral information to represent the various information classes present in the data. A common feature of all spectral data, is that samples belonging to the same resource exhibit some similarity among themselves and a dissimilarity with those patterns belonging to another resource. For example, if all the pixels of an image were plotted in vector space, there would be numerous clusters of feature vectors present. These clusters are the aggregation of similar feature vectors and may represent useful information classes. Unsupervised clustering algorithms seek out and define these clouds of data. This concept is the basic premise behind the unsupervised separation of digital information into its different classes.

There are basically two broad categories of clustering techniques. One is based on the hierarchical (heuristic) approach and the other is based on criterion functions. The hierarchical approach is a method whereby sets of rules (intuitions) guide the clustering, whereas in the other approach, a criterion function is identified and optimized in each iteration of the algorithm to cluster the data. The hierarchical method can be divided up into two primary groups, the agglomerative approach and the divisive approach. In clustering according to criterion functions, there are numerous functions developed and optimized for clustering. The following sections will discuss many found in the literature review.

2.6 Similarity Measures

A cluster is described by an aggregation of feature vectors that are spectrally similar. To cluster data it would seem obvious that a similarity metric must be identified. In multidimensional measurement space a good measurement of similarity (or dissimilarity) between two samples could be the distance between the two vectors. If this assumption is true, then the distance between sample vectors of the same resource would be less than the distance between pattern vectors that are not of the same resource. Maktav (1985) implements an unsupervised classification algorithm with a Euclidean distance criterion. Euclidean distance is a point to point distance metric and is utilized to measure the distance between two n - dimensional pixels, image vectors, y and z.

$$E = \left(\sum_{i=1}^{n} (y_i - Z_i)^2 \right)^{1/2}$$

Euclidean distance is the combined sum of the distance between the points in the individual bands. Maktav's algorithm is based on grouping pixels into the appropriate clusters based on the Euclidean distance between them, as a measure of similarity. Maktav does not make it clear where the estimates for the initial means come from to begin the clustering process.

2.7 Criterion Functions

A few of the criterion functions that Duda and Hart (1973) discusses are the sum of squared error criteria, related minimum variance criteria and the scatter criteria.

The most widely used criterion function is the sum of the squared error (variance) criterion. It is defined by the equation

$$J_c = \sum_{i=1}^{c} \sum_{x \in X_i} || x - m_i ||^2 .$$

For cluster X_i the mean vector m_i is the best representation of the samples in the distribution in the sense that it minimizes the sum of the squared lengths of the 'error' vector $x - m_i$ (Duda and Hart,1978). The value J_c depends on the variance of the samples; how close they are aggregated. 'c' is the number of clusters that are identified. The optimal partitioning minimizes J_c, the total squared error incurred by representing n samples x_1 ...

x_n, by c cluster centers m_1 ... m_c. Clustering of this type is called minimum variance partitioning.

The related minimum variance criterion eliminates the need for mean vector values found in the calculation of the minimum variance partitioning.

$$J_c = 1/2 \sum_{i=1}^{c} n_i s_i$$

where

$$S_i = \frac{1}{n_i^2} \sum_{x \epsilon X_i} \sum_{x' \epsilon X'_i} || x - x' ||^2$$

This criterion involves Euclidean distances between samples in the individual clusters as a measure of similarity; J_c is extremized (minimized) when the distance between the samples in the distribution are minimized.

The scattering criteria are a class of criterion functions based on the scatter matrices used in multiple discriminant analysis: scatter matrices, trace criterion, and determinant criterion.

The scatter matrices are equations representing the within-cluster scatter and the between-cluster scatter.

The within- and between-cluster scatter depend on the partitioning that takes place; minimizing the within scatter will tend to maximize the between cluster scatter.

The other two criteria that involve the scatter matrix of the distribution, deal with the size of the cluster. The trace and the determinant criterion entail minimizing the scalar measurement of the size of the scatter, having an implication on the variance of the distributions.

2.8 Iterative Optimization in Clustering

For optimal partitioning of the data into unique spectral classes the above criterion functions must be extremized. One approach to optimal partitioning is iterative optimization (Duda and Hart,1973). The process involves separating the data into initial partitions, and moving samples from one partition to another in order to optimize the values of the criterion function describing the clusters. This process has been related to 'hill climbing' in general, different starting points (initial partitions) can lead to different solutions (Duda and Hart,1973).

In the literature, many clustering techniques involve multiple iterations, such as in Armstrong (1977). The algorithm takes some initial guesses as to where the initial cluster centers are located and pattern vectors from the image are then merged to the nearest cluster. After each iteration the means are recalculated according to the new distributions. The process is repeated until a stable assignment of pixels is achieved, which is attained when the mean values between two consecutive iterations fall within a tolerance determined by the user. There is no guarantee that the clusters would not overlap in measurement space. So a distance measurement was devised to examine the separability of the derived clusters.

$$D_{ij} = \frac{1}{n} \sum_{k=1}^{n} \frac{(x_{ik} - x_{jk})}{\sigma_{ik} - \sigma_{jk}}$$

D_{ij} is the statistical distance between the clusters i and j. 'x' and 'σ' are the means and standard deviations for these clusters, respectively, and 'k' indicates the different spectral bands of the image. ISODATA (Story et al., 1984), ISOCLS (Werth, 1981) and CLUSTER (Colwell, 1984)are acronyms for clustering algorithms that take on the same basic approaches as that discussed above by

Armstrong; clusters grow around seedpoints (initial cluster centers) by aggregating pattern vectors to the nearest cluster. This style of clustering has also been called the 'K means' method (Gowda, 1984; Lillesand and Kiefer, 1987).

The Coalescence Clustering Algorithm (Ince, 1981) is different from previously mentioned techniques in that data points are clustered based on the attractive force between them. The clustering takes place in feature space, that is the algorithm operates on a multidimensional histogram array of all the image pixels for segmenting the data points into spectral classes. The attractive force is a gravitational one with a range limitation on attraction of ±r cells in each dimension. 'r' is a parameter that the user must specify, and identifies the range of the neighborhood to be considered for the force calculations.

$$F_i = \sum_{k=1}^{m} \frac{h_i h_k}{S_{ik}^2} \qquad i \neq k.$$

h_i is the frequency, number of occurrence, for that cell, or pattern vector i. S_{ik} is the Euclidean distance between cells i and k. F is the net force on cell i due to all the cells in its neighborhood. K is the total number of data points in the neighborhood of cell h_i.

$$a_i = F_i / h_i$$

a_i is the acceleration on cell i due to all the cells in the neighborhood. If the acceleration is greater than a designated threshold value then the mass (frequency) in that cell moves into the neighboring cell in the direction of the acceleration, emptying the original cell. At each iteration the histogram array is modified and the merging process ceases when all non-empty cells are beyond the neighborhood r of each other.

Another different clustering technique involves a convexity testing method (Vasseur and Postaire,1980). The convexity testing method is a mode (peak) detection procedure, since it assumes that in a multimodal probability density function (such as in a multidimensional histogram of all the digital data in the image) each mode corresponds to one clust3r. Modes can be characterized by the convexity of the underlying probability density function.

Mode seeking procedures identify convex areas of samples which are considered to be the nuclei of the individual classes, of which the remaining data are merged.

Similar clustering schemes using multi-dimensional histograms have been researched by Goldberg and Shlien (1977), Wharton (1983) and Leboucher et al. (1976).

In Goldberg and Shlien (1977) a 4 dimensional histogram (4 spectral bands) is used in clustering the data. It is a table listing the frequencies of the pattern vectors of the image. Peaks in the histogram (vectors with a high frequency of occurrence) are assumed to be associated with different resources. The method involves isolating these peaks and merging the associated vectors to create unimodal clusters. A threshold is chosen to divide the intensity vectors of the image into two sets, those that occur with at least this threshold frequency, and those that occur with less. The former group is separated into clusters corresponding to the peaks. The latter group of vectors are assigned to the closest peak.

The peaks are delineated by grouping those pixels, of the first group, that occur at a frequency equal to or greater than the threshold according to their connectedness. Pattern vectors are said to be connected if the intensity in each of the bands do not differ from one another by more than one. After the peaks are identified, the remaining pixels in the second group are assigned to a cluster according to the ‘connected to a cluster’ rule. Any vector remaining unclassified is then merged with a cluster by a Euclidean minimum distance rule.

Wharton (1983) follows four steps in his approach in identifying the peaks in the multi-dimensional histogram. First, the algorithm computes a list of neighboring vectors in the histogram. Second, after examining the list of neighbors a directed link or pointer connects each vector and its immediate neighbor having the maximum positive density gradient. The gradient is calculated by the difference in the two vectors divided by the distance between the two vectors. The distance measure is the city block measure

$$\text{Dist}\ (x,y) = \sum_{i=1}^{k} |\ x_i - y_i\ |.$$

k is equal to the number of spectral bands. In the third step, the cluster centroids are identified by locating the peaks in the histogram. A peak is defined to be a vector whose frequency count is greater than the frequency count of all its neighbors. A peak will form a nucleus of which other vectors can be grouped. In the fourth step, the directed links, discussed before, are used to merge the remaining non-peak vectors to the appropriate cluster. The directed links in theory should form a path leading to the centroid of the cluster. Because these paths are directed toward higher density (frequency) neighbors, adjacent clusters should be separated by low density valleys. All vectors in the paths should be assigned to the proper cluster thus delineating unimodal distributions from the histogram.

2.9 Hierarchical Approach

As was stated before, the hierarchical approach to clustering can be divided into two groups: the agglomerative and divisive methods. In the agglomerative method, each pattern vector is considered a cluster center and based on certain rules they are interactively merged together to form larger clusters. Whereas, the divisive method starts with a single cluster, a (multi-dimensional) distribution of all the data in the image, and splits it into smaller more meaningful clusters. The computation involved in the agglomerative procedure is usually simpler; however, if there are many pattern vectors and the analyst is only interested in separating a few clusters, the divisive method would be more efficient (Duda and Hart,1973).

2.9.1 Agglomerative Method

The multiple iteration clustering algorithms mentioned above by Armstrong (1977) and others, such as ISODATA, ISOCLS, and CLUSTER, can all be considered agglomerative techniques. These clustering methods start out with initial estimations for the cluster centers and group samples to these clusters according to similarity based on a distance measurement. Two more examples of agglomerative methods are the nearest neighbor and the furthest neighbor algorithms (Duda and Hart, 1973). The theories of these two techniques are explained in the following paragraphs.

The nearest neighbor algorithm connects the nearest neighbor data points according to a minimum distance measurement function. Since only distinct clusters and vectors are linked, the resulting clusters are never closed loops but grow in an open-ended fashion, as a tree. Because the algorithm uses a minimum distance metric it generates what is termed a ‘minimum spanning tree’ (Duda and Hart,1973). Sometimes a few points are positioned such that their presence causes two clusters to be linked forming an elongated cluster, a ’chaining effect’ (Duda and Hart, 1973). This can be advantageous if the clusters are elongated.

In the furthest neighbor algorithm a maximum distance function is used between points and the growth of elongated clusters is often discouraged. This method creates clusters with all the samples connected, unlike the nearest neighbor algorithm that produces chains. The furthest neighbor approach increases the diameter of a cluster as little as possible with each clustering iteration. A diameter is defined as the largest distance between points in the cluster. "True clusters are compact and roughly equal in size" (Duda and Hart,1973).

Processes involving distance functions based on the two extremes, maximum and minimum, often tend to be sensitive to "mavericks", "sports", "outliers" or "wildshots" (Duda and Hart,1973). The average and mean distance functions discussed in Duda and Hart (1973) are natural compromises to this problem, in which the structure of the mean function permits it to be the simplest to compute.

Besides an iteration technique explained above, Armstrong (1977) also describes a 'chain algorithm' which implements a nearest neighbor agglomerative method. Samples are joined to clusters based on a distance measure for similarity. The distance between the points must meet an analyst-defined threshold vᵃlue before it can be considered a cluster.

2.9.2 Divisive Method

An example of a divisive approach is presented by Chandrasekhar (1983), in which a single cluster is continuously split with successive iterations. The furthest two points used in the measurement of the diameter of the initial cluster are chosen as cluster centers. Pattern vectors are then assigned to these two centers according to similarity measures. The cluster with the largest diameter is then split by the same procedure mentioned above. This process is repeated until the desired number of clusters is obtained. Euclidean distance was implemented to measure similarity.

2.10 SEARCH Algorithm

Most of the unsupervised clustering processes above simultaneously classify the digital images during the clustering operation. Very few methods, such as the SEARCH algorithm presented by Pearson (1977), cluster the data yielding statistical parameters describing the spectral classes. These parameters as described before, represent training information that can be implemented into a statistical classifier (such as maximum likelihood) to segment the digital image.

The SEARCH algorithm's approach to clustering is similar in nature to the agglomerative heuristic processes, but instead of merging individual vectors to cluster, a group

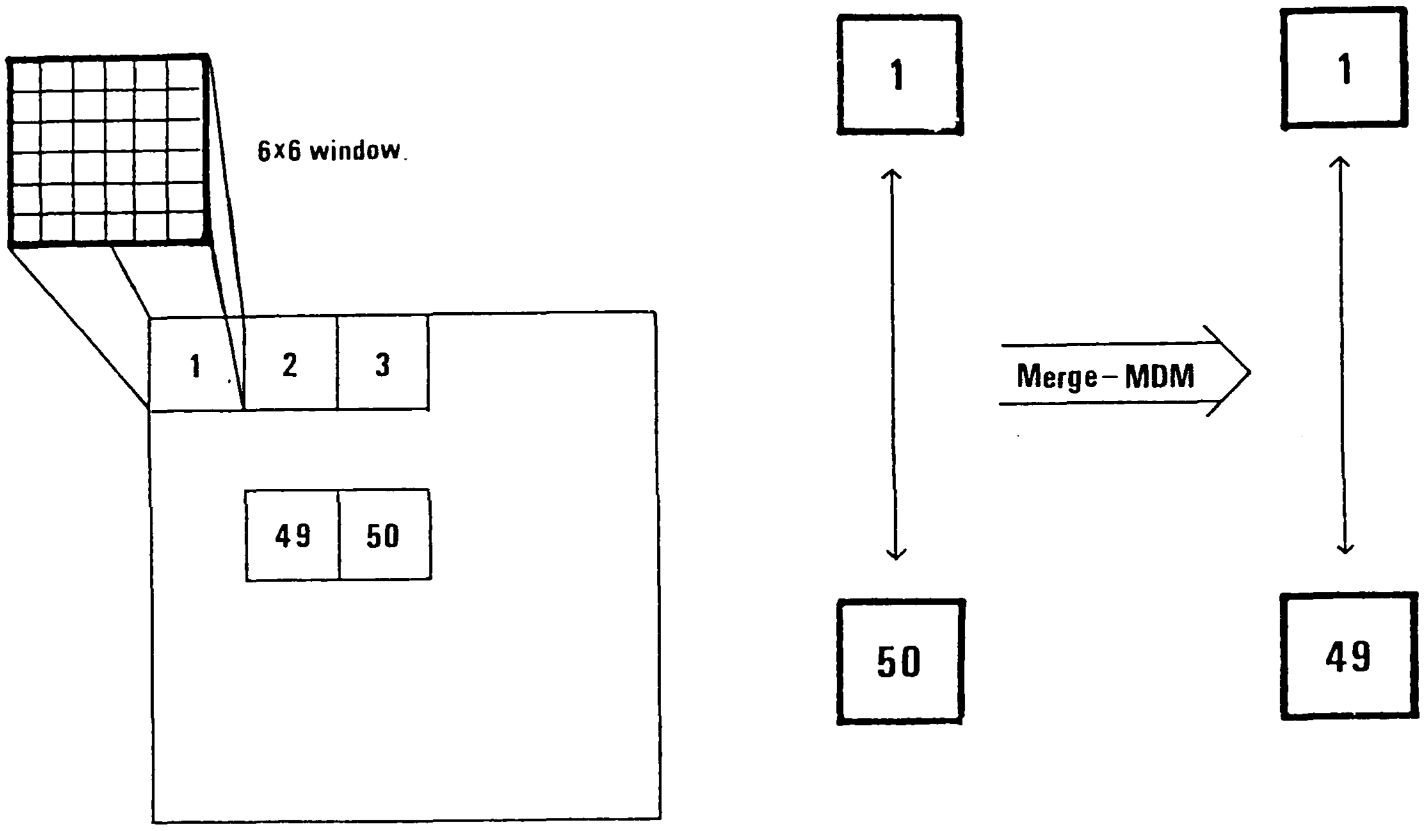


Figure 7a. SEARCH algorithm.

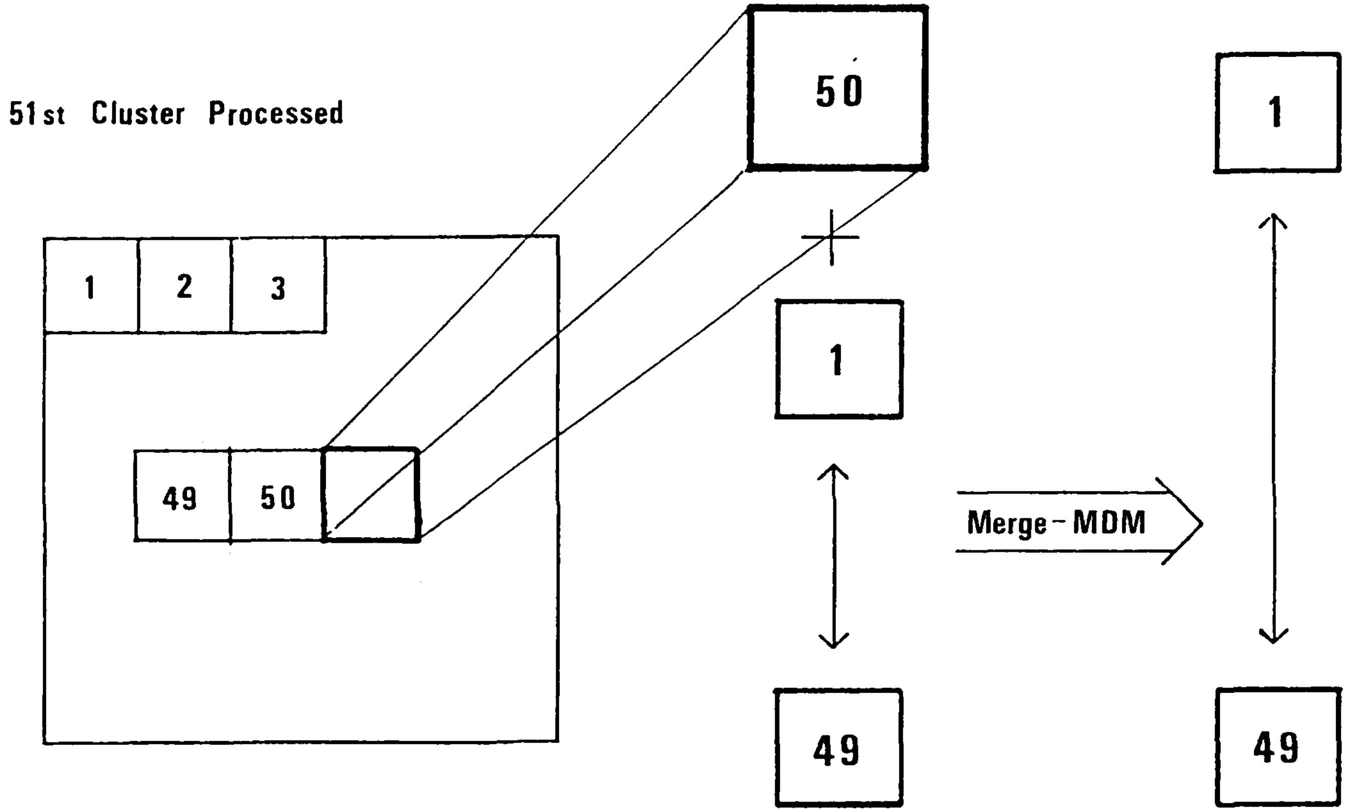


Figure 7b. SEARCH algorithm (continued).

of similar pixels is aggregated to a cluster center. Windows, 6 pixels by 6 pixels in dimension, are sequentially analyzed in the image as possible training samples (Figure 7a). Windows that appear to be homogeneous are stored as training samples, or signatures. Once the first 50 signatures are found the 2 clusters with the smallest pairwise divergence are merged reducing the total number by one. The next signature identified (Figure 7b) is considered the 50th cluster and again the 2 most similar clusters are merged according to a pairwise divergence metric. This process continues until all the 6 x 6 windows in the image are analyzed. The resulting clusters are merged down to the analyst-defined number of clusters by a pairwise divergence calculation. The pixels in a window are considered to be homogeneous by meeting the user designated limits on lower and upper bounds of the standard deviation for each spectral channel. The lower bound limit avoids extremely peaked clusters that may have high divergence when merged with another cluster. And the upper bound insures a homogeneous cluster.

2.11 FINDSET Algorithm

The FINDSET algorithm is the unsupervised clustering algorithm presently used at the Environmental Remote Sensing Center (ERSC) at the University of Wisconsin - Madision. It is based on the SEARCH algorithm detailed in the previous section. FINDSET identifies a maximum of 50 training sets, or clusters, in an image. The algorithm looks at 3 x 3 windows in the image to identify homogenous spectral clusters with a user defined variance threshold. The variance threshold places an upper bound on the maximum sum of the variance of the spectral bands of the data. If the sum of the variance exceeds this, it is not considered homogeneous. The clusters are merged by the following statistical distance metric

$$d = \sum_{n=1}^{k} \left\{ \frac{(x_1 - x_2)^2}{\sigma_1^2 + \sigma_2^2} \right\}^{1/2}$$

The n is the number of spectral bands. Symbols 'x' and 'σ' are the means and standard deviations of the indicated clusters. This measure is similar to other statistical similarity measures that have been discussed previously.

Possible Problems with FINDSET

Within the last few years the researchers at ERSC have recognized some possible problems with the FINDSET algorithm.

One is the potential inability to gather spectral statistics for all the land cover classes contained within an image, creating a bias with respect to the section of the image in which the initial 50 clusters are found. The 3 x 3 window begins searching for clusters in the upper left corner of the image. If the first 50 clusters are found in the upper portion of the image and this section contains one dominant land cover type, such as forest or water, then many of the training sets resulting could be dominated by the statistics from this cover type. This bias may prevent the creation of training sets for other land cover classes in an image, such as built-up areas or water, that may be located elsewhere within the image. It is also reasonable to assume that homogeneous clusters describing other land cover classes could be mixed with the forest training clusters, for this example, by the minimum distance calculation, contaminating the statistics. In short, there could be a bias towards the first 50 clusters identified and a possible creation of mixed training sets.

The above discussion about clustering bias involves the analogy stated in Duda and Hart (1973) about "hill climbing" (see section 2.8). Different starting points (initial partitions of the data) can lead to different solutions. "If an unfortunate sequence of samples is encountered, the error in classifying the unlabeled samples can drive the classification the wrong way"(Duda and Hart,1978).

Another problem associated with "hillclimbing", discussed above, is that an arbitrary scheme like this could result in the initial cluster centers being outside the clouds of the data points that represent the real sample (Armstrong,1977). This could result in mismerging, a contamination of the clusters with resources that have similar spectral characteristics. Story et al. (1984) found that mismerging between water and shaded forest is a possible scenario as a result of the misidentification of the initial cluster centers and subsequent mismerging of spectral data.

Another difficulty involves the resolution of the remotely sensed data classified. Resources that are less than 3 pixels in size, length or width, may be inadequately trained on by the clustering algorithm and representative training statistic never acquired. Spectrally homogenous 3 x 3 windows are delineated and analyzed as training information. If an earth resource is digitally represented by less than 3 pixels, it would not fill a 3 x 3 window and a good training sample may never be acquired. This could happen to roads, that are linear features, comprised of 1 or 2 pixels, in satellite images when the resolution of the data is 30 meters, for example. However if a color infrared aerial photograph was digitized at a spot size that would allow roads to be represented by 4 or more pixels, adequate training samples could be acquired in the FINDSET approach. This problem was also described in literature by Story et al. (1984) when he discussed a similar algorithm.

The issue of statistical independence between adjacent 3 x 3 windows, also referred to as autocorrelation between adjacent windows (Ahearn and Lillesand, 1986) is another discrepancy that may be present in the FINDSET algorithm. Ahearn demonstrated that autocorrelation is more prominent between sample windows touching at the sides rather than those touching at the corners. The FINDSET algorithm possibly reduces the validity of the statistical calculations by comparing adjacent windows that may not be independent samples.

Another difficulty that can be attributed to many unsupervised clustering algorithms pertains to the requirement of an a priori estimate of the number of clusters that should be found in the image. Unsupervised processes may require several attempts to request the appropriate number of classes that produce an informative classification (Armstrong,1977).

2.12 Choosing a Classification Algorithm

An appropriate concluding section on the literature review of classification techniques might be a discussion on the advantages and disadvantages of a classification algorithm that should be known before the analyst chooses one for a specific application. There are 4 points of concern that were outlined by Story et al. (1984): 1) the accuracy that can be attained with a given technique, 2) the ‘ease of use’, in other words, how "user friendly" is the system 3) the amount of input information that the analyst must supply, such as training set statistics or a variance threshold and 4) the ‘CPU’ time, or the time it takes to run the program on the digital image of interest.

2.13 Post-classification of Remotely Sensed Data

In the literature, post-classification processes seemed to involve two applications: 1) the reassessment of classified images for monitoring change detection and 2) post-classification spatial smoothing algorithms. Neither approach is similar to the automated reclassification according to ancillary data as developed in the Second Stage reclassification approach discussed in section 4.6. Nor did any of the research involving post-classification operations, in the literature, involve a similar process.

In monitoring change detection, post-classification involved multitemporal analysis (Goldberg et al.,1982;

Weismiller et al., 1977; Wickware and Howarth,1981). Digital images from two different dates were independently classified using automated techniques to segment the identical number of classes. In a post-classification approach the two classified images were compared and the changes recorded. Most of the time the classes monitored were visually analyzed pixel by pixel, but other times an automated comparator was implemented. The comparator was a processing program that compared the two classified images producing resulting images that indicated the areas of change. Also contingency tables identified the changes that occurred and the nature of the changes in the output.

Post-classification spatial filtering has been implemented to smooth classified images, thereby possibly increasing the classification accuracy (Moreira et al., 1986). In some areas land cover classifications resulted in the misclassification of some pixels in what should be homogeneously classified regions. A spatial filtering operation reclassified many of these anomalously labelled pixels, reducing the overall misclassification. In spatial filtering operations, windows of the data are accessed in various sizes and numerical calculations are applied. For classified images a majority filter is often used. This type of algorithm considers all the classified pixels in the window and determines which label has the highest frequency of occurrence and assigns that value to the center element of the window. This operation continues throughout the image until the whole image is processed. Spatial filtering implemented by Moreira et al. (1986) computed a threshold for each window that required the majority label in the window to meet a required frequency threshold before the central pixel was changed. Spatial filtering, in general, deemphasizes the high frequency components of the classified image, often referred to as a 'salt and pepper appearance' or noise (Lillesand and Kiefer, 1987). Low frequencies are deemphasized yielding an image that is smoothed in nature.

2.14 Hybrid Classifiers

The Semi-Supervised clustering analysis can be considered a hybrid approach to segmenting multispectral data. In the literature hybrid clustering and classification algorithms have been employed in numerous research projects.

Hybrid classifiers are classification techniques that involve both supervised and unsupervised analysis or any combination of different classification methods. Moreira et al. (1986) used unsupervised spectral classification technique with a post-classification spatial filtering process to increase the accuracy of wheat area estimates. In Story et al. (1984) the ISODATA algorithm requires the analyst to identify the initial cluster centers, in a supervised fashion, before it groups the remaining data to similar spectral classes in an unsupervised manner.

Swain and Davis (1978) summarizes a hybrid procedure for analyzing remotely sensed data. The process begins by using an unsupervised classifier to enhance the raw data by deriving some useful classes. The results of this classification assist in a supervised training of the area. After a supervised acquisition a cluster separability analysis is implemented to derive unimodal training set distributions. The statistical training information is then implemented into a maximum likelihood classifier tc segment the digital image.

The first stage of the Semi-Supervised Two Stage classification technique involves a clustering algorithm, whereas the second stage involves the application of ancillary data in a post-classification reclassification routine to increase the accuracy of the spectral classification. Ancillary information has been implemented in image segmentation with various degrees of success.

Chapter III

Ancillary Data in the Classification Process

3.1 Introduction

This chapter discusses the application of ancillary information in the automated classification process. Ancillary data can be defined as additional information that is accessed to assist in making or justifying a quantitative decision or analysis. Ancillary data that is implemented into the classification process is any digitally amenable data that can be helpful in describing the land cover classes of an image. There are many different types of digital ancillary data. For example, texture, digital topographical information (DEM), ratioed images, vegetation index ratios, digitized map overlays, digitized soil maps, and geographical information data bases, just to name a few. This chapter will detail the application of texture in the classification process, to support the reasoning for the author's reclassification approach discussed in the next chapter.

3.2 Texture

Texture is a descriptive attribute that is numerically quantified for applications in digital image processing.

Texture is an innate property of all objects that are around us. Visually, every surface has an arrangement of tonal values. The texture of surfaces can be qualitatively described as rippled, mottled, irregular, limited, striated, etched or any of a myriad of other terms. Essentially, the textural properties of objects are not independent of the tonal variations. Tone refers to the brightness or darkness of a surface. Texture can be defined as an arrangement of an elementary pattern, variation in tone, that is present over an area larger than the pattern itself. The photo interpreter relies on the combined principles of texture and tone to analyze aerial photographs. Texture is considered a useful descriptor in manual and automated land cover classifications, for just as surfaces with uniform spectral reflectance are considered to be objects, regions with homogeneous texture may also be considered objects.

Texture has been implemented in automated classifications to segment classes of interest of a digital image (Jensen and Toll, 1982; Shih and Schowengerdt, 1983; Hsu, 1978). Texture must be represented numerically to be employed in an automated classification process. Quantitatively, texture is the local spatial tonal variation within an image; also termed the coarseness of the data. A rough texture, high spatial frequency, involves large deviation of total variance; and smooth texture, low spatial frequency, minimal tonal variation in the data. These spatial frequencies are determined numerically with a texture algorithm. Texture algorithms are often area calculations indicating local tonal variance around an image point, based on a mathematical operation on a window of pixels. A window is an m by m array of pixels, where m is the size designated by the analyst.

3.3 Texture Algorithms

In the literature, there are basically two approaches to texture algorithm development: statistical analysis and the Fourier-based approach. In a study by Weszka and Rosenfeld (1975) it was concluded that statistical features perform much better than Fourier-based procedures for generalized land cover mapping. Since, the consensus from the literature is that resource textures are more appropriately modeled by statistical measures, the Fourier transform approach will not be discussed in this chapter.

Statistical Analysis of Texture

Statistic analysis of texture can be divided into two levels of development: 1) the first order statistics in spatial domain and 2) the second order grey level statistics in spatial domain.

First order textural statistics are derived from neighborhood calcılations. Local statistical values such as mean, variance and standard deviation can be computed for the pixels in a window that sequentially passes through the image. These statistical values produced for each window are assigned to the center pixel of the window, resulting in a textural image of textural measures for each pixel based on the tonal variation of its neighbors. Variance can be calculated for a matrix of pixels (window) by the following operation,

$$T = \sum_{i} \sum_{j} (s_{ij} - s)^2.$$

This formula is called 'variance with respect to the center' or 'variance with respect to the average', where s can either be the value of the center pixel of the window, or the average of all the matrix elements, respectively. 'i' and 'j' indicate row and column addresses for the neighboring pixels of the window.

Difference methods also indicate local textural properties. There are three different operations: horizontal, vertical and diagonal.

The horizontal algorithm compares the tonal variation on each side of a given pixel according to the given equation,

$$T_h (r,c) = | s(r,c-1) - s(r,c+1) |.$$

's' is the tonal signal at each row (r) and column (c) pixel address in the window, center pixel being s(r,c). The vertical method compares the variation above and below, relative to the image, a pixel with the equation

$$T_v\ (r,c) = |\ s(r+1,c) - s(r-1,c)\ |.$$

The diagonal approach compares the tonal variation of the neighbors at the corners of a given pixel by the equation,

$$T_{d1}\ (r,c) = |\ s(r+1,c+1) - s(r-1,c-1)\ |$$

or

$$T_{d2}\ (r,c) = |\ s(r+1,c-1) - s(r-1,r+1)\ |.$$

The diagonal computation is directional as well as diagonal. Unfortunately, the difference methods are often more of an indication of edges in a scene than texture.

Texture transforms are another approach for local measures that can be considered first order statistical operations (Hsu,1978; Irons and Peterson,1981). Many of the textural transforms proposed by Hsu and Irons are more statistically rigorous than the ones mentioned above. Hsu detailed 17 local descriptors in which four operations evaluated the four central moments (mean, standard deviation, skewness, and kurtosis) of a the distribution of grey levels in a 3 x 3 window. Various others were discussed, such as mean contrast grey level differencing among nearest neighbors, and a measure of the mean above and below 3 datum planes in the data: 50, 100 and 150. Irons developed several more descriptors on the basis of Hsu's works that involved differencing of the maximum and minimum values for the grey level distributions, and also the equation discussed above involving variance were embellished through normalizing and maximizing and minimizing operations which modified the equations.

