

**UNIVERSITY OF SOUTHAMPTON**

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

School of Geography

**Land Cover Classification: Refining Training Requirements for  
Support Vector Machine Classification using Remotely Sensed Data**

by

**Ajay Mathur**

Thesis for the degree of Doctor of Philosophy

September 2005

**ABSTRACT**

FACULTY OF ENGINEERING, SCIENCE & MATHEMATICS

SCHOOL OF GEOGRAPHY

Doctor of Philosophy

**Land Cover Classification: Refining Training Requirements for Support Vector Machine Classification using Remotely Sensed Data**

Ajay Mathur

Conventional approaches to training a supervised image classification aim to fully describe all of the classes spectrally. To achieve this, a large training set is typically required. Much of the literature on training data are based on the classical view of classification process emphasizing a large training set. It is not, however, always necessary to have training statistics that can potentially provide a complete and representative description of the classes, especially if using non-parametric classifiers. For classification by a support vector machine (SVM), only the training samples that are support vectors, which lie on part of the edge of the class distribution in feature space, are required; all other training samples provide no contribution to the classification analysis and can effectively be discarded without compromising the accuracy of the classification. The work presented here mainly focuses on the issue of reducing the training data requirements by exploiting the potential of SVM classifier.

First, an SVM analysis was evaluated against a series of classifiers with particular regard to the effect of training set size on classification accuracy. For each classification, accuracy was positively related with training set size. In general, the most accurate classifications were derived from the SVM approach, and with the largest training set the SVM classification were more accurate (93.75%) than that derived from the discriminant analysis (90.00%), decision tree (90.31%) and artificial neural networks (92.18 %). The SVM classifier used about 50 per cent of the training data as support vectors.

If the regions likely to furnish support vectors could be identified prior to the classification, it may be possible to intelligently select useful training samples. This was explored for the classification of agricultural crops in Feltwell area of U.K. The support vectors of one of the crops, wheat, were mainly derived from peat soils. Thus the ability to target useful training samples, in this case, based on soil type may allow accurate classification from small training sets in case the analysis is repeated in future.

The training data requirements may be reduced if there is a prior knowledge or some ancillary information that can be used to identify/locate training sites to regions from which the most informative training samples, the support vectors can be derived. This allows an intelligent training acquisition scheme to be devised in advance of training acquisition process and should include the variables affecting the spectral response of the classes. This was demonstrated for agricultural classes in south western part of Punjab state of India. Considering all the growth stages of the crops and background properties (water and soil) of the training sites provided appropriate support vectors central to the establishment of SVM classifier. The scheme was successful in its intent to capture support vectors directly from field as 70 % of the training samples collected were used by SVM as support vectors as compared to 47.7 % for conventional training scheme. The intelligent scheme of training data acquisition was cheaper by 26.09 per cent over the conventional scheme of training data acquisition because of reduced training set size.

The training data requirements can also be reduced when the concern is to map accurately only one class from the many land cover classes available in the study area. In such instances training data should be limited to the class of interest and classes facing the class of interest in feature space. This was demonstrated here in accurately mapping cotton crop.

The research thus illustrates the potential to direct training data acquisition strategies to target the most useful training samples to allow efficient and accurate image classification by SVM.

# CONTENTS

<b>List of Tables</b> .....	<i>v</i>
<b>List of Figures</b> .....	<i>xii</i>
<b>Declaration of Authorship</b> .....	<i>xvi</i>
<b>Acknowledgements</b> .....	<i>xviii</i>
<b>Abbreviations</b> .....	<i>xx</i>
<b>CHAPTER 1 - Introduction</b> .....	<b>1</b>
1.1 Introduction .....	1
1.2 Thesis Overview .....	5
<b>CHAPTER 2 - Literature Review</b> .....	<b>7</b>
2.1 Introduction.....	7
2.2 Preprocessing .....	9
2.2.1 Feature Reduction.....	9
2.2.2 Radiometric Preprocessing .....	10
2.2.3 Geometric Correction .....	10
2.3 Classification.....	11
2.3.1 Unsupervised Classification .....	12
2.3.2 Supervised Classification .....	12
2.3.2.1 Parametric Classifiers.....	13
2.3.2.2 Non-parametric Classifiers .....	13
2.3.3 Stages in Supervised Classification.....	13
2.3.3.1 Training Stage .....	15
2.3.3.1.1 Design of Training Strategy.....	15
2.3.3.1.1.1 Time of Sampling.....	16
2.3.3.1.1.2 Sample Type.....	16
2.3.3.1.1.3 Number of Training Samples .....	17
2.3.3.1.1.4 Sample Design.....	18
2.3.3.2 Allocation .....	22
2.3.3.3 Accuracy Assessment.....	23
2.3.3.3.1 Design of Sampling for Reference Data Acquisition.....	24
2.3.3.3.2 Number of Testing Data Samples .....	25
2.3.3.3.3 Error Matrix .....	26
2.3.3.3.3.1 Comparison of Error Matrices.....	28
2.3.4 Problems in Land Cover Classification .....	29
2.3.4.1 Issues Related to Characteristics of Remote Sensing Data.....	29
2.3.4.2 Nature of the Classes.....	30
2.3.4.3 Methods used in the Analysis.....	31
2.4 Maximum-likelihood .....	32
2.4.1 Design of MLC Classifier .....	33
2.4.1.1 Thresholding.....	34

2.4.2	Limitations of Maximum-likelihood Classifier .....	35
2.5	Artificial Neural Networks.....	36
2.5.1	Multi-layer Perceptrons .....	36
2.5.1.1	Training of Multi-layer Perceptron .....	37
2.5.2	Limitations of Artificial Neural Network .....	42
2.6	Decision Trees.....	44
2.6.1	Classification of Decision Trees.....	45
2.6.1.1	Univariate Decision Tree.....	45
2.6.1.2	Multivariate Decision Tree.....	46
2.6.1.3	Hybrid Decision Tree .....	46
2.6.2	Design of a Decision Tree .....	47
2.6.2.1	Bottom-up Approach.....	48
2.6.2.2	Top-down Approach.....	49
2.6.2.2.1	Selection of Node Splitting Rules .....	49
2.6.2.2.1.1	Information Gain and Information Gain Ratio .....	50
2.6.2.3	Hybrid Approach.....	51
2.6.3	Pruning of Decision Trees .....	51
2.6.3.1	Pessimistic Error Pruning.....	53
2.6.4	Limitations of Decision Tree Classification .....	55
2.7	Support Vector Machines.....	56
2.7.1	Design of Support Vector Machines .....	56
2.7.1.1	Linearly Separable Case.....	57
2.7.1.2	Linearly Non-separable Case .....	62
2.7.1.3	Decision Surfaces .....	63
2.7.2	Multi-Class Support Vector Machine.....	65
2.7.3	Limitations of Support Vector Machine .....	67
2.8	Conclusions.....	68
<b>CHAPTER 3 - Relative Evaluation of Multi-Class Image Classification by SVM.....</b>		<b>70</b>
3.1	Introduction.....	70
3.1.1	Study Area and Data Used .....	70
3.1.1.1	Characteristics of Training Data.....	71
3.1.2	Methodology .....	73
3.1.2.1	Training Set .....	74
3.1.2.2	Algorithms Used .....	74
3.1.2.3	Testing Set.....	75
3.1.3	Accuracy Assessment.....	77
3.1.4	Results .....	78
3.1.4.1	Discriminant Analysis (DA).....	79
3.1.4.2	Artificial Neural Network.....	80
3.1.4.3	Decision Tree .....	81
3.1.4.4	Support Vector Machine .....	83
3.1.5	Results and Discussion.....	85

3.1.6	Conclusions .....	87
<b>CHAPTER 4 - Reducing Requirement of Training Data by Relating Support Vectors with Soil Type of Training Fields .....</b>		
<b>92</b>		
4.1	Introduction.....	92
4.1.1	Data and Methods of Classification .....	92
4.1.2	Results and Discussion .....	95
4.2	Conclusions.....	100
<b>CHAPTER 5 - Intelligently Reducing Training Requirements of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification.....</b>		
<b>101</b>		
5.1	Introduction.....	101
5.2	Study Area .....	103
5.2.1	Description of Study Area.....	104
5.2.1.1	Methodology of the CAPE Project.....	111
5.2.1.1.1	Drawbacks of the CAPE project .....	114
5.3	Data.....	115
5.3.1	Conventional Training Data Scheme.....	117
5.3.2	Intelligent Training Data Scheme.....	121
5.4	Methodology of Classification.....	134
5.5	Results and Discussions .....	136
5.5.1	An Assessment of Ability to Intelligently Identify Most Useful Training Samples (Support Vectors) Directly from Field.....	136
5.5.2	Relative Accuracy .....	138
5.5.2.1	Discriminant Analysis .....	138
5.5.2.2	Decision Tree .....	139
5.5.2.3	Artificial Neural Networks.....	141
5.5.2.4	Support Vector Machine .....	142
5.5.2.5	Discussion .....	143
5.5.2.5.1	Analysis of Classifications Trained with Conventional Scheme of Training Data Acquisition .....	144
5.5.2.5.2	Analysis of Classifications Trained with Intelligent Scheme of Training Data Acquisition .....	144
5.5.2.5.3	Comparison between Classifications Trained with Conventional and Intelligent Scheme of Training Data Acquisition .....	146
5.5.3	Identifying Support Vectors with Ground Attributes of the Training Sites.....	147
5.5.4	Financial Implication of Reducing the Requirement of Training Data .....	150
5.5.5	Reduced Requirement of Training Data for Classifying Accurately only One Class.....	152
5.5.6	Summary .....	158
5.5.7	Conclusions .....	162
<b>CHAPTER 6 - Summary and Conclusions.....</b>		
<b>164</b>		
6.1	Summary.....	164
6.2	Conclusions.....	169

6.3 Future Work.....	171
<b>APPENDIX.....</b>	<b>172</b>
<b>REFERENCES.....</b>	<b>188</b>

## LIST OF TABLES

<b>Table 2.1:</b> Minimum sample size necessary per category (after Van Genderen <i>et al.</i> , 1978).	25
<b>Table 2.2:</b> Error matrix of a classification.	27
<b>Table 3.1:</b> Statistics of training data showing minimum (MIN), maximum (Max), mean and standard deviation (Sd) of digital numbers of training data of the six classes in the three bands.	73
<b>Table 3.2:</b> Variables studied in the study.	73
<b>Table 3.3:</b> Combination of training and testing set size for analysis. The training and testing set size has been abbreviated with prefix to N as the number of pixels and suffix as the iteration number.	77
<b>Table 3.4:</b> 2x2 error matrix to calculate the statistical significance of differences in classification accuracy based on M <sup>c</sup> Nemar test for related samples.	77
<b>Table 3.5:</b> Overall and class wise accuracy (%) on testing data using discriminant analysis for case A (all available testing data) analyses.	79
<b>Table 3.6:</b> Overall and class wise accuracy (%) on testing data using discriminant analysis for case B (17 pixels per class in testing data) analyses.	80
<b>Table 3.7:</b> Overall and class wise accuracy (%) on testing data using artificial neural network for case A (all available testing data) analyses.	80
<b>Table 3.8:</b> Overall and class wise accuracy (%) on testing data using artificial neural network for case B (17 pixels per class in testing data) analyses.	81
<b>Table 3.9:</b> Overall and class wise accuracy (%) on testing data using decision tree algorithm for case A (all available testing data) analyses.	82
<b>Table 3.10:</b> Overall and class wise accuracy (%) on testing data using decision tree for case B (17 pixels per class in testing data) analyses.	82
<b>Table 3.11:</b> Overall and class wise accuracy (%) on testing data using support vector machine for case A (all available testing data) analyses.	83
<b>Table 3.12:</b> Overall and class wise accuracy (%) on testing data using support vector machine for case B (17 pixels per class for testing data) analyses.	84
<b>Table 3.13:</b> Overall accuracy for case A analysis.	89
<b>Table 3.14:</b> Overall accuracy for case B analysis.	90
<b>Table 3.15:</b> Significance value (Z) of differences between accuracies of testing set obtained when the classifiers were trained with smallest training set size of 15 pixels and largest available size of 100 pixels per class at 95 % confidence level. Differences significant at the 95% confidence level ( $Z \geq 1.96$ ) are	91

highlighted in bold with positive values indicating higher accuracy when training data was 100 pixels/class.

