The Dissertation Committee for Joseph Troy Morgan Certifies that this is the approved version of the following dissertation:

ADAPTIVE HIERARCHICAL CLASSIFICATION WITH LIMITED TRAINING DATA

Committee:

Melba M. Crawford, Supervisor

J. Wesley Barnes

Joydeep Ghosh

John J Hasenbein

Elmira Popova

ADAPTIVE HIERARCHICAL CLASSIFICATION WITH LIMITED TRAINING DATA

by

Joseph Troy Morgan, B.S., M.S.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May, 2002

Dedication

Laura and the boys.

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Melba M. Crawford. Without her support and drive, I would not have been able to accomplish this dissertation. Secondly, I am in deep gratitude to Colonel Michael A. Schiefer, who has been a mentor to my Air Force career and has given me advice on how to deal with demanding advisors. I would also like to thank Alexandre Henneguelle, with whom I worked closely, and the entire faculty in the OR/IE department, where I've never found a door closed to my questions. To the friends I've made at the Center for Space Research, I wish I was continuing on. Lastly, my wife, Laura Jeffords Morgan, has had the more difficult and rewarding job of being a "single" mother to our two beautiful children, Joseph Paul and Nicholas Alexander. Thank you for understanding that I left early and came home late but still found a way to show me that you were very much looking forward to a non "Ph.D. student" life.

ADAPTIVE HIERARCHICAL CLASSIFICATION WITH LIMITED TRAINING DATA

Publication No._____

Joseph Troy Morgan, Ph.D.

The University of Texas at Austin, 2002

Supervisor: Melba M. Crawford

This research focused on the development of a hierarchical approach for classification that is robust with respect to training data that are limited both in quantity and spatial extent. Many difficult classification problems involve a high dimensional input and output space (candidate labels). Due to the "curse of dimensionality", it is necessary to reduce the size of the input space when there is only a limited quantity of training data available. While a significant amount of research has focused on transforming the input space into a reduced feature space that accurately discriminates between the classes in a fixed output space, traditional approaches fail to capitalize on the domain knowledge and flexibility gained by transforming the feature space and the output space simultaneously. A new approach is proposed that utilizes domain knowledge, which is automatically discovered from the data, to combat the "small sample size" problem.

Spatially limited training data can result in poor inference concerning the true populations. The detrimental impact that can result if this issue is ignored is explored and demonstrated. Transferal of information that was previously acquired is used to update the signatures with the new clusters if the hypothesis that the new clusters are indeed just deformed versions of what already exists in the spectral library is accepted.

Independent of limited training data, both in terms of the spatial implications and limited quantity, different sampling subsets of the same ground truth may result in slightly different classifiers. This issue has not been addressed rigorously. The advantages gained by using an ensemble of classifiers built from sub-samples of training data are widely acknowledged but have not previously been used in the context of a hierarchical classifier for remote sensing data or for hyperspectral data in general. The ensemble of classifiers is used to identify a suitable level of the tree for situations where the resolution of the output space cannot be supported. Further decisions of how the classification structure should be adapted and at what level need to be made are explored. Furthermore, pseudo-labeled data are utilized to improve classification results at that level of resolution.

Table of Contents

List of Tables xii

List of Figures xiv

Chapter 1: Introduction 1

1.1 The classification problem 1

1.2 Motivation 2

1.2.1 Estimation problems with limited amount of training data 4

1.2.2 Samples do not fully characterize the population 6

1.3 Problem statement: Limited training data and dynamic classification methods 8

Chapter 2: Background and Related Work 11

2.1 The classification framework 12

2.2 Statistical pattern classification 13

2.2.1 High dimensional hyperspectral input space 14

2.2.2 Dimensionality of output space dependent upon resolution 16

2.3 Pairwise classifier framework 16

2.3.1 Feature selection/extraction 18

2.3.2 Hard and soft Bayesian pairwise classification with Gaussian(s) 19

2.3.3 Hard and soft combiners 20

2.4 Best-bases feature extraction 21

2.4.1 Previous work 21

2.4.2 Top-down Generalized Local Discriminant Bases 22

2.4.3 Bottom-up Generalized Local Discriminant Bases 24

2.5 Binary hierarchical classifier framework 25

2.5.1 Fisher feature extraction 27

2.5.2 Bottom-up BHC 28

2.5.3 Top-down BHC 29
2.5.4 Combining in BHC 30

Chapter 3: Adaptive Best-Basis Bayesian Hierarchical Classifier 31
3.1 Related Work With Limited Training Data 34
3.1.1 Parameter stabilization techniques 34
3.1.2 Improve ratio of training data to input dimensionality 36
3.1.3 Subsampling/Combining schemes 38
3.2 Best Basis Bayesian Hierarchical Classifier (BB-BHC) 39
3.2.1 Best-Basis and the Binary Hierarchical Classifier framework 40
3.2.2 Adaptive feature space for the BB-BHC 42
3.2.3 Best-Basis and Limited Data 44
3.3 Application of Adaptive BB-BHC to Classification of Hyperspectral Data 45
3.3.1 Bolivar Peninsula 45
3.3.1.1 Classification accuracies across decreasing sampling percentages 48
3.3.1.2 Domain knowledge and image evaluation 51
3.3.2 Kennedy Space Center 52
3.3.2.1 Classification accuracies across decreasing sampling percentages 56
3.3.2.2 Domain knowledge and image evaluation 58
3.3.2.3 Intrinsic dimensionality 60

Chapter 4: Spatially Limited Ground Truth and Knowledge Transfer 63
4.1 Motivation 64
4.1.1 Shortcomings of traditional classification 65
4.1.2 Dynamic application area 66
4.2 Impact of limited spatial coverage of the ground truth 69
4.2.1 Transferring the classifiers 70
4.2.1.1 Bolivar Peninsula Classification Accuracies 71

4.2.1.2 Cape Canaveral Classification Accuracies 74
4.2.2 Combined classification results 75
4.3 Knowledge Transfer of Trees and Fisher Projections 77
4.3.1 Background 77
4.3.2 Updating parameter estimates by pseudo-labeled data 78
4.3.3 Performance 80
4.4 conclusions 82
Chapter 5: Ensembles and Output Space Precision 83
5.1 Constructing a Master Tree 84
5.2 Transferring the Master Tree and Identifying an Appropriate Output Space Precision 87
5.2.1 Distance measure between the pseudo-labeled clusters 88
5.2.2 Purity of the ensemble at each partition 89
5.3 Application to Hyperspectral Data 90
Chapter 6: Concluding Remarks 94
6.1 Summary of Contributions 94
6.1.1 Limited quantity of training data 94
6.1.3 Spatially limited training data 95
6.1.4 Ensemble of classifiers 95
6.2 Future Work 96
6.2.1 Feature selection 97
6.2.2 Unsupervised clustering 97
6.2.3 Deformable models 98

Appendix A 100
Appendix B 104
Appendix C 108
Appendix D 109
Appendix E 115
Appendix F 123
Appendix G 139
Appendix H 143
Appendix I 145
Bibliography 149
Vita 158

List of Tables

Table 3.1: Number of observations per class for Bolivar Peninsula at two different areas used for testing 47

Table 3.2: Classes for Bolivar Peninsula, Site 1, and the quantity of training data per class by sampling percentage 48

Table 3.3 Number of observations per class for Cape Canaveral at two different areas used for testing 54

Table 3.4: Classes for Cape Canaveral and the quantity of training data per class by sampling percentage 55

Table 4.1: Bolivar Peninsula average training and test set accuracies when classifier is applied to the site at which the ground truths were acquired 71

Table 4.2: Bolivar Peninsula classification accuracies when classifier is applied to the alternate site at which the ground truths were acquired 72

Table 4.3: KSC average training and test set accuracies when classifier is applied to the site at which the ground truths were acquired. 74

Table 4.4: KSC classification accuracies when classifier is applied to the alternate site at which the ground truths were acquired 75

Table 4.5: Bolivar Peninsula and KSC average training and test set accuracies when classifier is applied to the ground truths combined from both sites at which the ground truths were acquired 76

Table 4.6: Bolivar Peninsula classification accuracies when classifier is updated using pseudo-labeled data to estimate the new parameters and applied to the alternate site at which the ground truths were acquired 80
Table 4.7: KSC classification accuracies when classifier is updated using pseudo-labeled data to estimate the new parameters and applied to the alternate site at which the ground truths were acquired 82
Table 5.1: Bolivar Peninsula classification accuracies when classifier is updated using pseudo-labeled data to estimate the new parameters and applied to the alternate site at which the ground truths were acquired 91
Table 5.2 KSC classification accuracies when classifier is updated using pseudo-labeled data to estimate the new parameters and applied to the alternate site at which the ground truths were acquired 92

List of Figures

Figure 2.1: The general classification framework 12

Figure 2.2: An example of hyperspectral data where d=224 15

Figure 2.3: Pairwise classifier framework: $\binom{C}{2}$ pairwise classifiers with respective extractors, feature spaces, and classifiers 17

Figure 2.4: Example of an arbitrary binary tree obtained from TD-GLDB. The dark blocks are the bases 23

Figure 2.5: Example of an arbitrary binary tree obtained from BU-GLDB. The dark blocks are the bases 25

Figure 2.6: An example of a Binary Hierarchical Classifier (BHC) with C classes. Each internal node n comprises of a feature extractor, a classifier, a left child $2n$, and a right child $2n+1$. Each node n is associated with a meta-class Ω_n 26

Figure 3.1: An overview image of the two test sites at Bolivar Peninsula TX ... 46

Figure 3.2: Classification (test set) accuracies for Bolivar Peninsula 50

Figure 3.3: An overview image of the two test sites at Kennedy Space Center; Cape Canaveral Florida 53

Figure 3.4: Classification (test set) accuracies for KSC, Cape Canaveral FL 57

Figure 4.1: Spectral signatures for Class 8 (General Upland) at Bolivar Peninsula for the two different test sites 67

Figure 4.2: Spectral signatures for Hardwood Swamp at the two different Kennedy Space Center sites 68

Figure 4.3: Deformation of the meta-class distributions in the Fisher projected space calculated from the ground truth acquired at Bolivar Peninsula Site 2 73
Figure 5.1: An example of 6 possible hierarchies involving 4 classes 86

Chapter 1: Introduction

1.1 THE CLASSIFICATION PROBLEM

The classification of a pattern – identifying the "label" of an observation - is an essential task humans perform every day [1-2, 13, 14, 65]. Almost effortlessly, individuals identify each other by sight or touch, food by its smell, or animals by their sound. However, the development of statistical pattern recognition algorithms to automate many of these seemingly simple tasks continues to be an active area of research. Specifically, a classification problem arises when an observation from some pattern needs to be identified, but no given label is available. Instead, a label must be assigned to the observation based upon a vector of measurements. For example, it may be necessary for a fish-packing plant to separate the incoming fish according to species. Although no label is available, it may be possible to classify the fish as the correct species with a high level of accuracy based on measurements such as length, color, width, number and shape of fins, position of the mouth, and weight [1].

When classifying an observation, the number of labels from which to choose is typically assumed to be finite. Each label can be characterized by the probability distribution of the observations associated with that label, and therefore, each observation can be considered as a random observation from a label population. In this context, the labeling of observations can be resolved by using 'statistical decision functions' in which there are a number of hypotheses; each hypothesis purports that the probability distribution of the population from which the observation is acquired is that of a given label [2]. Given a random

observation with a vector of measurements, one of the hypotheses (labels) must be selected in favor of the others. Pattern classification and recognition problems are often categorized as "supervised" when it is assumed that a set of samples has been acquired that is representative of the patterns to be classified. Typically, a randomly selected portion of the sample is used to estimate the parameters of the label specific probability distributions for classification (training), and the remaining labeled samples are used to estimate the classification accuracy (testing). Necessary revisions to the classification model could be made, and the training and testing cycle repeated. With supervised classification, there is a training and test set for which the state of nature (class label) for each sample is known [1, 3, 5]. Therefore, for supervised approaches, the classification/recognition of an input/pattern is essentially the task of identifying the predefined class to which it belongs, where the user defines the classes. The proposed research is based on this statistical perspective of the classification problem.

1.2 MOTIVATION

With most difficult classification problems, it is advantageous to make full use of the domain knowledge specific to the application area. In particular, the problem specific characteristics of the data, quantity of data available for analysis, and information about the data acquisition process can be extremely important. One such problem, land cover classification, is an important application that can potentially benefit from remotely sensed data acquired by space-based and airborne platforms. Of particular interest is the potential of new hyperspectral

sensors that simultaneously acquire information at hundreds of wavelengths. They characterize the response of targets (spectral signatures) in greater detail than traditional sensors and thereby can improve discrimination between targets [8, 9, 51]. Data from these sensors are classified to create maps required for monitoring many critical earth resources problems.

A classification problem arises when an observation of a pattern, for which no specified label is given, needs to be identified. Instead, a vector of measurements $(x_1, x_2, ..., x_d)$ for the observation is used to assign a label to the pattern, where *d* is the dimensionality of the data. Typically the number of labels (L_i) from which to select is assumed to be a finite number *C* such that $(L_1, L_2, ..., L_C)$. Each label is a random variable that can be characterized by the joint probability distribution of the observations associated with that label, and therefore, each vector valued data point can be considered as a random observation from a label-conditional probability density function $P(x_1, x_2, ..., x_d | L_i)$, the probability density function (pdf) for $\boldsymbol{x}$ given that the label is L_i. The estimates of the parameters of these pdfs are used to form the decision rules that divide the feature space, F , into *C* decision regions, each representative of a label [1, 2, 5, 14, 21, 22]. Where possible, the most appropriate land cover label is determined using supervised classification in which training data (labeled "ground truth") $\boldsymbol{X}$ are used to estimate the label-conditional probability density functions $P(x_1, x_2, ..., x_d | L_i), i = 1, ..., C$. Typically, for practical reasons, the training data are collected only at a limited number of sites and, unfortunately, in a limited quantity $|\boldsymbol{X}|$.

In many cases, the criteria for selection of training and test samples are dictated by factors that are independent of the statistical analysis. Even if the sampling scheme has been developed using rigorous statistical methods, it may be impractical or impossible to implement these plans in real world applications. For example, land cover classification is typically performed based on the spectral response of land cover classes from "pockets" of a region, whereas the goal is to classify the entire region and possibly to even utilize the information for classification of other data acquired over regions for which ground truth cannot be acquired. When obtaining "ground truth" for land cover classes over an extensive region, it would be impractical, in terms of time and cost, to obtain samples at randomly chosen locations. A much more realistic scenario would involve acquisition of samples at reasonably accessible sites, with the quantity of samples (the number of labeled pixels available for the researcher) being dependent upon many external factors. In addition to the limitations of time and money, other issues, such as physical access and cloud cover during the acquisition of the remotely sensed data, can only serve to further reduce the quantity of usable data and the variety of sites. This research is motivated by the shortcomings of traditional classification methods for dealing with land cover classification in the context of limited quantities of non-global samples.

1.2.1 Estimation problems with limited amount of training data

Many difficult classification problems, such as land cover classification, involve a high dimensional input space and a large number of candidate labels. However, due to the "curse of dimensionality" and the Hughes phenomena [1, 3,

6, 7, 13, 56], it is necessary to reduce the dimensionality of the input space (I) when only limited quantities of training data are available. It is particularly problematic that almost all conventional statistical approaches require computing and inverting covariance matrices $\Sigma = \mathrm{Cov}(\boldsymbol{X}) = E(\boldsymbol{X} - \mu)(\boldsymbol{X} - \mu)'$, which are typically unknown and are estimated by the sample covariance matrices $S = \frac{1}{|\boldsymbol{X}| - 1} \sum_{j=1}^{|\boldsymbol{X}|} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'$. For example, Fisher's linear discriminant function is commonly used in linear discriminant analysis and is defined in terms of the within class covariance matrix and the between class covariance matrix. Any classifier using Fisher's linear discriminant function requires inversion of the within class covariance matrix. For the covariance matrix of *d*-dimensional data, there are $d(d+1)/2$ parameters to estimate and, at a minimum, $d+1$ observations are required to ensure a non-singular/invertible sample covariance matrix [57].

Numerous studies have considered what the minimum number of training samples should be, in relation to the dimensionality, for trustworthy estimation of a covariance matrix [3, 5, 19, 26]. In general, the literature recommends having 4-10 times the number of observations as the dimensionality for linear classifiers. While quadratic classifiers may perform better than linear classifiers in certain situations, the recommended number of observations is related to the square of the dimensionality [24, 25, 27]. This relationship is even worse for non-parametric classifiers, where it has been estimated that the required quantity of training data increases exponentially as the dimensionality increases in order to accurately

estimate the multivariate densities [3]. Regardless of the actual classifier being used, while hyperspectral data provide the opportunity for improved discrimination between different land cover types, the problems with limited training data in relation to dimensionality become more critical and can have a significant impact on classification accuracies [19, 25, 53]. Hereafter, this dilemma is referred to as the "small sample size" problem.

1.2.2 Samples do not fully characterize the population

For this research, classification focused on development of statistical methods for classification of objects for which there is a large number of descriptive attributes, thereby yielding a vector of inputs that is of high dimension. The methods are demonstrated on problems in land cover mapping using sensors that acquire data simultaneously in a large number of windows of the electromagnetic spectrum. The recent advancement in airborne and space based-sensors have made it possible to acquire hyperspectral data in over 200 bands. Bands of hyperspectral data are narrow, contiguous windows of the electromagnetic spectrum. Unlike the traditional multi-sensors, which use recorded electromagnetic spectral responses in a smaller number of wide bands, often of at least 100 nm, these new hyperspectral sensors can acquire spectral responses in hundreds of narrow windows of 5-10 nm width. Where possible, the most appropriate land cover label is determined using supervised classification. Typically, methods are trained and tested for classification accuracy from data in a "closed world"; all of the training and testing data are selected from a

contiguous subset(s) of the region. Furthermore, for practical reasons, the labeled data are collected only at a limited number of sites.

It is common to randomly select 50% of the labeled data for training, resulting in "testing" points that are neighboring - or even surrounded by - points on which the classifier was trained. Due to the limited spatial extent of the available data, it is possible that the training data are not representative of the entire population. In essence, inferences are being made about the underlying characteristics of the population based on local information. Consequently, the resulting classifier probably performs poorly on the other "segments" of the population where no labeled data are available. The true classification performance is compromised, and the generalization accuracy (the classification accuracy of labeled data not in the training set) may also be deceptive, inaccurate, and inflated. Thus, these results likely result in a poor representation of both the populations that are known to exist in an area and even worse characterization of the even more difficult problem: where the data have not been "seen".

A classifier that is "trained" on the responses of land cover types from certain sites may ultimately be used to classify the land cover types of other areas where no immediate data are available. Natural variation within classes during an acquisition is increased by the impact of spatially varying characteristics such as in soil composition, elevation, and environmental factors. While the issue of within sample variance has been recognized, the problem of spatial non-stationarity has not been studied. The limited spatial context of the acquired training data can result in a biased estimate of the true underlying populations.

Not only is the likely outcome diminished classification accuracy over the entire scene, but the researcher may also be misled that the accuracies are very high based upon test set results.

1.3 PROBLEM STATEMENT: LIMITED TRAINING DATA AND DYNAMIC CLASSIFICATION METHODS

This research focused on the development of a hierarchical approach for classification that is robust with respect to training data that are limited both in quantity and spatial extent. Because the general problem of robust classification is quite broad, the research focuses on specific problems. The problems are outlined here, in order of increasing difficulty and decreasing domain knowledge, and presented completely in Chapters 3-5.

1. Many difficult classification problems involve a high dimensional input and output space (candidate labels). Due to the "curse of dimensionality", it is necessary to reduce the size of the input space when there is only a limited quantity of training data available. While a significant amount of research has focused on transforming the input space into a reduced feature space that accurately discriminates between the classes in a fixed output space, traditional approaches fail to capitalize on the domain knowledge and flexibility gained by transforming the feature space and the output space simultaneously. A new approach is proposed in this research that utilizes domain knowledge, which is

automatically discovered from the data, to combat the "small sample size" problem.

2. Spatially limited training data can result in poor inference concerning the true populations. The detrimental impact that can result if this issue is ignored is explored and demonstrated. Transferal of information that was previously acquired is used to update the signatures with the new clusters if the hypothesis that the new clusters are indeed just deformed versions of what already exists in the spectral library is accepted.
3. Independent of limited training data, both in terms of the spatial implications and limited quantity, different sampling subsets of the same ground truth may result in slightly different classifiers. This issue has not been addressed rigorously. The advantages gained by using an ensemble of classifiers built from sub-samples of training data are widely acknowledged but have not previously been used in the context of a hierarchical classifier for remote sensing data or for hyperspectral data in general. The ensemble of classifiers is used to identify a suitable level of the tree for situations where the resolution of the output space cannot be supported. Further decisions of how the classification structure should be adapted and at what level need to be made are explored. Furthermore, pseudo-labeled data are utilized to improve classification results at that level of resolution.

This research focuses on the development of a hierarchical approach for classification that is robust with respect to populations that exhibit spatial variability in the context of limited –both in quantity and spatially - training data. The framework on which this research is developed is reviewed in Chapter 2. The specific problems that are addressed were outlined in Section 1.3 and are presented more completely in Chapters 3-5 followed by some concluding comments.

Chapter 2: Background and Related Work

Statistical classification of high dimensional input and output problems has been studied intensively for the past decade. This has been motivated by acquisition of higher dimensional information and made possible by advances in computational hardware. This chapter contains a review of the literature related to both feature extraction and the classification problem. A detailed description of the hierarchical classifier, which provided the foundation for the research reported here, is provided, including a description of the pairwise classifier framework upon which it was based. The presentation of the Bayesian Pairwise Classifier framework entails a description of the feature selection/extraction process by way of pairwise Fisher projection and forward feature selection, the Bayesian Pairwise Classifier itself, and the process of combining the pairwise classifiers using "hard" and "soft" techniques. Domain knowledge specific to the application area of land cover classification with hyperspectral data motivated the development of a best-bases feature extraction algorithm, to be used within the Bayesian Pairwise Classifier, so it too is discussed in detail. The explanation of the best-bases algorithm highlights the strengths of a "top-down" search and a "bottom-up" search for best bases. Scalability issues with the Bayesian Pairwise Classifier, as well as the desire to automatically discover and use domain knowledge, motivated the development of the Binary Hierarchical Classifier framework that ultimately provided the foundation for the new methods developed in this research. Thus, the Binary Hierarchical Classifier is presented

in detail, inclusive of an explanation of the "top-down" and "bottom-up" building approaches and feature extraction.

2.1 THE CLASSIFICATION FRAMEWORK

The general classification problem includes the following steps (depicted in Figure 2.1).

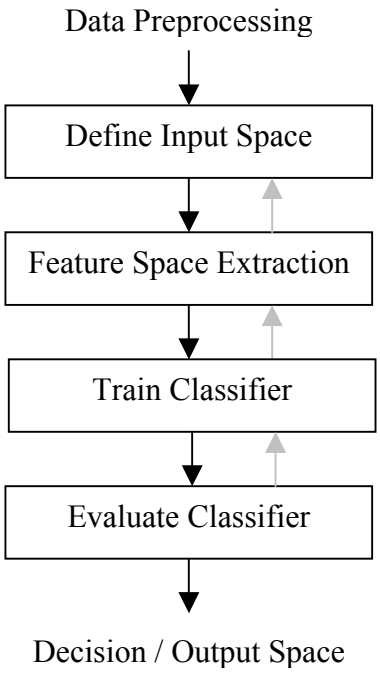


Figure 2.1: The general classification framework

Data Preprocessing involves radiometric, geometric, and atmospheric correction of the data and conversion to a format that is usable by a classifier. The selection of the set of ground cover types that is used by the classifier is also assumed to occur during this step.

Define Input Space $\left(\boldsymbol{I}\left(\boldsymbol{x, y, l}\right); x = 1, n_1; y = 1, n_2; l = 1, d\right)$ is assumed to be the preprocessed d-dimensional data points, each associated with a location $\left(x, y\right)$ on a regular grid.

Feature Space $\left(\boldsymbol{F}\right)$ **Extraction** is the transformation of the input space I into the feature space F. The feature space is selected using domain knowledge or statistical techniques such that the classes are more easily discriminated.

Train Classifier stage uses the training data to estimate the parameters of the probability density functions representing the responses of the individual classes.

Evaluate Classifier includes labeling/classifying every pixel in the image as one of the given ground cover types by using the trained classifier. The assessment of the classifier performance is accomplished by any combination of options such as accuracy tables, confusion matrices, and expert opinions.

Decision / Output Space $\left(\boldsymbol{O}\right)$ is the set of all "observed" classes. Domain knowledge helps determine the appropriate application dependent output space

2.2 STATISTICAL PATTERN CLASSIFICATION

As noted in Chapter 1, a classification problem arises when an observation of a pattern needs to be identified, but no given label is available. Instead, one of C labels (L_i) must be assigned to the observation based on a vector of

measurements $(x_1, x_2, ..., x_d)$ obtained from the d dimensional data. Each label can be considered a random variable characterized by a probability distribution of the observations associated with that label. Likewise, each observation can be considered as a random observation from a label-conditional pdf $P(x_1, x_2, ..., x_d | L_i)$. The estimates of the parameters of these pdfs are used to form the decision rules that divide F into C decision regions, each representative of a label [1, 2, 5, 14, 21, 22]. Selection of a "good" feature space F is application dependent. Further, there is also a class of these problems where the inputs represent sequential measurements over some domain – such as the spectrum.

2.2.1 High dimensional hyperspectral input space

The recent advancement in airborne and space borne sensors have made it possible to acquire hyperspectral data in over 200 bands. Bands of hyperspectral data are narrow, contiguous windows of the electromagnetic spectrum. Unlike the traditional multi-sensors, which recorded spectral responses in wide bands, often of at least 100 nm, modern sensors can acquire spectral responses in hundreds of narrow windows of 5-10 nm width. In the remote sensing community, these instruments are referred to as hyperspectral sensors. The data consist of a three-dimensional array: $I(x, y, d)$ - where (x, y) denotes the location of the pixel in an image and 'd' represents the spectral band. The value stored in the (x, y, d) location is the spectral response from that particular pixel. Figure 2.2 illustrates the data composition. The potential increased utility of hyperspectral data is being investigated for land cover mapping. Data from hyperspectral sensors characterize the response of targets (spectral signatures) with greater

detail than traditional sensors and thereby can potentially improve discrimination between targets [8, 9, 51, 72]. When possible, "supervised" classification methods are used where label-conditional probability density functions

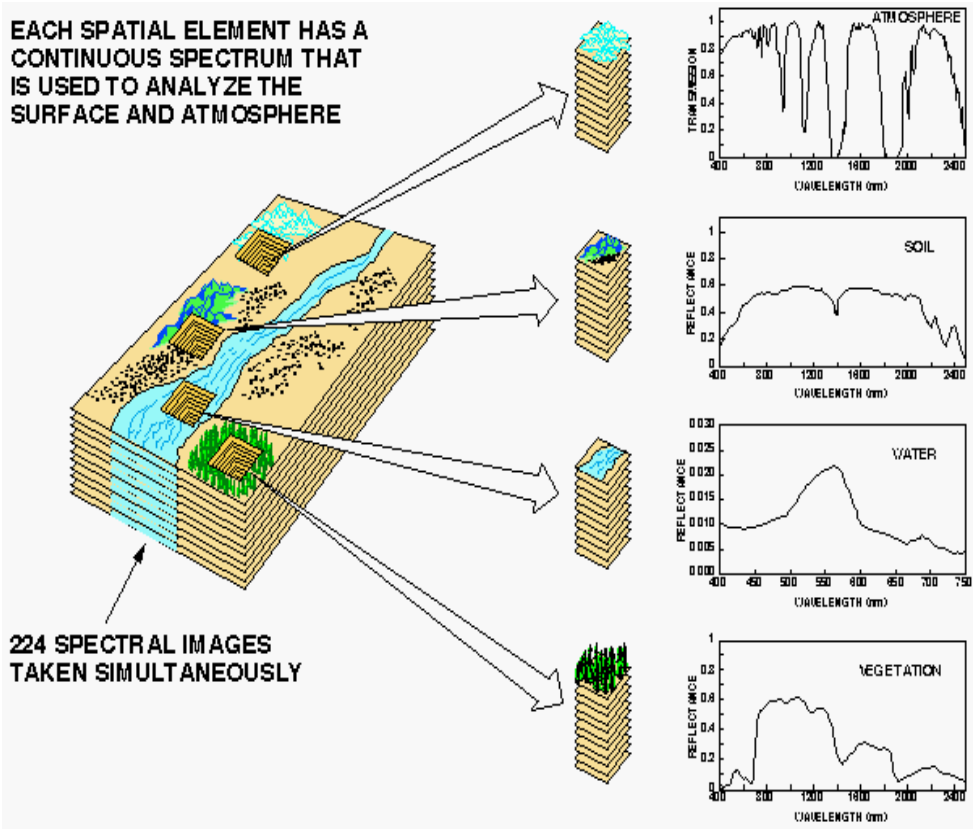


Figure 2.2: An example of hyperspectral data where d=224

$P(x_1, x_2, ..., x_d \mid L_i), i = 1, ..., C$ are estimated by available labeled training data ("ground truth") $\boldsymbol{X}$ to determine the most appropriate land cover label. Practical

limitations may only allow the acquisition of a limited quantity $|\boldsymbol{X}|$ of training data collected at a limited number of locations.

2.2.2 Dimensionality of output space dependent upon resolution

In many application areas, the dimensionality of the output space (the number of candidate labels from which to choose) is clearly defined. An example would be the classification of the letters in the English alphabet. For example, nearly everyone, if informed that they also need to discriminate between upper case and lower case letters, would identify fifty-two possible labels. That is not the case with land cover classification. The dimensionality of the output space will depend upon the researcher's judgment and the spatial and spectral usage of the data coupled with the intended usage of the information. While land cover classification schemes have been devised as an attempt to standardize mapping across multiple scales, the specific selection of classes remains task dependent. The decision on how fine or coarse the "resolution" on the general classes will be, and the resulting dimensionality of the output space, is important and has a significant impact on the resulting classification accuracies.

2.3 PAIRWISE CLASSIFIER FRAMEWORK

To maximize classification accuracy, features must be used that provide the best discrimination between classes. For high dimensional input problems, feature extraction is required prior to classification - to reduce both the computational burden and the effect of highly correlated inputs. Because different groups of classes are best distinguished by different sets of features

(typically spectral bands in remote sensing applications), it is often desirable that the feature extractors extract group-specific features [10, 11]. Additionally, each feature space selected to discriminate between the class pairs should be less complex than the feature space necessary for the entire C-class problem. These issues motivated the development of the Pairwise Classifier (PC) framework of Kumar *et. al* [8, 10, 15-17]. In the PC framework of Figure 2.3, the original C-class problem is decomposed into $\binom{C}{2}$ 2-class problems. Feature selection is

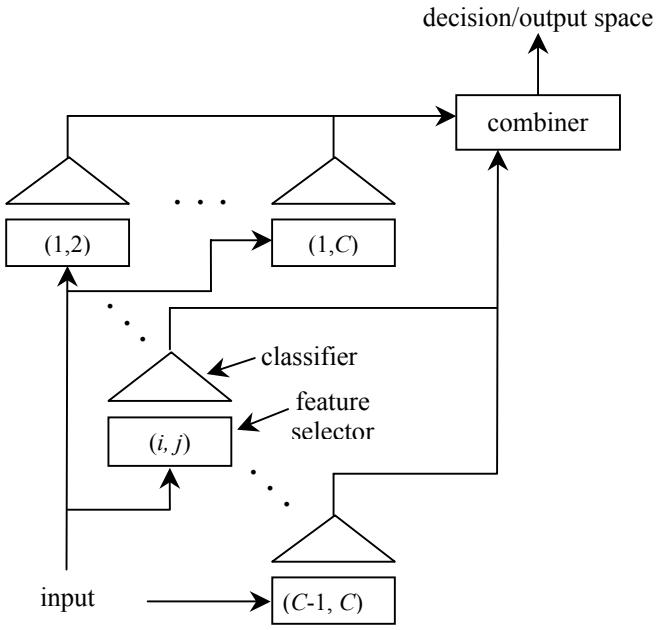


Figure 2.3: Pairwise classifier framework: $\binom{C}{2}$ pairwise classifiers with respective extractors, feature spaces, and classifiers

accomplished for each class pair such that the classifier for each pair (L_i, L_j) has an associated feature extractor $\Psi_{ij} : I \rightarrow F_{ij}$ where I is the input (vector) space

and F_{ij} is the extracted feature vector for each pair $\left(L_i, L_j\right)$ [10-11, 48]. Each of the two-class problems is solved independently, using a Bayesian pairwise classifier, and the results are combined by either the 'voting' method or the MAP (maximum a posteriori probability) method [6, 66].

2.3.1 Feature selection/extraction

In [8, 10, 15] Kumar et al. proposed a generic relevance measure,

$$\mathrm{J}_{ij}\left(F_{ij}\right)=\frac{1}{\left|X_i\right|}\sum_{\mathbf{x}\in X_i}\log\frac{\hat{P}_{ij}\left(L_i\middle|\Psi_{ij}\left(\mathbf{x}\right)\right)}{\hat{P}_{ij}\left(L_j\middle|\Psi_{ij}\left(\mathbf{x}\right)\right)}+\frac{1}{\left|X_j\right|}\sum_{\mathbf{x}\in X_j}\log\frac{\hat{P}_{ij}\left(L_j\middle|\Psi_{ij}\left(\mathbf{x}\right)\right)}{\hat{P}_{ij}\left(L_i\middle|\Psi_{ij}\left(\mathbf{x}\right)\right)}, \tag{2.1}$$

based on the estimated log-odds of posterior probabilities. Since this relevance measure does not assume anything about the data set or the nature of the classifier, it can be used in any case where the output of the pairwise classifier is the estimated posterior $\hat{P}_{ij}\left(L_i\middle|\Psi_{ij}\left(\mathbf{x}\right)\right)$. This relevance measure (2.1) is a filter type goodness measure according to Langley's taxonomy, where the idea is to exploit the differences of the posterior probabilities when a point belongs to one class versus the other [39]. Following Baye's rule:

$$\hat{P}_{ij}\left(L_k\middle|\Psi_{ij}\left(\mathbf{x}\right)\right)=\frac{\hat{P}_{ij}\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_k\right)\hat{P}_{ij}\left(L_k\right)}{\hat{P}_{ij}\left(\Psi_{ij}\left(\mathbf{x}\right)\right)},\quad k=i,j. \tag{2.2}$$

By substituting (2.2) into (2.1), and estimating the priors by the class proportionality, the estimated log-odds of posterior probabilities reduces to:

$$\mathrm{J}_{ij}\left(F_{ij}\right)=\frac{1}{\left|X_i\right|}\sum_{\mathbf{x}\in X_i}\log\frac{\hat{P}_{ij}\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_i\right)}{\hat{P}_{ij}\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_j\right)}+\frac{1}{\left|X_j\right|}\sum_{\mathbf{x}\in X_j}\log\frac{\hat{P}_{ij}\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_j\right)}{\hat{P}_{ij}\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_i\right)}. \tag{2.3}$$

While the mapping $\Psi_{ij} : I \to F_{ij}$ could be any feature extraction transforming the input space I into a feature space F_{ij} that is more suitable for discriminating the class pair $\left(L_i, L_j\right)$, Kumar used a greedy forward feature selection algorithm [8] where, for each pair $\left(L_i, L_j\right)$, the first band selected is that which maximizes (2.3). Additional bands are added in order of their respective incremental contributions if the corresponding increase in (2.3) is greater than a specified threshold.

2.3.2 Hard and soft Bayesian pairwise classification with Gaussian(s)

Although any classifier whose output can be used to infer the respective probabilities of class occurrence satisfy the requirements for the 2-class classifiers, Kumar et al. investigate both a single and mixture of Gaussians to model the probability density functions $\hat{P}_{ij}\left(\Psi_{ij}\left(\mathbf{x}\right)\middle| L_k\right),\ k = i, j$ for use in Bayesian classifiers [10, 15]. Once the pdfs are obtained for each pair in the feature space F_{ij} resulting from the $\Psi_{ij}\left(\mathbf{x}\right)$ transformation, the Bayesian pairwise classifier (BPC) framework uses Baye's rule to obtain the posteriors (2.2). In the single Gaussian (BPC_1) formulation, each pdf $\hat{P}_{ij}\left(\Psi_{ij}\left(\mathbf{x}\right)\middle| L_k\right),\ k = i, j$ is modeled as a Gaussian $\left(\mu_k^{(i,j)} \in \Re^{d_{ij}\times 1}, \Sigma_k^{(i,j)} \in \Re^{d_{ij}\times d_{ij}}\right)$ in the $d_{ij} = \left|F_{ij}\right|$ reduced dimensional space. The sample means and covariances for any feature extractor $\Psi_{ij}\left(\mathbf{x}\right)$ transformation are given by:

$$\mu_k^{(i,j)} = \frac{1}{\left|X_k\right|}\sum_{\mathbf{x}\in X_k}\Psi_{ij}\left(\mathbf{x}\right),\ \ k = i, j, \tag{2.4}$$

$$\Sigma_k^{(i,j)} = \frac{1}{\left|X_k\right|}\sum_{\mathbf{x}\in X_k}\left(\Psi_{ij}\left(\mathbf{x}\right) - \mu_k^{(i,j)}\right)\left(\Psi_{ij}\left(\mathbf{x}\right) - \mu_k^{(i,j)}\right)^{\mathrm{T}},\ \ k = i, j. \tag{2.5}$$

Alternatively, in the mixture of Gaussian (MOG) formulation (BPC_n), the class conditional pdfs are represented as

$$\hat{P}\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_k\right)=\sum_{\alpha=1}^{n_k^{(i,j)}}\pi_{k,\alpha}^{(i,j)}\mathrm{G}\left(\Psi_{ij}\left(\mathbf{x}\right);\mu_{k,\alpha}^{(i,j)},\Sigma_{k,\alpha}^{(i,j)}\right),\qquad k=i,j, \tag{2.6}$$

where $n_k^{(i,j)}$ is the number of Gaussians in the mixture for L_k, and $\left\{\mu_{k,\alpha}^{(i,j)}\left(\in F_{ij}\right),\Sigma_{k,\alpha}^{(i,j)}\right\}$ are the mean vector and covariance matrix for the α^{th} Gaussian in the mixture for class L_k of the $\left(L_i,L_j\right)$ classifier. The multivariate Gaussian pdf Γ is given by:

$$\mathrm{G}\left(\Psi_{ij}\left(\mathbf{x}\right);\mu,\Sigma\right)=\frac{1}{\sqrt{\left(2\pi\right)^{|F_{ij}|}\left|\Sigma\right|}}\exp\left[-\tfrac{1}{2}\left(\Psi_{ij}\left(\mathbf{x}\right)-\mu\right)^T\Sigma^{-1}\left(\Psi_{ij}\left(\mathbf{x}\right)-\mu\right)\right]. \tag{2.7}$$

Additionally, a "growing and pruning" algorithm selects the number of Gaussians and the parameters for the resulting mixtures, while simultaneously greedily selecting the feature space via the forward feature selection algorithm [8]. For either a single Gaussian or a MOG, the estimated class conditional pdfs are used to obtain the classifier output by Baye's rule:

$$P_{ij}\left(L_k\middle|\Psi_{ij}\left(\mathbf{x}\right)\right)=\frac{P\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_k\right)P_{ij}\left(L_k\right)}{P\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_i\right)P_{ij}\left(L_i\right)+P\left(\Psi_{ij}\left(\mathbf{x}\right)\middle|L_j\right)P_{ij}\left(L_j\right)},\quad k=i,j. \tag{2.8}$$

2.3.3 Hard and soft combiners

The PC framework generates $\binom{C}{2}$ outputs per pixel, one for each of the pairwise classifiers. In a "hard" combiner, a simple voting scheme proposed by Friedman [6] is followed, whereby the pixel is assigned the class label that occurs

the maximum number of times in the $\binom{C}{2}$ classifiers. Conversely, in a "soft" combiner approach, estimates of the true posteriors $P\left(L_i \middle| \left(x_1, x_2, ..., x_n\right)\right)$ i=1,…,C of the classes are obtained using a hill-climbing algorithm proposed by Hastie and Tibshirani [66]. The estimated posteriors can then be used in the MAP methodology where the pixel is assigned the class label that corresponds to the maximum estimated posterior probability.

2.4 BEST-BASES FEATURE EXTRACTION

While the PC framework reduces the dimensionality of the input space, it ignores correlation between individual inputs. From the domain knowledge in this field, it is already known that the original input features, called bands, are narrow, contiguous windows of the electromagnetic spectrum and that bands that are "spatially close" to each other tend to be highly correlated. Kumar et al. developed "best-bases" feature extraction algorithms which tie the feature space selection to the classification process by exploiting the correlation between adjacent spectral inputs for use in the PC framework [8, 16, 17]. The bottom-up and top-down algorithms they developed for combining subsets of adjacent bands are presented in the remainder of this section.

2.4.1 Previous work

Jia and Richards previously investigated band-combining algorithms and proposed a feature extraction technique for hyperspectral data based on Segmented Principal Components Transformation (SPCT) [62, 63]. With SPCT, image processing-based edge detection algorithms are used to transform the *D*

individual bands into subsets of adjacent bands that are highly correlated based upon the estimated population correlation matrix. From each subset, a small number of principal components that capture most of the variance in the data are selected to yield a feature vector that is of significantly lower dimension than the original input dimensionality. Although this approach utilizes the highly correlated adjacent bands in hyperspectral data to reduce the input feature space, it does not guarantee good discrimination capability because the principal component transform focuses on extracting orthogonal linear combinations with maximum variance rather than maximizing discrimination among classes. Additionally, the segmentation approach of SPCT is based upon the correlation matrix over all of classes and, because the interband correlation can vary significantly among the *C* classes, a good band-combining algorithm should exploit the class-conditional correlation matrices. Furthermore, estimating the correlation matrix in the original high dimensional input space requires an adequate amount of training data, and this dependence is not addressed.

2.4.2 Top-down Generalized Local Discriminant Bases

Saito and Coifman [67] developed a local discriminant bases (LDB) algorithm for classification of signals and images. In LDB, a binary tree of bases is searched for complete bases, called the *best bases*, which maximize a discrimination information function. Although this algorithm takes advantage of the correlation between adjacent bands, it is severely limited because its binary structure does not allow flexible` lengths for the bases. Additionally, LDB searches for bases that help discriminate all the classes simultaneously rather

allowing for class dependencies. These limitations motivated the development of a top-down procedure generalized from LDB referred to as TD-GLDB.

For decomposing bands $[l,u]$ $(1 \le l \le u \le d)$, the relevance measure (2.1), a log-odds ratio of the posterior probabilities, was extended to determine which bands should be split [10], [15]:

$$\mathrm{J}\ (l,u) = \frac{1}{|X_i|} \sum_{\mathbf{x} \in X_i} \log \frac{\hat{P}_{ij}\left(L_i \middle| \mathrm{M}\ (\mathbf{x}|l,u)\right)}{\hat{P}_{ij}\left(L_j \middle| \mathrm{M}\ (\mathbf{x}|l,u)\right)} + \frac{1}{|X_j|} \sum_{\mathbf{x} \in X_j} \log \frac{\hat{P}_{ij}\left(L_j \middle| \mathrm{M}\ (\mathbf{x}|l,u)\right)}{\hat{P}_{ij}\left(L_i \middle| \mathrm{M}\ (\mathbf{x}|l,u)\right)}. \tag{2.9}$$

A function, $\mathrm{M}\ (\mathbf{x}|l,u)$, is necessary to represent each group-band. For simplicity, the mean of the bands is used (2.10):

$$\mathrm{M}\ (\mathbf{x}|l,u) = \frac{1}{u-l+1} \sum_{i=l}^{u} x_i. \tag{2.10}$$

Initially, $l=1$, $u=d$, and $\tilde{k}$ is sought using (2.11):

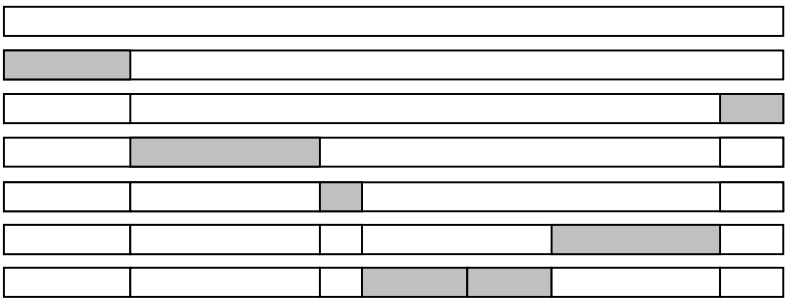


Figure 2.4: Example of an arbitrary binary tree obtained from TD-GLDB. The dark blocks are the bases.

$$\tilde{k} = \arg \max_{l \le k \le u} \max \left\{ \mathrm{J}\ (l,k), \mathrm{J}\ (k+1,u) \right\}. \qquad (2.11)$$

If $\mathrm{J}\ \left(l,\tilde{k}\right) > \mathrm{J}\ (l,u)$, $\left[l,\tilde{k}\right]$ is decomposed and if $\mathrm{J}\ \left(\tilde{k}+1,u\right) > \mathrm{J}\ (l,u)$, $\left[\tilde{k}+1,u\right]$ is decomposed. Group-bands such as those depicted in Figure 2.4 are generated, and a subset of the group-bands that best discriminates between $\left(L_i, L_j\right)$ is selected using a forward feature selection algorithm [68].