The other group, termed second order grey level statistical measures, base their higher order operations on grey-tone spatial-dependence matrices, computed from various angular relationships and distances between neighboring resolution cell pairs in the image; also refereed to as nearest neighbor grey-tone spatial-dependent matrices (Haralick et al., 1973; Haralick and Schanmugam, 1974). Each texture feature is derived from these angular relationships; close related measures of the matrix's unnormalized frequencies quantized to 45° intervals. Appendixes in both of Haralick's publications have the algorithms for 28 textural features. Wiersma and Landgrebe (1976) review four of Haralick's texture measures in closer detail: angular second moment, contrast, correlation, and entropy.

3.4 Texture Implemented in the Classification Process

Texture is considered as another band of digital information when implemented with spectral data in generalized land cover classifications. For example, if 3 spectral bands and a texture image were used, the training sets statistics resulting would be represented in 4 dimensional measurement space. Spectral and textural information are then used simultaneously in a classification process that accesses the image pixel by pixel. This is where texture and spectral information differ regarding the relative resolution of each. While each spectral pixel value presents information about its own spectral reflectance, each texture value, of an individual pixel, is the result of an area calculation measuring the tonal variations of its neighbors. Each textural value is a consideration of an area of the spectral data and has a relatively larger resolution significance.

The neighboring pixels, composed of the window could span land cover class boundaries. Land cover borders can and are misrepresented by texture. As the window moves over the boundary, the combined textures from each land cover creates textural heterogeneity and often misleads the calculations causing a smearing of the textural information over the land cover interface. This is analogous to the mixed pixels in spectral data, but it depends on the window size and can be much more severe. Such 'boundary smearing' is not permissible for detailed digital land cover classifications when it is precisely the boundary between the cover types that the analyst is trying to discern (Jensen,1979). The smearing of the boundaries is a function of the size of the window. Hsu (1978) selected window sizes of 3x3 pixels rather than 5x5 for generating his final decision maps because of misclassifications along the land cover boundaries. In Jensen (1979), use of a variance coding calculation of texture distorted class boundaries and decreased classification accuracy, especially at the land and water interface. Problems in classifications could also result when a land cover class is smaller or narrower with respect to the window size (Shih and Schowengerdt, 1983).

Selection of the size of window is also dependent upon the resolution of the image. High resolution imagery can have more spectral variability per unit area than a similar scene with lower resolution. Therefore, there could be a significant amount of textural information in a small window in high resolution images as compared to the same window size in an image with a larger resolution ground cell. This is indicated in a study by Irons and Peterson, (1981). Using various textural measures, generated by textural transforms similar to Hsu (1978), Irons classified Landsat Multispectral Scanner (MSS) image data which has a resolution of 79 x 79 meters. Hsu classified digitized high and low altitude black and white aerial photographs at resolutions of 17.3 x 17.3 meters and 2.67 and 2.67 meters respectively. Using the same window sizes, Irons concludes, "high resolution remotely sensed data may result in more useful information for the thematic mapping of land cover." Textural patterns evident in low resolution data may be larger than the standard 3 x 3 or 5 x 5 windows implemented in the reviewed texture research. Low resolution images may require larger windows to analyze a significant amount of the repetitious pattern, if it is present, for its tonal variation; this unfortunately, inevitably increasing textural boundary smearing.

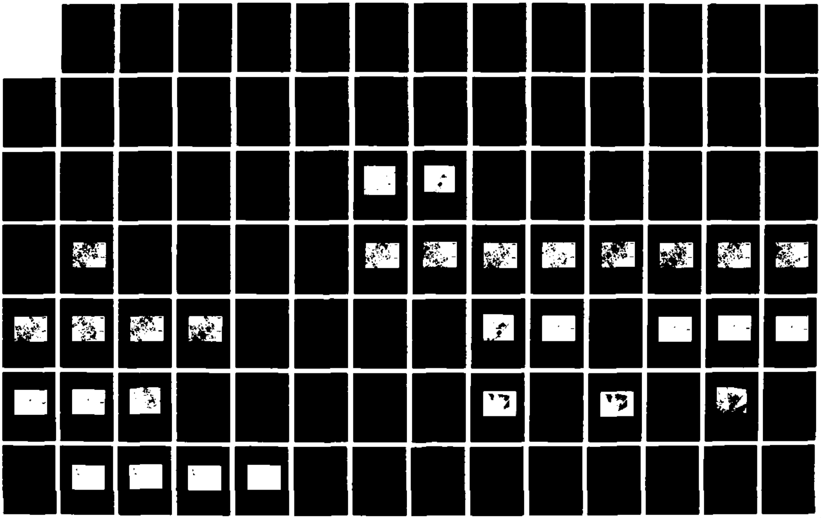
Another problem could be the assumption that an individual land cover is represented by a homogeneous texture pattern. Certain land cover types could contain several different textures or textures similar to other classes (Shih and Schowengerdt, 1983). For example, texture may be very efficient at detailing the continuous texture of a forest canopy, but the texture for an urban land cover class may not have as consistent and continuous a texture throughout the class. If this textural information is implemented as a fourth dimension in the multivariate statistical training set data, inconsistent texture values for the urban area may contaminate the unimodal multivariate probability density function of a perfectly good training set. Although results from Jensen's (1982) classification of certain areas on the urban fringe, have been very promising, it is unclear how many unsuccessful projects have not been documented. In general, texture for urban areas is not as predictable as that for a forest canopy. The converse of this argument is the potential advantage of textural information when classes are spectrally similar but are texturally distinct (Shih and Schowengerdt,1983).

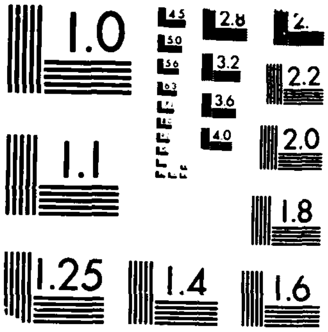
There have been many studies involving land cover classifications using only textural information and the combined spectral and textural information. From most of the studies it is evident that textural data by itself is not as successful as the combined information. And often it has been found that textural information, when compared to spectral-only classifications, has been selectively accurate in describing only some cover types of an image while failing to adequately describe others.

Haralick and Schanmugam (1974) segmented a Landsat MSS scene using texture-only with a classification accuracy according to test samples of 67.5%. Spectral-only was more accurate with an assessment of 77%. The spectral-texture classification resulted in a classification accuracy of 83.5%. Jensen and Toll (1982) found that combined spectral and textural data provided complementary information in describing cover types in an image.

AD-A192 525 SEMI-SUPERVISED TWO STAGE CLASSIFICATION TECHNIQUE(U) 2/3
ARMY MILITARY PERSONNEL CENTER ALEXANDRIA VA
D A TOOMEY 31 JUL 87

UNCLASSIFIED F/G 14/5 NL





MICROCOPY RESOLUTION TEST CHART
BUREAU OF STANDARDS 1963 A

However, as mentioned before, not all classifications involving texture analysis are so successful. Irons and Peterson (1981) did not offer a quantitative assessment of accuracy in their results, but did state that the various combinations of the texture transformations failed to provide useful texture classes, and that this sharply contrasted with Hsu's (1978) success. It should be noted that the classifications that were done by Hsu and Irons were texture-only.

3.5 The Benefits of Texture Measures

A benefit inherent in textural measures is that it can be derived from the original spectral information through algorithm transformations. A question is then posed, which spectral band should be selected to generate the most descriptive textural measures. In many instances in the literature the red spectral band (.6 - .7 μm) was chosen (Haralick and Schanmugam,1974; Jensen,1979; Wessman,1984). The red wavelengths are selected because they delineate boundaries between natural and manmade features; primarily because the red channel is a major chlorophyll absorption band. Red light is absorbed by photosynthetically active vegetation (Wessman,1984).

3.6 Selective Contribution of Texture

Texture has been implemented simultaneously with spectral data for classifying all the land cover in an image and also certain select land cover classes of interest.

Textural analysis has been used for classification of selected land cover within images. In Shih and Schowengerdt (1983), texture was implemented to discriminate between geologic classes that have spectral overlap but are texturally distinct. Such classes were varnished bedrock slopes and desert pavement, and also lightly covered bedrock and alluvial surfaces. In Jensen and Toll (1982) texture was used to detect five different levels of residential land-use development at the urban fringe. In the cited literature, although the whole image was classified with the assistance of textural information, only select land cover types of interest were evaluated in the experiments. It is not clear in these studies whether texture was effective in classifying all cover types in an image. Wessman (1984) postulated, from her results, that the contribution of textural information is selectively better for classifying some classes than for others. An assumption could be made, from the previous research, that texture may be more useful when it is implemented in a discriminant manner, for a specific purpose.

The selective contribution of texture in classifications and the intent to classify certain cover types with increased accuracy using texture, are the basic reasons for the investigation of a different way to implement ancillary data in a second stage post-classification approach. A post-classification introduction of texture to reclassify an image in a discriminant manner is discussed in the next chapter.

3.7 Alternate Ancillary Data

Other ancillary digital data bases mentioned at the beginning of this chapter could also be useful in image segmentation, but are seldom implemented. For example, vegetation index ratios are mostly used in vegetation analyses and change detections studies as an estimation of biomass present in the land cover (Tucker, 1979). Vegetation index ratio equations involve the red and near infrared spectral bands of a multispectral image: 1) normalized ratio (IR - Red / IR + Red), 2) transformed vegetation index [sqrt (VI + 0.5)], or 3) a simple ratio (IR/Red). Forest land cover and various gradations of vegetation can be level sliced from a vegetation index ratio image. Vegetation index ratios are derived from the original spectral data, and could also possibly be more beneficial in automated classification when implemented in a post-classification approach.

Shih and Schowengerdt (1983) discussed the use of spectral band ratios in classifications. As with the vegetation index ratios, these ratios can delineate certain features of interest in an image. Therefore ratios could be implemented to verify or reclassify a spectrally classified image, a technique detailed in chapter 4.

Geographic information system data bases are digitally amenable data that could also be utilized in a discriminant manner in classifications.

A method which implements ancillary data in a discriminant manner is detailed in the following chapter.

Chapter IV

Semi-Supervised Two Stage Classifier

4.1 Introduction

This chapter discusses the development of a two stage Semi-Supervised classification technique. The proposed Semi-Supervised clustering algorithm combines the advantages of an unsupervised method with the direction (guidance) attributed to a supervised approach.

The Semi-Supervised Second Stage classifier revolves around two central ideas. First is the creation of a Semi-Supervised clustering process that requires initial guidance from the user. Second is the development of a post-classification method with the capability to access useful information from ancillary data, such as texture, to assist in reclassifying certain cover types in an an image.

Although the title implies that these two processes are sequential in application, they are not. The Semi-Supervised approach does not have to implement a reclassification; and the post classification does not require a spectral classification that results from a Semi-Supervised process.

The combined classifier is a complete process involving:

1) A clustering technique (Semi-Supervised)
2) Intermediate step - Maximum Likelihood classifier or a Box-Preprocessor Maximum Likelihood Classifier
3) Post-classification acquisition of Second Stage ancillary information

4.2 *Design of the Classification Technique*

The classification process is composed of two stages. First, training sets are identified with a Semi-Supervised clustering scheme. These training sets are used by a maximum likelihood or a box-preprocessor maximum likelihood classifier to spectrally segment the scene. Secondly, a Second Stage of digital information is introduced through a technique that integrates ancillary information in a discriminant manner to assist in increasing classification accuracy of a scene.

4.3 Part One : Semi-Supervised Clustering

The process begins with the analyst designating relatively large training "areas" (with polygons) over locations containing the various spectral diversities of the individual land cover categories (i.e., a forest training "area" polygon may delineate an area containing stands of different tree species). The user then designates the number of spectral clusters for each land cover training area. One or two training areas are defined for each land cover category (for example forested land, urban areas or agricultural areas). These training areas cannot be equated with subimages, since the spectral diversity within each polygon is intended to represent the one or two land cover types and not that of the whole image. An unsupervised clustering aigorithm looks into each polygon with a 3 x 3 window (Figure 8). The clustering algorithm is based on the FINDSET approach detailed in section 2.11. With a user defined threshold for the sum of the variance of the spectral bands the algorithm searches for the defined number of clusters (Figure 9). The window moves through each polygon twice, in a checkerboard pattern; on the second pass it looks at those areas that it skipped over the first time. The checkerboard sampling pattern results in the windows touching only at the corners. This should reduce the effects of autocorrelation if it is significant. After the first 50 clusters are identified, the two clusters that are most similar are merged, according to the same statistical distance metric used in FINDSET, reducing the number by one. This continues until the polygon area is processed. The total number of clusters is then merged down to the user defined number of spectral classes for that area.

In short, a user defined number of clusters are identified to describe the various classes within each land cover category delineated by a polygon. Such a Semi-Supervised technique could be described as a stratified unsupervised clustering technique. Since it entails both supervised and unsupervised methods it is considered a hybrid approach.

This technique is capable of reducing the bias, inherent in clustering algorithms, of selecting clusters that cannot adequately classify the whole image. Spectral training statistics for all land cover classes in a scene

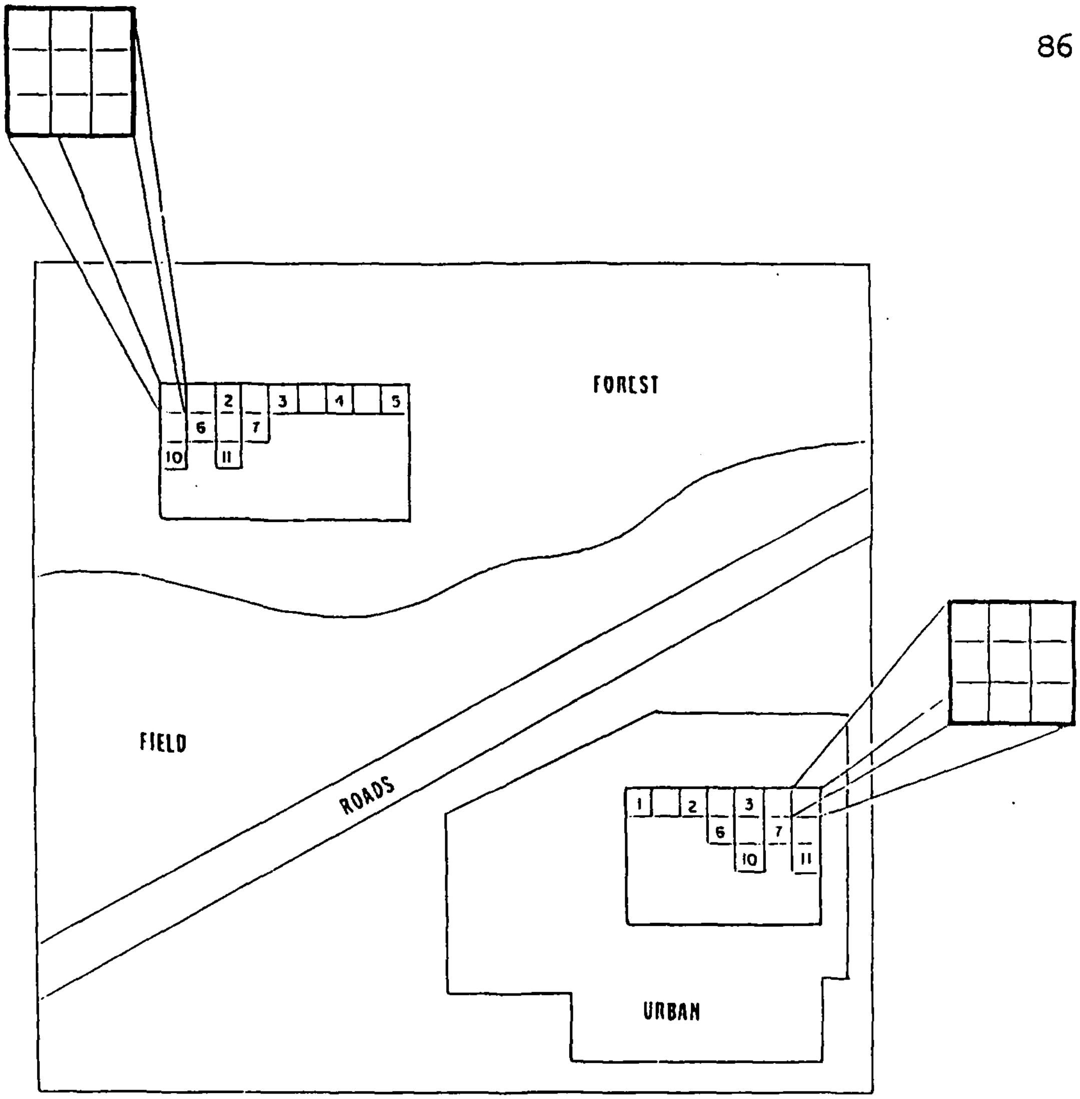


Figure 8. Unsupervised clustering within the training areas.

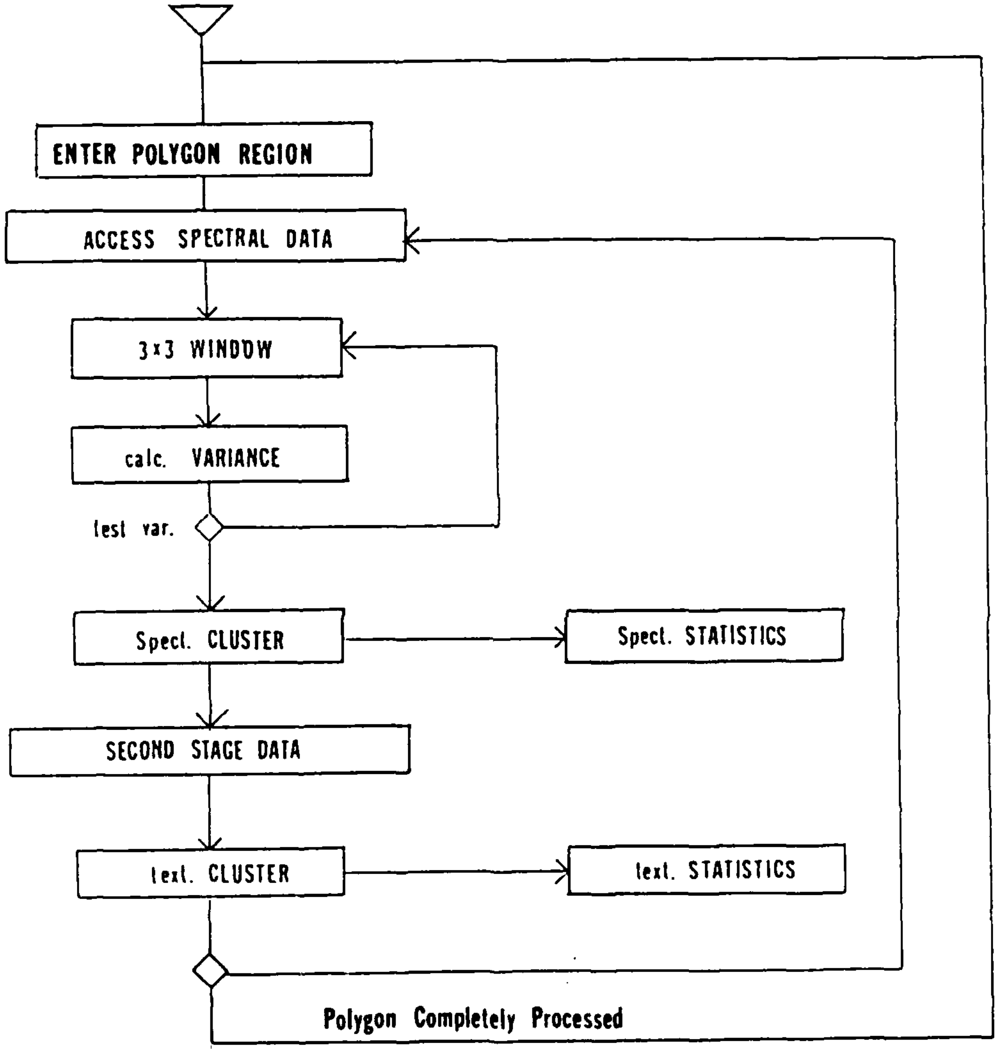


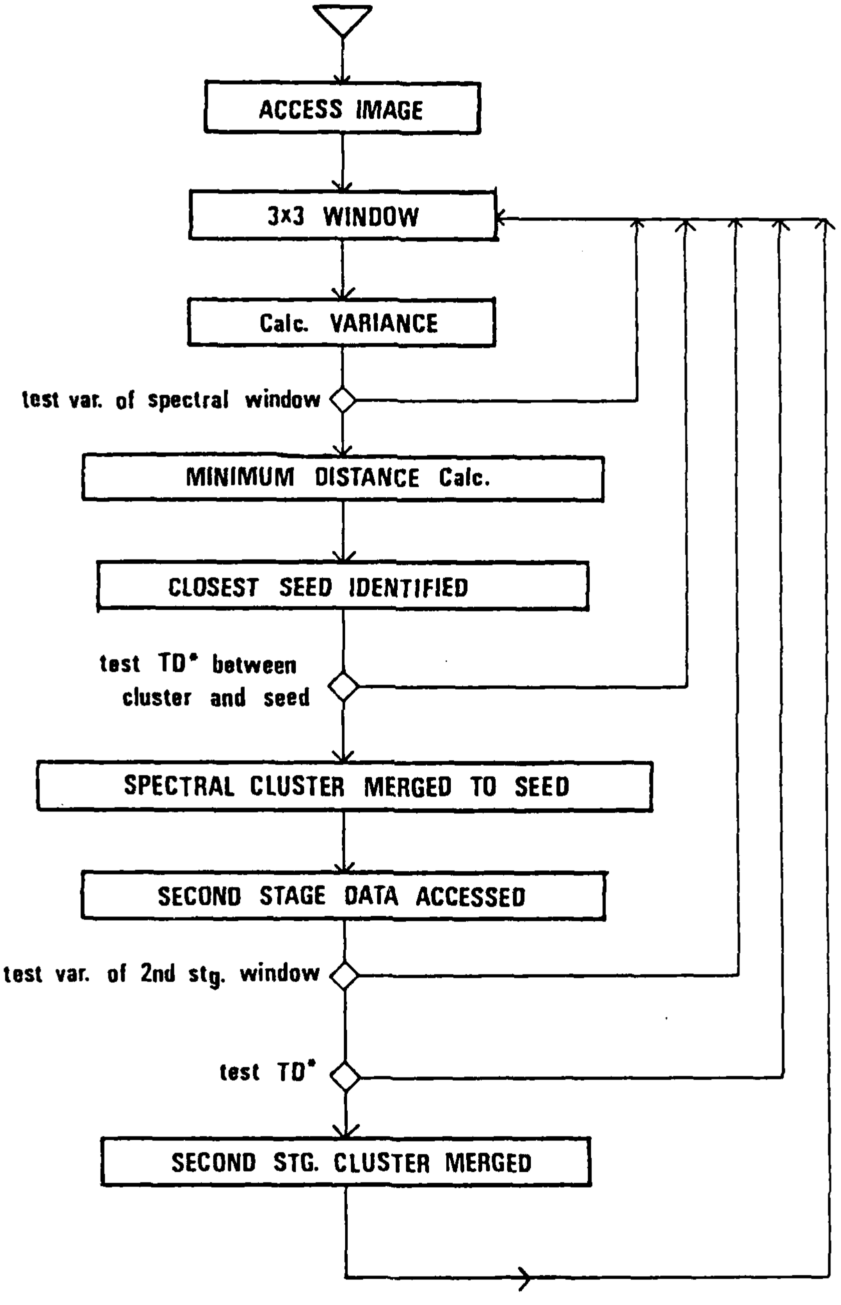
Figure 9. Semi-Supervised clustering of the spectral seeds.

that the user is interested in classifying, could be identified with a Semi-Supervised clustering algorithm. Spectral diversities within each category, could be identified by this Semi-Supervised clustering procedure, rather than through a multitude of training sets laboriously pinpointed by an analyst.

The number of pixels collected for each spectral cluster, from the training areas, may not be sufficient to represent statistically valid training sets for their prospective spectral category, therefore more spectral information from throughout the scene should be accumulated. An adequate sample population would be 10n to 100n pixels, where n is the number of spectral bands used in describing the data. Thus, the clusters are then implemented as seeds to pool spectral statistics from throughout the scene to produce statistically valid training sets.

A 3x3 window is passed across the image, in a checkered pattern, to test areas for spectral homogeneity, according to a user defined variance threshold (Figure 10). Pixels in the homogeneous areas are merged with the most spectrally similar seed cluster according to a minimum statistical distance operation. The seed clusters will not be merge with each other. A transformed divergence requirement must be met before a cluster from the image is merged with a seed. A second pass is not made throughout the scene, to test areas skipped while implementing a checkered pattern, due to the increased run time involved. There is an option given to the user to classify the scene with the incipient training sets, from the polygons, without fulfilling the seed option of the program.

Use of the transformed divergence operation as a measure of spectral separability should prevent mismerging of spectral data. Mismerging occurs when the cluster to be merged possesses spectral characteristics that are not represented by any of the seed clusters. The Semi-Supervised training approach directs the unsupervised clustering algorithm to train on certain areas. These areas designated by the user may not contain all the spectral diversity for a land cover type, thus when the whole image is accessed these spectral signatures could be mismerged with the improper training seeds. Transformed divergence should prevent mismerging. Still, the resulting training sets would only be from the areas designated


ACCESS IMAGE
3x3 WINDOW
Calc. VARIANCE
test var. of spectral window
MINIMUM DISTANCE Calc.
CLOSEST SEED IDENTIFIED
test TD* between cluster and seed
SPECTRAL CLUSTER MERGED TO SEED
SECOND STAGE DATA ACCESSED
test var. of 2nd stg. window
test TD*
SECOND STG. CLUSTER MERGED

• TD= Transformed Divergence

Figure 10. Merging of spectral data from the entire image with the spectral seeds.

by the analyst; it is possible for the analyst to omit certain spectral signatures resulting in unclassified areas in the classification.

The resulting training sets are placed in a statistical classifier to segment the image into the directed number of land cover categories. The analyst can then determine the relative significance of the individual spectral classes within each category. As explained before, an example would be different species in a stand of trees.

During the acquisition of spectral information, Second Stage data may be simultaneously accessed and formulated into statistics that represent the land cover of each spectral class described in the clusters (Figure 9). For each spectral window obtained in the polygon areas a 3 x 3 array of pixels in the Second Stage is also identified and stored statistically. However, when the initial Second Stage clusters are implemented as seeds, groups of pixels in the Second Stage that do not meet a user defined variance threshold for the Second Stage, are not merged (Figure 10). Also a transformed divergence criterion must be fulfilled.

Again, it is emphasized that the simultaneous acquisition of Second Stage information is not required, if there are no intentions of reclassification or a different approach to reclassification is performed.

4.4 Transformed Divergence

Divergence is a measure of statistical separability between patterns (clusters); a distance measurement sensitive to the means and variances of the distribution (Lillesand and Kiefer, 1987). It can be written as an equation involving the means and covariance matrices for the two pattern distributions i and j, such as,

$$\begin{aligned} D_{ij} &= 1/2\ \mathrm{Tr}[(\Sigma_i-\Sigma_j)(\Sigma_j^{-1}-\Sigma_i^{-1})] \\ &\quad + 1/2\ \mathrm{TR}[1/2(\Sigma_i^{-1}+\Sigma_j^{-1})(U_i-U_j)(U_i-U_j)^T] \end{aligned}$$

(Swain and Davis,1978).

This computed divergence value has an unlimited range, 0 to infinity, but the transformed divergence is a metric that has a minimum and maximum value that can be defined, 0 to 2000.

$$D_{ijT} = 2000 \times [1-e^{(D_{ij}/8)}].$$

A transformed divergence solution of 0 - 1500 indicates that the clusters are spectrally similar (Lillesand and Kiefer, 1987) and can be merged. A value above the 1500 threshold indicates a 90% probability of being statistically separate. Therefore, the homogeneous clusters that are merged to the seeds in Semi-Supervised approach must meet this range of transformed divergence measurements before merging takes place.

4.5 Second Stage Reclassification Approach

4.5.1 Introduction

After the image has been classified spectrally, a Second Stage of digital information is introduced to potentially increase the accuracy by verifying and reclassifying the spectral segmentation. The terminology 'Second Stage' is synonymous with digital ancillary data in this study. In the discussion of the theory of the reclassification, texture will often be referred to as the Second Stage due to its relevancy, in the discussion, but other ancillary data is amenable in the process.

In the past when spectral and texture features were combined for scene segmentation joint multivariate training statistics from both spectral and textural information were employed (Jensen and Toll, 1982; Shih and Schowengerdt, 1983). However, ancillary data may be more beneficial if it were introduced, in a limited manner, such as in the post-classification process which will now be discussed.

Probably not all the land cover classes spectrally classified should be reclassified with regards to the Second Stage. Second stage data such as texture can be relatively important in the final classification for only a few land cover classes. Studies previously mentioned, in chapter 3, demonstrate texture to be most useful in limited operations.

In a post-classification process, the user designates the appropriate land cover types, of a spectrally classified image, to be reclassified according to textural information. This prevents land cover types from being indiscriminately classified by texture; only the appropriate land cover types are reclassified.

Reclassification among certain land cover types could produce favorable results. For instance, relatively high textural values can be found within tree canopies that are unlike the texture for grassy open fields or moderately textured brush. Access of Second Stage textural information should alleviate misclassification of grassland and other such vegetation with forested land that may be spectrally similar but texturally distinct. Similarly, the Second Stage should be competent in dealing with other classes. The user must designate the land cover categories that are to be reclassified according to the Second Stage of information.

There are two different approaches to reclassification: statistically based procedures and thresholding procedures. The statistical method permits the implementation of Second Stage statistical files, such as textural statistics. The thresholding procedures allow threshold amenable data to be input into the reclassification. A binary mask or a geographical information system data base are ancillary data that are amenable to threshold reclassifications.

4.5.2 Statistical Reclassification

The statistical based approach will be discussed with texture as the Second Stage data base.

The same training areas used in identifying the spectral clusters will be utilized to find training set statistics in the Second Stage (see Figures 9 and 10 and also section 4.3). These Second Stage training sets should describe the individual land cover classes with ancillary data. For each pixel classified spectrally there is a corresponding textural value from the same spatial location within the Second Stage image. If the Second Stage information for the spectrally classified pixel is similar to the Second Stage statistics compiled for that class, then the pixel is considered correctly classified. Conversely, when a pixel that is spectrally classified as a certain class has a different textural value than that associated with that class, then it is considered unclassified from that class.

When a pixel becomes unclassified it must be reclassified according to the Second Stage data. Once again, in this step the analyst would designate the land cover classes to be reclassified, if misclassified, and the classes into which each can be reclassified, including priority. This process obviously requires some a priori knowledge of the scene or experience on the part of the user.

The user designated specifications will create a reclassification table that will be used by the algorithm to conduct the reclassification among the appropriate classes. Spectrally classified pixels, of the classes designated by the user, will be reclassified according to the reclassification table. This table will be an arrangement of conditional statements, representing the decision rules for reclassification (Figure 11).

The reclassification table will contain the land cover class to be reclassified and the classes into which it can be reclassified. Also the user must designate a probability threshold (standard deviation), so that the dis-

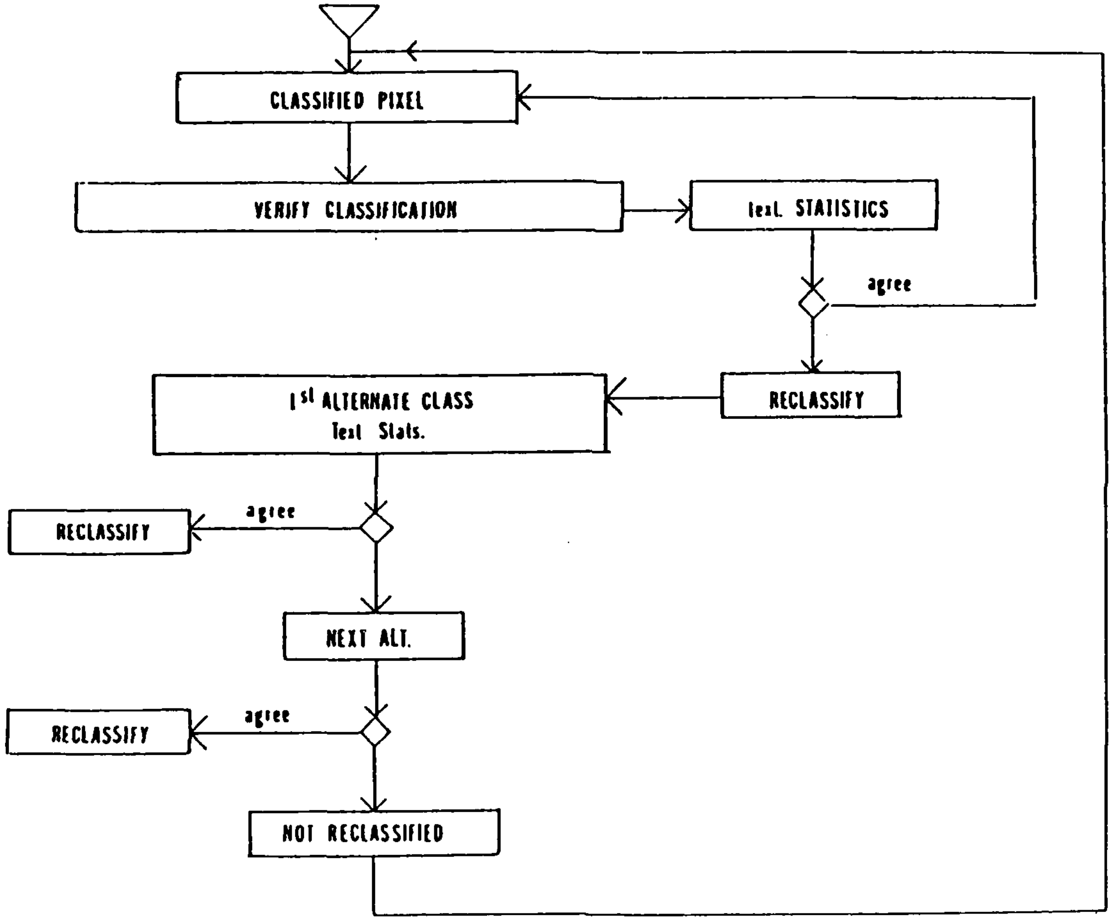


Figure 11. Statistical Second Stage reclassification approach.

criminant function (box-filter) can be utilized to reject a classification, if the probability is too low.