<b>Table 3.16:</b> Comparison of classification accuracy statements. The classifications derived with each method (SVM = support vector machine, DA = discriminant analysis, DT = decision tree and ANN =artificial neural network) at each size of training set for case A (all testing set), defined by the number of cases of each class, were compared using a M <sup>c</sup> Nemar test. Differences significant at the 95% confidence level ( $Z \geq 1.96$ ) are highlighted in bold with positive values indicating that the first named classifier had the higher accuracy.	91
<b>Table 3.17:</b> Comparison of classification accuracy statements. The classifications derived with each method (SVM = support vector machine, DA = discriminant analysis, DT = decision tree and ANN = artificial neural network) at each size of training set for case B (testing set comprising of 17 pixels per class), defined by the number of cases of each class, were compared using a M <sup>c</sup> Nemar test. Differences significant at the 95% confidence level ( $Z \geq 1.96$ ) are highlighted in bold with positive values indicating that the first named classifier had the higher accuracy.	91
<b>Table 4.1:</b> Results of 5n cross-validation on training data for optimal selection of parameters $C$ and $\gamma$ . The value in the bold gives the highest accuracy obtained on training data with parameters $C=2^4$ and $\gamma=2^{-8}$ respectively.	96
<b>Table 4.2:</b> Support vectors when training data comprised of pixels from both type of soils with parameter settings of $C$ and $\gamma$ of $2^4$ and $2^{-8}$ respectively deduced from 5n cross validation. The first column shows the $\alpha$ values of each support vector followed by spectral values of the support vector (training data) in the three bands under B3, B2 and B1.	97
<b>Table 4.3:</b> Support vectors with $\alpha$ values.	98
<b>Table 4.4:</b> Confusion matrix of testing set for both the analysis (classifier trained with or without wheat pixels from peat soils).	99
<b>Table 5.1:</b> Comparative area under rice and cotton in Punjab state (Source: Director of Land Records, Punjab, 2004).	107
<b>Table 5.2:</b> Statistics of training data showing minimum (Min), maximum (Max), Mean and standard deviation (Sd) of digital numbers of the training data of the five classes in the three bands.	119
<b>Table 5.3:</b> Variables considered in the intelligent scheme of training data collection for cotton crop.	125
<b>Table 5.4:</b> Variables considered in the intelligent scheme of training data collection for rice basmati crop.	126
<b>Table 5.5:</b> Variables considered in the intelligent scheme of training data collection for rice local crop.	127



<b>Table 5.6:</b> Statistics of training data showing minimum (Min), maximum (Max), Mean and standard deviation (Sd) of digital numbers of the training data of the five classes in the three bands.	134
<b>Table 5.7:</b> Variables completed in the study.	135
<b>Table 5.8:</b> Error matrix of testing set for the classification derived from the discriminant analysis (DA) trained by data acquired under conventional scheme.	138
<b>Table 5.9:</b> Error matrix of testing set for the classification derived from the discriminant analysis (DA) trained by data acquired under intelligent scheme.	138
<b>Table 5.10:</b> Error matrix of testing set for the classification derived from the decision tree (DT) trained by data acquired under conventional scheme.	139
<b>Table 5.11:</b> Error matrix of testing set for the classification derived from the decision tree (DT) trained by data acquired under intelligent scheme.	139
<b>Table 5.12:</b> Error matrix of testing set for the classification derived from the artificial neural network (ANN) trained by data acquired under conventional scheme.	141
<b>Table 5.13:</b> Error matrix of testing set for the classification derived from the artificial neural network (ANN) trained by data acquired under intelligent scheme.	141
<b>Table 5.14:</b> Error matrix of testing set for the classification derived from the support vector machine (SVM) trained by data acquired under conventional scheme.	142
<b>Table 5.15:</b> Error matrix of testing set for the classification derived from the support vector machine (SVM) trained by data acquired under intelligent scheme.	142
<b>Table 5.16:</b> Significance value ( $Z$ ) of differences between accuracies of testing set obtained when the classifiers were trained with training data collected under conventional and by intelligent scheme of training data collection. Differences significant at the 95% confidence level ( $Z \geq  1.96 $ ) are highlighted in bold with positive values indicating higher accuracy when classifier trained with training data collected under conventional scheme.	146
<b>Table 5.17:</b> Comparison of classification accuracy statements for classifications trained with conventional and intelligent scheme of training data acquisition (SVM = support vector machine, DA = discriminant analysis, DT = decision tree and ANN =artificial neural network). Differences significant at the 95% confidence level ( $Z \geq  1.96 $ ) are highlighted in bold with positive values indicating that the first named classifier had the higher accuracy.	147
<b>Table 5.18:</b> Parameters used to model the classifiers. The architecture in ANN describes the input layers, the nodes in the middle layer and the output layers. The network's architecture were defined from an evaluation of several hundreds of candidate networks.	147





<b>Table 5.19:</b> Comparison of expenditure incurred on classification process based on conventional with intelligent scheme of training data acquisition. The cost has been calculated in Indian Rupees with approximate rates prevalent in India for the work, though part of the analysis has been carried in United Kingdom too. (1 US dollars (USD) = Rs 43.40 and 1 United Kingdom Pounds (GBP) = Rs 80.26 as on 15 May 2005).	152
<b>Table 5.20:</b> Error matrix of testing set for the classification derived from the support vector machine (SVM) trained by data acquired under intelligent scheme for all classes except rice basmati.	155
<b>Table A.1:</b> Error matrix for the classification derived from the DA trained with training set 5n (containing 15 cases of each class) for case A analysis.	172
<b>Table A.2:</b> Error matrix for the classification derived from the DA trained with training set 10n (containing 30 cases of each class) for case A analysis.	172
<b>Table A.3:</b> Error matrix for the classification derived from the DA trained with training set 15n (containing 45 cases of each class) for case A analysis.	172
<b>Table A.4:</b> Error matrix for the classification derived from the DA trained with training set 20n (containing 60 cases of each class) for case A analysis.	172
<b>Table A.5:</b> Error matrix for the classification derived from the DA trained with training set 25n (containing 75 cases of each class) for case A analysis.	173
<b>Table A.6:</b> Error matrix for the classification derived from the DA trained with training set 30n (containing 90 cases of each class) for case A analysis.	173
<b>Table A.7:</b> Error matrix for the classification derived from the DA trained with the largest training set (containing 100 cases of each class) for case A analysis.	173
<b>Table A.8:</b> Error matrix for the classification derived from the DA trained with training set 5n (containing 15 cases of each class) for case B analysis.	173
<b>Table A.9:</b> Error matrix for the classification derived from the DA trained with training set 10n (containing 30 cases of each class) for case B analysis.	174
<b>Table A.10:</b> Error matrix for the classification derived from the DA trained with training set 15n (containing 45 cases of each class) for case B analysis.	174
<b>Table A.11:</b> Error matrix for the classification derived from the DA trained with training set 20n (containing 60 cases of each class) for case B analysis.	174
<b>Table A.12:</b> Error matrix for the classification derived from the DA trained with training set 25n (containing 75 cases of each class) for case B analysis.	174
<b>Table A.13:</b> Error matrix for the classification derived from the DA trained with training set 30n (containing 90 cases of each class) for case B analysis.	175
<b>Table A.14:</b> Error matrix for the classification derived from the DA trained with the largest training set (containing 100 cases of each class) for case B.	175

<b>Table A.15:</b> Error matrix for the classification derived from the DT trained with training set 5n (containing 15 cases of each class) for case A analysis.	175
<b>Table A.16:</b> Error matrix for the classification derived from the DT trained with training set 10n (containing 30 cases of each class) for case A analysis.	175
<b>Table A.17:</b> Error matrix for the classification derived from the DT trained with training set 15n (containing 45 cases of each class) for case A analysis.	176
<b>Table A.18:</b> Error matrix for the classification derived from the DT trained with training set 20n (containing 60 cases of each class) for case A analysis.	176
<b>Table A.19:</b> Error matrix for the classification derived from the DT trained with training set 25n (containing 75 cases of each class) for case A analysis.	176
<b>Table A.20:</b> Error matrix for the classification derived from the DT trained with training set 30 (containing 90 cases of each class) for case A analysis.	176
<b>Table A.21:</b> Error matrix for the classification derived from the DT trained with the largest training set (containing 100 cases of each class) for case A analysis.	177
<b>Table A.22:</b> Error matrix for the classification derived from the DT trained with training set 5n (containing 15 cases of each class) for case B analysis.	177
<b>Table A.23:</b> Error matrix for the classification derived from the DT trained with training set 10n (containing 30 cases of each class) for case B analysis.	177
<b>Table A.24:</b> Error matrix for the classification derived from the DT trained with training set 15n (containing 45 cases of each class) for case B analysis.	177
<b>Table A.25:</b> Error matrix for the classification derived from the DT trained with training set 20n (containing 60 cases of each class) for case B analysis.	178
<b>Table A.26:</b> Error matrix for the classification derived from the DT trained with training set 25n (containing 75 cases of each class) for case B analysis.	178
<b>Table A.27:</b> Error matrix for the classification derived from the DT trained with training set 30n (containing 90 cases of each class) for case B analysis.	178
<b>Table A.28:</b> Error matrix for the classification derived from the DT trained with the largest training set (containing 100 cases of each class) for case B analysis.	178
<b>Table A.29:</b> Error matrix for the classification derived from the ANN trained with training set 5n (containing 15 cases of each class) for case A analysis.	179
<b>Table A.30:</b> Error matrix for the classification derived from the ANN trained with training set 10n (containing 30 cases of each class) for case A analysis.	179
<b>Table A.31:</b> Error matrix for the classification derived from the ANN trained with training set 15n (containing 45 cases of each class) for case A analysis.	179