2.4.3 Bottom-up Generalized Local Discriminant Bases

A bottom-up algorithm, also generalized from the LDB algorithm and hence named GLDB-BU, was also developed by Kumar et al.[17]. Rather than searching for bases that help discriminate all the classes simultaneously - such as in LDB - GLDB-BU restricts its search for best bases specific to each pair of classes (groupings). An additional improvement is that any set of adjacent bands can be merged versus the recursive binary split of the bands as in LDB. For each class pair $\left(L_i, L_j\right)$ the estimated correlation matrix **q** and covariance matrix **Q** between all pairs of bands are given by (2.12):

$$\mathbf{q} = \left[q_{i,j} \right] = \frac{Q_{i,j}}{\sqrt{Q_{i,i} Q_{j,j}}} \in [0,1] \qquad \mathbf{Q} = \left[Q_{i,j} \right] = \frac{1}{|\boldsymbol{X}|-1} \sum_{\mathbf{x} \in d} (\mathbf{x} - \mu)(\mathbf{x} - \mu)^{\mathrm{T}}. \qquad (2.12)$$

The correlations between the bands, as well as the discrimination between the two classes when the bands are projected in the Fisher direction, are used as the criteria to group the bands $[l,u]$ $(1 \le l \le u \le d)$. The goal is to merge highly correlated bands in a way that also yields good discrimination between $\left(L_i, L_j\right)$. A 'correlation measure' $\mathbf{C}(l,u)$ is defined as the minimum pairwise correlation

$q_{i,j}$ over the $[l,u]$ subset. Similarly, a 'discrimination measure' $\mathbf{D}(l,u)$ is defined as the Fisher discriminant over the $[l,u]$ subspace. The algorithm searches for the maximal value of the product $\mathbf{J}(l,u)=\mathbf{C}(l,u)*\mathbf{D}(l,u)$, (restricted to pairs of bands/group-bands only). The pair is merged, the corresponding basis given by the Fisher projection over the subset is added, and the algorithm continues unless the resulting $\mathbf{J}(l,u)$ is less than that of both previous subsets [$\mathbf{J}(l)$ and $\mathbf{J}(u)$]. As in GLDB-TD, after the algorithm generates the group-bands such as those depicted in Figure 2.5, forward feature selection is used to choose an appropriate number of the bases. When used in

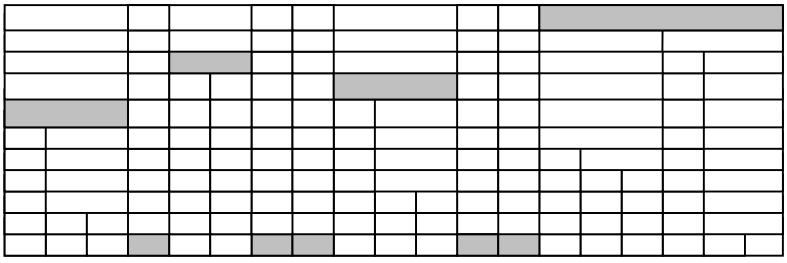


Figure 2.5: Example of an arbitrary binary tree obtained from BU-GLDB. The dark blocks are the bases.

conjunction with these best-bases feature extraction algorithms, the PC framework provides features that are of lower dimension and result in high classification accuracy [8].

2.5 BINARY HIERARCHICAL CLASSIFIER FRAMEWORK

In the single stage PC framework, the original *C*-class problem is decomposed into $\binom{C}{2}$ simpler 2-class problems where each meta-class is a

unique class. By reducing the size of the output space to a 2-class problem, the feature extractor is allowed the flexibility of obtaining a more group-specific feature space that should be less complex than the feature space necessary for comparable results for the entire *C*-class problem [10, 11].

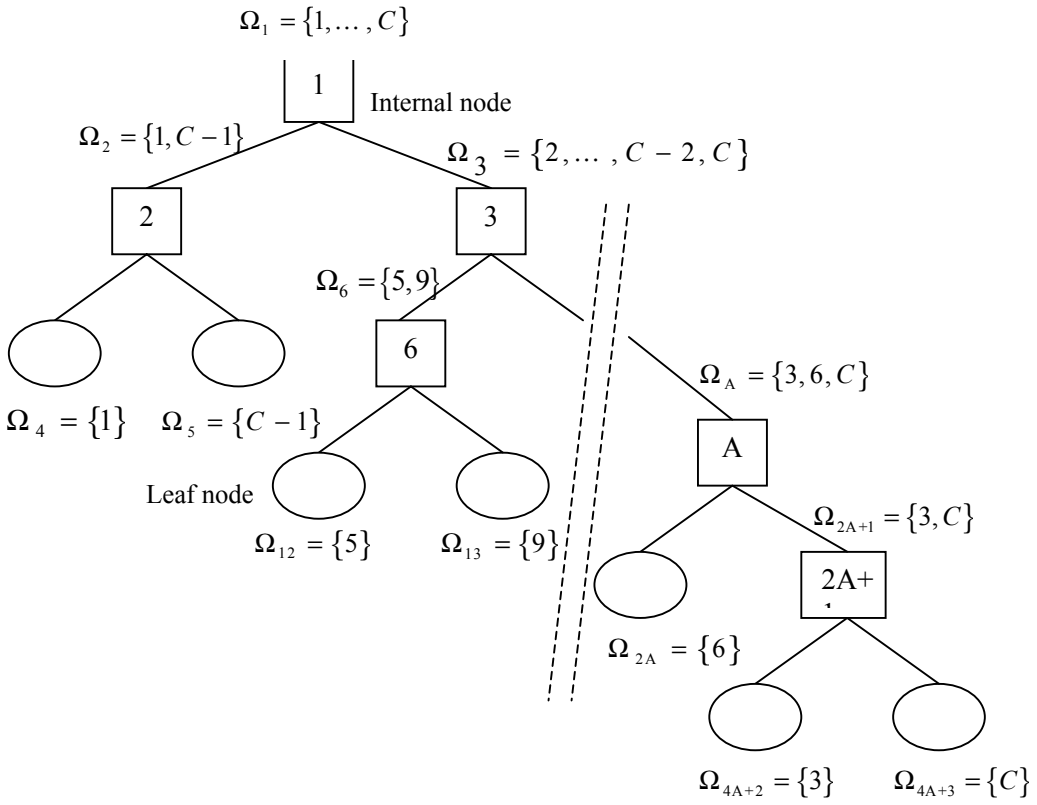


Figure 2.6: An example of a Binary Hierarchical Classifier (BHC) with *C* classes. Each internal node *n* comprises of a feature extractor, a classifier, a left child 2*n*, and a right child 2*n*+1. Each node *n* is associated with a meta-class Ω_n.

However, the number of pairwise classifiers grows as $O(C^2)$. Furthermore, many of these pairwise classifiers are not directly applicable to the

true underlying label, and they pass on the discriminatory burden to the "combiner". These issues motivated the multistage Binary Hierarchical Classifier (BHC) framework that creates a binary tree structured hierarchical classifier with C leaf nodes and C-1 internal nodes [8, 71, 73]. Not only is the number of 2-class problems reduced from $O(C^2)$ to $O(C)$, but the tree structure also allows the more natural and easier discriminations to be accomplished earlier [14]. Figure 2.6 depicts an example of a C-class BHC. New observations are directed down the tree into the leaf nodes representative of the C labels. In this section, both a bottom-up (combining meta-classes) and a top-down (splitting meta-classes) method for building the hierarchical trees are presented. Fisher's linear discriminant function is used as the feature extractor at each internal node of the BHC.

2.5.1 Fisher feature extraction

In order to determine where to split (top-down) or merge (bottom-up) a set of meta-classes, some measure of the distance between the meta-classes in a discriminatory feature space is necessary. Kumar [8] proposed using the Fisher discriminant, which is commonly used in linear discriminant analysis, to accomplish this task. The Fisher discriminant is not only used for constructing the tree, but also as the feature extractor at each internal node of the BHC. For a d-dimensional input space, there is a maximum of $\mathrm{M} = \min\{d, C\text{-}1\}$ Fisher's sample linear discriminants [57]. At each partition of the BHC, let $\{\mu_{\Omega_{2n}}, \mu_{\Omega_{2n+1}}\}$, $\{S_{\Omega_{2n}}, S_{\Omega_{2n+1}}\}$, and $\{P(\Omega_{2n}), P(\Omega_{2n+1})\}$ denote the respective mean vectors, sample

covariance matrices, and priors for the meta-class pair $\{\Omega_{2n}, \Omega_{2n+1}\}$. For $\{\Omega_{2n}, \Omega_{2n+1}\}$, Fisher's linear discriminant function is defined as

$$\mathbf{v}_{\Omega_{2n},\Omega_{2n+1}} = \arg\max_{\mathbf{v}\in\Re^{d\times 1}} \frac{\mathbf{v}'\mathbf{B}\mathbf{v}}{\mathbf{v}'\mathbf{W}\mathbf{v}} \tag{2.13}$$

and Fisher's discriminant measure is defined as

$$D\left(\Omega_{2n}, \Omega_{2n+1}\right) = \frac{\mathbf{v}'_{\Omega_{2n},\Omega_{2n+1}} \mathbf{B}\, \mathbf{v}_{\Omega_{2n},\Omega_{2n+1}}}{\mathbf{v}'_{\Omega_{2n},\Omega_{2n+1}} \mathbf{W}\, \mathbf{v}_{\Omega_{2n},\Omega_{2n+1}}}. \tag{2.14}$$

In (2.13) and (2.14), $\mathbf{B} = \mathbf{B}_{\Omega_{2n},\Omega_{2n+1}} = \left(\mu_{\Omega_{2n}} - \mu_{\Omega_{2n+1}}\right)\left(\mu_{\Omega_{2n}} - \mu_{\Omega_{2n+1}}\right)'$ is the between class covariance matrix and $\mathbf{W} = \mathbf{W}_{\Omega_{2n},\Omega_{2n+1}} = P(\Omega_{2n})S_{\Omega_{2n}} + P(\Omega_{2n+1})S_{\Omega_{2n+1}}$ is the within class covariance matrix. Since two meta-classes are used to define the within and between class covariance, Fisher (1) is the first, and only, sample discriminant. This projection is always 1-dimensional because of the between class covariance matrix is of rank one.

2.5.2 Bottom-up BHC

The Bottom-Up Binary Hierarchical Classifier (BU-BHC) is based on a basic agglomerative clustering algorithm in which each input is initially considered to be a unique cluster and, at each step, the two most "similar" clusters are merged and thereafter considered as a single cluster. In BU-BHC, each class represents one of the initial clusters, so the tree is built by merging the two most "similar" meta-classes until only one meta-class remains. Fisher's discriminant is used as the distance measure D for determining the order in which the classes are merged. A disadvantage of the BU-BHC algorithm is that it is $O\left(C^2\right)$ as the

distance between all pairs of classes must be computed at the very first stage, with each subsequent stage being $O(C)$.

2.5.3 Top-down BHC

In [18], Kumar et al. extended the work of Rose et al. [58-59] and proposed Generalized Associative Modular Learning Systems (GAMLS) for class decomposition through soft partitioning of the training set. In the unsupervised GAMLS framework, decomposition is achieved by softly associating data with different "sub-classes". The GAMLS framework motivated the Top-Down Binary Hierarchical Classifier (TD-BHC) algorithm. In TD-BHC, all the classes are initially considered to be in one meta-class and, as the algorithm iterates, one of the meta-classes Ω_n is partitioned into two meta-classes $(\Omega_{2n}, \Omega_{2n+1})$. In the meta-class Ω_n being partitioned, each class $L \in \Omega_n$ is initially associated (**A**) with Ω_{2n} and Ω_{2n+1} equally. The association is defined as the posterior probability $P(\Omega_\rho | L_i)$ of a class L_i belonging to a particular meta-class Ω_ρ, $\rho \in \{2n, 2n+1\}$ and the "completeness constraint" of the GAMLS framework implies that $P(\Omega_{2n} | L_i) + P(\Omega_{2n+1} | L_i) = 1 \ \forall L_i \in \Omega_n$. After a randomly selected class is selected to be associated with only one of the partitions, the feature extractor $\Psi(\boldsymbol{X}|\mathbf{A})$: $I \to \mathrm{F}\ (\Omega_i, \Omega_j)$ that maximally discriminates between Ω_{2n} and Ω_{2n+1} is sought using the Fisher's linear discriminant function (1). This feature space is used to estimate the log-likelihood of class $L \in \Omega$:

$$L\ (L|\Omega_\rho) = \frac{1}{N_L} \sum_{\mathbf{x} \in X_L} \log p\left(\Psi(\mathbf{x}|\mathbf{A}) \middle| \Omega_\rho\right), \quad \rho \in \{i, j\}, \quad \forall L \in \Omega \tag{2.15}$$

The estimated log-likelihood (2.15) is used as the basis for updating the association rules until a threshold is reached, at which time each class in the meta-class being partitioned is assigned entirely to Ω_{2n} or Ω_{2n+1}.

2.5.4 Combining in BHC

Either a hard combiner or a soft combiner can be used with the BHC framework. The Fisher projection(s) $\Psi(\mathbf{x}|\mathbf{A})$ between two meta-classes $\{\Omega_i, \Omega_j\}$ are used to model the pdfs $p(\Psi(\mathbf{x}|\mathbf{A})|\Omega_k)$, $k = i, j$. The hard combiner follows the same procedure as a decision tree where each new observation starts at the root node and is pushed to the meta-class child into which it is classified until it reaches a leaf node whose label it is assigned [69]. Therefore, the hard combiner requires each observation to go through at most $(C-1)$ classifiers. Alternatively, the soft combiner estimates the posteriors of the labels $P(L_i|\Psi_{ij}(\boldsymbol{x}))$ i=1,…,C and then applies the MAP rule to assign each observation a label. The posteriors for each label are estimated by Baye's rule by taking the product of posterior probabilities of all the internal node classifiers on the path to that particular label's leaf node [8]. As a result, when the soft combiner is used, each observation is processed by exactly $(C-1)$ classifiers.

An important, yet sometimes overlooked, issue central to the classification process is the dependency upon the labeled data for training. The quantity of training data, with respect to the dimensionality, is critical for accurate parameter estimation. Furthermore, even when an adequate amount of training data is available, if the population on which the classifier is being applied is not properly represented, then a high level of classification accuracy cannot be expected.

Chapter 3: Adaptive Best-Basis Bayesian Hierarchical Classifier

An important application area for the dimensional, narrow spectral band, hyperspectral data is land cover classification. Data from hyperspectral sensors characterize the response of targets (spectral signatures) with greater detail than traditional sensors and thereby can improve discrimination between targets [8, 9, 51, 72]. The labeled ground truth for training are typically limited both in quantity of data $|\boldsymbol{X}|$ and variety of sites at which they are collected. These data are used to estimate the label-conditional probability density functions $P(x_1, x_2, ..., x_d | L_i), i = 1, ..., C$ that are essential for determining the most appropriate land cover label. Because the data are of such high dimension, and there is only a limited quantity of ground truth, the "small sample size" problem is prevalent and must be dealt with foremost in the classification process. A new approach is proposed in this chapter that utilizes domain knowledge, which is automatically discovered from the data, to combat the "small sample size" problem.

While many difficult classification problems, including land cover classification with hyperspectral data, involve a high dimensional input space and a large number of candidate labels, it is necessary to reduce the dimensionality of the input space (I) when only limited quantities of training data are available because of the Hughes phenomena and the "curse of dimensionality" [1, 3, 6, 7, 13, 56]. To improve classification results beyond those obtainable by using a "mean distance classifier", more information is needed and, because of this fact,

almost all conventional statistical approaches use the sample covariance matrices $S = \frac{1}{|\boldsymbol{X}|-1}\sum_{j=1}^{|\boldsymbol{X}|}\left(\boldsymbol{X}_j - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_j - \bar{\boldsymbol{X}}\right)'$. For example, Fisher's linear discriminant function is defined in terms of the within class covariance matrix $\mathbf{W}_{i,j} = P(i)\Sigma_i + P(j)\Sigma_j$ and the between class covariance matrix $\mathbf{B}_{i,j} = \left(\mu_i - \mu_j\right)\left(\mu_i - \mu_j\right)'$. Any classifier using Fisher's linear discriminant function (2.13) or discriminant distance measure (2.14) would require the inversion of the within class covariance matrix $\mathbf{W}_{i,j}$. For the covariance matrix of *d*-dimensional data, there are $d\left(d+1\right)/2$ parameters to estimate and, at a minimum, there needs to be $d+1$ observations to ensure a non-singular/invertible sample covariance matrix [57]. Therefore, to utilize the improved spectral signature estimates provided by hyperspectral sensors compared to traditional sensors, the increased dimensionality must be taken into consideration.

Numerous studies have considered what the minimum number of training samples should be, in relation to the dimensionality, for trustworthy estimation of a covariance matrix [3, 5, 19, 26]. In general, literature recommends having 4-10 times the number of observations as the dimensionality for linear classifiers. While quadratic classifiers may perform better than linear classifiers in certain situations, rather than a linear relationship, the recommended number of observations is related to the square of the dimensionality [24, 25, 27]. This relationship is even worse for non-parametric classifiers, where it has been estimated that the required quantity of training data increases exponentially as the dimensionality increases in order to accurately estimate the multivariate densities

[3]. Regardless of the actual classifier being used, while hyperspectral data provide a greater opportunity for discrimination between different land cover types, the problems with limited training data in relation to dimensionality become more relevant and can have a significant impact on classification accuracies [19, 25, 53]. This dilemma will be referred to hereafter as the "small sample size" problem.

A significant amount of research has focused on transforming the input space into a reduced feature space that accurately discriminates between the classes in a fixed output space, thus helping mitigate the small sample size problem. To a lesser extent, decomposition of the output space (reducing the number of candidate labels) has also been considered in view of the fact that the decision boundaries for a reduced output space should be easier to learn and model than the original output space. As a shortcoming, traditional approaches fail to capitalize on the flexibility gained by transforming the feature space and the output space simultaneously or the domain knowledge specific to hyperspectral data. The simultaneous transformation of the input space and the output space in search of a good feature space had not been rigorously explored until Kumar et al. decomposed a $(C > 2)$-class problem into a binary hierarchy of $(C-1)$ simpler 2-class problems, each with its own feature space and classifier that are independently trained using the labeled training data [8, 10, 11, 15-18]. However, that classification framework, the Binary Hierarchical Classifier, does not address the classifier's dependency upon an adequate quantity of training data or make full use of the domain knowledge specific to hyperspectral data. A new

approach, used within a classification framework extended from the work of Kumar et al., is proposed in this dissertation that automatically makes use of domain knowledge to combat the "small sample size" problem.

3.1 Related Work With Limited Training Data

The "small sample size" problem and the associated degradation of classification accuracy are widely acknowledged. Generally, previous work that addresses the small sample size problem follows one of three general approaches [30, 38]. The first, parameter stabilization techniques, try to improve the parameter estimates directly. Some other methods seek to avoid the problem by improving the ratio of training data to dimensionality. While these two general approaches attempt to stabilize the parameter estimates, the third method, using an ensemble of classifiers, attempts to improve classification by considering a combination of "weaker" classifiers. The three approaches are reviewed in the following sub-sections.

3.1.1 Parameter stabilization techniques

A widely used technique for stabilizing the estimated covariance matrix directly consists of weighting the sample covariance matrix as well as "supplemental" matrices and is generally referred to as "regularization" or "shrinkage". In particular, when the sample covariance matrix is "shrunk" towards the identity matrix, it is referred to as the ridge estimate of the covariance matrix and is the basis for regularized discriminant analysis [29, 38]. Similar to the ridge estimates, the covariance matrix can be "shrunk" toward values other

than diagonal weights, and there are hybrids that give weights to sample covariance (normal and diagonal) and a pooled covariance (normal and diagonal) matrix [12, 32]. When training data are limited, using the pooled estimate of the covariance can yield better results than using the class dependent covariance matrices [29]. However, if the class dependent covariance matrices are accurately estimated, they provide important discrimination not available from a pooled estimate [25]. Additionally, while the variance of the parameter estimates has been reduced, the bias of the parameter estimates may increase dramatically, depending upon the differences between the true parameter values and those towards which they are being shrunk. Furthermore, when the covariance matrix is severely shrunk toward the identity matrix, the classifier would simply assign each new observation to the class whose mean vector is closest in terms of Euclidean distance.

Rather than stabilizing the covariance matrix directly, the pseudo-inverse of the covariance matrix can be used instead of the true inverse. Pseudo-inversion, based upon singular value decomposition, utilizes the non-zero eigenvalues of the covariance matrix [34, 38]. However, in addition to poor performance when the ratio of training data to dimensionality is very small, the pseudo-inverse has a "peaking effect" in its performance. It has been shown that the pseudo-inverse performs best when $|\boldsymbol{X}| = d/2$ and that the performance degrades as $|\boldsymbol{X}|$ approaches d [30, 37]. Thus, the pseudo-inverse has the undesirable characteristic that there are situations where, counter-intuitively, if the quantity of training data is increased, the classification accuracy could actually be

reduced. Therefore, if possible, it would be advantageous to exploit the "sweet spot" characteristic of the pseudo-inverse. Additionally, for $|\boldsymbol{X}| > d$, the pseudo-inverse is the same as the traditional inverse, even though the covariance matrix is most likely poorly estimated.

3.1.2 Improve ratio of training data to input dimensionality

Because it is often not possible to acquire more training data, methods have attempted to improve the ratio of labeled data to dimensionality either by transforming the input space into a reduced feature space or by artificially increasing the quantity of labeled data. One such method is feature extraction. These methods, which include Principle Component Analysis, Fisher's linear discriminant function, and MNF transforms, can be used to project the original data into an adequately reduced feature space [23, 34, 100]. The transformations, however, are data dependent. Furthermore, not only can the limited data result in poor estimation of the transformations, but also these techniques do not address estimation of the covariance matrix in the original feature space with limited data, which is generally required for estimating the projections. Lastly, a great deal of interpretability is lost when the original features are no longer being used. This is important for many applications.

Alternatively, the selection of critical features is often used to reduce the size of the input space. Interpretability is preserved because a subset of the features is used instead of the original feature space such that the features can still be directly related to the information content of the data [39]. While this method combats the "small sample size" problem, it should be noted that sample

estimates of the "effectiveness" of the feature subsets are being used. Therefore, poor estimates due to a small sample size could also be reflected in the subset [19, 52]. Also, this can be computationally very expensive because all subsets must be investigated to find the "best" subset (there are 2^d subsets), so it is only practical when the original dimensionality d is small. Furthermore, aside from the difficulties associated with determining the optimal subset size and performing feature selection, this approach would not be beneficial for combating the small sample size problem unless the size of the best subset were small compared to the original dimensionality of the data. Finally, much like feature extraction techniques, any feature selection technique that utilizes the covariance matrix in the original d dimensional feature space must address its estimation in the context of limited training data.

Assuming it is not possible to acquire more training data, if the ratio of quantity of training data $|\boldsymbol{X}|$ to dimensionality d is to be improved by increasing $|\boldsymbol{X}|$, then it must be accomplished artificially. One such method augments the original training data by using pseudo-labeled data, which are usually identified by a classifier constructed from the original training data. Specific techniques for identifying and augmenting the existing training data with unlabeled data already exist and have been shown to enhance strictly supervised classification [4, 20, 41-46]. However, not only can convergence of the updating scheme be problematic, but the method is also affected by selection of the initial training samples and by outliers. Therefore, even if the limited training data are sufficient to design a classifier for pseudo-data identification, the poor initial parameter estimates due

to the small sample size problem may lead to incorrect pseudo-labels and poor updates.

3.1.3 Subsampling/Combining schemes

The ensemble of "weaker" classifiers approach does not explicitly address the diminished accuracy of an individual classifier. Rather, multiple classifiers are designed with the hope that an assessment of the group's aggregated output will result in higher accuracies. These methods, such as simple random sampling without replacement, bagging, and arcing, involve selecting subset samples for the original data and generating a classifier specific to each sub-sample [31, 33, 35, 36, 49, 70]. While the aggregate output of the ensemble can be combined in many different ways, the two most popular approaches are the 'voting' method, where the pixel is assigned the class label that occurs the maximum number of times, and the MAP (maximum a posteriori probability) method, where the pixel is assigned the class label that corresponds to the maximum estimated posterior probability [6, 50, 54, 55]. However, because these methods are based upon altering the sample distribution by selecting subsets of the available data, and the context being considered here already involves problems with limited data, the sub-sampling approach is very problematic and may not even be an option because the quantity of training data is effectively reduced for each classifier in the ensemble. Additionally, when the quantity of data does allow sub-sampling, these methods may still have problems since the degradation in individual classifier performance (because of the reduced data) cannot be compensated for by the gains from using an ensemble [70]. Furthermore, due to the large quantity

of data involved in land cover classification, the additional computation for an ensemble may be prohibitive for certain applications such as on-board target identification via sensors on unmanned platforms.

3.2 BEST BASIS BAYESIAN HIERARCHICAL CLASSIFIER (BB-BHC)

While traditional approaches fail to capitalize on the domain knowledge specific to hyperspectral data or the flexibility gained by transforming the feature space and the output space simultaneously, the new methodology developed in this study exploits both of these practices, while specifically addressing the small sample size problem. This method decomposes a $(C > 2)$-class problem into a binary hierarchy of $(C-1)$ simpler 2-"meta-class" problems, each with its own feature space and classifier that are independently trained using the labeled training data. The size of the feature space is dependent upon the quantity of labeled data available, specific to each 2-meta-class problem. While other methods have sought to combat the small sample size problem by using parameter stabilization techniques, improving the ratio of training data to dimensionality, or by sub-sampling and combining schemes, this method focuses on a feature reduction rule specific to hyperspectral data that avoids the drawbacks characteristic of traditional feature selection or feature extraction. Bands are aggregated in a manner that not only allows for an analytic evaluation and intuitive understanding of the feature space, but also adds domain knowledge in the process. Additionally, this methodology actually reduces the quantity of data that must be stored, since there is no need to retain all of the original features.

From the domain knowledge in this field, it is already known that the original input features, the bands of hyperspectral data, that are "spectrally close" to each other tend to be highly correlated. While the original BHC classification framework accomplished the simultaneous transformation of the input space and the output space in search of a good feature space, using a discriminant function on the original input space failed to exploit the correlation structure of the bands or the fact that the bands are ordered. It also failed to leverage this information with respect to the quantity of training data available. This domain knowledge must be utilized in order to take advantage of the fact that hyperspectral sensors characterize the spectral signature of targets with greater detail than traditional sensors. The new "adaptive best-basis" method utilizes a best-basis band-combining algorithm in conjunction with the BHC framework. This approach has not been previously investigated and it both utilizes the domain knowledge specific to hyperspectral data and has value in terms of acquiring additional domain knowledge.

3.2.1 Best-Basis and the Binary Hierarchical Classifier framework

Unlike the BHC, the BB-BHC performs a band-combining algorithm prior to the partitioning (TD-BB-BHC) or combining (BU-BB-BHC) of meta-classes. Unlike the original BHC algorithm, this algorithm partitions the spectrum and maximizes the discrimination among classes in two stages. It is assumed that highly correlated ordered bands "behave" most similarly, relative to other combinations of bands, so that combining them has the least detrimental effect on the potential for discrimination among classes. To capitalize on this characteristic

of the data, this feature extraction technique is comprised of two independent stages. In the first stage, a band reduction algorithm intelligently generates a set of customized bands to discriminate among the classes. Then, in the second stage, once the number of "group" bands is small enough with respect to the amount of training data, the algorithm maximizes the discrimination between the classes. The TD-GLDB (partitioning) and BU-GLDB (combining) algorithms of Kumar et al., which also address feature extraction specifically for hyperspectral data, utilize the ordering of the bands and yield excellent discrimination [11, 17]. However, they were not intended to be used in a two-stage feature extractor and have proved to be very computationally intensive due to repetitive calculation of a discrimination function for all candidate splits (TD-GLDB) or merges (BU-GLDB). Additionally, the quality of the discrimination functions, and thus the structure of the resulting feature space, is affected by the amount of training data and this critical issue is not addressed.

Performing the feature extraction in two stages allows for the band-combining algorithm to focus more on preserving the most distinct characteristics of the data while discovering domain knowledge without constraining the results to also account for the discrimination between classes. Because the correlation between bands varies among classes, the band reduction algorithm must be class dependent. In order to estimate the "correlation" for a group of bands (meta-bands) $B = [p:q]$ over a set of classes Ω, the correlation measure $Q(B)$ is defined as the minimum of all the correlations within that group:

$$Q(B) = \min_{L_k \in \Omega} \min_{p \le i < j \le q} Q_{i,j}^{L_k} = \min_{L_k \in \Omega} \min_{p \le i < j \le q} \frac{S_{i,j}^{L_k}}{\sqrt{S_{i,i}^{L_k} S_{i,j}^{L_k}}} \tag{3.1}$$

Also, $S_{i,j}^{L_k}$ is defined as the row *i*, column *j* element of the sample covariance matrix for class L_k. The correlation measure (3.1) is used to determine which set of adjacent meta-bands should be merged at each successive step of the algorithm. Therefore, this band-combining algorithm works with the class specific correlation matrices and is not hindered by the additional objective of discriminating between the classes. BB-BHC was designed to either construct a "best-basis" for the entire BHC structure or have the basis determined for each partition of the BHC.

3.2.2 Adaptive feature space for the BB-BHC

The BB-BHC framework, like other statistical approaches, relies on the inversion of class-specific covariance matrices for all classes. The Fisher discriminant, which is used for constructing the tree and as the feature extractor at each internal node of the BHC, requires the inversion of the within class covariance matrix $\mathbf{W}_{\Omega_{2n},\Omega_{2n+1}}$ and, when the data are *d*-dimensional, the training samples must include at least *d*+1 independent samples in order for the sample covariance matrix to be nonsingular. Furthermore, even if there are $|\boldsymbol{X}| > (d+1)$ training samples, the $d(d+1)/2$ parameter estimates may be very poor. Initial research with the BB-BHC indicated that while the BB-BHC has comparable classification performance to the traditional BHC classifier when there are large quantities of training data, the degradation in performance for the BHC was much more pronounced than that of the BB-BHC when the quantity of training data was

reduced. This motivated the design of an adaptive feature space, using the best-basis algorithm at each partition of the BHC, whose size is directly dependent upon the quantity of labeled training data available at each partition. Therefore, rather than using a threshold on the correlation measure to determine whether bands or group-bands should be merged, the new algorithm focuses on preserving as many of the original bands as possible, dependent upon an adequate amount of training data. If band reduction is necessary, the band-combining algorithm ensures that the least amount of discriminatory information is lost to achieve a satisfactory ratio of training data to dimensionality. Because literature recommends different thresholds for the minimum $\alpha_{\text{ratio}} \leq \frac{|\boldsymbol{X}|}{d}$, it was allowed to be a user-defined input. In pseudo-code, the adaptive band-combining algorithm that is performed before partitioning or merging meta-classes is:

1. $d^* = \min\left(d, \frac{|\boldsymbol{X}|}{\alpha_{\text{ratio}}}\right)$
2. Initialize $l = 0$, $N_0 = d$, and $B_l^{\,k} = [k:k]$, $\forall\ k = 1,...,d$
3. If $N_l > d^*$ then continue. Otherwise, stop.
4. Find the best pair of band to merge: $K = \arg\max_{k=1,...,N_l-1} Q\left(B_l^{\,k} \cup B_l^{\,k+1}\right)$
5. Update band structure:
   - $l = l+1$, $N_l = N_{l-1} - 1$
   - If $K > 1$ then $B_l^{\,k} = B_{l-1}^{\,k}$, $\forall\ k = 1,...,K-1$
   - $B_l^{\,K} = B_{l-1}^{\,K} \cup B_{l-1}^{\,K+1}$
   - If $K < N_l$ then $B_l^{\,k} = B_{l-1}^{\,k+1}$, $\forall\ k = K+1,...,N_l$
6. Return to step 3.

3.2.3 Best-Basis and Limited Data

When constructing a basis specific to each split in the BB-BHC, the quality of the correlation measure, computed from the class condition covariance matrices, is dependent on the quantity of training data available to estimate the meta-class covariance matrices. This will become even more relevant for the "low branches" of the BB-BHC as the meta-classes become smaller in cardinality and the amount of training data strictly decreases. In particular, the class specific correlation matrices $Q_{i,j}^{L_k} = \frac{S_{i,j}^{L_k}}{\sqrt{S_{i,i}^{L_k} S_{i,j}^{L_k}}}$ are required in (3.1) to estimate the correlation measure $Q(B)$. However, if the label specific S^{L_k} covariance matrices are not suitable for inversion, failure to stabilize their estimates before constructing the basis unsatisfactorily passes the disadvantage of the small sample size from the estimate of Fisher's disciminant and linear discriminant function to the basis construction. Therefore, the label specific sample covariance matrices need to be stabilized. The ancestor sample covariance matrix S^{Anc} is defined as the sample covariance matrix which is estimated from at least $\alpha_{\text{ratio}} |\boldsymbol{X}|$ observations and is most closely related to L_k based upon the BB-BHC structure. Because the trees are constructed using both top-down and bottom-up approaches, the search for S^{Anc} is performed uniquely for each type. In the top-down framework, if meta-class Ω_k is being considered for partitioning, then $S^{\Omega_k} = \sum_{L_i \in \Omega_k} P(L_i) S^{L_i}$ is the first candidate for S^{Anc}. However, if $\left|\boldsymbol{X}_{\Omega_k}\right| < \alpha_{\text{ratio}} d$, then the BB-BHC tree structure is climbed in search of a meta-class where $\left|\boldsymbol{X}_{\Omega_k}\right| \geq \alpha_{\text{ratio}} d$. With the bottom-up framework, if $\{\Omega_{2n}, \Omega_{2n+1}\}$ are being considered for agglomeration, the first candidate for S^{Anc} is

$S^{\text{Pooled}} = P\left(\Omega_{2n}\right)S^{\Omega_{2n}} + P\left(\Omega_{2n+1}\right)S^{\Omega_{2n+1}}$. However, because the BB-BHC is being constructed bottom-up, the structure cannot be climbed in search of a suitable S^{Anc}. Therefore, if $\left|\boldsymbol{X}_{\Omega_i+\Omega_j}\right| < \alpha_{\text{ratio}} d$, than $S^{\text{Anc}} = \sum_{i=1}^{C} P\left(L_i\right)S^{L_i}$. Note that this estimate for S^{Anc} is used, even when the total quantity of training data available is less than $\alpha_{\text{ratio}} d$. The stabilized estimates of the label specific covariance matrices are defined as:

$$\hat{S}^{L_k} = \frac{\left|\boldsymbol{X}_{\Omega_L}\right| S^{L} + \left|\boldsymbol{X}_{\Omega_{\text{Anc}}}\right| S^{\text{Anc}}}{\left|\boldsymbol{X}_{\Omega_L} + \boldsymbol{X}_{\Omega_{\text{Anc}}}\right|} \qquad (3.2)$$

These stabilized class dependent covariance matrices (3.2) are used to estimate the correlation measure (3.1).

3.3 Application of Adaptive BB-BHC to Classification of Hyperspectral Data

Hyperspectral data acquired over two sites were used to evaluate the proposed algorithms. The adaptive BB-BHC algorithm was evaluated on airborne hyperspectral data acquired over Bolivar Peninsula, located at the mouth of Galveston Bay, Texas and NASA's John F. Kennedy Space Center (KSC) in Florida.

3.3.1 Bolivar Peninsula

Bolivar Peninsula, part of the low relief barrier island system on the Texas Gulf coast, is an area of interest due to the shoreline changes that occur as a result of sedimentary processes such as high-energy wave and low-energy tidal and wind processes. The University of Texas Bureau of Economic Geology closely

monitors shoreline dynamics in this area. An overview map with a mosaic image of the two test sites considered in this research are depicted in Figure 3.1 The area contains two general vegetation types, wetlands and uplands, with the marsh area further characterized in terms of sub-environments defined by the wetland maps. For classification purposes, 11 classes representing the various land cover

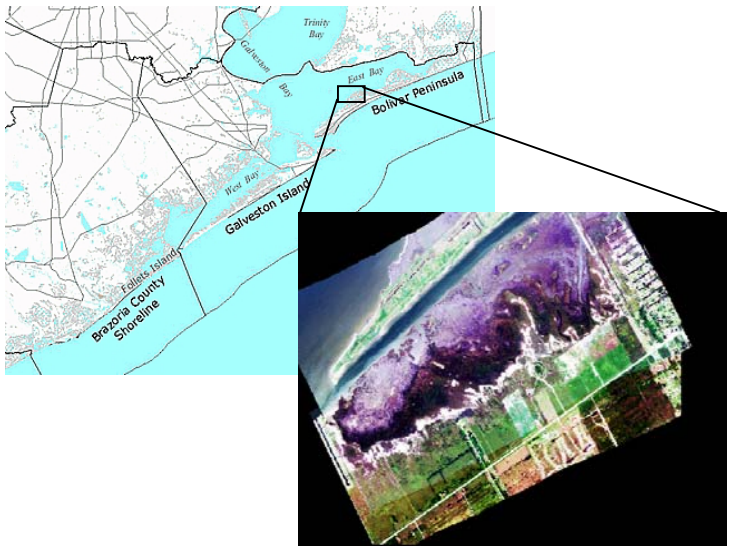


Figure 3.1: Overview image and hyperspectral (HyMap) images of Area 1 and Area 2 at Bolivar Peninsula

types that occur in this environment have been identified for the site. These include: water, wetlands (low proximal marsh, high proximal marsh, high distal marsh, and pure salicornia) and uplands (trees, general uplands, two agricultural classes, sand flats, and a transition zone) [48]. The low proximal marsh corresponds to tidal flats comprised of *Spartina alterniflora*, which experiences frequent flooding. The high proximal marsh, which is composed of a mixture of

Spartina alterniflora and *Salicornia virginica*, is flooded less frequently and has more continuous vegetation cover. The high distal marsh, which is inundated even less frequently than the proximal marshes, contains *Spartina patens*, *Salicornia virginica* and *Juncus roemerianus*. Adjacent to the high distal marsh, a small highly saline region of sand flats surrounded by pure *Salicornia virginica* delineates the boundary between the wetlands and uplands. The quantity of ground truth available for each class at the two test sites is given in Table 3.1.

Class	Name	Area1	Area2	Total Obs
1	Water	1019	4529	5548
2	Low Proximal Marsh	1127	647	1774
3	High Proximal Marsh	910	1083	1993
4	High Distal Marsh	752	494	1246
5	Sand Flats	148	112	260
6	Ag 1 (pasture)	3073	2454	5527
7	Trees	222	238	460
8	General Uplands	704	534	1238
9	Ag 2 (bare soil)	1095	1127	2222
10	Transition Zone	114	210	324
11	Pure Silicornia	214	129	343
	TOTAL	9378	11557	20935

Table 3.1: Number of observations per class for Bolivar Peninsula at two different areas used for testing

The topography of these areas is mainly a function of sedimentary processes such as high-energy wave and low-energy tidal and wind processes. As a result, the frequency of the inundation, soil salinity, and vegetation cover all

depend on this topography [48]. HyMap (Hyperspectral Mapper) acquired data over Bolivar Peninsula on September 17, 1999, at a spatial resolution of 5m. HyMap, an airborne hyperspectral optical sensor developed in Australia, acquired the data in 126 bands with almost contiguous spectral coverage over the wavelength range of 0.44-2.48 [47]. For this particular acquisition, only four bands {63, 64, 95, 126} were dominated by water absorption, resulting in a low signal to noise ratio, and therefore not considered in subsequent analysis. In this case, the practical dimensionality *d* is 122.

3.3.1.1 Classification accuracies across decreasing sampling percentages

Multiple experiments were performed on data from Site 1 using stratified (class specific) sampling at percentages of: 75, 50, 30, 15, 5, and 1.5. The quantity of ground truth for each class, indicated by sampling percentage, is listed in Table 3.2. It is interesting to note that even at the sampling percentage of 75,

Class	Name	Total Obs	75%	50%	30%	15%	5%	1.5%
1	Water	1019	764	510	306	153	51	15
2	Low Proximal Marsh	1127	845	564	338	169	56	17
3	High Proximal Marsh	910	683	455	273	137	46	14
4	High Distal Marsh	752	564	376	226	113	38	11
5	Sand Flats	148	111	74	44	22	7	2
6	Ag 1 (pasture)	3073	2305	1537	922	461	154	46
7	Trees	222	167	111	67	33	11	3
8	General Uplands	704	528	352	211	106	35	11
9	Ag 2 (bare soil)	1095	821	548	329	164	55	16
10	Transition Zone	114	86	57	34	17	6	2
11	Pure Silicornia	214	161	107	64	32	11	3
		9378	**7035**	**4691**	**2814**	**1407**	**470**	**140**

Table 3.2: Classes for Bolivar Peninsula, Site 1, and the quantity of training data per class by sampling percentage

the amounts of training data for classes 5 and 10 are still less than *d* (sand flats $\left|\boldsymbol{X}_{L_5}\right| = 111$ and transition zone $\left|\boldsymbol{X}_{L_{10}}\right| = 86$). For all sampling percentages except for 1.5%, the value of $\alpha_{\text{ratio}} = 5$ was used. At 1.5%, the value was reduced to $\alpha_{\text{ratio}} = 1.5$ to ensure there were at least two observations per label L_i. Ten experiments using simple random sampling, were performed at each percentage for the bottom-up and top-down frameworks of the traditional BHC [TD-BHC, BU-BHC], the traditional BHC using the pseudo-inverse for tree construction (estimating Fisher's discriminant as a distance measure) and feature extraction (calculating Fisher's linear discriminant function) [TD-P-BHC, BU-P-BHC], and the adaptive best-basis BHC [TD-BB-BHC, BU-BB-BHC]. Additionally, results from two "nearest distance to mean" classifiers are also presented, one using the Euclidean distance [EUCL-D] and the other using the sum of squared deviations from the respective class means [SQRD-D]. These two classifiers are intended to establish a baseline for the results one could expect by simply assigning labels to pixels based upon their proximity to the mean spectral signatures of the class training data, regardless of their variance structure. The results are presented in Figure 3.2.