To illustrate the function of the reclassification table, the following example is provided (Figure 11). Assume the classes into which the forest canopy can be reclassified are brush (rangeland), grass, and marsh vegetation; listed according to priority. (these are chosen as examples and in the flow chart are called alternate classes). The algorithm will go into the image and find pixels spectrally classified as forest and test its registered textural value. If this textural value is considered to be of this forest class, according to the statistics describing the texture for this class, and surpasses the probability threshold designated by the user, it is not reclassified. But if it is not described by the textural box type filter, or it is described, but falls below the probability threshold the pixel will be reclassified. The textural value is then compared to the textural statistics of the first alternate class designated by the user for reclassification. An identical type of test is performed for this class and subsequent alternate classes designated until the forest class is reclassified. If the forest pixel is not reclassified when the designated classes are exhausted, then it remains classified as forest.

4.5.3 Threshold Reclassification

Any ancillary data that is amenable to thresholding procedures can be implemented to reclassify certain classes of the spectrally classified image. Thresholding involves the identification of certain ranges or individual digital values in an image for enhancing certain features in the image or delineating certain cover types in the scene. The latter is the purpose behind the threshold reclassification approach.

A polygon mask will be employed as the Second Stage image file to illustrate this approach.

As with the statistical option, threshold reclassification will only involve the appropriate classes in the classified image. The threshold technique permits reclassification according to individual values or ranges in the Second Stage instead of statistics. The pixel is reclassified into the class indicated by the analyst.

For example, a body of water in a scene has specular reflection that is spectrally classified as bare soil and wet grassland because of its spectral signatures. A polygon mask, of the water body, is generated by delineating the water with a polygon and using a poly-masking program. The masked image file has "zeros" in place of the water (polygon) and unchanged values everywhere else in the image.

With the threshold reclassification technique the analyst could declare all bare soil and grass in the image to be tested against the binary file. Every pixel that is classified bare soil and grass and has a zero in the Second Stage data file will be reclassified as water. This could adequately remove the specular reflection.

Not all threshold amenable files are binary. Some will contain more than two levels of data. For example, geographic information system data bases have levels for every category in the file. A zoning class data base may have a level for each zoning category: different groups of industrial, commercial and residential areas. A county soil survey file may have levels for individual soil textures.

To verify or reclassify certain land cover classes in a spectrally classified image, an analyst may be interested in comparing the classified data with certain ranges or individual digital values in the Second Stage. Two options are available for threshold reclassification. The first one involves reclassifying classified pixels according to a threshold level. This would reclassify a pixel into a designated class if the value in the Second Stage were less than or equal to the threshold value. The second, reclassifying classified pixels according to a threshold range, would reclassify a pixel into a class designated by the analyst if the value in the Second Stage were within this range.

The analyst input to the reclassification process would be the classes to be reclassified, the option (by threshold value or by threshold range), the threshold value(s) and the class into which it is permitted to be reclassified.

To illustrate the function of threshold reclassification see the flow chart, figure 12. The designated spectrally classified pixel is compared to the corresponding Second Stage value according to the option 1 or 2 decision rule.

Option 1: For each classified pixel designated, if the Second Stage data base value is less than or equal to the threshold value, defined by the user, then it is reclassified into another class, also defined by the analyst. (There is also an option to reclassify or not to reclassify if this requirement is met.)

$$\text{Second Stage value} \leq \text{threshold}$$

Option 2: For each classified pixel designated, if the Second Stage data base value is within the range, described by the analyst, then it is reclassified into another class, defined by the user. (There is also an option to reclassify or not to reclassify if this requirement is met.)

$$\text{threshold1} \leq \text{Second Stage value} \leq \text{threshold2}$$

4.6 Summation

In summation, the Semi-Supervised Two Stage Classification Technique is a combination of two procedures that can be executed independently. The Semi-Supervised clustering algorithm can be used to classify a scene that will not be reclassified. And an image classified with procedures other than the Semi-Supervised technique can be reclassified. The title indicates Second Stage statistical

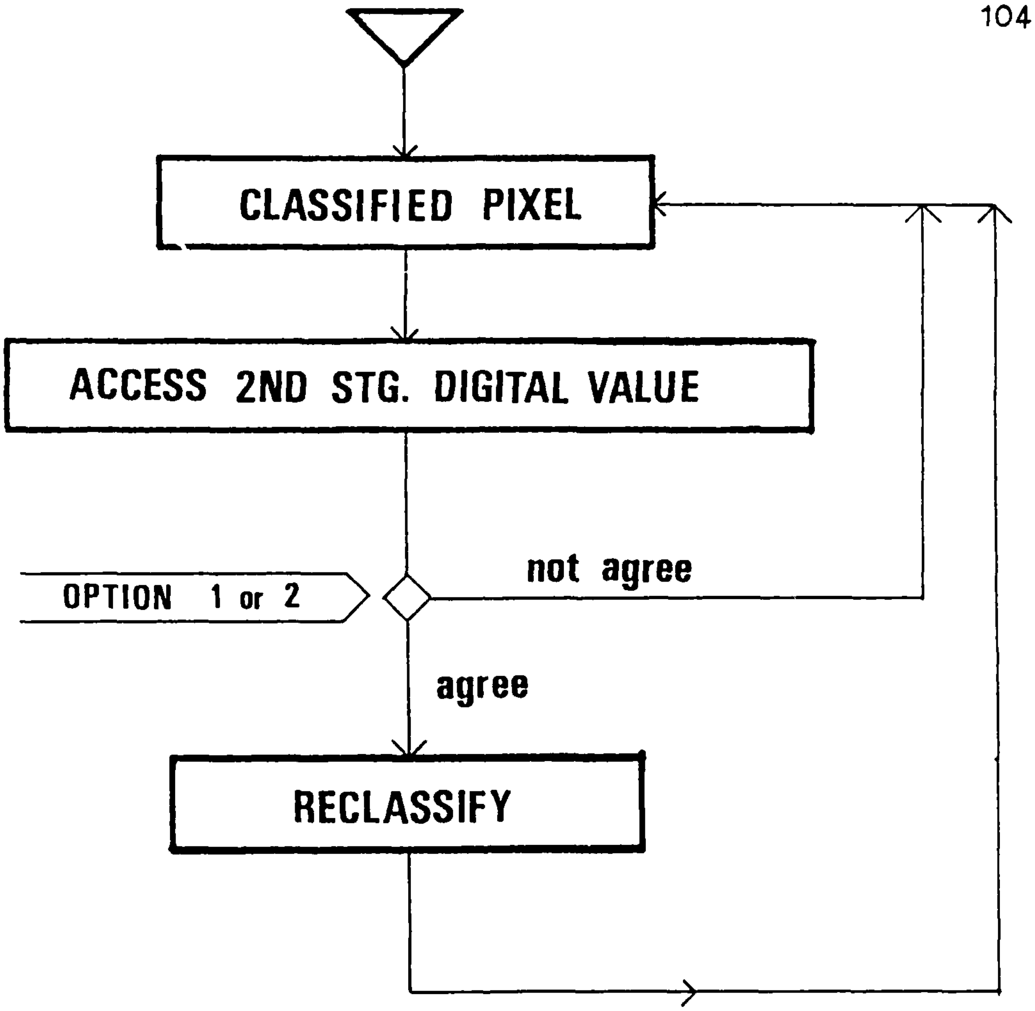


Figure 12. Threshold Second Stage reclassification approach.

information for the spectral classes can be simultaneously acquired with the Semi-Supervised clustering algorithm.

The Semi-Supervised (directed) clustering approach may potentially reduce the clustering bias of the FINDSET algorithm. Also with the transformed divergence modification mismerging of spectral clusters should be avoided when the seeds acquire more spectral information from the scene.

The post-classification reclassification technique can be used to reassess certain classes, in a spectrally classified image, according to ancillary data. This discriminant process has two methods of operation: statistical reclassification and threshold level reclassification. The Second Stage statistical information can be obtained in a Semi-Supervised manner or in a supervised acquisition; the only thing that is required is that there are Second Stage statistics for each spectral class to be verified or into which it will be reclassified. The Second Stage data base for the second option must be amenable to thresholding, such as binary masks, map overlays, digitized soil surveys, GIS data bases and even images that have been ratioed, if the relevant digital values are known. Any digitally amenable ancillary data could be implemented in a post-classification reclassification process.

Chapter V

Methodology

5.1 Introduction

Evaluation of the Semi-Supervised Two Stage Classification Technique involves two sets of experimentation. One is designed to evaluate the Semi-Supervised clustering process and compare to the FINDSET unsupervised approach. The second is intended to demonstrate the utility of the Semi-Supervised clustering accompanied by a Second Stage reclassification *technique*.

5.2 Study Sites

There are two study sites evaluated in this thesis.

The first is a subimage from a SPOT satellite image located over Madison, Wisconsin. It was acquired on June 3, 1986 at about 11:00 AM Central Daylight Time (Figure 13). The subimage is 512 columns by 480 rows and is located over the community of Middleton area of the Greater-Madison area. Figure 13 is false color composite of the three spectral bands recorded by the SPOT satellite: green, red, and infrared. The western edge of Lake

Mendota can be viewed in the bottom right of the scene. The land cover types in the image are primarily agricultural in nature with 3 wetland areas: Waunakee Marsh (upper left), Dorn Creek Marsh (center right), and the Pheasant Branch Creek Marsh (bottom right). The image was taken early in the growing season, therefore most row crops are primarily depicted as bare soil and emergent vegetation. Perennial alfalfa fields are present throughout the image. With a pixel ground resolution of 20 meters, the area covers approximately 10 kilometers (approximately 6 miles) in each direction.

The second is a digitized aerial photograph taken over the Chesapeake Bay Region, near Edgewood, Maryland (Figure 14). The prominent ground features in the image are Watson Creek, and two partially reclaimed landfills surrounded by thick forest canopies. It is a color infrared aerial photograph taken on June 24, 1981. The image is 467 columns by 400 rows, with a pixel resolution of 0.50 meters. Each pixel is a point sample of approximately every 3 meters on the ground (every sixth pixel of the original digital image). The scale of the digital image as viewed on the screen is 1:8000 and covers an area of

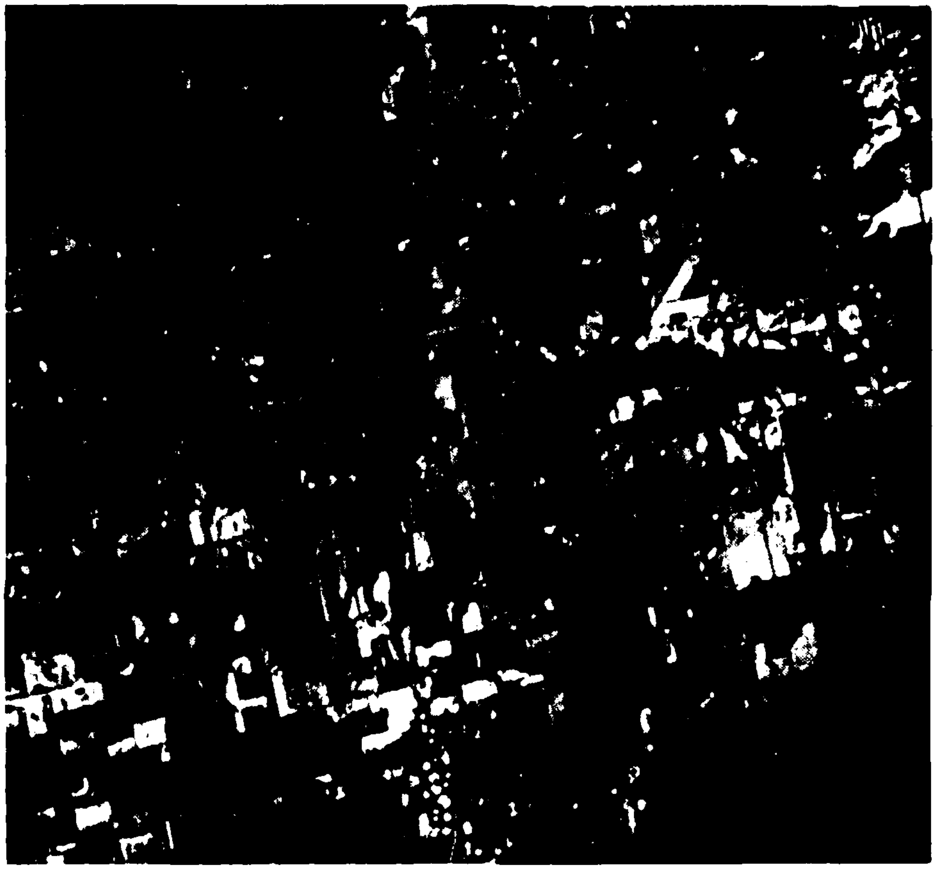


Figure 13. Middleton area study site, SPOT satellite sub-scene.

Figure 14. Chesapeake Bay region study site (original scale 1:10,000).

1.3 km by 1.1 km. The site contains 5 basic land cover categories to be segmented: forest, grass area, bare soil (dry/disturbed land), man-made features and water.

5.3 Clustering Evaluation

The Middleton subimage will be used in the clustering evaluation of the algorithms. In evaluating the Semi-Supervised clustering process, the resulting classified images from the Semi-Supervised clustering and FINDSET clustering will be compared, and accuracy analysis will be based on a supervised classification of the same scene. To analyze clustering bias 4 different rotations of the image will undergo clustering analysis and classification. No spectral information is lost in the rotation.

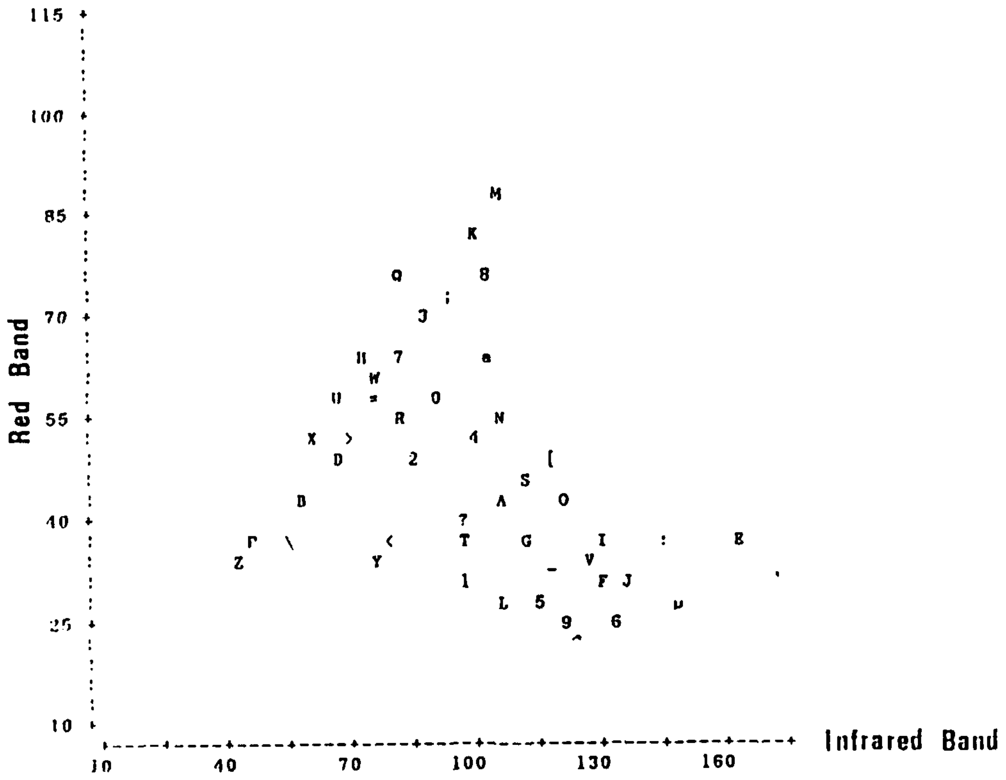
A threshold variance must be selected for the clustering algorithms. A variance threshold of 30 was selected to identify homogeneous 3x3 windows in the Madison scene. This threshold has been selected in accordance with experiments in the previous months that evaluated clustering at different logical threshold values. It was concluded that the suggested threshold of 30 (Ahearn,1986) for the sum of the variances in the spectral bands, adequately segmented the image.

The selection of the variance threshold depends on the land cover classes one is interested in describing. To acquire training set statistics for uniform spectral resources such as agricultural fields a lower variance threshold should be selected. A selection of a higher variance threshold would also permit training sets to be identified for resources, such as forested areas, which by nature have a high variation of spectral response.

The Semi-Supervised approach requires the analyst to designate a transformed divergence threshold value to prevent mismerging of spectral data. A value between 0 and 2000 can be selected. Statistically, a value equal to or greater than 1500 indicates that there is 90% confidence that the clusters being compared are spectrally separate and should not be merged. In these studies a value of 1500 was used in the clustering.

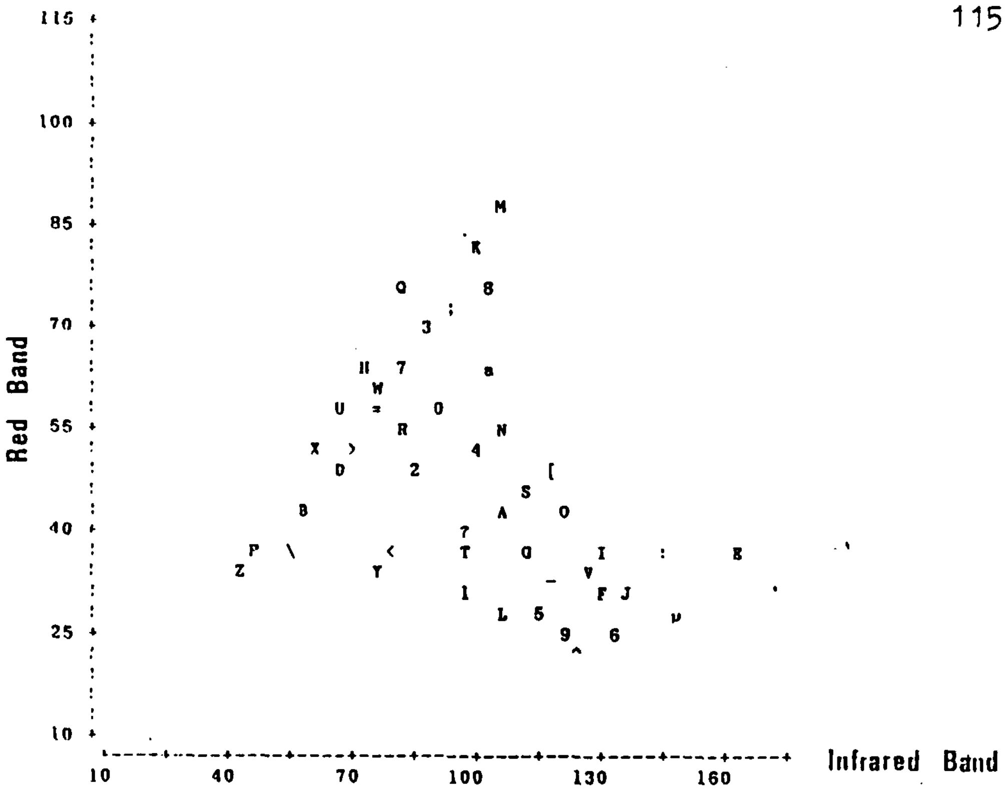
At the Environmental Remote Sensing Center (ERSC) at the University of Wisconsin-Madison there are two programs that employ the SEARCH clustering method (see section 2.10): FINDSET and FINDCLASS. FINDSET was written first and did not calculate a covariance matrix for the training sets. FINDCLASS was written second and calculates a covariance matrix. The algorithms are also different in the way each accesses the image. In FINDSET 240 columns of the image at a time are processed, stepping over 240 columns sequentially until the entire image is accessed. FINDCLASS evaluates all the columns in the image at once, sequentially evaluating every three rows.

Through preliminary analysis it was discovered that FINDSET did not access more than 240 columns in the image. The statistics that were gained from a subimage of 240 columns were identical to those of an image of 512 columns. Meanplots for 50 clusters from both images, of 240 and 512 columns, can be viewed in figure 15 and 16 respectively. No water bodies are located in the first 240 columns of the image. If clusters describing the water were identified their means would plot in the lower left hand corner of the meanplots, an indication of low



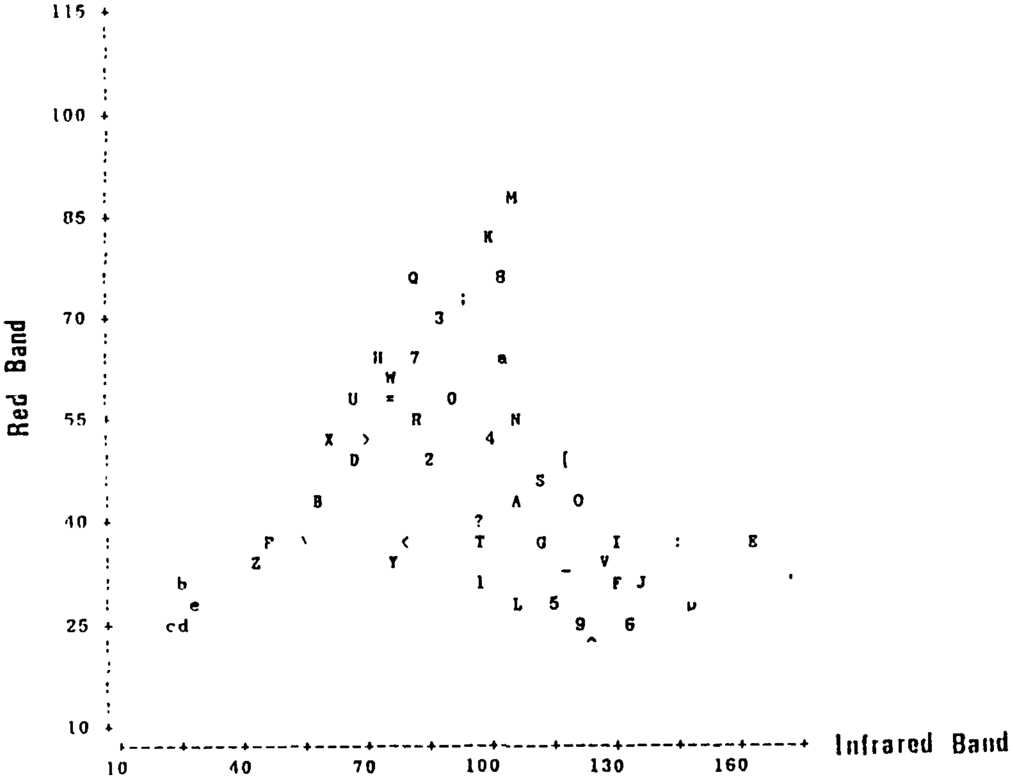
```
Key:

Class #:
                    1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9

Symbol:
1 2 3 4 5 6 7 8 9 : ; < = > ? μ A B C D E F G H I J K L M N O P Q R S T U V W

Class #:
4 4 4 4 4 4 4 4 4 4 5
0 1 2 3 4 5 6 7 8 9 0

Symbol:
X Y Z [ \ ] ^ _ ' a b

Symbol: 0 represents classes:  19  45
```

Figure 15. Spectral meanplot of FINDSET 50 clusters for a 240 column image.



```
Key:

Class #:
                  1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9

Symbol:
 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J K L M N O P Q R S T U V W

Class #:
 4 4 4 4 4 4 4 4 4 4 5
 0 1 2 3 4 5 6 7 8 9 0

Symbol:
 X Y Z [ \ ] ^ _ ` a b

Symbol: 0 represents classes:  19  46
```

Figure 16. Spectral meanplot of FINDSET 50 clusters for a 512 column image.



```
Key:

Class #:
                    1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3
  1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9

Symbol:
  1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J K L M N O P Q R S T U V W

Class #:
  4 4 4 4 4 4 4 4 4 4 5 5 5 5
  0 1 2 3 4 5 6 7 8 9 0 1 2 3

Symbol:
  X Y Z [ \ ] ^ _ ' a b c d e

Symbol: 0 represents classes:  19  45
```

Figure 17. Spectral meanplot of FINDSET 50 clusters and supervised training sets for water in a 512 column image.

spectral reflectance in the red and the infrared bands. The vertical axis is the red spectral response while the horizontal axis is the infrared. To verify the possibility that training set statistics for water were never clustered, supervised training sets were acquired for the four different water bodies in the image. The statistics were plotted with the other 50 clusters in figure 17. Note the differences in the lower left section of the meanplots. Training sets for water were never acquired. FINDCLASS was evaluated in a similar study ensuring it accessed all the rows of an image, and was employed as the 'FINDSET' algorithm in which the Semi-Supervised approach would be compared.

All classified results from the clustering algorithms will be compared to the same thematic map that was segmented using a supervised analysis. This thematic map in figure 18 was generated by the Environmental Monitoring Practicum Analysis class (Blohm, et al., 1987). The author of this thesis was one of the primary classifiers in segmenting the Middleton image during the practicum analysis. After months of spectrally analyzing the image and several field verifications, this scene was classified with 72 training

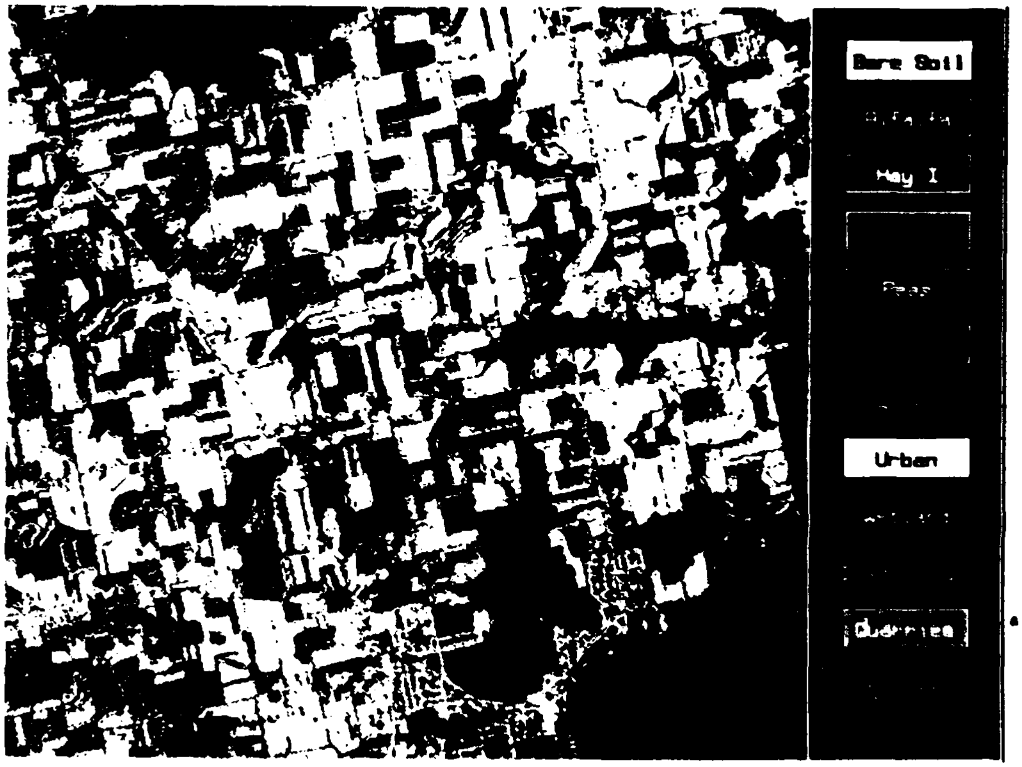


Figure 18. Supervised classification of the Middleton study area.

sets yielding a per pixel accuracy of 88% according to 95 test sites selected in the image. For the practicum analysis the image was classified into 12 land cover classes. Unfortunately, information classes do not necessarily represent spectral classes, which facilitate the evaluation of a clustering algorithm. Therefore, the supervised classification was renumbered accordingly. One of the classes modified was the row crop category, which contained 3 separate spectral classes: bare soil, hay I, and hay II. They indicate 3 different levels of vegetation growth, in which both hay classes are titled as such from ground truth information in the Middleton area. To properly evaluate mismerging of spectral data by unsupervised clustering algorithms it is appropriate to analyse the segmentation of such spectral classes, rather than the previous information classes. The resulting thematic map involves 12 land cover classes: wetland (emergent vegetation), bare soil (row crop), hay I, hay II, peas, disturbed vegetation, water, quarries (disturbed land), urban (roads), forested areas, and an unclassified category. Forests, wetland and disturbed vegetation are spectrally similar and were very difficult to segment in many instances. Disturbed vegetation is an information class describing over-grazed pastures dominated by low growing herbaceous vegetation. This class was selected because it was consistently misclassified as wetlands and forest in fields within the image. It must be noted that most training set polygons for these classes were not acquired within this sub-scene. The study area for the practicum analysis was 1440 rows by 1094 columns.

5.4 Clustering of the Rotated Images

To analyze the Semi-Supervised and FINDCLASS clustering algorithm for clustering bias, the Middleton image was classified at different orientations. There are four basic orientations, the original orientation (0 degrees), seen in figure 13, and three relative rotations: 90, 180 and 270 degrees. Both algorithms were required to identify 50 clusters in each of the four orientations, and 27 clusters in the 0 and 90 degree rotations. The resulting classifications segmented with FINDCLASS can be viewed in figures 19 through 24, and those classified by the Semi-Supervised approach can be seen in figures 25 to 30.

In the four rotated Middleton images, FINDCLASS was directed to identify 50 clusters, and in the typical orientation and the image rotated 180 degrees 27 clusters were requested. In figure 19, the FINDCLASS algorithm found 48 clusters and classified the image into 12 land cover classes: wetland (emergent vegetation), bare soil (also row crops), alfalfa, hay I, hay II, peas, disturbed vegetation, water, quarries (disturbed land), urban (roads), forest and an unclassified category. The FINDSET and FINDCLASS algorithms identify 49 clusters when 50 clusters are requested because there is a final merging between two of the 50 clusters after the last cluster is identified. In the case of figure 19, 49 training sets were identified and one could not be implemented into a statistical classification program because the covariance matrix was not acceptable (see section 6.3 in chapter 6 for details).

Figures 20, 21 and 22 involve the same image that has been rotated from its original orientation in figure 19, analyzed with the FINDCLASS clustering algorithm with 50 clusters, and classified. Figure 20 shows the classified image rotated 90 degrees; figure 21 rotated 180 degrees, and figure 22 rotated 270 degrees. Figures 23 and 24 are the same orientations as in figures 19 and 21, but the FINDCLASS program was asked to identify only 27 clusters.

The Semi-Supervised clustering method was employed and yielded the 6 classified images in figures 25 through 30. Figure 25 illustrates the classified image in its original orientation; figure 26 rotated 90 degrees; figure 27 rotated 180 degrees; and figure 28 rotated 270 degrees. Figures 25 to 28 are segmented with 48 to 49 clusters. Figure 29 is the original orientation classified with 26 Semi-Supervised clusters, and figure 30 is rotated 180 degrees and classified with 27 clusters. The classified images of figures 25 and 28 were classified with only 48 Semi-Supervised clusters to retain an equal comparison with the FINDCLASS output in figures 19 and 22.

Figures 19 to 30 were classified using a maximum likelihood classifier with a box pre-processor filter (see section 2.4.5). Pixels lying outside of the box filter were unclassified. This classification algorithm will classify only the image pixels described by the training sets.