<b>Table A.32:</b> Error matrix for the classification derived from the ANN trained with training set 20 (containing 20 cases of each class) for case A analysis.	179
<b>Table A.33:</b> Error matrix for the classification derived from the ANN trained with training set 25n (containing 75 cases of each class) for case A analysis.	180
<b>Table A.34:</b> Error matrix for the classification derived from the ANN trained with training set 30n (containing 90 cases of each class) for case A analysis.	180
<b>Table A.35:</b> Error matrix for the classification derived from the ANN trained with the largest training set (containing 100 cases of each class) for case A analysis.	180
<b>Table A.36:</b> Error matrix for the classification derived from the ANN trained with training set 5n (containing 15 cases of each class) for case B analysis.	180
<b>Table A.37:</b> Error matrix for the classification derived from the ANN trained with training set 10n (containing 30 cases of each class) for case B analysis.	181
<b>Table A.38:</b> Error matrix for the classification derived from the ANN trained with training set 15n (containing 45 cases of each class) for case B analysis.	181
<b>Table A.39:</b> Error matrix for the classification derived from the ANN trained with training set 20n (containing 60 cases of each class) for case B analysis.	181
<b>Table A.40:</b> Error matrix for the classification derived from the ANN trained with training set 25n (containing 75 cases of each class) for case B analysis.	181
<b>Table A.41:</b> Error matrix for the classification derived from the ANN trained with training set 30n (containing 90 cases of each class) for case B analysis.	182
<b>Table A.42:</b> Error matrix for the classification derived from the ANN trained with the largest training set (containing 100 cases of each class) for case B analysis.	182
<b>Table A.43:</b> Error matrix for the classification derived from the SVM trained with training set 5n (containing 15 cases of each class) for case A analysis.	182
<b>Table A.44:</b> Error matrix for the classification derived from the SVM trained with training set 10n (containing 30 cases of each class) for case A analysis.	182
<b>Table A.45:</b> Error matrix for the classification derived from the SVM trained with training set 15n (containing 45 cases of each class) for case A analysis.	182
<b>Table A.46:</b> Error matrix for the classification derived from the SVM trained with training set 20n (containing 60 cases of each class) for case A analysis.	183
<b>Table A.47:</b> Error matrix for the classification derived from the SVM trained with training set 25n (containing 75 cases of each class) for case A analysis.	183
<b>Table A.48:</b> Error matrix for the classification derived from the SVM trained with training set 30n (containing 90 cases of each class) for case A analysis.	183

<b>Table A.49:</b> Error matrix for the classification derived from the SVM trained with the largest training set (containing 100 cases of each class) for case A analysis.	183
<b>Table A.50:</b> Error matrix for the classification derived from the SVM trained with training set 5n (containing 15 cases of each class) for case B analysis.	184
<b>Table A.51:</b> Error matrix for the classification derived from the SVM trained with training set 10n (containing 30 cases of each class) for case B analysis.	184
<b>Table A.52:</b> Error matrix for the classification derived from the SVM trained with training set 15n (containing 45 cases of each class) for case B analysis.	184
<b>Table A.53:</b> Error matrix for the classification derived from the SVM trained with training set 20n (containing 60 cases of each class) for case B analysis.	184
<b>Table A.54:</b> Error matrix for the classification derived from the SVM trained with training set 25n (containing 75 cases of each class) for case B analysis.	185
<b>Table A.55:</b> Error matrix for the classification derived from the SVM trained with training set 30n (containing 90 cases of each class) for case B analysis.	185
<b>Table A.56:</b> Error matrix for the classification derived from the SVM trained with the largest training set (containing 100 cases of each class) for case B analysis.	185
<b>Table A.57:</b> Summary of 76 support vectors resulting from SVM analysis using training data acquired under intelligent scheme of training data acquisition. The $\alpha$ values in SVM are between a pair of classes as SVM is basically a binary classifier and, therefore, there are four $\alpha$ values for the five classes in the analysis. The four columns follow some particular order based on class label, for example, the $\alpha$ values for cotton class (label 3) has $\alpha$ values in column 1, column 2, column 3 and column 4 with respect to class 1, 2, 4 and 5 respectively.	186

## LIST OF FIGURES

<b>Figure 2.1:</b> Stages in supervised classification.	14
<b>Figure 2.2:</b> Sampling techniques.	20
<b>Figure 2.3:</b> Graph of a semi-variogram.	21
<b>Figure 2.4:</b> Multi-layer perceptron.	37
<b>Figure 2.5:</b> Error surface.	40
<b>Figure 2.6:</b> Energy surface showing local and global minima.	41
<b>Figure 2.7:</b> Over-fitting of training data.	42
<b>Figure 2.8:</b> Classification of a forest using a decision tree. Each box is a node with root at the top which contains all the data (Infrared values (IR)). Splitting rules based on the value of IR values are applied at each node to make the data in the child nodes homogeneous. The forest is classified as Oak, Maple, Pine and Deodar as shown by the leaf of the decision tree.	45
<b>Figure 2.9:</b> A hybrid decision tree classifier, using splitting rules as linear discriminant function (LDF) at root, k nearest neighbour on the left tree and a univariate decision tree on the right.	47
<b>Figure 2.10:</b> The two classes (  and  ) can be separated by a number of decision surfaces (shown by black lines in between the two classes). However, there is only one decision surface called the optimal separating hyperplane (shown by dark blue line), that is expected to generalize accurately on unseen cases as compared to other decision surfaces.	57
<b>Figure 2.11:</b> Optimal hyperplane (dark black line) with parameters $w$ and $b$ . Parallel planes P1 and P2 contains support vectors relevant in the formulation of optimal hyperplane.	58
<b>Figure 2.12:</b> Training data cannot be separated by single linear separating hyperplane.	63
<b>Figure 2.13:</b> The two classes (  and  ) cannot be separated by linear decision surface but after transformation to a high dimensional feature space through a function $\phi$ , the data can be separated by a linear decision surface (Vapnik, 1995).	64
<b>Figure 3.1:</b> Location of training data in feature space.	71
<b>Figure 3.2:</b> Histograms of training data for the six classes in bands 4, 6, 9.	72

<b>Figure 4.1:</b> Soil map of Feltwell area. The black box represents the bounds of study area. The sand soil comprised of humic gleyic rendzinas (346), brown rendzinas (343g), typical brown sands (551g), typical humic-sandy gley soils (861b), whereas, peat soils comprised of earthy eutro-amorphous peat soil (1024a and b) and earthy eu-fibrous peat soils (1022a) (source: Soil Survey of England and Wales).	93
<b>Figure 4.2:</b> SPOT HRV FCC of study area demarcated into sandy and peat soils by the yellow line. The study area was dominated by winter wheat and barley crops. (Data, courtesy NERC). The area west of yellow line is covered by peat soils and that to its east by sandy soils.	95
<b>Figure 4.3:</b> The distribution of training data of winter wheat and barley class in feature space. The lone support vector of winter wheat class from peat soil is encircled and its $\alpha$ value of 4.32 highlighted.	97
<b>Figure 5.1:</b> Study area shows the districts (1. Bathinda 2. Muktsar 3. Faridkot 4. Moga 5. Part of Ludhiana district) of Punjab state.	104
<b>Figure 5.2:</b> An unlined canal.	106
<b>Figure 5.3:</b> Salt affected land due to waterlogging.	107
<b>Figure 5.4:</b> Wilted rice as a result of waterlogging.	108
<b>Figure 5.5:</b> No watering (dry conditions) of cotton crop after flowering stage. The exposed soil is dry as farmers do not water cotton crop after flowers appear.	110
<b>Figure 5.6:</b> FCC of raw IRS-1D satellite data (date of acquisition 16-09-2002) of selected segments A, B and C of the Muktsar district for cotton area estimation under CAPE project. These segments constitute 15 per cent of Muktsar district in area. The figure also shows that some of the B type segments (marked as C in the FCC) definitely belong to C type. The problem arises because the segments are selected once in 4 -5 years and the area under cotton is fluctuating.	113
<b>Figure 5.7:</b> The newspaper report about waterlogging in the study area.	116
<b>Figure 5.8:</b> Procedure followed for training data acquisition under conventional scheme.	118
<b>Figure 5.9:</b> Spectral distribution of training data collected under conventional training data collection scheme.	119
<b>Figure 5.10:</b> Histograms of the training data collected under conventional training data scheme.	120
<b>Figure 5.11:</b> Procedure followed for training data acquisition under intelligent scheme. The scheme was tested on testing data acquired under conventional scheme.	123

<b>Figure 5.12:</b> Rice fields showing matured crop in far end with nearer fields still green and healthy. This variation can be exploited to capture support vectors.	122
<b>Figure 5.13:</b> Farmers being consulted in field about the crop status in the area.	124
<b>Figure 5.14:</b> Cotton crop in saline land. The white patches of salt due to waterlogging can be seen on the exposed soil. This was expected to increase the spectral response in all three bands.	128
<b>Figure 5.15:</b> Basmati rice was very young and green throughout the study area as such no variability could be observed in field by the naked eye.	128
<b>Figure 5.16:</b> The canopy of basmati rice does not permit soil to be exposed to sky. Thus the contribution of the soil in the spectral response of the crop could be considered only due to its contribution in the growth of the crop.	129
<b>Figure 5.17a:</b> Very matured local rice. NIR values would be low and Red higher as compared to a young healthy crop. Likewise MIR value would be higher as the crop was dry.	129
<b>Figure 5.17b:</b> Very matured local rice NIR values would be low and Red higher as compared to a young healthy crop. Likewise MIR value would be higher as the crop was dry.	130
<b>Figure 5.17c:</b> Very matured local rice adjoining a canal. Water reduces spectral response especially in MIR band.	130
<b>Figure 5.18a:</b> Matured local rice near canal. The leaves have started yellowing and grain formation has set in. Water reduces spectral response especially in MIR band.	131
<b>Figure 5.18b:</b> Matured local rice. Grain formation has taken place and yellowing of leaves has also started. The spectral values would be between young healthy and very matured local rice in similar conditions.	131
<b>Figure 5.19a:</b> Young local rice. The grain formation is there but leaves are still green. The NIR values would be high and Red very low as compared to matured crop.	132
<b>Figure 5.19b:</b> Young local rice. The grain formation is there but leaves are still green. The values in NIR would be high, low in Red and low in MIR (leaves were moist) as compared to matured crop.	132
<b>Figure 5.20:</b> Local rice in area affected by waterlogging. The white salt can be seen on the soil (lower right corner of the photograph). The pump in the field is to drain out water due to waterlogging from the field out into surrounding drain.	133
<b>Figure 5.21:</b> Top view of local rice. The canopy does not expose soil to sky.	133
<b>Figure 5.22:</b> spectral distribution of training data collected under intelligent scheme.	134



**Figure 5.23:** Training data of conventional scheme overlaid by that captured under intelligent scheme. The prefix SV in labels in the legend refers to training data collected by intelligent scheme. The 'A' refers to training data from site with very matured local rice collected under intelligent scheme. 137

**Figure 5.24:** Tree structure when DT was trained by training data collected under conventional scheme. Each box is a node with root at the top which contains all the training data. Splitting rules used the values in the three input bands (Red, NIR and MIR) at nodes to make the data purer in the child nodes. For example, the left node after the root splits the data into child nodes based on values of Red band. Thus the splitting rule  $\text{Red} < 77$  qualifies training data with values less than 77 in Red band for this branch of the tree. The terminal node (last node) circular in shape refers to the classified output with numbers 1 to 5 corresponding to classes Built-up, sand, Cotton, Local rice and Basmati rice in the study. 156

**Figure 5.25:** Tree structure when DT was trained by training data collected under intelligent scheme 157

## ACKNOWLEDGEMENTS

A journey is easier and enjoyable when travelled in a group and this holds good for my thesis as well. This thesis is the result of three years of work supported by many people. It is a wonderful moment that I now have the opportunity to express my gratitude to all of them.

