

5-1-1987

# Methods for Multisource Data Analysis in Remote Sensing

Jon Atli Bendsiktsson  
*Purdue University*

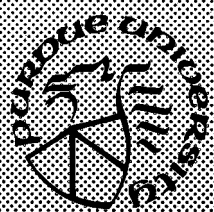
Philip H. Swain  
*Purdue University*

Follow this and additional works at: <https://docs.lib.purdue.edu/ecetr>

---

Bendsiktsson, Jon Atli and Swain, Philip H., "Methods for Multisource Data Analysis in Remote Sensing" (1987). *Department of Electrical and Computer Engineering Technical Reports*. Paper 570.  
<https://docs.lib.purdue.edu/ecetr/570>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact [epubs@purdue.edu](mailto:epubs@purdue.edu) for additional information.



# Methods for Multisource Data Analysis in Remote Sensing

Jon Atli Benediktsson  
Philip H. Swain

TR-EE 87-26  
May 1987

Laboratory for Applications of  
Remote Sensing  
and

School of Electrical Engineering  
Purdue University  
West Lafayette, Indiana 47907

# **METHODS FOR MULTISOURCE DATA ANALYSIS IN REMOTE SENSING**

**Jon Atli Benediktsson and Philip H. Swain**  
**TR-EE 87-26**  
**May 1987**

**Laboratory for Applications of Remote Sensing**  
**and**  
**School of Electrical Engineering**  
**Purdue University**  
**West Lafayette, Indiana 47907**

**This work was jointly supported by NASA Grant No. NAGW-925 "Earth Observation Research - Using Multistage EOS-like Data" (Principal Investigators: David A. Landgrebe and Chris Johannsen) and NSF Grant No. ECS-800324.**

## ACKNOWLEDGMENTS

The financial support of the National Aeronautics and Space Administration (NASA) through Grant No. NAGW-925 and the National Science Foundation through Grant No. ECS-800324 is gratefully acknowledged. We also would like to thank Hakil Kim for his help in digitizing a ground reference map.

## TABLE OF CONTENTS

	Page
LIST OF TABLES .....	v
LIST OF FIGURES.....	vii
ABSTRACT .....	viii
CHAPTER 1 - INTRODUCTION.....	1
CHAPTER 2 - PREVIOUS WORK .....	3
2.1 A Few Early Methods .....	3
2.2 Statistical Multisource Analysis.....	5
2.3 Dempster-Shafer Theory .....	9
2.3.1 Fundamentals in Dempster-Shafer Theory .....	10
2.3.2 Decision Rules .....	14
2.3.4 Example of Multisource Classification Using Dempster-Shafer Theory.....	15
2.4 Fuzzy Reasoning .....	17
2.5 Comparison of Multisource Classification Methods for Use in Processing of Remotely Sensed Data.....	19
CHAPTER 3 - THE APPROACH.....	22
3.1 General Concepts .....	22
3.2 Separability Estimates .....	28
3.2.1 Jeffries-Matusita Distance .....	28
3.2.2 The Transformed Divergence .....	29
3.3 The Method.....	30
CHAPTER 4 - EXPERIMENTAL RESULTS .....	33
4.1 General Remarks.....	33
4.2 Experiment 1: Unsupervised Analysis .....	37

	Page
4.3 Experiment 2: Supervised Analysis .....	47
4.4 General Observations .....	54
<b>CHAPTER 5 - CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH .....</b>	<b>57</b>
5.1 Discussion .....	57
5.2 Directions for Further Research.....	59
<b>LIST OF REFERENCES.....</b>	<b>63</b>

## LIST OF TABLES

Table	Page
4.1 Normalized Separability of Information Classes.....	35
4.2 Reliability Factors Determined from the Separability Measures for Classification of the Near-Infrared and Visible Sources.....	39
4.3 Reliability Factors Determined from the Separability Measures for Classification of the Near-Infrared and Thermal Sources .....	39
4.4 Reliability Factors Determined from Overall Classification Accuracy for Classification of the Near-Infrared and Visible Sources in Experiment 1 .....	40
4.5 Results of Experiment 1: Classification of the Near-Infrared and Visible Sources and Their Composite with Various Values of "Reliability" .....	41
4.6 Results of Experiment 1: Classification of the Near-Infrared and Thermal Sources and Their Composite with Various Values of "Reliability" .....	44
4.7 Reliability Factors Determined from Overall Classification Accuracy for Classification of the Near-Infrared and Thermal Sources in Experiment 1 .....	45
4.8 Reliability Factors Determined from Overall Classification Accuracy for Classification of the Near-Infrared and Visible Sources in Experiment 2 .....	47

Table	Page
4.9 Results of Experiment 2: Classification of the Near-Infrared and Visible Sources and Their Composite with Various Values of "Reliability" .....	50
4.10 Results of Experiment 2: Classification of the Near-Infrared and Thermal Sources and Their Composite with Various Values of "Reliability" .....	52
4.11 Reliability Factors Determined from Overall Classification Accuracy for Classification of the Near-Infrared and Thermal Sources in Experiment 2 .....	53



**LIST OF FIGURES**

<b>Figure</b>	<b>Page</b>
<b>3.1 The Family of Curves <math>p^a</math>.....</b>	<b>32</b>
<b>4.1 The Supervised Training Procedure.....</b>	<b>48</b>

## ABSTRACT

Methods for classifying remotely sensed data from multiple data sources are considered. Special interest is in general methods for multisource classification and three such approaches are considered: Dempster-Shafer theory, fuzzy set theory and statistical multisource analysis. Statistical multisource analysis is investigated further. To apply this method successfully it is necessary to characterize the "reliability" of each data source. Separability measures and classification accuracy are used to measure the reliability. These reliability measures are then associated with reliability factors included in the statistical multisource analysis. Experimental results are given for the application of statistical multisource analysis to multispectral scanner data where different segments of the electromagnetic spectrum are treated as "different" sources. Finally, a discussion is included concerning future directions for investigating reliability measures.

## CHAPTER 1

### INTRODUCTION

Computerized information extraction from remotely sensed imagery has been applied successfully over the last two decades. The data used in the processing has mostly been multispectral data and the statistical pattern recognition (multivariate classification) methods are now widely known. Within the last decade advances in space and computer technologies have made it possible to amass large amounts of data about the Earth and its environment. The data are now more and more typically not only spectral data but include, for example, forest maps, ground cover maps, radar data and topographic information such as elevation and slope data. We may therefore have many kinds of data from different sources regarding the same scene. These are called multisource data.

We are interested in using all these data to extract more information and get more accuracy in classification. However the conventional multivariate classification methods cannot be used satisfactorily in processing multisource data. This is due to several reasons. One is that the multisource data need not be just spectral; they can for example be elevation ranges or even non-numerical data such as ground cover classes or

soil types. The data are also not necessarily in common units and therefore scaling problems may arise. It is also desirable to determine the reliability of each source, because all the sources are in general not equally reliable. This all implies that other methods than the conventional multivariate classification have to be used to classify multisource data.

Various ad hoc methods have been proposed to classify multisource data. However, we are interested in developing more general methods which can be applied to classify any type of data. In particular, our attention is focused on statistical multisource analysis by means of a method based on Bayesian classification theory which was proposed recently by Swain, Richards and Lee [1]. An extension of this method will be developed in this report.

Our objective is to modify the method to take into account the relative reliabilities of the sources of data involved in the classification. This requires a way to quantify the reliability of a data source. Its importance becomes apparent when we look at the combination of information. The foundation of the method for combination from various sources consists essentially of multiplication of source-specific posterior probabilities from all the sources involved in the classification. If any of the sources are unreliable they can affect the outcome of the multiplication disproportionately and consequently increase classification error.

The goal of this report is to investigate methods to determine the reliability and define a corresponding reliability factor for each data source. The reliability factors are then included in the classification process. Experimental results will be given.

## CHAPTER 2

### PREVIOUS WORK

#### 2.1 A Few Early Methods

Several methods have been used in the past to classify multisource data. One method is the "ambiguity reduction" where the data are classified based on one or more of the data sources, the results from the classification are assessed, and other sources are then resorted to in order to resolve the remaining ambiguities. The ambiguity reduction can be achieved by logical sorting methods. Hutchinson has used this method successfully [2].

A second method is supervised relaxation labeling derived by Richards et al. [3] in order to merge data from multiple sources. This method, like other relaxation methods, tries to develop consistency among a collection of observations by means of an iterative numerical "diffusion" process. So far this method has not been fully investigated on multiple sources and its iterative nature makes it computationally very expensive.

A third method is to subdivide the data based on a subset of the data sources and then analyze each subdivision based on the remaining sources. In this method the data are subdivided in such a way that variation within

each subdivision is minimized or eliminated, due to some of the subdividing variables. An example of this method can be found in Strahler et al. [4].

None of the methods described above is a general approach in multisource classification and all of them depend heavily on the user. They all deal with the various sources of data independently. In contrast the fourth method mentioned here is a general approach which does not deal with the data sources independently. This method is the stacked-vector approach, i.e., formation of an extended vector with components from all of the data sources and handling the compound vector in the same manner as data from a single source. This method is the most straightforward and the simplest of the methods. It works very well if the data sources are similar and the relations between the variables are easily modeled [5]. However, the method is not applicable when the various sources cannot be described by a common model, e.g., the multivariate Gaussian model. Another drawback is that when the multivariate Gaussian model is used, the computational cost grows as the square of the number of dimensions. This makes the computational cost severe if the number of sources is large.

All the methods discussed up to this point have significant limitations as general approaches for multisource classification. Our goal is to develop a general method which can be used to classify complex data sets, containing both multispectral, topographic and other forms of geographic data. Three such methods are discussed below. First we discuss statistical multisource analysis, a probabilistic method which is based on Bayesian decision theory and was developed recently by Swain, Richards and Lee [1]. Then we address two non-probabilistic approaches for combining sources,

methods based on Dempster-Shafer theory and fuzzy set theory. We will review the main concepts of these three approaches and then pursue the one we think is most applicable in multisource classification of remotely sensed data.

## 2.2 Statistical Multisource Analysis

As noted previously, this method was proposed recently by Swain, Richards and Lee [1]. It is a general method which extends well-known concepts used for classification of multispectral images when only one data source is involved. In this method the various data sources are handled independently and each data source can be modeled by any appropriate model. The main concepts in the theory are addressed below.

Assume there are  $n$  separate data sources, each providing a measurement  $x_s$  ( $s = 1, \dots, n$ ) for each of the pixels of interest. If any of the sources is multidimensional, the corresponding  $x_s$  will be a measurement vector. Let there be  $M$  user-specified information classes in the scene (not necessarily a property of the data) denoted  $\omega_j$  ( $j = 1, \dots, M$ ). The pixels are to be classified into these classes.

Each data source is at first considered separately. For a given source, an appropriate training procedure can be used to segment or classify the data into a set of classes that will characterize that source. We could for example use clustering for this purpose. The data types are assumed to be very general, e.g., both topographic and multispectral data. We therefore refer to the source-specific classes or clusters as data classes, since they are defined from relationships in a particular data space.

The data classes are for instance spectral classes in the case of spectral data while for topographic data they may for example be elevation ranges. In general there may not be a simple one-to-one relation between the user-desired information classes and the set of data classes available. It is one of the requirements of a multisource analytical procedure to devise a method by which inferences about information classes can be drawn from the collection of data classes.

The  $i$ -th data class from the  $s$ -th source is denoted by  $d_{si}$  ( $i = 1, 2, \dots, m_s$ ), where  $m_s$  is the number of data classes for source  $s$ . The measurement vectors are associated with data classes according to a set of data-specific membership functions,  $f(d_{si}|x_s)$ . This means that for a given measurement from the  $s$ -th source,  $f(d_{si}|x_s)$  gives the strength of association of  $x_s$  with data class  $d_{si}$  defined for that source.