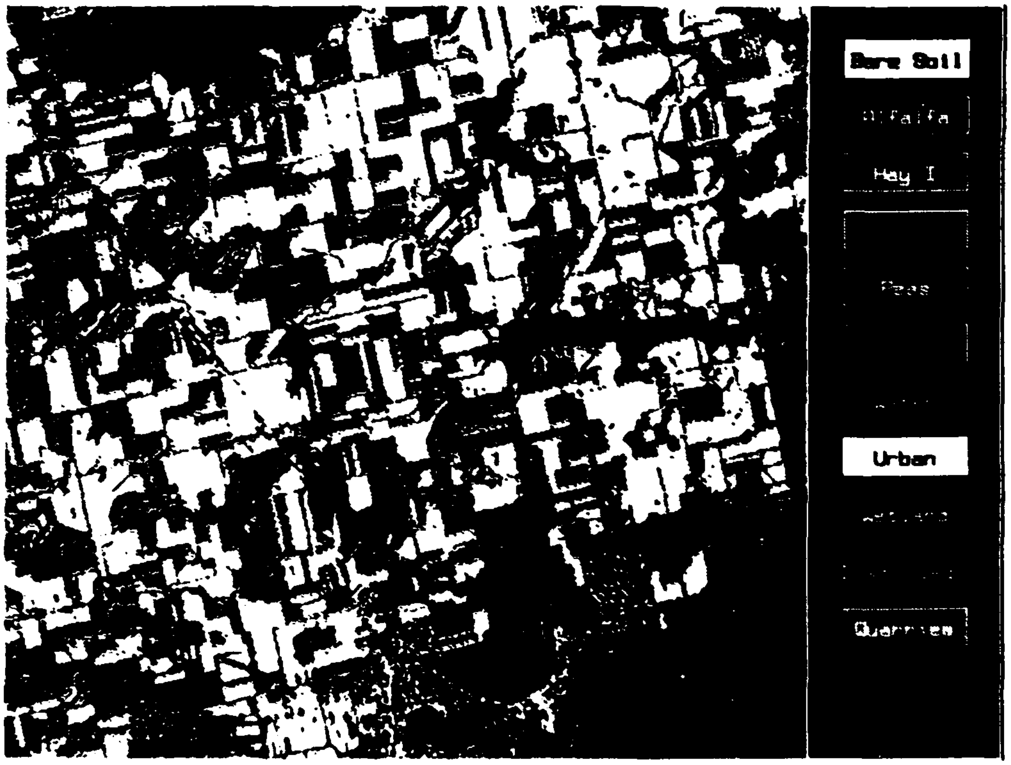


Figure 19. Resulting classification from 48 FINDCLASS clusters, original orientation.

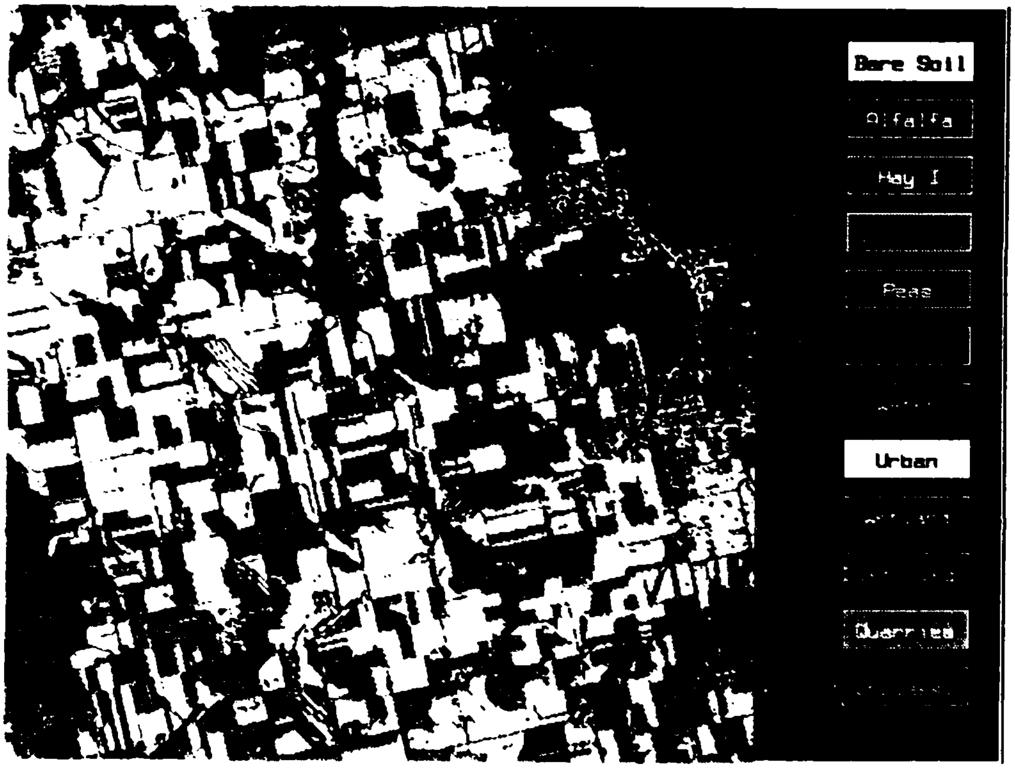


Figure 20. Resulting classification from 49 FINDCLASS clusters, rotated 90 degrees.

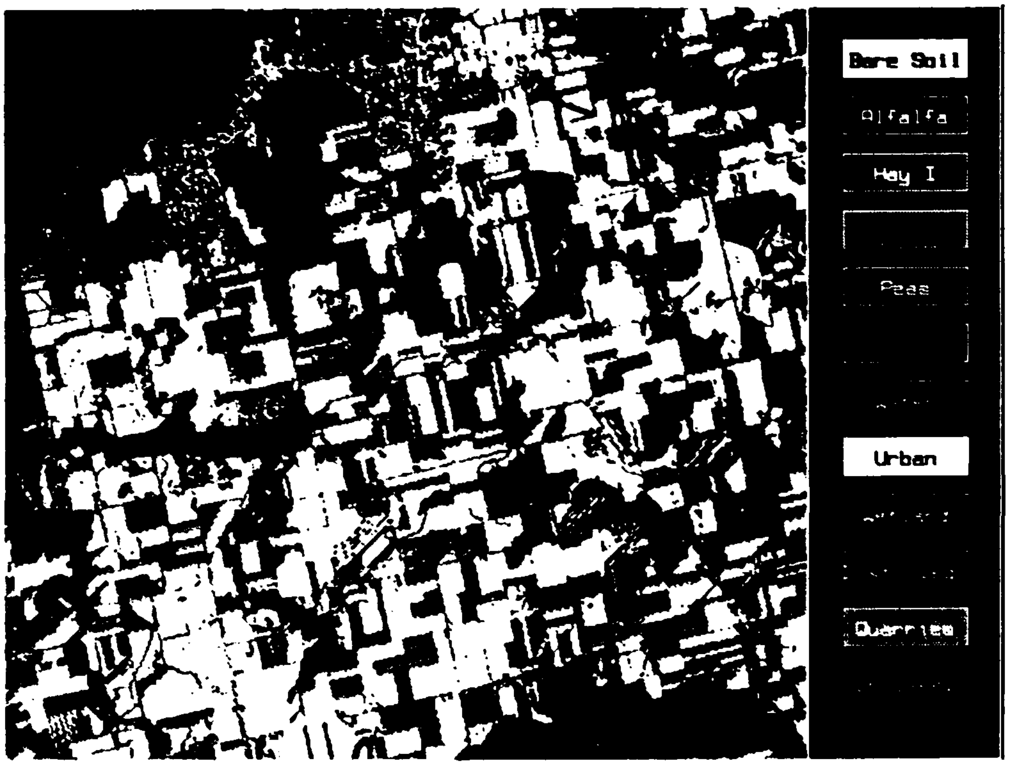


Figure 21. Resulting classification from 49 FINDCLASS clusters, rotated 180 degrees.

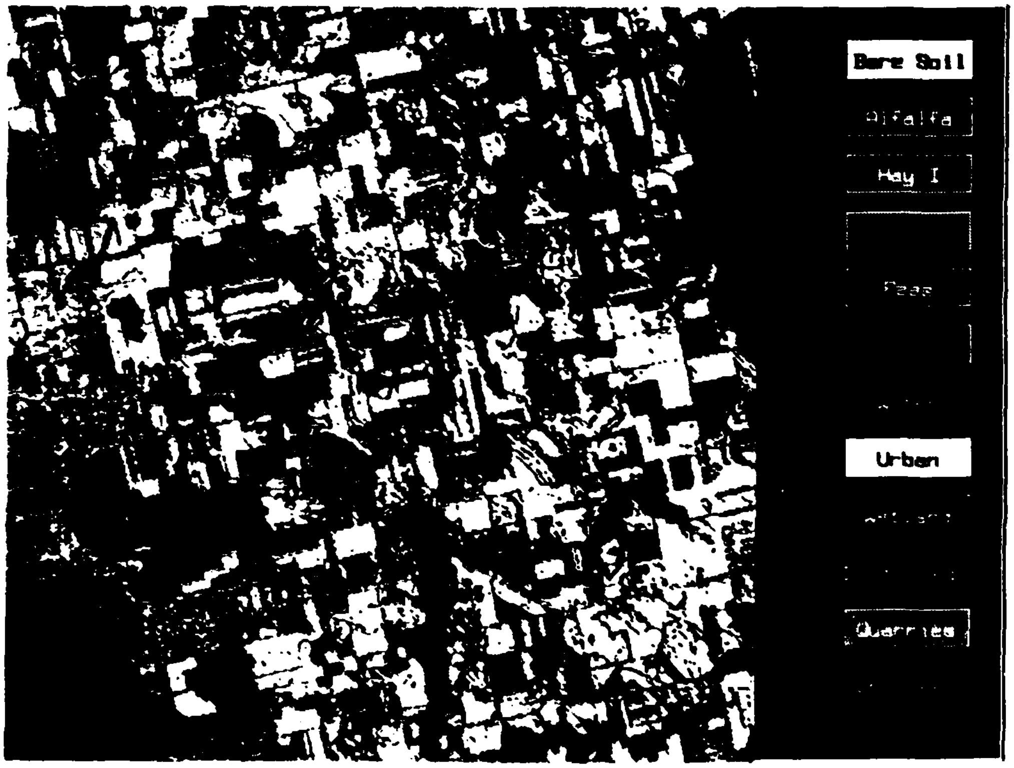


Figure 22. Resulting classification from 48 FINDCLASS clusters, rotated 270 degrees.

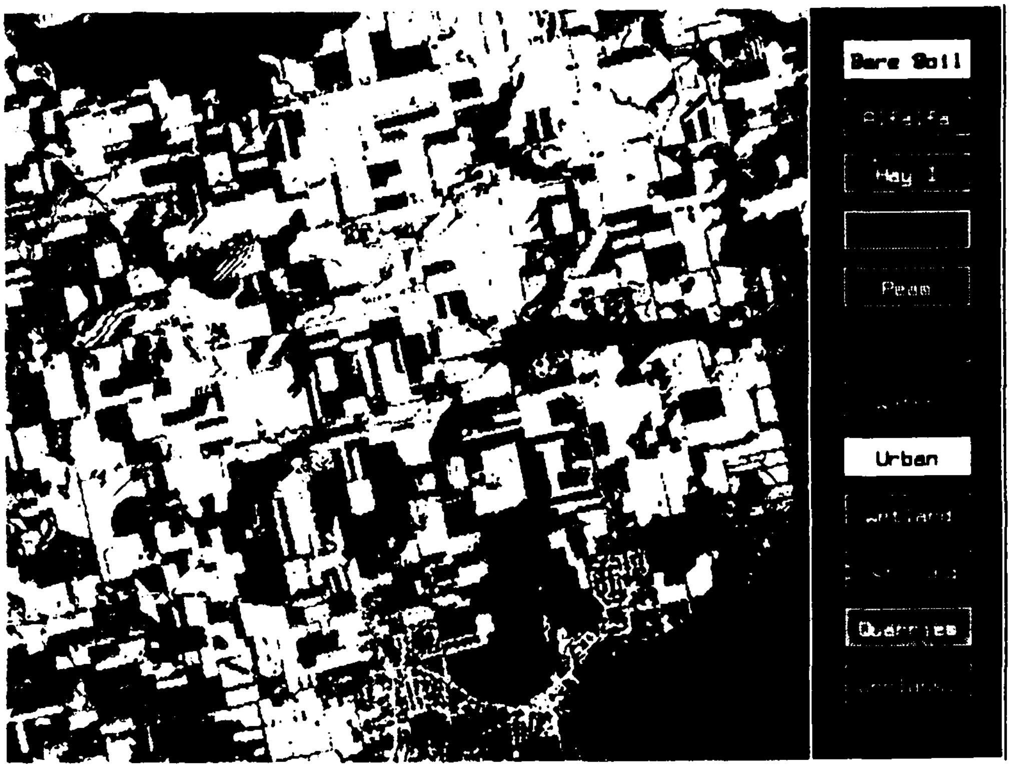


Figure 23. Resulting classification from 26 FINDCLASS clusters, original orientation.

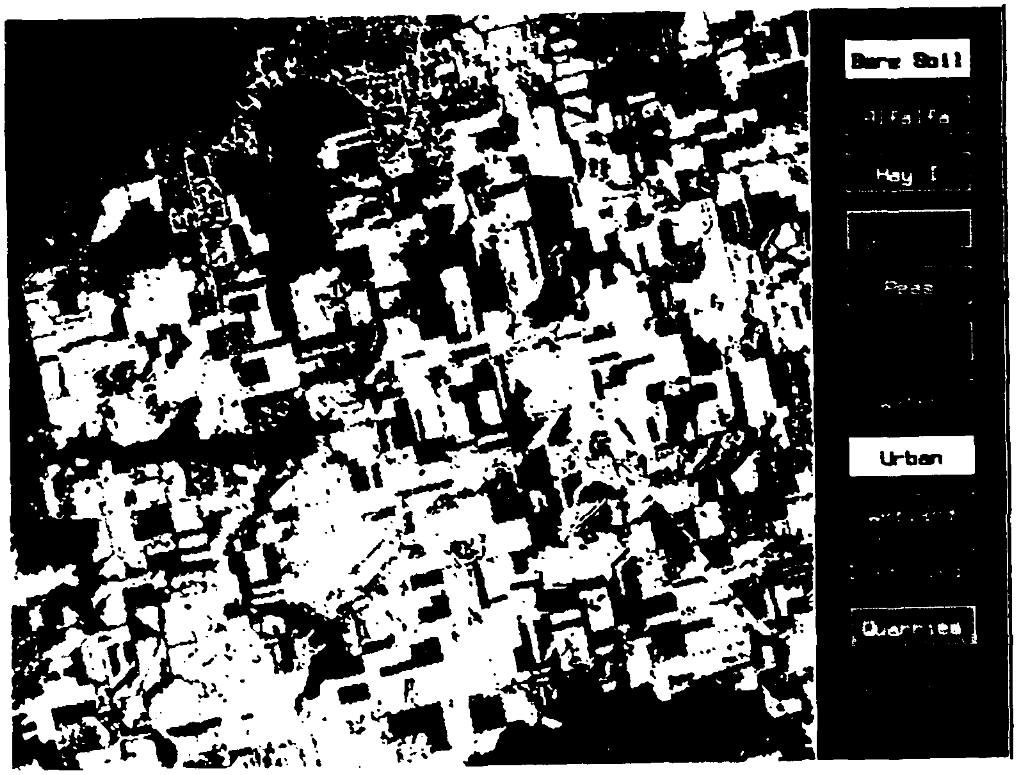


Figure 24. Resulting classification from 27 FINDCLASS clusters, rotated 180 degrees.

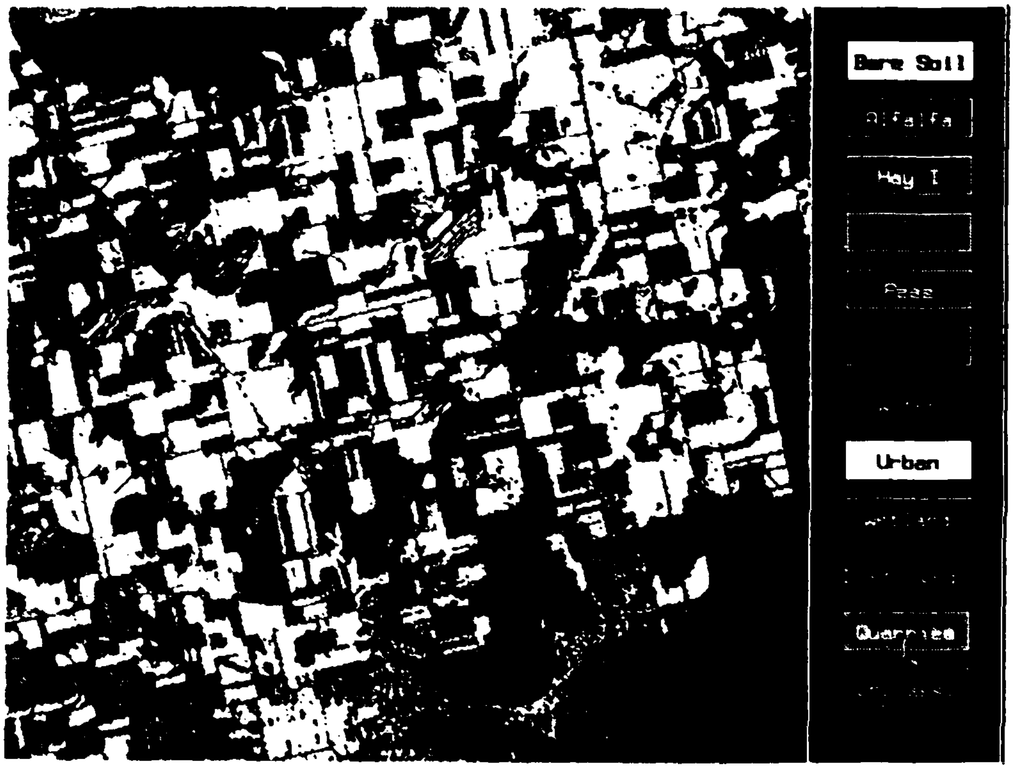


Figure 25. Resulting classification from 48 Semi-Supervised clusters, original orientation.

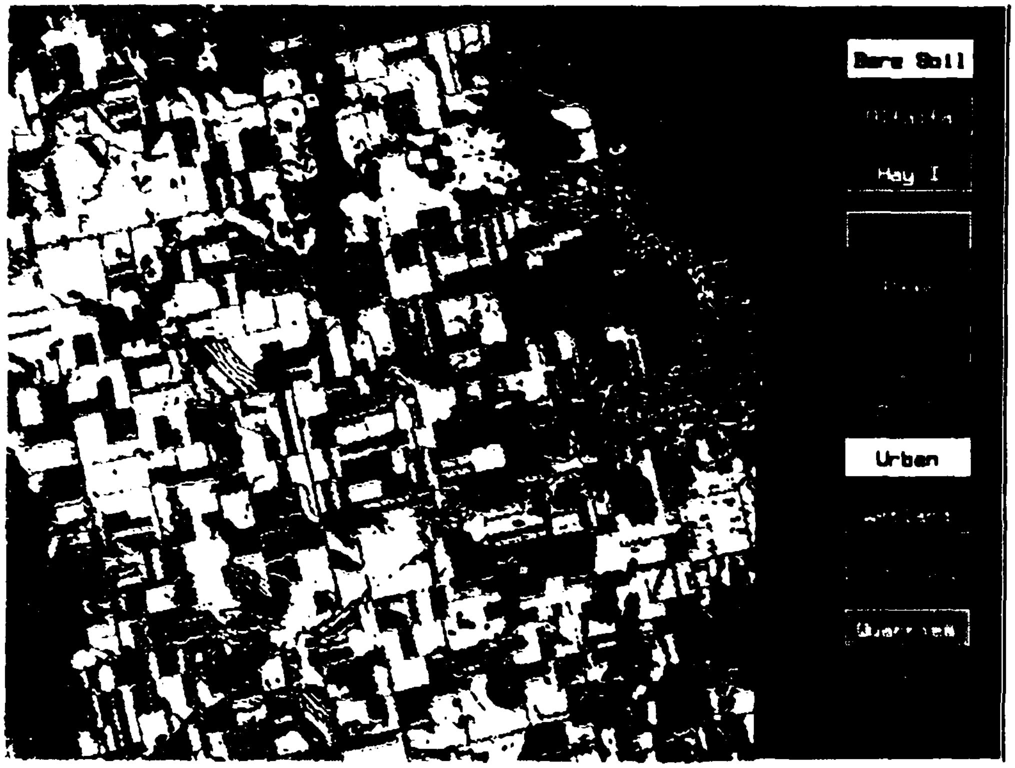


Figure 26. Resulting classification from 49 Semi-Supervised clusters, rotated 90 degrees.

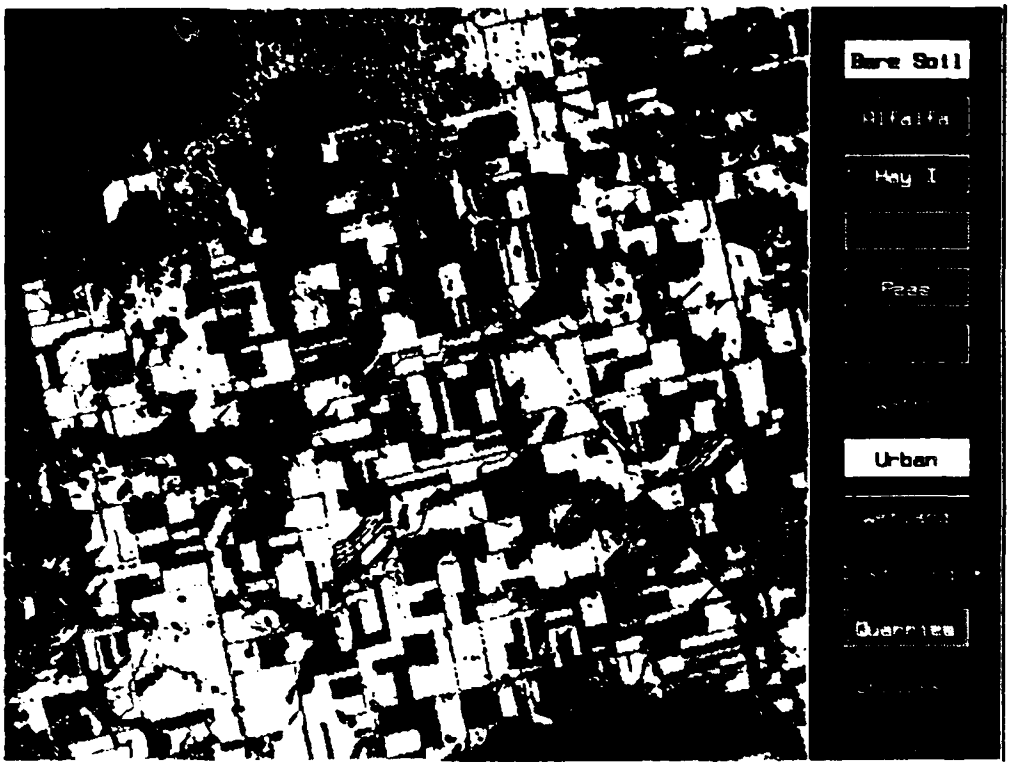


Figure 27. Resulting classification from 49 Semi-Supervised clusters, rotated 180 degrees.

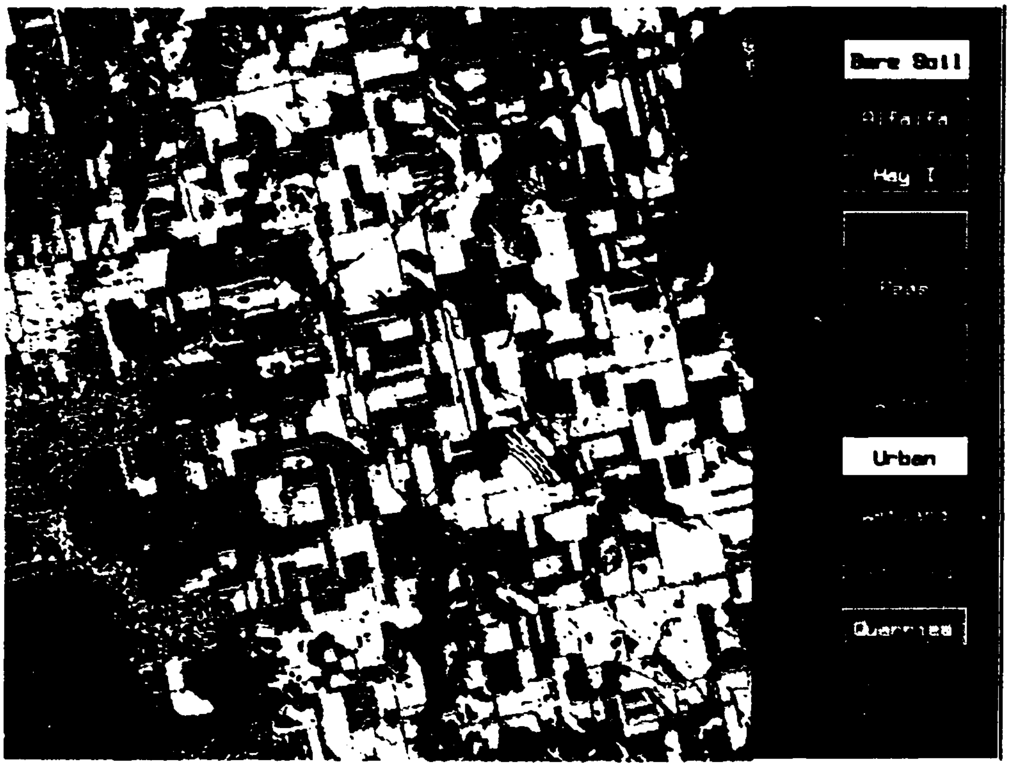


Figure 28. Resulting classification from 48 Semi-Supervised clusters, rotated 270 degrees.

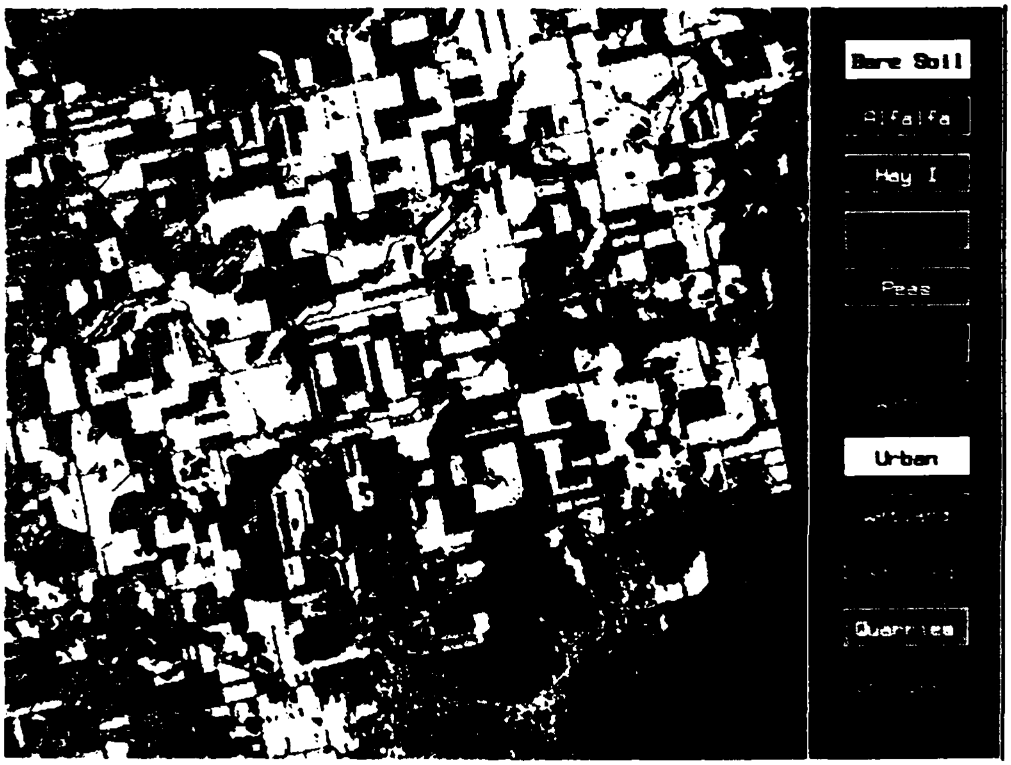


Figure 29. Resulting classification from 26 Semi-Supervised clusters, original orientation.

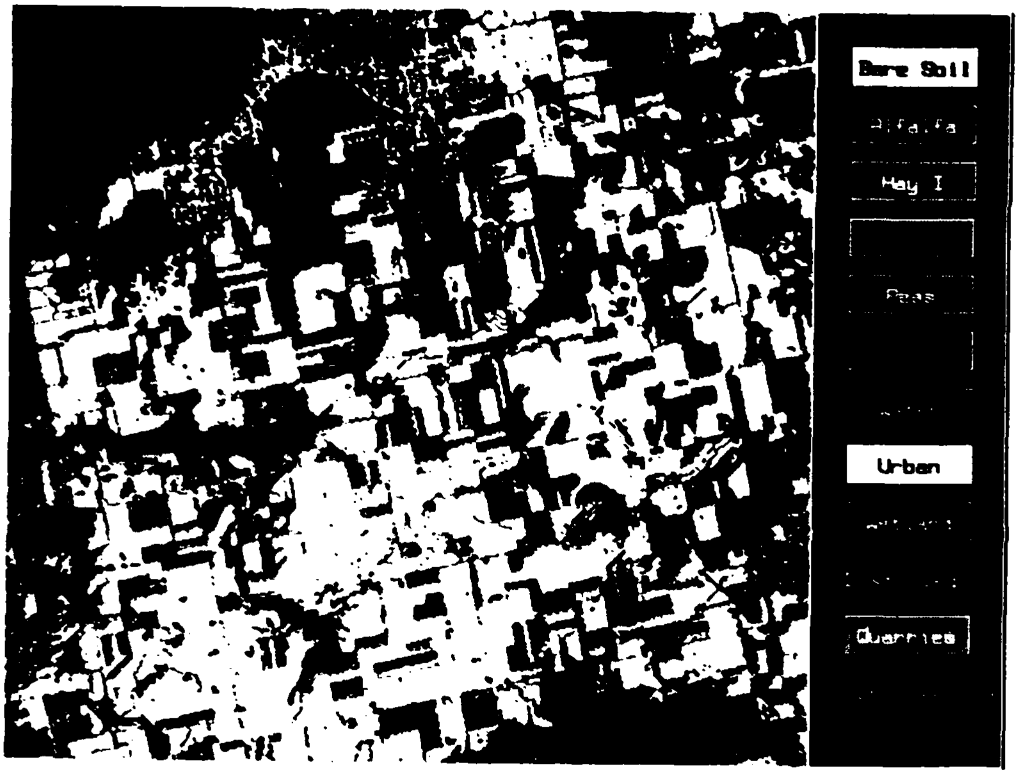


Figure 30. Resulting classification from 27 Semi-Supervised clusters, rotated 180 degrees.

5.5 Spectral Mismerging Evaluation

To evaluate spectral mismerging, the rotated classifications discussed in the previous section, figures 19 through 30, underwent accuracy assessments. Test sites were delineated in the images and the resulting contingency tables were analyzed.

Mismerging of spectral data during the clustering process can be analyzed by assessing the classification accuracy of certain cover types that possess clearly defined boundaries in the supervised thematic map of Middleton. Test sites are placed in these spectral classes, in the classified images resulting from the FINDCLASS and Semi-Supervised clustering analysis, and contingency tables are produced to evaluate the errors of omission and errors of commission.

Classes that possess clearly defined spectral boundaries are bare soil, alfalfa, hay I, hay II, and some of the forested areas. The remaining classes in the image do not possess boundaries that can be considered spectrally absolute, such as disturbed vegetation, wetland and some of the forest class. These categories have boundaries that interfinger throughout the wetland areas and other locations in the Middleton area. To quantitatively evaluate the clustering algorithms for mismerging, test sites were placed in areas known to be clearly defined classes. Test sites were not placed in the wetland and disturbed vegetation classes. Besides not having discrete boundaries, supervised training sets statistics for these classes were not easily acquired and most were not even attained from this sub-scene. Therefore, an unsupervised clustering algorithm cannot be expected to cluster quality training set statistics for these classes that may not be present. Test sites were not placed in the urban areas, disturbed lands or the roads. These are not prominent features in the image and good training sets statistics for these classes were not acquired in this sub-scene during the supervised training process.

Test sites were drawn in the supervised classified image in areas described as bare soil, alfalfa, hay I, hay II, and forest. Table 1 tabulates the site allocations.

Table 1

TEST SITES EMPLOYED

Land Cover Class	Number of Test Sites
Bare Soil	7
Alfalfa	5
Hay I	5
Hay II	5
Forest	2

5.6 Demonstration of the Semi-Supervised Two Stage Reclassification

To demonstrate the Semi-Supervised Two Stage process, the Chesapeake Bay image (Figure 14) was classified using the Semi-Supervised clustering method and then reclassified according to a Second Stage of ancillary data.

The Semi-Supervised clustering algorithm acquired 25 training sets defining the spectral diversity present in the forest canopy, grassland, water, soil and man-made features. Three different reclassifications were implemented separately according to texture, vegetation index ratios, and a polygon mask, to exemplify the 3 different options for reclassification: statistical, threshold range and threshold value, respectively.

As can be seen in figure 31, polygons were placed over regions that represented the spectral diversity of each land cover type. Four polygons were used: two in the forest canopy, one over the grassland and bare soil (this polygon is difficult to discern but is within the oval landfill), and one over the water body, Watson creek.

The Chesapeake Bay image has the potential to demonstrate the fact that cover types which are spectrally similar are often misclassified and can be reclassified using ancillary data. The shaded areas in the forest canopy and the dark wet grassy areas, assumed to be marsh, are spectrally similar and classified as forest according to the training set from the forest canopy. All the pixels in the image were classified with a maximum likelihood routine. In figure 32, the areas misclassified as forest are located primarily in the marshy grassland. More spectral statistics could be gathered from the image to possibly eliminate this problem or a reclassification with ancillary data could be attempted. To demonstrate the reclassification method the later was selected.