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I want to thank my parent department Punjab Remote Sensing Centre (PRSC), Ludhiana, India for giving me permission to commence this thesis in the first instance, to do the necessary field work and to use departmental data. I am grateful to the Commission of European Community for the ATM data, which were acquired as part of the European AgriSAR campaign, Natural Environmental Research Centre (NERC) for the SPOT HRV data and Soil Survey of England and Wales for the soil map used in the study.

The neural networks used were based on Trajan software. The decision tree was based on CTree algorithm developed by Angshuman Saha. The support vector machines were constructed with BSVM (version 2.01) software developed by Chih-Wei Hsu and Chih-Jen Lin, National University, Taipei. I have further to thank the Commonwealth scholarship commission (U.K) for providing financial support in this great endeavour.

I have great appreciation for my colleagues from my parent department, PRSC in India for all their help and support. Especially I am obliged to Dr. P.K Sharma, Mr. P.K Litoria and Dr. D.C. Loshali for their kind help. I am also indebted to fellow research colleague Mr. Jadunandan Dash for friendly help.

I am also indebted to Agricultural departments located in the study area, newspaper reports and farming community at the field level for providing necessary information to make this work a success.

Indeed words at my command are inadequate for expressing my gratitude to my supervisor Prof. G.M. Foody from the University of Southampton, Southampton, U.K whose help, stimulating suggestions and above all wit encouraged me all through the

research and produce the work in its present form. My thanks are also due to Prof. Paul Curran for critical suggestions about the work during MPhil to Ph.D. upgrade examination.

I also want to thank my parents, who taught me the value of hard work by their own example. I would like to share this moment of happiness with my parents and other family members. Especially, I would like to give my special thanks to my wife Anjly for consistent support. My young children Arpit and Ashima contributed by not disturbing me in my studies at home and are deeply acknowledged here.

## ABBREVIATIONS

ANN	Artificial neural network
ATM	Airborne thematic mapper
DA	Discriminant analysis
DN	Digital number
DT	Decision tree
GCP	Ground control point
GIS	Geographical information systems
IRS	Indian Remote Sensing
LISS	Linear Image Self Scanning
MLC	Maximum-likelihood classifier
MLP	Multi-layer perceptron
OHP	Optimal hyperplane
SVM	Support vector machine

# CHAPTER 1 - Introduction

## 1.1 Introduction

The availability of accurate and up-to-date land-cover maps is crucial for many applications including agriculture, environment and forestry. Remote sensing is one of the efficient tools as compared to conventional methods of surveying in terms of providing land cover information at frequent intervals.

Despite the considerable potential of remote sensing as a source of land cover information many problems are encountered and the accuracy of the derived land cover information is sometimes viewed as insufficient by the user community (Foody, 2002). There are many factors responsible for this situation including: the nature of the classes being studied, properties of sensing system used (*e.g.* spatial and spectral resolutions) to acquire the imagery and the techniques used to extract thematic information from the imagery, the classification techniques (Pal and Mather, 2003).

Supervised classification is one of the widely used approaches in extracting information from remotely sensed data. Supervised classification comprises of three stages: training, allocation and testing. In the training stage generally the areas of known ground identity (training areas) are identified on the image. The spectral response of the training areas (training data) may be used to generate descriptive statistics for the land cover classes such as mean and standard deviation to inform the second stage (allocation stage) of the classification. The accuracy of the classification is evaluated in the testing stage, usually on a sample of cases not used in the training stage.

The value of the classified output generated is typically a function of the accuracy of the classification (Hashemain *et al.*, 2004). The accuracy of supervised classification is generally dependent on the first two stages of the classification over which the analyst has considerable control. As a consequence, means to increase the accuracy of classifications

derived from remotely sensed data have been widely researched. Much research, has, for example, focused on the allocation stage, the classifiers used to classify the data.

Achieving an optimal classification is, however, a challenging and open problem (Ho *et al.*, 1994). The accuracy of classification is also dependent to a large extent on the quality of the training data used to train the classifier. Indeed the nature of the training stage can have a larger impact on classification accuracy than the classification technique used (Hixon *et al.*, 1980; Campbell, 2002). Much research, therefore, focused on issues related with the design of the training stage of a supervised image classification. This includes sampling design (Campbell, 2002), training set size (Congalton, 1991), spacing of training data (Atkinson, 1991) and time of sampling with respect to image acquisition time (Justice and Townshend, 1981). However, the size of training set, the number of samples for training the classifier, has been the core focus as it is costly in terms of time and finance to acquire large training sets (Buchheim and Lillesand, 1989; Jackson and Landgrebe, 2001).

The design of the training stage is often guided by the classical statistical view of the classification process, generally considering a probabilistic algorithm such as the maximum-likelihood classifier (MLC). Statistical classifiers are based on statistical description generated from training data and require a complete description of each class in feature space. For this, a large training set, spread over the entire study area is often required to capture the spectral variability of the classes.

In general, studies have shown that classification accuracy tends to be positively related to training set size (Pal and Mather, 2003; Zhang *et al.*, 1994). Fewer training samples or inappropriate placement of training samples produces statistics which may not be able to characterize the land cover classes. The requirement for large sample sizes is, therefore, not unusual and penalties on classification accuracy for using small training sets can in some cases be severe (Curran and Williamson, 1985). Conventional training data acquisition schemes, therefore, aims to capture a large training set spread all over the study area.

Much research has focused on the potential to reduce the training data requirements without compromising the accuracy of the classification so as to reduce the cost of the classification process. This includes selecting non-autocorrelated (spatially independent) training data by using semi-variograms (Atkinson, 1991; Chen and Stow, 2002), signature extension, establishing permanent ground data sites, reducing the dimensionality of the data to avoid Hughes phenomenon (for finite training samples, accuracy first increases with dimensionality and then decreases)(Melgani and Bruzzone, 2004).

There are many recommendations made as to the required size of the training set, typically based on the classical statistical view of the classification process. For example, Lillesand *et al.*, 2004 related the requirement of training data with spectral bands used per class and proposed a minimum of 10 to 100 times the discriminatory bands used. However, such recommendations are general and are based without any regard to the study area or the complexity of the classes therein or the classifier to be used.

Different classifier often produces different results even with the same training sets (Huang *et al.*, 2002). This can be attributed to the way the classifiers partition the feature space. For example, parametric classifiers like MLC are based on an assumed parametric model and, therefore, requires, a large training sample for wider coverage to ensure that the statistical parameters are able to describe the classes. However, non-parametric classifiers like decision tree (DT) and artificial neural networks (ANN) are not based on any parametric model but use the training data directly for training. Foody (1999) has shown that with MLP neural network, the training samples that lie at the edge of class distribution in feature space are most informative for an accurate classification than those that lie away in the feature space. This indicates that some training samples are more useful than others.

The objective in classification is to get as accurate a map as possible using, if possible, a small number of training sets to make the classification process economical. An optimal classifier would be one that generalizes accurately on unseen cases as compared to

other classifiers and at the same time needs a small training set. SVM is potentially one such classifier.

An SVM classification aims to fit an optimal separating hyperplane (OSH) between classes by focusing on the training samples that lie at the edge of the class distributions, the support vectors. The OSH is a hyperplane oriented in feature space such that it is placed at maximum distance between the two classes. It is because of this orientation that SVM is expected to generalize more accurately on unseen cases as compared to classifiers that aim to minimize the training error such as neural networks. Thus for a SVM, the training data are not equally informative and those lying near the hyperplanes are most informative for SVM classification.

Sample size or number of data points within a sample is not simply a matter of “bigger the better” (Mather, 1999). Every data has a cost attached to it. With a SVM, only the training samples that lie at the edge of the class distributions in feature space (support vectors) are relevant in the establishment of the OSH. Data other than support vectors can effectively be discarded without compromising the accuracy of the classification.

The main aim of the research reported in this thesis was to reduce the training data requirements by exploiting the potential of SVM that of using only training samples that are potential support vectors. The research reported first investigates the effect of training set size on classification accuracy using discriminant analysis (DA), ANN, DT and SVM. The thesis then focuses on means to enhance classification studies by intelligent training site selection. Here attention was focused especially on SVM classifier and was based on the hypothesis that if there is prior knowledge or ancillary information that can be used to identify/locate training sites to regions from which the most informative training samples, the support vectors can be derived, it may be possible to acquire a small intelligently selected training set that can be used to accurately classify the data. In addition, the research focuses on means to reduce the training data requirements if the same classification analyses are repeated in future. This was with the understanding that the



knowledge gained about the relationship of support vectors derived from SVM classification with ancillary information can be exploited in case the analysis is repeated in future to focus the training data acquisition process to the regions most likely to furnish support vectors.

Thus the research essentially explores the means for reducing the training data requirements. A procedure to reduce training set size requirements of SVM is outlined and tested in this research.

## **1.2 Thesis Overview**

Chapter 2 reviews the literature on different classification techniques, with special regard to supervised classification that are used in remote sensing. The chapter focuses on training data issues, details of supervised classifiers used in the research reported later and accuracy assessment.

Chapter 3 reports on the effect of training set size on classification accuracy using DA, ANN, DT and SVM classifiers. The results show that in general the classification accuracy was positively related with training set size. The SVM used in the analyses was more accurate as compared to other classifiers in most of the cases and used only a fraction of training data called as support vectors.

Chapter 4 reports on the procedure to exploit the potential of SVM to reduce the training data requirements. It is shown that with information on soil type, training sample acquisition can be focused to regions most likely to furnish support vectors.

Chapter 5 details the procedure for the acquisition of a small intelligently selected training data that provided appropriate support vectors for an SVM classification directly from field. The intelligent scheme was compared against a conventional scheme in which a large training set was acquired. It is also shown that with ancillary information on soil and water status of the training sites, training sample acquisition can be focused to regions most likely to furnish support vectors. The work also shows that for accurately mapping

only one class from the many land cover classes available in the study area, training data for all the land cover classes are not required.

Chapter 6 discusses the conclusions that arise from the research detailed in chapters 3, 4 and 5.

# CHAPTER 2 - Literature Review

## 2.1 Introduction

Land cover affects our climate by influencing energy, water and gas exchanges with the atmosphere and through acting as a source and sink in biogeochemical cycles (Betts *et al.*, 1996). Accurate information on land cover is, therefore, required to aid the understanding and management of the environment. The term land cover refers to natural entities like vegetation, water bodies, rock/soil, whereas land use refers to the use of land by human beings. The terms, land cover and land use are both closely related and, therefore, are commonly used interchangeably (Campbell, 2002).

Information on land cover is central to scientific studies that link many parts of the human and physical environments. Accurate and up-to-date information on land cover is required for a plethora of applications, including land resource planning, studies of environmental change and biodiversity conservation. The researchers often want land cover data in map form (Marcal *et al.*, 2005).