By adapting the size of the feature space to reflect the amount of training data available, a high level of classification accuracy is preserved for an extremely small sample size. At the 50$^{\text{th}}$ percentage of sampling, the value typically used to separate data sets into training and testing, the BB-BHC actually performs slightly better that the BHC. Importantly, even though using the pseudo-inverse does not improve the results at the 50$^{\text{th}}$ percentage, because there

are at least $d+1$ observations per L_i, the results indicate that the covariance matrices are still poorly estimated, so the estimates of corresponding inverses will also be poor. Not only do the BB-BHC methods perform the best at every

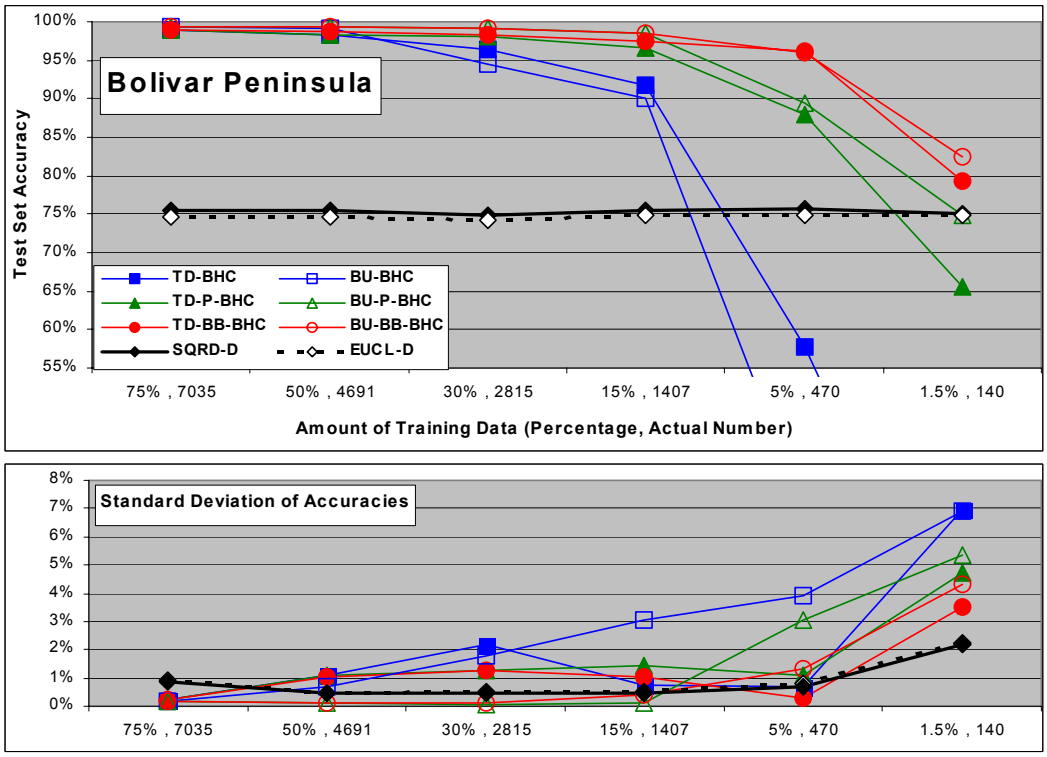


Figure 3.2: Classification (test set) accuracies for Bolivar Peninsula

sampling percentage relative to the other TD and BU classifiers, but also the accuracies are generally more stable (smaller variation of the classification accuracies). This is important because different investigators may select somewhat different ground truth. Combating the limited training data by using the correlation matrix for feature reduction helps retain the information necessary for successful land cover prediction. Classification accuracies of over 80% were

still achieved, even with only 140 total labeled samples and only 2 labeled pixels available for classes 5 (sand flats) and 10 (transition zone). Furthermore, retaining the use of the class specific covariance matrices helped increase the accuracies dramatically for these experiments. Overall, the adaptive BB-BHC classifiers use the retained covariance information to achieve much higher accuracies, and comparable variation, at every level except the 1.5% sampling rate, at which point the accuracies are still higher for the adaptive BB-BHCs but the variation is worse.

3.3.1.2 Domain knowledge and image evaluation

Classified images were examined to study how well scene information was retained as the quantity of training data was reduced. A classified image obtained using all of the available data was compared to images obtained at each of the sampling percentages (75, 50, 30, 15, 5, and 1.5) for both the adaptive BB-BHC (%Y) and the BHC using the pseudo-inverse (%N).

Examining the results for the BU classifiers [Appendix A], the quality of the images obtained by both techniques remains high for sampling rates $\geq 15\%$ [Figure A.1]. However, at the 5% sampling rate, the adaptive BB image [Figure A.2] is markedly better than that obtained using pseudo-inversion. While the overall results for the adaptive BB still appear better than it's pseudo-inverse counterpart at the 1.5% sampling rate, they both have deteriorated markedly from the results obtained using 100% of the available ground truth. Analyzing the images obtained with the TD classifiers suggests similar results except that the differences at the 5% sampling rate are even more distinctive than those with the

BU classifiers and is more comparable to the image obtained using 3 times as much data and pseudo-inversion [Figure A.4]. This result that is also suggested by a smaller difference in classification accuracies for the BU classifiers at that sampling percentage [Figure 3.2].

3.3.2 Kennedy Space Center

The wetlands of the Indian River Lagoon system, located on the western coast of the Kennedy Space Center (KSC), are part of the closely monitored Merritt Island National Wildlife Refuge. Over 1,000 plant species have been identified on the 140,000-acre Refuge and 16 of the more than 500 species of wildlife have been federally listed as either threatened or endangered. Accurate classification and mapping of upland vegetation is important for monitoring this critical habitat for species of waterfowl and aquatic life. An overview image depicting the two test sites for research in the area is shown in Figure 3.3. The test sites for this research consist of a series of impounded estuarine wetlands of the northern Indian River Lagoon (IRL) that reside on the western shore of the Kennedy Space Center. The impoundments were created during the 1950's and 1960's for the purpose of mosquito control. The marshes along the IRL contain both high and low marsh communities. The three dominant marsh groups that comprise the high marsh communities are cabbage palm savanna, sand cordgrass, and black rush. The cabbage palm savanna consists of isolated canopies of Cabbage Palm (*Sabal palmetto*) and a graminoid layer of sand cordgrass (*Spartina bakerii*) and black rush marsh (*Juncus roemerianus*). Salt tolerant grasses and halophytes dominate the low marsh communities. The primary salt

tolerant grass is *Distichilis spicata.* Halophytes typically include *Batis maritima* and *Salicornia virginica* Upland vegetation is also mapped, as it is adjacent to the impounded wetlands. The majority of the upland vegetation at KSC is oak scrub and saw palmetto scrub. Other upland communities include slash pine (*Pinus elliottii*) and hardwood swamps that are dominated by deciduous trees such as Red Maple (*Acer rubrum*). Dense hammocks of Cabbage Palm (*S. palmetto*) and Live Oaks (*Quercus virginiana*) are also common [40]. Classification of land cover for this environment is difficult due to the similarity of spectral signatures

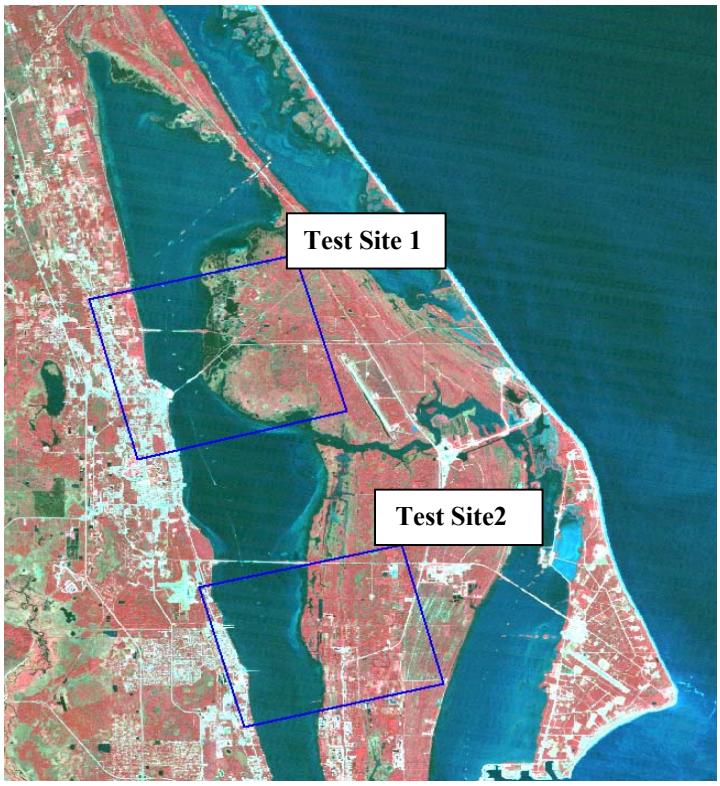


Figure 3.3: An overview image of the two test sites at Kennedy Space Center; Cape Canaveral Florida.

for certain vegetation types.

The NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) spectrometer was used to acquire data over the Kennedy Space Center, Florida on March 23, 1996. AVIRIS acquires data in 224 bands of 10 nm widths in the reflected visible and near infrared spectrum (400 - 2500 nm). The data, acquired

Class	Name	Area 1	Area 2	Total Obs
1	Scrub	761	422	1183
2	Willow Swamp	243	180	423
3	CP Hammock	256	431	687
4	CP/Oak Hammock	252	132	384
5	Slash Pine	161	166	327
6	Oak/Broadleaf Hammock	229	274	503
7	Hardwood Swamp	105	248	353
8	Graminoid Marsh	420	453	873
9	Spartina Marsh	520	241	761
10	Cattail Marsh	396		396
11	Salt Marsh	419	156	575
12	Mud Flats	447		447
13	Water	927	1392	2319
14	Slash Pine (Dense)		393	393
15	Citrus		269	269
16	Slash Pine/Oak Hammock		142	142
	TOTAL	5136	4899	10035

Table 3.3: Number of observations per class for Cape Canaveral at two different areas used for testing

from an altitude of approximately 20km, have a spatial resolution of 18 m [42]. Forty-eight bands collected by the AVIRIS sensor are dominated by water absorption, which results in a low signal noise ratio, and are not considered in

subsequent analysis. In this case, *d*=176 bands of AVIRIS data are used {bands 1-4, 102-116, 151-172, and 218-224 have been removed}. For classification purposes, 16 classes representing the various land cover types that occur in this environment have been defined for the two combined test sites, with some of the classes not present in each specific site. The amount of ground truth acquired for each class is given in Table 3.3. KSC is a more difficult application area than Bolivar Peninsula, due to the complexity of the land cover, the mixed classes, and the spatial resolution of the data.

Class	Name	Total Obs	75%	50%	30%	15%	5%	1.5%
1	Scrub	761	571	381	228	114	38	11
2	Willow Swamp	243	182	122	73	36	12	4
3	CP Hammock	256	192	128	77	38	13	4
4	CP/Oak Hammock	252	189	126	76	38	13	4
5	Slash Pine	161	121	81	48	24	8	2
6	Oak/Broadleaf Hammock	229	172	115	69	34	11	3
7	Hardwood Swamp	105	79	53	32	16	5	2
8	Graminoid Marsh	420	315	210	126	63	21	6
9	Spartina Marsh	520	390	260	156	78	26	8
10	Cattail Marsh	396	297	198	119	59	20	6
11	Salt Marsh	419	314	210	126	63	21	6
12	Mud Flats	447	335	224	134	67	22	7
13	Water	927	695	464	278	139	46	14
		5136	**3852**	**2572**	**1542**	**769**	**256**	**77**

Table 3.4: Classes for Cape Canaveral and the quantity of training data per class by sampling percentage

The primary research site at Kennedy Space Center is Test Site 1 and, for each of the 13 identified classes at that site Table 3.4 indicates the quantity of available training data for the classes at each of the sampling percentages that were used for the experiments. While the entire scene of Test Site 1 includes a

portion of the Indian River Lagoon (IRL) and the IRL's west embankment, it is important to note that the area of interest is the western shore of the Kennedy Space Center and that the west embankment, as well as a large section of the IRL itself, are masked not being analyzed. The AVIRIS data being used have a spatial resolution of 18m and a practical dimensionality *d* of 176.

3.3.2.1 Classification accuracies across decreasing sampling percentages

Multiple experiments were performed on this site using stratified (class specific) sampling at percentages of: 75, 50, 30, 5, and 1.5. At the sampling percentage of 75, the amounts of training data for classes 5, 6, and 7 are still less than *d* and, at the 50th percentage, so are classes 2, 3, and 4. A threshold $\alpha_{\text{ratio}} = 5$ was used for all sampling percentages except for 1.5 ($\alpha_{\text{ratio}} = 1.5$). Ten experiments, using simple random sampling, were performed at each percentage for the bottom-up and top-down frameworks of the traditional BHC [TD-BHC, BU-BHC], the traditional BHC using the pseudo-inverse for tree construction (estimating Fisher's discriminant) and feature extraction (calculating Fisher's linear discriminant function), [TD-P-BHC, BU-P-BHC], and the adaptive best-basis BHC [TD-BB-BHC, BU-BB-BHC]. The results are presented in Figure 3.4.

The test set accuracies for KSC are very similar to those of Bolivar Peninsula, except that the pseudo-inverse classifiers perform better at the 1.5% sampling rate, with the accuracies for the pseudo-inverse BHC classifiers maintaining the accuracy level that had been achieved at the 5% sampling rate. At the lower sampling percentages, the covariance matrices are very poorly estimated in the full dimensional space, yet the accuracies are still fairly high

using pseudo-inversion, indicating that the differences in class means is the main reason the level of discrimination is being maintained. This result is also reflected by the standard deviations of the accuracies, which "spike" in the 15%-30% sampling rate range for the pseudo-inverse classifiers where the covariance matrices are still helping maintain a higher level of classification accuracy (than in the 1.5%-5% range), though unstable. The lower classification accuracies of

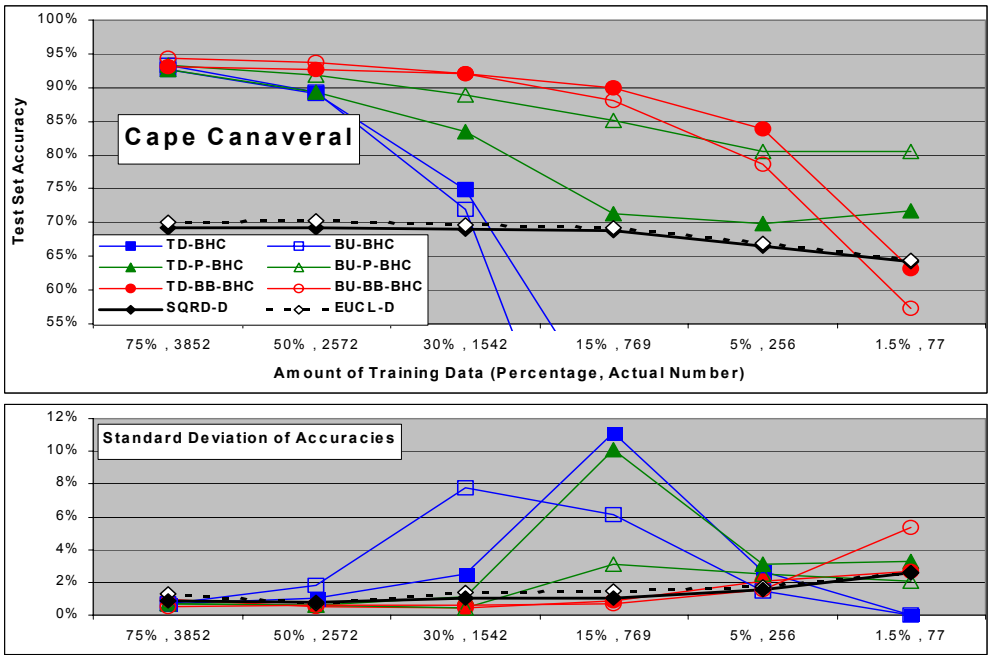


Figure 3.4: Classification (test set) accuracies for KSC, Cape Canaveral FL

the BB-BHC at the 1.5% sampling rate might be explained by a minimum requirement, the "intrinsic dimensionality" [26, 60, 61], for the number of bands, after which the results degrade sharply. While the benefit relative to "nearest

distance to mean" classifiers is significant at higher sampling rates, the use of the retained covariance information may be unwarranted at the 1.5% sampling rate because the quantity of training data, relative to the original dimensionality, is severely limited. As noted in [64], when the quantity of training data is very small, a simpler classifier (such as the mean distance classifiers) may perform better than one that attempts to properly estimate the covariance structure. However, using the adaptive BB-BHC classifiers has served to extend the range of training data quantity over which the covariance information is useful for improving the classification results.

3.3.2.2 Domain knowledge and image evaluation

Classified images were studied to determine, at least qualitatively, how the adaptive BB-BHC classifier performed in the context of the entire areas the quantity of training data was reduced. Images obtained using the adaptive BB-BHC (%, BB) were compared to those obtained using the BHC with pseudo-inversion (%, Pseudo) at each of the percentages (75, 50, 30, 15, 5, and 1.5) for both the TD and BU methodologies. Analyzing the images obtained using the BU algorithms indicates that the algorithm still performed well overall, at 30% sampling rate [Appendix B]. However, at the 15% and 5% sampling rates the results of the adaptive BB image [15 BB] are slightly better than those from the pseudo-inverse approach, whose output shows classes that are not spatially cohesive. [15 Pseudo, 5 Pseudo]. However, at the 1.5% sampling rate, the adaptive BB image is quite poor, whereas the results shown in the one obtained using pseudo-inversion have not deteriorated much from the one obtained at the

5% sampling rate. This insight is consistent with the classification accuracy performance of the classifiers [Figure 3.1].

Examining the results from the TD framework, it is interesting to note that both techniques have trouble discriminating the shallow water areas where reflectance from the sand bottom is visible in the imagery [Appendix B]. While some researches may find this disconcerting, others may be alarmed that the BU algorithms didn't identify these shallow water areas. This research focuses on discrimination of land cover classes and prior studies in this area had entirely masked out the Gulf waters. Comparing the images at the 30% sampling rate and with their respective images obtained using 100% of the data, the adaptive BB image is slightly superior to the one obtained using pseudo-inversion [Figure B.1], which exhibits isolated classes associated with overlapping distributions that cannot be well characterized. Further yet, the adaptive BB image is noticeably better at the 15% sampling rate [Figure B.2]. While the adaptive BB images at the 5% and 1.5% sampling rate appear to be slightly better than the pseudo-inverse counter-parts, the quality is poor. This is a very interesting result due to the fact that the pseudo-inverse based classifier performs better on the labeled data than the adaptive BB classifier, both for the TD and the BU frameworks, at the 1.5% sampling rate. However, the differences in accuracies at the 1.5% sampling rate between the TD classifiers is slightly greater than 5%, whereas the difference is more than 20% for the BU classifiers. This indicates that in situations where the pseudo-inverse technique yields better accuracies for training and test data, but where the difference is not too large, the adaptive BB algorithm

may still perform better on the entire scene due to over-training of the pseudo-inverse classifier.

3.3.2.3 Intrinsic dimensionality

Further investigation was performed to determine how the possible violation of the "intrinsic dimensionality" of the data by the adaptive BB-BHC might have impacted the classification results. To achieve this, the advantages of the adaptive BB-BHC and the pseudo-inverse were both utilized. For partitions of the BHC where the ratio of data to dimensionality would result in a number of bands less than the threshold, the pseudo-inverse was utilized. Because of the peaking performance of the pseudo-inverse, when $|\boldsymbol{X}| < \alpha_{\text{ratio}} d$ at a partition of the BB-BHC structure, the algorithm sets $d^* = \min\left(d, 2|\boldsymbol{X}|\right)$, achieves this dimensionality of the data using BB, and then uses the pseudo-inverse [30, 37]. Based on some preliminary results, thresholds of 10 and 50 were used. Additionally, the pseudo-inverse optimized to the "sweet spot" at each partition was investigated. The results for the TD-BHC and the BU-BHC are presented in Appendix C. Using a threshold on the number of bands helps improve the classification accuracies at the lower sampling percentages, but they still fail to do better than the full dimensional model with the pseudo-inverse. While it extends the advantage of the BB-BHC, there will still be a point at which a nearest Euclidean mean classifier is a better option than trying to get a stable covariance estimate. It appears that the intrinsic dimensionality is somewhere between 10 and 50: performance is better with a 50 band minimum for the 5 and 1.5 percentages but it is better at the 0 and 10 band minimum for the other

percentages. Furthermore, for both the TD and BU frameworks, the optimized pseudo-inverse classifier performs worse than the non-optimized pseudo-inverse classifier at the 1.5% sampling rate. The degradation of performance of the "optimized" pseudo-inverse classifier indicates that the improvement of the pseudo-inverse classifiers over the adaptive BB-BHC classifiers at that quantity of data is due to an increased emphasis on the differences between class means rather than an improved preservation of the covariance information.

Summary

The dependency of classification accuracy upon an adequate quantity of training data, as compared to the dimensionality of the data, is widely noted and needs to be addressed during the design of a classifier. While the advent of hyperspectral sensors has provided unique opportunities in the application area of land cover classification, the increased dimensionality of the data necessitates that researchers pay even more attention to the classifier dependence on the quantity of training data. Our proposed multi-classifier framework utilizes the flexibility gained by transforming the output space and input space simultaneously to combat the small sample size problem. By reducing the size of the feature space in a directed manner, dependent upon the quantity of training data available in the binary hierarchy of meta-classes, a high level of classification accuracy is preserved even when faced with low quantities of training data for some of the classes.

Further research is necessary to support the lower bounds on the intrinsic dimensionality of the feature space. Additionally, the algorithm currently uses the

average spectral response for representing the spectral response of the group-bands, a more sophisticated representation of the group-bands may result in better performance, particularly when the group-bands grow quite large. While combating the small sample size problem with the dynamic best-basis algorithm helps preserve the interpretability of the data, using Fisher's linear discriminant function as the feature extractor at each internal node of the BHC diminishes this attractive characteristic. While the discriminant function weights on each band/group-band could be analyzed to determine the respective band's importance, the interpretation and insight would be less complicated if feature selection was performed rather than feature extraction. Therefore, feature selection rather than feature extraction, and the likely trade-off between classification accuracy and retention of domain knowledge, should be investigated. Investigation has also been performed using two different pair-wise correlation measures for constructing the basis rather than trying to stabilize the estimates with the pooled covariance from an "ancestor" meta-class. While, in preliminary experiments, the computation time was greater and performance was degraded, more investigation is required. Additionally, a researcher should consider using a nearest Euclidean mean distance classifier instead of the pseudo-inverse for partitions where the amount of training data available does not support the intrinsic dimensionality.

Chapter 4: Spatially Limited Ground Truth and Knowledge Transfer

The adaptive Best-Basis BHC demonstrated a remarkable ability to preserve a high level of classification accuracy when only limited quantities of data are available. However, other "limitations" on the data may also exist. Another difficult problem is encountered when it is necessary to "transfer" a classifier to a population that is not properly represented by the training data. In essence, the data are of limited (poor) quality. This may happen for many reasons, including lack of adequate information on the populations of interest and changes in populations from the set where the training data are acquired to a set on which the classifier is to be applied. Furthermore, there may be domains, such as remote sensing, where there is a spatial context.

Both Bolivar Peninsula and Kennedy Space Center are environmentally important, inaccessible areas that contain dynamic classes (spatially varying within class variation), some of which are very hard to separate. Further, there is concern that if the training data sites do not adequately cover the regions, characteristics of certain classes that are not exhibited for all "pockets" of those classes may be overlooked or, even worse, entire classes may not be included in the results. However, due to time and financial limitations, it is not possible to field verified ground truth that covers the entire region. Therefore, it is often necessary to acquire ground truth data over a spatially limited area and assume that the data contained in this "closed world" is representative of the entire region of interest. While the classification and mapping of land cover types for these

environments is difficult due to the similarity of spectral signatures for certain vegetation types, discrimination between the land cover types is reduced even further when ground truth is limited. A novel approach for preserving the classification accuracies when the researcher is faced with a limited quantity of ground truth was presented in Chapter 3. A different limitation occurs when training data are acquired on only a limited portion of the test site: results can be biased and discrimination reduced. The dramatic impact this limitation can have on classification results and image analysis is investigated in this chapter. Furthermore, it is demonstrated that this problem can be viewed as a (spatially) limited training data problem. An adaptive approach for combating this problem is presented in which previously acquired information is transferred and applied for classification of regions for which spatially proximal ground truth cannot be acquired.

4.1 MOTIVATION

In many cases the criteria for selection of training and test samples are dictated by factors that are independent of the statistical analysis. Even if the sampling scheme has been developed using rigorous statistical methods, it may be impractical or impossible to implement these plans in real world applications. A much more realistic scenario involves acquisition of samples at reasonably accessible sites, with a large number of samples collected at each site; in practice, this is exactly how the labeled ground truth is acquired. Furthermore, land cover classification is typically performed based on the spectral response of the ground truth from spectrally homogeneous "pockets" of a region, whereas the goal is to

classify the entire region and possibly to even utilize the information for classification of other data acquired over regions for which ground truth cannot be acquired.

4.1.1 Shortcomings of traditional classification

Current land cover classification methods are not designed to deal with spatially dynamic classes – exhibiting trends within classes from region to region - in the context of non-global samples. Typically, methods are trained and tested for classification accuracy from data in a "closed world"; all of the training and testing data are selected from a contiguous subset(s) of the region and, for practical reasons, the labeled data are collected only at a limited number of sites. Random selection of labeled data for training can result in "testing" points that are neighboring - or even surrounded by - points on which the classifier was trained. Due to the limited spatial extent of the available data, it is possible that the training data are not representative of the entire population. In essence, inferences are made about the underlying characteristics of the population based upon local information. Consequently, the resulting classifier usually performs poorly on the other "segments" of the population where no labeled data are available. The true classification performance is compromised, and the implied generalization accuracy (the classification accuracy of labeled data not in the training set) may also be deceptive, inaccurate, and inflated. Thus, these results likely result in a poor representation of both the populations that are known to exist in an area and even worse characterization of the even more difficult problem: where the data have not been "seen".

4.1.2 Dynamic application area

The natural variation of the spectral signatures within each class is increased by the impact of environmental characteristics that are spatially non-stationary such as soil composition, terrain, and weather. Because pixels in an image often are not entirely pure due the resolution of the sensors [5 meter at Bolivar Peninsula and 20m at KSC], the presence of mixed and small classes, can also impact the spectral signature of individual pixels. While the issue of "within sample variance" has been recognized, the problem of between sample spatial variability of class spectral signatures has not been formally addressed within the remote sensing community. Natural differences can be further exacerbated by factors related to the acquisition of the data. For example, an airborne sensor cannot typically acquire data over an entire study area on one flight line because of limited swath width. Differences can result from the sun angle and bi-directional response of the targets observed on different flight lines. An example where the natural variation is possibly confounded with differences related to multiple acquisitions and the radiance-to-reflectance transformation would be the flight lines of Bolivar Peninsula, the mosaic of which is depicted in Figure 3.1. The HyMap sensor, which is usually flow on light, twin-engine aircraft platform, has an operational altitude of 2000-5000 m Above Ground Level (AGL), a swath width of 60-70 degrees, and, in general, a spatial resolution of 2-10 m. At low altitudes, swaths are narrow, with the maximum width being about 5 km [47]. As a result multiple acquisitions were required to achieve full coverage of the two focus areas on Bolivar Peninsula. Figure 4.1 depicts the spectral signatures for

the ground truth of the General Uplands class on Bolivar Peninsula. While the visible range of the spectrum (bands 1-20: 0.45 – 0.89□m) appears very similar for both sites, the spectral responses acquired at the second test site (labeled blue) are, in general, noticeably higher than those for the same class at test site one (labeled red) for bands 21-62 (NIR: 0.89 – 1.35□m), slightly higher for bands

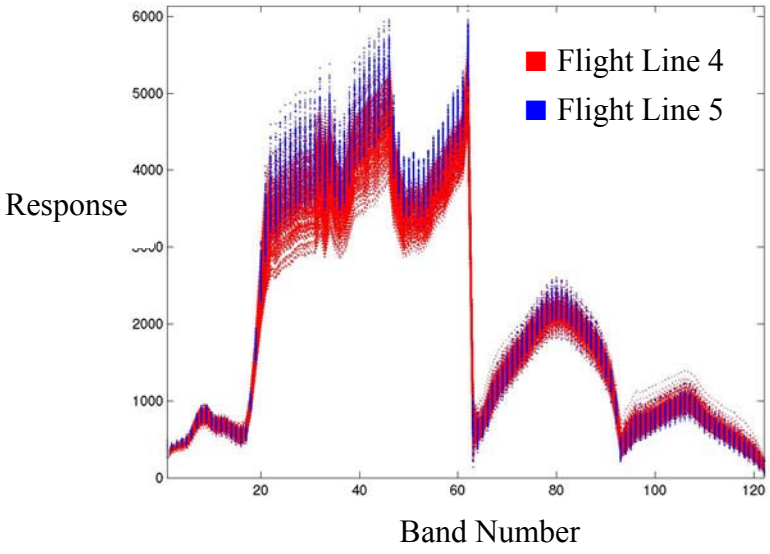


Figure 4.1 Spectral signatures for Class 8 (General Upland) at Bolivar Peninsula for the two different test sites

63-94 (SWIR1: 1.40 – 1.80□m), but slightly lower for bands 95-122 (SWIR2: 1.95 – 2.48□m). The similarity of the spectral signature of this class to other classes determines the amount of classification degradation. The spectral signatures, with the respective test sites identified, for each of the 11 classes used at Bolivar Peninsula are in Appendix D.

An example where most of the "controllable" factors are negated would be the KSC, Cape Canaveral Florida test sites depicted in Figure 3.3. The two test sites are from the same acquisition, with the time difference just being that of the travel time of the ER-2 whose speed is about 730 km/hr [42]. AVIRIS is able to acquire a larger swatch (approximately 11km) than HyMap because it flies at a much higher altitude (20km vs 5km) and acquires data at a coarser spatial resolution (20m vs 10m). However, even when it is possible to control for non-natural spatial variability, dramatic differences can be present in supposedly

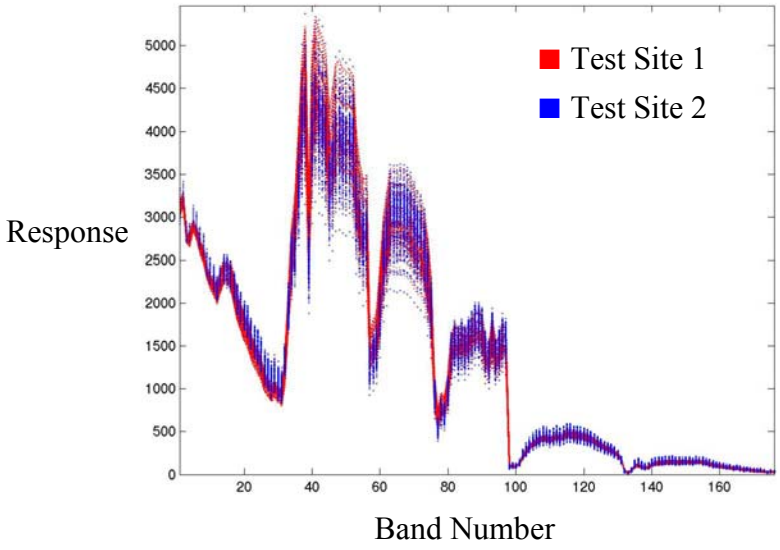


Figure 4.2 Spectral signatures for Hardwood Swamp at the two different Kennedy Space Center sites

"pure" ground truth collected at different locations as depicted for the Hardwood Swamp class at KSC [Figure 4.2]. Much like the HyMap data of Bolivar

Peninsula, the differences in the spectral signatures between the two test sites at KSC are not always dramatic, but they differ enough to impact identification of the true land cover label. The spectral signatures for the 16 classes used at Kennedy Space Center are in Appendix E. Additionally, five of the land cover classes identified at KSC are only found at one or the other test site but not both sites. A researcher that used a classifier trained and tested on one of the test sites to classify the other test site, without any modifications, would fail to identify most of the original classes and would not detect the new class.

4.2 IMPACT OF LIMITED SPATIAL COVERAGE OF THE GROUND TRUTH

A classifier that is "trained" on the ground truth of land cover types from certain sites is ultimately used to classify the land cover types of other areas where no immediate data are available. Multiple experiments were performed in which the ground truth was partitioned into training and testing subsets in order to evaluate the "performance" of the trained classifier. Here, the average classification accuracy on the test sets is generally used as the measure for how well the classifier identifies the correct label, although the fully classified images are evaluated qualitatively. In regions where no immediate ground truth is available, it is obvious that there is a lack of adequate domain knowledge to determine whether a classifier is performing poorly. Therefore, even if the available ground truth is not representative of the populations on which the classifier will be applied, the test set accuracies could mislead a researcher to assume a high level of performance. The "purity" of the available ground truth is highly dependent on the selection process, which can be very subjective, and will

directly impact the test set accuracies. The concern here is what happens when the possible impact of the spatially limited ground truth is not accounted for and how that coincides with the implied performance obtained on the ground truth available.

4.2.1 Transferring the classifiers

In addition to noting the overall classification accuracy rate, the confusion matrix that depicts the number of pixels for each class given a specific label is also evaluated. While a confusion matrix can be helpful for succinctly presenting the overall results, it does not effectively utilize the available domain knowledge. Because a hierarchical classifier is being used, it would be advantageous when analyzing the results to be able to evaluate what types of misclassifications are occurring in relation to the structure of the classifier being used. While the information in the confusion matrix is useful, it is not helpful in evaluating the locations in a tree where the mistakes are made. With a "precision tree", which was motivated during this research by the shortcomings of confusion matrices, the tree is structurally identical to the hierarchical classifier being used. However, at each partition of the applicable BHC structure, an accuracy measure is given that is indicative of how well the classifier performs at the meta-class level. This measure is the percentage of correctly labeled pixels for each meta-class. Therefore, the top node "precision" will always be 100% since it contains all classes, whereas the leaf node precisions are reflective of the purity of the pixels labeled that specific class.

4.2.1.1 Bolivar Peninsula Classification Accuracies

The average classification accuracies, based on 10 experiments with different seeds, for the TD and BU Adaptive BB-BHC classifiers were obtained by using a 50% stratified (by class) random partition of the data sets, these results are listed in Table 4.1. Based upon the high level of classification accuracy being

Region	Classifier	Training Set Accuracy	Test Set Accuracy
Bolivar Site 1	TD	0.9925	0.9923
Bolivar Site 1	BU	0.9956	0.9894
Bolivar Site 2	TD	0.9987	0.9973
Bolivar Site 2	BU	0.9990	0.9981

Table 4.1: Bolivar Peninsula average training and test set accuracies when classifier is applied to the site at which the ground truths were acquired.

obtained, it would be plausible to build a classifier based on all available ground truth to classify the entire image and possibly use it for areas where no immediate ground truth are available. To increase the diversity of the observed sample the BHC structures based on all the available ground truth are used to evaluate the performance of the classifier when it is transferred to the alternate site.

Whereas the practitioner may expect to achieve highly accurate results, when the classifiers are transferred from one site to the other, the accuracies are actually reduced dramatically as shown in Table 4.2. Furthermore, analysis of the resulting precision trees and confusion matrices indicates that a large number of

From Region	To Region	Classifier	Classification Accuracy
Bolivar 1	Bolivar 2	TD	0.7229
Bolivar 1	Bolivar 2	BU	0.6259
Bolivar 2	Bolivar 1	TD	0.3524
Bolivar 2	Bolivar 1	BU	0.4020

Table 4.2: Bolivar Peninsula classification accuracies when classifier is applied to the alternate site at which the ground truths were acquired.

the errors are occurring at early partitions of the hierarchy [Appendix F]. For instance, for the TD classifier built on Site 2, the first partition of the hierarchy discriminates $\Omega_2 = [1,2,3,10]$ from $\Omega_3 = [4,5,6,7,8,9,11]$ which can qualitatively be described as separating the water type classes from the land type classes. However, when this classifier is applied to Site 1, almost half of the ground truth pixels labeled as Ω_2 are incorrect [Figure F.4]. The distributions in the Fisher projected space for the training data (Site 2) and the testing data (Site 1) are depicted in Figure 4.3. The classifier trained on one region and tested on the

other is failing because the spectral signatures deformations have resulted in a subsequent change in the distributions in the projected space.

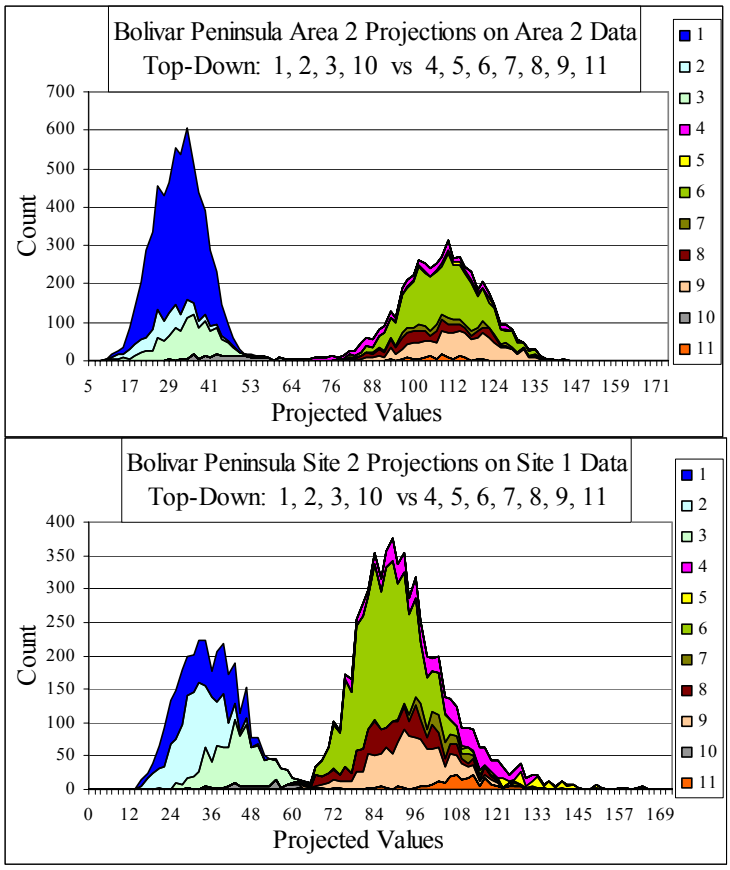


Figure 4.3: Deformation of the meta-class distributions in the Fisher projected space calculated from the ground truth acquired at Bolivar Peninsula Site 2

Image Analysis

Analysis of the images reinforces the fact that the transferred classifier performs much differently than expected given the performance of the classifiers

on the training sites [Appendix G]. Additionally, the images reinforce the implied disparity between transferring the classifiers from Site 1 to Site 2 versus from Site 2 to Site 1: while the images of Site 2 are quite poor, those of Site 1 are even worse.

4.2.1.2 Cape Canaveral Classification Accuracies

While not as high as those obtained at Bolivar Peninsula, accuracies for KSC data are very good [Table 4.3]. However, when the classifiers are trained on

Region	Classifier	Training Set Accuracy	Test Set Accuracy
KSC Site 1	TD	0.9659	0.9258
KSC Site 1	BU	0.9729	0.9369
KSC Site 2	TD	0.9282	0.8633
KSC Site 2	BU	0.9242	0.8530

Table 4.3: KSC average training and test set accuracies when classifier is applied to the site at which the ground truths were acquired.

one of the sites and then applied to the alternate site, the accuracies are extremely poor, even for classes that exist in both regions [Table 4.4].

From Region	To Region	Classifier	Accuracy on Identical Classes	Accuracy on All Classes
KSC 1	KSC 2	TD	0.4720	0.3946
KSC 1	KSC 2	BU	0.4901	0.4097
KSC 2	KSC 1	TD	0.5141	0.4297
KSC 2	KSC 1	BU	0.5213	0.4357

Table 4.4: KSC classification accuracies when classifier is applied to the alternate site at which the ground truths were acquired.

Image Analysis

Similar to the results obtained at Bolivar Peninsula, if it is assumed that visual evaluation of the fully classified data sets obtained by using all of the ground truth specific to each site are the standard by which to judge the classification of the land cover types, the fully classified data sets obtained by directly transferring the classifiers from one site to the other are poor [Appendix G].

4.2.2 Combined classification results

The classifiers were developed using the combined data sets to investigate whether the classes from the alternate sites were truly different or whether training data sets are too small (unrepresentative). Classification accuracies obtained from the combined training data are high, which indicate that the problem can be attributed to (spatially) limited training data [Table 4.5]. Although the level of accuracy retained at KSC is not as high as that observed for

Bolivar Peninsula, the land cover classification problem at KSC is more difficult for the combined data than the site specific data because the combined output space size has 16 classes, whereas individually there are only 13 classes at Site 1 and 14 classes at KSC Site 2.

Region	Classifier	Training Set Accuracy	Test Set Accuracy
Bolivar 1 and 2	TD	0.9875	0.9863
Bolivar 1 and 2	BU	0.9857	0.9832
KSC 1 and 2	TD	0.9063	0.8506
KSC 1 and 2	BU	0.8893	0.8352

Table 4.5: Bolivar Peninsula and KSC average training and test set accuracies when classifier is applied to the ground truths combined from both sites at which the ground truths were acquired.

Furthermore, analysis of the classified images provides visual evidence that the classifiers built on the combined data are applicable for both areas [Appendices H]. Whereas the images obtained using the classifiers trained on one site and applied to the other are dramatically different from those obtained using the classifier built on each of the respective site's ground truth, the images obtained using the classifiers built with the combined training data are very similar to the site specific images. This supports the formulation of the problem as still being one of a small sample size where, in this context, the "smallness" refers to the limited spatial coverage of the ground truth.

4.3 KNOWLEDGE TRANSFER OF TREES AND FISHER PROJECTIONS

If no training data are available for the new area, current land cover classification methods do not address how a trained classifier can be applied to the "new" region. While the BHC classifiers do not specifically address the problem of spatial variability in the signatures of given classes, the framework is advantageous for performing this task.

4.3.1 Background

When the underlying labels of the pixels are not known, the similarity of the spectral signatures and/or spatial characteristics of the images can be used to form groupings or clusters. Basic unsupervised clustering algorithms group pixels into N different clusters, C_i, $i = 1,\ldots,N$, based on the spectral characteristics of the pixels. A commonly used algorithm in pattern classification and image analysis for partitioning the data is *k*-means, which moves data points from one cluster to another to improve on a predetermined criterion such as sum-of-squared-error [1, 28]. Extensive research has been completed in the classification/clustering area in the context of no training data (referred to as "unsupervised") with applications to spectral data [85-87, 95, 96]. Additionally, the basic ideas of unsupervised clustering have been extended to account for spatial relationships within neighborhoods of data points [92-94]. While accounting for the spatial information in clustering on a grid is a desired characteristic, the added computational effort can be quite prohibitive. An additional, and nontrivial, concern is the determination of the number of clusters (cluster validation) [1, 88-91]. These topics are outside the scope of this work.

4.3.2 Updating parameter estimates by pseudo-labeled data

Finding a feature space in which the underlying labels are easily discriminated is central to clustering. When training data are available for the BHC framework, the projected values, based upon the Fisher projection(s) $\Psi\left(\mathbf{x}\middle|\mathbf{A}\right)$ between two meta-classes $\left\{\Omega_i,\Omega_j\right\}$, are used to model the pdfs $p\left(\Psi\left(\mathbf{x}\middle|\mathbf{A}\right)\middle|\Omega_k\right)$, $k=i,j$. However, in the context of this problem, the spectral signatures may be "deformed" due to spatial variation and hence the pdfs "learned" in the old area may no longer be strictly applicable. However, rather than trying to cluster the pixels into class homogenous clusters and then determining their underlying label without any prior information, the domain knowledge acquired from the old area should be utilized as much as possible. The BHC structure is conducive because the hierarchy has already been discovered in which the easiest discriminations are performed first. Iteratively, it is much easier to discriminate between two classes at a time rather than attempting to identify the label as 1 of C while simultaneously solving the complicated issue of cluster validation. Furthermore, the space in which the discrimination should be attempted has already been found based upon the characteristics of the Fisher projections that are generally robust to moderate changes in distributions.

While the idea of transferring classifiers, or "knowledge reuse", is not new for problems in which there are limited data and/or long training times, previous work focused on transferring classifiers or data sets based upon the assumptions that the old trained classifier will be able to correctly identify the labels for a

portion of the new data or it is assumed that there is at least a small quantity of training data available [97-99]. Here, an alternative method is proposed to adaptively update and transfer the BHC framework to regions where there is no immediate ground truth available, and the old classifier is not assumed to necessarily correctly identify the label of new data. It is assumed that the deformations in the spectral signatures are similar across the classes, and therefore the domain knowledge acquired previously is still relevant for discrimination. The BHC framework learned in an area where there is ground truth available can be transferred in the following manner:

1. For the meta-class pair $\{\Omega_i, \Omega_j\}$, project the unlabeled data into the Fisher space by applying the projection $\Psi(\mathbf{x}|\mathbf{A})$ learned from the "old" area, specific to that split, to the new data **y**: $\mathbf{z} = \mathbf{y} \cdot \Psi(\mathbf{x}|\mathbf{A})$
2. Cluster the projected values **z** in the Fisher space using k-means to form two clusters $\{C1, C2\}$. Record the cluster means $\{\mu_{C1}, \mu_{C2}\}$, variances $\{\sigma_{C1}^2, \sigma_{C2}^2\}$, and membership.
3. Assign the meta-class labels $\{\Omega_i, \Omega_j\}$ to the members of each cluster such that the distance measure $[(\mu_{\Omega_i} - \mu_{Ci})^2 + (\mu_{\Omega_j} - \mu_{Cj})^2]$ is minimized.
4. Return to Step 1 if any remaining meta-classes have not been partitioned down to the leaf (specific label) node level and if there are pseudo-labeled data available to cluster.