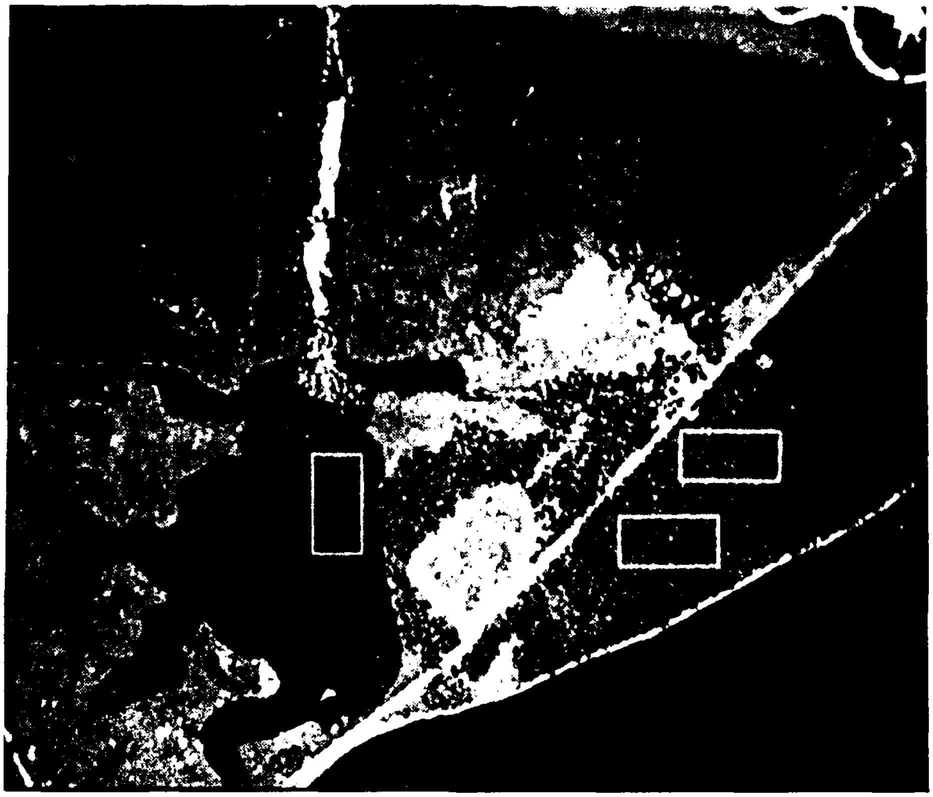


Figure 31. Chesapeake Bay study area with Semi-Supervised traininc areas delineated.

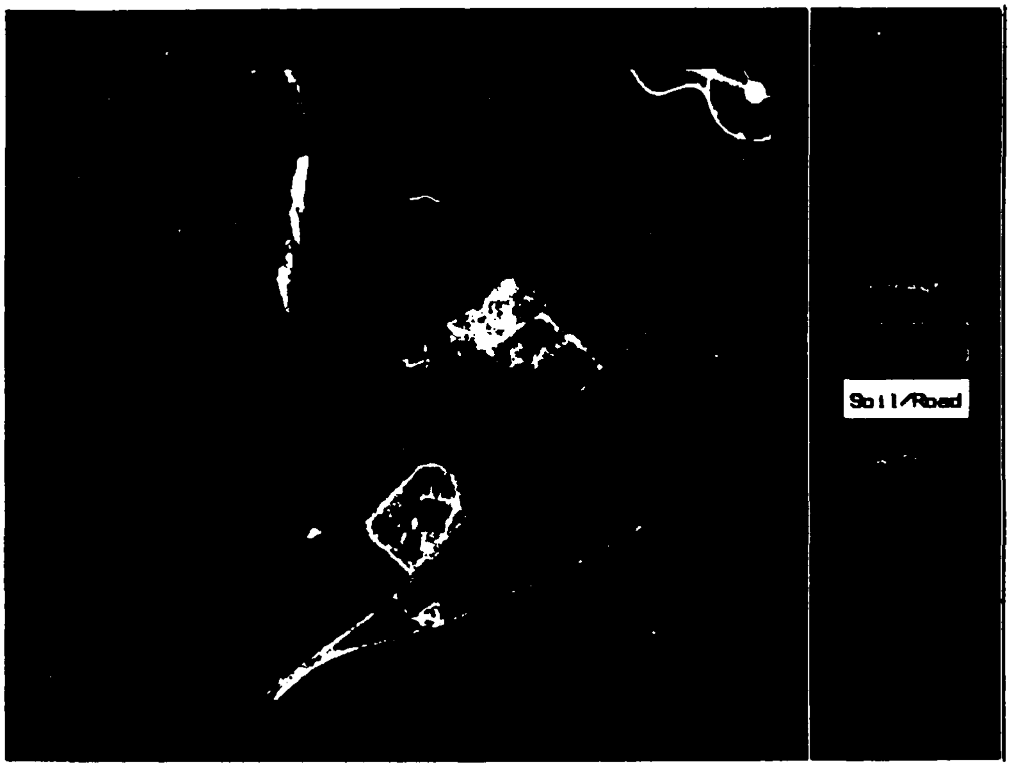


Figure 32. Chesapeake Bay Semi-Supervised spectral classification.

Reclassifications were accomplished employing textural information, vegetation index ratios and a derived polygon mask, as ancillary Second Stage data, and can be seen in figures 33 to 37. The resulting reclassified images are listed in table 2. The Second Stage data bases such as texture, vegetation index ratios and a polygon mask are described in the following sections.

Table 2

LIST OF RECLASSIFICATIONS

Figure	Type of Second Stage
33	Texture
34	Vegetation Index Ration (VI)
35	VI (smoothed)
36	Polygon Mask
37	Polygon Mask and VI

5.6.1 Texture - Second Stage

Textural statistics for each spectral training set are simultaneously acquired in the Semi-Supervised process from the textural image in figure 38.

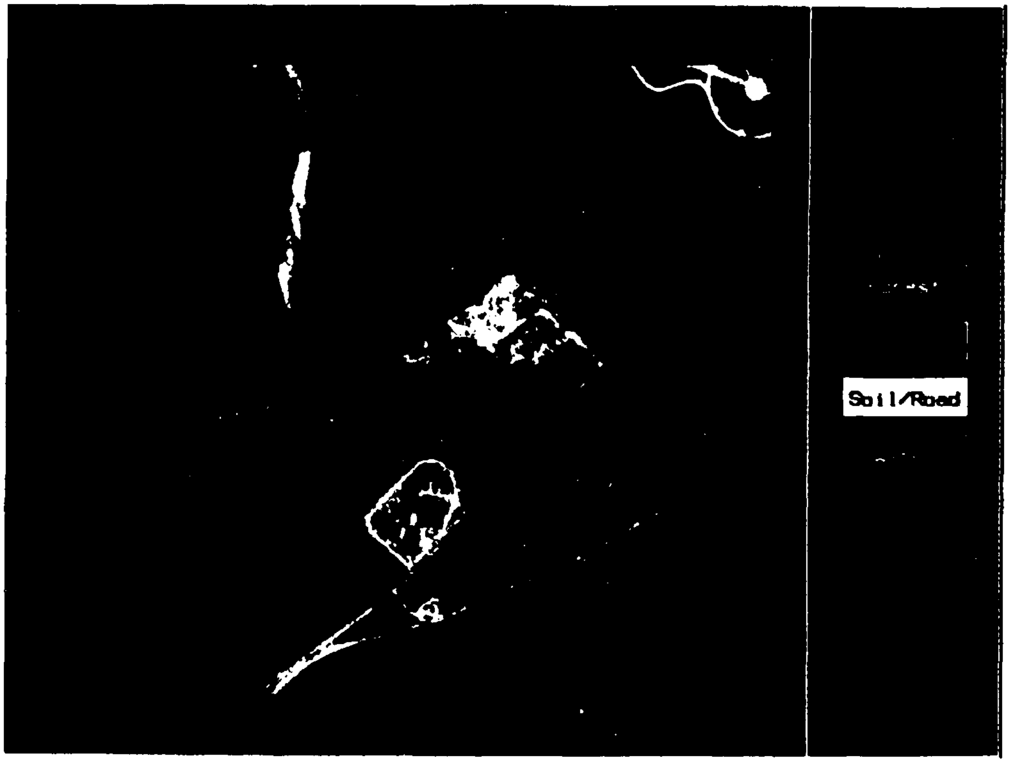


Figure 33. Chesapeake Bay area reclassification with texture.

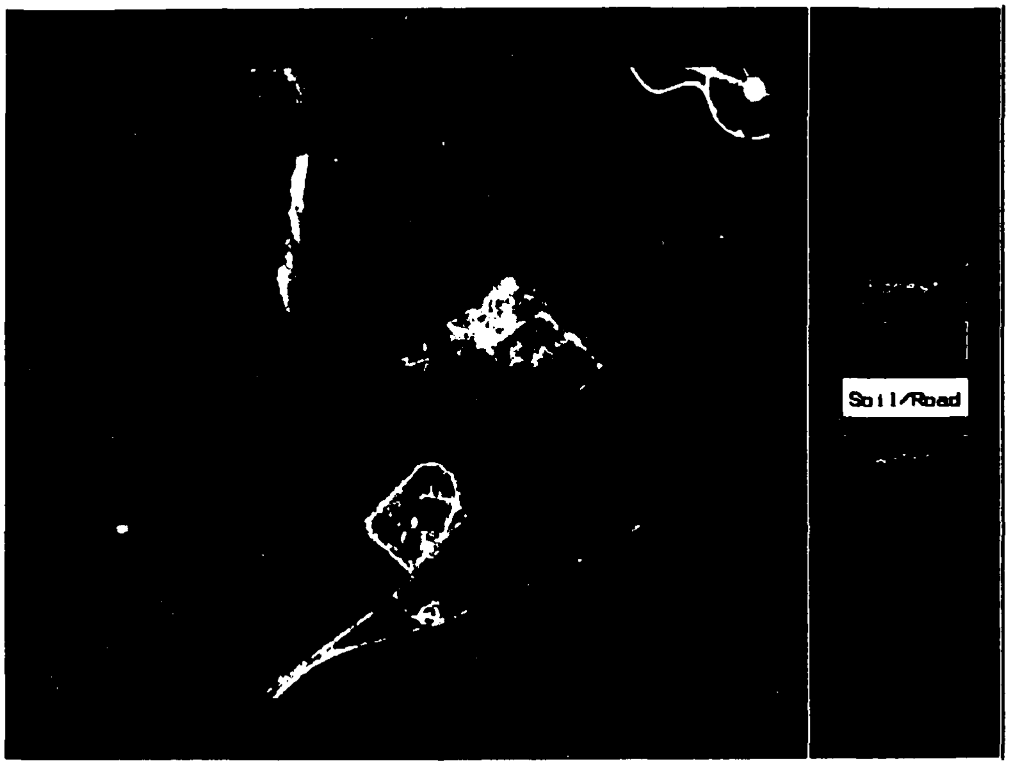


Figure 34. Chesapeake Bay area reclassification with a vegetation index ratio.

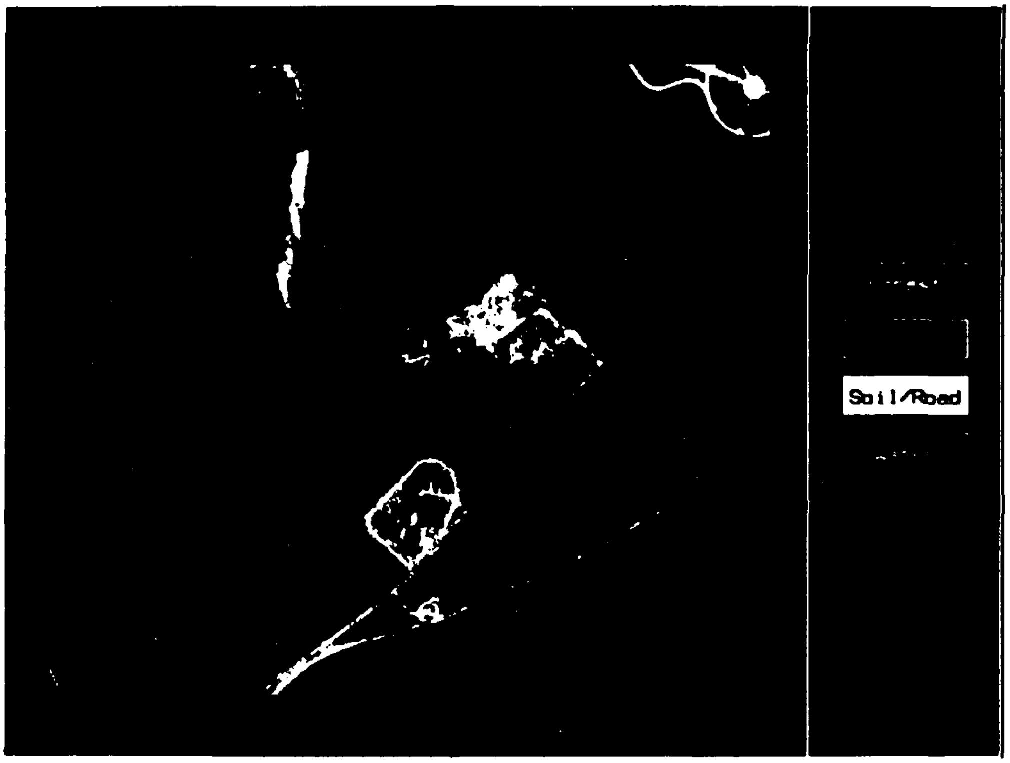


Figure 35. Chesapeake Bay area reclassified with a smoothed vegetation index ratio.

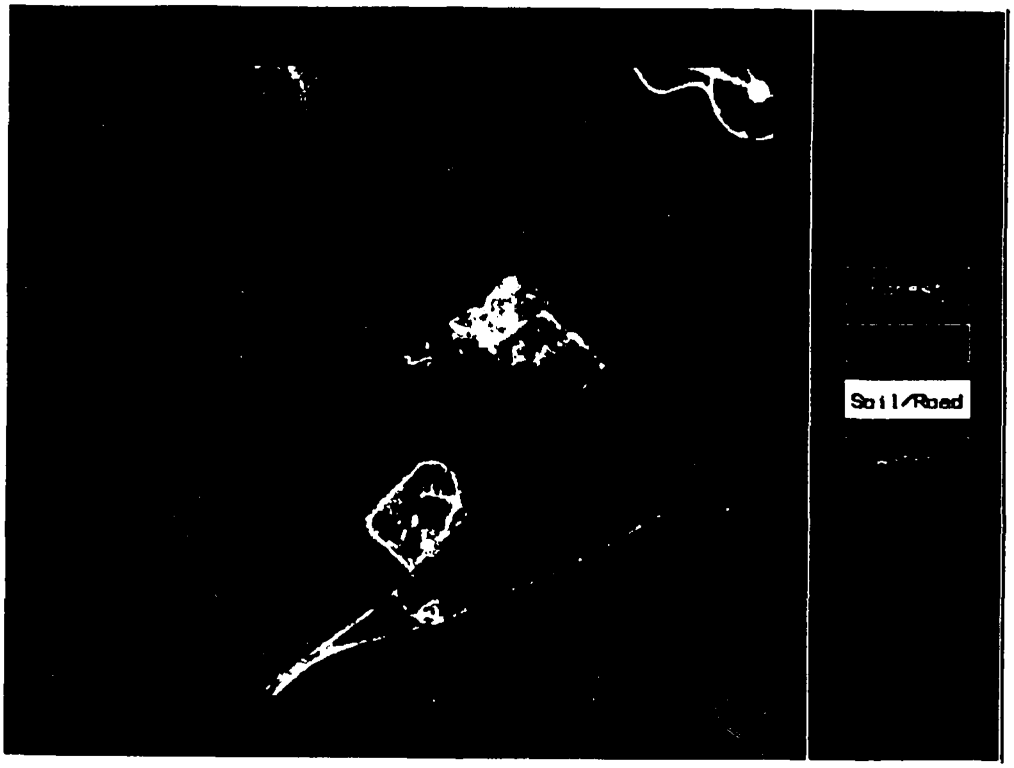


Figure 36. Chesapeake Bay area reclassified with a polygon mask.

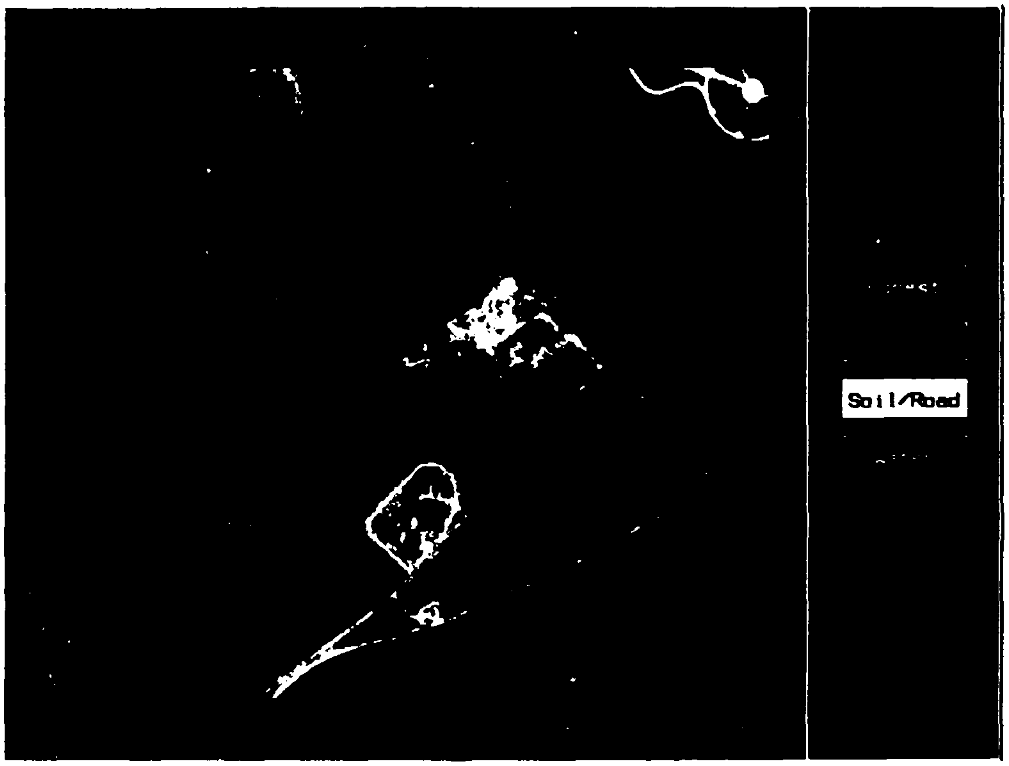


Figure 37. Chesapeake Bay area reclassified with both a vegetation index ratio and a polygon mask.

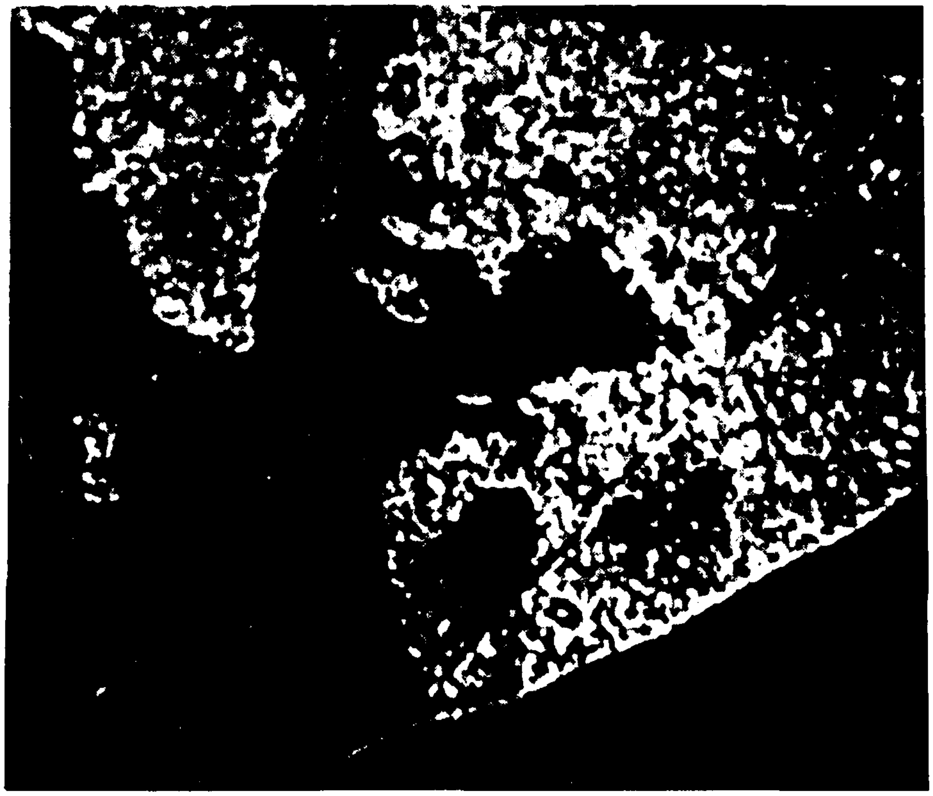


Figure 38. Chesapeake Bay texture image.

The textural algorithm that will produce the textural information should be very responsive to the textural differences between forested and non-forested land cover.

The literature used to develop the textural algorithms on the IBM AT's, at the Environmental Remote Sensing Center Micro-processing Laboratory, suggested that four algorithms are relatively ideal in discriminating between forested and non-forested land cover (Weszka et al., 1976). These are the 1st and 2nd order grey level statistic algorithms called Mean and Contrast. Of these four, the 1st order Mean and Contrast were compared, to evaluate which one was the prime discriminator. The 2nd order algorithms were not tested since (1) their computer processing time was longer than the 1st order, and (2) the difference in the textural output of the 1st and 2nd order Mean and Contrast programs was insignificant considering the extra time it takes to run the 2nd order algorithms.

The 1st order Mean program was proven more responsive to the texture than the Contrast algorithm. An effective window size, 9 x 9, was selected after many experiments. As a smaller window was implemented, the textural values within the forest canopy became patchy, salt-and-peppered gray tones, and the variance among the different textural values increased. A window size of 9 x 9, the largest one that can be selected in the program, produced more consistent textural values within the canopy, with less variance. The direction values, in the algorithm, were chosen according to the appearance of the output they produced. Using a 9 x 9 window as the delta values decrease, from 'delta row' = 5 and 'delta column' = -5, to 'delta row' = 1 and 'delta column' = -1, the output became stringy, with the appearance of noodles on a dinner plate.

During a previous project in which this area was classified with spectral and textural information, this author found that a 9x9 window caused smearing of textural information resulting in misclassification along the boundary. Since the Mean program requires a 9x9 window to completely describe the texture of the canopy, another algorithm was evaluated.

An assessment of the two 1st order textural operations, standard deviation with respect to the average, and the standard deviation with respect to the center, yielded positive results. Both algorithms were compared using a 3x3 window and the 'standard deviation with respect to the average' appeared to have the most continuous texture for describing the forest canopy. The textural image was smoothed twice with an average spatial filter to ensure as continuous and consistent a texture as could be attained (Figure 38).

In figure 38, bright pixels indicate high textural values, also referred to as a rougher texture. As discussed in chapter 3, texture may enhance edges in an image, such as the high textural values along the water and grassland interface around Watson Creek. In general, the forest canopy is represented by digital values ranging from about 70 to 119. Grass contains textural measures in the 20's. Meanwhile, the edge anomalies in the data along Watson creek have measures in the 60's and the 70's, not representative of either the grass or the water. Edge enhancement also occurs on the landfill, in the center of the image to the right, and also along the sinusoidal roads at the top of the image. A 3x3 window was used to create this textural image to reduce these edge enhancements and texture boundary smearing, in general.

A spectral variance threshold of 300 and a textural variance threshold of 70 were designated for the Semi-Supervised clustering of the Chesapeake image. A variance threshold of 300 permits the clustering of training sets with high standard deviations which are typical for spectral statistic of digitized aerial photographs with this resolution. A value of 70 was found to be a reasonable threshold for texture, affording clusters with high variance to be identified. High variance is often associated with textural information.

The polygons in figure 32 were simultaneously drawn over areas that contained textural information representative of the spectral class. Therefore, the texture statistics were successful in representing the cover type identified by each spectral class.

The red band was selected as the channel in which texture measures were computed for this image (for background information see section 3.5).

5.6.2 Vegetation Index Ratios - Second Stage

Two vegetation index rations were calculated from the spectral data: (1) the infrared band divided by the red band (IR/RED), and (2) RED/IR. After the ratio is computed, the resulting values in the image are stretched linearly to normalize the values from 0 - 255. In the RED/IR ratio low digital values 0 - 6 represent the forest canopy. This is because low digital values in the red band are divided by high digital values in the infrared band resulting in small numbers, that are still low ranging numbers when stretched linearly over a range of 0 to 255. The IR/RED ratio resulted in high values representing the forest. These threshold ranges were identified using an interactive program that level slices ranges or values of an image on the computer monitor. Through visual analysis it was determined that the IR/RED ratio was slightly more descriptive of the canopy, with digital values 21 - 255 selected and displayed in figure 39.

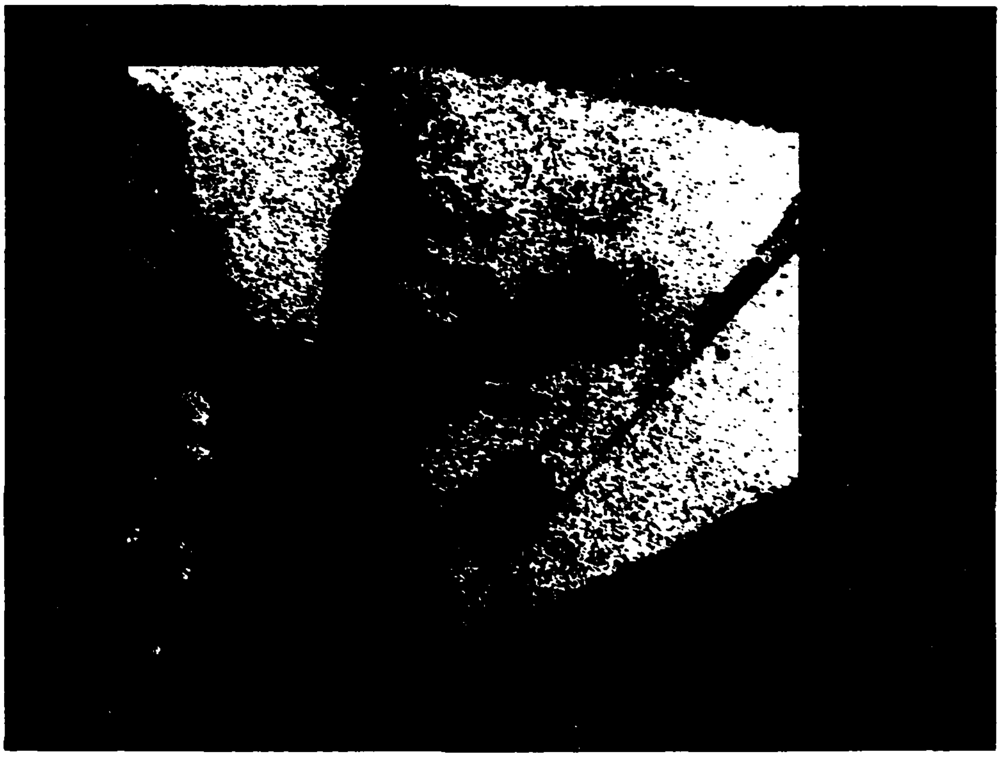


Figure 39. Chesapeake Bay vegetation index ratio.

To produce a more continuous representation of the forest canopy in the vegetation index ratio the image was smoothed employing an average spatial filtering operation (Figure 40).

As an aside, in the vegetation index ratio, image values 0 - 9 identified water bodies, 10 - 20 grass, soil and manmade features, and 21 - 255 represented the forest canopy. The threshold range 10 - 20 could be broken up into ranges that would separate grass and manmade features if the analyst were interested in doing so.

Both ratio images figure 39 and 40 were employed in the reclassification process.

5.6.3 Polygon Mask - Second Stage

The specular reflection in Watson Creek can be reclassified by implementing a polygon mask.

A polygon mask is created by altering all the values in an image delineated by a polygon to a specific digital number, with the assistance of an image-processing program.

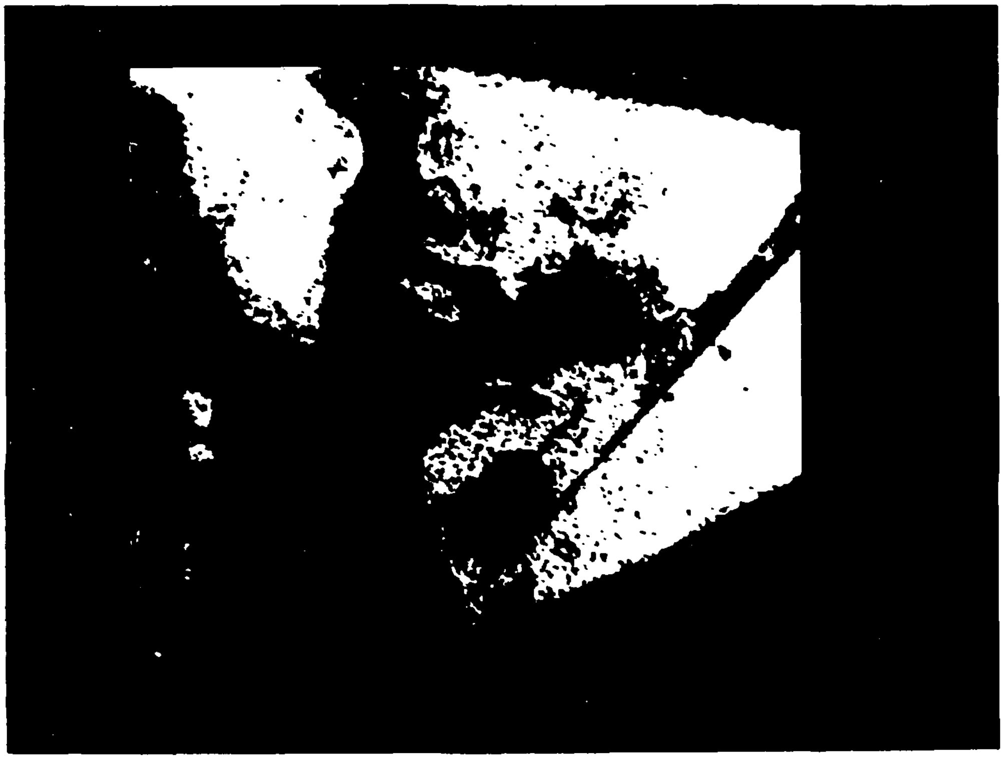


Figure 40. Chesapeake Bay smoothed vegetation index ratio.

A polygon mask can be placed into any band of data; in this experiment, the red band was selected. With an interactive program a cursor was used to draw a polygon around the area in Watson Creek that contains the specular reflection. The values within the polygon were altered to zero and are displayed in black in figure 41.

5.6.4 The Second Stage Reclassification

Three spectral forest classes were identified in the Chesapeake Bay classification for reclassification. The classes describe the shaded areas in the forest canopy and the dark wet grassland; and the spectrally vivid red found in both the forest canopy and areas in the grassland. The textural statistics for these classes represented the forest canopy; and therefore, the pixels classified in the grassland and marshes were considered misclassified and candidates for reclassification according to textural information. The textural values for these misclassified pixels matched those of the textural statistics for grassland, and therefore should be reclassified into the grassland category with little difficulty.

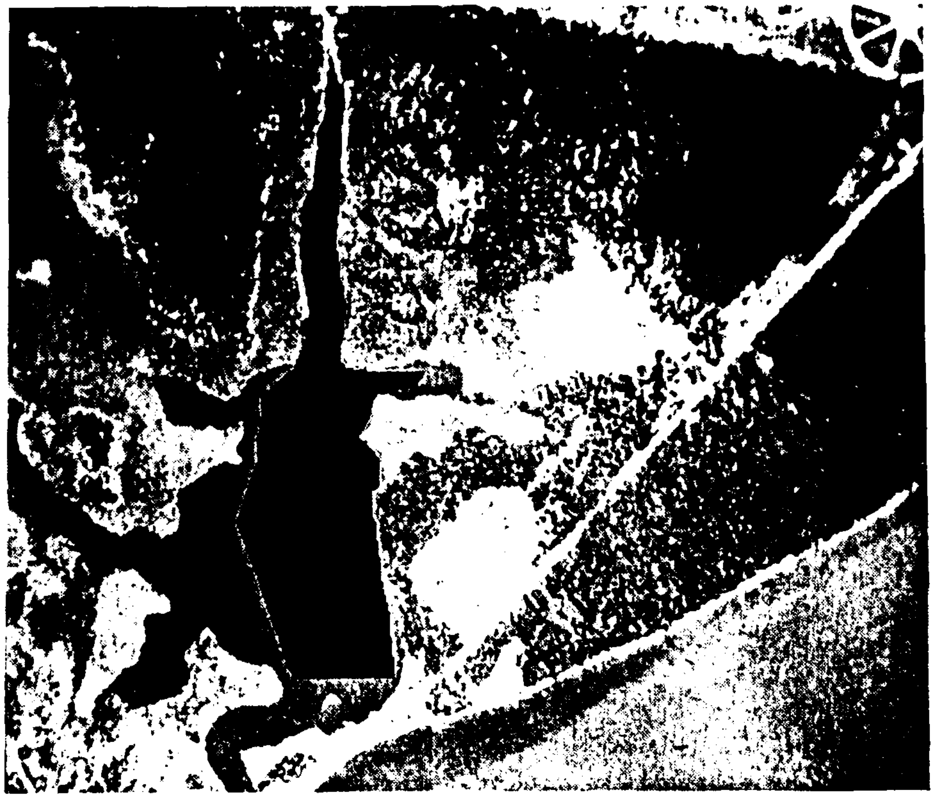


Figure 41. Chesapeake B~y polygon mask.