Land cover maps are generally not readily available or are difficult to acquire (DeFries and Townshend, 1994; Foody, 2002). Even if the maps are available, they are often outdated and in need for updating due to change of time. The frequency of updating required, however, depends upon the land cover categories under consideration. For example, a general map of global land cover may be required every ten years but a crop map may be required on annual basis.

A thematic land cover map can be generated by conventional (ground surveys) or by remote sensing techniques. The conventional ground techniques are time consuming, laborious, and expensive. However, remote sensing techniques are not only quick but also may be used for inaccessible areas.

Remote sensing has been used worldwide in a number of land cover projects, especially for crop inventory. For example, large area crop inventory experiment (LACIE) (Pinter *et al.*, 2003; Olthof *et al.*, 2005), Monitoring agriculture with remote sensing (MARS) program (Gallego, 1999) and Crop acreage and production estimation (CAPE) (Navalgund *et al.*, 1991).

Initially remote sensing was limited to the use of aerial photographs taken from balloons or cameras onboard an aeroplane. However, the launch of Landsat satellite in 1972 brought a new milestone in the development of remote sensing (Campbell, 2002). Remote sensing satellite sensors provided repetitive and systematic observations of Earth's surface. This led to ready acceptance of satellite remote sensing data.

A thematic map can be generated using remote sensing data by a process called as image classification (Foody, 2004; Tatem *et al.*, 2004). Remote sensing can provide multi-spectral, multi-spatial, multi-temporal data useful for land cover mapping by both visual interpretation and quantitative (digital) techniques.

The advancement in computer technology and the availability of inexpensive computer hardware and software (Townshend and Justice, 1981; Mather, 1999) has brought to fore the development in digital remote sensing. The ability to hold and manipulate voluminous data consistently and in a format ready for integration in geographical information systems (GIS) for geographic analysis has resulted in extensive use of digital image classification techniques for remote sensing data (Campbell, 2002).

However, it is not yet possible to map land cover accurately from remotely sensed data. There are many reasons for this and one key issue is the errors inherent in the remotely sensed data. The data, therefore, needs to be pre-processed before the classification can be undertaken.

## **2.2 Preprocessing**

The raw remote sensing data contains error in geometry and in the measured spectral response or brightness values of pixels (Richards and Jia, 1998). The errors in the brightness are called as radiometric errors, whereas errors in the image geometry are called as geometric errors. The operations that are carried on the raw image before the main analysis (*e.g.*, classification) are called as preprocessing. Preprocessing is essentially carried to remove unwanted radiometric and geometric distortions to the data that may otherwise impact negatively on later analyses. Key preprocessing operations are feature reduction, radiometric correction and geometric correction.

### **2.2.1 Feature Reduction**

There are two approaches to feature reduction, feature selection and feature extraction.

The objective of feature selection is to identify spectral bands out of all the available bands that contain the most important information, almost as much as all the bands put together. Generally the discarded data contains noise and errors present in the original data (Campbell, 2002). Thus feature selection reduces the number of bands which in turn reduces the cost of analysis (Melgani and Bruzzone, 2004).

Often the data acquired in some bands are redundant as they are strongly correlated with data in other bands. Correlation between bands helps to identify such redundant bands. High correlation between pair of bands reflects that the two bands are closely related and only one band be retained for the analysis with the other removed. This removal may have only a minor loss of information but gives the analyst advantages that will be apparent later.

Reduction in bands can also be achieved by a process called as feature extraction. In feature extraction, the spectral space is altered unlike feature selection. Principal component analysis (PCA) is one of the important techniques of feature extraction (Han *et al.*, 2004).

The process decorrelates the data by transforming the spectral response around sets of new multi-spaced axes. The process generates a number of principal components that is equal to the original number of input bands. The first principal component explains the maximum information. The first few principal components should be chosen as they carry the most information. Thus a reduced dataset is obtained, which in turn reduces the time required for analysis.

### **2.2.2 Radiometric Preprocessing**

Radiometric preprocessing influences the spectral response (*e.g.*, brightness value or DN) of pixels by removing undesirable influences such as those associated with atmospheric interference, system noise and sensor motion.

The brightness recorded by the sensor is generally a result of reflectance from Earth's surface and that by atmospheric scattering. The undesirable atmospheric component can be removed through an atmospheric correction such as dark object subtraction (Chavez, 1988). The method involves identifying very dark objects (*e.g.*, very deep water) in the image. The spectral response of such a dark object should be zero or nearly zero but if it has some brightness value, it can be attributed basically to the effects of atmospheric scattering. The brightness value of a dark object should, therefore, be subtracted from each pixel on that band to reduce atmospheric effects.

### **2.2.3 Geometric Correction**

The transformation of a remotely sensed image so that the resulting image has the projection properties of a map is called as geometric correction (Mather, 1999). The

remotely sensed images undergo geometric distortions due to a number of factors like rotation and curvature of the Earth, wide field of view of some sensors *etc.*, (Richards and Jia, 1998). The image is usually geometrically corrected in case the information has to be integrated with other map data.

The geometric correction can be achieved by establishing a mathematical relationship between the locations of pixels of an image with their corresponding location in reality (ground). This can be achieved by making use of a map, which can provide ground control points (GCPs). The GCPs are features that can be identified both on image and map and thus helps in deriving the mathematical relationship between their location on image and on Earth.

### **2.3 Classification**

According to the Chambers twentieth century dictionary of the English language (Mather, 1999), classification is defined as the act of forming into classes. Classification in remote sensing terminology can be defined as the process of assigning pixels or other defined spatial unit of an image to various categories to which they belong. In the context of this thesis, the focus of study is pixel as a unit. Classification is one of the most often used methods for information extraction in remote sensing. The classification algorithm uses the spectral response of various features to determine class membership.

The classification can be either hard or soft (Foody, 2002). A hard classifier assumes that pixels are pure that is they are composed of one cover type only. In reality, the pixels may not represent only one class but more than one class especially at the edges of two classes (*e.g.*, pixels lying on the border of built-up and agriculture). The problem of mixed pixel is, therefore, most pronounced with coarse spatial resolution data (*e.g.*, 1.1 km. Advanced Very High Resolution Radiometer (AVHRR) data) (Atkinson *et al.*, 1997). In a soft classification, each pixel is allowed to belong to more than one class (Foody, 2002).

The present study is limited only to hard classification and, therefore, the discussion would focus on hard classification.

The process of digital classification can be performed using either unsupervised or supervised approach.

### **2.3.1 Unsupervised Classification**

Unsupervised classification is a process by which pixels in an image are assigned to spectral classes without the user having the fore knowledge of the existence or names of those classes (Richards and Jia, 1998; Boles *et al.*, 2004).

With an unsupervised classification technique, the classification algorithm groups the pixel data into different spectral classes using a clustering method (Duda and Canty, 2002, Han *et al.*, 2004). The user specifies the number of clusters based on his experience of the study area covered by the image. The analyst then assigns these spectral classes into information classes, based on ground data/reference data. ISODATA and AMOEBA are some of the well known unsupervised classification algorithms.

Since the classes are not selected beforehand, this method is called unsupervised classification. The unsupervised classification is undertaken if ground data/reference data are not sufficient or the analyst is not sure whether the classes proposed can be spectrally discriminated and is , therefore, not the focus of this thesis.

### **2.3.2 Supervised Classification**

Supervised classification is more closely controlled by the analyst than unsupervised classification. The supervised classification requires more input by the analyst as compared to unsupervised classification. The general procedure is to identify homogeneous, representative samples of classes of interest called training areas. The training areas are usually selected by consulting maps, aerial photographs or from field visits. Infact, the selection of training data is more of an art than a science (Lillesand *et al.*,



2004), which requires close interaction between the image analyst and the data to be used. The analyst then demarcates these training sites on the digital image using interactive graphics device such as light pen, joystick, mouse *etc.*, which helps to extract spectral information of the classes in all the bands for the pixels comprising the training sites. The spectral information, thus generated are used to train the classification algorithm to recognize spectrally similar areas for each class using algorithms (of which there are several variations) specifically tailored for classification purpose. Since the analyst guided the learning process, the procedure is called as “supervised classification”. Infact, supervised classification procedure is most often used for quantitative analysis of remote sensing image data (Arora and Foody, 1997; Richards, 1996).

In general, supervised classification algorithms lie within one of the two types, parametric and non-parametric (Emrahoglu, 2003).

### **2.3.2.1 Parametric Classifiers**

Parametric classifiers assume that the training data obtained for each class in each band follows some distribution, usually a normal Gaussian distribution. For example, the maximum-likelihood classifier (MLC) is a parametric classifier assuming normal distribution of training data. The MLC is the most popular statistical algorithm and is widely accepted as a standard approach (Emrahoglu, 2003).

### **2.3.2.2 Non-parametric Classifiers**

A non-parametric classifier does not assume any distribution for the training data (*i.e.*, it is distribution-free) (Kavzoglu and Mather, 2003). Commonly used non-parametric classifiers are ANN, DT, and SVM.

### **2.3.3 Stages in Supervised Classification**

Supervised classification whether parametric or non-parametric has three broad stages (Figure 2.1). The first stage is the training stage in which the pixels or other defined

areas of known class membership are identified on the image called as training areas. The spectral response of the training areas may be used to generate statistics depending upon the requirements of the classifier to be used. For example, minimum distance classifier requires mean of the classes in all the bands used. However, some classifiers like the ANN and DT does not require any statistical parameters but use the spectral response of the training areas directly to train the classifier. In the second stage, the trained classifier classifies the image into various information classes. Thirdly, the accuracy of the classification is evaluated in the testing stage. The accuracy should ideally be tested for a dataset not used in the training stage of classification.

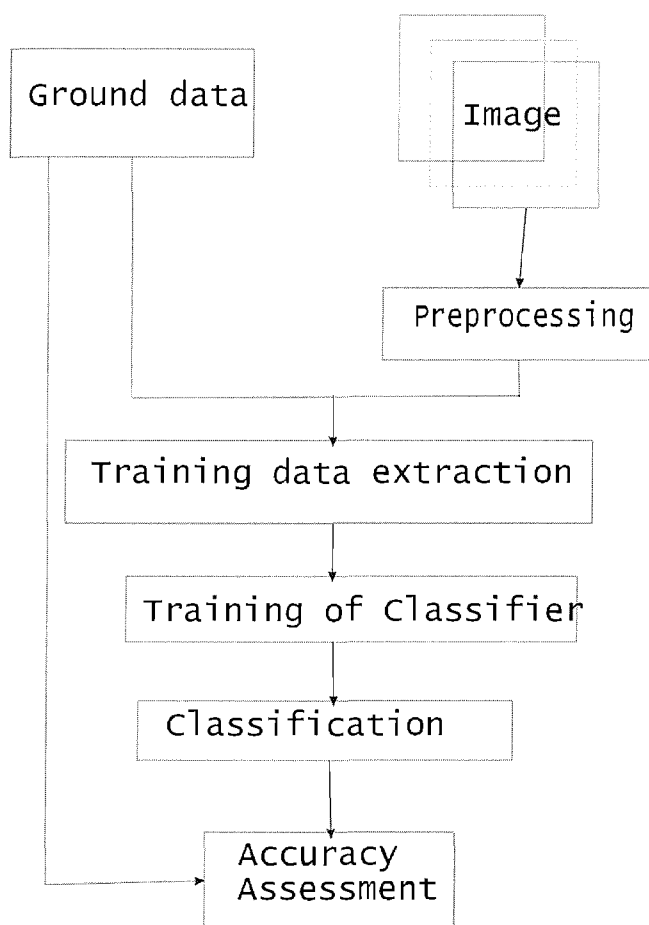


Figure 2.1: Stages in supervised classification.