The information classes  $\omega_j$  are related to the data classes from a single source by means of a set of source-specific membership functions  $f(\omega_j|d_{si}(x_s))$ , for all  $i, j, s$ , where  $f(\omega_j|d_{si}(x_s))$  is the strength of association of data class  $d_{si}$  with information class  $\omega_j$ , possibly influenced by the value of  $x_s$ . This expression is different from previous approaches for single source classification, where it is often assumed in the analysis that there is a unique correspondence between spectral and information classes, once prior probabilities have been determined.

Now a set of global membership functions is defined, that collect together the inferences concerning a single information class from all of the data sources (as represented by their data classes). The membership function  $F_j$  for class  $\omega_j$  is of the general form:



$$F_j = F_j[f(\omega_j | d_{si}(x_s)), r_s] \quad (i=1,2, \dots, m_s \quad s=1,2, \dots, n) \quad (2.1)$$

where  $r_s$  is the quality or reliability factor of the  $s$ -th source and is defined to weight the various sources, reflecting the perceived or measured reliabilities of the various sources of data. This is very important because it may be known that all the sources are not equally reliable and therefore the analyst is allowed to take into account his confidence in the recommendation of each of the individual sources of data available.

Finally a pixel  $X = [x_1, \dots, x_n]^T$  is classified according to the usual maximum selection rule, i.e., it is decided that  $X$  is in class  $\omega^*$  for which

$$F^* = \max_j F_j \quad (2.2)$$

Now the membership functions are defined specifically. From experience with Bayesian classification theory a natural choice for the global membership function is the joint-source posterior probabilities.

$$F_j(X) = p(\omega_j | X) = p(\omega_j | x_1, x_2, \dots, x_n) \quad (2.3)$$

If we make the assumption that the data sources are statistically independent, the global membership function may be written [1]:

$$F_j(X) = [p(\omega_j)]^{1-n} \prod_{s=1}^n p(\omega_j | x_s) \quad (2.4)$$

It may be argued that independence between two unrelated sources is unlikely and the independence assumption may therefore introduce errors. On the other hand there are mainly two reasons why use of the independence assumption is desirable in this case. First, it is clear that interactions between two data sources can be very complex and consequently

hard to model. To make use of dependence between sources these interactions have to be modeled, but we are either unable or unwilling to do that. Secondly, taking dependence into account will increase the computational complexity of the classification procedure and may impose considerable burden on the computer resources available. Using this reasoning, independence between data sources is justified in the global membership function.

Now consider the individual source-specific membership functions which appear here explicitly as source-specific posterior probabilities. These can be expressed as:

$$p(\omega_j | \mathbf{x}_s) = \sum_{i=1}^{m_s} p(\omega_j | d_{si}, \mathbf{x}_s) p(d_{si} | \mathbf{x}_s) \quad (2.5)$$

where the source-specific membership functions appear explicitly as  $p(\omega_j | d_{si}, \mathbf{x}_s)$  and the data-specific membership functions as  $p(d_{si} | \mathbf{x}_s)$ . Another way to write (2.5) is:

$$p(\omega_j | \mathbf{x}_s) = \sum_{i=1}^{m_s} p(\mathbf{x}_s | \omega_j, d_{si}) p(d_{si} | \omega_j) p(\omega_j) / p(\mathbf{x}_s) \quad (2.6)$$

Implementation of the classification technique involves using (2.5) or (2.6) to determine the posterior probabilities in (2.4) and then (2.2) is used for the decision. In turn the quantities in (2.5) or (2.6) as appropriate have to be estimated. It is now interesting to look at equations (2.5) and (2.4) taken together. In (2.5) we are just looking at one source at a time. There we see explicitly the relation between the data vectors and the data classes and the information classes, demonstrating the role of data classes as intermediaries. Equation (2.4) then aggregates the information from all the

sources of data for each specific information class.

As seen above, statistical multisource analysis is an extension of one source Bayesian classification. We now turn away from the Bayesian framework and look at combination of sources using Dempster-Shafer theory and fuzzy set theory.

### **2.3 Dempster-Shafer Theory**

Several approaches for dealing with the problem of quantifying uncertainty have been proposed in the literature. One approach comes from the works of Dempster and Shafer in connection with a mathematical theory of evidence. The theory as described in Shafer [6] is a departure from the traditional Bayesian approach in that mass is assigned to some subsets, whereas uncertainty is spread over all subsets.

In this respect the traditional Bayes approach has been rejected by many authors because [7,8]:

- 1) Knowledge is conditional on the past and this requires large amounts of statistical data.
- 2) It is difficult to ensure and maintain consistency in a collection of interrelated propositions. This also stems from the need to assign point probability values even when the underlying models from which these values are derived are incapable of supplying such precise data.

- 3) Uncertainty about a proposition implies near certainty about the negation of that proposition, i.e., Bayesian theory cannot distinguish between the lack of belief and disbelief.

### 2.3.1 Fundamentals in Dempster-Shafer Theory

The idea is to use a number between zero and one to indicate the degree of support a body of evidence provides for a proposition. The fundamental concept in Dempster-Shafer theory is the basic probability assignment  $m$ . For a set  $A$ ,  $m(A)$  measures the belief that is committed exactly to  $A$  alone. It can be defined in the following way:

Definition: Assume  $m$  is a set mapping from subsets of the finite set  $X$  into the unit interval, i.e.,

$$m : 2^X \rightarrow [0,1]$$

such that:

- 1)  $m(\phi) = 0$  (where  $\phi$  is empty)
- 2)  $\sum_{A \subseteq X} m(A) = 1$

$m$  is then called a basic probability assignment. It is worthwhile to note that:

- 1)  $m(X)$  is not necessarily one.
- 2)  $A \subseteq B$  does not necessarily imply  $m(A) \leq m(B)$
- 3) It is allowed that belief not be committed to either  $A$  or  $A^c$ .

This quantity  $m(A)$  measures the belief that one commits exactly to  $A$ , not the total belief that one commits to  $A$ . To obtain the measure of the total belief committed to  $A$ , one must add to  $m(A)$  the quantities  $m(B)$  for all proper subsets  $B$  of  $A$ . Then a belief function can be defined in the following way:

Definition: Given a basic probability assignment  $m$ , define the belief function:

$$\text{Bel} : 2^X \rightarrow [0,1]$$

such that for any  $A \subseteq X$ :

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B) \quad (2.7)$$

The evidence for a proposition  $A$  is described by a subinterval  $[s(A), p(A)]$  of the unit interval  $[0,1]$ , where

$$s(A) = \text{Bel}(A) \quad (2.8)$$

$$p(A) = 1 - s(A^c) \quad (2.9)$$

The lower value,  $s(A)$ , represents the "support" for the proposition and sets a minimum value for its likelihood. The upper value,  $p(A)$ , denotes the "plausibility" of that proposition and establishes a maximum

likelihood. Support may be interpreted as the total positive effect that a body of evidence has on a proposition, while plausibility represents the total extent to which a body of evidence fails to refute a proposition. The degree of uncertainty about the actual probability value for a proposition corresponds to the width of its evidential interval; i.e.,  $p(A) - s(A)$ . If this difference is zero for all propositions, the system is Bayesian [8].

For example if we represent a proposition A using the notation  $A_{[s(A),p(A)]}$ , then [8]:

$A_{[0,1]}$	There is no knowledge at all about A.
$A_{[0,0]}$	A is false.
$A_{[1,1]}$	A is true.
$A_{[.20,1]}$	Evidence provides partial support for A.
$A_{[0,.80]}$	Evidence provides partial support for $A^c$ .
$A_{[.20,.80]}$	Probability of A is between .20 and .80. Evidence provides simultaneously support for both A and $A^c$ .

An important part of Shafer's theory involves the combination of belief functions to form a composite belief function, i.e., combining various sources of evidence. Shafer accomplishes this by use of Dempster's rule of combination, sometimes called Dempster's orthogonal sum. This gives the aggregated mass that can be assigned to the labeling proposition X.

$$m(X) = \frac{\sum_{A \cap B = X} m_1(A)m_2(B)}{1 - \sum_{A \cap B = \phi} m_1(A)m_2(B)} \quad (2.10)$$

We may call  $\text{Bel}_1 \oplus \text{Bel}_2$  the orthogonal sum of  $\text{Bel}_1$  and  $\text{Bel}_2$ . Because of the commutativity and associativity of the belief functions:

$$\text{Bel}_1 \oplus \text{Bel}_2 = \text{Bel}_2 \oplus \text{Bel}_1 \quad (2.11a)$$

$$(\text{Bel}_1 \oplus \text{Bel}_2) \oplus \text{Bel}_3 = \text{Bel}_1 \oplus (\text{Bel}_2 \oplus \text{Bel}_3) \quad (2.11b)$$

we form pairwise sums and combine two functions at a time to accomplish the combination.

To illustrate use of Dempster-Shafer theory further we give a simple example using two sources of evidence. In this example the sets  $A$  and  $A^c$  are subsets of the set  $\Theta$  which is usually referred to as the "frame of discernment."

For source # 1 we have:

$$A = \{a\} \quad A^c = \{b, c\} \quad \Theta = \{a, b, c\}$$

We assign the basic probability assignments in the following way:

$$m(A) = 0.6 \quad m(A^c) = 0.3 \quad m(\Theta) = 0.1$$

Then we can calculate the support and plausibility for each set by using equations (2.8) and (2.9). This calculation gives:

$$s(A) = 0.6 \quad s(A^c) = 0.3 \quad s(\Theta) = 0.6 + 0.3 + 0.1 = 1$$

$$p(A) = 1 - 0.3 = 0.7 \quad p(A^c) = 1 - 0.6 = 0.4 \quad p(\Theta) = 1 - 0 = 1$$

We can therefore write:

$$A_{[.6,.7]} \quad A^c_{[.3,.4]} \quad \Theta_{[1,1]}$$

Now for source # 2 we have the same sets:

$$A = \{a\} \quad A^c = \{b,c\} \quad \Theta = \{a,b,c\}$$

However, the basic probability assignments are different:

$$m(A) = 0.3 \quad m(A^c) = 0.7 \quad m(\Theta) = 0.0$$

Using these data we now get:

$$s(A) = 0.3 \quad s(A^c) = 0.7 \quad s(\Theta) = 0.3 + 0.7 = 1$$

$$p(A) = 1 - 0.7 = 0.3 \quad p(A^c) = 1 - 0.7 = 0.3 \quad p(\Theta) = 1$$

We can now write:

$$A_{[.3,.3]} \quad A^c_{[.7,.7]} \quad \Theta_{[1,1]}$$

To calculate the aggregated mass from these two sources we can now use Dempster's rule (equation (2.10)). That calculation gives:

$$m(A) = \frac{0.6 \cdot 0.3 + 0.3 \cdot 0.1}{1 - (0.6 \cdot 0.7 + 0.3 \cdot 0.3)} = 0.43$$

$$m(A^c) = \frac{0.3 \cdot 0.7 + 0.7 \cdot 0.1}{1 - (0.6 \cdot 0.7 + 0.3 \cdot 0.3)} = 0.57$$

### 2.3.2 Decision Rules

In statistical pattern recognition methods there is usually a straightforward way to select a decision rule to use in deciding the preferred label among a range of options. For maximum likelihood



algorithms the rule is usually expressed in terms of the most favored label. This is also the case for the multisource statistical technique described above in which class membership is decided on the basis of maximizing the global membership function.

This is not the case, though, with evidential methods, where an evidential interval bounded by support and plausibility rather than a single value is attached to candidate class labels. In that case one has a number of options potentially to choose among for a decision rule [9].