These same three classes could be reclassified according to ranges designated in the vegetation index ration Second Stage. The range delineating the forest canopy could be employed. If a pixel classified as one of the above three classes did not have a Second Stage value in the forest canopy range, then it would be reclassified into a designated grass category.

To reclassify the specular reflection with a polygon mask Second Stage, the classes that describe the specular reflection are designated for reclassification. A threshold value of zero is chosen for the Second Stage. The identified categories in the classified image are reclassified into a designated water class if the pixel's Second Stage digital value is less than or equal to the threshold.

5.6.5 Assessment of the Reclassified Images

The analysis of the reclassified images, figures 33 to 37, is both quantitative and qualitative in nature. Histograms were analyzed to evaluate the significance of the changes. Comparison image files were created to visually show the pixels that were reclassified in each image. The comparison images for each reclassification are listed in table 3.

Table 3

LIST OF COMPARISON IMAGES

Figure	Images Compared
42	Figures 32 and 33
43	Figures 32 and 34
44	Figures 32 and 35
45	Figures 32 and 36

The comparison image in figure 42 depicts the pixels reclassified by texture. Figures 43 and 44 identifies those pixels reclassified by a vegetation index ratio and smoothed vegetation ratio, respectively. Figure 45 illustrates the reclassified pixels comprising the specular reflection in the Chesapeake Bay image.

The results of the experimentation, discussed in the above sections, are detailed in the following chapter.

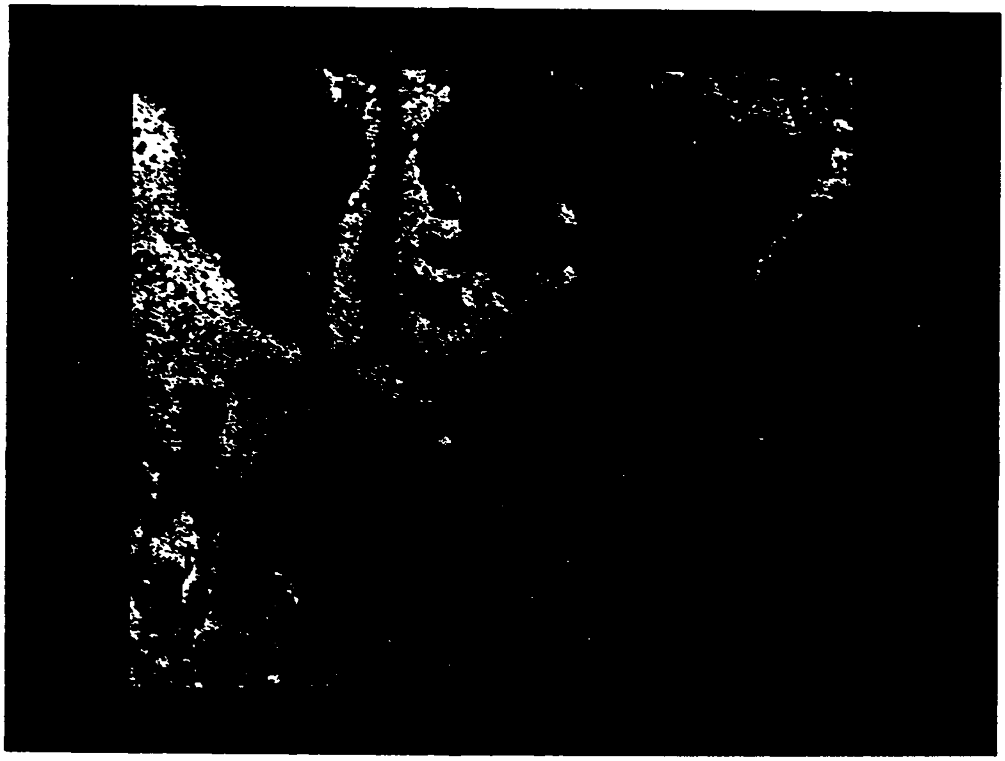


Figure 42. Image pixels in the Chesapeake Bay region that were reclassified according to texture.

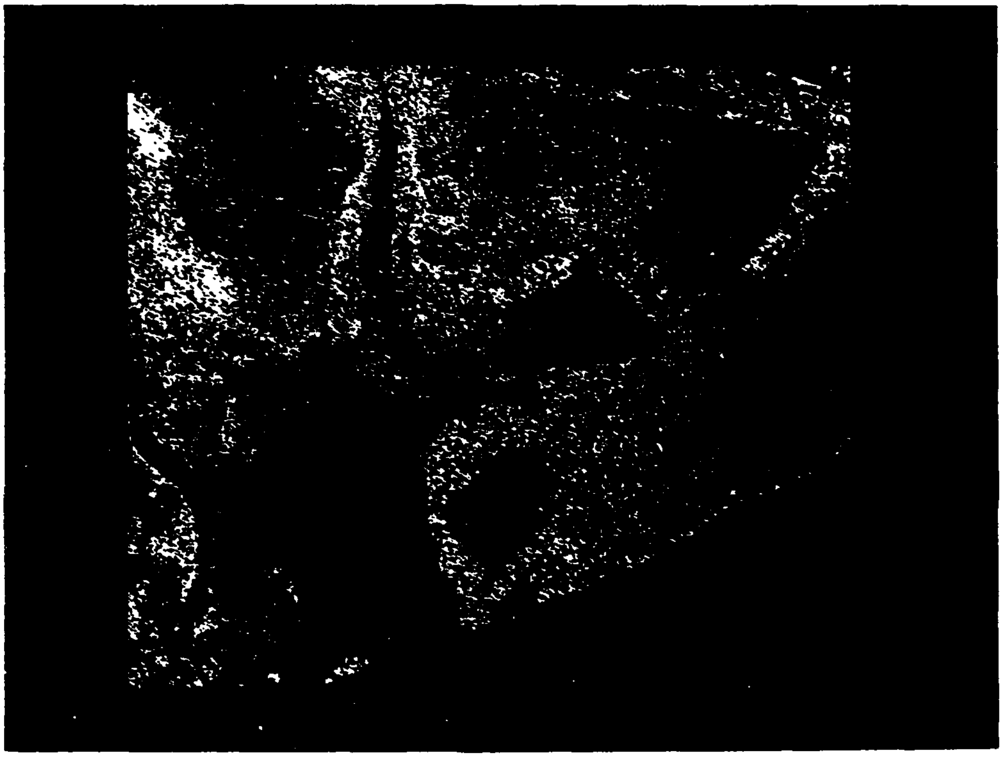


Figure 43. Image pixels in the Chesapeake Bay region that were reclassified according to an unsmoothed vegetation index ratio.

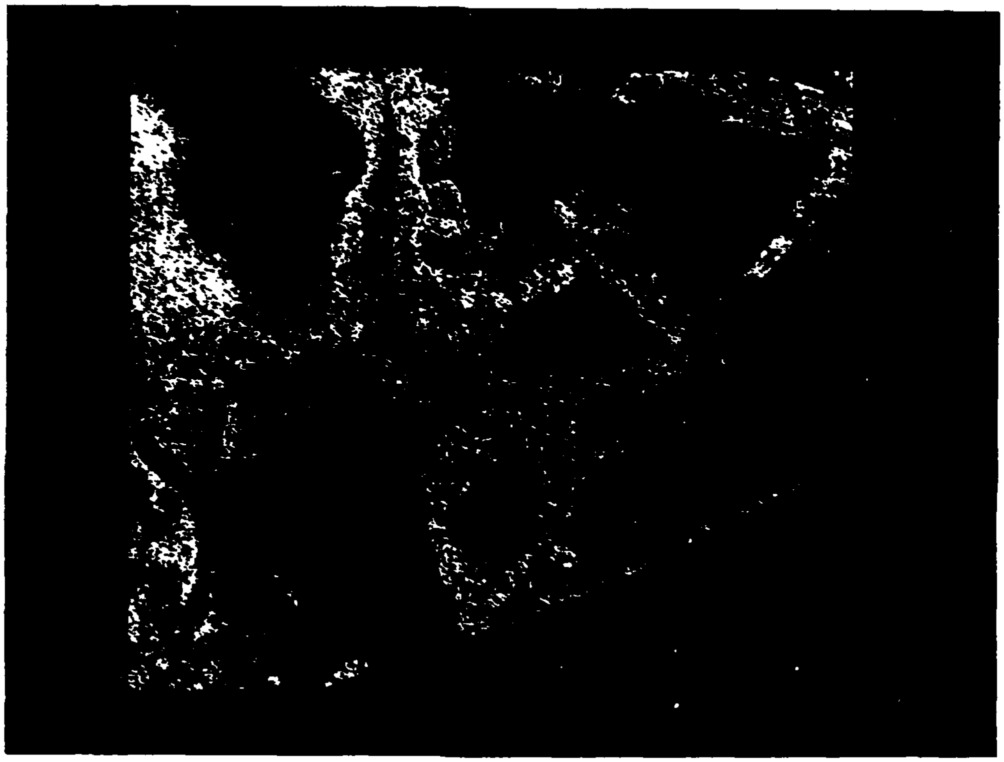


Figure 44. Image pixels in the Chesapeake Bay region that were reclassified according to a smoothed vegetation index ratio.

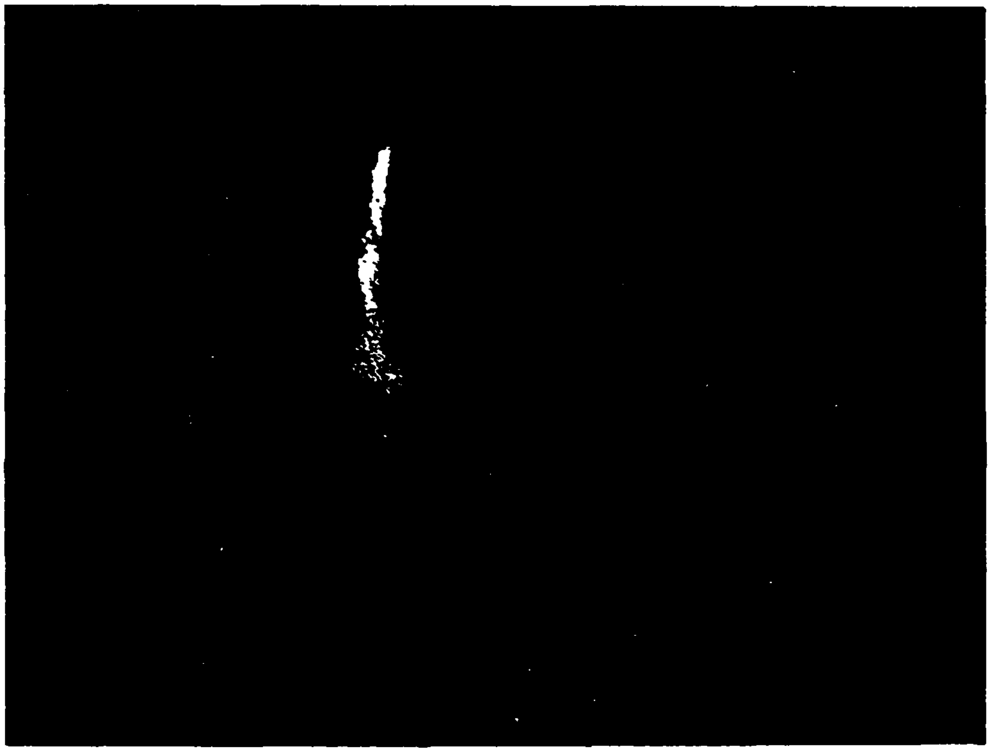


Figure 45. Image pixels in the Chesapeake Bay region that were reclassified according to a polygon mask.

Chapter VI

Results

6.1 Introduction

The Semi-Supervised clustering technique was employed to identify training set statistics in two images, the Middleton image (Figure 13) and Chesapeake Bay region (Figure 14). The post-classification Second Stage reclassification was implemented to increase the classification accuracy of the Chesapeake Bay image. In the following chapter, results involving these two experiments will be discussed. But first, the effects of the transformed divergence criterion, in the Semi-Supervised algorithm, and the statistical validity of the resulting training sets, from the FINDCLASS and Semi-Supervised clustering, will be detailed.

6.2 Effects of the Transformed Divergence Calculation

The transformed divergence (TD) criterion was implemented to prevent homogeneous windows of pixels in the image from being mismerged with the seed clusters. The principle being, that the initial seeds represent resources of interest that the analyst has directed the Semi-Supervised algorithm to describe. Since it is an analyst- directed operation, some of the resources in an image may not be represented by an initial seed. Therefore, if every cluster that was considered homogeneous were merged with one of the seeds according to a statistical distance metric, mismerging is a likely result. A transformed divergence criterion would not permit a cluster to be merged unless it met a user defined transform divergence threshold value.

To evaluate the difference that the transformed divergence function has on cluster merging, Semi-Supervised seeds were merged with clusters using a transformed divergence requirement and not implementing the specification. The resulting training set statistics were analyzed. Although it is not evident that all the statistical results were improved because of the transformed divergence calculation, the results were supportive of this theory.

In both cases, the training set statistics were printed out before and after clusters from the whole image were merged with the seeds. In some instances, fewer clusters were merged to some seeds when the transformed divergence computation was employed. In comparing the statistics before and after, the mean values of the initial clusters were often altered less significantly, and the standard deviations (an indication of variance) were lower or more similar to those of the seeds when employing the transformed divergence calculation.

This is demonstrated by two extreme cases; on the average, the changes were not as significant, or made little difference in the small subset image on which this experiment was executed. The examples are from a training set document of two digital bands of data (Figure 46a and 46b).

The seed cluster number 7 (Figure 46a) had 18 pixels at the start; and when merging took place without the TD function, the number of pixels increased to 90 and the means in both bands were significantly altered. However, when the TD calculation was implemented, the number of pixels resulting reduced to 63 and the means were not altered as significantly as before. The standard deviations were lower with the TD computation indicating that

Initial Seed

```
:Set #:              7

:Set Name:  tree

:N Pixels:              18

:Means:
                    61.61111  144.16670

:Standard Deviations:
                    9.96547  12.48188

:Covariance Matrix:
                   99.31046  47.30392
                   47.30392  155.79410
```

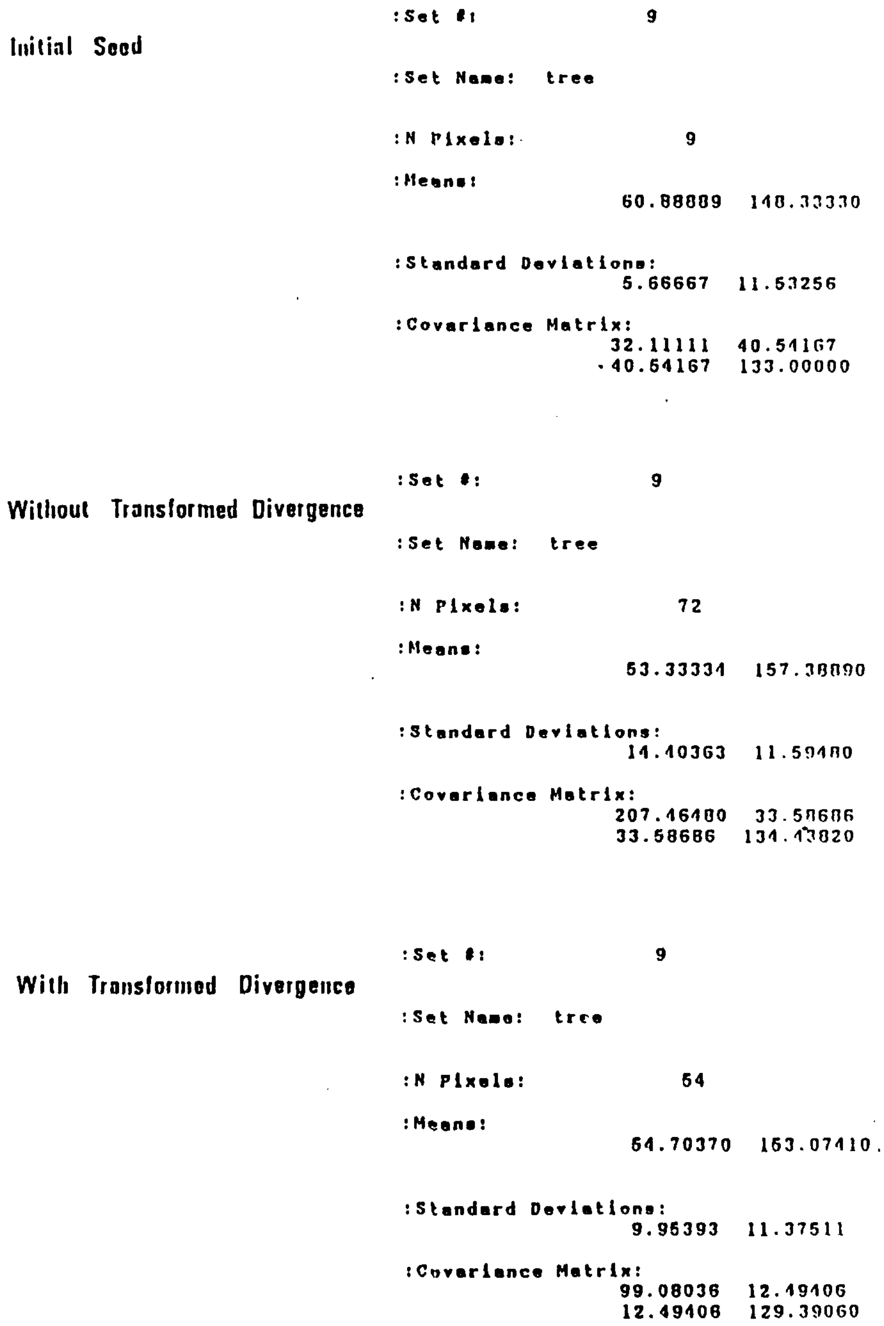
Without Transformed Divergence

```
:Set #:              7

:Set Name:  tree

:N Pixels:              90

:Means:
                    46.72222  152.40000

:Standard Deviations:
                    18.42876  11.31836

:Covariance Matrix:
                   339.61860  -47.66292
                   -47.66292  128.10790
```

With Transformed Divergence

```
:Set #:              7

:Set Name:  tree

:N Pixels:              63

:Means:
                    51.44444  138.80950

:Standard Deviations:
                    10.73701  11.27923

:Covariance Matrix:
                   115.28320  28.71505
                   28.71605  127.22120
```

Figure 46a. The first example of training set statistics created without the implementation of the transformed divergence calculation and with transformed divergence.



```
                                 :Set #:              9
Initial  Seed
                                 :Set Name:  tree

                                 :N Pixels:              9

                                 :Means:
                                                60.88889  148.33330

                                 :Standard Deviations:
                                                 5.66667  11.53256

                                 :Covariance Matrix:
                                                32.11111  40.54167
                                                40.64167  133.00000


                                 :Set #:              9
Without  Transformed Divergence
                                 :Set Name:  tree

                                 :N Pixels:              72

                                 :Means:
                                                53.33334  157.38890

                                 :Standard Deviations:
                                                14.40363  11.59480

                                 :Covariance Matrix:
                                               207.46480  33.58686
                                                33.58686  134.43820


                                 :Set #:              9
With  Transformed  Divergence
                                 :Set Name:  tree

                                 :N Pixels:              54

                                 :Means:
                                                54.70370  163.07410

                                 :Standard Deviations:
                                                 9.95393  11.37511

                                 :Covariance Matrix:
                                                99.08036  12.49406
                                                12.49406  129.39060
```

Figure 46b. The second example of training set statistics created without the implementation of the transformed divergence calculation and with transformed divergence.

mismerging was possibly reduced. As a side note, the first band is textural data and the second is a spectral band; both are from a digitized aerial photograph with high resolution, and high standard deviations are expected in the training set statistics. Similar evidence is presented by training set number 9 in figure 46b.

In the previous chapter, it is noted that there is a spectral class representing the forest canopy, in the Chesapeake image, that is spectrally similar to the marsh areas. The initial texture statistics represented high texture values of the forest canopy before the merging began. After the merging process the textural statistics gathered without the TD calculation were changed and represented a smoother texture. This is probably due to mismerging of the smoother textural statistics representing the marsh areas. With the TD calculation, the mismerging was prevented and the final textural statistics still represented the forest canopy. These statistics could then be used to reclassify the spectrally misclassified pixels of this class in the image.

6.3 The Statistical Validity of Training Sets

Statistical validity involves the number of pixels sampled to represent a given population statistically. A training set is considered statistically significant if it has at least 10n pixels. The n refers to the number of digital bands of data. Since 3 bands of data were used, in this instance, a statistically significant number of pixels would be 30.

The statistical validity of the training set output from the Semi-Supervised and FINDCLASS clustering algorithms was analyzed.

The FINDCLASS algorithm consistently produces numerous clusters that have only 9 pixels. These classes are obviously statistically invalid. The Semi-Supervised approach creates fewer training sets containing a statistically invalid number of pixels. The number of training sets that resulted in 18 pixels or less were tabulated. Table 4 lists the results from training sets files for two images.

Table 4

NUMBER OF CLASSES WITH 18 PIXELS OR LESS

Image	Algorithm	9 pixels	18 pixels
1	FINDCLASS	6	7
	Semi-Supervised	0	3
2	FINDCLASS	6	3
	Semi-Supervised	0	3

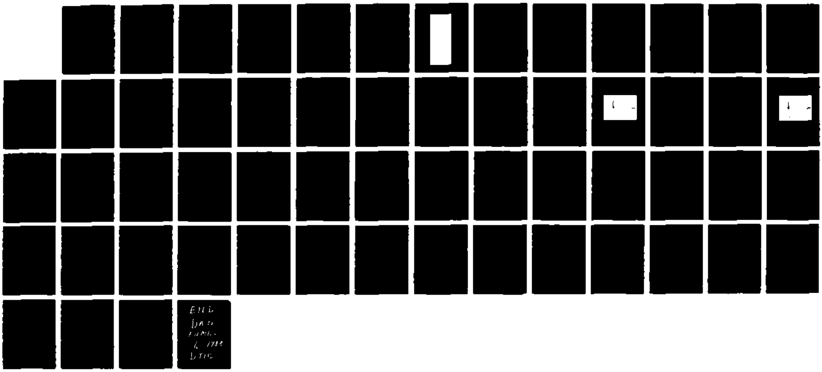
The FINDCLASS algorithm can be considered 'top heavy' regarding the number of pixels in the training sets it identifies first. Figure 47 contains the number of pixels per training set from the FINDCLASS and Semi-Supervised clustering approaches. As FINDCLASS processes the image, each new cluster identified is considered the fiftieth cluster, and the two most similar clusters are merged. This continues throughout the whole image, and therefore it is possible that the final clusters identified in the image are represented by the training sets that contain 9 or 18 pixels. These sets may not have had the opportunity to merge with spectral data from the entire image, whereas the earlier clusters have done so. One training set identified by FINDCLASS to classify figure 23 contained 18% of the image pixels, 45794 pixels.

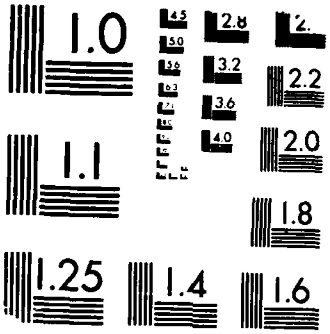
FINDCLASS		Semi-Supervised	
Cluster #	# Pixels	Cluster #	# Pixels
1	23238	1	1728
2	10332	2	585
3	9495	3	99
4	2331	4	117
5	4590	5	864
6	6723	6	459
7	6165	7	367
8	3681	8	1818
9	3024	9	1008
10	7767	10	657
11	126	11	1359
12	3717	12	216
13	2430	13	621
14	342	14	540
15	666	15	261
16	513	16	333
17	216	17	36
18	2016	18	72
19	9	19	711
20	3789	20	1368
21	720	21	531
22	819	22	2070
23	243	23	207
24	729	24	468
25	63	25	279
26	27	26	603
27	36	27	117
28	4275	28	369
29	126	29	783
30	2502	30	387
31	121	31	99
32	36	32	351
33	54	33	72
34	27	34	441
35	333	35	828
36	18	36	756
37	9	37	243
38	18	38	423
39	45	39	306
40	36	40	126
41	9	41	99
42	18	42	63
43	9	43	558
44	9	44	27
45	18	45	126
46	18	46	2259
47	18	47	2115
48	18	48	1620
49	9	49	63

Figure 47. The number of pixels per training set from the FINDCLASS and Semi-Supervised clustering approaches.

AD-A192 525 SEMI-SUPERVISED TWO STAGE CLASSIFICATION TECHNIQUE(U) 3/3
ARMY MILITARY PERSONNEL CENTER ALEXANDRIA VA
D A TOOMEY 31 JUL 87

UNCLASSIFIED F/G 14/5 NL





MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

The Semi-Supervised technique produces more training sets with valid distributions, indicating that each cluster represents spectral samples from the entire image, as can be seen in columns 3 an 4 of figure 47. The Semi-Supervised approach creates training set seeds in the polygons directed by the user, and each seed grows by collecting similar spectral signatures from throughout the image.

Another disquieting occurrence was noted regarding FINDCLASS. It identified a training set in two of the rotated Middleton images that could not be implemented into the maximum likelihood classification algorithm. Two problems emerged, the training sets covariance matrices could not be inverted and a negative determinant was calculated. The two training sets are presented in figure 48. Curiously enough, both training sets have only 9 pixels but the covariances between the 1st and 3rd and 2nd and 3rd channel indicates an uncommonly vast distribution of spectral data. These statistics were not used in the classification of the image and could not have been implemented in any event.

```
:Set #:        1

:N Pixels:     9

:Means:
                     51.556       36.222        0.000

:Standard Deviations:
                      1.740        2.635        4.475

:Covariance Matrix:
                      3.028        4.486     6287.000
                      4.486        6.944     4414.375
                   6287.000     4414.375       20.028



:Set #:        49

:N Pixels:     9

:Means:
                     51.556       36.222       19.889

:Standard Deviations:
                      1.740        2.635        4.475

:Covariance Matrix:
                      3.028        4.486     5133.444
                      4.486        6.944     3603.903
                   5133.444     3603.903       20.028
```

Figure 48. Anomalous training sets within the statistical output of the FINDCLASS algorithm.

6.4 Improper Application of the FINDSET Algorithm

In section 5.3, it was mentioned that the FINDSET program did not completely access the image. Before this error was discovered, however, FINDSET was employed to cluster data in rotated images to evaluate clustering bias discussed in section 2.11. The Middleton image in figure 13 was rotated and analyzed. Since the program only accessed the first 240 columns of the image, when the image was in the typical orientation, as in figure 18, no training sets for water were identified. However, when the image was rotated to the orientation in figure 21, nine training sets were found.

The resulting statistics were used in a minimum distance to mean classification algorithm which classifies all the pixels in an image. Water would be classified whether or not there were spectral statistics describing this land cover. At first, it appeared that the FINDSET algorithm was biased towards the dominant land cover type of the image, water, when it was the first class the algorithm analyzed in the upper left hand corner. This was realized to not be the case, since the algorithm never accessed the water in the image (Figure 13 orientation) and never acquired statistics for water, but the minimum distance to mean classifier still classified the water with an organic soil spectral training set ('z' in Figure 16). The FINDCLASS algorithm discussed in section 5.3 accesses the entire image and is the algorithm comparatively evaluated within this thesis research.

6.5 Evaluation of Clustering Bias

The Middleton image was selected because of its diverse land cover types in the entire image accompanied by one dominant land cover class in the corner of the scene. The dominant cover type could be positioned in all four corners of the image, permitting the evaluation of clustering bias.

FINDCLASS identified 8 or 9 training sets for water in each of the four rotations depicted in figures 19 through 22. Table 5 summarizes the number of training sets identified for water in each image.

Table 5

NUMBER OF FINDCLASS TRAINING SETS FOR WATER

FINDCLASS - 50 clusters

Figure	Number of Training Sets for Water
19	9
20	9
21	8
22	8

The number of training sets that described water were not significantly different, suggesting that there is no dependence upon the rotation of the image. Therefore, from this image it cannot be concluded that clustering bias is a significant problem in the FINDCLASS approach to clustering. Another experiment was devised to retest this possibility.

Another area within the SPOT satellite scene over the Cherokee marsh in Madison was selected, 200 columns by 480 rows (Figure 49). In its original orientation, the upper half of the image is agricultural crops and a river which leads to the lower half of the scene which is entirely water. This image underwent clustering analysis at four different rotations. The number of clusters describing the water, from 50 clusters requested for each rotation

Figure 49. Cherokee marsh sub-scene, SPOT satellite image.

are listed in table 6. The orientations in table 6 have the following meaning: number 1 is the original orientation seen in figure 49, 2 is the original image rotated 180 degrees, 3 is the original rotated 90 degrees counter-clockwise, and number 4 is the original rotated 90 degrees clockwise.

Table 6

NUMBER OF WATER CLASSES - CHEROKEE MARSH

	Orientation	Number of Water Classes
1	Water Bottom Section	15
2	Water Top Section	11
3	Water Right Half	12
4	Water Left Half	13

Again, a consistent pattern indicating clustering bias is not present in the Cherokee image. In this case, the water body dominated over half of the image in each rotation but the results were still inconclusive.

Although the number of classes describing water in the 4 rotations of the Middleton image were nearly identical, the fact remains that there were an overabundant number of classes (at least 8) devoted to segmenting a relatively simple cover type. Nine training sets are not required to classify water accurately. In the practicum analysis, only 5 spectral classes were used to segment water in an image six times the size of the Middleton image.

The Semi-Supervised approach adequately classified water by directing that 3 to 4 clusters define the water class (Table 7).

Table 7

NUMBER OF SEMI-SUPERVISED CLASSES FOR WATER

	Semi-Supervised - 50 Clusters
Figure	Number of Training Sets for Water
25	3
26	4
27	4
28	4

Three Semi-Supervised training sets were requested from a polygon drawn in the water; the fourth training set for water was found in a polygon placed over the agricultural areas. In figures 26, 27 and 28 the round pond near the center of the scene was inadvertently contained within the agricultural polygon.

The Semi-Supervised approach permits the analyst to designate the number of training sets that result in segmenting earth resources, such as water, whereas the FINDCLASS algorithm does not have this potential.

The disadvantage of FINDCLASS identifying 9 training sets for water is that there are fewer clusters remaining to describe cover types that may be more spectrally diverse. The reduced number of clusters remaining facilitates the potential for mismerging of spectral information. This point will be discussed in the following section, which discusses the overall results of the FINDCLASS and Semi-Supervised clustering analysis.

6.6 Accuracy Assessment of the Resulting Classifications

In general, the resulting classifications, for the Middleton area, of the FINDCLASS and Semi-Supervised algorithms are very similar. Table 8 lists the accuracy assessment of each image according to the test sites discussed in section 5.5. As was discussed in section 5.5, the disturbed vegetation, wetland, and many of the areas classified as forest, are not quantitatively evaluated because they lack absolute classification boundaries because of their spectral similarities. "Absolute" means that the boundaries in the thematic map are considered distinct, accurately classified and known. An appendix is attached containing the resulting contingency tables associated with the classification accuracies in table 8.

Table 8

ACCURACY ASSESSMENT OF CLASSIFIED IMAGES

Figure	Overall Accuracy (5 Classes)	5 Class Average Accuracy (5 Classes)
	FINDCLASS - 50 Clusters	
19	98.9 %	98.3 %
20	99.8 %	99.7 %
21	98.5 %	98.3 %
22	89.8 %	95.2 %
	Semi-Supervised - 50 Clusters	
25	97.2 %	95.8 %
26	98.2 %	97.4 %
27	97.4 %	96.7 %
28	98.2 %	98.2 %
	FINDCLASS - 27 Clusters	
23	83.0 %	76.4 %
24	77.7 %	57.5 %
	Semi-Supervised - 27 Clusters	
29	97.2 %	94.3 %
30	77.9 %	68.7 %

The accuracy assessment in table 8 cannot stand alone in describing the overall benefits or disadvantages of either clustering algorithm.

At 49 clusters, the overall accuracy for the Semi-Supervised approach is independent of the 4 rotations; whereas, in the FINDCLASS algorithm, the fourth rotation, figure 22, results in a spectral mismerging of the clusters *resulting* in a reduced classification accuracy. In figure 22, there is a misclassification of Hay II and the bare soils. This is not the case in the Semi-Supervised approach in figure 28.