This methodology uses the BHC structure to "push" pseudo-labeled pixels to a leaf (specific label) node. The pseudo-labeled data are then used to update the parameter estimates necessary for classification.

4.3.3 Performance

In general, the updating scheme improved the classifier's transferability. On average, the classification accuracy improved by nearly 25% (from 52.58% to 76.17%) [Table 4.6]. Additionally, it is important what "type" of errors where made. Comparison of the precision trees obtained with and without the parameter

From Region	To Region	Classifier	Classification Accuracy
Bolivar 1	Bolivar 2	TD	0.7492
Bolivar 1	Bolivar 2	BU	0.7576
Bolivar 2	Bolivar 1	TD	0.9102
Bolivar 2	Bolivar 1	BU	0.6299

Table 4.6: Bolivar Peninsula classification accuracies when classifier is updated using pseudo-labeled data to estimate the new parameters and applied to the alternate site at which the ground truths were acquired.

updating scheme indicates that the classifiers using the updated parameters make errors that would be "more acceptable", in terms of the class hierarchy indicated by the respective BHC framework, then the non-updated counter-part classifier. For instance, for the BU classifier trained on Site 1 and tested on Site 2, the

updated classifier only identifies Class 2 correctly 19% of the time whereas the original classifier has an accuracy of 42.4%. However, for the meta-class $\Omega = \{1, 2\}$ the updated classifier has an accuracy level of 99.4% versus 98.3% for the non-updated classifier. For the other classifiers, for which the updated classifiers have a higher classification rate than the corresponding non-updated classifiers, the superiority of the precision matrices is even more pronounced.

Unfortunately, the updated classifier did not yield clearly superior classification results for the KSC sites. In fact, in all but one case (TD BHC trained on Site 1 and tested on Site 2), the classification accuracy was slightly worse [Table 4.7]. Furthermore, analysis of the precision trees between the updated and non-updated classifiers indicates that those obtained from the non-updated classifier are comparable to the updated counterparts [Figures F.5-F.8, F.13-F.16]. These results highlight one of the major drawbacks of the updating scheme: if a class exists that has not previously been seen, or if a previous class is no longer present, the updating scheme fails from that level of the BHC hierarchy downward as future partitions are dependent upon the quality of their "ancestors" in the hierarchy.

From Region	To Region	Classifier	Accuracy on Identical Classes	Accuracy on All Classes
KSC 1	KSC 2	TD	0.4886	0.4085
KSC 1	KSC 2	BU	0.4481	0.3746
KSC 2	KSC 1	TD	0.4500	0.3762
KSC 2	KSC 1	BU	0.4682	0.3914

Table 4.7: KSC classification accuracies when classifier is updated using pseudo-labeled data to estimate the new parameters and applied to the alternate site at which the ground truths were acquired.

4.4 CONCLUSIONS

The impact of spatial variability of class signatures on classification accuracy was demonstrated to be quite significant. While it would be convenient to simply attribute the deterioration of classification accuracy to a change in classes and that the problem is in fact a "new" one, it is also demonstrated that the problem can be framed in the context of limited training data. While the proposed parameter updating methodology performs fairly well for problems where the output space does not change, the presence or absence of classes during the knowledge transferal process poses too difficult of a situation for the updating scheme.

Chapter 5: Ensembles and Output Space Precision

Recently, sub-sampling methods have been investigated as means to create an ensemble of classifiers for use with classification trees similar to the BHC framework [33, 49, 50, 79], with the goal of improving overall classification accuracy. However, with applications such as land cover classification with hyperspectral data that already suffer from a poor ratio of quantity of training data to dimensionality, the use of sub-sampling has gone largely unexplored and uninvestigated. This issue has not been addressed rigorously, and approaches that mitigate the effect have not been developed.

The Adaptive BB-BHC framework, which preserves classification accuracies when the available training data are limited, can also be used with sub-sampling techniques. Independent of spatial variation and limited training data, different sampling subsets of the same ground truth may result in slightly different classification results due to differences in the parameter estimates. In Chapter 4, the serious classification problems that can arise if a classifier trained on one area is "blindly" applied to a seemingly very similar area where no immediate ground truth is available were demonstrated. To mitigate this problem, a new "classifier transferal" method for updating the parameters of the meta-class conditional distributions was proposed. While the application of the method to the Bolivar Peninsula and KSC data sets indicated that the updating scheme has potential and performs relatively well at the meta-class levels, the accuracies can still be quite poor in terms of the class-specific precision. Additional domain knowledge

gained from transferring classifiers with different hierarchies indicates that certain splits of the hierarchy can be more advantageous for the transferability process than others. For instance, when the classifiers were transferred from Site 2 to Site 1 at Bolivar Peninsula, the TD structure outperformed the BU structure by nearly 30% [Table 4.6]. In this chapter, the methodology is outlined for constructing a single hierarchy, a "master tree" T_M, when multiple hierarchies are available due to different combinations of samples and classification algorithms. The ability to identify a single framework that incorporates information from all the structures, rather than having several, such as both a TD BHC and a BU BHC classifier, is advantageous because it helps preserve interpretability of the class hierarchy and should be more robust for transferal than any of the individual structures. The T_M structure is utilized in two different ways: the structure can be used as the hierarchy to be transferred by training it with all the available ground truth, or it can be used as an evaluation tool for the aggregate output of the individual classifiers. Furthermore, techniques for determining the appropriate precision of the output space are developed specific to each type of application of the master tree. These methods are applied to the transferal of classifiers between the different test sites at Bolivar Peninsula and KSC.

5.1 CONSTRUCTING A MASTER TREE

Different classification methods, as well as different samples of ground truth, can produce different hierarchies. Previous research has investigated the search for exact tree structures or matching partitions within different structures [80-82]. None of this previous research addresses the problem of consolidating

the structurally different trees T_i, $i = 1,\ldots S$ that result from the S combinations of different samples and classifier. However, the domain knowledge inherent in the S structures can be used to make inferences about inter-class relationships, such as which classes are always separated first and which ones' locations are the least "stable". A distance based approach for utilizing this information is proposed.

The method involves tabulating tree structures for each of the hierarchies in terms of the meta-class tree structure and using a greedy "bottom-up" agglomerative clustering algorithm to form the collective association rules. The average number of internal nodes $\bar{O}_{i,j} = \left(\sum_{S} O_{i,j}^{T_i} \right) \Big/ S$ that exist between the leaf nodes for label pairs $\{L_i, L_j\}$, $\forall i < j$ and the meta-class Ω_k that contains them both ($\{L_i, L_j\} \in \Omega_k$) is determined. This "distance", a measure of how far the labels are from the meta-class in which they are partitioned, is not symmetric and must be calculated in both directions ($L_i \to \Omega_k$ and $L_j \to \Omega_k$). As a minimum, if there is a meta-class $\Omega_k = \{L_i, L_j\}$ present in all of the hierarchies, $\bar{O}_{i,j} = 2$. Conversely, the maximum $\bar{O}_{i,j} = (C/2)$ is realized if, for each of the trees, the leaf nodes of L_i and L_j cannot be structurally connected without traveling through all the internal nodes in the hierarchy. As an example, six possible BHC tree configurations are depicted in Figure 5.1. For this example, $\bar{O}_{1,2} = \frac{3+3+3+2+3+4}{6} = 3$ because $O_{1,2}^{T_1} = 3$ $(L_1 \to \Omega_2 = 2, L_2 \to \Omega_2 = 1)$, $O_{1,2}^{T_2} = 3$ $(L_1 \to \Omega_2 = 2, L_2 \to \Omega_2 = 1)$, $O_{1,2}^{T_3} = 3$ $(L_1 \to \Omega_2 = 1, L_2 \to \Omega_2 = 2)$, $O_{1,2}^{T_4} = 2$ $(L_1 \to \Omega_5 = 1, L_2 \to \Omega_5 = 1)$, $O_{1,2}^{T_5} = 3$ $(L_1 \to \Omega_2 = 2, L_2 \to \Omega_2 = 1)$, and $O_{1,2}^{T_6} = 4$ $(L_1 \to \Omega_1 = 2, L_2 \to \Omega_1 = 2)$.

The greedy algorithm for constructing the master tree T_M from the distance measures $\bar{O}_{i,j}$ is initialized by selecting the class pair $\{L_i, L_j\}$ to

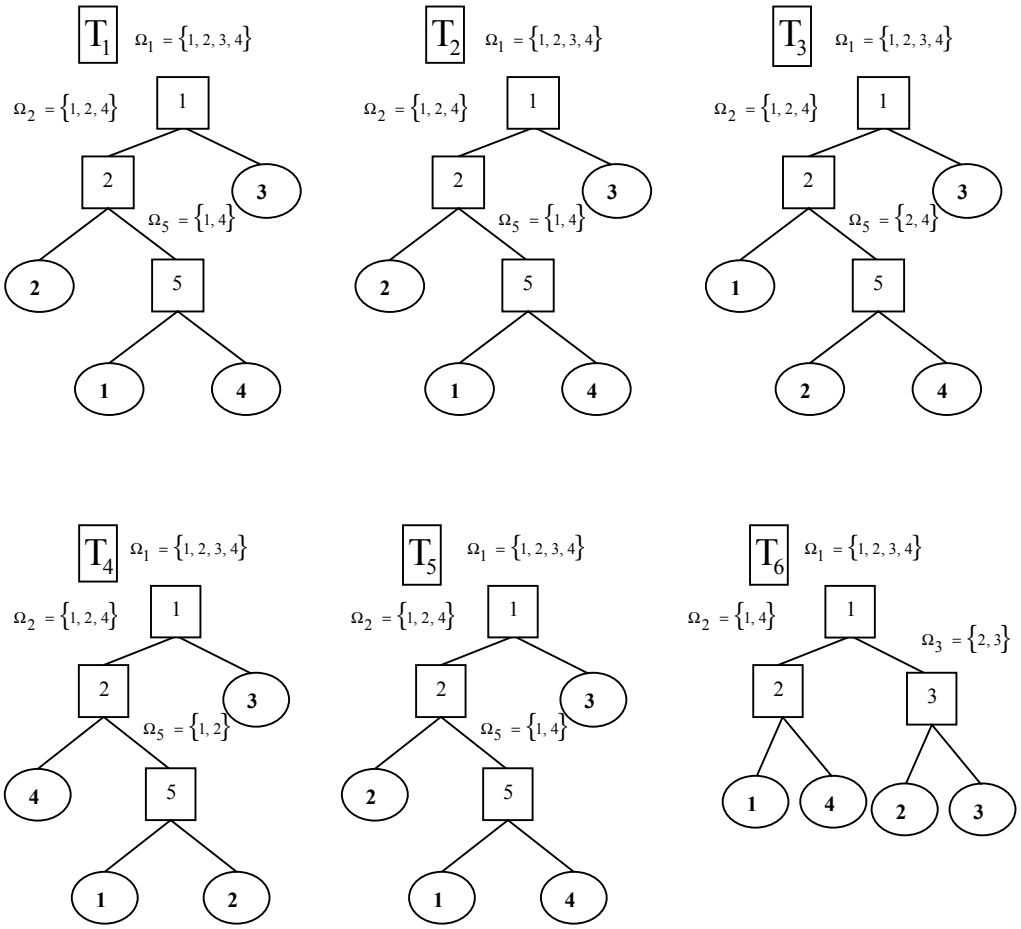


Figure 5.1: An example of 6 possible hierarchies involving 4 classes.

combine such that $\bar{O}_{i,j}$ is the minimum individual value. The distance measure to the meta-class $\Omega = \{L_i, L_j\}$ that results from combination is calculated for the remaining classes (or meta-classes) by merging M the distance measures

$\left\{\bar{O}_{i,k}, \bar{O}_{j,k}\right\} \forall k \neq i, j$ to form the new distance measure $\bar{O}_{\{i,j\}k} = M \left\{\bar{O}_{i,k}, \bar{O}_{j,k}\right\} \forall k \neq i, j$. The merging can be based upon the average, the minimum, or the maximum of the pair. Investigation of the different merge functions indicated that using the minimum results in class hierarchies that were most appealing. Class and meta-class pairs are combined in this manner until a single meta-class, the top node in the hierarchy, remains and the resulting Master Tree can be constructed by recreating the binary combinations. An algorithm was also developed that identifies a master tree based upon the most common partitions. It was not as robust in terms of dealing with a variety of hierarchies whose structures are not very similar and did not perform well when *S* was not large, so it was not pursued further.

5.2 Transferring the Master Tree and Identifying an Appropriate Output Space Precision

The Master Tree can be utilized in two different manners. Rather than using the classification algorithm to determine the appropriate hierarchy, the T_M structure can be used during the training process. Conversely, rather than having a single re-trained classifier with the T_M structure, each classifier in the ensemble can be retained. The aggregate "vote" of the ensemble can then be represented and evaluated in terms of the T_M structure. Recent research related the classification accuracies to the precision with which classes are defined and the complexity of the classifier algorithm being used [64]. Another factor that dramatically impacts all three problems investigated here is the precision of the ultimate output space. For example, it may be much easier to identify a tree in

general versus specific types of trees such as Slash Pine or Oak Hammock. Two different methods, each specific to the two different usages of the T_M, are presented here that are valuable tools for a researcher evaluating the appropriate output space precision for transferring a classifier. One is based on distance measures between pseudo-labeled clusters and the other on the purity (diversity) of the ensemble of classifiers.

5.2.1 Distance measure between the pseudo-labeled clusters

When the Master Tree is transferred to the region in which there is no immediate ground truth and it is applied to the data, clusters are formed at each partition of the T_M framework, matched to the original meta-classes, and then used to update the parameter estimates at each partition. While no labeled data are available, if it is assumed at each partition that the pseudo-labels are correct, a distance measure of the "separation" of the two classes can be compared to the same measure on the previously acquired ground truth. If the distance measure differs greatly from what would be expected based upon the known ground truth, it would signal a potential failure at that level of the T_M framework to correctly identify the pseudo-labeled data. The researcher should consider redefining the output space at the corresponding meta-class level and conduct further investigation. A popular distance measure of the separation between two classes $\{L_i, L_j\}$ is the Bhattacharyya distance (5.1) [1,65]:

$$D = \frac{1}{8}\left[\boldsymbol{\mu}_{L_i} - \boldsymbol{\mu}_{L_j}\right]' \left[\left(\Sigma_{L_i} + \Sigma_{L_j}\right)/2\right]^{-1} \left[\boldsymbol{\mu}_{L_i} - \boldsymbol{\mu}_{L_j}\right] + \frac{1}{2}\ln \frac{\left|\left(\Sigma_{L_i} + \Sigma_{L_j}\right)/2\right|}{\sqrt{\left|\Sigma_{L_i}\right|\left|\Sigma_{L_j}\right|}} \qquad (5.1)$$

where the first term is a measure of the separation due to differences in the mean

vectors $\{\boldsymbol{\mu}_{L_i}, \boldsymbol{\mu}_{L_j}\}$ and the second terms accounts for separation due to differences in the covariance matrices $\{\Sigma_{L_i}, \Sigma_{L_j}\}$. For any distance measure that utilizes the class conditional covariance matrices, an adequate ratio of data quantity to dimensionality must be achieved for reliable calculation. When the ratio is inadequate, the Best-Basis algorithm is utilized to reduce the data dimensionality. This method can use the distance measure with the Master Tree because the T_M structure is being used, and therefore all of the partitions are applicable.

5.2.2 Purity of the ensemble at each partition

Rather than retraining a classifier with a forced T_M structure, each the ensemble of individual classifiers can be evaluated simultaneously by considering how each of the classifiers in the ensemble agrees or disagrees with the voted label. This "purity" P can be measured in terms of classifier agreement on the meta-class $\Omega = \{L_i, L_j\}$ levels. For example, if the ensemble votes are $\Omega = \{L_1, L_2, L_1, L_3, L_4, L_1\}$ then the vote label would be L_1 and the purity for this particular observation would be evaluated for all of the meta-classes containing L_1. For instance, if there is a meta-class $\Omega = \{L_1, L_2, L_3\}$ in the hierarchy, then the purity at that meta-class for that pixel is $P\left(\Omega = \{L_1, L_2, L_3\}\right) = (1+1+1+1+0+1)/6 = 0.8333$, where the zero indicates that the 5th classifier in the ensemble did not indicate the pixel was of that particular meta-class (it's vote was for L_4 which is not with-in the meta-class). The purity never decreases as the applicable meta-class grows. Both measures, the comparison of distance measures and the ensemble purity, can be used to

determine the appropriate level of the hierarchy at which they are comfortable with the results.

5.3 Application to Hyperspectral Data

The Master Tree approach was applied to both KSC and Bolivar Peninsula data for transferring an ensemble of classifiers to regions where ground truth is not immediately available. For both data sets, the ensemble consists of 21 different classifiers. Using multiple 50% stratified samples of the available ground truth, 10 TD Adaptive BB-BHC classifiers and 10 BU Adaptive BB-BHC classifiers were trained and then transferred, with parameter updates, for classification of the alternate test sites from which they were trained. Lastly, all of the ground truth was used to train the Master Tree, and it was also transferred for classification of the alternate site. The sub-sample classifiers were combined using the voting method to obtain a unique predicted label. The voted prediction was compared to the vote of the Master Tree and, if they were in disagreement, the vote of the Master Tree was adopted only if the posterior probability for that observation was higher than the percentage of the sub-sample classifier that had agreed on the voted prediction.

Bolivar Peninsula

Both ensembles, when transferred from Site 1 to Site 2 and when transferred from Site 2 to Site 1, have a higher classification rate than the similar TD or BU classifiers that did not utilize the ensemble and Master Tree method. The accuracies are reported in Table 5.1. The classifier trained on Site 2 and tested on Site 1 performs really well, correctly identifying the label over 93% of the time. Unfortunately, the classifier trained on Site 1 and tested on Site 2 still has a difficult time with a classification rate of just over 75%. The precision trees, the confusion matrices, and the "separation" distance and purity measures

From Region	To Region	Classifier	Classification Accuracy
Bolivar 1	Bolivar 2	TD	0.7492
Bolivar 1	Bolivar 2	BU	0.7576
Bolivar 1	Bolivar 2	Master Tree	0.7585
Bolivar 2	Bolivar 1	TD	0.9102
Bolivar 2	Bolivar 1	BU	0.6299
Bolivar 2	Bolivar 1	Master Tree	0.9342

Table 5.1: Bolivar Peninsula classification accuracies when classifier is updated using pseudo-labeled data to estimate the new parameters and applied to the alternate site at which the ground truths were acquired.

were evaluated to gain additional insight [Appendix I]. For Site 1, the precision tree can be used to quickly identify the partition of L_1 (Water) and L_2 (Low Proximal Marsh) as being difficult. Fortunately, analysis of the distance measure and the purities probably would lead a researcher to find this problem because the

perceived separation between the two classes has dropped by over 50%, and the purity of both classes is ~50% [Figure I.1]. Unfortunately, while the partitioning of L_1 and L_2 is the source of almost all the error at Site 2, the distance measures fail to indicate this may be the case as the lowest purity measure is over 82 percent and the biggest decrease in the distance measure is less than 40% [Figure I.2]

Kennedy Space Center

The classification accuracies obtained at KSC are contained in Table 5.2. For this application area, the ensemble results do not outperform the original TD

From Region	To Region	Classifier	Accuracy on Identical Classes	Accuracy on All Classes
KSC 1	KSC 2	TD	0.4886	0.4085
KSC 1	KSC 2	BU	0.4481	0.3746
KSC 1	KSC 2	Master	0.4029	0.3386
KSC 2	KSC 1	TD	0.4500	0.3762
KSC 2	KSC 1	BU	0.4682	0.3914
KSC 2	KSC 1	Master	0.4577	0.3826

Table 5.2: KSC classification accuracies when classifier is updated using pseudo-labeled data to estimate the new parameters and applied to the alternate site at which the ground truths were acquired.

or BU transferred classifiers. However, analysis of the distance and purity measures indicates that this approach is still very beneficial. For example, at Site

1 [Figure I.3], the distance measure for the partitioning of $\Omega = \{L_{13}, L_{15}\}$ is just 4% of the distance measured on the training data. This would indicate some severe problems and may lead a researcher to discover the absence of L_{15} from Site 1. Similar insight can be gained at Site 2 [Figure I.4], the distance measure from partitioning $\Omega = \{L_{10}, L_{11}, L_{13}\}$ has increase 1500%, indicating a very severe problem that can be attributable to the absence of L_{10} from Site 2. Furthermore, if $\Omega = \{L_{10}, L_{11}, L_{13}\}$ is considered as a unique label rather than 3 sub-labels, the accuracy improves by over 13.5%.

5.4 CONCLUDING REMARKS

While these sub-sampling methods are not new, they have not previously been applied to a hierarchical classifier in the context of the limited training data problem. However, the new Adaptive BB-BHC framework makes it possible to use sub-sampling techniques to create an ensemble of classifiers. Furthermore, the Master Tree structure helps preserve the interpretability of inter-class relationship, an important factor in this domain, and is a useful tool, along with the distance and purity measures, for identifying when the resolution of the output classes being sought is too fine for knowledge transfer.

Chapter 6: Concluding Remarks

Many classification problems involve a high dimensional input space and several possible output classes. While such problems are challenging theoretically, numerically, and computationally, they potentially provide flexibility via decomposition that can be utilized by classification methods to achieve a high level of accuracy. Classification algorithms must accommodate these issues to fully exploit the additional information a high dimensional input space provides. While it is desirable to develop classification techniques in as general a framework as possible so that they can be applied to a wide variety of domains, the quantity and quality of available data for training and testing must be a primary factor during model development. Ultimately, high levels of accuracy may be achieved consistently on these difficult problems only if extensive domain knowledge is incorporated. Generalization can be improved by techniques that seek to automatically discover this critical information.

6.1 Summary of Contributions

This research focused on the development of a hierarchical approach for classification that is robust with respect to training data that is limited both in quantity and spatial extent.

6.1.1 Limited quantity of training data

Many difficult classification problems involve a high dimensional input space. Due to the "curse of dimensionality", it is necessary to reduce the size of the input space when there is only a limited quantity of training data available. A

new approach was developed that preserves as much of the discriminatory "power" of the data as possible, conditioned on the actual quantity of data available. The ability of this technique to mitigate the degradation of classifier performance when the quantity of training data is reduced was demonstrated.

6.1.3 Spatially limited training data

Spatially limited training data can result in poor inference concerning the true populations. The detrimental impact that can result, if this issue is ignored, is explored and demonstrated. Furthermore, it is shown that the problem can be viewed as one of a "spatially limited" acquisition of training data. This insight is critical for achieving successful classification of areas where no training data are available. Rather than beginning the classification process afresh without any training data, which would be required if the two populations were totally different, viewing two samples as just spatially deformed versions from the same population indicates that the classifier trained on one sample may be of use on the alternate sample. This discovery led to the development of a dynamic algorithm that automatically updates parameter estimates. Transferal of information that was previously acquired, such as the discriminatory feature space and output space, is used to form clusters representative of the deformed classes which in turn are used to update the parameter estimates of the transferred classifier.

6.1.4 Ensemble of classifiers

Independent of limited training data, both in terms of the spatial implications and limited quantity, different sampling subsets of the same ground

truth may result in slightly different classifiers. This issue was not previously addressed rigorously. The advantages gained by using an ensemble of classifiers built from sub-samples of training data are widely acknowledged but have not previously been used in the context of a hierarchical classifier for remotely sensed data. The ability of the adaptive BB-BHC now makes sub-sampling a viable option. Using this technique, an ensemble of classifiers is used to identify a Master Tree that provides interpretability to the ensemble and is more robust for classifier transferal. Furthermore, tools are provided that help identify a suitable meta-class level of the transferred classifier for situations where the full resolution of the output space may not be appropriate.

6.2 Future Work

The difficult and exciting application area of classification offers a wide variety of research problems. Within the application area of land cover/land use mapping and monitoring, classification is of growing importance due to heightened interest in global ecological monitoring that must be performed using remote sensing technologies. Hyperspectral data can potentially contribute to capability for discriminating between targets that has heretofore been impossible. Although it is operationally a new technology, the quantity and variety of hyperspectral data available are growing rapidly. While most of these data are still being acquired from airborne platforms, a hyperspectral sensor (Hyperion) is now successfully acquiring hyperspectral data, and other missions are planned internationally. New methodology will be required to effectively extract

information from these sensors. Several extensions to the current work could be pursued in the immediate future.

6.2.1 Feature selection

In the BHC framework, feature selection has not been explored. Feature selection at each partition of the BHC could be accomplished similarly to the way features were selected for the BPC framework. Feature selection, rather than feature extraction, would help preserve the interpretability of the feature space. The method for transferring classifiers would be identical to that described for the Fisher projected space considered in this research. Clusters could be formed in the reduced feature space and matched to the respective meta-classes so that a measure of total distance between the clusters and old classes, in that feature space, is minimized. However, as more features are required to attain a higher level of classification accuracy, the updating will become more challenging and it will be difficult to visualize the results during algorithm development.

6.2.2 Unsupervised clustering

Unsupervised clustering on the entire image is necessary when no ground truth is available. A large amount of research has been completed in the unsupervised classification/clustering area with applications to spectral data [85-87, 95, 96]. Additionally, the basic ideas of unsupervised clustering have been extended to account for spatial relationships of the pixels [92-94]. While accounting for spatial information is a desired characteristic, the added computational effort may be prohibitive, and selection of the specific approach

should compromise between the computational requirements and the goal of obtaining spatially and spectrally contiguous clusters. An additional concern is determining the number of clusters (cluster validation) [88]. However, given the intended usage of the clusters, it is more advantageous to allow more clusters in hopes that they will be more "class pure" rather than a smaller number of "mixed" clusters. Furthermore, at this stage it may be advantageous to eliminate or at least hold in reserve some of the *N* clusters from immediate consideration in classifier development either due to their size or their perceived potential to be "class impure".

6.2.3 Deformable models

The methodology of using the features that were determined to be highly discriminatory in the training data to form clusters and update parameter estimates in a new testing site may not be very robust. Features that were the "best" in the old region may no longer be suitable for discrimination. Having access to a larger number of data sets would allow the investigation of which features may not be the best for any one area, but are suitable for all of the areas. This additional domain knowledge could be utilized by comparing and matching the signatures obtained from the "new" region to the existing spectral library derived from previous training data. Automation of the matching problem has been addressed by a significant amount of research in the fields of signal processing and time-series data [101-104]. However, despite the high degree of correlation between adjacent bands, no research has investigated the possibility that the hyperspectral bands could be treated as points in a signal/time series and applied pattern-

matching algorithms to identify new "training" data. A simple approach to this problem would be to develop a distance measure to estimate the similarity of the old spectral label signatures and the new clusters. However, construction of an effective distance measure can be highly non-trivial, very problem-dependent, inflexible to "deformations" in the patterns, and cannot quantify the certainty associated with a "match".

Various distance-based methods have been modified to be more "flexible" but they still suffer from the problems inherent in distance measures [102, 103]. In the *probabilistic generative modeling* approach proposed by Ge and Smyth [101], a model M_Q is constructed for pattern Q that typically consists of both a mean shape and a distribution function that describes variation about the mean shape. Therefore, the similarity of new patterns to the original pattern can be computed by $p\left(Q_{\text{new}} \middle| M_Q\right)$ and this approach could be applied for identifying deformed spectral signatures by comparison to existing spectral libraries.

Appendix A

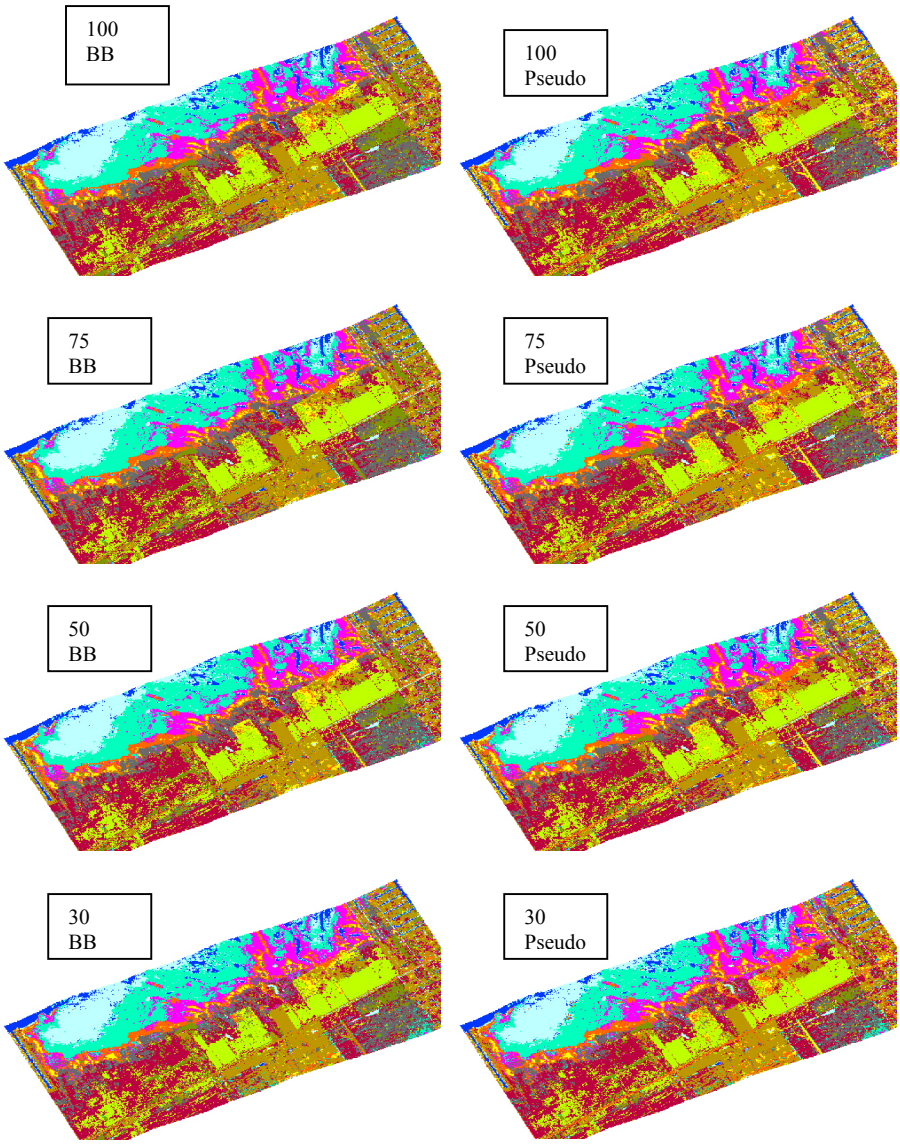


Figure A.1: Bolivar Peninsula BU-BHC classified images (sampling %, BB vs Pseudo)

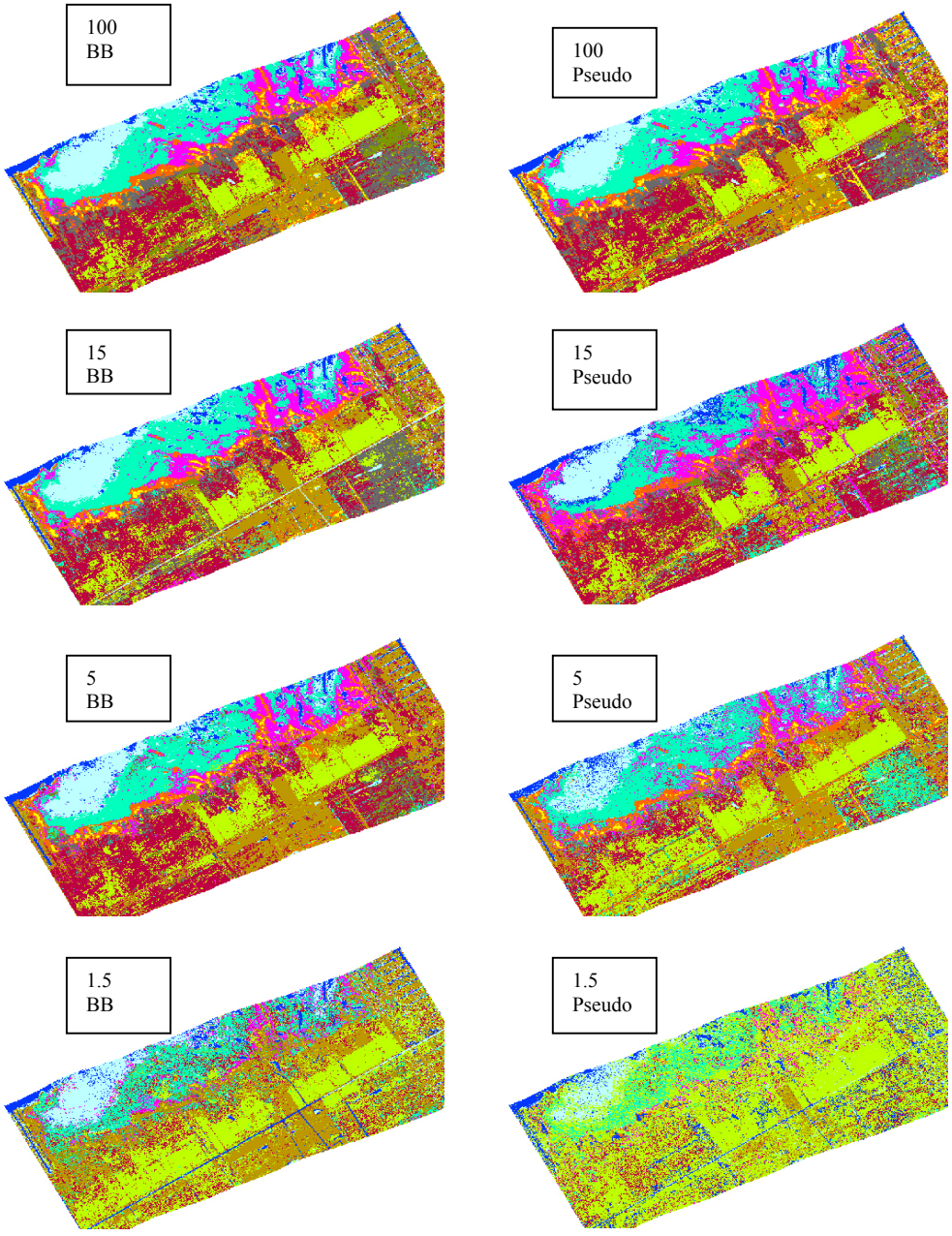


Figure A.2: Bolivar Peninsula BU-BHC classified images (sampling %, BB vs Pseudo)

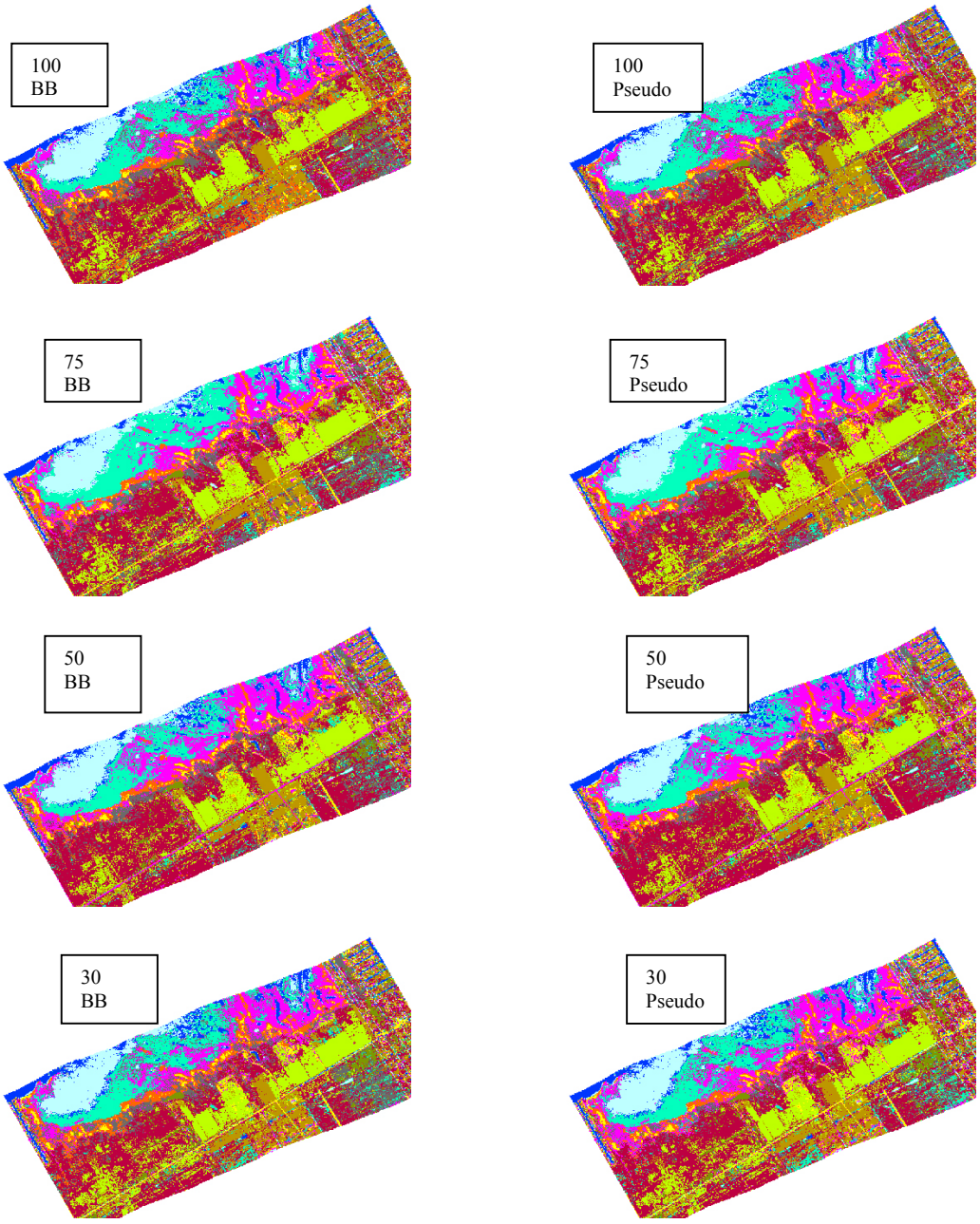


Figure A.3: Bolivar Peninsula TD-BHC classified images (sampling %, BB vs Pseudo)

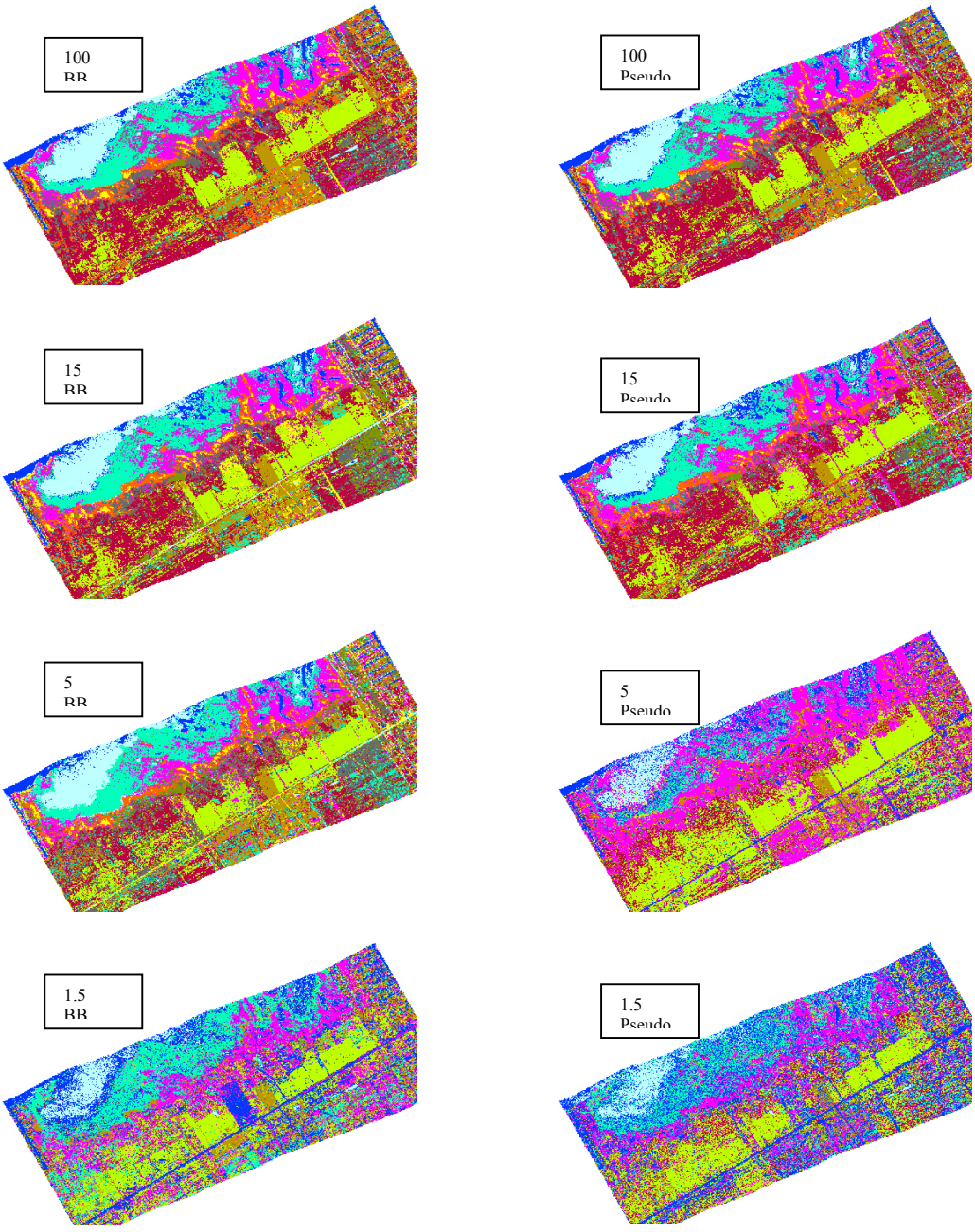


Figure A.4: Bolivar Peninsula TD-BHC classified images (sampling %, BB vs Pseudo)

Appendix B

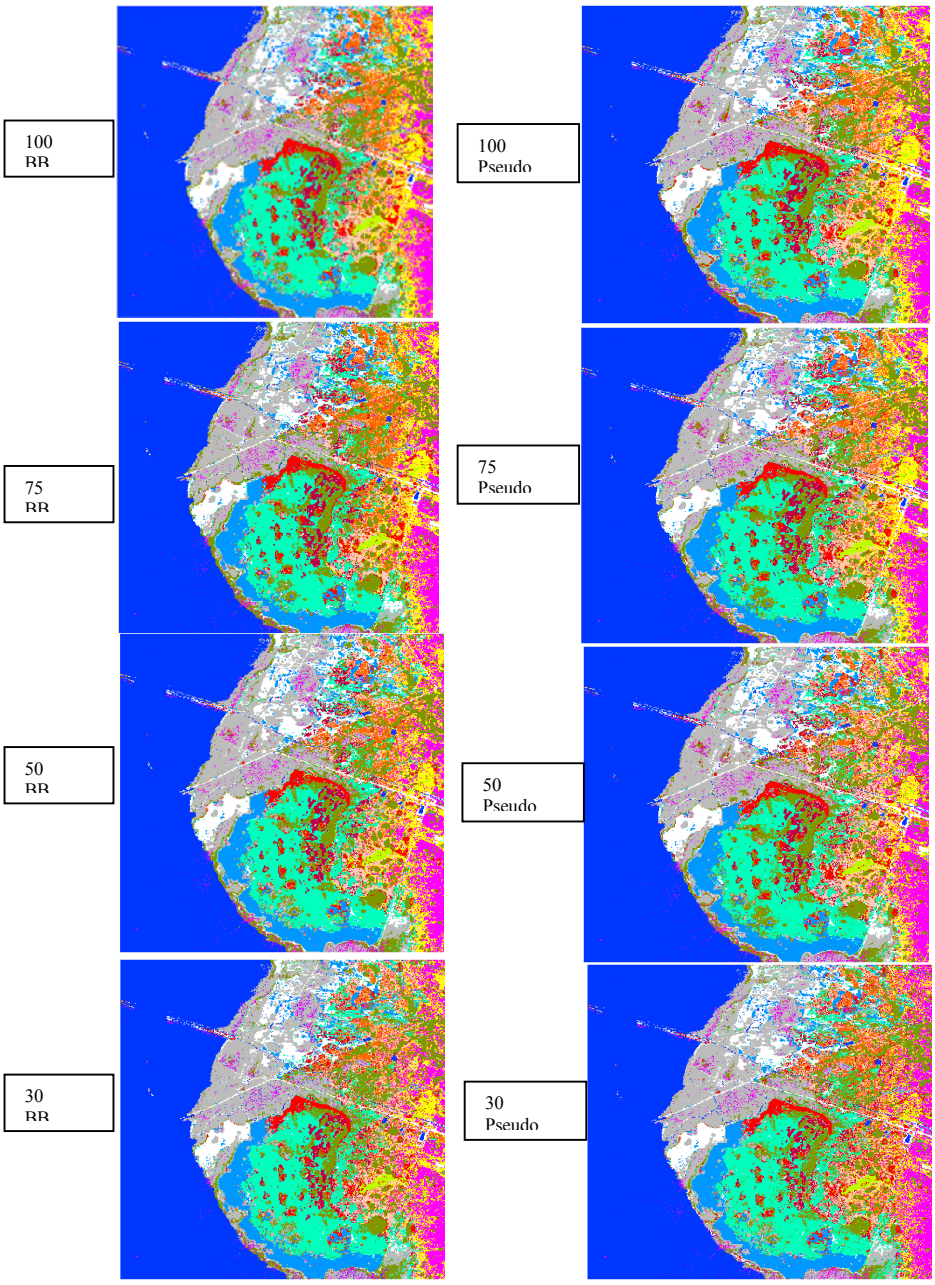


Figure B.1: KSC BU-BHC classified images (sampling %, BB vs Pseudo)