Some of the candidates are:

- 1) A maximum support rule, where the labeling proposition with the highest support is chosen.
- 2) A maximum plausibility rule, where the proposition with the highest plausibility is chosen.
- 3) An absolute rule, where the proposition whose support exceeds all other plausibilities is chosen. If the width of the evidential interval is larger than the difference between the two highest supports, this rule will not give a decision.
- 4) A maximum support and plausibility rule, where the label chosen has both the highest support and plausibility.

#### **2.3.4 Example of Multisource Classification Using Dempster-Shafer Theory**

Kim et al. [10] have applied Dempster-Shafer theory to multisource data. They use a distance measure as the weight of evidence for data

classification to determine the degrees of support based on the multispectral, digital elevation and digital slope data. In their work the Mahalanobis distance is used to take into account correlation and dispersion of samples.

They define the measure of support for a certain class  $\omega_i$  as:

$$B_i(z') = 1 - P_i(Z \leq z') = 1 - F_Z(z') \quad (2.12)$$

where  $z'$  denotes the distance from the mean vector of  $\omega_i$  to a given observation vector  $X$ .  $P_i(Z \leq z')$  is the probability of the event  $(Z \leq z')$  for samples in  $\omega_i$  and  $F_Z(z')$  is the cumulative distribution function of  $Z$ .

It is easy to see that the function  $B_i(\cdot)$  has the properties:

- 1)  $B_i: [0, \infty] \rightarrow [0, 1]$
- 2)  $B_i$  is nonincreasing.
- 3)  $B_i(0) = 1$  and  $B_i(\infty) = 0$

Properties (2) and (3) correspond to the human intuition that the disbelief in the hypothesis  $X$  belonging to class  $\omega_i$  increases as the distance between the mean and  $X$  increases. Thus  $1 - F_Z(z')$  may be considered as the measure of support for the hypothesis.

Kim et al. use  $B_i$  to find the support for the proposition that pixel  $X$  in source  $s$  belongs to class  $\omega_i$ . They calculate this for each source and then use Dempster's rule to combine the evidence from all the sources, so the pixel can be classified using any appropriate decision rule.

## 2.4 Fuzzy Reasoning

Aside from Dempster-Shafer theory another way to deal with uncertainty is to apply the notion of fuzzy or monotonic measures which initially comes from the work of Zadeh [11]. In fuzzy theory a fuzzy set is a class of objects with a continuum of grades of membership. Such a set is characterized by a membership function which assigns to each object a grade of membership ranging between zero and one. Therefore for a fuzzy subset  $A$  of the universe set  $A_0$ , with membership function  $\mu_A(x)$ , we have:

$$\sum_A \mu_A(a_i) \leq 1 \quad \text{for all } a_i \quad (2.13)$$

This is very different from conventional ("crisp") set theory where we have an "on/off" membership function that takes only values 0 or 1, i.e., we place our full confidence in an element being a member of particular set or not [12]. To illustrate this concept further, we know for conventional sets that the Bayesian probability of the subset  $A$  is:

$$P(A) = \sum_{a_i \in A} p(a_i) \quad (2.14a)$$

On the other hand in fuzzy set theory the corresponding probability is:

$$P(A) = \sum_{a_i \in A} \mu_A(a_i) p(a_i) \quad (2.14b)$$

where  $p(\ )$  is the probability density and  $\mu_A(\ )$  is the membership function.

In combining evidence from multiple sources, fuzzy theory has been used in combination with Dempster-Shafer theory. Ishizuka [13] and Ishizuka et al. [14] have extended Dempster-Shafer theory to include fuzzy sets. They define the degree that a fuzzy subset  $A_1$  is included in another

fuzzy subset  $A_2$  of the same universe set  $A_0$  as:

$$I(A_1 \subseteq A_2) = \frac{\min_a(1, 1 - \mu_{A_1}(a) + \mu_{A_2}(a))}{\max_a(\mu_{A_1}(a))} \quad (2.15)$$

where  $\mu_{A_1}$  and  $\mu_{A_2}$  are the membership functions of  $A_1$  and  $A_2$  respectively. The denominator is called the height of the fuzzy subset and equation (2.15) takes the value 1 if  $A_1$  is completely included in  $A_2$  and 0 if  $\mu_{A_2}(a) = 0$  at the point where  $\mu_{A_1}(a)$  takes its peak value.

They also define the degree of intersection of two fuzzy subsets  $A_1$  and  $A_2$  as:

$$J(A_1, A_2) = \frac{\max_a(\mu_{A_1 \cap A_2}(a))}{\min(\max_a(\mu_{A_1}(a)), \max_a(\mu_{A_2}(a)))} \quad (2.16)$$

where the membership function of the intersection  $A_1 \cap A_2$  is defined in fuzzy set theory as:

$$\mu_{A_1 \cap A_2}(a) = \min(\mu_{A_1}(a), \mu_{A_2}(a)) \quad (2.17)$$

The denominator of (2.16) is 1 if the fuzzy subsets  $A_1$  and  $A_2$  are normalized, i.e., iff for all  $a \in A$ :

$$\mu_{A_1}(a) = \mu_{A_2}(a) = 1$$

The degree that the intersection of  $A_1$  and  $A_2$  is  $\phi$  (empty) is defined as:

$$1 - J(A_1, A_2)$$

If now an extended Dempster-Shafer probability assignment  $m(A)$  is defined for each fuzzy subset  $A$  characterized by  $\mu_A(a)$  then equations (2.15) and (2.16) can be used to define a belief function and a combination rule

which are direct extensions of the ones in Dempster-Shafer theory. The belief function is then:

$$\text{Bel}(A_i) = \sum_{A_j} I(A_j \subseteq A_i) m(A_j) \quad (2.18)$$

The combination rule is an extension of Dempster's rule:

$$m(A_k) = \frac{\sum_{A_{1i} \cap A_{2j} = A_k} J(A_{1i}, A_{2j}) m_1(A_{1i}) m_2(A_{2j})}{\sum_{A_{1i}, A_{2j}} (1 - J(A_{1i}, A_{2j})) m_1(A_{1i}) m_2(A_{2j})} \quad (2.19)$$

This extension of Dempster-Shafer theory makes it possible to use the decision rules described in 2.3.2.

Several other methods of combining fuzzy sets have been addressed in the literature. Two of them are listed below but will not be discussed any further here.

- 1) Taking minimum and maximum of the membership functions [15].
- 2) Using linguistic probability [16].

## 2.5 Comparison of Multisource Classification Methods for Use in Processing of Remotely Sensed Data

We have now described methods used for classification of multisource data. As said earlier, we are only interested in general methods, not in ad hoc methods. There were three general approaches discussed in this chapter.

Dempster-Shafer theory deals with uncertainty in the data measurements and is widely recognized and studied. It has been examined in expert systems [17] and is now being used in geographic information processing [18]. This approach has some problems, which include how to give values to the basic probability assignment and what decision rule to choose. These problems are highly application-specific in nature.

Fuzzy set theory deals with uncertainty, but in a different way, and has not been used extensively in classification of remotely sensed data. Some authors have examined clustering with fuzzy techniques [19,20] and other have addressed combination of evidence using fuzzy sets as described in section 2.4. The problems with this approach are similar to the ones using Dempster-Shafer theory. Here we have to specify a membership function for each set and it is not evident what is the best way to do that.

It is interesting to note here that although Dempster-Shafer theory and fuzzy set theory have more mechanism to handle uncertainty than Bayesian decision theory does, Bayesian statisticians do not think very highly of these theories. Berger for example views them either as unnecessary elaborations on robust probabilistic analysis or as insufficiently complicated representations of reality [21]. On the other hand we do have much more experience with Bayesian classification theory when processing remotely sensed data. Statistical methods such as the maximum likelihood method have been used for a long time in conventional one-source classification. The statistical multisource method by Swain, Richards and Lee is an extension of such methods. It is therefore a reasonable choice in our analysis. This method also does not have any of the problems associated

with the two approaches above. However, the method as presented by Swain, Richards and Lee does not provide a mechanism to account for varying degrees of reliability of different sources as do Dempster-Shafer theory or fuzzy set theory. It is our belief that this problem can be overcome if we assign reliability factors to each source involved in the classification. For these reasons we will investigate a modified version of the statistical multisource analysis by Swain, Richards and Lee by means of which reliability analysis is added to the classification process.

## CHAPTER 3

### THE APPROACH

#### 3.1 General Concepts

From the Swain, Richards and Lee approach we have the global membership function [1]:

$$\begin{aligned}
 F_j(X) &= [p(\omega_j)]^{1-n} \prod_{i=1}^n p(\omega_j | x_i) \\
 &= [p(\omega_j)]^{1-n} p(\omega_j | x_1) p(\omega_j | x_2) \dots p(\omega_j | x_n)
 \end{aligned} \tag{3.1}$$

We want to associate reliability factors with the sources as discussed in chapter 2, i.e., to express quantitatively our confidence in each source, and use them for classification purposes. This is very important because we need to increase the influence of the "more reliable" sources, i.e., the sources we have more confidence in, on the global membership function and consequently decrease the influence of the "less reliable" sources in order to improve the classification accuracy. The need for reliability factors becomes apparent if we look at equation (3.1) where the global membership function is a product of posterior probabilities related to each source. Each probability has value in the interval from 0 to 1. If any one of them is near zero it will carry the value of the membership function close to zero and therefore downgrade drastically the contribution of information from other



sources, although the particular source involved may have little or no reliability.

From above it is clear that we have to put weights (reliability factors) on the sources which will influence their contributions to classification. Since we have a product of posterior probabilities this weight has to be involved in such a way that when the reliability of a source is low it must discount the influence of that source and when the reliability of a source is high it must give the source relatively high influence. One possible choice for this kind of analysis is to put reliability factors as exponents on the posterior probabilities of each source. Then equation (3.1) would be written in the following form:

$$\begin{aligned} F_j(X) &= [p(\omega_j)]^{1-n} p(\omega_j | x_1)^{a_1} \dots p(\omega_j | x_n)^{a_n} \\ &= [p(\omega_j)]^{1-n} \prod_{i=1}^n p(\omega_j | x_i)^{a_i} \end{aligned} \quad (3.2)$$

Equation (3.2) can also be written in a logarithmic form as:

$$\log F_j(X) = (1 - n) \log p(\omega_j) + \sum_{i=1}^n a_i \log p(\omega_j | x_i) \quad (3.2a)$$

where the reliability factors are expressed as the coefficients in the sum. These coefficients act like weights in the sum and control the influence of a source on the global membership function. If a coefficient is high compared to the other coefficients, the source it represents will have greater influence on the global membership function. If on the other hand a coefficient is low compared to other coefficients, it will decrease the influence of its source. Another way to see this is to look at the sensitivity of the global

membership function to changes in one of the posterior probabilities which can be expressed as [9]:

$$\frac{\delta F_j(X)}{F_j(X)} = a_i \frac{\delta p(\omega_j | x_i)}{p(\omega_j | x_i)}$$

We select the  $a_i$ 's ( $i = 1, \dots, n$ ) in the interval  $[0,1]$  because of the following reasons. If source  $i$  has no reliability ( $a_i=0$ ) it will not have any influence on (3.2) because  $p(\omega_j | x_i)^0 = 1$ , and if source  $i$  has the highest reliability then it will give a full contribution to (3.2) because  $p(\omega_j | x_i)^1 = p(\omega_j | x_i)$ . It is also worthwhile to note that this method of putting exponents on the posterior probabilities does not change the decision for a single-source classification because the exponential function  $p^a$  is a monotonic function of  $p$ .