The classification accuracy for the Semi-Supervised thematic map in figure 25 is reduced because some of the pixels in the hay I test sites were labelled unclassified, which is indicated in the contingency table in figure 50. Except for these unclassified pixels, the accuracy assessment for the same orientation clustered with FINDCLASS, in figure 19, is very similar, as can be seen in the contingency table in figure 51.

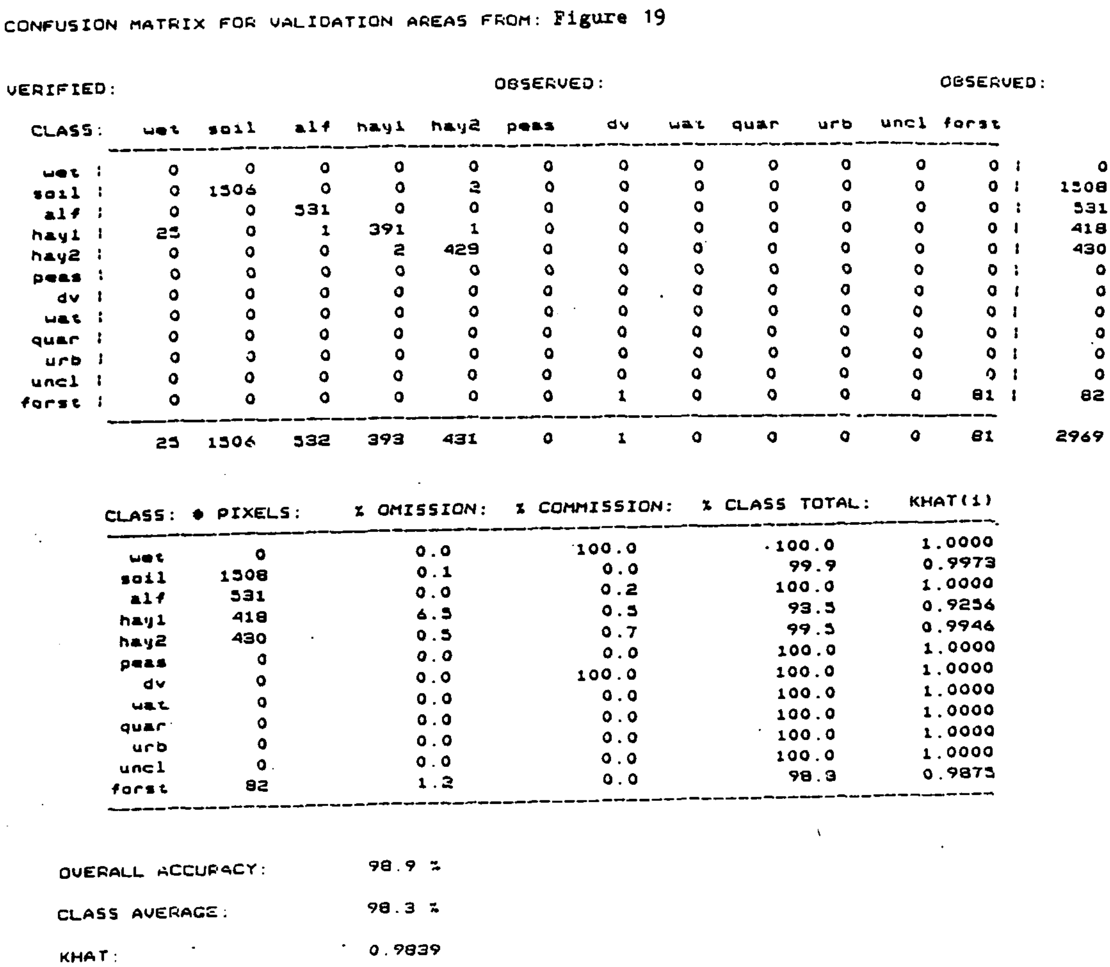
CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 25

VERIFIED: CLASS:	wet	soil	alf	hay1	hay2	peas	dv	wat	quar	urb	uncl	forst	
	OBSERVED:												OBSERVED:
wet	0	0	0	0	0	0	0	0	0	0	0	0	0
soil	0	1508	0	0	0	0	0	0	0	0	0	0	1508
alf	0	0	531	0	0	0	0	0	0	0	0	0	531
hay1	14	0	0	336	2	0	0	0	0	0	66	0	418
hay2	0	0	0	0	430	0	0	0	0	0	0	0	430
peas	0	0	0	0	0	0	0	0	0	0	0	0	0
dv	0	0	0	0	0	0	0	0	0	0	0	0	0
wat	0	0	0	0	0	0	0	0	0	0	0	0	0
quar	0	0	0	0	0	0	0	0	0	0	0	0	0
urb	0	0	0	0	0	0	0	0	0	0	0	0	0
uncl	0	0	0	0	0	0	0	0	0	0	0	0	0
forst	1	0	0	0	0	0	0	0	0	0	0	81	82
	15	1508	531	336	432	0	0	0	0	0	66	81	2969

CLASS:	# PIXELS:	% OMISSION:	% COMMISSION:	% CLASS TOTAL:	KHAT(1)
wet	0	0.0	100.0	100.0	1.0000
soil	1508	0.0	0.0	100.0	1.0000
alf	531	0.0	0.0	100.0	1.0000
hay1	418	19.6	0.0	80.4	0.7788
hay2	430	0.0	0.5	100.0	1.0000
peas	0	0.0	0.0	100.0	1.0000
dv	0	0.0	0.0	100.0	1.0000
wat	0	0.0	0.0	100.0	1.0000
quar	0	0.0	0.0	100.0	1.0000
urb	0	0.0	0.0	100.0	1.0000
uncl	0	0.0	100.0	100.0	1.0000
forst	82	1.2	0.0	98.8	0.9875

OVERALL ACCURACY:	97.2 %
CLASS AVERAGE:	95 8 %
KHAT:	0 9584

Figure 50. Accuracy assessment for the classified image in figure 25.



```
CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 19

VERIFIED:                              OBSERVED:                          OBSERVED:

  CLASS:   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
         ------------------------------------------------------------------------
    wet :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   soil :    0  1506     0     0     2     0     0     0     0     0     0     0 :   1508
    alf :    0     0   531     0     0     0     0     0     0     0     0     0 :    531
   hay1 :   25     0     1   391     1     0     0     0     0     0     0     0 :    418
   hay2 :    0     0     0     2   429     0     0     0     0     0     0     0 :    430
   peas :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
     dv :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
    wat :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   quar :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
    urb :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  forst :    0     0     0     0     0     0     1     0     0     0     0    81 :     82
         ------------------------------------------------------------------------
            25  1506   532   393   431     0     1     0     0     0     0    81     2969

         CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(i)
         ------------------------------------------------------------------------
           wet       0           0.0          100.0           100.0        1.0000
          soil    1508           0.1            0.0            99.9        0.9973
           alf     531           0.0            0.2           100.0        1.0000
          hay1     418           6.5            0.5            93.5        0.9256
          hay2     430           0.5            0.7            99.5        0.9946
          peas       0           0.0            0.0           100.0        1.0000
            dv       0           0.0          100.0           100.0        1.0000
           wat       0           0.0            0.0           100.0        1.0000
          quar       0           0.0            0.0           100.0        1.0000
           urb       0           0.0            0.0           100.0        1.0000
          uncl       0           0.0            0.0           100.0        1.0000
         forst      82           1.2            0.0            98.3        0.9875
         ------------------------------------------------------------------------

       OVERALL ACCURACY:          98.9 %

       CLASS AVERAGE:             98.3 %

       KHAT:                      0.9839
```

Figure 51. Accuracy assessment for the classified image in figure 19.

Unclassified areas in the resulting thematic maps are significant in that they indicate the resources that the clustering analysis were unable to describe. In the FINDCLASS thematic maps, quarries, urban features and edge pixels are often unclassified. The 3x3 window biases the algorithm against linear features and other resources that cannot completely fill the window. In a satellite image, such as SPOT, the resolution does not afford the roads in the Middleton area to be wider than 2 pixels. Roads of 3 or 4 pixel widths usually entail mixed pixels which often do not constitute homogeneous regions. The roads in the Middleton image predominantly comprised of mixed pixels, are classified as soils or hay, since training sets for the roads are unattainable. This was found to be the case in the supervised classification process until an adequate road training set was described; in the practicum analysis, this training set was not identified in the Middleton area.

The FINDCLASS thematic maps classified quarries as both unclassified and urban. On the other hand, in the Semi-Supervised analysis, the quarries were completely unclassified. Semi-Supervised training was not directed to cluster training sets for quarries, since quarries were not in any of the polygons. Also pea fields were not within the polygon areas for the classification in figure 25; subsequently, peas were accurately unclassified. In figures 26, 27, and 28, pea fields were in one of the training areas and were accurately classified. This indicated that the Semi-Supervised approach prevents the mismerging of spectral information describing the peas and quarries with the other clusters. Mismerging is prevented by the transformed divergence calculation which ensures that resources not described by the initial seeds remain unclassified. There are three pea fields in the Middleton scene.

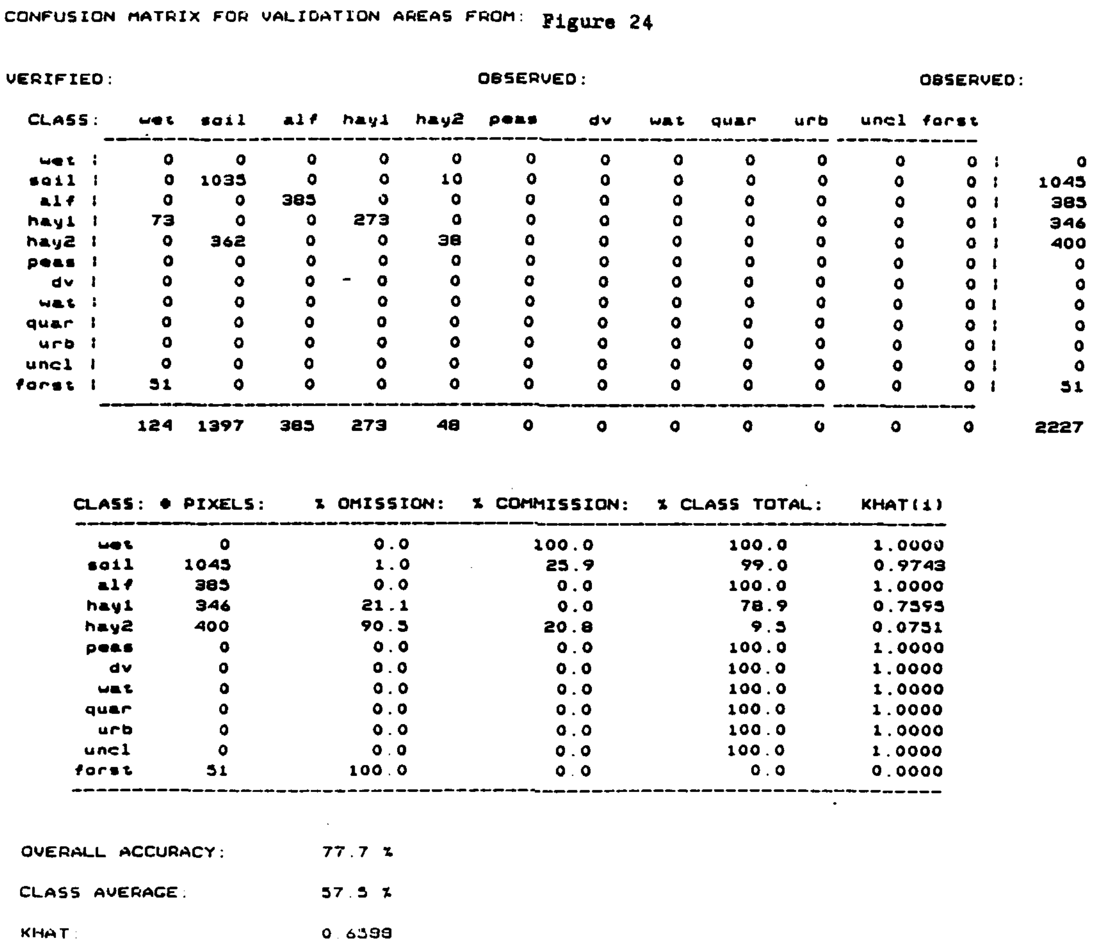
Identifying fewer than 49 clusters resulted in misclassification by both clustering algorithms for different reasons, but the Semi-Supervised algorithm performance is far more acceptable. In all cases, the misclassification was less in the Semi-Supervised thematic maps.

The Semi-Supervised approach finds 27 initial seed clusters, for example, in a directed manner, and then implements these clusters to collect spectral information from throughout the image. The FINDCLASS algorithm identifies 49 clusters in an image, and then merges down to the user-defined number of clusters requested, 27 for example. These clusters are merged according to a minimum statistic distance rule, possibly mismerging clusters that are spectrally different. For mismerging to occur in the Semi-Supervised approach, it will take place during the acquisition of the initial clusters in each polygon. The seeds are identified with a FINDCLASS operation in each polygon. Therefore, to ensure that mismerging does not occur, in the Semi-Supervised approach, the user must designate a number of spectral classes to be found that is greater than the spectral classes expected to result from that training area. This should prevent the mismerging of spectrally different clusters in the polygons. A transformed divergence calculation, as discussed in section 6.2, prevents the mismerging of spectral information after the seeds are created.

In figure 23, only 27 clusters were requested, and FINDCLASS mismerged clusters, biasing the classification toward the bare soils and hay I over hay II. Visually assessing the classification, it is also noted that the disturbed vegetation class has been mostly classified as wetland. These results are much different for the thematic map in figure 29, that was clustered with a Semi-Supervised algorithm at 27 clusters (see Table 8). In figure 29, there is slight mismerging among the bare soil and hay II because of the reduced number of clusters requested in the agricultural training area. Visually, disturbed vegetation and wetland classes still remain with very little misrepresentation, unlike figure 23. Peas still remain unclassified in figure 29. Overall, the mismerging is reduced considerably by the Semi-Supervised technique, relatively speaking.

The Middleton image was rotated 180 degrees and 27 clusters were identified by both algorithms. The resulting classifications can be seen in figures 24 and 30, FINDCLASS and Semi-supervised respectively. According to the accuracy assessment in table 8, both performed similarly. However, the FINDCLASS clustering process merged all the clusters representing wetland, disturbed vegetation, forest areas, and some clusters describing alfalfa and hay I into one training set which is represented as wetland in figure 24. Alfalfa fields not designated as test sites for accuracy assessment were included in this class. Also, hay I is no longer classified in the northern section of the Pheasant Branch Creek marsh. All of the forested areas are misclassified by this class. It must be emphasized that one spectral training set described all of these areas that were previously described by several classes. This can be noted in the contingency table for this image in figure 52. Pixels in the test sites for hay I and forest were misclassified as wetland. Comparing these results, for figure 24, to the same image orientation in figure 21 at 49 FINDCLASS clusters, a significant mismerging of training sets occurred when reducing the output from 50 clusters to 27. Also, as indicated in the accuracy assessment in figure 52, mismerging occurred among the bare soil and hay II classes.

The Semi-Supervised output for the same rotation can be seen in figure 30. Relatively speaking, the mismerging between bare soil and hay II is slightly reduced but still occurs. This is due to the mismerging in the agricultural polygon area with the unsupervised clustering algorithm, because of the reduced number of clusters requested and a



```
CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 24

VERIFIED:                             OBSERVED:                          OBSERVED:

  CLASS:   wet  soil   alf  hay1  hay2  peas   dv   wat  quar   urb  uncl forst
         ---------------------------------------------------------------------
    wet :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   soil :    0  1035     0     0    10     0     0     0     0     0     0     0 :  1045
    alf :    0     0   385     0     0     0     0     0     0     0     0     0 :   385
   hay1 :   73     0     0   273     0     0     0     0     0     0     0     0 :   346
   hay2 :    0   362     0     0    38     0     0     0     0     0     0     0 :   400
   peas :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
     dv :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    wat :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   quar :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    urb :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  forst :   51     0     0     0     0     0     0     0     0     0     0     0 :    51
         ---------------------------------------------------------------------
           124  1397   385   273    48     0     0     0     0     0     0     0    2227


     CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(i)
     -----------------------------------------------------------------------
       wet        0           0.0          100.0          100.0       1.0000
      soil     1045           1.0           25.9           99.0       0.9743
       alf      385           0.0            0.0          100.0       1.0000
      hay1      346          21.1            0.0           78.9       0.7595
      hay2      400          90.5           20.8            9.5       0.0751
      peas        0           0.0            0.0          100.0       1.0000
        dv        0           0.0            0.0          100.0       1.0000
       wat        0           0.0            0.0          100.0       1.0000
      quar        0           0.0            0.0          100.0       1.0000
       urb        0           0.0            0.0          100.0       1.0000
      uncl        0           0.0            0.0          100.0       1.0000
     forst       51         100.0            0.0            0.0       0.0000
     -----------------------------------------------------------------------

  OVERALL ACCURACY:        77.7 %

  CLASS AVERAGE:           57.5 %

  KHAT:                    0.6598
```

Figure 52. Accuracy assessment for the classified image in figure 24.

different order of acquisition of the polygon area, since the image is now upside down. This coincides with the previous conclusion that the results of the FINDCLASS style of unsupervised clustering is dependent upon the orientation of the image. An important point not evidenced by the accuracy assessment, is that the Semi-Supervised analysis reduced the mismerging of wetland, forest, disturbed vegetation, alfalfa, and hay I that is present in the FINDCLASS output in figure 24. Spectral diversities of these classes were still described by the Semi-Supervised approach. This is somewhat indicated by the contingency table, in figure 53, for the hay I and forest test sites, as compared to the contingency table in figure 52 discussed above.

In summation, in all cases, the mismerging found in the FINDCLASS algorithm has been reduced in the Semi-Supervised technique. There is consistently a larger number of pixels unclassified in the Semi-Supervised thematic maps than in the FINDCLASS, an indication that the Semi-Supervised approach is directed in nature. And often, the unclassified areas represent land cover types that were not trained on with the directed clustering

```
CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 30

VERIFIED:                            OBSERVED:                              OBSERVED.

 CLASS.   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
        ----------------------------------------------------------------------
   wet !    0     0     0     0     0     0     0     0     0     0     0     0 !     0
  soil !    0  1045     0     0     0     0     0     0     0     0     0     0 !  1045
   alf !    0     0   385     0     0     0     0     0     0     0     0     0 !   385
  hay1 !   49     0     0   229    29    33     4     0     0     0     2     0 !   346
  hay2 !    0   358     0     0    42     0     0     0     0     0     0     0 !   400
  peas !    0     0     0     0     0     0     0     0     0     0     0     0 !     0
    dv !    0     0     0     0     0     0     0     0     0     0     0     0 !     0
   wat !    0     0     0     0     0     0     0     0     0     0     0     0 !     0
  quar !    0     0     0     0     0     0     0     0     0     0     0     0 !     0
   urb !    0     0     0     0     0     0     0     0     0     0     0     0 !     0
  uncl !    0     0     0     0     0     0     0     0     0     0     0     0 !     0
 forst !    2     0    15     0     0     0     0     0     0     0     0    34 !    51
        ----------------------------------------------------------------------
           51  1403   400   229    71    33     4     0     0     0     2    34    2227


      CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(1)
      --------------------------------------------------------------------
        wet        0            0.0          100.0           100.0    1.0000
       soil     1045            0.0           25.5           100.0    1.0000
        alf      385            0.0            3.8           100.0    1.0000
       hay1      346           33.3            0.0            66.2    0.6231
       hay2      400           89.5           40.8            10.5    0.0755
       peas        0            0.0          100.0           100.0    1.0000
         dv        0            0.0          100.0           100.0    1.0000
        wat        0            0.0            0.0           100.0    1.0000
       quar        0            0.0            0.0           100.0    1.0000
        urb        0            0.0            0.0           100.0    1.0000
       uncl        0            0.0          100.0           100.0    1.0000
      forst       51           33.3            0.0            66.7    0.6615
      --------------------------------------------------------------------

  OVERALL ACCURACY:        77.9 %

  CLASS AVERAGE.           68.7 %

  KHAT:                    0.6603
```

Figure 53. Accuracy assessment for the classified image in figure 30.

analysis.

6.7 Assessment of the Semi-Supervised Two Stage Reclassification

The results of the reclassification of the Chesapeake Bay classified image (Figure 32) with ancillary data, such as texture, vegetation index ratio and polygon masks will now be discussed. The reclassified images can be viewed in figures 33 through 37.

6.7.1 Texture Reclassification

The discriminant reclassification of selected classes, in the Chesapeake Bay image, according to textural information was very successful, and many of the details discussed about texture and classifications can be related to the results in figure 33. But first, it is important that the reader visualize the outcome of classifying an image with spectral and textural information indiscriminately (Figure 54). In figure 54, the texture image was implemented as a fourth band of digital information for this classification. One training set classified the area

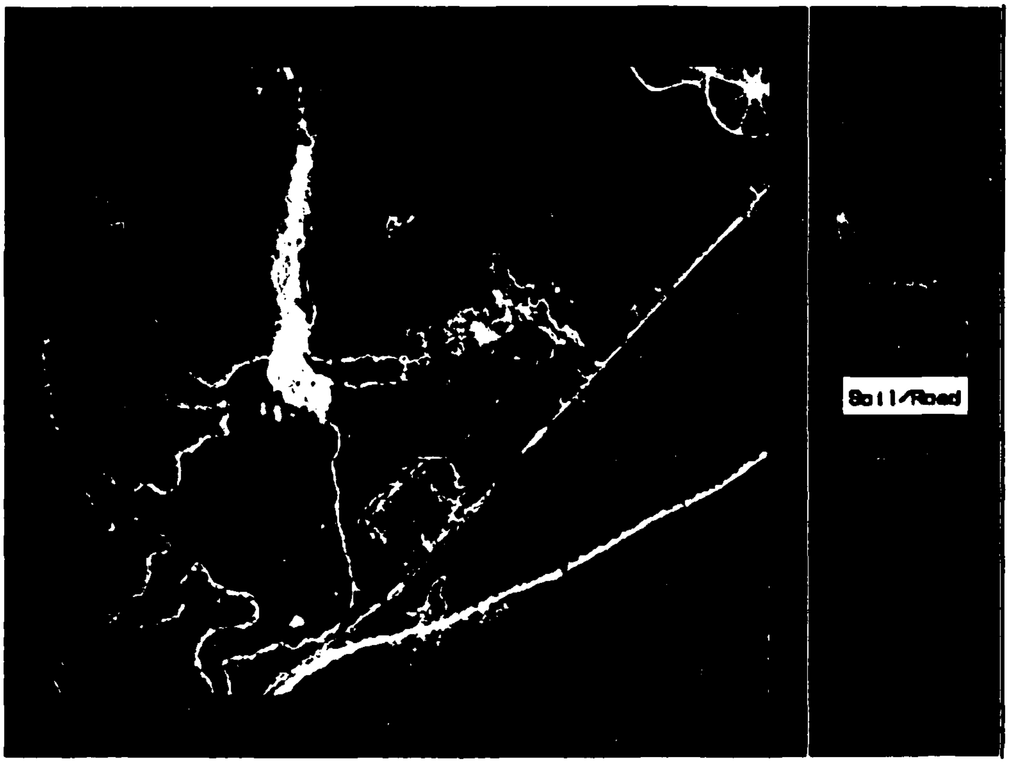


Figure 54. Classification of the Chesapeake Bay according to spectral and textural information.

displayed in black in figure 55. This supervised training set was taken in the less vivid forest canopy and is spectrally similar to the grass. Areas denoted in black are grassland which possess these spectral characteristics and high textural values. These high texture areas are the result of texture values smearing over the boundary of the forest canopy and are also enhancements of edges on the landfill and along Watson Creek. In figure 54, the soil/road class is also misclassified because of the addition of texture. A road training set placed on a section of the road that has a bluish tint and a high textural component, classified the dark wet grass near the edge of the inlet where texture is high because of the edge enhancement of the water/grass boundary. Because of texture, the specular reflection and much more of the surrounding water are classified into bare soil. Also, blobs of the water are classified as bare soil because of texture. It would be inviting to rename this training set to water, but, unfortunately, the roads, the grass around Watson creek, and the beach would all be misclassified. The soil/road category is unaffected by texture when it is implemented in a discriminant manner. Throughout the following discussion, this image will be referred to regarding the results.

The Second Stage reclassification according to texture in figure 33 had very positive results. The actual description of the forest canopy was to be maintained while the misclassified pixels surrounding the forest in the grassland (Figure 32) would be reclassified. This has been accomplished barring some small details. The pixels that were reclassified can be viewed in figure 42.

Forest pixels still remain incorrectly classified in the grassland, where the textural values were similar to the forest canopy; texture values are high along the edges of land cover features. Forest pixels still remain along the upper part of the Watson Creek; but these remaining pixels are not as unacceptable, as the entire misclassification of the border along the Watson as in figure 54. In figure 33, Forest pixels remaining on the left side of the image in the grassland should have been reclassified since there appears to be the smooth texture present. There are two reasons for this: (1) the texture in these areas matched the textural statistics of the forest canopy, or (2) the texture in these areas matched neither the forest textural

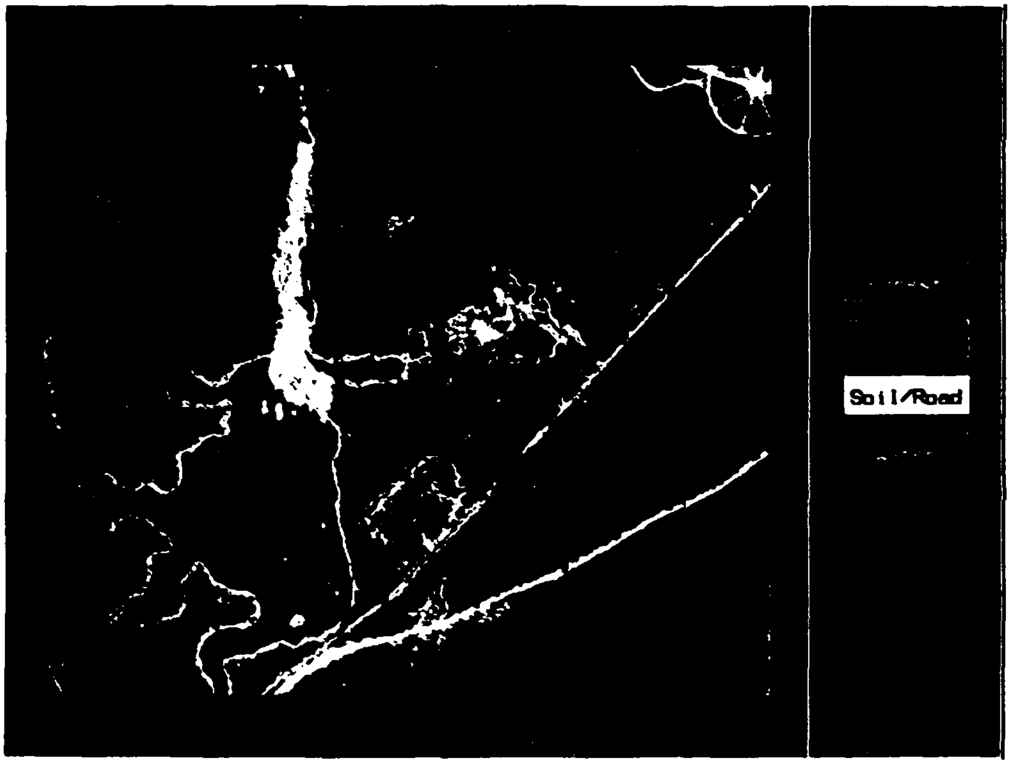


Figure 55. Spectral-texture classification of the Chesapeake study area with a selected training set enhanced in black.

statistics nor the texture of the grassland, and therefore were not reclassified. A small stand of trees to the left of Watson Creek still remained classified correctly. The small stand of trees at the mouth of Watson creek above the Chesapeake still remained classified as forest, but it appears in the texture image that texture representing a forest canopy is not present. This stand probably survived the reclassification because the texture did not match either forest or grassland as was mentioned above.

The detail of the main forest canopy that prevails in the discriminant reclassification contrasts the appearance of blobs describing the boundary of the canopy in the spectral textural classification in figure 54. In the reclassified forest, in figure 33, there are open spaces in the canopy that identify the gaps where shrubs and brush are present in image (Figure 14). These inlets are not present in figure 54.

The reclassification according to texture changed 6.8 % of the image to grassland (Figure 42).

6.7.2 Vegetation Index Ratio Reclassification

There were two executions of the Second Stage reclassification technique to reclassify misclassified forest pixels, in the Chesapeake Bay region, with a vegetation index ratio; first with an unsmoothed ratio (Figure 34) and the second time with a smoothed version (Figure 35).

The smoothed vegetation index ratio in figure 40 was more continuous and complete in describing the shape and area of the forest canopy than the unsmoothed version in figure 39. After the reclassification with the unsmoothed ratio, the forest canopy around the two landfills was very scanty and not completely described. The pixels that were reclassified can be viewed in figure 43; note: many pixels in the forest canopy were reclassified into grassland. The small stand of forest near the mouth of Watson Creek in the lower section of the image is almost non-descript. The smoothed vegetation index ratio prevented some of the main forest canopy around the landfills from being reclassified (Figure 44); but again the values in the ratio describing the small stand of trees at the mouth of the Watson were not present. The small stand in the resulting reclassification is hardly described.

The smoothed vegetation index ratio reclassified 13.2% of the image as compared to texture, which reclassified 6.8% of the image. This is due to a few reasons. The vegetation index ratio reclassified pixels along the edge of the roads and the upper portion of Watson Creek that prevailed when the texture Second Stage was implemented. Also, texture tends to smear the boundary of the forest, whereas the ratio is more descriptive of the canopy's perimeter, resulting in more reclassified pixels along the edge of the canopy in the grassland. More of the forested pixels in the grassland on the left side of the image and the top section of the image are reclassified correctly according to the ratio.

Four percentage fewer pixels were reclassified by the smoothed vegetation index ratio than with the unsmoothed version, primarily because the unsmoothed version did not prevent many forest pixels in the main canopy and surrounding the landfills from being reclassified. Figure 43 shows the pixels reclassified by the unsmoothed vegetation index ratio.

6.7.3 Polygon Mask Reclassification

The Second Stage reclassification implementing a polygon mask to reclassify the specular reflection, in the Chesapeake Bay image, was very successful (Figure 36). All but a few pixels (not visible on the color reproduction) representing the specular reflection were reclassified into the water class. These appear to be two classes that were not designated for reclassification. Zero point eight percentage (0.8 %) of the image was reclassified.

In figure 37, a two part reclassification is demonstrated. A vegetation index ratio reclassifies the misclassified grassland, and the polygon mask reclassified the specular reflection.

6.8 Closing Discussion

Secticn 2.12 mentions that the time required to execute a program is one of the characteristics that should be noted by the analyst in selecting the appropriate clustering algorithm or classification.

The Semi-Supervised clustering algorithm processed 50 clusters in a 512 column by 480 row image in 30 minutes on an 8 MHZ (megahertz) IBM PC-AT microcomputer and 43 minutes on a 4 MHZ machine. Second Stage statistics were simultaneously acquired.

The FINDCLASS algorithm completely processed the same image in 22 minutes on a PDP 1145 minicomputer. Second Stage statistics could not be acquired with this program. It should be noted that a minicomputer is much faster than a portable microprocessor, and the FINDCLASS algorithm is not an acquisition-directed approach like the Semi-Supervised. The Semi-Supervised clustering algorithm is essentially a supervised approach with an added unsupervised twist. Therefore, it may be more appropriate to compare the Semi-Supervised technique to the supervised training process.

Three or four training areas can be identified in an image in less than five minutes, and the Semi-Supervised clustering analysis can grind out spectral training statistics from these areas in 30 minutes while the analyst is working on another project. It is questionable that an image analyst can draw 50 polygons in selected homogeneous areas, to represent the wide range of spectral diversities in the land cover types of an image, in less than 35 minutes. And these polygons must then be input into two programs that collect the pixel values delineated within the polygons, and compute training sets statistics; this could take another 5 to 10 minutes for 50 training sets. And in the end, the training set statistics from both the Semi-Supervised clustering algorithm and supervised approach, are placed in a statistical classification program to evaluate how well the image was classified with these statistics. And, still, the retraining process may have to be done all over again. In the end, the results from both processes are similar, therefore it should be noted, that the Semi-Supervised approach allows a person time to do other things.

The Second Stage reclassification is extremely fast. Three forest classes were reclassified, according to Second Stage textural statistics in a 467 column by 400 row image, in 5 minutes. The reclassification according to a threshold range, takes less than 5 minutes. And the reclassification by a threshold value, takes about two-and-one-half minutes.

Concluding remarks are discussed in the next chapter followed by an appendix of the program source codes.