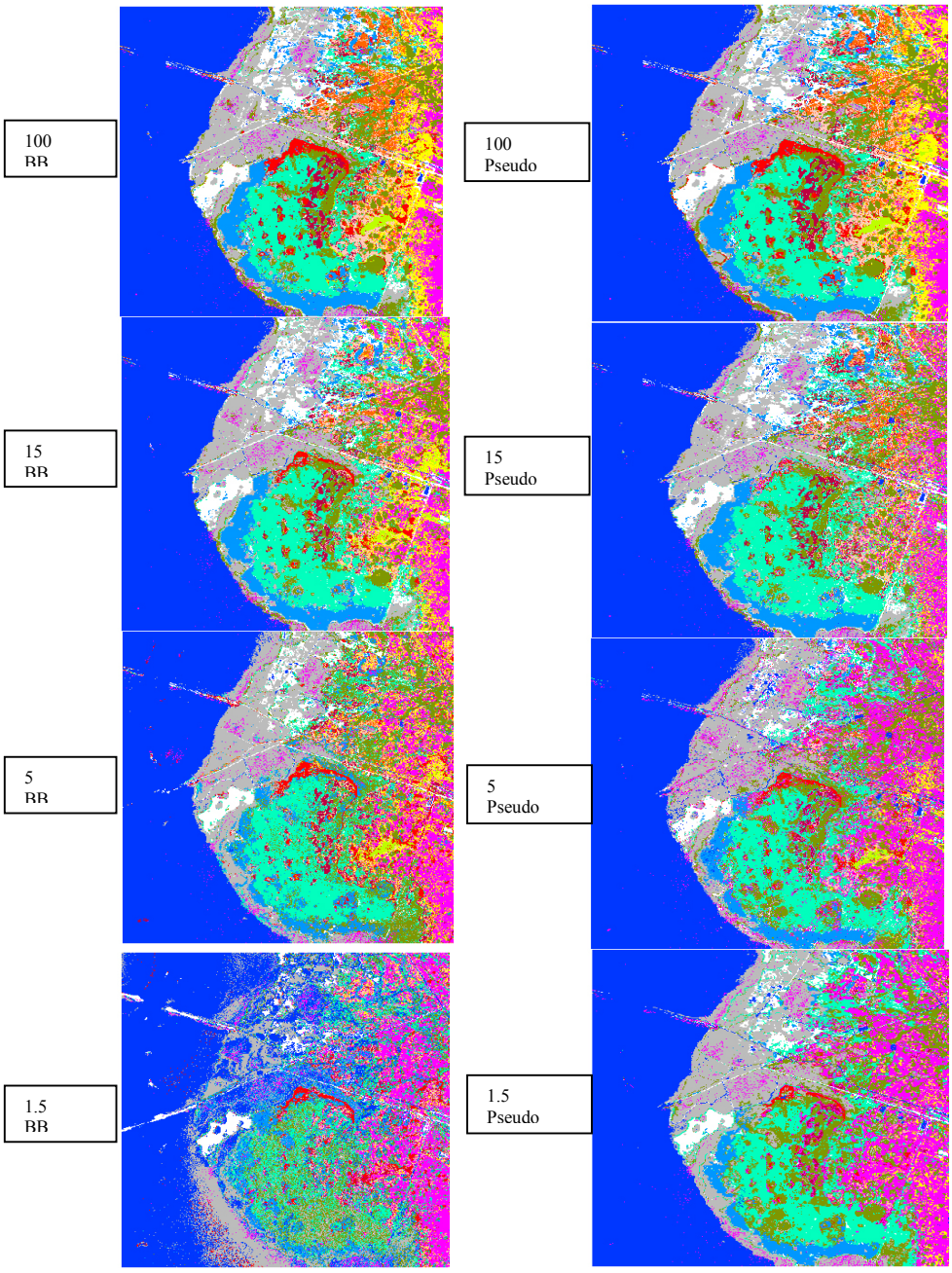


Figure B.2: KSC BU-BHC classified images (sampling %, BB vs Pseudo)

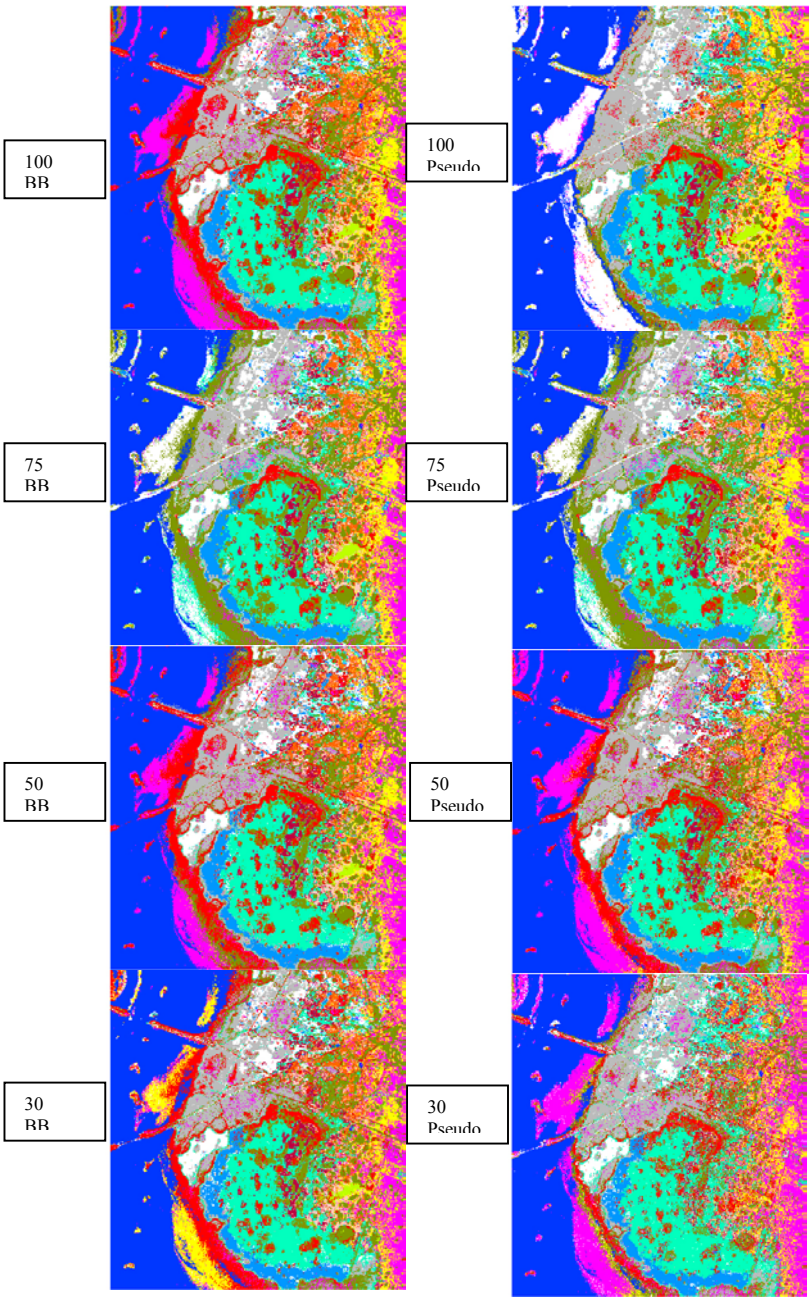


Figure B.3: KSC TD-BHC classified images (sampling %, BB vs Pseudo)

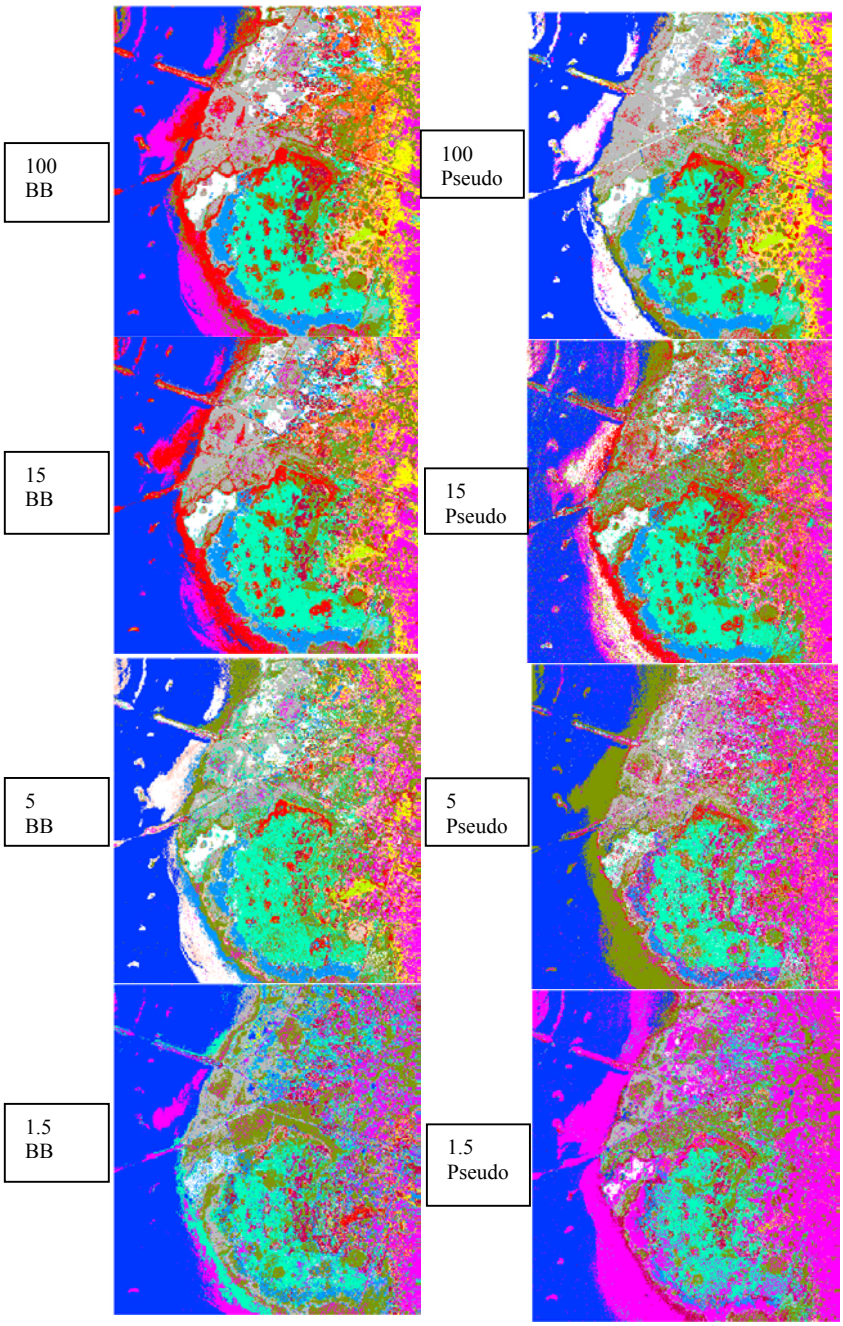
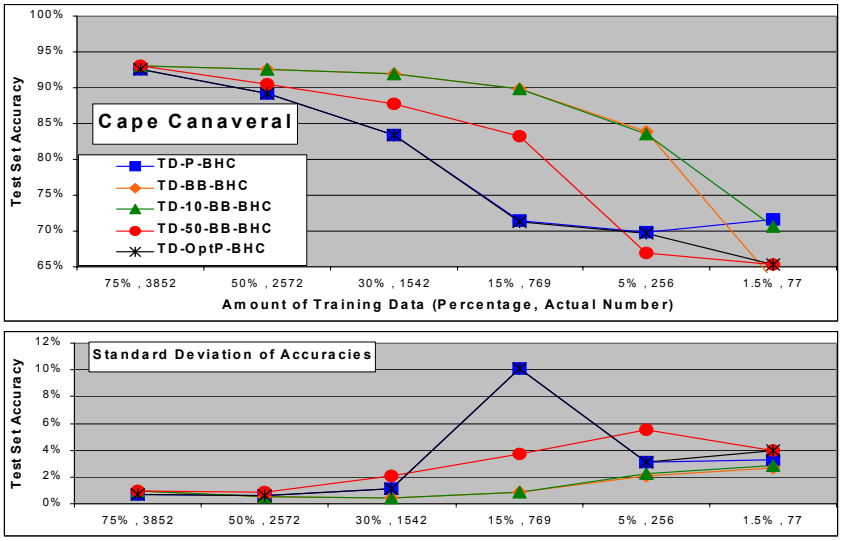
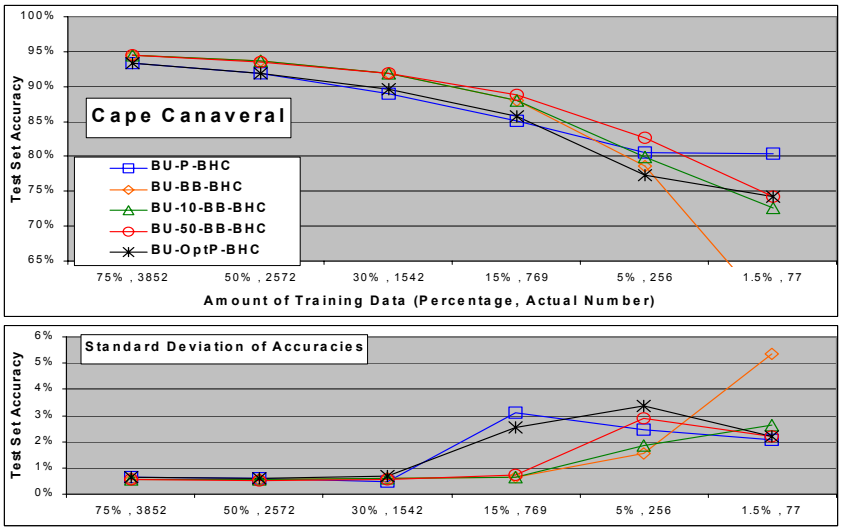


Figure B.4: KSC TD-BHC classified images (sampling %, BB vs Pseudo)

Appendix C

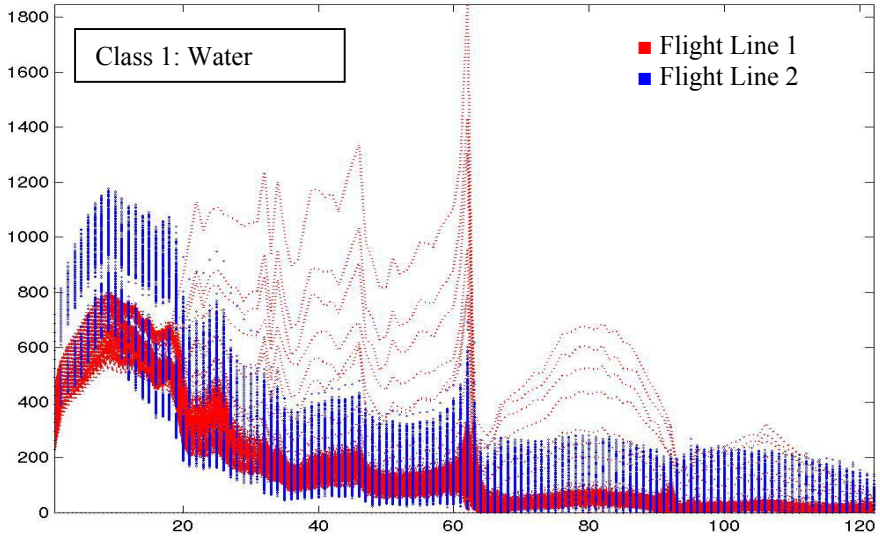


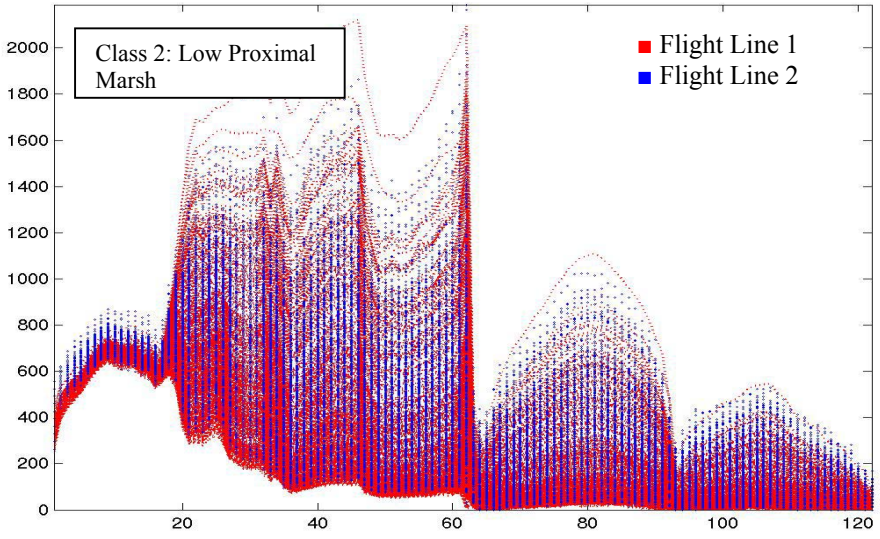
Classification (test set) accuracies for KSC, Cape Canaveral FL with adaptive Bottom-Up BHC and Pseudo

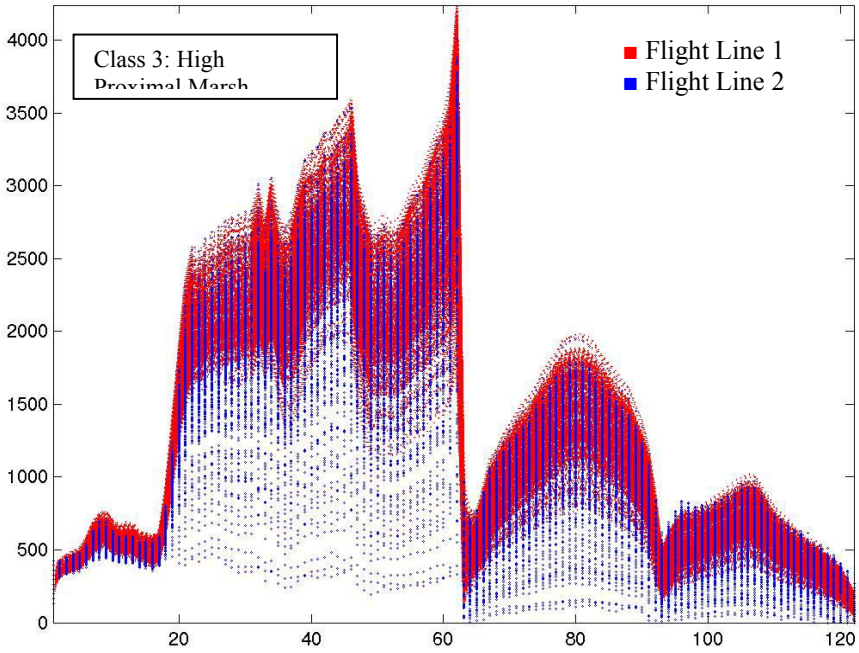


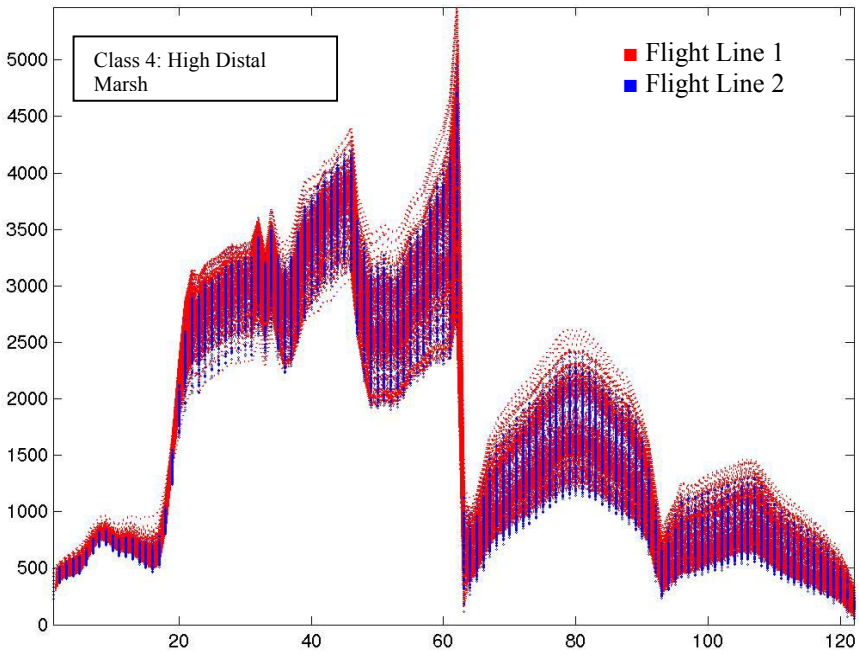
Classification (test set) accuracies for KSC, Cape Canaveral FL with adaptive Top-Down BHC and Pseudo

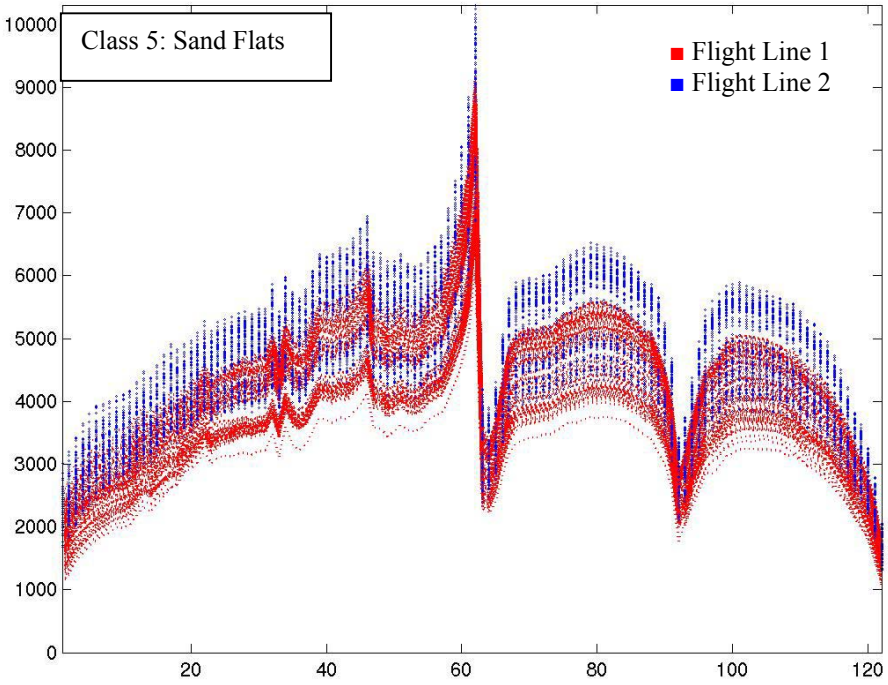
Appendix D

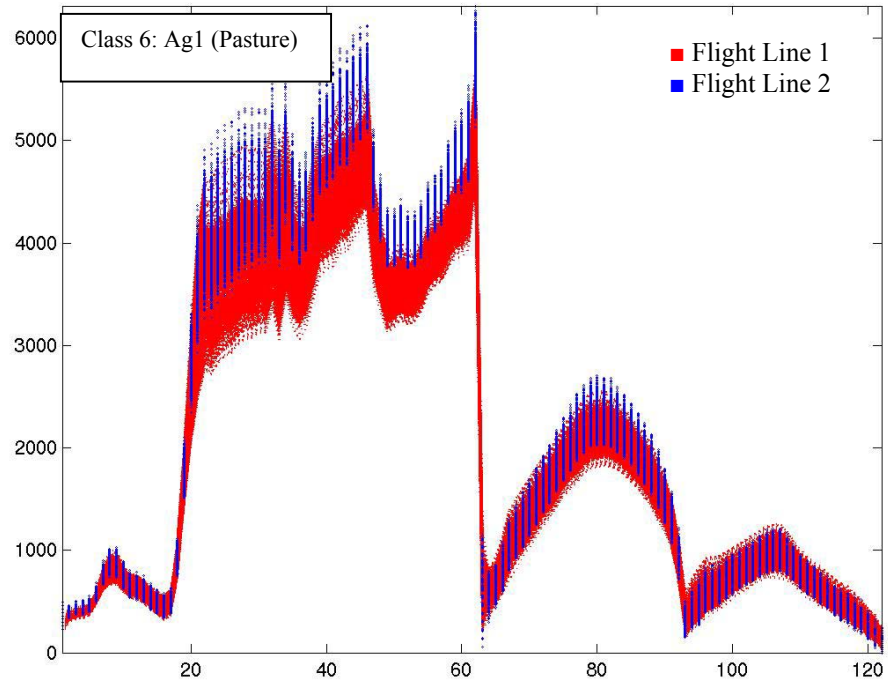

Class 1: Water
Flight Line 1
Flight Line 2
1800
1600
1400
1200
1000
800
600
400
200
0
20
40
60
80
100
120

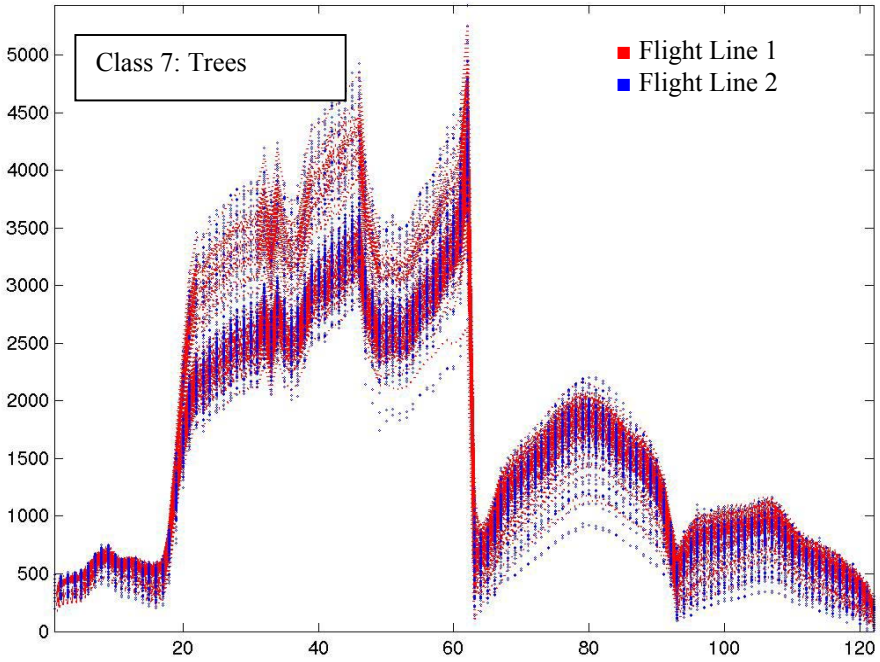

Class 2: Low Proximal Marsh
Flight Line 1
Flight Line 2
2000
1800
1600
1400
1200
1000
800
600
400
200
0
20
40
60
80
100
120

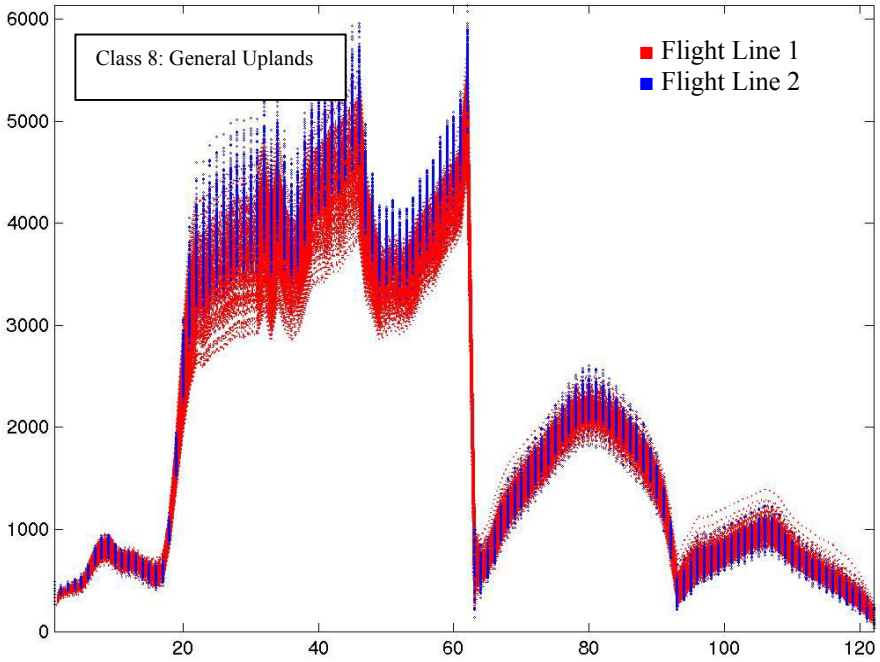

Class 3: High
Proximal Marsh
Flight Line 1
Flight Line 2
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120

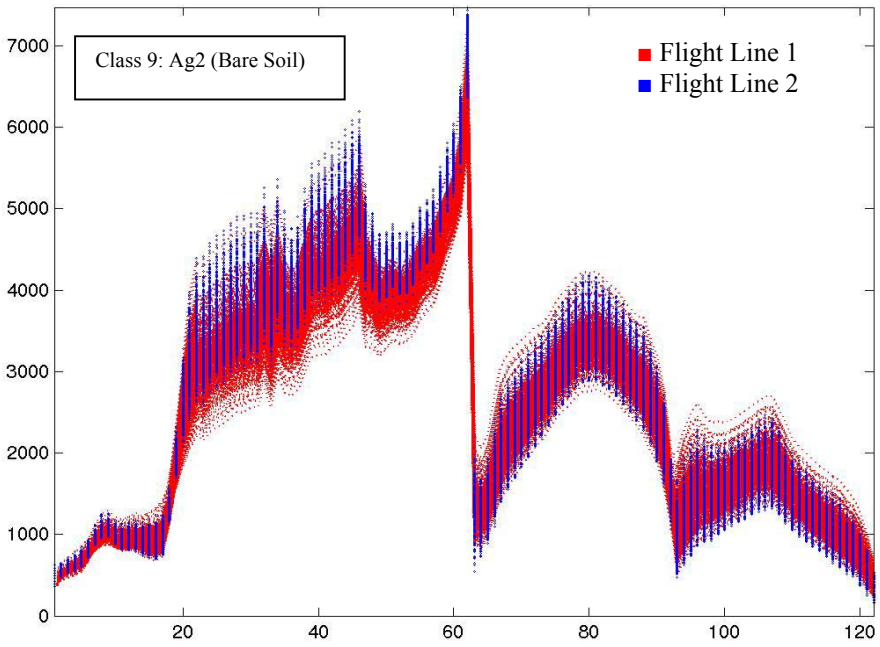

Class 4: High Distal
Marsh
Flight Line 1
Flight Line 2
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120

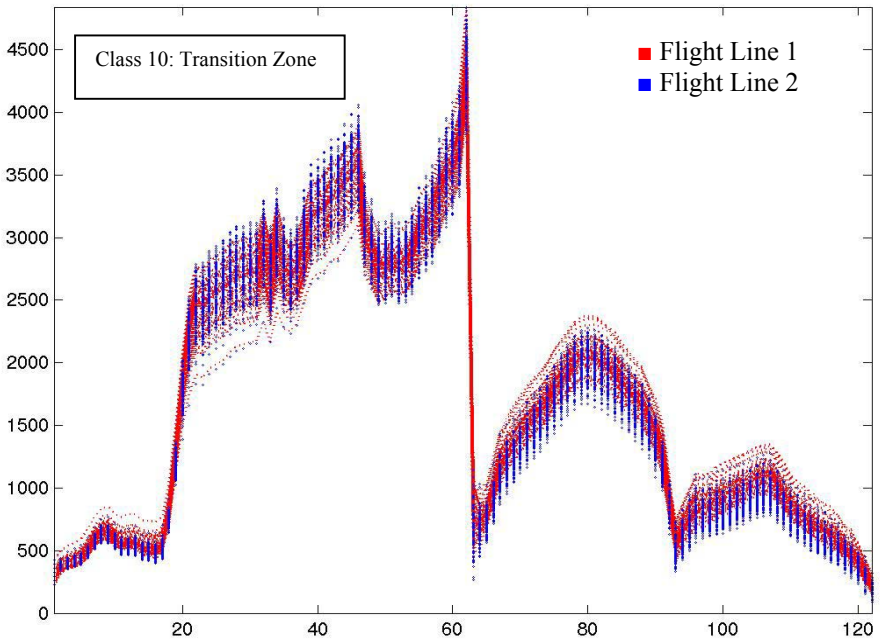

Class 5: Sand Flats
Flight Line 1
Flight Line 2
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
20
40
60
80
100
120

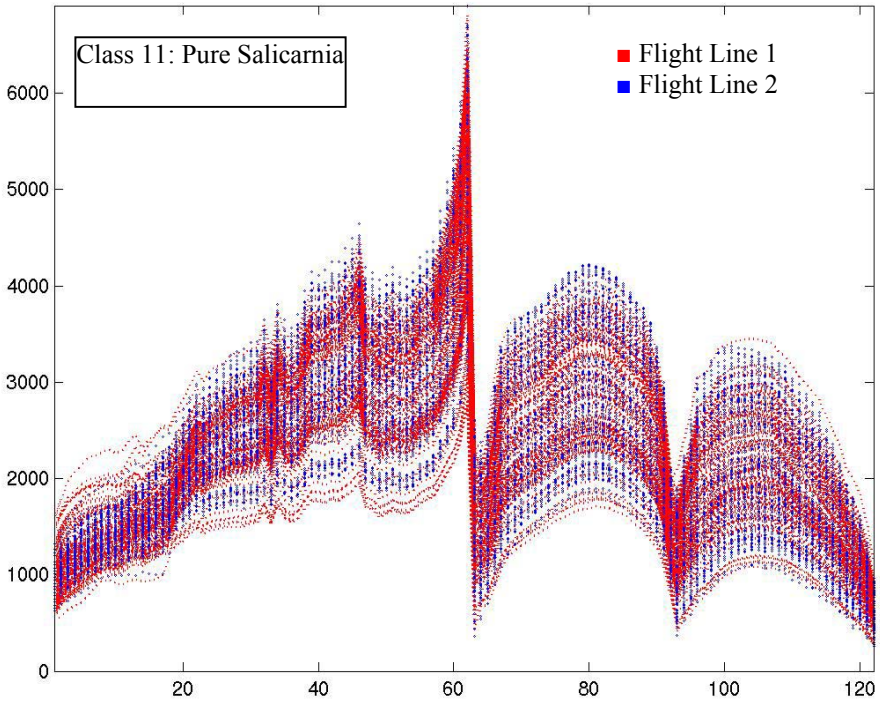

Class 6: Ag1 (Pasture)
Flight Line 1
Flight Line 2
6000
5000
4000
3000
2000
1000
0
20
40
60
80
100
120


Class 7: Trees
Flight Line 1
Flight Line 2
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120

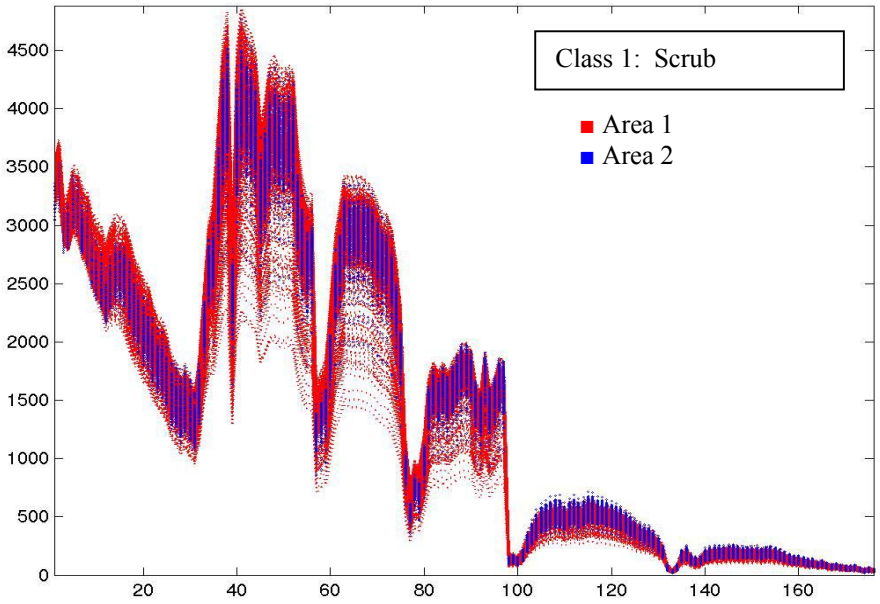

Class 8: General Uplands
Flight Line 1
Flight Line 2
6000
5000
4000
3000
2000
1000
0
20
40
60
80
100
120

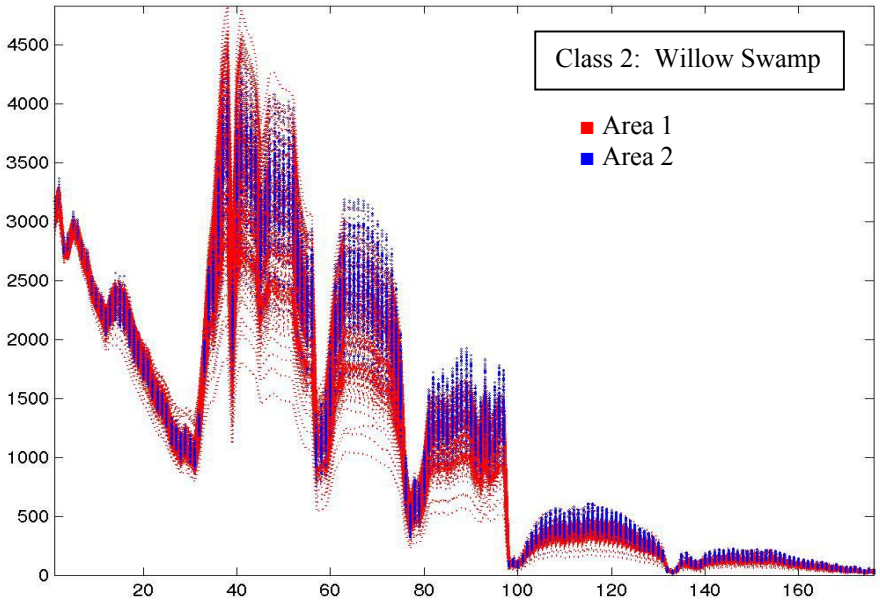

Class 9: Ag2 (Bare Soil)
Flight Line 1
Flight Line 2
7000
6000
5000
4000
3000
2000
1000
0
20
40
60
80
100
120

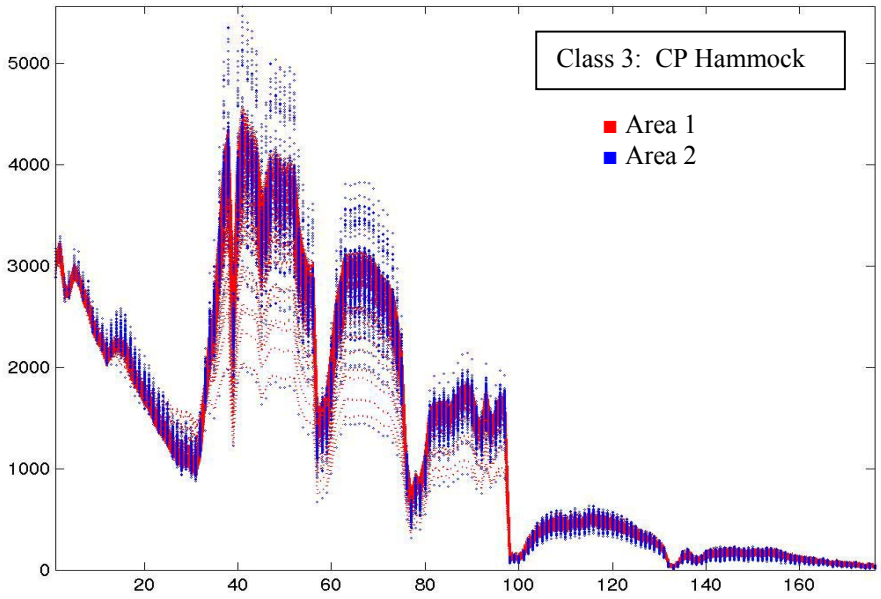

Class 10: Transition Zone
Flight Line 1
Flight Line 2
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120

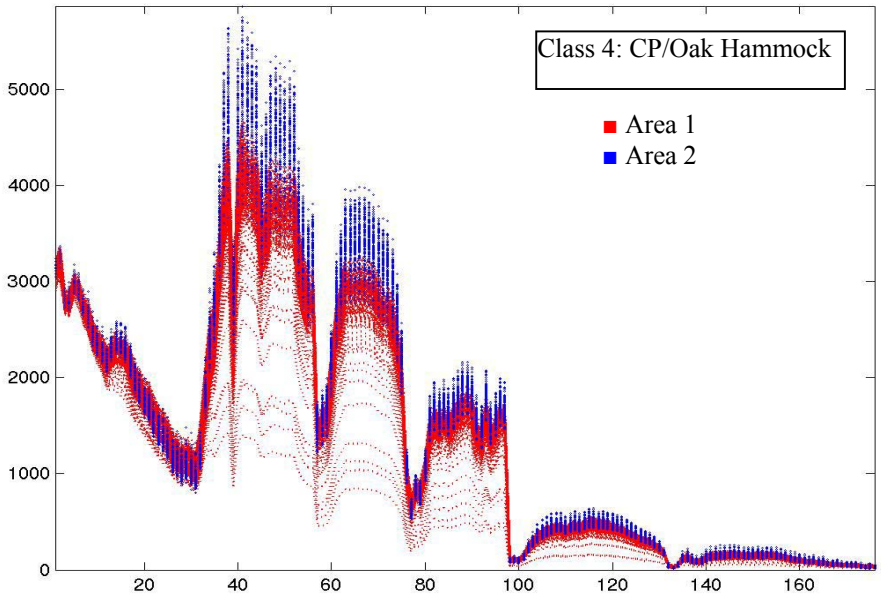

Class 11: Pure Salicarnia
Flight Line 1
Flight Line 2
6000
5000
4000
3000
2000
1000
0
20
40
60
80
100
120