To illustrate the last point, consider a simple example. In this example assume that we have one source, that  $a$  is a number in the interval  $(0,1]$ , and that we have just two information classes  $\omega_1$  and  $\omega_2$ . We are observing one ground element  $x$  and the global membership functions  $F_1$  and  $F_2$  are of the form in (3.1):

$$F_1(x) = p(\omega_1 | x) \tag{3.3a}$$

$$F_2(x) = p(\omega_2 | x) \tag{3.3b}$$

Assume now that  $p(\omega_1 | x) > p(\omega_2 | x)$ . Using the maximum selection rule we decide  $x$  belongs to  $\omega_1$ . Now applying the exponent method above, the global membership functions will be of the form in equation (3.2):

$$F_1'(x) = p(\omega_1|x)^a \quad (3.4a)$$

$$F_2'(x) = p(\omega_2|x)^a \quad (3.4b)$$

Keeping in mind that  $p(\omega_1|x)$  and  $p(\omega_2|x)$  are numbers in the interval  $[0,1]$ ,  $a$  is a number in the interval  $(0,1]$  and  $p(\omega_1|x) > p(\omega_2|x)$  we get:

$$p(\omega_1|x)^a > p(\omega_2|x)^a$$

Therefore the decision is the same for this particular  $x$ , i.e., we classify  $x$  to  $\omega_1$ . This of course applies for all ground-elements  $x$  while  $a \in (0,1]$ . If  $a = 0$  we get no decision, but in case we are considering multisource data this source will have no influence on (3.2) and the decision will depend on the other sources. When we combine two or more sources, the global membership function becomes more complex to analyze because it consists of a product of posterior probabilities with different reliability factors and this product is normalized by the priori probabilities.

The problem is to determine the  $a_i$ 's based on the reliability of the sources. We think  $a$  of source as being reliable if its contribution to the combination of information from various sources is "good", i.e., if we increase the classification accuracy substantially or extract more information by using this particular source. Using this understanding of a reliable source we apply two measures to determine the reliability of a source: weighted average separability and overall classification accuracy.

It is our belief that we can call a source reliable if the separability of the information classes is high for the source. If on the other hand the separability of the information classes is low, we can assume that the source

is not very reliable. Therefore one possibility for reliability evaluation is to use the average separability of the information classes in each source, e.g., average Jeffries-Matusita (JM) distance, average transformed divergence or any other separability function. What kind of average is used depends on what we are after in the multisource classification. For instance if we are trying to improve the overall classification accuracy we use the weighted overall average. If, however, we are concentrating just on specific classes, the weighted average separability of those information classes is used.

Another way to measure reliability of a data source is to use the classification accuracy of the source. In this case we call a source reliable if the classification accuracy for the source is high, but if the accuracy is low we call the source unreliable. This approach is related to the method of using separability measures in that increased separability gives higher accuracy.

As said earlier we want the reliability factors to have values in the interval  $[0,1]$ . We also want to associate the reliability factors to values of some separability measure or to the classification accuracy. If we choose to use the values of the separability measures to determine reliability factors, we know that some separability functions have saturating behavior as functions of normed distance, e.g., the transformed divergence and JM-distance. We know beforehand that they take values in some interval  $[\min, \max]$  and we simply have to norm them by division and/or subtraction. Thus for separability function  $f(x)$  we calculate:

$$a = \frac{f(x - \min)}{\max - \min}$$

so  $a$  takes value in  $[0,1]$ . Some separability estimates, e.g., the divergence, do not have this saturating behavior and increase with increased normed distance. In that case we have to specify a cutoff point somewhere on the curve as our maximum value to saturate the function. This means that every value higher than this cutoff will be mapped to the cutoff value. This saturation is done to limit the influence or dominance of "very separable" classes on the weighted average of the separability. We choose a specific cutoff value which reflects our belief that the information classes which have separability higher than this value are "separable enough." We then use this "saturated" curve in the same manner as described above.

It remains to be shown whether the simple mapping described above is sufficient to produce appropriate values for the reliability factor. That will be discussed further in section 3.3. We shall now look more closely at separability estimation.

### 3.2 Separability Estimates

In this research we look at two separability estimates, the JM-distance and the transformed divergence.

#### 3.2.1 Jeffries-Matusita Distance

The JM distance between two classes  $\omega_i$  and  $\omega_j$  is defined formally as:

$$J_{ij} = \left\{ \int_{\mathbf{X}} [\sqrt{p(\mathbf{X}|\omega_i)} - \sqrt{p(\mathbf{X}|\omega_j)}]^2 d\mathbf{X} \right\}^{1/2} \quad (3.6)$$

It is roughly speaking a measure of the average difference between the two class density functions [22,23].

In classification of remotely sensed data we assume most often that the classes have normal density functions, i.e.,

$$p(\mathbf{X}|\omega_i) = N(\mathbf{U}_i, \Sigma_i)$$

$$p(\mathbf{X}|\omega_j) = N(\mathbf{U}_j, \Sigma_j)$$

With this assumption (3.6) reduces to:

$$J_{ij} = [ 2 ( 1 - e^{-b_{ij}} ) ]^{1/2} \quad (3.7)$$

where  $b_{ij}$  is the Bhattacharyya distance:

$$\begin{aligned} b_{ij} = & \frac{1}{8}(\mathbf{U}_i - \mathbf{U}_j)^T \left( \frac{\Sigma_i + \Sigma_j}{2} \right)^{-1} (\mathbf{U}_i - \mathbf{U}_j) \\ & + \frac{1}{2} \log_e \left[ \frac{|\frac{\Sigma_i + \Sigma_j}{2}|}{|\Sigma_i|^{1/2} |\Sigma_j|^{1/2}} \right] \end{aligned} \quad (3.8)$$

And the average class separability is:

$$J_{ave} = \frac{1}{1 - \kappa} \sum_{i=1}^M \sum_{j=1}^M p(\omega_j) p(\omega_i) J_{ij} \quad (3.9)$$

where:

$$\kappa = \sum_{i=1}^M p(\omega_i)^2 \quad (3.10)$$

The  $J_{ave}$  has the saturating behavior and has a maximum value of  $\sqrt{2}$ . Therefore we can normalize  $J_{ave}$  to lie in the interval  $[0,1]$  by division by  $\sqrt{2}$ .

### 3.2.2 The Transformed Divergence

The divergence of two classes  $\omega_i$  and  $\omega_j$  is defined formally as:

$$D_{ij} = E[L_{ij}(X)|\omega_i] + E[L_{ji}(X)|\omega_j] \quad (3.11)$$

where  $L_{ij}(X)$  is the logarithmic-likelihood ratio:

$$L_{ij}(X) = \log_e p(X|\omega_i) - \log_e p(X|\omega_j) \quad (3.12)$$

If we assume as before that the class density is normal,  $D_{ij}$  reduces to:

$$D_{ij} = \frac{1}{2} \text{tr}[(\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1})] + \frac{1}{2} \text{tr}[(\Sigma_i^{-1} + \Sigma_j^{-1})(U_i - U_j)(U_i - U_j)^T] \quad (3.13)$$

$D_{ij}$  is not bounded as a function of normalized distance. i.e., it is monotonically increasing with increasing distance. To use the divergence we could specify some cutoff value and apply the approach described in section

3.1. However a saturating function of divergence, called transformed divergence, can also be used. This function is defined as:

$$D_{ij}^T = 2[1 - \exp(-\frac{D_{ij}}{8})] \quad (3.14)$$

The average separability using  $D_{ij}^T$  is:

$$D_{ave}^T = \frac{1}{1 - \kappa} \sum_{i=1}^M \sum_{j=1}^M p(\omega_i) p(\omega_j) D_{ij}^T \quad (3.15)$$

where  $\kappa$  is:

$$\kappa = \sum_{i=1}^M p(\omega_i)^2 \quad (3.16)$$

$D_{ave}^T$  has 2.0 as its maximum value. We can therefore normalize  $D_{ave}^T$  by 2.0 for use in our global membership function (3.2).

### 3.3 The Method

In the statistical multisource analysis, each source is first classified separately. When the reliability factor evaluation is added we use the classification accuracy or calculate the average separability for each source by any appropriate separability estimate. One thing which is important here is that we are discounting the sources by putting reliability factors on each source-specific posterior probability  $p$  in the global membership function. If we look at the family of curves  $p^a$  as a function of  $a$ , where  $a$  has value in  $[0,1]$  as shown in Fig. 3.1, we see that the functions are more discriminable as  $a$  increases. This leads us to the point that the separability estimates and the classification accuracy should only be used to



measure reliability. The source that has the "highest reliability" should be given the highest reliability factor and the others should be given reliability factors relative to this value. One way to accomplish this is to scale the values of the reliability measure as described below.

Assume we have  $n$  sources and we have calculated the reliability for each source  $i$  by some measure and its value is  $R_i$ . We give the source with the highest reliability the highest reliability factor  $a_{\max}$ . If the smallest possible reliability measure is  $\min$  we can calculate the reliability factors for the sources according to:

$$a_i = \frac{R_i - \min}{\max_{j=1,n}\{R_j - \min\}} a_{\max} \quad (3.17)$$

These values are then used as reliability factors in the global membership function (3.2). From there on we continue as described in section 2.2.

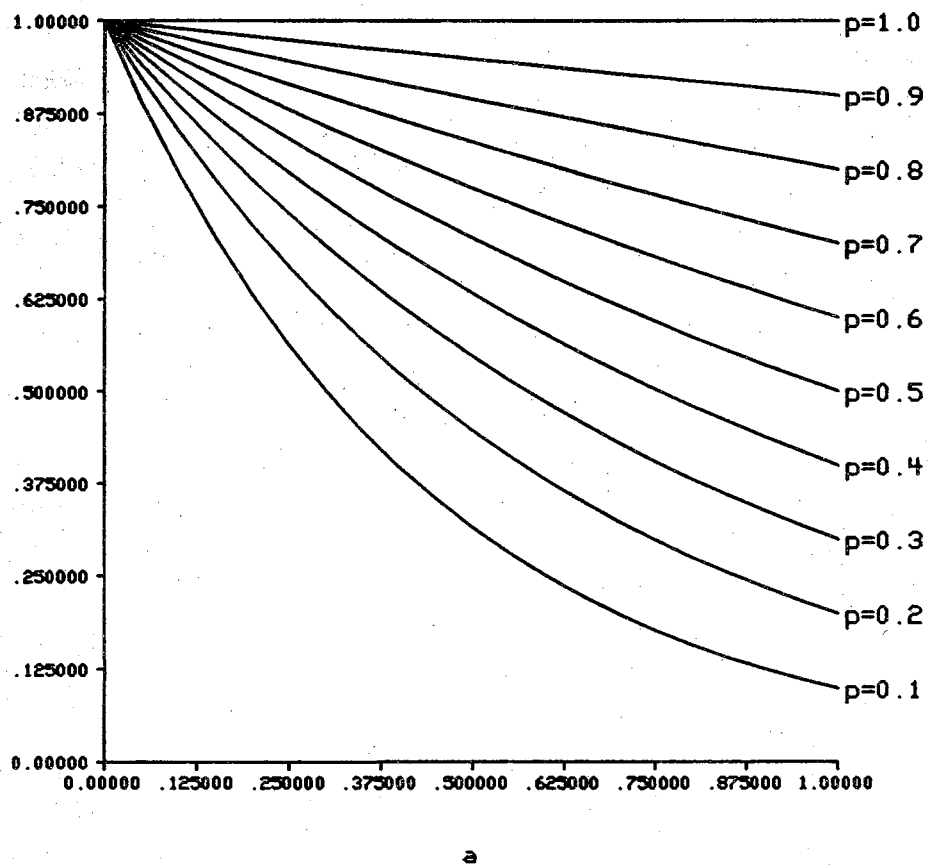


Figure 3.1 The Family of Curves  $p^a$

## CHAPTER 4

### EXPERIMENTAL RESULTS

#### 4.1 General Remarks

Our objective is to apply the statistical multisource analysis with varying levels of "reliability." To explore the method we would prefer a data set which contained several geometrically registered sources of data, e.g., Landsat Multispectral Scanner or Thematic Mapper data, aircraft multispectral scanner data, radar data, digital topographic data and a digital reference map for the particular area involved. Unfortunately we have not had a suitable data set of this kind available. Therefore to get preliminary results, the algorithm was applied to 12 channel aircraft multispectral scanner data, treating different regions of the electromagnetic spectrum (visible, near IR, ...) as different "sources." The data set chosen for experiment is a portion of flight line 210 from the 1971 Corn Blight Watch Experiment conducted by the Laboratory for Applications of Remote Sensing (LARS) at Purdue University, NASA and the U.S. Department of Agriculture. The portion of the data set used is 140 x 220 pixels and covers an agricultural area in Tippecanoe County, Indiana. A reference photograph and a ground cover reference map were available for this area.