Chapter VII

Conclusions

7.1 Semi-Supervised Clustering

The Semi-Supervised clustering algorithm performed successfully, as anticipated. Overall, it performed better than FINDCLASS in all aspects of the thesis evaluation. FINDCLASS, however, performed better than anticipated. There was no evidence of clustering bias in the FINDCLASS algorithm toward the dominant land cover class in the image. In the thesis hypothesis it was stated that the Semi-Supervised clustering algorithm would reduce some of the biases inherent in the FINDSET algorithm. From these results, also discussed in section 6.5, it cannot be concluded that the Semi-Supervised process reduces the clustering bias. It was found that the Semi-Supervised technique was able to control the number of clusters that would describe the dominant land cover class of the image, whereas the FINDCLASS algorithm found an unpredictable number of clusters for such a category, and the training sets were often redundant. The spectral mismerging identified in the FINDCLASS algorithm was reduced in the Semi-Supervised clustering process; this is evidenced when fewer than 50 clusters were requested. The

Semi-Supervised training set statistics were found to be more statistically valid regarding the number of pixels in each class, on whole, than those described by FINDCLASS. This is because the training sets resulting from the Semi-Supervised clustering involve spectral seeds that were permitted to merge with spectral data from throughout the image. The addition of the transformed divergence calculation prevents mismerging of spectral clusters in the Semi-Supervised approach, permitting the analyst to direct the clustering of spectral information of certain land cover classes of interest. In summation, the Semi-Supervised approach offers the analyst a priori knowledge as to the utility of the resulting clusters, because it is in fact guided by the user. The Semi-Supervised approach is comparable to the supervised training process but requires less input from the analyst, allowing the user to economize his time.

7.2 Second Stage Reclassification

The hypothesis also stated that the application of ancillary data, as a Second Stage, implemented in a discriminant manner would improve the classification accuracy. It has been found that the discriminant reclassification of spectrally classified images with ancillary data did improve the classification accuracy. This Second Stage post-classification reclassification is able to recategorize only the classes the analyst designates, permitting the ancillary data to be applied to the applicable land cover types. This is a notable benefit since ancillary data often selectively contributes to the classification of digital images.

Misclassified grassland areas were reclassified with textural statistics and a vegetation index ratio. The resulting classification with the vegetation index ratio was slightly more descriptive of the actual shape and area of the forest canopy and reclassified more misclassified pixels in the grassland than the textural reclassification. Reclassifications can be made, based on Second Stage statistics, such as for texture, and can also be based on threshold ranges in a Second Stage image file.

The specular reflection in the water of a digitized aerial photograph was successfully reclassified as water using a polygon mask. The classes detailing the specular reflection were reclassified if they fell inside the designated region of the Second Stage polygon mask. This type of threshold value reclassification is based on a Second Stage image file that is threshold amenable.

7.3 Summary

The Semi-Supervised Two Stage Classification technique was found to be a viable method for classifying remotely sensed digital imagery. A Semi-Supervised clustering algorithm analyzes multispectral data to gather spectral training set statistics from directed regions of the image under the guidance of the analyst. This clustering algorithm may also simultaneously acquire statistical information from a Second Stage of information. Resulting spectral statistics are then implemented into a statistical classifier to segment the multispectral image.

The Second Stage involves a reclassification of a spectrally classified image based on ancillary data. There are three different styles of reclassification: (1) a statistical approach, (2) a threshold range approach and (3) a threshold value approach.

The Semi-Supervised Two Stage Classification Technique has proven to be a viable hybrid classification process for the clustering, classification and reclassification of remotely sensed data.

SELECTED BIBLIOGRAPHY

Ahearn. S. C., 1986. Toward an Expert System for the Analysis of High Resolution Satellite Imagery. Ph.D thesis. University of Wisconsin-Madison.

Ahearn, S. C. and T. M. Lillesand, 1986. A Proposed Hotelling T^2 Based Unsupervised Procedure as Input to a Bayesian Classifier. Annual Convention of American Society of Photogrammetry and Remote Sensing, March, pp. 350 - 359.

Armstrong, A. C., 1977. The Relative Performance of Some Unsupervised Clustering Techniques for the Per-Field Classification of Landsat Data. British Interplanetary Society Journal, vol. 30, no. 5, May, pp. 168 - 171.

Blohm, J., W. Fraczek, H. Hardjakusumah, R. Maki, S. Maselli, D. Mossman, P. Northcutt, H. Ping and D. Toomey, 1987. *A Preliminary Assessment* of SPOT-1 Satellite Data. University of Wisconsin-Madison 105 p.

Chandrasekhar, S., 1983. Unsupervised Classification of Multispectral Data Using Divisive Algorithm. 1983 Proceedings of the International Conference on Systems, Man and Cybernetics, vol. 1, pp. 376 - 377.

Colwell, R. N., 1984. Analysis of the Quality of Image Data Required by the Landsat-4 Thematic Mapper and Multispectral Scanner. National Aeronautics and Space Administration, Washington, D.C. December, 151 p.

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

Fruend, J. E., 1971. Mathematical Statistics. Prentice-Hall, Englewood Cliffs, N.J. 561 p.

Goldberg, M., D. Schlaps, M. Alvo and G. Karam, 1982. Monitoring and Change Detection with Landsat Imagery. Proceedings of the 6th International Conference on Pattern Recognition, vol. 1, pp. 523 - 526.

Goldberg, M. and S. Shlien, 1977. A Four-Dimensional Histogram Approach to the Clustering of Landsat Data. 4th Machine Processing of Remotely Sensed Data Symposium, June 21 -23, pp. 250 - 257.

Gowda, K. C., 1984. A Feature Reduction and Unsupervised Classification Algorithm for Multispectral Data. Pattern Recognition, vol. 17, no. 6, pp. 667 -676.

Haralick, R. M., K. Schanmugam and I. Dinstein, 1973. Texture Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, vol smc-3, no. 6, pp. 610 - 621.

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

Hsu, S., 1977. A Texture-Tone Analysis for Automated Land-Use Mapping with Panchromatic Images. American Society of Photogrammetry Proceedings, vol 43, pp. 203 - 215.

Hsu, S., 1978. Texture-Tone Analysis for Automated Land-Use Mapping. Photogrammetric Engineering and Remote Sensing, vol. 44, no. 11, pp. 1393 - 1404.

Ince, F., 1981. The Application of the Coalescence Clustering Algorithm to Remotely Sensed Multispectral Data. Pattern Recognition, vol. 14, nos. 1 - 6, pp. 121 - 130.

Irons J. R. and G. W. Peterson, 1981. Texture Transforms of Remotely Sensing Data. Remote Sensing of Environment, vol. 11, pp. 359 - 370.

Jensen, J. R., 1979. Spectral and Textural Features to Classify Elusive Land Cover at the Urban Fringe. Professional Geographer, vol. 31, no. 4, pp. 400 - 409.

Jensen, J. R. and D. L. Toll, 1982. Detecting Residential Land-Use Development at the Urban Fringe. Photogrammetric Engineering and Remote Sensing, vol. 48 , pp. 629 - 643.

Kanlensky, Z. D., W. C. Moore, G. A. Campbell, D. A. Wilson and A. J. Scott, 1981. Summary Forest Resource Data from Landsat Images. Canadian Forestry Service Environment Canada (Information Report PI-X-5), Chalk River, Ontario. 24 p.

Leboucher, G., B. E. Lowitz, E. Matra, 1976. What Can a Histogram Really Tell the Classifier. 3rd International Joint Conference on Pattern Recognition, November 8 - 11, Coronado, California, pp. 689 - 695.

Lillesand T. M. and R. W. Kiefer, 1987. Remote Sensing and Image Interpretation, 2nd Edition. Wiley, New York. 721 p.

Maktav, D., 1985. The Study of the Natural Geographic Differences in the Coastal Areas of Water Covered Parts of Marmara Region in Turkey with the Help of Landsat-4 MSS Data Using an Unsupervised Classification Algorithm with Euclidean Distance. 11th Annual Machine Processing of Remotely Sensed Data Symposium, June 25 - 27, pp. 122 - 127.

Moreira, M. A., S. C. Chen and A. M. de Lima, 1986. Evaluation of Spatial Filtering on the Accuracy of Wheat Area Estimate. International Symposium on Remote Sensing of Environment Papers, March, 28 p.

NASA/ERL, 1981. ELAS Earth Resources Laboratory Applications Software, NSTL Rept. 183, NSTL. ms.

Nelson, C. A., D. E. Meisner and B. Smekofski, 1981. Techniques to Update a Land Management Information System with Landsat. 7th Annual Machine Processing of Remotely Sensed Data Symposium, June 23 - 26, pp. 505 - 517.

Pearson, R. W., 1977. SEARCH - An Efficient, Automatic Training Sample Selection Algorithm. 4th Annual Machine Processing of Remotely Sensed Data Symposium, June 21 - 23, pp. 309.

Pratt, W. K., 1978. Digital Image Processing. California. 750 p.

Richards, J. A., 1986. Remote Sensing Digital Image Analysis, An Introduction. Springer-Verlag, Germany. 281 p.

Schowengerdt, R. A. 1983. Techniques for Image Processing and Classification in Remote Sensing. Academic Press, Orlando, Florida. 249 p.

Shih, E. H. H. and R. A. Schowengerdt, 1983. Classification of Arid Geomorphic Surfaces Using Landsat Spectral and Textural Features. Photogrammetric Engineering and Remote Sensing, vol. 49, pp. 337 - 347.

Story, M. H., J. B. Campbell and G. Best, 1984. An Evaluation of the Accuracies of Five Algorithms for Machine Classification of Remotely Sensed Data. The Ninth William T. Pecora Memorial Remote Sensing Symposium, Sioux Falls, South Dakota, October 2 - 4, pp. 399 - 405.

Swain, P. H. and S. M. Davis, 1978. Remote Sensing: A Quantitative Approach. McGraw-Hill, New York. 396 p.

Toomey D. A. and F. L. Scarpace, 1987. A Proposed Semi-Supervised Two Stage Classification Technique. Annual Convention of the American Society for Photogrammetry and Remote Sensing. vol 6, March, pp. 1 - 6.

Tucker, C. J., 1979. Red Photograph Linear Combinations for Monitoring Vegetation. Remote Sensing of the Environment. vol. 8, pp. 127 - 150.

Vasseur, C. P. A. and J. G. Postaire, 1980. A Convexity Testing Method for Cluster Analysis. IEEE Transactions on Systems, Man, and Cybernetics, vol. smc-10, no. 3, March, pp. 145 - 149.

Wharton, S. W., 1983. A Generalized Histogram Clustering Scheme for Multidimensional Image Data. Pattern Recognition, vol. 16, no. 2, pp. 193 - 199.

Weismiller, R. A., S. J. Kristof, D. K. Scholz, P. E. Anuta, and S. M. Momin, 1977. Evaluation of Change Detection Techniques for Monitoring Coastal Zone Environments. NASA Earth Resources Survey Program, Washington, D.C., June, 24 p.

Werth, L. F., 1981. An Evaluation of ISOCLS and Classy Clustering Algorithms for Forest Classification in Northern Idaho. National Aeronautics and Space Administration, Washington, D.C., September, 21 p.

Weszka, J. S. and A. Rosenfeld, 1975. A Comparative Study of Texture Measures for Terrain Classification. IEEE Proceedings of the Conference on Computer Graphics, Pattern Recognition and Data Structure, pp. 62 - 64.

Weszka, J. S., C. R. Dyer and A. Rosenfeld, 1976. A Comparative Study of Texture Measures for Terrain Classification. IEEE Transactions of Systems, Man, and Cybernetics, vol. 6, no. 4, pp. 269 - 285.

Wessman, C. A., 1984. The Utility of Color Infrared Photography and Synthetic Aperture Radar for Vegetation Type Discrimination in the Tropics. Master of Science Thesis, University of Wisconsin-Madison, 170 p.

Wickware G. M. and P. J. Howarth, 1981. Change Detection in the Peace-Athabasca Delta Using Digital Landsat Data. Remote Sensing of Environment, vol. 11, pp 9 - 25.

Wiersma, D. J. and D. Landgrebe, 1976. The Use of Spatial Characteristics for the Improvement of Multispectral Classification of Remotely Sensed Data. Proceedings of the 1976 Symposium on Machine Processing of Remotely Sensed Data, West Lafayette, Indiana, June 29 - July 1, pp. 2A-18 - 2A-26.

APPENDICES

CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 19

VERIFIED:	OBSERVED:												OBSERVED:
CLASS:	wet	soil	alf	hay1	hay2	peas	dv	wat	quar	urb	uncl	forst	
wet	0	0	0	0	0	0	0	0	0	0	0	0	0
soil	0	1506	0	0	2	0	0	0	0	0	0	0	1508
alf	0	0	531	0	0	0	0	0	0	0	0	0	531
hay1	25	0	1	391	1	0	0	0	0	0	0	0	418
hay2	0	0	0	2	429	0	0	0	0	0	0	0	430
peas	0	0	0	0	0	0	0	0	0	0	0	0	0
dv	0	0	0	0	0	0	0	0	0	0	0	0	0
wat	0	0	0	0	0	0	0	0	0	0	0	0	0
quar	0	0	0	0	0	0	0	0	0	0	0	0	0
urb	0	0	0	0	0	0	0	0	0	0	0	0	0
uncl	0	0	0	0	0	0	0	0	0	0	0	0	0
forst	0	0	0	0	0	0	1	0	0	0	0	81	82
	25	1506	532	393	431	0	1	0	0	0	0	81	2969

CLASS:	# PIXELS:	% OMISSION:	% COMMISSION:	% CLASS TOTAL:	KHAT(1)
wet	0	0.0	100.0	100.0	1.0000
soil	1508	0.1	0.0	99.9	0.9973
alf	531	0.0	0.2	100.0	1.0000
hay1	418	6.5	0.5	93.5	0.9256
hay2	430	0.5	0.7	99.5	0.9946
peas	0	0.0	0.0	100.0	1.0000
dv	0	0.0	100.0	100.0	1.0000
wat	0	0.0	0.0	100.0	1.0000
quar	0	0.0	0.0	100.0	1.0000
urb	0	0.0	0.0	100.0	1.0000
uncl	0	0.0	0.0	100.0	1.0000
forst	82	1.2	0.0	98.3	0.9875

OVERALL ACCURACY:	98.9 %
CLASS AVERAGE.	98.3 %
KHAT:	0.9839

CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 20

```
VERIFIED:                          OBSERVED:                            OBSERVED:

 CLASS:   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
        -----------------------------------------------------------------------
   wet :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  soil :    0  1185     0     0     0     0     0     0     0     0     0     0 :   1185
   alf :    0     0   370     0     0     0     0     0     0     0     0     0 :    370
  hay1 :    0     0     0   363     5     0     0     0     0     0     0     0 :    368
  hay2 :    0     0     0     0   343     0     0     0     0     0     0     0 :    343
  peas :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
    dv :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   wat :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  quar :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   urb :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
 forst :    0     0     0     0     0     0     0     0     0     0     0    53 :     53
        -----------------------------------------------------------------------
            0  1185   370   363   348     0     0     0     0     0     0    53     2319


 CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(1)
 ------------------------------------------------------------------------
    wet         0          0.0           0.0           100.0     1.0000
   soil      1185          0.0           0.0           100.0     1.0000
    alf       370          0.0           0.0           100.0     1.0000
   hay1       368          1.4           0.0            98.6     0.9839
   hay2       343          0.0           1.4           100.0     1.0000
   peas         0          0.0           0.0           100.0     1.0000
     dv         0          0.0           0.0           100.0     1.0000
    wat         0          0.0           0.0           100.0     1.0000
   quar         0          0.0           0.0           100.0     1.0000
    urb         0          0.0           0.0           100.0     1.0000
   uncl         0          0.0           0.0           100.0     1.0000
  forst        53          0.0           0.0           100.0     1.0000
 ------------------------------------------------------------------------

OVERALL ACCURACY:         99.8 %

CLASS AVERAGE.            99.7 %

KHAT.                     0.9963
```

CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 21

```
VERIFIED:                           OBSERVED:                                    OBSERVED:

 CLASS:   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
         ---------------------------------------------------------------------
   wet :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  soil :    0  1041     0     0     4     0     0     0     0     0     0     0 :   1045
   alf :    0     0   382     0     0     0     3     0     0     0     0     0 :    385
  hay1 :   20     0     0   326     0     0     0     0     0     0     0     0 :    346
  hay2 :    0     6     0     0   394     0     0     0     0     0     0     0 :    400
  peas :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
    dv :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   wat :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  quar :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   urb :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
 forst :    0     0     0     0     0     0     0     0     0     0     0    51 :     51
         ---------------------------------------------------------------------
           20  1047   382   326   398     0     3     0     0     0     0    51       2227


 CLASS: # PIXELS:     % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(i)
 ----------------------------------------------------------------------
   wet       0            0.0         100.0           100.0        1.0000
  soil    1045            0.4           0.6            99.6        0.9928
   alf     385            0.8           0.0            99.2        0.9906
  hay1     346            5.8           0.0            94.2        0.9323
  hay2     400            1.5           1.0            98.5        0.9817
  peas       0            0.0           0.0           100.0        1.0000
    dv       0            0.0         100.0           100.0        1.0000
   wat       0            0.0           0.0           100.0        1.0000
  quar       0            0.0           0.0           100.0        1.0000
   urb       0            0.0           0.0           100.0        1.0000
  uncl       0            0.0           0.0           100.0        1.0000
 forst      51            0.0           0.0           100.0        1.0000
 ----------------------------------------------------------------------

 OVERALL ACCURACY:        98.5 %

 CLASS AVERAGE:           98.3 %

 KHAT:                    0.9787
```

```
CONFUSION MATRIX FOR VALIDATION AREAS FROM:  Figure 22

VERIFIED:                                   OBSERVED:                                  OBSERVED:

  CLASS:   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
          ------------------------------------------------------------------------
    wet :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   soil :    0   714     0     0   211     0     0     0     0     0     0     0 :   925
    alf :    0     0   326     0     0     0     0     0     0     0     0     0 :   326
   hay1 :    0     0     0   449     1     0     1     0     0     0     0     0 :   451
   hay2 :    0     3     0     0   362     0     0     0     0     0     0     0 :   365
   peas :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
     dv :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    wat :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   quar :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    urb :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  forst :    0     0     0     0     0     0     0     0     0     0     0    45 :    45
          ------------------------------------------------------------------------
             0   717   326   449   574     0     1     0     0     0     0    45    2112


     CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(1)
     ----------------------------------------------------------------------
       wet        0           0.0            0.0          100.0       1.0000
      soil      925          22.8            0.4           77.2       0.6546
       alf      326           0.0            0.0          100.0       1.0000
      hay1      451           0.4            0.0           99.6       0.9944
      hay2      365           0.8           36.9           99.2       0.9887
      peas        0           0.0            0.0          100.0       1.0000
        dv        0           0.0          100.0          100.0       1.0000
       wat        0           0.0            0.0          100.0       1.0000
      quar        0           0.0            0.0          100.0       1.0000
       urb        0           0.0            0.0          100.0       1.0000
      uncl        0           0.0            0.0          100.0       1.0000
     forst       45           0.0            0.0          100.0       1.0000
     ----------------------------------------------------------------------

  OVERALL ACCURACY          89.8 %

  CLASS AVERAGE             75.2 %

  KHAT                      0.8608
```

CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 23

```
                                  OBSERVED:                                  OBSERVED:
VERIFIED:

 CLASS:   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
        ---------------------------------------------------------------------------
   wet :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  soil :    0  1508     0     0     0     0     0     0     0     0     0     0 :  1508
   alf :    0     0   531     0     0     0     0     0     0     0     0     0 :   531
  hay1 :   89     0     0   328     1     0     0     0     0     0     0     0 :   418
  hay2 :    0   349     0    65    16     0     0     0     0     0     0     0 :   430
  peas :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    dv :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   wat :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  quar :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   urb :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
 forst :    0     0     0     0     0     0     0     0     0     0     0    82 :    82
        ---------------------------------------------------------------------------
           89  1857   531   393    17     0     0     0     0     0     0    82    2969


   CLASS:  # PIXELS:     % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(1)
   ---------------------------------------------------------------------------
     wet       0             0.0          100.0          100.0         1.0000
    soil    1508             0.0           18.8          100.0         1.0000
     alf     531             0.0            0.0          100.0         1.0000
    hay1     418            21.5           16.5           78.5         0.7518
    hay2     430            96.3            5.9            3.7         0.0317
    peas       0             0.0            0.0          100.0         1.0000
      dv       0             0.0            0.0          100.0         1.0000
     wat       0             0.0            0.0          100.0         1.0000
    quar       0             0.0            0.0          100.0         1.0000
     urb       0             0.0            0.0          100.0         1.0000
    uncl       0             0.0            0.0          100.0         1.0000
   forst      82             0.0            0.0          100.0         1.0000
   ---------------------------------------------------------------------------

 OVERALL ACCURACY:          83.0 %

 CLASS AVERAGE:             76.4 %

 KHAT:                      0.7306
```

CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 24

```
VERIFIED:                               OBSERVED:                                OBSERVED:

 CLASS.   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
         ------------------------------------------------------------------------
   wet :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  soil :    0  1035     0     0    10     0     0     0     0     0     0     0 :   1045
   alf :    0     0   385     0     0     0     0     0     0     0     0     0 :    385
  hay1 :   73     0     0   273     0     0     0     0     0     0     0     0 :    346
  hay2 :    0   362     0     0    38     0     0     0     0     0     0     0 :    400
  peas :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
    dv :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   wat :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  quar :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   urb :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
 forst :   51     0     0     0     0     0     0     0     0     0     0     0 :     51
         ------------------------------------------------------------------------
          124  1397   385   273    48     0     0     0     0     0     0     0     2227


  CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(i)
  ---------------------------------------------------------------------------
    wet         0           0.0          100.0           100.0     1.0000
   soil      1045           1.0           25.9            99.0     0.9743
    alf       385           0.0            0.0           100.0     1.0000
   hay1       346          21.1            0.0            78.9     0.7595
   hay2       400          90.5           20.8             9.5     0.0751
   peas         0           0.0            0.0           100.0     1.0000
     dv         0           0.0            0.0           100.0     1.0000
    wat         0           0.0            0.0           100.0     1.0000
   quar         0           0.0            0.0           100.0     1.0000
    urb         0           0.0            0.0           100.0     1.0000
   uncl         0           0.0            0.0           100.0     1.0000
  forst        51         100.0            0.0             0.0     0.0000
  ---------------------------------------------------------------------------

 OVERALL ACCURACY:       77.7 %

 CLASS AVERAGE.          57.5 %

 KHAT.                   0.6599
```

CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 25

VERIFIED: OBSERVED: OBSERVED:

CLASS:	wet	soil	alf	hay1	hay2	peas	dv	wat	quar	urb	uncl	forst	
wet	0	0	0	0	0	0	0	0	0	0	0	0	0
soil	0	1508	0	0	0	0	0	0	0	0	0	0	1508
alf	0	0	531	0	0	0	0	0	0	0	0	0	531
hay1	14	0	0	336	2	0	0	0	0	0	66	0	418
hay2	0	0	0	0	430	0	0	0	0	0	0	0	430
peas	0	0	0	0	0	0	0	0	0	0	0	0	0
dv	0	0	0	0	0	0	0	0	0	0	0	0	0
wat	0	0	0	0	0	0	0	0	0	0	0	0	0
quar	0	0	0	0	0	0	0	0	0	0	0	0	0
urb	0	0	0	0	0	0	0	0	0	0	0	0	0
uncl	0	0	0	0	0	0	0	0	0	0	0	81	82
forst	1	0	0	0	0	0	0	0	0	0	0	0	
	15	1508	531	336	432	0	0	0	0	0	66	81	2969

CLASS:	# PIXELS:	% OMISSION:	% COMMISSION:	% CLASS TOTAL:	KHAT(i)
wet	0	0.0	100.0	100.0	1.0000
soil	1508	0.0	0.0	100.0	1.0000
alf	531	0.0	0.0	100.0	1.0000
hay1	418	19.6	0.0	80.4	0.7788
hay2	430	0.0	0.5	100.0	1.0000
peas	0	0.0	0.0	100.0	1.0000
dv	0	0.0	0.0	100.0	1.0000
wat	0	0.0	0.0	100.0	1.0000
quar	0	0.0	0.0	100.0	1.0000
urb	0	0.0	0.0	100.0	1.0000
uncl	0	0.0	100.0	100.0	1.0000
forst	82	1.2	0.0	98.8	0.9875

OVERALL ACCURACY: 97.2 %

CLASS AVERAGE: 95.8 %

KHAT: 0.9584

CONFUSION MATRIX FOR VALIDATION AREAS FROM Figure 26

```
VERIFIED:                              OBSERVED:                          OBSERVED:

  CLASS:   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
         ---------------------------------------------------------------------
    wet :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   soil :    0  1185     0     0     0     0     0     0     0     0     0     0 :  1185
    alf :    0     0   369     0     0     0     1     0     0     0     0     0 :   370
   hay1 :    0     0     0   328     0    38     0     0     0     0     2     0 :   368
   hay2 :    0     0     0     0   343     0     0     0     0     0     0     0 :   343
   peas :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
     dv :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    wat :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   quar :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    urb :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  forst :    0     0     0     0     0     0     1     0     0     0     0    52 :    53
         ---------------------------------------------------------------------
             0  1185   369   328   343    38     2     0     0     0     2    52    2319


        CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(i)
        ---------------------------------------------------------------------
          wet         0           0.0            0.0           100.0     1.0000
         soil      1185           0.0            0.0           100.0     1.0000
          alf       370           0.3            0.0            99.7     0.9968
         hay1       368          10.9            0.0            89.1     0.8734
         hay2       343           0.0            0.0           100.0     1.0000
         peas         0           0.0          100.0           100.0     1.0000
           dv         0           0.0          100.0           100.0     1.0000
          wat         0           0.0            0.0           100.0     1.0000
         quar         0           0.0            0.0           100.0     1.0000
          urb         0           0.0            0.0           100.0     1.0000
         uncl         0           0.0          100.0           100.0     1.0000
        forst        53           1.9            0.0            98.1     0.9807
        ---------------------------------------------------------------------

     OVERALL ACCURACY:          98.2 %

     CLASS AVERAGE.             97.4 %

     KHAT                       0.9729
```

```
CONFUSION MATRIX FOR VALIDATION AREAS FROM:  Figure 27

VERIFIED:                          OBSERVED:                              OBSERVED:

  CLASS:   wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
          ---------------------------------------------------------------------
    wet :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   soil :    0  1042     0     0     3     0     0     0     0     0     0     0 :  1045
    alf :    0     0   382     0     0     0     2     0     0     0     1     0 :   385
   hay1 :    3     0     0   293    15    33     0     0     0     0     2     0 :   346
   hay2 :    0     0     0     0   400     0     0     0     0     0     0     0 :   400
   peas :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
     dv :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    wat :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   quar :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    urb :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  forst :    0     0     0     0     0     0     0     0     0     0     0    51 :    51
          ---------------------------------------------------------------------
             3  1042   382   293   418    33     2     0     0     0     3    51      2227


   CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:    KHAT(1)
   --------------------------------------------------------------------------
     wet         0           0.0          100.0           100.0       1.0000
    soil      1045           0.3            0.0            99.7       0.9946
     alf       385           0.8            0.0            99.2       0.9906
    hay1       346          15.3            0.0            84.7       0.8236
    hay2       400           0.0            4.3           100.0       1.0000
    peas         0           0.0          100.0           100.0       1.0000
      dv         0           0.0          100.0           100.0       1.0000
     wat         0           0.0            0.0           100.0       1.0000
    quar         0           0.0            0.0           100.0       1.0000
     urb         0           0.0            0.0           100.0       1.0000
    uncl         0           0.0          100 0           100 0       1.0000
   forst        51           0.0            0 0           100 0       1.0000
   --------------------------------------------------------------------------

 OVERALL ACCURACY:         97 4 %

 CLASS AVERAGE:            96.7 %

 KHAT:                     0 9619
```

```
CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 28

VERIFIED:                          OBSERVED:                          OBSERVED:

  CLASS:   wet. soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
          ----------------------------------------------------------------------
    wet :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   soil :    0   925     0     0     0     0     0     0     0     0     0     0 :    925
    alf :    0     0   323     0     0     0     3     0     0     0     0     0 :    326
   hay1 :    0     0     0   415     2    29     0     0     0     0     5     0 :    451
   hay2 :    0     0     0     0   365     0     0     0     0     0     0     0 :    365
   peas :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
     dv :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
    wat :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   quar :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
    urb :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
   uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :      0
  forst :    0     0     0     0     0     0     0     0     0     0     0    45 :     45
          ----------------------------------------------------------------------
             0   925   323   415   367    29     3     0     0     0     5    45     2112

         CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(1)
         ---------------------------------------------------------------------------
           wet       0           0.0           0.0           100.0         1.0000
          soil     925           0.0           0.0           100.0         1.0000
           alf     326           0.9           0.0            99.1         0.9891
          hay1     451           8.0           0.0            92.0         0.9007
          hay2     365           0.0           0.5           100.0         1.0000
          peas       0           0.0         100.0           100.0         1.0000
            dv       0           0.0         100.0           100.0         1.0000
           wat       0           0.0           0.0           100.0         1.0000
          quar       0           0.0           0.0           100.0         1.0000
           urb       0           0.0           0.0           100.0         1.0000
          uncl       0           0.0         100.0           100.0         1.0000
         forst      45           0.0           0.0           100.0         1.0000
         ---------------------------------------------------------------------------

      OVERALL ACCURACY:        98.2 %

      CLASS AVERAGE:           98.2 %

      KHAT.                    0.9741
```

CONFUSION MATRIX FOR VALIDATION AREA FROM Figure 29

```
VERIFIED                                  OBSERVED                                   OBSERVED

 CLASS     wet  soil   alf  hay1  hay2  peas    dv   wat  quar   urb  uncl forst
        -----------------------------------------------------------------------
   wet .     0     0     0     0     0     0     0     0     0     0     0     0 :      0
  soil :     0  1507     0     0     1     0     0     0     0     0     0     0 :   1508
   alf :     0     0   529     0     0     0     0     0     0     0     0     2 :    531
  hay1 :     7     0     0   360     3    20     0     0     0     0    28     0 :    418
  hay2 :     0    13     0     0   417     0     0     0     0     0     0     0 :    430
  peas :     0     0     0     0     0     0     0     0     0     0     0     0 :      0
    dv :     0     0     0     0     0     0     0     0     0     0     0     0 :      0
   wat :     0     0     0     0     0     0     0     0     0     0     0     0 :      0
  quar :     0     0     0     0     0     0     0     0     0     0     0     0 :      0
   urb :     0     0     0     0     0     0     0     0     0     0     0     0 :      0
  uncl :     0     0     0     0     0     0     0     0     0     0     0     0 :      0
 forst :     0     0     0     0     0     0     9     0     0     0     0    73 :     82
        -----------------------------------------------------------------------
             7  1520   529   360   421    20     9     0     0     0    28    75     2969


 CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:    KHAT(i)
 -------------------------------------------------------------------------
   wet       0            0.0          100.0          100.0        1.0000
  soil    1508            0.1            0.9           99.9        0.9986
   alf     531            0.4            0.0           99.6        0.9954
  hay1     418           13.9            0.0           86.1        0.8421
  hay2     430            3.0            1.0           97.0        0.9648
  peas       0            0.0          100.0          100.0        1.0000
    dv       0            0.0          100.0          100.0        1.0000
   wat       0            0.0            0.0          100.0        1.0000
  quar       0            0.0            0.0          100.0        1.0000
   urb       0            0.0            0.0          100.0        1.0000
  uncl       0            0.0          100.0          100.0        1.0000
 forst      82           11.0            2.7           89.0        0.8374
 -------------------------------------------------------------------------

 OVERALL ACCURACY:        97.2 %

 CLASS AVERAGE:           94.3 %

 KHAT                     0.9533
```

CONFUSION MATRIX FOR VALIDATION AREAS FROM: Figure 30

```
VERIFIED:                              OBSERVED:                         OBSERVED:

 CLASS:   wet  soil   alf  hay1  hay2  peas   dv   wat  quar  urb  uncl forst
        ---------------------------------------------------------------------
   wet :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  soil :    0  1045     0     0     0     0     0     0     0     0     0     0 :  1045
   alf :    0     0   385     0     0     0     0     0     0     0     0     0 :   385
  hay1 :   49     0     0   229    29    33     4     0     0     0     2     0 :   346
  hay2 :    0   358     0     0    42     0     0     0     0     0     0     0 :   400
  peas :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
    dv :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   wat :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  quar :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
   urb :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
  uncl :    0     0     0     0     0     0     0     0     0     0     0     0 :     0
 forst :    2     0    15     0     0     0     0     0     0     0     0    34 :    51
        ---------------------------------------------------------------------
           51  1403   400   229    71    33     4     0     0     0     2    34    2227


     CLASS: # PIXELS:    % OMISSION:  % COMMISSION:  % CLASS TOTAL:   KHAT(i)
     ---------------------------------------------------------------------------
       wet         0            0.0          100.0           100.0    1.0000
      soil      1045            0.0           25.5           100.0    1.0000
       alf       385            0.0            3.8           100.0    1.0000
      hay1       346           33.3            0.0            66.2    0.6231
      hay2       400           89.5           40.8            10.5    0.0755
      peas         0            0.0          100.0           100.0    1.0000
        dv         0            0.0          100.0           100.0    1.0000
       wat         0            0.0            0.0           100.0    1.0000
      quar         0            0.0            0.0           100.0    1.0000
       urb         0            0.0            0.0           100.0    1.0000
      uncl         0            0.0          100.0           100.0    1.0000
     forst        51           33.3            0.0            66.7    0.6615
     ---------------------------------------------------------------------------


  OVERALL ACCURACY:        77.9 %

  CLASS AVERAGE            68 7 %

  KHAT.                    0 6603
```

The source code for the Semi-Supervised Two Stage Classification Technique may be found in the thesis copy in Memorial Library.