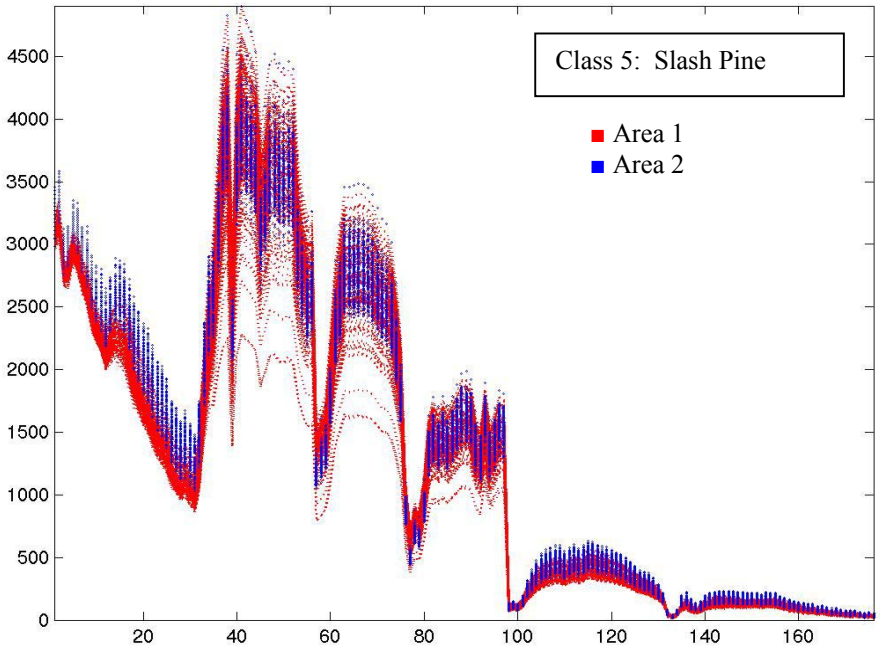
Appendix E

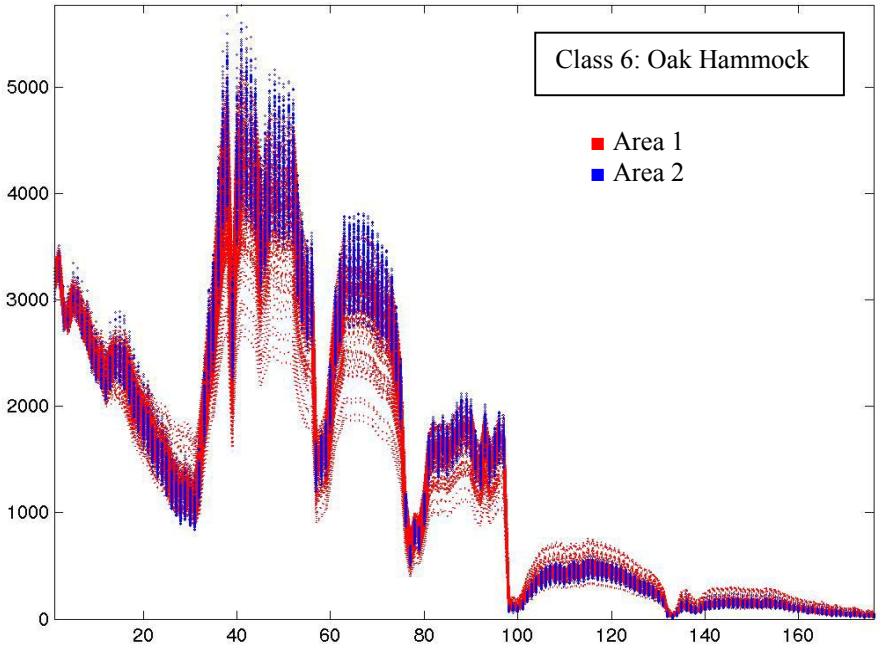

Class 1: Scrub
Area 1
Area 2
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

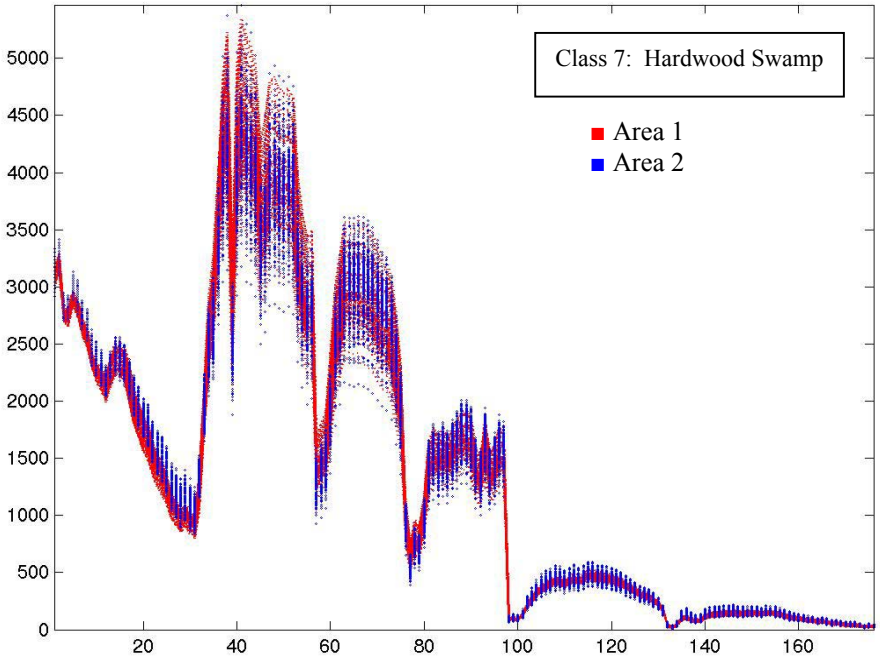

Class 2: Willow Swamp
Area 1
Area 2
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

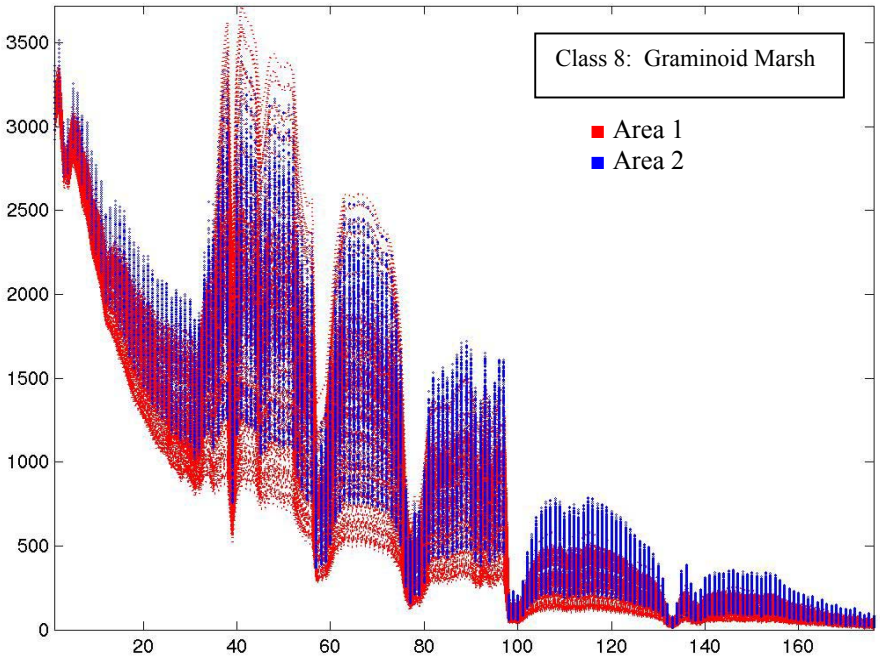

Class 3: CP Hammock
Area 1
Area 2
5000
4000
3000
2000
1000
0
20
40
60
80
100
120
140
160

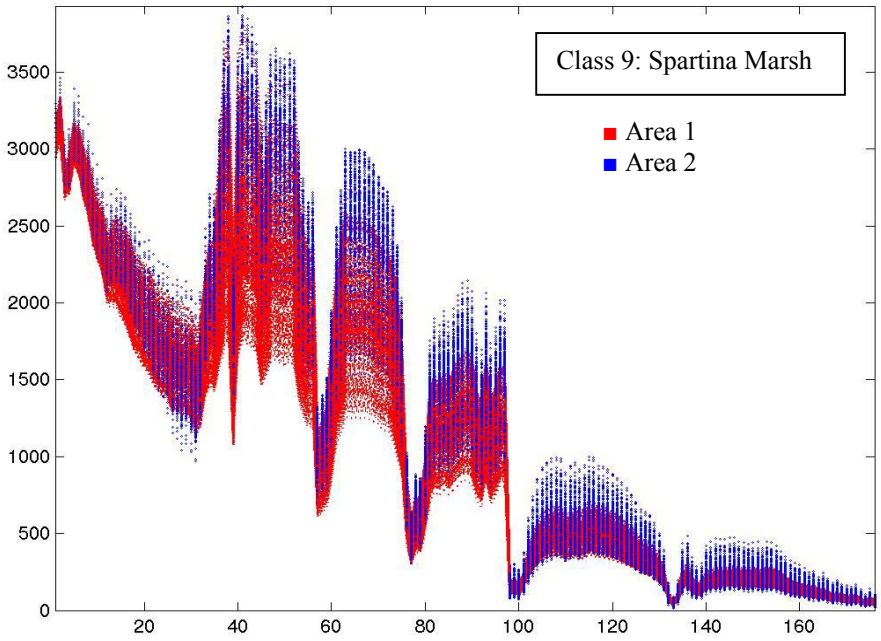

Class 4: CP/Oak Hammock
Area 1
Area 2
5000
4000
3000
2000
1000
0
20
40
60
80
100
120
140
160

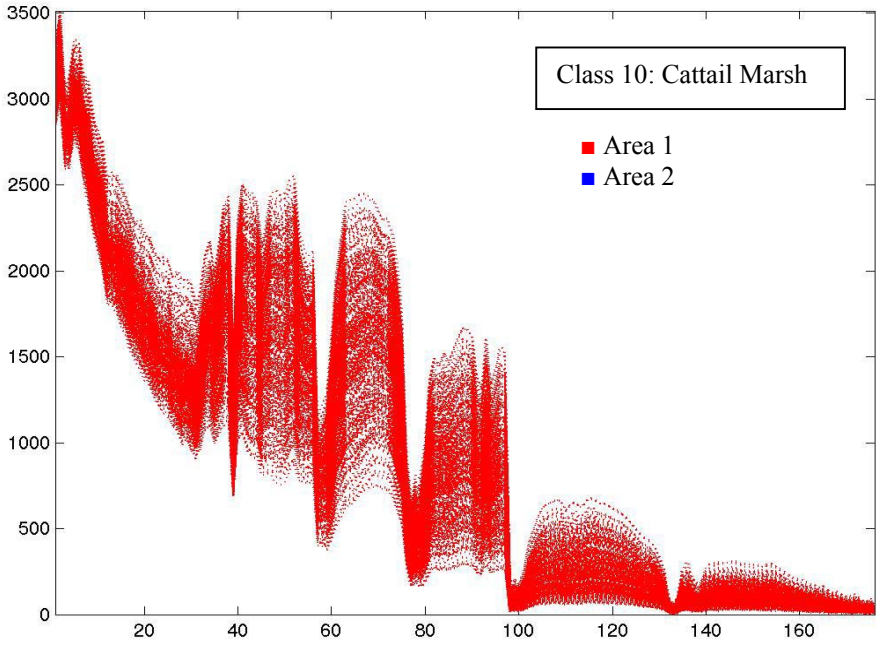

Class 5: Slash Pine
Area 1
Area 2
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

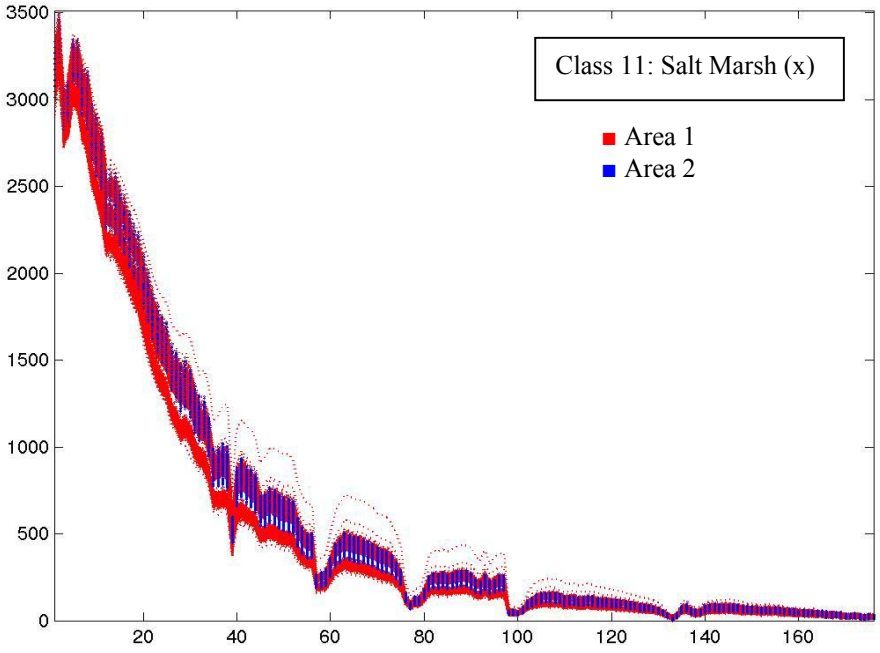

Class 6: Oak Hammock
Area 1
Area 2
5000
4000
3000
2000
1000
0
20
40
60
80
100
120
140
160

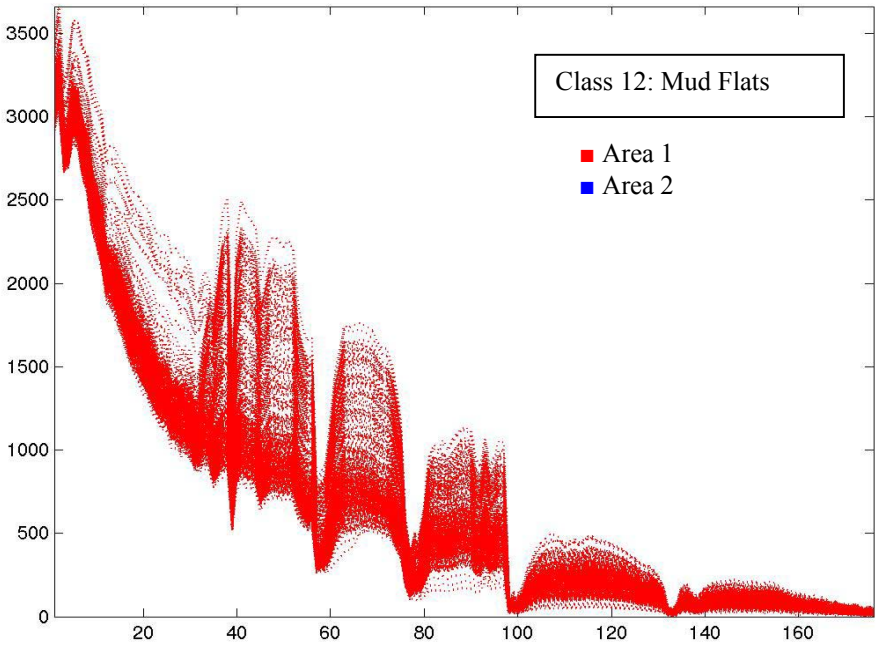

Class 7: Hardwood Swamp
Area 1
Area 2
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

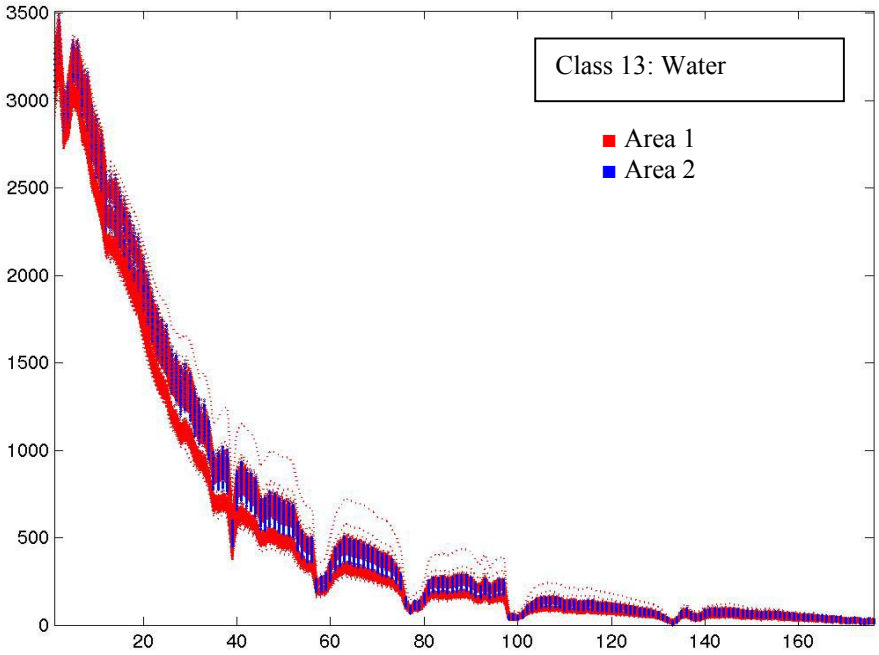

Class 8: Graminoid Marsh
Area 1
Area 2
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

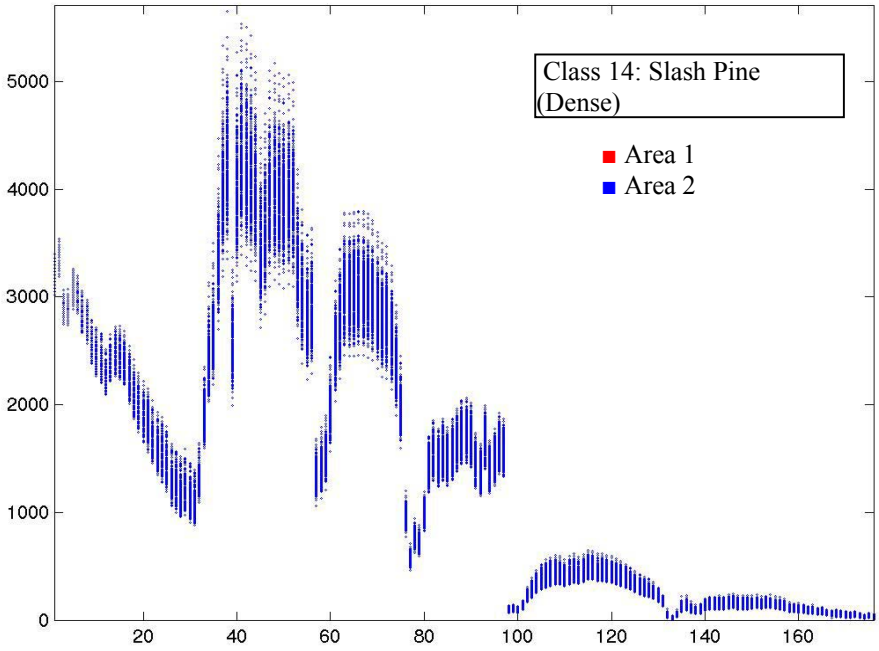

Class 9: Spartina Marsh
Area 1
Area 2
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

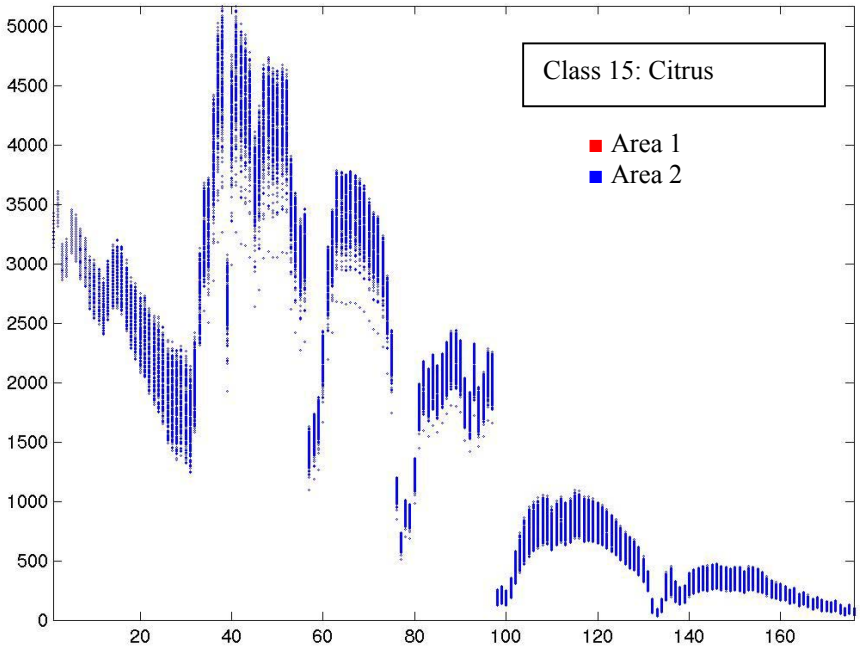

Class 10: Cattail Marsh
Area 1
Area 2
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

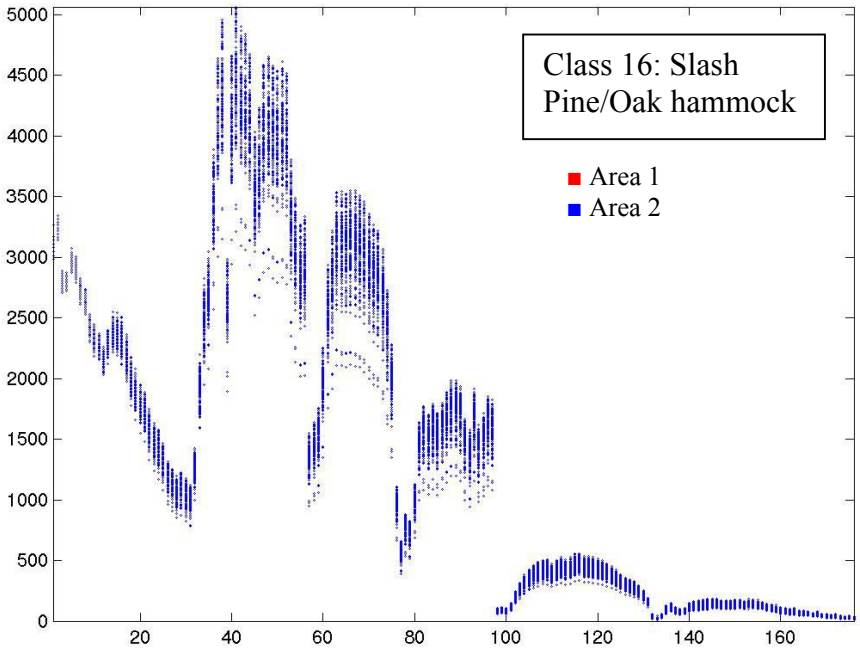

Class 11: Salt Marsh (x)
Area 1
Area 2
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160


Class 12: Mud Flats
Area 1
Area 2
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

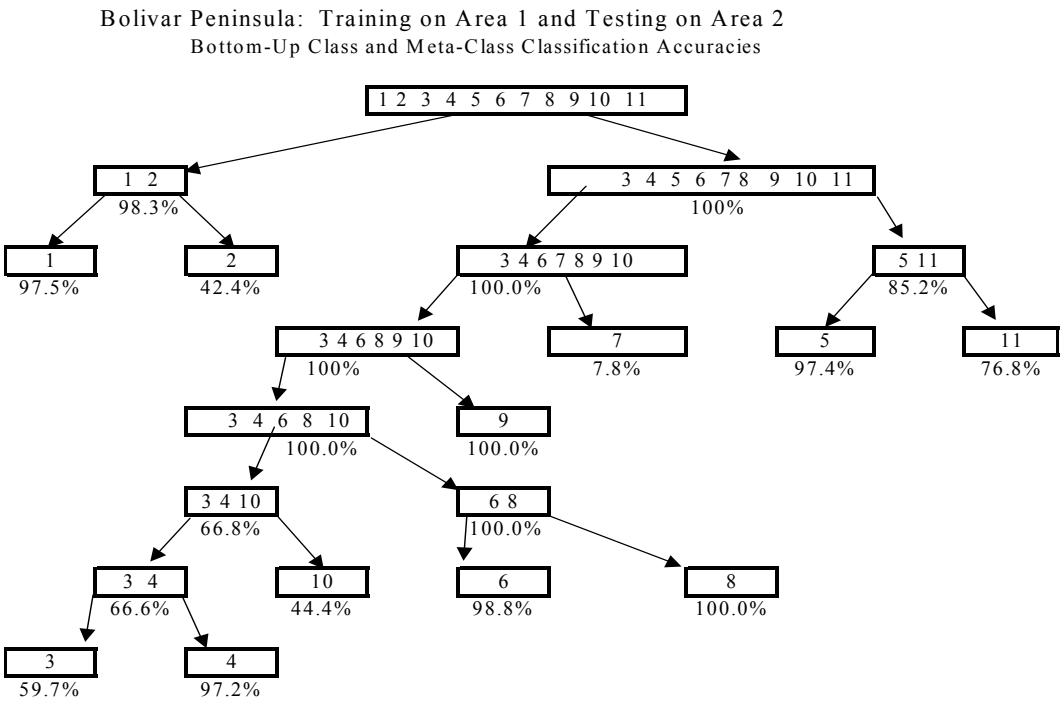

Class 13: Water
Area 1
Area 2
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160


Class 14: Slash Pine (Dense)
Area 1
Area 2
5000
4000
3000
2000
1000
0
20
40
60
80
100
120
140
160


Class 15: Citrus
Area 1
Area 2
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160


Class 16: Slash Pine/Oak hammock
Area 1
Area 2
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
120
140
160

Appendix F



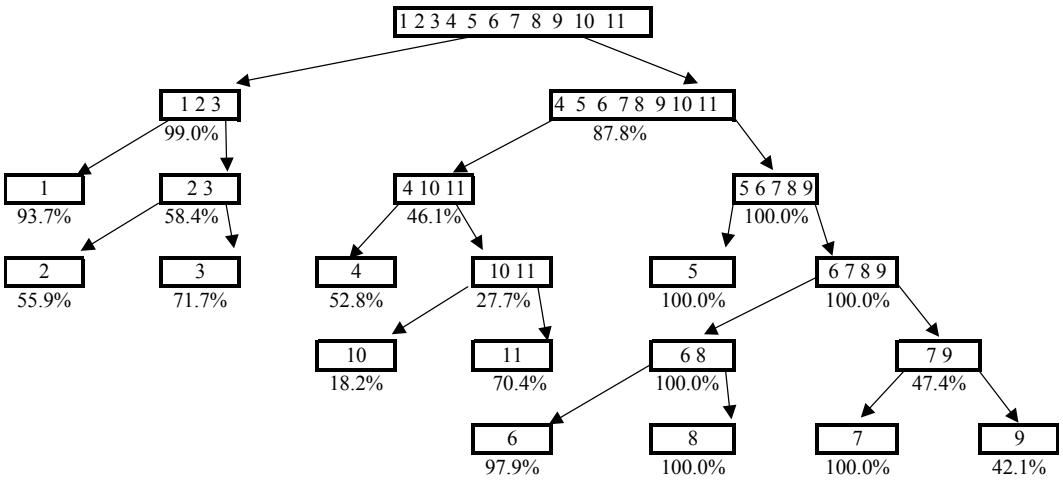
PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	
1	3764	765	0	0	0	0	0	0	0	0	0	4529
2	52	595	0	0	0	0	0	0	0	0	0	647
3	44	44	713	0	0	0	282	0	0	0	0	1083
4	0	0	28	173	1	0	292	0	0	0	0	494
5	0	0	0	0	112	0	0	0	0	0	0	112
6	0	0	307	2	0	1241	884	0	0	20	0	2454
7	0	0	0	0	2	0	236	0	0	0	0	238
8	0	0	113	3	0	15	313	35	0	55	0	534
9	0	0	0	0	0	0	920	0	176	0	31	1127
10	0	0	34	0	0	0	108	0	0	60	8	210
11	0	0	0	0	0	0	0	0	0	0	129	129
	3860	1404	1195	178	115	1256	3035	35	176	135	168	11557

Figure F.1 Bolivar Peninsula Precision Tree and Confusion Matrix for BU-BHC, trained on Site 1 and tested on Site 2

Bolivar Peninsula: Training on Area 1 and Testing on Area 2

Top-Down Class and Meta-Class Classification Accuracies



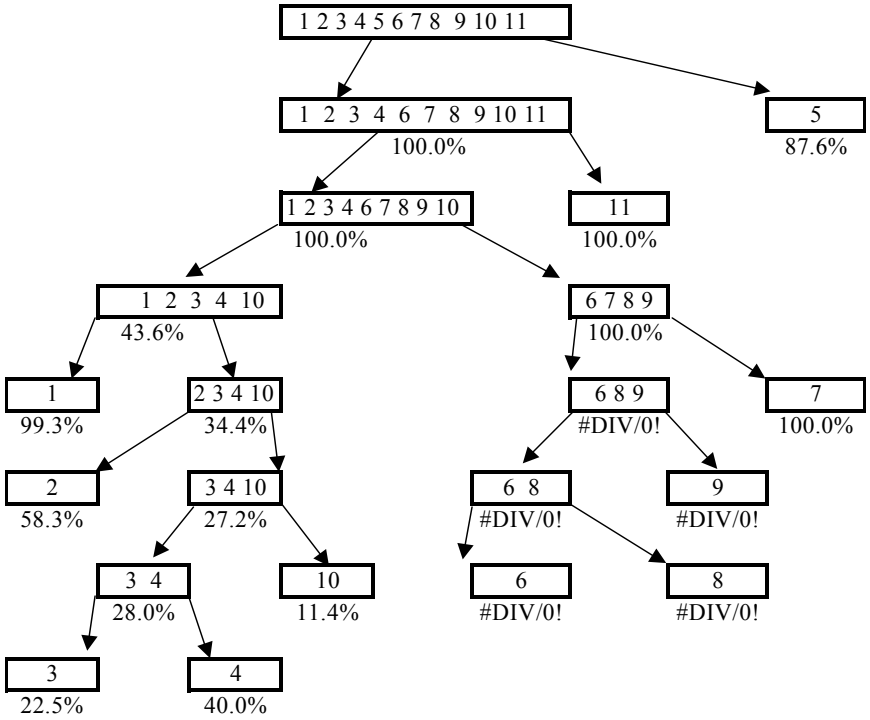
PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	
1	4048	471	0	10	0	0	0	0	0	0	0	4529
2	49	598	0	0	0	0	0	0	0	0	0	647
3	222	0	142	125	0	0	0	0	0	584	10	1083
4	0	0	0	494	0	0	0	0	0	0	0	494
5	0	0	0	10	102	0	0	0	0	0	0	112
6	0	0	17	7	0	1285	0	0	1145	0	0	2454
7	0	0	2	11	0	0	225	0	0	0	0	238
8	0	0	25	53	0	27	0	268	161	0	0	534
9	0	0	12	128	0	0	0	0	950	0	37	1127
10	0	0	0	80	0	0	0	0	0	130	0	210
11	0	0	0	17	0	0	0	0	0	0	112	129
	4319	1069	198	935	102	1312	225	268	2256	714	159	11557

Figure F.2 Bolivar Peninsula Precision Tree and Confusion Matrix for TD-BHC, trained on Site 1 and tested on Site 2

Bolivar Peninsula: Training on Area 2 and Testing on Area 1

Bottom-Up Class and Meta-Class Classification Accuracies

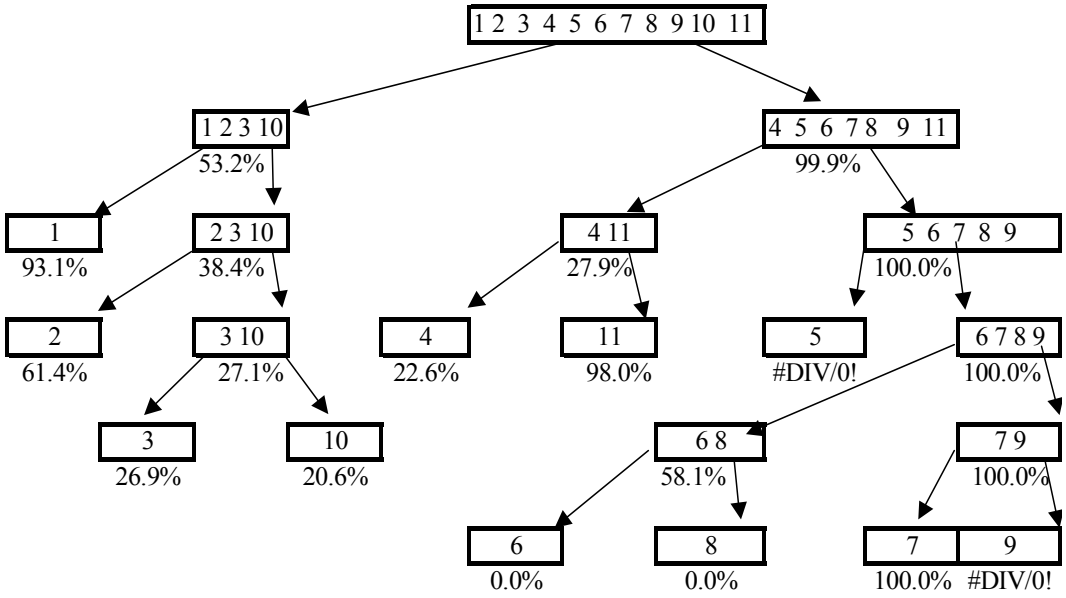


		PREDICTED											
		1	2	3	4	5	6	7	8	9	10	11	
ACTUAL	1	562	456	1	0	0	0	0	0	0	0	0	1019
	2	4	1123	0	0	0	0	0	0	0	0	0	1127
	3	0	0	910	0	0	0	0	0	0	0	0	910
	4	0	7	0	745	0	0	0	0	0	0	0	752
	5	0	0	0	0	148	0	0	0	0	0	0	148
	6	0	0	2509	402	0	0	0	0	0	162	0	3073
	7	0	0	5	185	0	0	24	0	0	8	0	222
	8	0	0	268	349	0	0	0	0	0	87	0	704
	9	0	336	332	156	0	0	0	0	0	271	0	1095
	10	0	0	22	24	0	0	0	0	0	68	0	114
	11	0	3	0	0	21	0	0	0	0	0	190	214
		566	1925	4047	1861	169	0	24	0	0	596	190	9378

Figure F.3 Bolivar Peninsula Precision Tree and Confusion Matrix for BU-BHC, trained on Site 2 and tested on Site 1

Bolivar Peninsula: Training on Area 2 and Testing on Area 1

Top-Down Class and Meta-Class Classification Accuracies



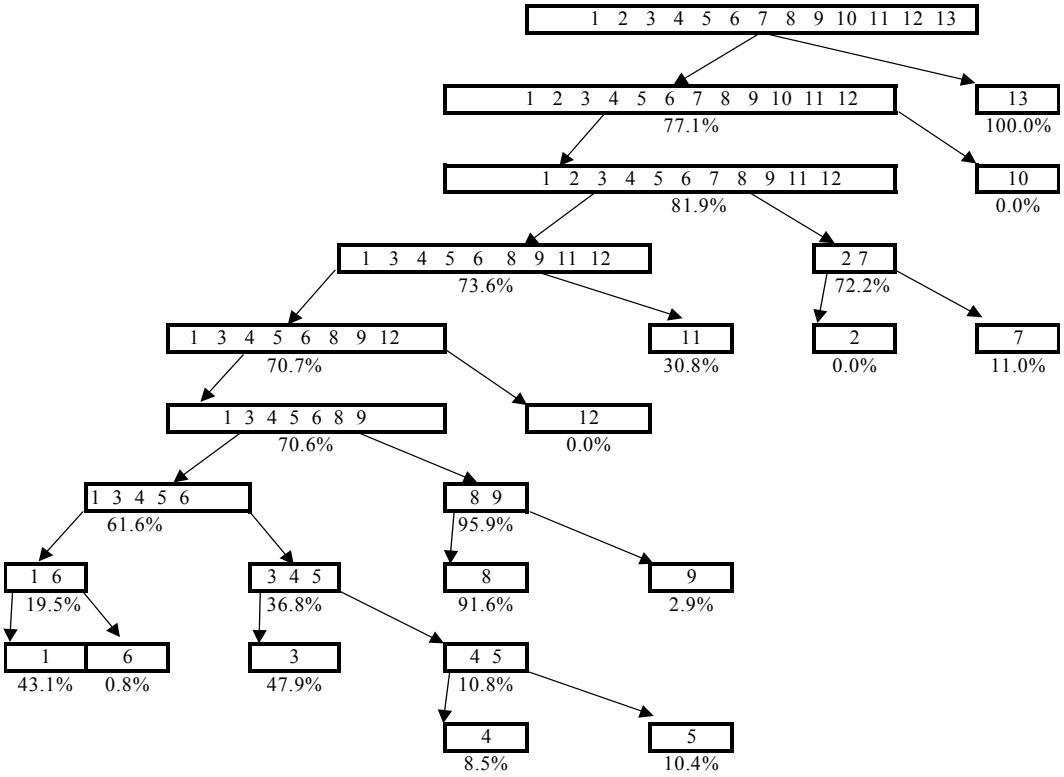
PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	
1	336	682	1	0	0	0	0	0	0	0	0	1019
2	3	1124	0	0	0	0	0	0	0	0	0	1127
3	0	0	910	0	0	0	0	0	0	0	0	910
4	0	0	22	725	0	0	0	0	0	5	0	752
5	0	0	0	148	0	0	0	0	0	0	0	148
6	0	0	2036	856	0	0	0	0	0	181	0	3073
7	0	0	2	76	0	27	87	1	0	29	0	222
8	0	0	359	229	0	50	0	0	0	66	0	704
9	0	19	24	1035	0	8	0	0	0	8	1	1095
10	0	1	33	5	0	0	0	0	0	75	0	114
11	22	6	0	138	0	0	0	0	0	0	48	214
	361	1832	3387	3212	0	85	87	1	0	364	49	9378

Figure F.4 Bolivar Peninsula Precision Tree and Confusion Matrix for TD-BHC, trained on Site 2 and tested on Site 1

KSC: Training on Area 1 and Testing on Area 2

Bottom-Up Class and Meta-Class Classification Accuracies



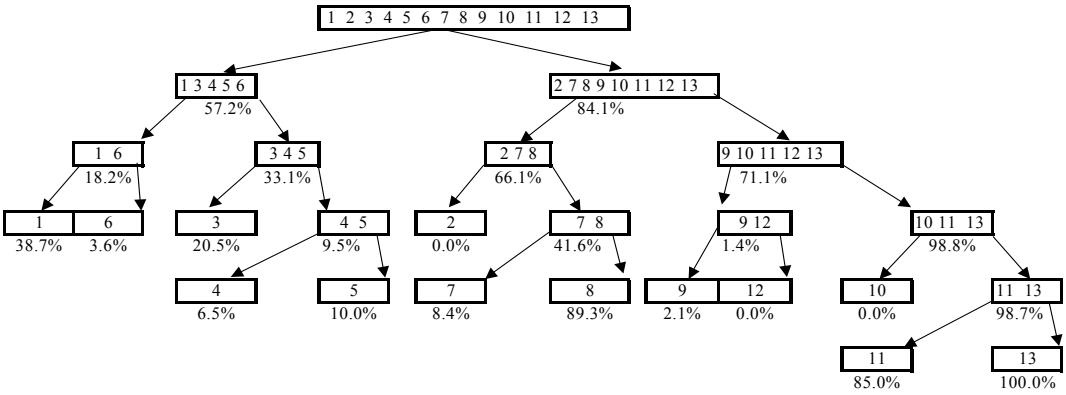
PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	81	0	0	11	294	26	1	9	0	0	0	0	0	0	0	0	422
2	0	0	4	16	3	5	142	0	0	3	0	7	0	0	0	0	180
3	0	0	45	332	50	0	3	0	0	0	0	1	0	0	0	0	431
4	0	0	5	72	12	30	4	0	0	9	0	0	0	0	0	0	132
5	82	0	0	10	71	1	0	0	0	0	2	0	0	0	0	0	166
6	0	0	0	108	113	3	7	0	0	43	0	0	0	0	0	0	274
7	5	0	4	130	30	34	24	0	0	21	0	0	0	0	0	0	248
8	1	7	1	4	2	0	3	174	97	0	151	13	0	0	0	0	453
9	0	0	35	65	6	0	3	7	3	0	113	9	0	0	0	0	241
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	5	0	0	0	0	0	0	3	6	142	0	0	0	0	0	156
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1392	0	0	0	1392
14	19	0	0	34	56	274	4	0	0	6	0	0	0	0	0	0	393
15	0	0	0	0	0	0	0	0	0	216	53	0	0	0	0	0	269
16	0	0	0	61	47	3	27	0	0	4	0	0	0	0	0	0	142
	188	12	94	843	684	376	218	190	103	308	461	30	1392	0	0	0	4899

Figure F.5 KSC Precision Tree and Confusion Matrix for BU-BHC, trained on Site 1 and tested on Site 2

KSC: Training on Area 1 and Testing on Area 2

Top-Down Class and Meta-Class Classification Accuracies



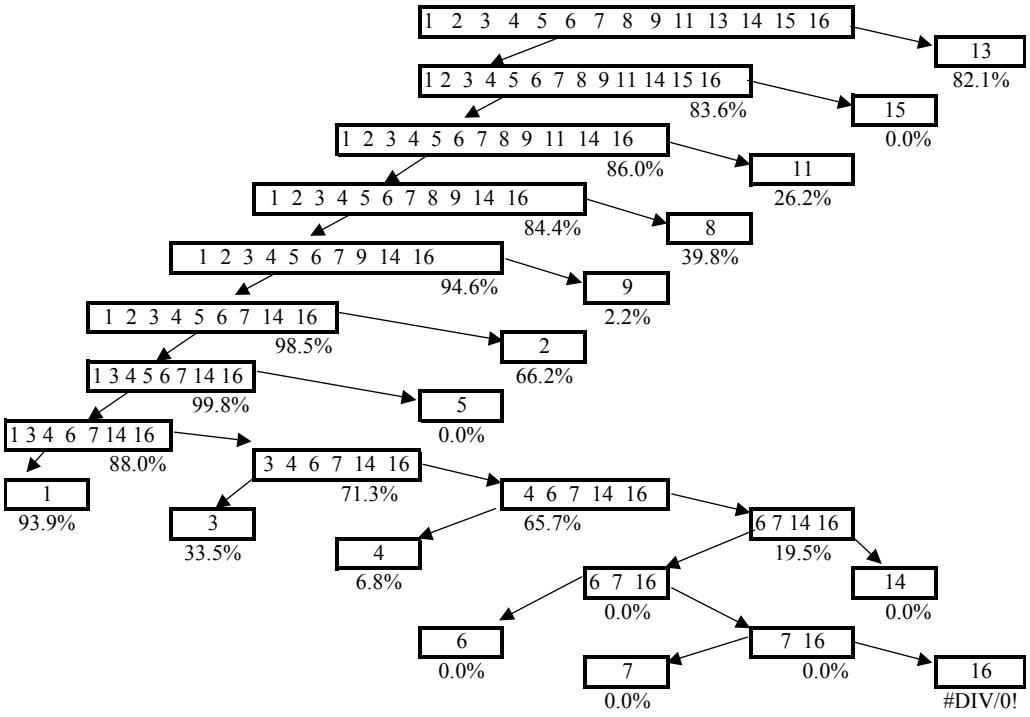
PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	67	0	0	7	292	22	22	4	4	0	0	4	0	0	0	0	422
2	0	0	4	22	6	2	107	1	30	0	0	8	0	0	0	0	180
3	0	0	47	308	66	0	3	0	6	0	0	1	0	0	0	0	431
4	0	0	9	50	10	29	28	0	4	0	0	2	0	0	0	0	132
5	82	0	0	10	73	0	0	0	0	0	0	1	0	0	0	0	166
6	0	1	0	107	121	14	2	0	28	0	0	1	0	0	0	0	274
7	4	0	9	125	42	34	22	0	12	0	0	0	0	0	0	0	248
8	0	4	2	1	1	0	6	151	263	0	18	7	0	0	0	0	453
9	1	0	146	61	10	1	4	8	9	0	1	0	0	0	0	0	241
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	5	11	0	0	0	0	5	25	2	108	0	0	0	0	0	156
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1392	0	0	0	1392
14	19	0	0	18	55	287	4	0	10	0	0	0	0	0	0	0	393
15	0	0	0	0	0	0	46	0	38	0	0	185	0	0	0	0	269
16	0	0	1	64	53	4	17	0	3	0	0	0	0	0	0	0	142
	173	10	229	773	729	393	261	169	432	2	127	209	1392	0	0	0	4899

Figure F.6 KSC Precision Tree and Confusion Matrix for TD-BHC, trained on Site 1 and tested on Site 2