The ground cover map was digitized and then geometrically registered to the multispectral scanner data.

From the 12 spectral bands three data "sources" were defined. The data set contained 7 visible bands; three of them were selected as the visible source (band 1: 0.46 - 0.49  $\mu\text{m}$ , band 4: 0.52 - 0.57  $\mu\text{m}$  and band 7: 0.61 - 0.70  $\mu\text{m}$ ). The data set has 3 bands in the near-infrared region (band 8: 0.72 - 0.92  $\mu\text{m}$ , band 9: 1.00 - 1.40  $\mu\text{m}$  and band 10: 1.50 - 1.80  $\mu\text{m}$ ) which were all selected to represent the near-infrared source. One band in the thermal region (band 12: 9.30 - 11.70  $\mu\text{m}$ ) was selected as the thermal source. It is known from a long history of experience with the data that the ground cover types have significantly different degrees of separability in these three spectral regions.

Two approaches were applied to determine reliability factors for the three sources. One used the weighted average separability of pairs of information classes in each source as a measure of reliability; the other measured the reliability by the overall classification accuracy in each source. Since the separabilities were calculated for the information classes as defined by the reference map, they do not depend on the signatures used for classification of a data source. Therefore, in our experiments, different training methods did not affect the values of the reliability factors determined from the weighted average separability of the information classes. The separability could thus be calculated before the individual sources were classified. In this research two types of separability estimates were used: JM - distance and transformed divergence. The values of these estimates for each data source are shown in Table 4.1. For the purpose of

comparison the values in the table are normalized to be in the range from 0 to 1.

As pointed out in Chapter 2 various training methods can be applied in statistical multisource analysis. In our experiments we used both unsupervised and supervised training. In the first experiment (unsupervised training) we used the data classes in each source; in the second experiment (supervised training) data classes were picked by selecting regions with distinctly different color on an image display. When the statistics for each source had been determined by applying the selected training procedure, each source was classified by maximum likelihood classification.

Table 4.1

Normalized Separability of Information Classes		
Source	JM - Distance	Transformed Divergence
Visible	0.7595	0.7461
Near-Infrared	0.8291	0.8166
Thermal	0.5715	0.4971

In order to apply equation (3.2), the source-specific probabilities were written in the following form:

$$p(\omega_j | x_i) = [p(x_i)]^{-1} \sum_{k=1}^{m_i} p(x_i | d_k, \omega_j) p(d_k, \omega_j) \quad (4.1)$$

Here  $m_i$  is the number of data classes for source  $i$  and  $p(x_i)$  is computed by:

$$p(x_i) = \sum_{j=1}^M \sum_{k=1}^{m_i} p(x_i | d_k, \omega_j) p(d_k, \omega_j) \quad (4.2)$$

where  $M$  is the number of information classes. For each source, the joint probabilities  $p(d_k, \omega_j)$  were tabulated in a joint occurrence matrix by comparing single-source data-class classifications to information classes in the reference map. To reduce considerably the computation and memory requirements, the class-conditional probabilities were computed independently of information classes, i.e., we set:

$$p(x_i | d_k, \omega_j) = p(x_i | d_k) \quad \text{for all } \omega_j$$

This approximation is valid if the distribution of a data class is the same regardless of information class. It is unlikely to hold exactly in the case of unsupervised classification, but the approximation is essential to the feasibility of carrying out the computations on a microcomputer (a PC/AT-based system was used). Using the approximation and equations (4.1) and (4.2), equation (3.2) can be written in the following form:

$$F_j(X) = [p(\omega_j)]^{1-n} \prod_{i=1}^n \left( \frac{\sum_{k=1}^{m_i} p(x_i | d_k) p(d_k, \omega_j)}{\sum_{j=1}^M \sum_{k=1}^{m_i} p(x_i | d_k) p(d_k, \omega_j)} \right) \exp a_i \quad (4.3)$$

All computer processing was done on an ERDAS image processing system based on an IBM PC/AT.

## 4.2 Experiment 1: Unsupervised Analysis

In this experiment the classifier training for each source was performed using an unsupervised approach. For this purpose a one-pass clustering algorithm called STATCL in the ERDAS software was used. This algorithm works as follows [24]:

A 3 x 3 window is moved over the multispectral image row by row and column by column. In each box the standard deviation of each band and the interband covariance matrix are calculated. The standard deviations are then compared to the user-specified upper and lower bounds on standard deviation in a cluster. If all of the standard deviations are within these bounds the covariances in the covariance matrix are compared to a fixed upper bound on covariance as specified by the user. If every covariance in the covariance matrix is less than this fixed covariance, the window becomes a cluster, otherwise not. In experiment 1 the default values in the algorithm were used, i.e., the lower bound on standard deviation was always set to be 0.1, the upper bound 1.2 and the upper bound on covariance was 12.

After the image has been scanned by the 3 x 3 window and all the clusters have been made they are merged according to a user-specified bound on the Mahalanobis distance. In the experiment this bound was always selected to be 3 (default). The output from the STATCL algorithm is the mean vector and the covariance matrix for each data class in the image.

When the STATCL algorithm had been run to define data classes for each source, all sources were classified independently by maximum likelihood classification. The clustering had identified 9 data classes in the visible source, 10 in the near-infrared source and 5 in the thermal source. The test

area contains 9 ground cover classes. The co-occurrence matrices showing the joint occurrences of the information and data classes for each source were computed by considering the whole test area. In practice we usually have just a small training area, which should be representative of the whole area, from which to calculate the joint occurrence matrix. At this point in testing the algorithm we want the joint occurrence matrices to be as accurate as possible and we therefore used the whole area.

In this experiment we combined two sources at a time. The separability of the information classes in the near-infrared source was the highest; therefore that source was combined first with the visible source and then with the thermal source. Since the near-infrared source had the highest separability according to both JM-distance and transformed divergence, its reliability factor determined from these separability measures was given the value 0.9. The reliability factors of the other sources were scaled relative to this value by using equation (3.17) and the values in Table 4.1. We selected 0.9 as the highest reliability factor ( $a_{\max}$ ) because the prior probabilities can be considered as a separate source in equation (3.2) with the reliability factor 1.0 (since the prior probabilities are computed from the reference map which is representative of the total area classified). The values of the reliability factors for both separability measures are shown in Table 4.2 and Table 4.3.

In order to get a baseline result and see how the values of the reliability factors affect the classification, the classification was also performed for a range of values of the reliability factor. While one source was given a constant reliability factor of 0.9 the reliability factor of the other source was



**Table 4.2**

**Reliability Factors Determined from the Separability Measures  
for Classification of the Near-Infrared and Visible Sources**

Source	JM - Distance	Transformed Divergence
Near-Infrared	<b>0.9000</b>	<b>0.9000</b>
Visible	<b>0.8244</b>	<b>0.8222</b>

**Table 4.3**

**Reliability Factors Determined from the Separability Measures  
for Classification of the Near-Infrared and Thermal Sources**

Source	JM - Distance	Transformed Divergence
Near-Infrared	<b>0.9000</b>	<b>0.9000</b>
Thermal	<b>0.6203</b>	<b>0.5478</b>

successively reduced from 0.9 in steps of 0.1. This was done for both sets of sources involved in the classification. The results are shown in Tables 4.5 and 4.6.

Table 4.5 shows the results of the classification of the visible and near-infrared sources. If we look at the individual classification of each data source we see that the clustering algorithm has isolated corn, soybeans, non-farm and pasture in both data sources. The near-infrared source does a much better job of classifying the soybeans but the visible source isolates additionally another information class which is sudex. The overall classification accuracy is slightly higher in the near-infrared source (78.7%) compared to the visible source (73.1%). These accuracies were used to calculate a set of reliability factors by applying equation (3.17). The reliability factors are shown in Table 4.4.

Table 4.4

Reliability Factors Determined from Overall Classification Accuracy for Classification of the Near-Infrared and Visible Sources in Experiment 1

Source	Classification Accuracy	Reliability Factor
Near-Infrared	78.7%	0.9000
Visible	73.1%	0.8360

Table 4.5

**Results of Experiment 1:  
Classification of the Near-Infrared and Visible Sources  
and Their Composite with Various Values of "Reliability"**

NIR VS	Percent Agreement with Reference for Class									OA
	1	2	3	4	5	6	7	8	9	
near-infrared	84.8	92.6	91.5	0.0	0.0	0.0	0.0	69.1	0.0	78.7
visible	81.4	88.2	73.4	0.0	0.0	0.0	0.0	49.0	86.1	73.1
100 100	89.2	94.1	90.0	0.0	0.0	3.6	0.0	45.8	82.9	82.6
90 90	90.1	94.0	89.8	0.0	0.0	19.0	0.0	48.2	83.7	82.8
90 83.6 (C)	89.9	94.0	89.6	0.0	1.2	19.9	0.3	49.5	83.3	82.8
90 82.4 (J)	89.9	94.0	89.6	0.0	1.2	19.9	0.4	50.0	83.0	82.8
90 82.2 (T)	89.9	93.9	89.6	0.0	1.6	19.9	0.4	50.0	83.0	82.8
90 80	89.9	93.9	89.6	0.0	2.1	20.2	0.4	51.0	82.8	82.7
90 70	89.8	93.9	89.2	0.0	3.8	22.0	1.5	57.9	81.2	82.7
90 60	89.5	93.7	89.0	0.0	4.0	22.0	3.2	63.9	80.7	82.6
90 50	88.6	93.5	88.4	0.0	6.9	22.6	9.8	65.2	78.3	82.4
80 90	90.6	94.0	89.5	0.0	0.0	23.2	0.2	48.4	83.8	82.8
70 90	91.2	93.6	89.4	0.0	0.0	44.3	0.3	48.2	84.6	83.0
60 90	92.1	93.3	88.0	0.0	2.6	47.0	0.3	47.9	84.7	82.4
50 90	92.9	92.7	86.3	16.3	2.3	57.1	1.1	47.9	84.8	82.0
# of pixels	2783	10543	12939	610	577	336	1167	382	1463	30800

NIR VS indicates the level of "reliability" assigned to the near-infrared (NIR) and the visible (VS) sources. (C) indicates weighting according to classification accuracy; (J) according to JM-distance; (T) according to transformed divergence.

Names of information classes:

- 1 - Non-farm
- 2 - Corn
- 3 - Soybeans
- 4 - Hay
- 5 - Oats
- 6 - Woods
- 7 - Wheat
- 8 - Pasture
- 9 - Sudex

When the sources are combined with full reliability (1.0) assigned to both of them we get a significant increase in overall classification accuracy compared to the classification of the individual sources. Assigning the reliability factors shown in Table 4.2 and Table 4.4 does not increase the overall accuracy very much. All these computed reliability factors give very similar results, an overall accuracy of 82.8%. This is not the highest overall accuracy in Table 4.5, however. The highest accuracy is, somewhat surprisingly, accomplished by giving the near-infrared source a lower value of reliability than the visible source (70,90). This result is surprising because we estimated the near-infrared source to be more reliable than the visible source.