### **2.3.3.1 Training Stage**

Remote sensing provides view of large areas of Earth's surface at one time and help extract useful information at a very attractive cost-benefit ratio. It does not however mean "absence from field" (Anonymous, "Ground Truthing", 2002, <http://www.ecoman.une.edu.au//BRMSAL/chapter.htm>). The ground data has to be collected to relate to digital image.

Ground data typically, refers to any reference data or ancillary information used in support of the analysis. Ground data play an important role in the training stage of the classification to identify areas of known ground identity. The training data can be collected by field visits or by ancillary means, if there are constraints of time, labour and money provided the source is current.

The training data collected should describe the classes under investigation. The quality of training data can, therefore, significantly influence the classification accuracy (Hixson *et al.*, 1980). Inappropriate placement or too few pixels in training site produces statistics which may not be able to characterize the land cover classes. Studies have shown that different training strategies can result in very different accuracy estimates for the final classification (Fitzpatrick-Lins, 1981; Congalton, 1988; Gong and Howarth, 1990). Therefore, several researchers have emphasized the importance of the training stage so that the classifier generalizes well for unseen cases (Hixson *et al.*, 1980; Foody, 1999).

#### **2.3.3.1.1 Design of Training Strategy**

The objective in designing a training strategy should be such that the training samples are able to describe the classes under investigation. Foody (1999) shows that the border training pixels (pixels lying away from mean of the spectral distribution of a category) classify unseen cases accurately as compared to the core pixels (pixels concentrated near the mean value of spectral distribution of the categories). Hence some

intelligent way of planning ground data are needed to collect only border training pixels for classes under consideration.

A range of factors, however, should be considered while designing a training strategy, such as:

1. Time of sampling
2. Sample type
3. Sample size
4. Sample design

#### **2.3.3.1.1.1 Time of Sampling**

Generally, the samples should be recorded at the time of remote sensing data acquisition. This is particularly the case if mapping rapidly changing phenomenon like soil moisture, erosion, and land cover resulting from floods or fire. But it is not always possible to collect the data at the time of satellite sensor overpass. Thus sample data may be collected at least for the same season as the satellite sensor data, even if the dates are different (Justice and Townshend, 1981) for categories which undergo seasonal changes like forest.

#### **2.3.3.1.1.2 Sample Type**

When planning a project involving remote sensing data, a classification scheme must be finalized in the beginning to fulfill the very objectives of the project. In many cases, an existing classification scheme such as Anderson classification system (1976) for United States Geological Survey (Lillisand and Kiefer *et al.*, 2004; Campbell, 2002) or The International Geosphere-Biosphere programme (IGBP) scheme (Hansen and Reed, 2000) of the U.S Geological survey can be used. Some other classification schemes are based upon it. The benefit of using an existing scheme is that the study can be compared with other similar projects completed using the same standard scheme. The classification

scheme chosen should ensure that the classes are mutually exclusive and defined exhaustively that is any particular parcel of land should fall into one category only.

The sampling scheme must be clearly understood at the beginning of the work, otherwise there is bound to be a great loss of time and much frustration at the end of the project (Congalton, 1991).

### 2.3.3.1.1.3 Number of Training Samples

To ensure that the sample data provide a representative statement of the spatial population, the sample size must be chosen with care (Curran and Williamson, 1985). However, sample size or number of data points within a sample is not simply a matter of “bigger the better” (Mather, 1999). The cost involved in collecting the training data is also an important factor. Each sample collected has a cost attached to it and thus it is very important that the sample size of training data should be kept to a minimum. The requirement for large sample sizes is however not unusual and penalties for collecting less in some cases can be severe (Curran and Williamson, 1985).

A number of relationships have been suggested to ascertain the minimum number of training samples. For example, Fitzpatrick-Lins (1981) argue that the number of samples ( $n$ ) can be computed by the following relationship:

$$n = \frac{Z^2 * A * Q}{E^2} \quad (2.1)$$

where:

Z = Z score and represents the number of standard deviations a data value falls above or below the mean of a normal distribution.

A = Expected accuracy (%)

Q = 100-A

E = allowable error

Hay (1979) suggested that as a general rule 50 sample points (pixels) per-category are required. Mather (1999) suggested that at least  $30n$  pixels per class should be selected, where  $n$  is the number of bands. Lillesand *et al.*, (2004) stated that a minimum of  $10n$  to  $100n$  pixels should be used for each class, where  $n$  is the number of spectral bands.

Congalton (1991) suggested that as a thumb rule minimum of 50 samples for each land cover category be chosen and if the area is very large (more than a million acres) or if there are large number of categories (more than 12 categories) minimum number of samples should be increased to 75. He further stressed that number of samples can be adjusted based on the spectral variability of the categories in the study (*e.g.*, fewer samples can be taken for water as it shows little spectral variability).

In general, the number of training samples should be sufficient to capture the spectral variability of the categories so that the classifier is able to generalize well for unseen cases.

#### **2.3.3.1.1.4 Sample Design**

The location of training areas of each class must be well distributed over the study area; otherwise, the training data would be biased and unrepresentative, thereby affecting the accuracy of classification, on unseen cases.

There are a number of ways in which an area can be sampled in a two dimensional space. These include simple random, stratified random and systematic random sampling (Figure 2.2). In unaligned sampling, each point is chosen randomly, that is both  $x$  and  $y$  coordinates of a point is chosen randomly, whereas in aligned sampling one of the two coordinates is fixed and the other is chosen randomly. Simple random sampling is one, where every distinct sample has an equal chance of being drawn. The scheme has a drawback that it may under sample or may not sample categories, which cover very small area. This drawback can be nullified by using stratified sampling or systematic sampling. In stratified sampling scheme, the area is divided into strata. These stratas should not

overlap and together should constitute the whole area under study. In systematic sampling, the area is divided into sections, usually rectangular or square in shape.

The above sampling designs are single stage but there are two stage sampling design also. In the two stage sampling, the area is divided into strata as in stratified sampling. Sampling is then restricted to only a limited number of randomly selected sub-areas called as primary units. The advantage of the sampling design lies in the fact, that the time to travel between the samples is reduced, as the samples are concentrated in a small area. However, the disadvantage is that the samples may not be representative of the area and, therefore, they may fail to capture the spectral variability of the categories under consideration. However, the number of samples can be increased because of the operational advantage of less travel.

With cluster sampling, a group of pixels are selected unlike individual pixels in simple random sampling, stratified random sampling *etc.*, as discussed above. However large clusters should be avoided as adjoining pixels may add very little information due to autocorrelation (Congalton, 1991). Cluster sampling is less costly as compared to other sampling designs discussed above.

However, an ideal sample distribution should derive non-autocorrelated (spatially independent) training data (Atkinson, 1991; Chen and Stow, 2002). Spatial dependence can be summarized as the expectation that observations close together are more likely to be similar than observations further apart. An estimate of spatial autocorrelation can be made via the semi-variogram (Curran and Williamson, 1985). The semi-variogram is a graph (Figure 2.3) of semi-variance of values given for pixels separated by different distances. The semi-variance represents the average of the squared difference in values separated by a specific lag distance. Semi-variogram  $\gamma(g)$  is given by:

$$\gamma(g) = \frac{\sum_{i=1}^a (x(i) - x(i + g))^2}{2a} \quad (2.4)$$

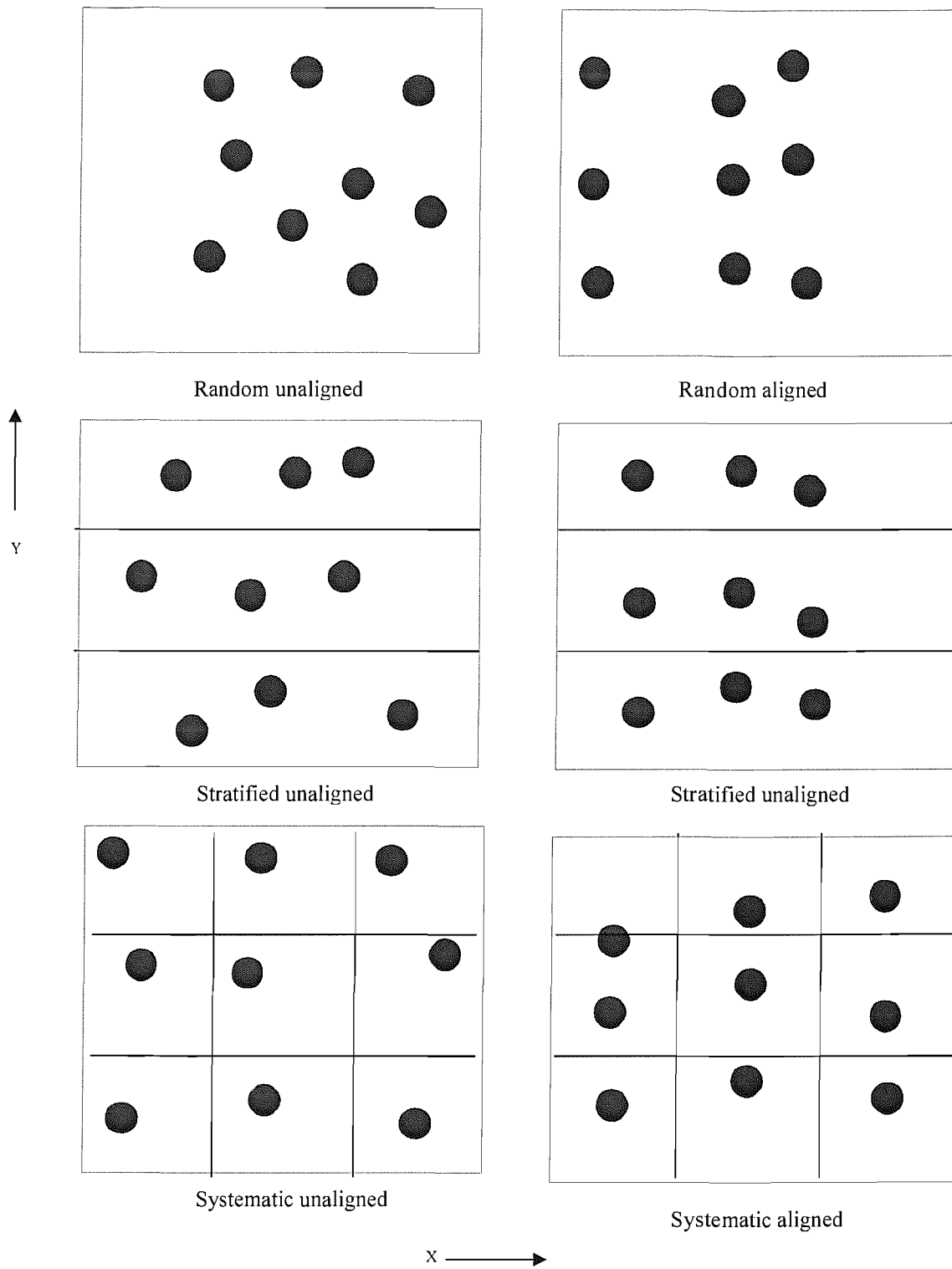


Figure 2.2: Sampling techniques.