KSC: Training on Area 2 and Testing on Area 1

Bottom-Up Class and Meta-Class Classification Accuracies

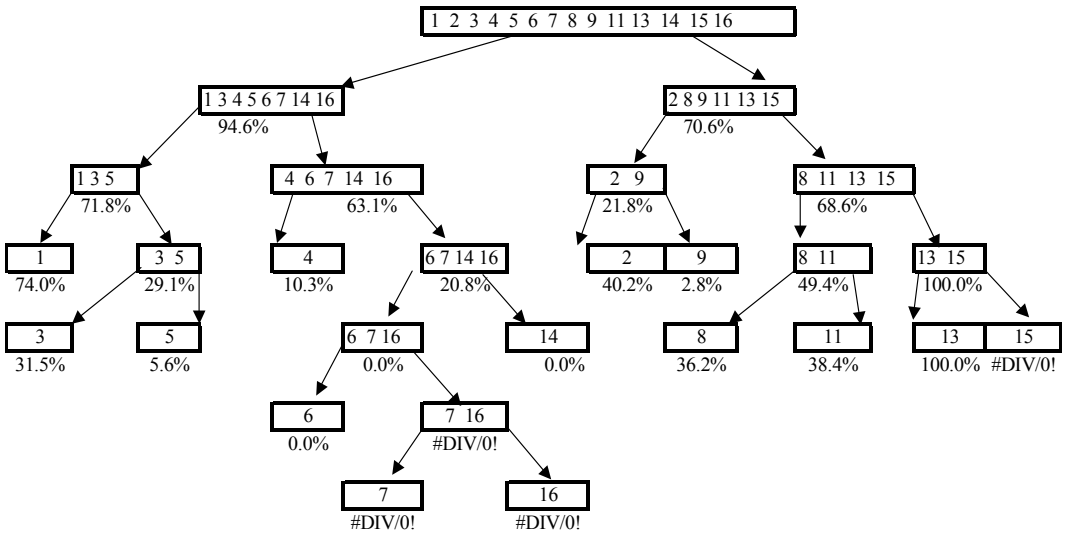


ACTUAL \ PREDICTED	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	294	12	25	39	286	0	1	5	19	0	13	0	0	67	0	0	761
2	0	208	0	1	0	0	0	2	25	0	7	0	0	0	0	0	243
3	0	0	218	0	0	0	0	1	32	0	5	0	0	0	0	0	256
4	0	0	221	16	0	0	0	6	8	0	0	0	0	1	0	0	252
5	7	0	142	1	0	1	0	2	8	0	0	0	0	0	0	0	161
6	12	1	40	141	4	0	0	0	5	0	9	0	0	17	0	0	229
7	0	67	2	36	0	0	0	0	0	0	0	0	0	0	0	0	105
8	0	10	2	0	0	0	0	303	14	0	91	0	0	0	0	0	420
9	0	0	0	0	0	0	0	88	4	0	428	0	0	0	0	0	520
10	0	0	0	0	0	0	0	3	18	0	189	0	186	0	0	0	396
11	0	0	0	0	0	0	0	83	23	0	268	0	15	0	30	0	419
12	0	16	0	0	0	0	0	269	30	0	14	0	1	0	117	0	447
13	0	0	0	0	0	0	0	0	0	0	0	0	927	0	0	0	927
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	313	314	650	234	290	1	1	762	186	0	1024	0	1129	85	147	0	5136

Figure F.7 KSC Precision Tree and Confusion Matrix for BU-BHC, trained on Site 2 and tested on Site 1

KSC: Training on Area 2 and Testing on Area 1

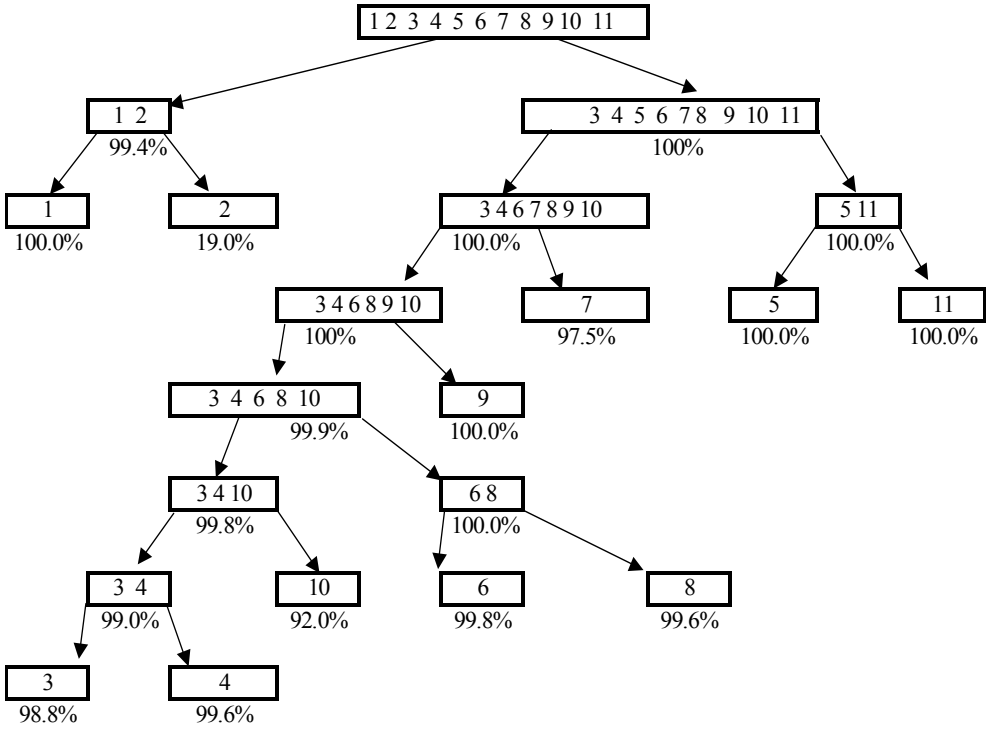
Top-Down Class and Meta-Class Classification Accuracies



		PREDICTED																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
ACTUAL	1	288	3	3	24	339	0	0	22	17	0	7	0	0	58	0	0	761
	2	0	222	0	0	0	0	0	18	0	0	3	0	0	0	0	0	243
	3	24	1	148	3	0	0	0	5	73	0	2	0	0	0	0	0	256
	4	4	0	192	22	9	1	0	5	18	0	0	0	0	1	0	0	252
	5	18	0	94	5	25	1	0	3	15	0	0	0	0	0	0	0	161
	6	6	1	9	121	60	0	0	0	10	0	6	0	0	16	0	0	229
	7	0	66	1	38	0	0	0	0	0	0	0	0	0	0	0	0	105
	8	2	1	3	0	0	0	0	401	8	0	5	0	0	0	0	0	420
	9	2	0	0	0	0	0	0	338	15	0	165	0	0	0	0	0	520
	10	12	252	16	0	5	0	0	0	60	0	51	0	0	0	0	0	396
	11	23	2	2	0	0	0	0	188	45	0	159	0	0	0	0	0	419
	12	10	4	2	0	10	0	0	129	276	0	16	0	0	0	0	0	447
	13	0	0	0	0	0	0	0	0	0	0	0	0	927	0	0	0	927
	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		389	552	470	213	448	2	0	1109	537	0	414	0	927	75	0	0	5136

Figure F.8 KSC Precision Tree and Confusion Matrix for TD-BHC, trained on Site 2 and tested on Site 1

Bolivar Peninsula: Training on Area 1 and Testing on Area 2
Parameters Updated Based Upon Pseudo-Labeled Data
Bottom-Up Class and Meta-Class Classification Accuracies

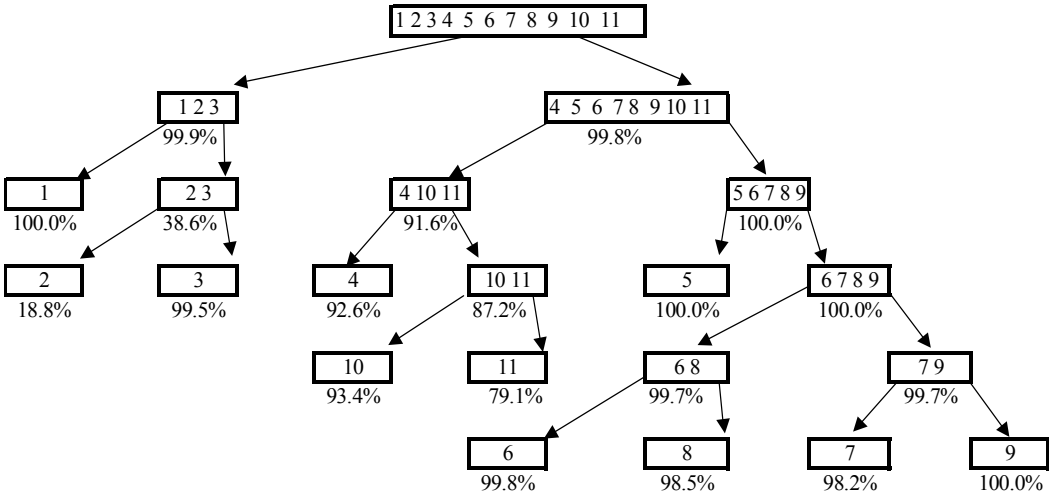


PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	
1	1806	2723	0	0	0	0	0	0	0	0	0	4529
2	0	647	0	0	0	0	0	0	0	0	0	647
3	0	32	1049	0	0	0	0	0	0	2	0	1083
4	0	0	0	482	0	0	0	0	0	12	0	494
5	0	0	0	0	112	0	0	0	0	0	0	112
6	0	0	0	0	0	2452	0	2	0	0	0	2454
7	0	0	0	0	0	1	234	0	0	3	0	238
8	0	0	0	0	0	5	6	523	0	0	0	534
9	0	0	0	0	0	0	0	0	1127	0	0	1127
10	0	0	13	2	0	0	0	0	0	195	0	210
11	0	0	0	0	0	0	0	0	0	0	129	129
	1806	3402	1062	484	112	2458	240	525	1127	212	129	11557

Figure F.9 Bolivar Peninsula Precision Tree and Confusion Matrix for BU-BHC, trained on Site 1 and tested on Site 2, Parameters Updated

Bolivar Peninsula: Training on Area 1 and Testing on Area 2
Parameters Updated Based Upon Pseudo-Labeled Data
Top-Down Class and Meta-Class Classification Accuracies

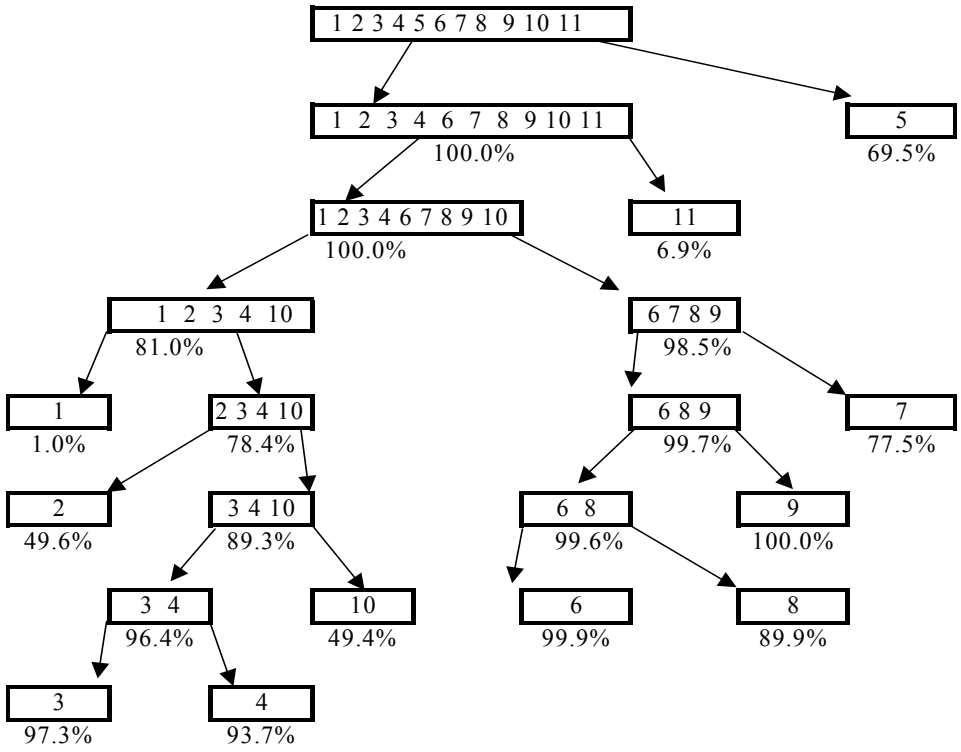


PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	
1	1806	2723	0	0	0	0	0	0	0	0	0	4529
2	0	647	0	0	0	0	0	0	0	0	0	647
3	0	66	1004	3	0	0	0	0	0	10	0	1083
4	0	0	1	490	0	0	0	0	0	3	0	494
5	0	0	0	3	98	0	0	0	0	0	11	112
6	0	0	0	0	0	2452	1	1	0	0	0	2454
7	0	0	0	7	0	1	222	7	0	1	0	238
8	0	0	0	9	0	5	3	517	0	0	0	534
9	0	0	0	9	0	0	0	0	1095	0	23	1127
10	0	0	4	8	0	0	0	0	0	198	0	210
11	0	0	0	0	0	0	0	0	0	0	129	129
	1806	3436	1009	529	98	2458	226	525	1095	212	163	11557

Figure F.10 Bolivar Peninsula Precision Tree and Confusion Matrix for TD-BHC, trained on Site 1 and tested on Site 2, Parameters Updated

Bolivar Peninsula: Training on Area 2 and Testing on Area 1
Parameters Updated Based Upon Pseudo-Labeled Data
Bottom-Up Class and Meta-Class Classification Accuracies

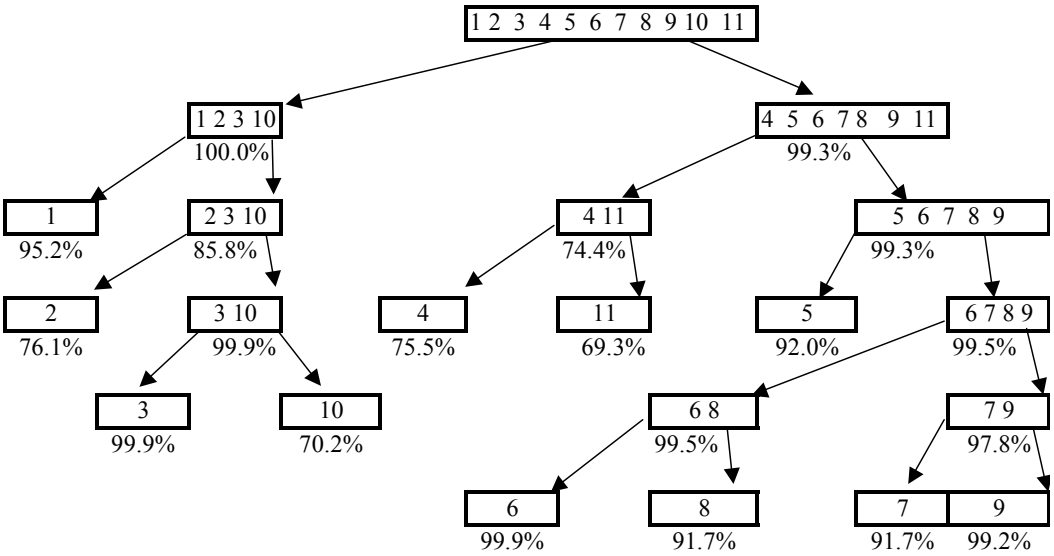


PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	
1	10	0	0	0	0	0	6	0	0	0	1003	1019
2	313	177	0	0	0	0	51	0	0	0	586	1127
3	127	0	571	0	0	0	0	0	0	10	202	910
4	351	0	0	178	0	0	0	13	0	0	210	752
5	0	0	0	0	148	0	0	0	0	0	0	148
6	27	0	4	0	0	2982	0	60	0	0	0	3073
7	0	0	0	0	0	0	220	1	0	0	1	222
8	9	0	10	4	0	2	7	657	0	15	0	704
9	114	180	2	4	0	0	0	0	729	63	3	1095
10	11	0	0	4	0	0	0	0	0	86	13	114
11	0	0	0	0	65	0	0	0	0	0	149	214
	962	357	587	190	213	2984	284	731	729	174	2167	9378

Figure F.11 Bolivar Peninsula Precision Tree and Confusion Matrix for BU-BHC, trained on Site 2 and tested on Site 1, Parameters Updated

Bolivar Peninsula: Training on Area 2 and Testing on Area 1
Parameters Updated Based Upon Pseudo-Labeled Data
Top-Down Class and Meta-Class Classification Accuracies

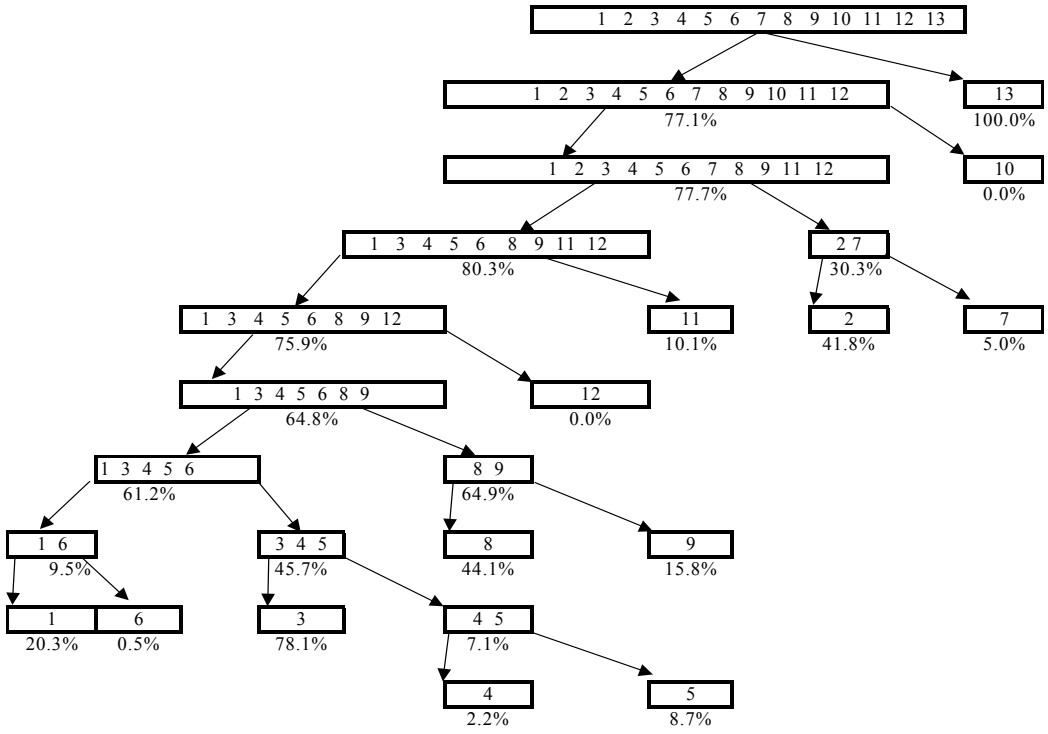


PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	
1	676	342	1	0	0	0	0	0	0	0	0	1019
2	34	1093	0	0	0	0	0	0	0	0	0	1127
3	0	0	857	16	0	0	0	0	0	37	0	910
4	0	0	0	727	0	0	10	13	0	0	2	752
5	0	0	0	0	69	0	0	0	0	0	79	148
6	0	0	0	190	0	2823	6	45	9	0	0	3073
7	0	0	0	0	0	0	221	1	0	0	0	222
8	0	0	0	1	0	2	4	697	0	0	0	704
9	0	0	0	4	0	0	0	0	1081	0	10	1095
10	0	1	0	22	0	0	0	4	0	87	0	114
11	0	0	0	3	6	0	0	0	0	0	205	214
	710	1436	858	963	75	2825	241	760	1090	124	296	9378

Figure F.12 Bolivar Peninsula Precision Tree and Confusion Matrix for TD-BHC, trained on Site 2 and tested on Site 1, Parameters Updated

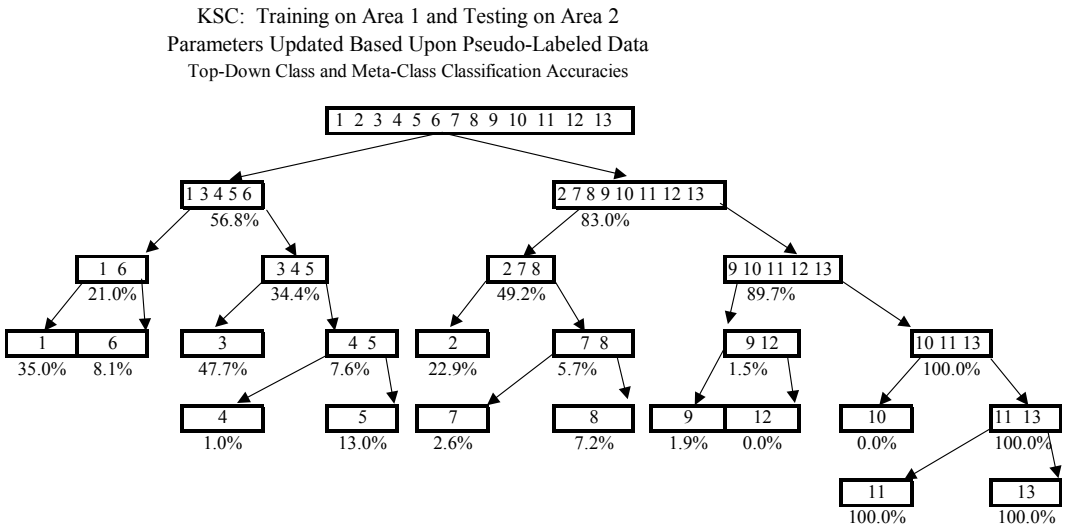
KSC: Training on Area 1 and Testing on Area 2
Parameters Updated Based Upon Pseudo-Labeled Data
Bottom-Up Class and Meta-Class Classification Accuracies



PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	28	0	13	10	117	5	0	15	0	99	5	130	0	0	0	0	422
2	0	110	0	0	0	0	0	0	0	64	0	6	0	0	0	0	180
3	0	7	196	30	13	1	1	12	3	18	4	146	0	0	0	0	431
4	0	19	14	3	0	8	30	0	0	58	0	0	0	0	0	0	132
5	49	2	2	3	20	2	0	1	0	7	12	68	0	0	0	0	166
6	0	20	9	29	43	1	21	1	0	139	0	11	0	0	0	0	274
7	9	40	9	34	6	17	14	0	0	109	0	10	0	0	0	0	248
8	0	6	0	0	0	0	0	26	29	19	206	167	0	0	0	0	453
9	0	3	2	1	0	0	0	2	6	70	117	40	0	0	0	0	241
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	3	0	0	0	0	0	0	0	114	39	0	0	0	0	0	156
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1392	0	0	0	1392
14	52	16	1	5	18	184	18	2	0	94	1	2	0	0	0	0	393
15	0	0	0	0	0	0	189	0	0	78	2	0	0	0	0	0	269
16	0	37	5	20	12	3	5	0	0	56	0	4	0	0	0	0	142
	138	263	251	135	229	221	278	59	38	925	386	584	1392	0	0	0	4899

Figure F.13 KSC Precision Tree and Confusion Matrix for BU-BHC, trained on Site 1 and tested on Site 2, Parameters Updated

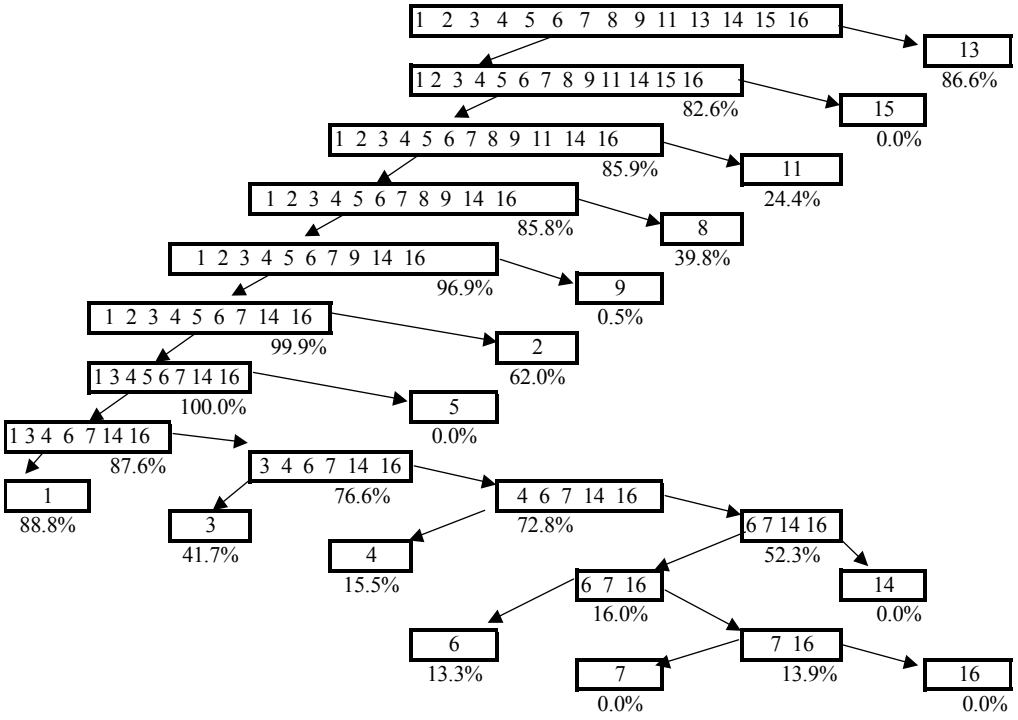


PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	108	20	4	13	259	13	0	5	0	0	0	0	0	0	0	0	422
2	1	131	13	19	0	11	1	1	1	0	0	2	0	0	0	0	180
3	0	4	250	144	30	1	0	0	0	0	0	2	0	0	0	0	431
4	0	19	22	5	0	61	23	2	0	0	0	0	0	0	0	0	132
5	85	0	5	9	63	4	0	0	0	0	0	0	0	0	0	0	166
6	1	31	10	96	80	36	20	0	0	0	0	0	0	0	0	0	274
7	20	39	19	119	11	37	3	0	0	0	0	0	0	0	0	0	248
8	2	263	4	1	0	0	0	17	113	0	0	53	0	0	0	0	453
9	8	14	170	39	0	6	0	0	3	0	0	1	0	0	0	0	241
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	16	20	0	0	0	0	0	45	26	4	45	0	0	0	0	156
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	1	0	0	0	0	0	0	0	10	0	0	1381	0	0	0	1392
14	84	7	0	11	16	260	15	0	0	0	0	0	0	0	0	0	393
15	0	2	0	0	0	0	53	211	0	0	0	3	0	0	0	0	269
16	0	26	7	70	26	13	0	0	0	0	0	0	0	0	0	0	142
	309	573	524	526	485	442	115	236	162	36	4	106	1381	0	0	0	4899

Figure F.14 KSC Precision Tree and Confusion Matrix for TD-BHC, trained on Site 1 and tested on Site 2, Parameters Updated

KSC: Training on Area 2 and Testing on Area 1
Parameters Updated Based Upon Pseudo-Labeled Data
Bottom-Up Class and Meta-Class Classification Accuracies

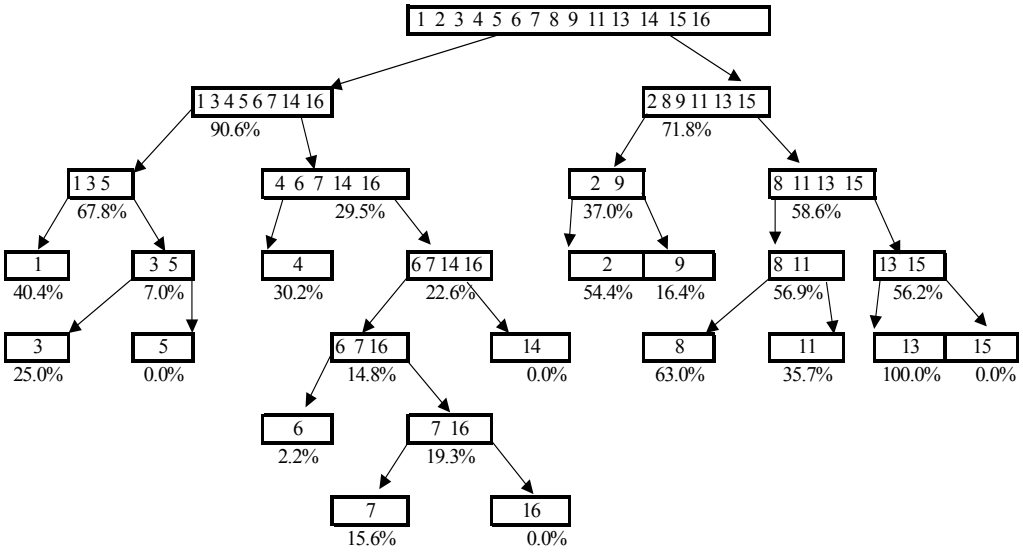


PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	174	8	5	3	426	7	13	11	64	0	15	0	0	34	0	1	761
2	0	150	0	0	0	0	0	2	75	0	6	0	0	0	10	0	243
3	0	0	165	0	0	1	0	3	82	0	5	0	0	0	0	0	256
4	0	1	142	15	0	11	7	7	61	0	0	0	0	1	0	7	252
5	9	0	77	5	0	20	0	4	45	0	0	0	0	0	0	1	161
6	13	6	7	50	13	6	2	1	22	0	9	0	0	100	0	0	229
7	0	76	0	24	0	0	0	0	0	0	0	0	0	0	0	5	105
8	0	1	0	0	0	0	0	311	17	0	91	0	0	0	0	0	420
9	0	0	0	0	0	0	0	90	2	0	428	0	0	0	0	0	520
10	0	0	0	0	0	0	0	4	15	0	241	0	135	0	1	0	396
11	0	0	0	0	0	0	0	104	2	0	260	0	8	0	45	0	419
12	0	0	0	0	0	0	0	244	27	0	11	0	0	0	165	0	447
13	0	0	0	0	0	0	0	0	0	0	0	0	927	0	0	0	927
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	196	242	396	97	439	45	22	781	412	0	1066	0	1070	135	221	14	5136

Figure F.15 KSC Precision Tree and Confusion Matrix for BU-BHC, trained on Site 2 and tested on Site 1, Parameters Updated

KSC: Training on Area 2 and Testing on Area 1
Parameters Updated Based Upon Pseudo-Labeled Data
Top-Down Class and Meta-Class Classification Accuracies

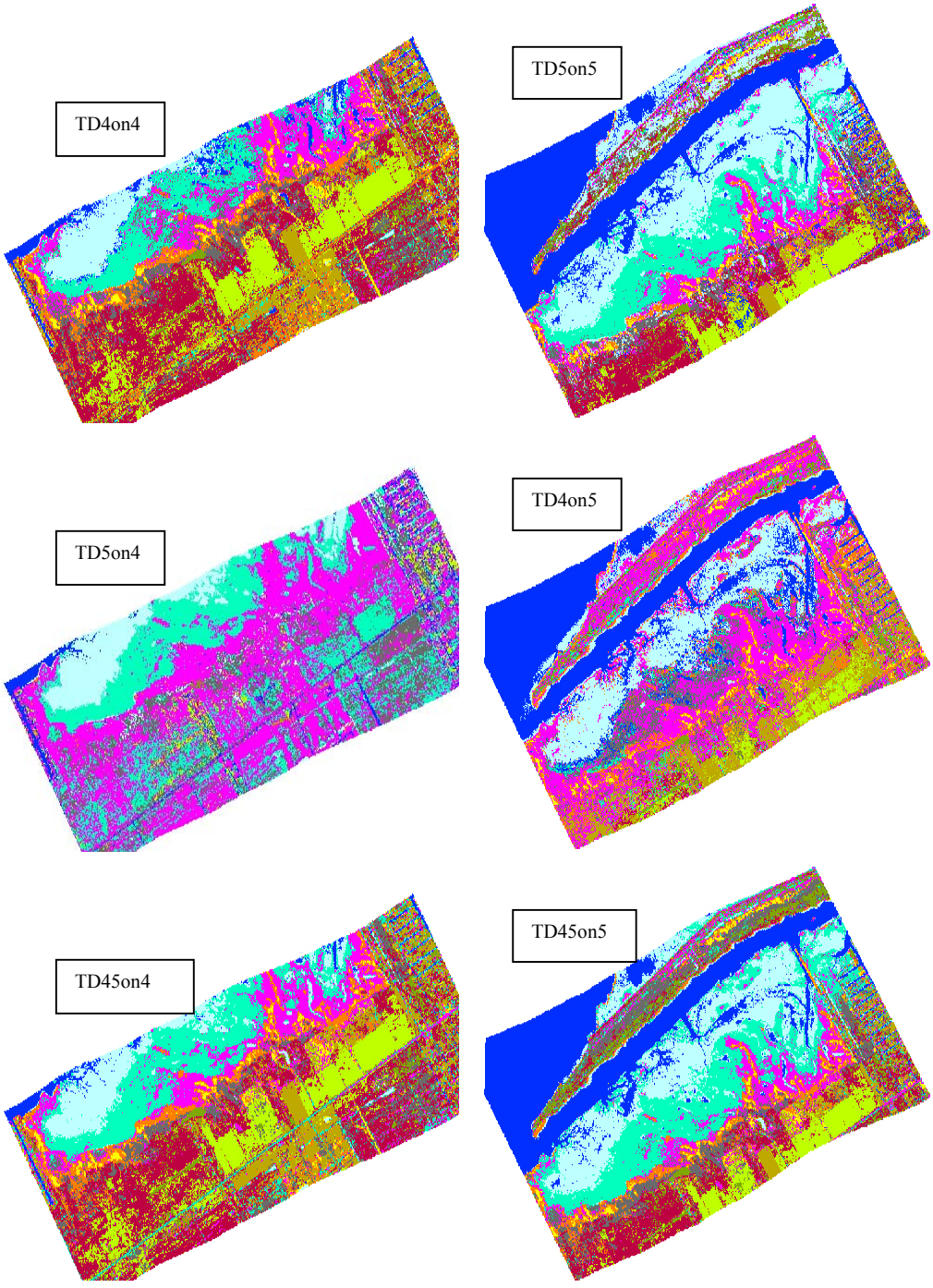


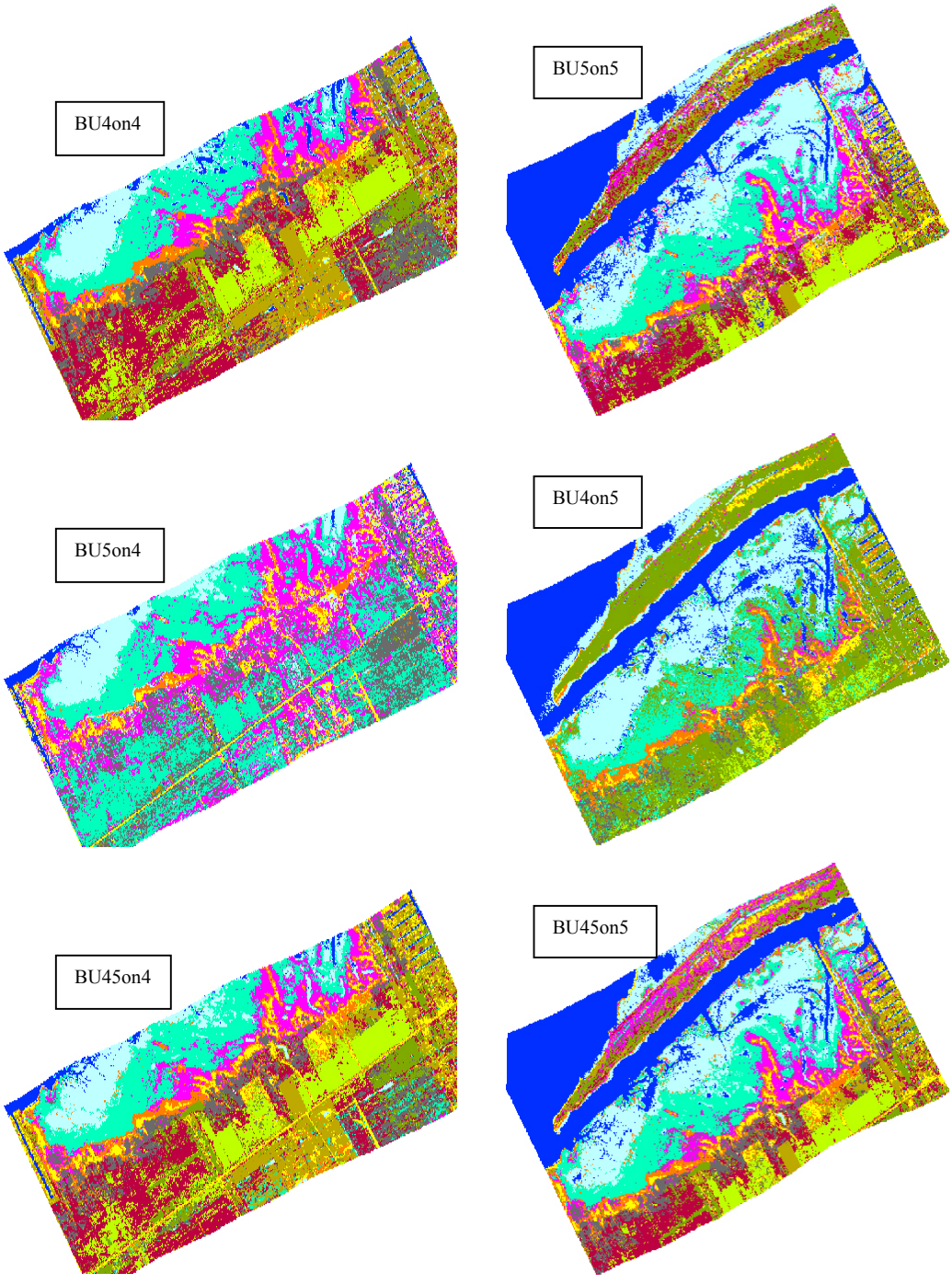
PREDICTED

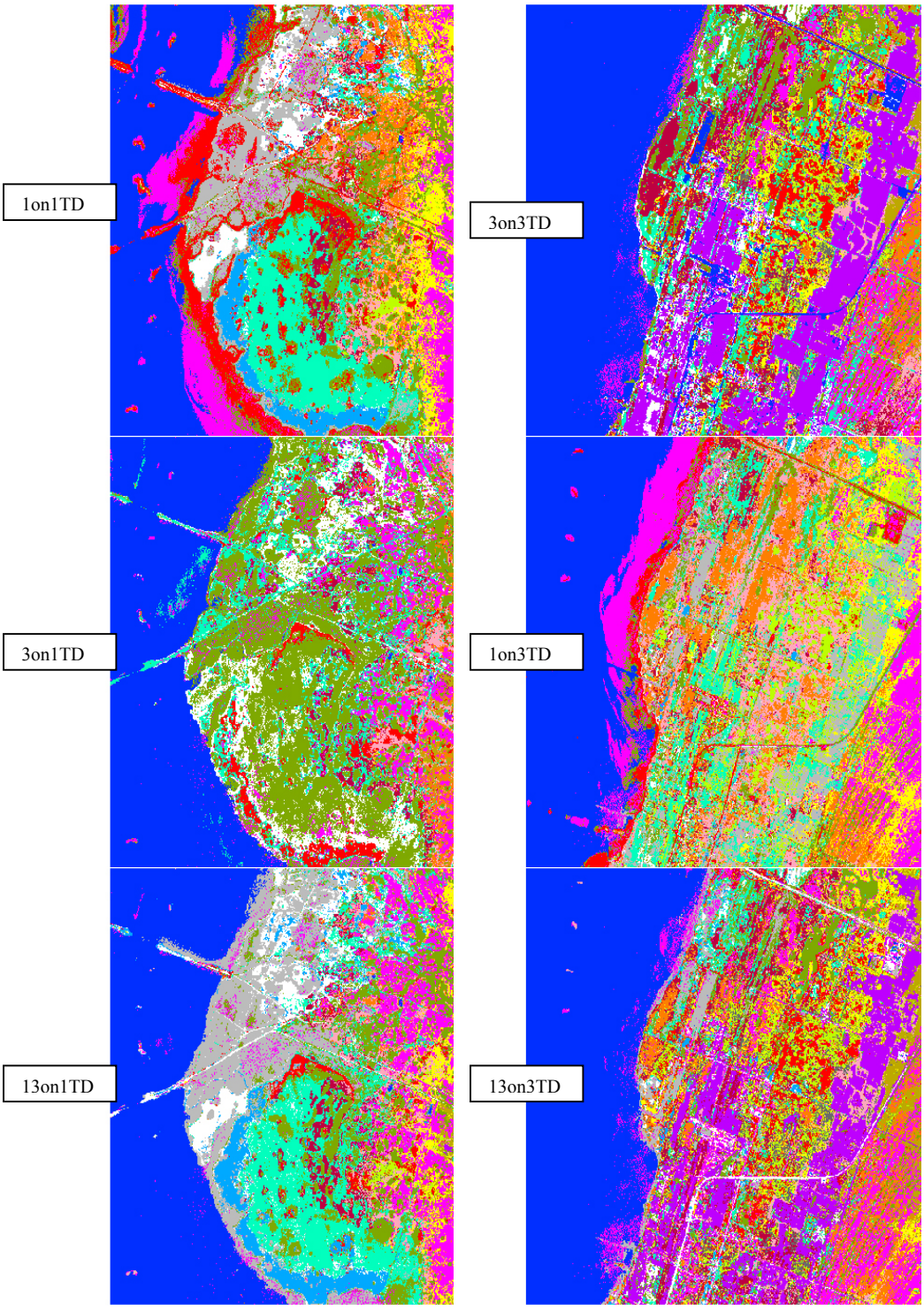
ACTUAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	84	10	9	24	236	50	30	1	2	0	3	0	0	311	0	1	761
2	0	211	0	9	0	6	0	0	0	0	0	0	0	0	14	3	243
3	0	3	19	181	0	19	0	3	14	0	1	0	0	0	0	16	256
4	0	1	10	159	0	20	24	1	5	0	1	0	0	2	0	29	252
5	3	0	3	40	0	80	4	1	2	0	0	0	0	7	0	21	161
6	1	0	2	77	2	4	6	0	1	0	1	0	0	129	0	6	229
7	0	33	0	34	0	3	12	0	0	0	0	0	0	1	0	22	105
8	16	15	7	2	0	0	1	344	17	0	17	0	0	0	0	1	420
9	103	0	26	0	0	0	0	115	54	0	222	0	0	0	0	0	520
10	0	89	0	0	0	0	0	0	0	0	21	0	0	0	286	0	396
11	0	0	0	0	0	0	0	42	167	0	155	0	0	0	55	0	419
12	1	26	0	0	0	0	0	39	30	0	13	0	0	0	338	0	447
13	0	0	0	0	0	0	0	0	37	0	0	0	890	0	0	0	927
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	208	388	76	526	238	182	77	546	329	0	434	0	890	450	693	99	5136