The increase in overall accuracy using different levels of reliability is so small that it is hard to draw conclusions from these results. But the main reason for the small increase in overall accuracy is that we do not get much increase in accuracy contribution from the small classes. In the area there are two dominating information classes, corn and soybeans, covering 76.2% of the area. To get a substantial increase in overall accuracy by changing the levels of reliability we have to get high accuracy for these classes and also some increase in accuracy for the smaller classes. When we get the highest accuracy (83.0%) we accomplish this but the difference in accuracy contribution from the smaller classes other than sudex is very small.

However, we can see that changes in the reliability factors significantly affect the classification accuracy of the individual information classes. For example the classification accuracy of pasture increases substantially when the value of the reliability factor for the visible source is decreased. Similar

things happen for woods and hay when the reliability factor for the near-infrared source is decreased. This leads us to conclude that it is possible to optimize the classification accuracy of single information classes by adjusting the reliability factors. One possible way to determine the reliability factors in this case would be to base them on the weighted average separability of a single information class versus all other information classes in each source.

Another point which is interesting to note is how well information classes are discriminated by a source. The "strength of discrimination" of information classes is a possible reason why we get the peak in overall accuracy when we discount the near-infrared source. Although classification accuracy for corn and soybeans is higher in the near-infrared source, the classification accuracy of these classes decreases only slightly when the near-infrared source is discounted. We can therefore assume that these classes are very well discriminated by the near-infrared source. We discuss this further below when we look at the results in Table 4.6 where we have combined the near-infrared and the thermal sources.

In Table 4.6 we see that the clustering of the thermal source does not isolate one of the large classes (corn) but does isolate wheat which is not isolated by the near-infrared source. Since corn is never classified correctly by the thermal source alone, the overall classification accuracy for the source is only 49.2%. The reliability factors calculated from the overall classification accuracy of the sources are shown in Table 4.7.

When the sources are combined with full reliability (1.0) assigned to both, we get a substantial increase in overall accuracy compared to the overall accuracy of the classification of the thermal source but no increase

Table 4.6

**Results of Experiment 1:  
Classification of the Near-Infrared and Thermal Sources  
and Their Composite with Various Values of "Reliability"**

NIR TH	Percent Agreement with Reference for Class									OA
	1	2	3	4	5	6	7	8	9	
near-infrared	84.8	92.8	91.5	0.0	0.0	0.0	0.0	69.1	0.0	78.7
thermal	58.1	0.0	97.5	0.0	0.0	0.0	77.6	0.0	0.0	49.2
100 100	81.7	93.4	88.7	0.0	0.0	0.0	40.4	47.4	0.0	78.7
90 90	79.9	93.0	88.6	0.0	0.3	0.0	52.6	61.5	0.0	79.0
90 80	79.0	92.8	88.6	0.0	0.3	0.0	55.0	61.8	0.0	79.0
90 70	78.0	92.7	88.5	0.5	0.0	0.0	56.3	63.6	0.2	78.9
90 62.0 (J)	77.8	92.7	88.4	0.7	0.0	0.0	56.6	68.6	0.3	78.9
90 60	77.8	92.7	88.4	0.7	0.0	0.0	56.6	68.6	0.3	78.9
90 56.3 (C)	76.8	92.7	88.4	0.7	0.0	0.0	57.4	69.1	1.8	78.9
90 54.8 (T)	76.8	92.7	88.3	0.7	0.0	0.0	57.9	69.4	1.8	78.9
90 52	76.6	92.7	88.3	0.7	0.0	0.0	59.0	69.9	2.0	78.9
90 50	76.6	92.7	88.2	1.3	0.0	0.0	59.0	70.9	12.4	79.4
90 40	73.9	92.7	69.7	1.8	0.0	0.0	59.0	78.8	55.7	73.6
80 90	77.7	92.7	88.6	0.0	0.9	0.0	60.0	61.5	0.0	79.0
70 90	76.2	92.3	88.5	0.0	3.6	0.0	69.1	61.1	0.2	79.1
60 90	74.1	92.1	88.2	0.0	7.8	0.0	74.2	61.8	0.3	79.0
50 90	70.5	91.5	88.2	0.0	8.0	0.0	79.9	63.6	0.0	78.7
40 90	64.4	90.4	87.6	0.0	11.8	0.0	88.4	67.3	0.0	78.0
# of pixels	2783	10543	12939	610	577	336	1167	382	1463	30800

NIR TH indicates the level of "reliability" assigned to the near-infrared (NIR) and the thermal (TH) sources. (C) indicates weighting according to classification accuracy; (J) according to JM-distance; (T) according to transformed divergence.

Names of information classes:

- 1 - Non-farm
- 2 - Corn
- 3 - Soybeans
- 4 - Hay
- 5 - Oats
- 6 - Woods
- 7 - Wheat
- 8 - Pasture
- 9 - Sudex

compared to the overall accuracy of the classification of the near-infrared source. When the reliability factors are assigned we get the overall accuracy as high as 79.4%. This increase in overall accuracy is caused by an increase in the accuracy of source is discounted while the classification accuracy of corn and soybeans does not decrease by much. The reliability factors in Table 4.3 and Table 4.7 all give an overall accuracy of 78.9%. These reliability factors apparently do not discount the thermal source enough.

Table 4.7

Reliability Factors Determined from Overall Classification Accuracy for Classification of the Near-Infrared and Thermal Sources in Experiment 1

Source	Classification Accuracy	Reliability Factor
Near-Infrared	78.7%	0.9000
Thermal	49.2%	0.5626

Looking at the results in Table 4.6 there are still other things which are interesting. For example when we decrease the reliability of the near-infrared source in which the information classes are much more separable than in the thermal source, the overall accuracy goes up to the high of 79.1%. The accuracy of the large classes corn and soybeans goes down just a bit. This is interesting because the clustering of the thermal source does not isolate corn. Therefore we can conclude that soybeans are so well

discriminated by the near-infrared source that we can reduce the reliability factor to as little as 0.4 without affecting the accuracy of the classification by much. We can generalize this by saying that if information classes are well discriminated by a source, their classification accuracy will be relatively independent of the value of the reliability factor specified for the source. The reliability factor can then be specified to maximize the classification accuracy of other information classes.

It is also interesting to note in Table 4.6 that the classification accuracy of sudex increases significantly as we decrease the value of the reliability factor of the thermal source. This is interesting because sudex is not isolated by the clustering in either source. The experimental results indicate though that the near-infrared source gives some support to this information class.

Since we did not get much improvement in the classification accuracy in this experiment by using our reliability measures, we wanted to do another experiment differently on the same data set. In this experiment some information classes were not isolated by the clustering and a high overall classification accuracy was not accomplished. These results indicated that the signatures used were not representative and we consequently questioned the training of the data sources. We therefore chose to train the data sources differently. Since a supervised approach is likely to overcome the shortcomings described above, a supervised approach was defined to train the data sources.



### 4.3 Experiment 2: Supervised Analysis

In this experiment we trained each source using a supervised approach. For each source, data classes were picked by selecting regions with distinctly different color on a color monitor. The training samples were classified, a confusion matrix and the JM - distance were calculated and "non-separable" training samples were merged as shown in Fig. 4.1. This procedure identified 22 data classes in the visible source, 24 classes in the near-infrared source but only 5 in the thermal source. A few of the information classes were not isolated by this training approach because they were not separable from the other information classes. This was especially the case for the smaller information classes (woods, oats and hay). Apart from the training the experiment was conducted in the same manner as Experiment 1. The reliability factors calculated from classification accuracies are shown in Tables 4.8 and 4.11. The experimental results are shown in Tables 4.9 and 4.10.

Table 4.8

Reliability Factors Determined from Overall Classification Accuracy for Classification of the Near-Infrared and Visible Sources in Experiment 2

Source	Classification Accuracy	Reliability Factor
Near-Infrared	79.3%	0.9000
Visible	76.7%	0.8705

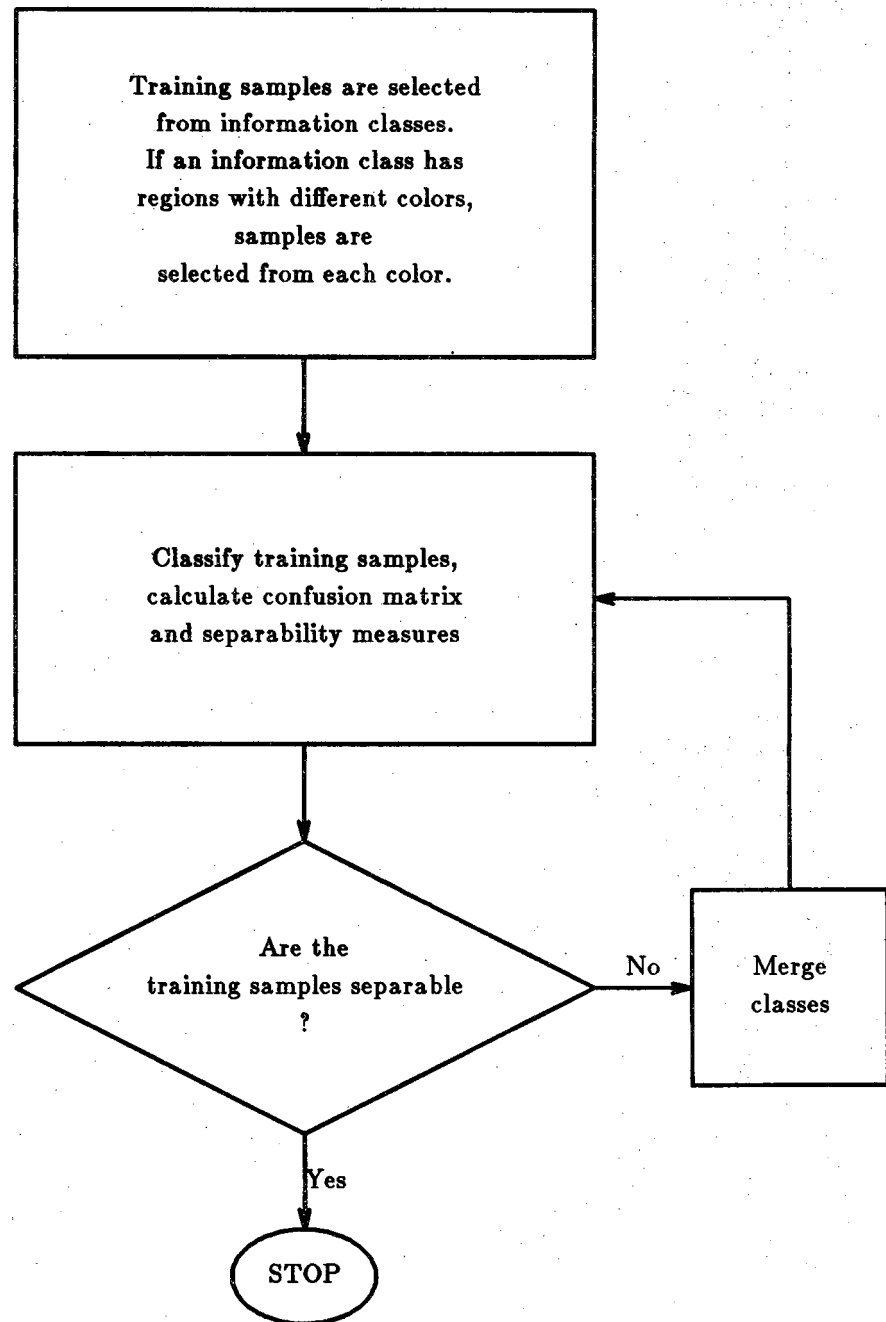


Figure 4.1 The Supervised Training Procedure

In Table 4.9 we see the classification results for the combination of the near-infrared source and the visible source. In the near-infrared source 6 information classes are isolated and the overall classification accuracy for this source is 79.3%. The classification of most of these classes is more accurate in the near-infrared source than in the visible source but 2 more information classes are isolated in the visible source and the overall classification accuracy for the visible source is 76.7%.