Figure 2.3: Semi-variogram.

where,  $a$  is the number of pixel pairs separated by a distance  $g$ , and  $x(i)$  and  $x(i + g)$  are pixel values at  $i$  and  $i + g$  respectively. As distance between pixels becomes larger, the difference in pixel values between pixel pairs generally becomes larger, and at some distance, the semi-variogram develops a flat region which is the limit of auto-correlation that is it indicates the distance over which the values sampled are similar.

Investigators have reported that the use of variogram resulted in a 3.5 to 9 fold reduction in sample size for a given level of error (McBratney and Webster, 1983). However to take advantage of such reductions in sample effort, an investigator needs to pre-sample in order to construct the semi-variogram (Curran and Williamson, 1986).

### 2.3.3.2 Allocation

In the allocation stage, the trained classifier classifies the image into various categories. Most attention has, however, been focused on the allocation phase of the classification particularly with regard to the development of new algorithms to increase the classification accuracy. A number of algorithms, both parametric and non-parametric have been developed over the years in remote sensing. A number of different classifiers having very different approaches used in the present research are discussed briefly hereafter with details following in section 2.4 to 2.7.

MLC is the most commonly used classifier. MLC is a parametric classifier based on statistical theory. It assumes that the distribution of the spectral classes can be described by Gaussian normal probability distribution. Generally, the spectral distribution of categories is not normally distributed, and in that case, it is statistically invalid to apply MLC algorithm.

The limitation of the parametric classifiers, and following advances in computer technology, alternative non-parametric classification like ANN, DT and SVM have been developed.

One of the alternatives to the statistical classifier is the ANN, particularly the feed-forward multi-layer perceptrons using back-propagation (Kanellopoulos and Wilkinson, 1997; Liu and Wu, 2005). There has been an explosion of interest in ANN by the remote sensing community. This includes their widespread use in supervised classification (Foody, 1999). One of the main advantages of the ANN for classification is that they are distribution-free, that is, no underlying model is assumed for the classes in the training data. The training phase in a neural network, unlike MLC, is not a one time calculation of statistical measures, but is an iterative process with the intention to achieve minimal error between the desired output (determined from training data) and actual output values of the network. The trained network may then be used to allocate pixels to classes based on the response at the output stage. ANN are very attractive for the classification of large data

sets because networks once trained provides very rapid processing (Foody and Arora, 1997).

A decision tree recursively partitions the data into smaller subsets on the basis of tests or thresholds applied at each node of the tree. Decision trees are non-parametric classifier that are computationally very efficient, unlike ANN, which take very long to undertake thousands of iterations before the network stabilizes and is ready for final classification. Feature reduction is inherent in the decision tree, which reduces the requirement of larger training data. The decision tree can be easily interpreted by the analyst as compared to ANN.

Support vector machines (SVM) are relatively new to the remote sensing community as compared to other classifiers like MLC, ANN or the DT. Key attraction of SVM based approach is that it seeks to fit an optimal hyperplane between classes. The optimal hyperplane is so chosen that it maximizes the margin between the categories and, therefore, should generalize well to unseen cases with least errors amongst all possible boundaries, separating the classes.

### **2.3.3.3 Accuracy Assessment**

After classification, the thematic map produced has to be checked for accuracy, as a key concern is that the land cover maps derived are often judged to be of insufficient quality for operational applications (Foody, 2002). The exercise is necessary for both users and producers of the maps to understand the utility of the map so produced. The producer would like to identify and correct the errors so as to increase the information content of the thematic map. The user on the other hand requires accuracy information of individual class or a number of classes to understand the suitability for a particular purpose (*e.g.*, a user may be interested only in wheat category out of all the categories represented by a thematic map). This inference is based on the comparison of the derived land cover map

with ground or other reference data. The accuracy of the classification refers to the degree to which the derived image classification agrees with reality or conforms to the truth (Jensen and Van der Wel, 1994; Smits *et al.*, 1999; Campbell, 2002; Foody, 2002).

Ideally, accuracy assessment should consider the entire population of pixel pairs from reference and classified images (Campbell, 2002) but in practice it is seldom possible. The collection of complete reference data may not be feasible because of time, labour and cost constraints or because of the inaccessibility of the area. As such, the accuracy assessment is usually conducted on a sample of ground/reference data.

#### **2.3.3.3.1 Design of Sampling for Reference Data Acquisition**

After classification, accuracy of classification should be tested on a data set independent to that used in training the classifier. In remote sensing applications, for accuracy assessment, reference data should be collected at the same time as that of training sample collection (Richards and Jia, 1998).

The data for testing accuracy can be collected after the remote sensing data has been received and classified. However, collecting data for testing accuracy of classification independent of training data is not economical. Further, some of the categories present at the time of data acquisition may not be available because of the time delay between classifying the data and visiting the field. For example, crops like wheat may be harvested by the time remote sensing data are acquired; analyzed and field visit is made.

The testing pixels can be randomly located on the thematic map after the remote sensing data has been classified. The problem is that minor categories (categories occupying a small area on thematic map) may be under-sampled or may not be sampled at all. Other sampling schemes as detailed in section 2.3.3.1.1.4 can also be followed. However, stratified random approach has the advantage in the sense that if the thematic map can be divided into strata based on categories, then all the categories can be represented in the accuracy assessment.

### 2.3.3.3.2 Number of Testing Data Samples

The testing data can be collected along with the training data or based on the thematic map generated as a result of classification. In case, the testing data are collected at the same time as training data, then testing data can be akin to the training data *i.e.*, equal in number to the training data as discussed in section 2.3.3.1.1.3.

However, the number of testing data can also be calculated based on the thematic map. The sample size can be calculated based on binomial statistics. If the accuracy of a class is  $\theta$ , then probability  $P$  that  $s$  pixels in a sample of  $u$  pixels belongs to that class is given by the binomial distribution (Schowengerdt, 1983);

$$P(s : u) = {}^u C_s \theta^s (1 - \theta)^{u-s} \quad s = 0, 1, \dots, u \quad (2.2)$$

Van Genderen *et al.*, (1977) presumed that if the sample is too small then there is a fair chance that all the pixels selected are correctly labelled, that is  $s=u$ . For example, if only one pixel is considered for a category, the accuracy may be 100 % if it is correctly classified (Richards and Jia, 1998). Then,

$$P(u : u) = \theta^u \quad (2.3)$$

Van Genderen *et al.*, (1977) evaluated the above expression as tabulated in Table 2.1.

Classification accuracy (%)	Sample size
0.95	60
0.90	30
0.85	20
0.80	15
0.60	10
0.50	5

Table 2.1: Minimum sample size necessary per category (after Van Genderen *et al.*, 1977).

### 2.3.3.3.3 Error Matrix

The error matrix also known as confusion matrix or contingency table is the most widely used means of accuracy assessment in remote sensing. It is a simple cross-tabulation of the mapped class label against those observed in the ground or in the reference data for a sample of cases at specified locations (Canters, 1997; Campbell, 2002; Foody, 2002). The confusion matrix provides the means to describe both the classification accuracy and confusion between the classes.

The columns of the error matrix may represent the correct (reference) data and rows, the classified as generated from remotely sensed data or vice-versa. The diagonal elements represent the correct classification (*i.e.*, classification is in agreement with ground/reference data). The non-diagonal elements represent the error in the classification.

The error matrix (Table 2.2) can be used to derive a number of metrics of classification accuracy, the most popular is the percentage of overall (sum total of all the classes) cases correctly allocated in the classification and is a measure of overall accuracy of classification. Likewise accuracy can also be calculated for individual categories, by comparing their correct allocation to the total number of cases in the respective categories.

For individual categories, the total number of correctly classified pixels can be divided either by total number of pixels in the corresponding row or by corresponding column. The total number of pixels of a category is divided by the total number of pixels of that category as derived from the reference data (Table 2.2). This accuracy is known as producer's accuracy as the producer of the map is more interested to understand how well a certain area on ground (reference data) can be classified (Congalton, 1991). However, if the total number of correct pixels in a category is divided by the total number of pixels classified in that category by the classifier (row total), the accuracy is known as user's accuracy. This reflects that a pixel classified on the map is in agreement with the ground.

The off diagonal elements in the error matrix are used to generate error of omission and commission. The error matrix must be diagnosed for non-diagonal elements to

		Actual class				
		A	B	C	D	$\Sigma$
Predicted class	A	$n_{AA}$	$n_{AB}$	$n_{AC}$	$n_{AD}$	$n_{A+}$
	B	$n_{BA}$	$n_{BB}$	$n_{BC}$	$n_{BD}$	$n_{B+}$
	C	$n_{CA}$	$n_{BC}$	$n_{CC}$	$n_{CD}$	$n_{C+}$
	D	$n_{DA}$	$n_{DB}$	$n_{DC}$	$n_{DD}$	$n_{D+}$
	$\Sigma$	$n_{+A}$	$n_{+B}$	$n_{+C}$	$n_{+D}$	$n$

$$\text{percentage correct} = \frac{\sum_{k=1}^q n_{kk}}{n} \times 100$$

$$\text{user's accuracy} = \frac{n_{ii}}{n_{i+}} \times 100$$

$$\text{producer's accuracy} = \frac{n_{ii}}{n_{+i}} \times 100$$

$$\text{kappa coefficient} = \frac{n \sum_{k=1}^q n_{kk} - \sum_{k=1}^q n_{k+} n_{+k}}{n^2 - \sum_{k=1}^q n_{k+} n_{+k}}$$

Table 2.2: Error matrix of a classification.

understand the interclass confusion. Error of omission (exclusion) refers to not assigning to correct class. It is 100-producer accuracy. Error of commission (inclusion) on the other hand, refers to assigning to incorrect class. It is 100-users accuracy.

The equation of various metrics extracted from confusion matrix depends upon the sampling scheme followed in collecting the sample data. The error matrix detailed in Table 2.2 corresponds to simple random sampling. If, however, stratified random sampling is followed, then calculations of various metrics are based on stratas employed by the sampling.

A major problem in classification accuracy is that some cases may be allocated to the correct class by chance (Hord and Brooner, 1976; Rosenfield and Fitzpatrick-Lins,

1986; Congalton, 1991; Pontiffs, 2000). The kappa coefficient of agreement,  $\kappa$  (Table 2.2) compensates for such a chance agreement. It can be defined (Campbell, 2002) as;

$$\kappa = \frac{\text{observed} - \text{expected}}{1 - \text{expected}} \quad (2.5)$$

Here, observed designates, the overall accuracy reported in the error matrix and expected refers to the correct classification that can be anticipated by chance agreement of both the reference and the remote sensing data. Kappa ( $\kappa$ ) has a range from 0 to 1. A value of 1 suggests perfect effectiveness of the classification and a value 0 suggests that the contribution of chance is equal to the effect of correct classification (Campbell, 2002).