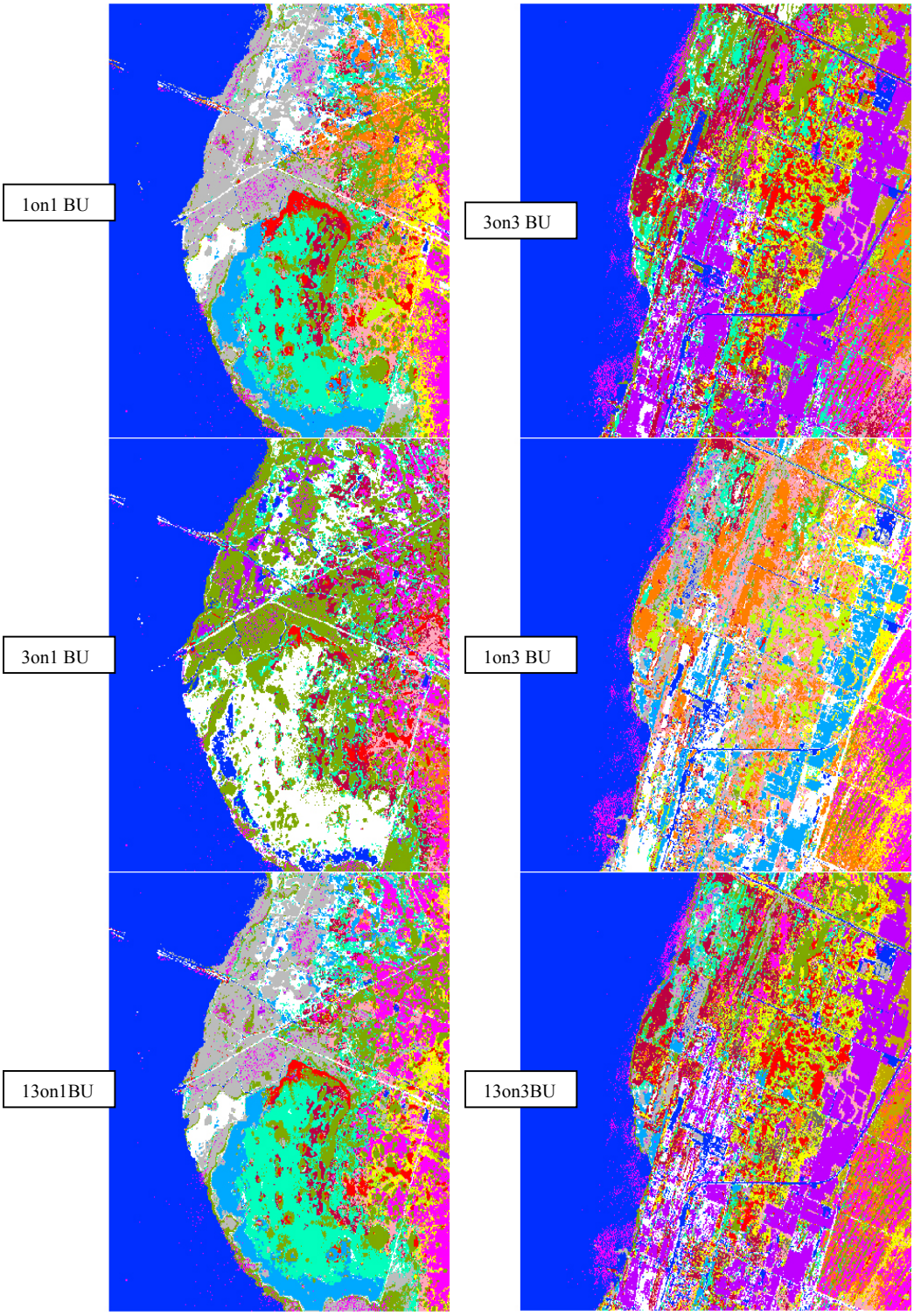
Figure F.16 KSC Precision Tree and Confusion Matrix for TD-BHC, trained on Site 2 and tested on Site 1, Parameters Updated

Appendix G

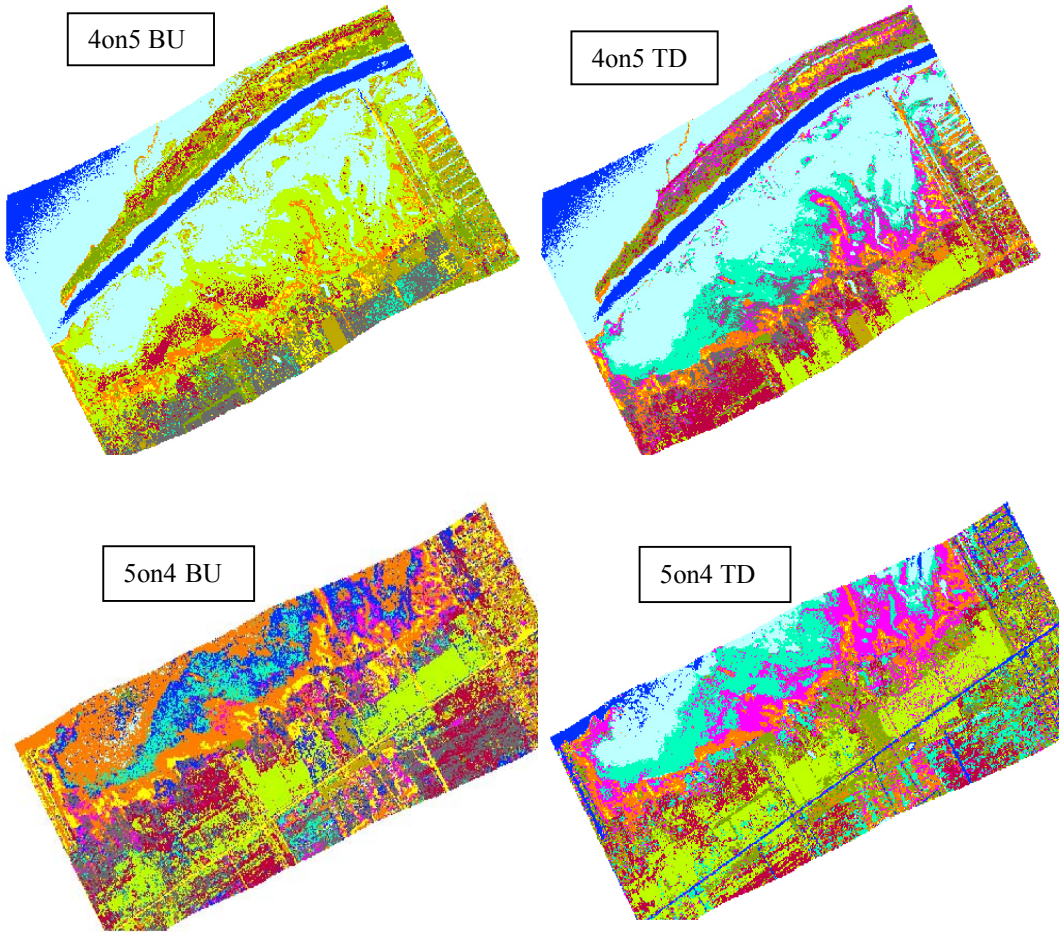


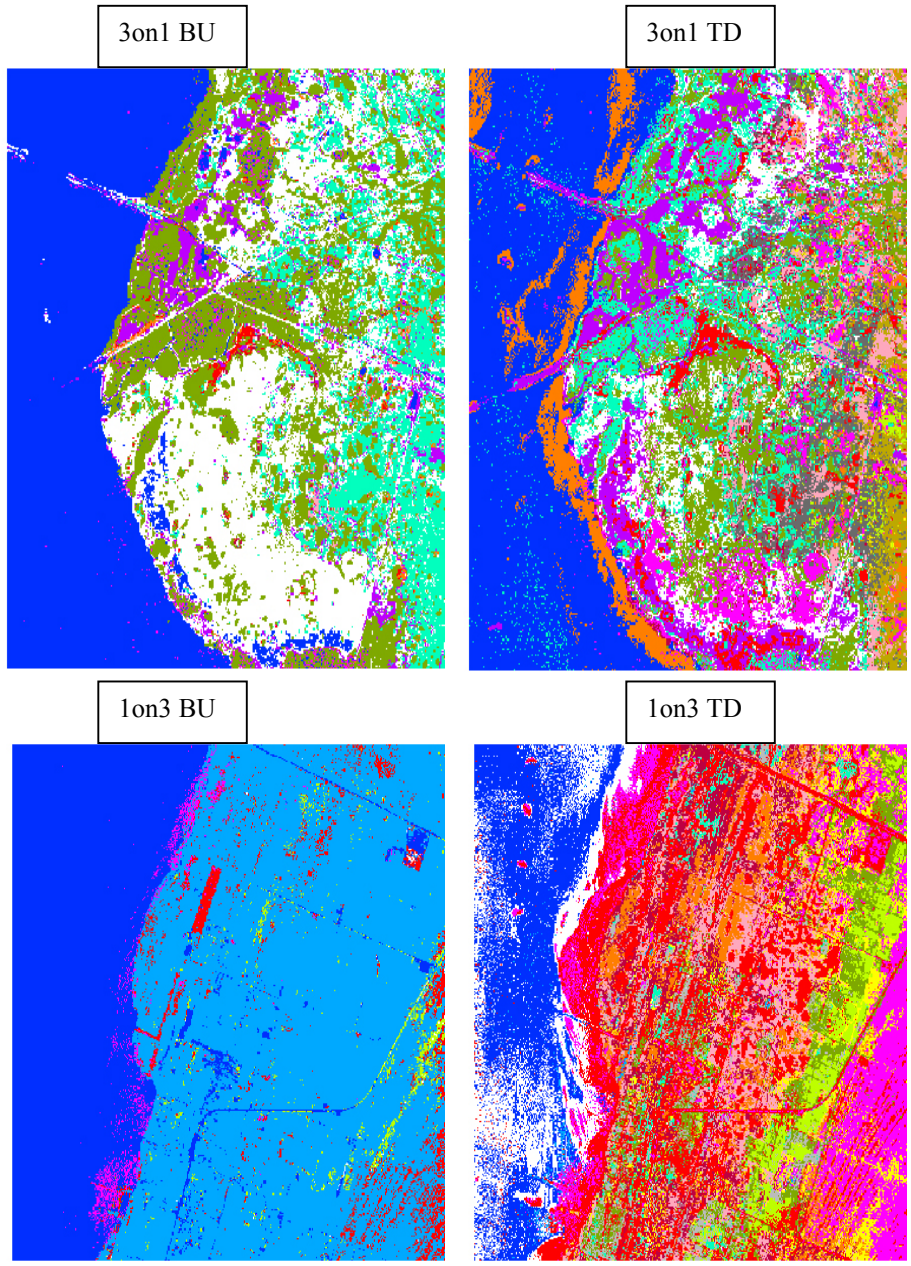

BU4on4
BU5on5
BU5on4
BU4on5
BU45on4
BU45on5


1on1TD
3on3TD
3on1TD
1on3TD
13on1TD
13on3TD


1on1 BU
3on3 BU
3on1 BU
1on3 BU
13on1BU
13on3BU

Appendix H

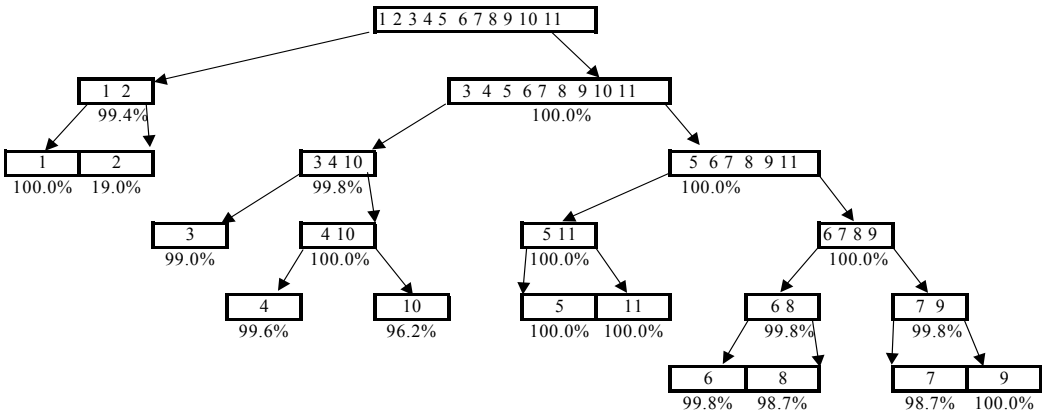



3on1 BU
3on1 TD
1on3 BU
1on3 TD

Appendix I

Bolivar Peninsula Ensemble Results: Training on Area 1 and Testing on Area 2

Master Tree Class and Meta-Class Classification Accuracies



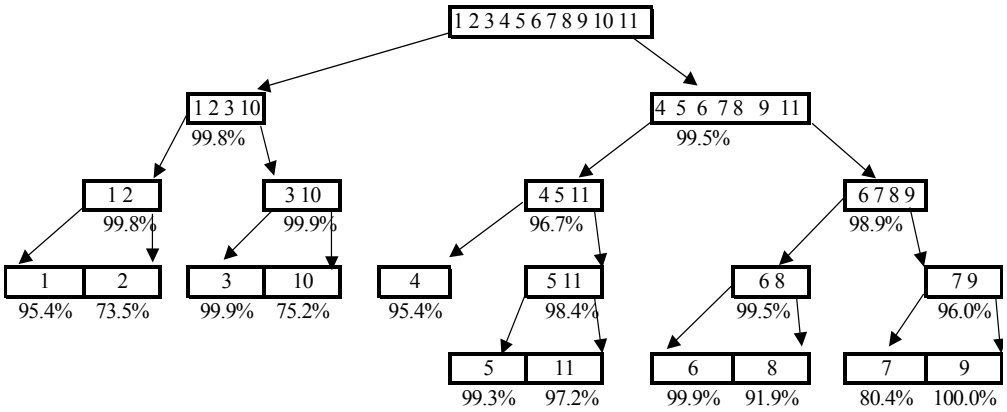
		PREDICTED											
		1	2	3	4	5	6	7	8	9	10	11	
ACTUAL	1	1806	2723	0	0	0	0	0	0	0	0	0	4529
	2	0	644	3	0	0	0	0	0	0	0	0	647
	3	0	31	1052	0	0	0	0	0	0	0	0	1083
	4	0	0	0	486	0	0	0	0	0	8	0	494
	5	0	0	0	0	112	0	0	0	0	0	0	112
	6	0	0	0	0	0	2452	0	2	0	0	0	2454
	7	0	0	0	0	0	1	232	5	0	0	0	238
	8	0	0	0	0	0	5	3	526	0	0	0	534
	9	0	0	0	0	0	0	0	0	1127	0	0	1127
	10	0	0	8	2	0	0	0	0	0	200	0	210
	11	0	0	0	0	0	0	0	0	0	0	129	129
		1806	3398	1063	488	112	2458	235	533	1127	208	129	11557

	Distance Measure			Purities	
Master Tree Partition	New	Old	Ratio	Left	Right
1 2 3 4 5 6 7 8 9 10 11 into 3 4 5 6 7 8 9 10 11 and 1 2	35.9211	41.2021	0.8718	0.9784	0.9976
3 4 5 6 7 8 9 10 11 into 5 6 7 8 9 11 and 3 4 10	13.6639	15.3359	0.8910	0.9889	0.9958
1 2 into 1 and 2	15.8391	12.6821	1.2489	0.9998	0.9668
5 6 7 8 9 11 into 5 11 and 6 7 8 9	44.4579	26.6221	1.6700	0.9166	0.9923
3 4 10 into 3 and 4 10	13.4598	15.9916	0.8417	0.9724	0.9884
5 11 into 5 and 11	21.3458	12.7528	1.6738	0.8259	0.9523
6 7 8 9 into 6 8 and 7 9	16.4928	16.9576	0.9726	0.9970	0.9770
4 10 into 4 and 10	5.2256	8.6687	0.6028	0.9664	0.9485
7 9 into 7 and 9	68.5864	96.2060	0.7129	0.9348	0.9859
6 8 into 6 and 8	11.5039	10.6350	1.0817	0.9948	0.9836

Figure I.1 Bolivar Peninsula Precision Tree and Confusion Matrix for Master Tree, trained on Site 1 and tested on Site 2

Bolivar Peninsula Ensemble Results: Training on Area 2 and Testing on Area 1

Master Tree Class and Meta-Class Classification Accuracies



PREDICTED

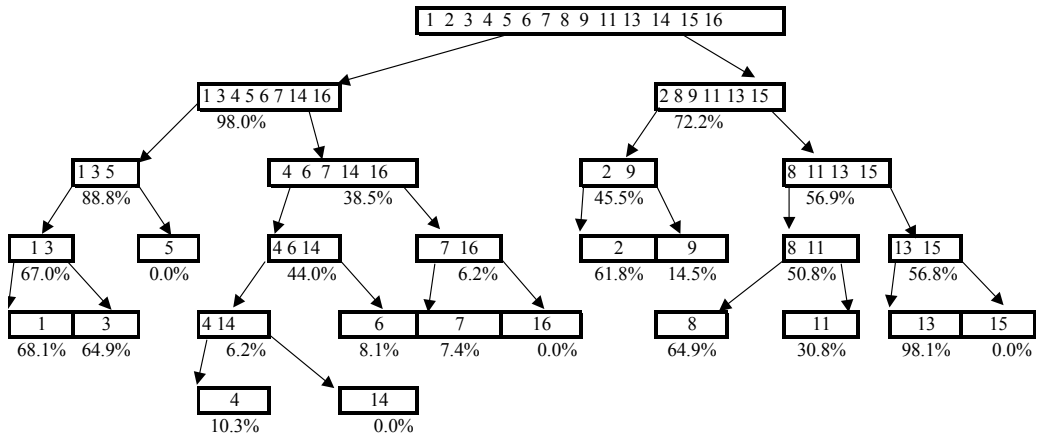
ACTUAL	1	2	3	4	5	6	7	8	9	10	11	
1	621	397	1	0	0	0	0	0	0	0	0	1019
2	25	1102	0	0	0	0	0	0	0	0	0	1127
3	0	0	869	6	0	0	4	0	0	31	0	910
4	5	0	0	706	0	0	28	12	0	0	1	752
5	0	0	0	0	148	0	0	0	0	0	0	148
6	0	0	0	13	0	3007	10	43	0	0	0	3073
7	0	0	0	0	0	0	221	1	0	0	0	222
8	0	0	0	0	0	2	6	696	0	0	0	704
9	0	0	0	3	0	0	0	0	1087	0	5	1095
10	0	0	0	9	0	0	6	5	0	94	0	114
11	0	0	0	3	1	0	0	0	0	0	210	214
	651	1499	870	740	149	3009	275	757	1087	125	216	9378

	Distance Measure			Purities	
Master Tree Partition	New	Old	Ratio	Left	Right
1 2 3 4 5 6 7 8 9 10 11 into 1 2 3 10 and 4 5 6 7 8 9 11	16.5312	20.5504	0.8044	0.5889	0.9442
4 5 6 7 8 9 11 into 4 5 11 and 6 7 8 9	12.7054	25.2349	0.5035	0.7344	0.8879
1 2 3 10 into 1 2 and 3 10	43.7629	28.9055	1.5140	0.4990	0.7020
4 5 11 into 4 and 5 11	29.2969	48.7344	0.6012	0.5995	0.8132
6 7 8 9 into 6 8 and 7 9	12.7092	16.1384	0.7875	0.8920	0.8645
3 10 into 3 and 10	4.6700	3.7950	1.2306	0.6765	0.7822
1 2 into 1 and 2	13.6638	33.2083	0.4115	0.4886	0.5008
5 11 into 5 and 11	11.8273	19.5537	0.6049	0.6951	0.6779
7 9 into 7 and 9	33.5644	58.2448	0.5763	0.9398	0.8489
6 8 into 6 and 8	10.7646	11.2133	0.9600	0.8817	0.8937

Figure I.2 Bolivar Peninsula Precision Tree and Confusion Matrix for Master Tree, trained on Site 2 and tested on Site 1

KSC Ensemble Results: Training on Area 2 and Testing on Area 1

Master Tree Class and Meta-Class Classification Accuracies



PREDICTED

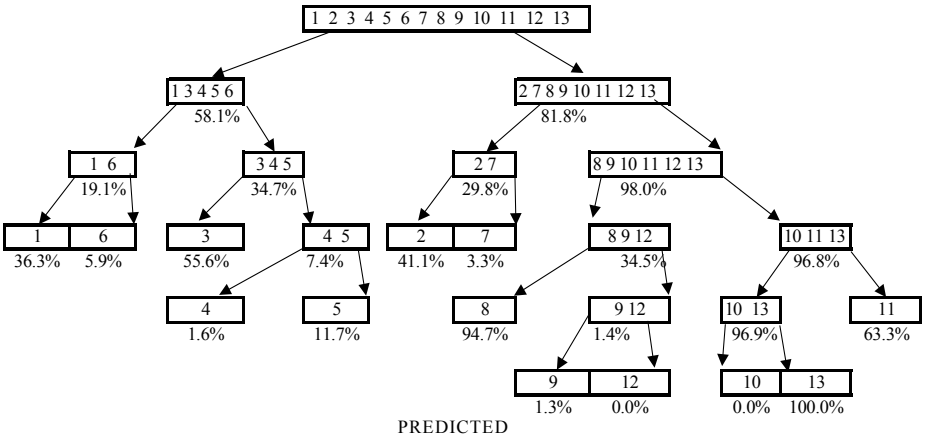
ACTUAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	64	10	1	84	350	5	98	2	4	0	3	0	0	139	0	1	761
2	0	235	0	0	0	0	2	0	0	0	0	0	0	0	5	1	243
3	0	3	63	3	1	30	3	3	21	0	3	0	0	0	0	126	256
4	0	0	22	29	1	50	48	3	7	0	0	0	0	0	0	92	252
5	4	0	2	11	0	50	17	2	6	0	0	0	0	0	0	69	161
6	4	0	1	131	3	12	19	0	0	0	1	0	0	50	0	8	229
7	0	49	0	23	0	1	15	0	0	0	0	0	0	0	0	17	105
8	2	4	2	0	0	0	1	396	2	0	13	0	0	0	0	0	420
9	19	0	6	0	0	0	0	102	29	0	364	0	0	0	0	0	520
10	0	68	0	0	0	0	0	1	1	0	51	0	18	0	257	0	396
11	0	0	0	0	0	0	0	27	118	0	195	0	0	0	79	0	419
12	1	11	0	0	0	0	0	74	12	0	3	0	0	0	346	0	447
13	0	0	0	0	0	0	0	0	0	0	0	0	927	0	0	0	927
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	94	380	97	281	355	148	203	610	200	0	633	0	945	189	687	314	5136

Master Tree Partition	Distance Measure New	Old	Ratio	Purities Left	Right
1-9 11 13-16 into 1 3 4 5 6 7 14 16 and 2 8 9 11 13 15	10.6805	10.6602	1.0019	0.9388	0.9443
2 8 9 11 13 15 into 2 9 and 8 11 13 15	2.9704	9.5979	0.3095	0.7294	0.8695
1 3 4 5 6 7 14 16 into 1 3 5 and 4 6 7 14 16	5.6812	5.0183	1.1321	0.6929	0.7516
8 11 13 15 into 8 11 and 13 15	10.0915	38.3694	0.2630	0.7798	0.7936
2 9 into 2 and 9	5.8057	10.6792	0.5437	0.7702	0.5040
4 6 7 14 16 into 4 6 14 and 7 16	2.6827	3.1096	0.8627	0.6988	0.6247
1 3 5 into 1 3 and 5	3.9032	3.1452	1.2410	0.5642	0.6460
13 15 into 13 and 15	12.8545	320.9514	0.0401	0.7304	0.7685
8 11 into 8 and 11	2.1803	4.9953	0.4365	0.7480	0.6513
7 16 into 7 and 16	1.9007	1.5275	1.2444	0.5227	0.5590
4 6 14 into 4 14 and 6	2.3620	2.6118	0.9044	0.6605	0.5066
1 3 into 1 and 3	2.4802	3.2954	0.7526	0.5731	0.5409
4 14 into 4 and 14	2.2243	2.7329	0.8139	0.6337	0.6120

Figure I.3 KSC Precision Tree and Confusion Matrix for Master Tree, trained on Site 2 and tested on Site 1

KSC Ensemble Results: Training on Area 1 and Testing on Area 3

Master Tree Class and Meta-Class Classification Accuracies



PREDICTED

ACTUAL	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	102	13	3	10	279	9	0	4	0	1	0	1	0	0	0	0	422
2	0	161	0	12	0	5	0	0	0	2	0	0	0	0	0	0	180
3	0	3	291	109	24	0	0	0	0	0	0	4	0	0	0	0	431
4	0	37	15	7	0	52	16	0	0	5	0	0	0	0	0	0	132
5	101	0	0	5	58	2	0	0	0	0	0	0	0	0	0	0	166
6	0	31	6	76	94	26	28	0	0	13	0	0	0	0	0	0	274
7	12	43	14	127	7	31	11	0	0	2	0	1	0	0	0	0	248
8	1	29	4	1	0	0	0	215	117	0	15	71	0	0	0	0	453
9	7	8	169	40	0	4	0	5	2	1	3	2	0	0	0	0	241
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	10	17	0	0	0	0	1	35	16	31	46	0	0	0	0	156
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	646	0	0	746	0	0	0	1392
14	58	16	0	6	8	297	6	0	0	2	0	0	0	0	0	0	393
15	0	0	0	0	0	0	267	2	0	0	0	0	0	0	0	0	269
16	0	41	4	58	24	11	1	0	0	3	0	0	0	0	0	0	142
	281	392	523	451	494	437	329	227	154	691	49	125	746	0	0	0	4899

	Distance Measure			Purities	
Master Tree Partition	New	Old	Ratio	Left	Right
1 - 13 into 1 3 4 5 6 and 2 7 8 9 10 11 12 13	8.4476	12.9524	0.6522	0.8025	0.9011
2 7 8 9 10 11 12 13 into 2 7 and 8 9 10 11 12 13	11.8797	15.5532	0.7638	0.7627	0.8274
1 3 4 5 6 into 1 6 and 3 4 5	5.3817	10.0524	0.5354	0.8002	0.7639
8 9 10 11 12 13 into 8 9 12 and 10 11 13	16.6825	10.3318	1.6147	0.5492	0.7839
2 7 into 2 and 7	3.3441	4.3648	0.7662	0.7114	0.7811
1 6 into 1 and 6	2.4795	3.3011	0.7511	0.6989	0.7447
3 4 5 into 3 and 4 5	3.0361	3.9611	0.7665	0.6629	0.7207
10 11 13 into 10 13 and 11	100.7571	6.4289	15.6724	0.7088	0.5397
8 9 12 into 8 and 9 12	4.1069	5.0674	0.8105	0.4793	0.5250
4 5 into 4 and 5	2.1861	2.6197	0.8345	0.6142	0.6931
10 13 into 10 and 13	2.1707	70.8571	0.0306	0.5014	0.6552
9 12 into 9 and 12	1.0726	9.4190	0.1139	0.6452	0.4917

Figure I.4 KSC Precision Tree and Confusion Matrix for Master Tree, trained on Site 1 and tested on Site 2

Bibliography

[1] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*. 2nd Ed., New York: John Wiley & Sons, 2001.

[2] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 1984.

[3] D. Landgrebe, "Information extraction principles and methods for multispectral and hyperspectral image data," *Information Processing for Remote Sensing*, ed. Chen, C.H., World Scientific Pub. Co, NJ, 1999.

[4] S. Tadjudin and D.A. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Trans on Geosci and RS*, vol. 38, no. 1, pp. 439-45, Jan. 2000.

[5] A.K. Jain, P.W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans on PAMI*, vol. 22, no. 1, pp. 4-37, 2000.

[6] Jerome H. Friedman, "On bias, variance, loss, and the curse of dimensionality," Technical report, Department of Statistics, Stanford University, 1996.

[7] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. New Jersey: Princeton University Press, 1961.

[8] Shailesh Kumar, Modular Learning Through Output Space Decomposition, Ph.D. Thesis, University of Texas at Austin, 2000.

[9] M.M. Lewis, V. Jooste, and A.A. Gasparis, "Discrimination of arid vegetation with airborne multispectral scanner hyperspectral imagery," *Proceedings Tenth Australasian Remote Sensing and Photogrammetry Conference*, Adelaide, Australia, 2000.

[10] S. Kumar, J. Ghosh, and M.M. Crawford, "A versatile framework for labeling imagery with large number of classes," *Proceedings of the International Joint Conference on Neural Networks*, Washington, D.C., 1999.

[11] S. Kumar, J. Ghosh, and M.M. Crawford, "Multiresolution feature extraction for pairwise classification of hyperspectral data," *Proc. Of SPIE: Applications of Artificial Neural Networks in Image Processing V, IS&T/SPIE's Electronic Imaging*, pp. 60-71, January 2000.

[12] S. Tadjudin and D.A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans on Geosci and RS*, vol. 37, no. 4, pp. 2113-8, 1999.

[13] J. H. Friedman, "An overview of predictive learning and function approximation," in V. Cherkassky, J.H. Friedman, and H. Wechsler, editors, *From Statistics to Neural Networks, Proc. NATO/ASI Workshop*, pp. 1-61, Springer Verlag, 1994.

[14] P.A. Devijver and J. Kittler (editors), *Pattern Recognition Theory and Application*. Springer-Verlag, 1987.

[15] M. M. Crawford, M. R. Ricard, A. Neunschwander, S. Kumar, and J. C. Gibeaut, "Fusion of airborne polarimetric and interferometric SAR data for classification of coastal environments," *IEEE Trans on Geosci and RS*, vol. 37, no. 3, pp. 1306-1315, 1999.

[16] S. Kumar, J. Ghosh, and M.M. Crawford, "Best Basis Feature Exaction Algorithms for Classification of Hyperspectral Data," *IEEE Trans on Geosci and RS*, vol. 39, issue 7, pp. 1368-79, July 2001.

[17] S. Kumar, J. Ghosh, and M. M. Crawford, "Classification of hyperspectral data using best-bases feature extraction algorithms," *Proc. of SPIE: Applications and Science of Computational Intelligence III*, vol. 4055, pp. 362-73, April 2000.

[18] S. Kumar and J. Ghosh, "GAMLS: A generalized framework for associative modular learning systems," (invited paper). In *Proceedings of the Applications and Science of Computational Intelligence II*, pp. 24-34, Orlando, Florida, 1999.

[19] S.J. Raudys and A.K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners", *IEEE Trans on PAMI*, vol. 13, no. 3, pp. 252–64, March 1991.

[20] Qiong Jackson and David Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set", *IEEE Trans on Geosci and RS*, vol. 39, issue 12, pp. 2664-79, Dec 2001.

[21] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 2nd ed., 1985.

[22] Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

[23] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 1984.

[24] L. Jimenez and D.A. Landgrebe, "Supervised classification in high dimensional space: geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Trans on System, Man, and Cybernetics*, vol. 28, part C, no. 1, pp. 39-54, Feb. 1998.

[25] C. Lee and D.A. Landgrebe, "Analyzing high dimensional multispectral data," *IEEE Trans on Geosci and RS*, vol. 31, no. 4, pp. 792-800, 1993.

[26] Andrew Webb, *Statistical pattern recognition.* London: Oxford University Press, 1999.

[27] K. Fukunaga and R.R. Hayes, "Effects of sample size in classifier design", *IEEE Trans on PAMI*, vol. 11, no. 8, pp. 873-85, 1989.

[28] J.A. Richards and Xiuping Jia, *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag, Berlin, Germany, 3rd ed., 1999.

[29] J.H. Friedman, "Regularized discriminant analysis", *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165-75, 1989.

[30] Marina Skurichina, "Stabilizing weak classifiers," Thesis, Vilnius State Univesity, 2001.

[31] L. Breiman, "Bagging predictors," *Machine Learning*, vol 24, no. 2, pp. 123-40, 1996.

[32] A. McCallum, R. Rosenfeld, T. Mitchell, and A.Y. Ng, "Improving text classification by shrinkage in a hierarchy of classes," Proc. 15th International Conf. on Machine, Madison, WI, Morgan Kaufmann, San Mateo, CA, pp. 359-67 1998.

[33] T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, 40(2): 139-58, 2000.

[34] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed, Boston, 1990.

[35] Bradley Efron, "The jackknife, the bootstrap, and other resampling plans," *Society for Industrial and Applied Mathematics VII, Regional conference series in applied mathematics*, 38, 92 p., Philadelphia, PA: 1982.

[36] Y. Freund and R.E. Schapire, "Experiments with a new boosting algorithm", *Proceedings of the 13th International Conference on Machine Learning*, pp. 148-56. Morgan Kaufmann, 1996.

[37] Sarunas Raudys and Robert P. W. Duin, "Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix," *Pattern Recognition Letters*, vol 19, pp 385-92, April 1998.

[38] M. Skurichina and R.P.W. Duin, "Stabilizing classifiers for very small sample sizes", Proc. 13th Int. Conf. on Pattern Recognition (Vienna, Austria, Aug.25-29) Vol. 2, Track B: Pattern Recognition and Signal Analysis, IEEE Computer Society Press, Los Alamitos, pp. 891-6, 1996.

[39] P. Langley, "Selection of relevant features in machine learning," In *Proceedings of AAAI Fall Symposium on Relevance*, AAAI, September 1994.

[40] *Center for Space Research, The University of Texas at Austin, Remote Sensing Program Home Page.* http://www.csr.utexas.edu/rs/research/ksc/index.html.

[41] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Proceedings of the 11th Annual Conf. Computational Learning Theory*, pp. 92-100, 1998.

[42] *Jet Propulsion Lab (JPL), California Institute of Technology, Home Page*. http://makalu.jpl.nasa.gov/.

[43] B. Jeon and D. Landgrebe, "Partially supervised classification using weighted unsupervised clustering," *IEEE Trans on Geosci and RS*, vol. 37, no. 2, pp. 1073-9, March 1999.

[44] T.M. Mitchell, "The role of unlabeled data in supervised learning," *Proc. Sixth International Colloquium on Cognitive Science*, 8pgs, 1999.

[45] V.R. de Sa, "Learning classification with unlabeled data," *Advances in Neural Information Processing Systems 6*, 1994.

[46] B.M. Shahshahani and D.A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans on Geosci and RS*, vol. 32, no. 5, pp. 1087-95, 1994.

[47] T. Cocks, R. Jenssen, A. Stewart, I. Wilson, and T. Shields, "The HyMap airborne hyperspectral sensor: the system, calibration and performance", Proc. 1st EARSeL Workshop on Imaging Spectroscopy (M. Schaepman, D. Schläpfer, and K.I. Itten, Eds.), Zurich, EARSeL, Paris, pp. 37-42, 6-8 October, 1998.

[48] Rasih Ustun, Spectral/Spatial Classification and Output-Based Fusion for Multisensor Remotely Sensed Image Data, MSE Thesis, University of Texas at Austin, 2000.

[49] C. Furlanello and S. Merler, "Boosting of tree-based classifiers for predictive risk modeling in GIS. In Roli F. (ed), *First International Workshop on Multiple Classifier Systems*, Lecture Notes in Computer Sciences, Springer Verlag, London, pp. 220-9, 2000.

[50] Eric Bauer and Ron Kohavi, "An empirical comparison of voting classification algorithms: bagging, boosting, and variants," *Machine Learning*, vol. 36, pp. 105-42, 1999.

[51] Christopher D. Elvidge and Zhikang Chen, "Comparison of broad-band and narrow-band red and near-infrared vegetation indices," *Remote Sensing of the Environment*, vol. 54, pp. 38-48, 1995.

[52] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance", *IEEE Trans on PAMI*, vol. 19, no. 2, pp. 153-8, 1997.

[53] L. Kanal, B. Chandrasekaran, "On dimensionality and sample size in statistical pattern classification," *Pattern Recognition*, vol.3, pp. 225-34, 1971.

[54] Tin Kam Ho, J.J. Hull, and S.N. Srihari, "Decision combination in multiple classifier systems", *IEEE Trans on PAMI*, vol. 16, no. 1, pp. 66-75, 1994.

[55] M. Skurichina, R.P.W. Duin, "The role of combining rules in bagging and boosting" in Advances in Pattern Recognition (Proc. Joint International Workshops SSPR 2000 and SPR 2000, Alicante, Spain, August/September 2000), F.J. Ferri, J.M. Inesta, A. Amin and P. Pudil (eds.), Lecture Notes in Computer Science, vol. 1876, Springer-Verlag, Berlin, pp. 631-40, 2000.

[56] G.F. Hughes, "On the mean accuracy of statistical pattern recognizers", *IEEE Trans on Information Theory*, vol. IT-14, no. 1, pp. 55-63, 1968.

[57] R.A. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*. 4th ed., New Jersey: Prentice-Hall, 1998.

[58] K. Rose, E. Gurewitz, and G. C. Fox, "Statistical mechanics and phase transitions in clustering," *Physical Review Letters*, 65(8), pp. 945-8, 1990.

[59] Kenneth Rose, E. Gurewitz, and G.C. Fox, "Vector quantization by deterministic annealing," *IEEE Transactions on Information Theory*, vol. 38, no. 4, pp. 1249-57, July 1992.

[60] J. Bruske and G. Sommer, “Intrinsic dimensionality estimation with optimally topology preserving maps”, *IEEE Trans on PAMI*, vol. 20, no. 5, pp. 572-5, 1998.

[61] Peter J. Verveer and Robert P.W. Duin, “An evaluation of intrinsic dimensionality estimators”, *IEEE Trans on PAMI*, vol. 17, no. 1, pp. 81-6, 1995.

[62] X. Jia, Classification Techniques for Hyperspectral Remote Sensing Image Data. PhD Thesis, Univ. College, ADFA, University of New South Wales, Australia, 1996.

[63] X. Jia and J.A. Richards, “Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification”, *IEEE Trans on Geosci and RS*, vol. 37, no. 1, pp. 538-42, 1999.

[64] D. Landgrebe, “On the relationship between class definition precision and classification accuracy in hyperspectral analysis”, *Proceedings of IEEE Geoscience and Remote Sensing Symposium*, Honolulu HI, July 24-8, 2000.

[65] S. Watanabe, *Pattern Recognition: Human and Mechanical*. New York: John Wiley & Sons, 1985.

[66] T. Hastie and R. Tibshirani, “Classification by pairwise coupling,” *Advances in Neural Information Processing Systems*, vol. 10, pp. 507-13, MIT Press, Cambridge Massachusetts, 1998.

[67] N. Saito and Ronald R. Coifman, “Local discriminant bases,” *In Mathematical Imaging: Wavelet Applications in Signal and Image Processing II, Proc. of SPIE*, vol. 2303, pp. 2-14, 1994.

[68] A. W. Whitney, “A direct method of nonparametric measurement selection,” *IEEE Transactions on Computers*, vol. 20, pp. 1100-3, 1971.

[69] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and regression trees*. Wadsworth, Belmont, 1984.

[70] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers" *Connection Science*, spl. issue on Combining, vol 8, no 3/4, pp. 385-404, Dec 1996.

[71] S. Kumar, J. Ghosh, and M. M. Crawford, "A hierarchical multiclassifier system for hyperspectral data analysis", First International Workshop on Multiple Classifier Systems, Sardinia, Italy, pp. 270-9, June 2000.

[72] David Landgrebe, "Hyperspectral image data analysis as a high dimensional signal processing problem," (Invited), Special Issue of the *IEEE Signal Processing Magazine*, vol 19, no 1 pp. 17-28, January 2002.

[73] S. Kumar, J. Ghosh and M. M. Crawford, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Pattern Analysis and Applications journal special issue on Classifier Fusion* (submitted).

[74] S.L. Lohr, *Sampling: Design and Analysis*. Brooks/Cole, Pacific Grove, CA, 1999.

[75] H. Drucker and C. Cortes, "Boosting decision trees," *Neural Information Processing 8*, Morgan-Kaufman, eds D.S. Touretzky, M.C. Mozer, and M.E. Hassemo, pp. 479-85, 1996.

[76] Y. Freund and R.E. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, 14(5), pp. 771-80, 1999.

[77] Bradley Efron, "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, vol 7, pp. 1-26, 1979.

[78] D. Rubin, "The Bayesian bootstrap," *The Annals of Statistics*, vol 9, ppl 130-4, 1981.

[79] R.J. Tibshirani and K. Knight, "Model search and inference by bootstrap 'bumping'". Technical report, Dept. of Statistics, University of Toronto, 1995.

[80] R. Cole, R. Hariharan, and P. Indyk, "Tree pattern matching and subset matching in deterministic $O(n \log^3 n)$-time," Technical report, NYU, 2000.

[81] J. T.-L. Wang, B.A. Shapiro, D. Shasha, K. Zhang, and K.M. Currey, "An algorithm for finding the largest approximately common substructures of two trees," *IEEE Trans. PAMI*, vol. 20, no. 8, pp. 889-95, 1998.

[82] J. T.-L. Wang, K. Zhang, K. Jeong, and D. Shasha, "A system for approximate tree matching," *IEEE Trans. Knowledge and Data Engineering*, vol. 6, no. 4, pp. 559-71, 1994.

[83] R.E. Schapire, "Theoretical views of boosting," *Computational Learning Theory: Fourth European Conference (EuroCOLT'99)*, pp. 1-10, 1999.

[84] Joe Hoffbeck and David A. Landgrebe, "Effect of radiance-to-reflectance transformation and atmosphere removal on maximum likelihood classification accuracy of high-dimensional remote sensing data," *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'94*), pp. 3289-94, Pasadena, Calif.

[85] T. Yamazaki and D. Gingras, "Unsupervised multispectral image classification using MRF models and VQ method," *IEEE Trans on Geosci and RS*, vol. 37, no. 2, pp. 1173-6, 1999.

[86] A. Ifarraguerri and C. Chang, "Unsupervised hyperspectral image analysis with projection pursuit," *IEEE Trans on Geosci and RS*, vol. 38, no. 6, pp. 2529-38, 1990.

[87] Shao-Shan Chiang, Chein-I Chang, and I.W. Ginsberg, "Unsupervised hyperspectral image analysis using independent component analysis," *Proceedings IGARSS 2000,* vol. 7, pp. 3136-8, 2000.

[88] D.A. Langan, J.W. Modestino, and J. Zhang, "Cluster validation for unsupervised stochastic model-based image segmentation," *IEEE Transactions on Image Processing*, vol 7, 180-95, 1998.

[89] R. Dave, "Validating fuzzy partitions obtained through c-cells clustering," *Pattern Recognition Letter,* vol. 17, pp. 613-23, 1996.

[90] I. Gath and B. Geva, "Unsupervised optimal fuzzy clustering", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773-80, 1989.

[91] M. Ramze Rezaee, B. P. F. Lelieveldt and J. H. C. Reiber, "A new cluster validity index for the fuzzy *c*-mean", *Pattern Recognition Letters*, vol. 19, pp. 237-46, 1998.

[92] S. Lee and M.M. Crawford, "Unsupervised classification for multi-sensor data in remote sensing using Markov random field and maximum entropy method," *IGARSS*, vol. 2, pp. 1200-2, 1999.

[93] Y. Rangsanseri, "A fuzzy clustering of multispectral images based on integrated spectral and spatial features," *Proceedings IGARSS 1999*, vol. 2, pp. 1306-8, 1999.

[94] J. Garcia-Consuegra, G. Cisneros, and A. Martinez, "Establishing spatially continuous patterns in a non-supervised way**,"** *Proceedings IGARSS 1999*, vol. 2, pp. 726-8, 1999.

[95] Pascale Masson and Wojciech Pieczynski, "SEM algorithm and unsupervised statistical segmentation of satellite images," *IEEE Trans on Geosci and RS*, vol. 31, no. 3, 1993.

[96] H. Ren, and C. Chang, "A generalized orthogonal subspace projection approach to unsupervised multispectral image classification," *IEEE Trans on Geosci and RS*, vol. 38, no. 6, pp. 2515-28, 2000.

[97] K. Bollacker and J. Ghosh, "Knowledge Reuse in multiclassifier systems", *Pattern Recognition Letters*, pp. 1385-90, Nov 1997.

[98] K. Bollacker and J. Ghosh, "A Supra-Classifier Architecture for Scalable Knowledge Reuse", Proc. Intl. Conf. on Machine Learning (ICML-98), Madison, WI, pp. 64-72, 1998.

[99] K. Bollacker and J. Ghosh, "Knowledge Reuse Mechanisms for Categorizing Related Image Sets", *Soft Computing and Image Processing*, S.K. Pal, A. Ghosh and M.K. Kundu (eds.), Physica-Verlag, Heidelberg, 2000.

[100] A.A. Green, M. Berman, P. Switzer, and M.D. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal", *IEEE Trans on Geosci and RS*, vol. 26, no. 1, pp. 65-74, 1988.

[101] X. Ge and P. Smyth, "Deformable Markov model templates for time-series pattern matching," Technical report, Department of Information and Computer Science, University of California, Irvine, 2000.

[102] D.J. Berndt, "Using dynamic time warping to find patterns in time series," *AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94)*, pp. 359-70, 1994.

[103] G. Das, D. Gunopulos, and H. Mannila, "Finding similar time series," *PKDD'97*, 1997.

[104] R. Hughey and A. Krogh, "Hidden Markov models for sequence analysis: extension and analysis of the basic method," *CABIOS,* vol. 12, no. 2, pp. 95-107, 1996.

Vita

Joseph Troy Morgan was born in Flint, Michigan, the second of four sons of Gary and Christina Marie Morgan. A 1991 graduate of Whittemore-Prescott High School, Whittemore, Michigan, he entered the United States Air Force Academy, Colorado Spring, Colorado that summer and graduated 31May95. He received the degree of Master of Science in Engineering from Arizona State University, Tempe, Arizona, in August 1996. In August 1998 he entered the Operations Research, Industrial Engineering Program at the University of Texas, Austin, Texas.

The views expressed in this article are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

Permanent address: 326 Dewitt Ct, O'Fallon, IL 62269

This dissertation was typed by the author.