When the sources are combined the overall accuracy goes up to 87.7%, which is a significant increase. The accuracy in all classes but three goes up compared to the classification accuracy in the individual sources. We get, for instance, over 90% classification accuracy for the three largest classes; soybeans, corn and non-farm. The increase in classification accuracy for non-farm is 29.9% compared to the classification accuracy of the near-infrared source and 43.0% compared to the classification accuracy of the visible source. We do not get higher accuracy after combination for oats in the visible source and wheat and pasture in the thermal source. However, in all those cases the classification accuracy is increased by the combination as compared to the classification accuracy of the other source.

When reliability factors are assigned we get a further increase in overall accuracy. Using the reliability factors in Table 4.2 and Table 4.8 we get the highest overall accuracy which is 88.1%. Varying the reliability factors has for most of the information classes the expected effect that when we discount the visible source the classification accuracy goes up for the classes which have higher accuracy in the near-infrared source. In particular we see that the classification accuracies of pasture and wheat increase compared to

Table 4.9

**Results of Experiment 2:  
Classification of the Near-Infrared and Visible Sources  
and Their Composite with Various Values of "Reliability"**

NIR VS	Percent Agreement with Reference for Class									OA
	1	2	3	4	5	6	7	8	9	
near-infrared	61.6	86.4	87.2	0.0	0.0	0.0	79.5	97.6	69.4	79.3
visible	48.5	81.8	86.6	6.2	74.5	0.0	48.2	81.7	76.2	76.7
100 100	91.5	91.8	92.5	17.2	38.5	5.4	75.5	93.7	84.3	87.7
90 90	91.5	91.2	91.9	28.4	43.8	19.3	77.5	95.8	84.6	87.8
90 87.1 (C)	91.9	91.0	91.9	28.7	44.0	38.4	78.3	97.6	84.6	88.1
90 82.4 (J)	92.0	91.0	91.6	29.2	43.3	43.2	78.3	99.5	84.7	88.1
90 82.2 (T)	92.1	91.0	91.6	29.2	43.5	43.2	78.8	99.5	84.7	88.1
90 81	92.3	91.0	91.5	29.3	43.0	43.5	79.4	99.7	84.8	88.1
90 80	92.4	91.0	91.5	29.5	42.8	43.8	79.5	99.7	84.8	88.1
90 78	92.7	91.0	91.4	29.8	43.0	43.8	79.7	99.7	84.8	88.1
90 70	92.2	90.3	90.7	31.1	42.5	43.5	80.3	99.7	84.6	87.5
90 60	91.4	88.9	89.5	32.1	41.1	46.7	79.3	99.7	84.5	86.5
90 50	90.4	87.5	88.1	33.4	40.7	47.0	78.1	99.7	84.5	85.3
80 90	90.6	90.5	91.2	36.6	50.4	48.2	77.0	97.1	84.6	87.8
70 90	87.0	89.6	90.2	44.4	55.8	53.6	72.8	96.9	84.2	86.9
60 90	82.9	88.2	88.9	49.3	61.5	56.5	68.0	95.8	83.3	85.5
50 90	79.6	86.7	87.6	54.3	63.3	59.2	63.2	95.0	82.2	84.0
# of pixels	2783	10543	12939	610	577	336	1167	382	1463	30800

NIR VS indicates the level of "reliability" assigned to the near-infrared (NIR) and the visible (VS) sources. (C) indicates the according to classification accuracy; (J) according to JM-distance; (T) according to transformed divergence.

Names of information classes:

- 1 - Non-farm
- 2 - Corn
- 3 - Soybeans
- 4 - Hay
- 5 - Oats
- 6 - Woods
- 7 - Wheat
- 8 - Pasture
- 9 - Sudex

the accuracy in classification of either source. This is also true for oats, i.e., when we discount the near-infrared source the classification accuracy of oats goes up.

It is also interesting to note that although woods is isolated by neither source in single source classification, its classification accuracy is much better than chance when the sources are combined and the accuracy increases when either of the two sources is discounted. This is especially true when the near-infrared source is discounted; as shown in Table 4.9, the classification accuracy of woods increases to over 55%. Another interesting observation is that the classification accuracy of hay goes up when we discount the visible source even though this class is isolated in the visible source but not in the near-infrared source. This shows that the near-infrared source gives some support to this class although it is not isolated in the source. This also demonstrates the strength of discrimination of hay by the visible source. Furthermore, the classification accuracy of hay increases still more when the near-infrared source is discounted. These two examples of changes in classification accuracy for hay and woods suggest the possibility of defining class-specific reliability factors to optimize classification of specific ground cover types. Similar effects are seen when we combine the near-infrared source and the thermal source, which we discuss below.

In Table 4.10 we have combined the near-infrared source and the thermal source. The thermal source has lower accuracy in classification for most of the information classes and two fewer classes are isolated than for the near-infrared source. The overall classification accuracy (67.7%) is

Table 4.10

**Results of Experiment 2:  
Classification of the Near-Infrared and Thermal Sources  
and Their Composite with Various Values of "Reliability"**

NIR TH	Percent Agreement with Reference for Class									OA
	1	2	3	4	5	6	7	8	9	
near-infrared	61.6	86.4	87.2	0.0	0.0	0.0	79.5	97.6	69.4	79.3
thermal	76.5	79.3	73.6	0.0	0.0	0.0	71.5	0.0	0.0	67.7
100 100	71.8	90.2	92.7	8.4	34.5	0.0	77.3	95.6	76.1	84.8
90 90	71.6	89.7	92.4	16.7	35.7	0.0	78.2	95.8	77.6	84.9
90 80	72.0	89.7	92.4	17.5	36.0	0.3	78.5	95.8	78.2	84.9
90 76.8 (C)	73.8	89.7	92.3	17.7	36.0	0.4	80.0	95.8	78.3	85.1
90 70	75.5	89.6	92.0	18.0	36.2	0.6	81.4	95.8	78.5	85.2
90 62.0 (J)	76.2	89.6	91.9	18.7	36.2	0.9	79.9	96.1	78.8	85.2
90 60	76.5	89.4	91.7	19.0	36.2	1.2	79.8	96.1	78.8	85.1
90 57	77.9	89.1	91.0	19.3	36.4	1.2	78.7	96.1	79.8	84.8
90 54.8 (T)	78.3	88.9	90.6	19.3	36.6	1.2	78.1	96.1	80.5	84.6
90 50	78.9	88.3	88.8	19.7	37.3	0.9	78.1	96.1	80.8	83.8
90 43	79.9	86.8	85.1	20.2	38.5	0.9	78.1	96.3	81.2	81.8
80 90	71.5	89.5	89.8	18.2	37.4	0.0	74.0	95.8	78.3	83.6
70 90	68.5	89.1	88.4	19.3	38.0	0.0	73.8	95.6	78.6	82.6
60 90	67.5	88.5	86.6	20.0	38.3	0.0	73.6	95.6	80.7	81.7
50 90	64.0	87.4	85.1	20.2	38.3	0.0	74.0	87.7	80.8	80.3
# of pixels	2783	10543	12939	610	577	336	1167	382	1463	30800

NIR TH indicates the level of "reliability" assigned to the near-infrared (NIR) and the thermal (TH) sources. (C) indicates weighting according to classification accuracy; (J) according to JM-distance; (T) according to transformed divergence.

Names of information classes:

- 1 - Non-farm
- 2 - Corn
- 3 - Soybeans
- 4 - Hay
- 5 - Oats
- 6 - Woods
- 7 - Wheat
- 8 - Pasture
- 9 - Sudex

much higher using the supervised approach than in the classification of the thermal source in experiment 1 (49.2%) because corn was not isolated by the clustering there. The reliability factors calculated from the overall classification accuracies of the near-infrared and thermal sources are shown in Table 4.11.

Table 4.11

Reliability Factors Determined from Overall Classification Accuracy for Classification of the Near-Infrared and Thermal Sources in Experiment 2

Source	Classification Accuracy	Reliability Factor
Near-Infrared	79.3%	0.9000
Thermal	67.7%	0.7683

When the sources are combined the overall accuracy goes up substantially. As in Table 4.9 there is an increase in accuracy for most of the information classes. When reliability factors are included in the global membership function the overall accuracy goes up to as much as 85.2%. Using the reliability factors from Table 4.3 we get this maximum with the reliability factors calculated from the JM - distance. The reliability factors calculated from the transformed divergence give only 84.8% overall accuracy, still quite close to the maximum. The reliability factors in Table 4.11 give 85.1% overall accuracy. The trend in classification accuracy in

Table 4.10 is similar to the trend in Table 4.9, i.e., when we discount the "more reliable" source the overall accuracy goes down and when we discount the "less reliable" source to a certain point the overall accuracy goes up.

The most significant increase in accuracy is for hay and oats which are not isolated by either source but, after the combination and changes in reliability factors, the accuracy in the classification of these classes increases to over 20% and 38%, respectively.

#### **4.4 General Observations**

Combination of data from various data sources using statistical multisource analysis provides in most of our experiments a significant increase in overall classification accuracy as compared to single-source analysis. Combining the near-infrared source and the visible source gives, for instance, 88.1% overall classification accuracy in experiment 2 when certain reliability factors are assigned to the sources. There were two approximations made in our experiments which could have introduced some error. First, we ignored dependence between data sources in the global membership function. The advantages of this approach are that it reduces the computational complexity of the classification procedure and provides the opportunity to update the classification based on additional sources without starting all over again. Secondly, we made the approximation that the distribution of the data in a data class is the same regardless of information class. This approximation is unlikely to hold exactly for the unsupervised case but it, too, reduces the complexity of the computations and memory requirements.



The results of the classification in experiment 2 are better than in experiment 1, consistent with the superiority of the supervised training over unsupervised training. Although there is not a large increase in overall accuracy achieved by assigning reliability factors in either experiment, the different levels of reliability often give a substantial increase in classification accuracy of individual classes, even for classes which are not isolated in the classification based on any of the individual sources. In our view, this justifies in part the use of reliability factors in equation (3.2) for the purpose of weighting the influence of the various sources.

Using separability analysis to estimate the reliability of a source seems to be a reasonable choice, especially when the assumption can be made that the information classes have normal distributions. In experiment 2 we had some success assigning reliability factors using the separability measures to achieve the highest overall accuracy. In experiment 1 we did not get the highest overall accuracy by applying this approach but that may be due to the STATCL algorithm and the possibility it did not provide representative statistics. But this also illustrates a shortcoming in this approach: we have to assume a particular distribution for the information classes in order to be able to calculate the separability. In these experiments we believe the Gaussian model was reasonable, but when handling different kinds of data the Gaussian assumption may be unsuitable for some of the sources.

On the other hand, using classification accuracy to measure the reliability of a source is a straightforward approach which is computationally inexpensive and overcomes some of the shortcomings of the separability approach. The reliability factors calculated from the

classification accuracy depend on the training of the data sources in contrast to the separability approach applied in this report. This might be an advantage of the classification approach, because if a source is badly trained it is likely to have lower reliability. In our experiments the results using the reliability factors calculated from the classification accuracy were very similar to the ones using the separability measures.

The main problem is how to associate reliability factors with the reliability measures. In this research we have assigned the highest reliability factor to the "most reliable" source, assumed a linear relationship between the reliability of the different sources and scaled them relative to the maximum value. This linearization is almost certainly a simplification of reality and consequently introduces errors in the reliability factor calculations in some cases. In the next chapter we will discuss this problem in conjunction with other ways of estimating the reliability of sources.

## CHAPTER 5

### CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

#### 5.1 Discussion

The objective of this research is to investigate methods of statistical multisource analysis. The proposed method has several advantages as a general approach in multisource classification, viz., it handles various sources of data independently, has the potential to treat non-numerical as well as numerical data and, with certain approximations, provides a way to update the classification based on new data sources without having to calculate everything all over again. We have investigated ways to estimate the reliability of individual sources and to include reliability in the global membership function of the statistical multisource analysis. The experimental results show that assigning reliability factors to the sources can either improve or degrade the overall classification accuracy. In our experiments, assigning reliability factors did not increase the overall accuracy very much. It was clear, however, that different levels of reliability can affect individual classes significantly, and we demonstrated the possibility of assigning reliability factors to optimize accuracy of individual

classes. This was especially interesting when, for instance, an information class was isolated by neither individual source. In that case it was possible to achieve a significant accuracy for this class by varying the reliability factors.

The problem of determining optimal reliability factors can be split into two parts. First we have to use some measure to assess the reliability of a source, and then we have to associate this measure with the reliability factors. In this report, two methods were proposed to determine reliability factors. One used the weighted average separability of the information classes for a source as its measure of reliability; the other used the overall classification accuracy for a source. Two separability measures were considered to explore the separability approach, the transformed divergence and the JM - distance. The separability measures and the classification accuracies were associated with the reliability factors by assigning the highest reliability factor to the source with the "highest reliability" and then scaling the measured reliability of the other sources according to this value by using equation (3.17). Applying the calculated reliability factors in the statistical multisource analysis gave the highest overall accuracy in experiment 2 (the reliability factors calculated from the JM - distance) but the results were not as good in experiment 1. The change in overall accuracy using the reliability factors was so small that it was hard to draw firm conclusions from the results. It is clear, however, that the linearity relation in equation (3.17) has some limitations. We know, for instance, that the separability functions are not linear and we have some difficulty in justifying this linearity relation for the classification accuracy.

Using the separability estimates to measure reliability has the disadvantage that we have to assume some probability distribution for the information classes. Although normal distributions can be assumed for spectral classes of corn and soybeans, we would not be able to assume such a probability distribution for elevation data. It may not in all cases be possible to calculate the separability measures even though they can be expressed in a nice closed form when normal distribution is assumed. Thus separability measures will not be suitable to estimate reliability factors in all cases.

Using the classification accuracy to measure reliability does not require any knowledge of the probability distribution of a source. This approach is computationally relatively inexpensive because each data source needs to be classified individually anyway in the statistical multisource analysis. We discuss below another method which could be investigated for reliability factor estimation. This method also does not assume anything about the probability distribution of information or data classes.

## 5.2 Directions for Further Research

One way to characterize reliability of a source would be to examine the correspondence between the information classes and the data classes, i.e., the conditional probabilities that we observe a specific information class given a data class. All these conditional probabilities can be computed by comparing the reference map to a classified map from a data source.

Assuming we have  $r$  information classes  $\{x_1, \dots, x_r\}$  and  $s$  data classes  $\{y_1, \dots, y_s\}$  we can write all the conditional probabilities as the  $s \times r$

correspondence matrix  $R$ , where  $R$  is:

$$R = \begin{bmatrix} p(x_1|y_1) & p(x_2|y_1) & \dots & p(x_r|y_1) \\ p(x_1|y_2) & p(x_2|y_2) & \dots & p(x_r|y_2) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_1|y_s) & p(x_2|y_s) & \dots & p(x_r|y_s) \end{bmatrix} \quad (5.1)$$

We can now define reliability in the following way: If a source is optimal in reliability there would be a specific information class corresponding to each data class. Therefore ideally one conditional probability in each row of  $R$  would be 1 and all the others would be zero. If a source were very unreliable, there would be no correspondence between the data classes and the information classes; in the worst case all the numbers in the matrix would be the same.

Now we would like to associate a number with the matrix  $R$  to characterize the reliability. Using information theoretic measures [25] we could think of the information classes as a transmitted signal and the data classes as a received signal which must be used to estimate the transmitted signal. Using this approach we can state that there is an uncertainty of  $\log[1/p(x_i|y_j)]$  about the information class  $x_i$  when we observe data class  $y_j$  in a data source.

We can calculate the average loss of information when we observe the data class  $y_j$ , which is given by [26,27]:

$$H(x|y_j) = \sum_i p(x_i|y_j) \log \frac{1}{p(x_i|y_j)} \quad (5.2)$$

Now we want to average the information loss over all observed data classes

$y_j$ . This is called the equivocation of  $x$  with respect to  $y$  and is denoted by  $H(x|y)$ :

$$\begin{aligned}
 H(x|y) &= \sum_j p(y_j) H(x|y_j) \\
 &= \sum_i \sum_j p(y_j) p(x_i|y_j) \left\{ \log \frac{1}{p(x_i|y_j)} \right\} \\
 &= \sum_i \sum_j p(x_i, y_j) \left\{ \log \frac{1}{p(x_i|y_j)} \right\} \quad (5.3)
 \end{aligned}$$

$H(x|y)$  represents the average uncertainty about an information class over all the data classes. Evidently,  $H(x|y)$  is the average loss of information per data class and therefore seems to be a reasonable term to associate with the reliability of a source. Since  $H(x|y)$  measures uncertainty, the higher value it has the more unreliable a source is. If we estimate this quantity for all the data sources, we could give the source with the lowest  $H(x|y)$  the highest reliability factor and then determine the reliability factors for the other sources accordingly.

To calculate  $H(x|y)$  is relatively inexpensive because all the probabilities needed can be computed easily from the reference map and the classified maps from the individual sources. This reliability measure also has the advantage that we do not need to know anything about the probability distributions of the information classes in any source. The only problem at this point is how to associate reliability factors with the uncertainty, a problem common to all the reliability measures discussed so far.

The global membership function which we are trying to optimize is a complex non-linear function. To include reliability factors in that function

is by no means easy, but several different approaches have been discussed to quantify the reliability. To associate the reliability factors with these measures is a complicated problem. We would prefer a linear relationship between the reliability measures and the reliability factors or at least have the relationship as a closed expression. In this research we used separability measures and classification accuracy to estimate the reliability and approximated the relation between these measures and the reliability factors by a linear function. It is hard to justify this approximation. Consequently this problem should be investigated further.



## **LIST OF REFERENCES**

## LIST OF REFERENCES

- [1] P.H. Swain, J.A. Richards and T. Lee, "Multisource Data Analysis in Remote Sensing and Geographic Information Processing," *Proceedings of the 11th International Symposium on Machine Processing of Remotely Sensed Data 1985*, West Lafayette, Indiana, pp. 211-217, June 1985.
- [2] C.F. Hutchinson, "Techniques for Combining Landsat and Ancillary Data for Digital Classification Improvement," *Photogrammetric Engineering and Remote Sensing*, vol. 48, no. 1, pp. 123-130, 1982.
- [3] J.A. Richards, D.A. Landgrebe and P.H. Swain, "A Means for Utilizing Ancillary Information in Multispectral Classification," *Remote Sensing of Environment*, vol. 12, pp. 463-477, 1982.
- [4] A.H. Strahler and N.A. Bryant, "Improving Forest Cover Classification Accuracy from Landsat by Incorporating Topographic Information," *Proceedings Twelfth International Symposium on Remote Sensing of the Environment*, Environmental Institute of Michigan, pp. 927-942, April 1978.
- [5] R.M. Hoffer and staff, "Computer-Aided Analysis of Skylab Multispectral Scanner Data in Mountainous Terrain for Land Use, Forestry, Water Resources and Geological Applications," *LARS Information Note 121275*, Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, IN 47907, 1975.
- [6] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, New Jersey, 1976.
- [7] M. Goldberg, D.G. Goodenough, M. Alvo and G.M. Karam, "A Hierarchical Expert System for Updating Forestry Maps with Landsat Data," *Proceedings of the IEEE*, vol. 73, no. 6, pp. 1054-1063, 1985.

- [8] T.D. Garvey, J.D. Lowrance and M.A. Fischler, "An Inference Technique for Integrating Knowledge from Disparate Sources," *7th International Joint Conference on Artificial Intelligence*, Vancouver, British Columbia, Canada, pp. 319-325, August 1981.
- [9] J.A. Richards, T. Lee and P.H. Swain, "Probabilistic and Evidential Approaches to the Multisource Data Analysis Problem in Remote Sensing and Geographic Information Systems," presented at Workshop on Analytical Methods in Remote Sensing for Geographic Information Systems, International Association for Pattern Recognition, Paris, France, October 1986.
- [10] H. Kim, P.H. Swain and J.A. Benediktsson, "A Method for Combining Multispectral and Ancillary Data in Remote Sensing and Geographic Information Processing," *Proceedings of the 1986 IEEE International Conference on Systems, Man and Cybernetics*, Atlanta, Georgia, pp. 1543-1547, October 1986.
- [11] L.A. Zadeh, "Fuzzy Sets," *Information and Control*, vol. 8, no. 3, pp. 338-353, 1965.
- [12] D. Dubois and H. Prade, *Fuzzy Sets and Systems - Theory and Applications*, Academic Press, New York 1980.
- [13] M. Ishizuka, "Inference Methods Based on Extended Dempster & Shafer's Theory for Problems with Uncertainty/Fuzzyness," *New Generation Computers*, vol. 1, pp. 159-168, 1983.
- [14] M. Ishizuka, K.S. Fu and J.T.P. Yao, "Inference Procedures Under Uncertainty for the Problem Reduction Method," *Information Sciences*, vol. 28, no. 3, pp. 179-206, 1982.
- [15] L.W. Fung and K.S. Fu, "An Axiomatic Approach to Rational Decision Making in a Fuzzy Environment," *Fuzzy Sets and Their Applications to Cognitive and Decision Processes*, edited by L.A. Zadeh and K.S. Fu et al., Academic Press, New York 1980.
- [16] L.A. Zadeh, "The Concept of Linguistic Variable and its Application to Approximate Reasoning," part III, *Information Sciences*, vol. 9, no. 1, pp. 43-80, 1975.
- [17] J.A. Barnett, "Computational Methods for a Mathematical Theory of Evidence," *Proceedings 7th International Joint Conference on Artificial Intelligence*, Vancouver, British Columbia, Canada, August 1981.

- [18] T.D. Garvey, "Evidential Reasoning for Land-Use Classification," presented at Workshop on Analytical Methods in Remote Sensing for Geographic Information Systems, International Association for Pattern Recognition, Paris, France, October 1986.
- [19] R.L. Cannon, J.V. Dave, J.C. Bezdek and M.M. Trivedi, "Segmentation of a Thematic Mapper Image Using the Fuzzy c-Means Clustering Algorithm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-24, no. 3, 1986.
- [20] M.M. Trivedi and J.C. Bezdek, "Low-Level Segmentation of Aerial Images with Fuzzy Clustering," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-16, no. 4, 1986.
- [21] J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer-Verlag, New York 1985.
- [22] P.H. Swain, "Fundamentals of Pattern Recognition in Remote Sensing," in *Remote Sensing - The Quantitative Approach*, edited by P.H. Swain and S. Davis, McGraw-Hill Book Company, New York 1978.
- [23] J.A. Richards, *Remote Sensing Digital Image Analysis - An Introduction*, Springer-Verlag, Berlin, W. Germany 1986.
- [24] Private communication with J. Dooley of ERDAS Inc.
- [25] C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Chicago 1963.
- [26] R.J. McEliece, *The Theory of Information and Coding*, Encyclopedia of Mathematics and its Applications, vol. 3, Addison-Wesley Publishing Company, Reading, Massachusetts 1977.
- [27] A. Rosenfeld and A.C. Kak, *Digital Picture Processing*, 2nd edition, vol. 1, Academic Press, New York 1982.