The kappa coefficient has been widely used in remote sensing for comparison of classification accuracy even though the approach may be inappropriate (Foody, 2004) because the assumption of independence of samples is violated. For instance, comparative studies using different classification algorithms (*e.g.*, ANN and SVM) generally use the same ground/reference data in assessing the accuracy of classification by each of the classifier. As such, the assumption of independence of samples is violated.

#### **2.3.3.3.1 Comparison of Error Matrices**

Error matrices permit comparison of different classifications due to analysis carried for data, acquired at different dates or classified by different algorithms. In such instances, comparison of error matrices provide means of judicious selection of factors that provide the highest accuracy of classification (*e.g.*, digital data with its date of acquisition, issues related with training data properties, choice of algorithms).

The direct comparison of matrices can be difficult in case of differing number of observations. Normalization of error matrix can be undertaken to circumvent such problem. Normalization may be achieved through an iterative procedure that brings the row and column sums to unity.



Examination of the normalized matrices affords very convenient comparison. However, in some instances normalized values are so small that they are neglected with the notion that they do not alter the interpretation. It is also impossible to derive original matrix from normalized version and, therefore, should not be attempted.

### **2.3.4 Problems in Land Cover Classification**

There has been considerable developments made recently in land cover classifications, but the accuracy with which thematic maps may be derived from remotely sensed data are, however, often still judged to be too low for operational use (Foody, 2002). Typically, the reasons for accurate land cover mapping include; characteristics of remote sensing data, the nature of the classes and the methods used in mapping (Foody, 1999).

#### **2.3.4.1 Issues Related to Characteristics of Remote Sensing Data**

The digital data provided by remote sensing is raster or grid based. The smallest element of a digital image, the pixel is, therefore, not a point entity but represents an area on the Earth's surface described by the spatial resolution or instantaneous field of view of the sensor. The spectral response of a pixel is, therefore, not representative of any point on Earth's surface but is an average of spectral response over the area described by the pixel on the Earth. The digital classifications are based on these values with the understanding that they are faithful representation of the Earth's entity.

The characteristics of remote sensors are, however, far from ideal (Cracknell, 1998). One of the drawbacks is that the spectral response of a pixel (*e.g.*, reflectance) is not only contributed by land cover inside the pixel, but also by adjoining pixels. Another drawback is that the sensors are centre biased such that the reflectance towards the centre of the pixel has the most influence on the reflectance value of the pixel. In other words, the reflectance information contained in a pixel tends to be more similar to the reflectance of

land cover located towards the centre of the pixel's ground area and least similar to cover, towards its edge. This effect of reflectance of one part of the field-of-view on the value recorded is not well understood, especially if the reflectance of that part is very different from the remainder (Fisher, 1997).

The problem in recording the spectral response can thus lead to the analyst observing the same reflectance for very different mixtures of sub-pixel classes. Such differences cannot be distinguished by the analyst. The contribution depends basically upon the area of the pixel covered by the categories. Hence, the representation of digital numbers of a pixel is truly speaking, not very faithful of the area represented by it on the ground.

The available sensor provides data in many different spatial resolutions, for example 1m (IKONOS), 1 km (NOAA AVHRR). Each pixel is commonly classified as belonging to only one land cover class with the assumption that land cover fits exactly into multiples of rectangular spatial units, and that such small areas are homogeneous up to the coarsest pixel (Fisher, 1997). Typically, the Earth does not fit into the concept of pure elemental squares of even the finest pixel. However, the analyst is always faced with the problem of extracting information, much smaller than the size of the pixel (*e.g.*, pixels lying on boundaries of two land cover classes). These give rise to mixed pixel problem.

There are other operational problems linked with cloud cover and topography of the area being sensed. The remote sensing data for land cover is not available in optical range under cloudy conditions. The effect of shadows because of obstruction (like mountains, high rise buildings) alters the digital values of the pixel.

#### **2.3.4.2 Nature of the Classes**

There is no guarantee that different land cover classes will have different spectral responses. Indeed, an accurate land cover class map assumes that each land cover class has unique spectral properties. Classes lacking such unique characteristics must either be

combined to give broader classes so that the resulting classes are spectrally unique or they should be separated using ancillary data or by multi-date analysis.

#### **2.3.4.3 Methods used in the Analysis**

There are a number of factors in land cover mapping controlled by the analyst. The accuracy of a classification is determined by a range of factors including the analyst's skill, judgement and familiarity with the study area (Foody, 1999). The factors include issues related to remote sensing data, the ancillary data and the choice of classifier to generate the land cover map.

The choice of sensor and date of image acquisition should be such that the spectral bands are able to discriminate the various land cover classes of interest and the spatial resolution is able to furnish the required details.

The quality of training data can significantly influence the accuracy of classification. The objective in designing the training strategy as discussed under section 2.3.3.1.1.4. should be such that the training samples are able to characterize the classes under investigation. The nature of the testing set can have a significant affect on the resulting accuracy statement (Congalton, 1988). The testing data should be representative of the classes.

The training and the testing set has to be extracted from digital image for training the classifier and subsequent accuracy assessment. This involves the registration of the remote sensing with ancillary data (usually a map), and if the two sets are not accurately co-registered, it will have detrimental effect on both the classification and the accuracy assessment.

At times, the land cover classes are spectrally overlapping, so multi-date analysis can help to segregate the confusing classes by carrying out analysis at different times of their growth period. Such an exercise may be very costly. Therefore, classifiers, which can

discriminate the land cover classes, based on single date remote sensing data, would be an optimum choice.

There are a number of classifiers with very different approaches available with the fact that the most accurate classifier is unknown. The land cover problem should, therefore, be submitted to various classifiers, and the classifier or an ensemble (combination) of classifiers (Steele, 2000) giving the highest accuracy under the given conditions should be chosen to generate the land cover map.

It is, however, unrealistic to suppose that a single optimal method can be devised for all classification tasks in all terrain types (Townshend and Justice, 1981). The analyst must be aware of how the different variables affect the accuracy of classification to enable maximum extraction of information from remotely sensed data. Then the analyst may select an approach, which is appropriate for a particular investigation (Arora and Foody, 1997).

The thesis is focused with the aim to study some of the variables affecting classification accuracy. To study the effect of classifiers on classification accuracy, four classifiers namely MLC, ANN, DT and SVM with very different approaches towards classification were used for analysis and are discussed in detail hereafter.

## **2.4 Maximum-likelihood**

Maximum-likelihood classification (MLC) is the most common supervised classification method used with remote sensing image data (Richards and Jia, 1998; Wang *et al.*, 2004). An important assumption in MLC is that each spectral class can be described by a normal Gaussian probability distribution in multi-spectral space. Such a distribution describes the chance of finding a pixel as belonging to any particular class at any given location in the multi-spectral space. Generally most pixels in a distinct cluster or spectral class would lie towards the centre and would gradually decrease away from the centre, thereby resembling a Gaussian probability distribution.

### 2.4.1 Design of MLC Classifier

The conditional probability that a pixel at a location  $x$  belongs to a class is given by:

$$p(w_i|x), \quad i = 1, \dots, q$$

where,  $q$  is the total number of classes.

The probability  $p(w_i|x)$ , therefore, gives the likelihood that the correct class is  $w_i$  for a pixel at position  $x$ . Classification is performed accordingly if:

$$x \in w_i \text{ if } p(w_i|x) > p(w_j|x) \quad \text{for all } j \neq i \quad (2.6)$$

that is the pixel at  $x$  belongs to class  $w_i$  if its likelihood  $p(w_i|x)$  is the largest amongst all the classes.

However,  $p(w_i|x)$  are unknown. The solution is to estimate a probability distribution  $p(x|w_i)$  for the cover types from training data.  $p(x|w_i)$  describes the chance of finding a pixel from class  $w_i$  at the position  $x$ . There would be as many probabilities  $p(x|w_i)$  as there are ground cover classes. In other words, the set of probabilities  $p(x|w_i)$  would give relative membership of a pixel with respect to all the available classes.

The desired probability  $p(w_i|x)$  and the computed likelihood's  $p(x|w_i)$  from training data are related by Bayes theorem as;

$$p(w_i|x) = p(x|w_i) p(w_i) / p(x) \quad (2.7)$$

where:

$p(w_i)$  is the probability of class  $w_i$  in the image.

$P(x) = \sum_i p(x|w_i) p(w_i)$ , is the probability of finding a pixel from any class

at location  $x$ . The  $p(w_i)$  are called a priori or prior probabilities. The classification rule of equation 2.6, using equation 2.7 becomes:

$$x \in w_i \text{ if } p(x|w_i) p(w_i) > p(x|w_j) p(w_j) \quad (2.8)$$

The rule given by equation 2.8 is more acceptable than that given by equation 2.6, since  $p(x|w_i)$  are known from training data and  $p(w_i)$  are either known or can be estimated, based on analyst experience. For mathematical convenience, taking logarithm of both side of equation 2.8:

$$\begin{aligned} f_i(x) &= \ln\{ p(x|w_i) p(w_i) \} \\ &= \ln p(x|w_i) + \ln p(w_i) \end{aligned} \quad (2.9)$$

Where,  $f_i(x)$  is referred to as discriminant function and  $\ln$  is the natural logarithm.

Equation 2.8 in terms of discriminant functions can be restated as:

$$f_i(x) > f_j(x) \quad \text{for all } j \neq i \quad (2.10)$$

Generally, the probability distribution for the classes are assumed to be multivariate normal, as the properties of such a distribution is well known. This is an assumption, rather than a practical property of natural spectral classes. It is because of this assumption of normality, that MLC is categorized as a parametric classifier. For A bands (Richards and Jia, 1998)

$$p(x|w_i) = (2\pi)^{-A/2} |\Sigma_i|^{-1/2} \exp \left\{ -1/2 (x-m_i)^t \sum_i^{-1} (x-m_i) \right\} \quad (2.11)$$

where,

$m_i$  and  $\sum_i^{-1}$  are the mean and covariance matrix of the data in class  $w_i$ .

#### 2.4.1.1 Thresholding

The output of MLC is a set of likelihood values for each pixel, one likelihood value for each class considered in training the classifier. The pixel is then classified as belonging to the class for which it has the greatest likelihood value. Hence in MLC all the pixels are classified into one of the classes for which the classifier was trained.

In many remote sensing applications, the main operational objective is the identification of a specific land cover class or of a few land cover classes of interest in a

geographical area. Hence pixels of categories not considered in the training stage will have some likelihood values and will be classified, irrespective of how small the likelihood values are, to the class with which it has maximum likelihood. Thus mis-classification can result, if all the classes constituting the land cover are not considered. The problem can be solved by applying threshold to the discriminant functions. The decision rule given by equation 2.10 after incorporation of threshold becomes

$$x \in w_i \text{ if } f_i(x) > f_j(x) \text{ for all } j \neq i \quad (2.12)$$

$$\text{and } f_i(x) > U_i$$

where,

$U_i$  is the threshold specified by the user for all the classes under consideration.

The threshold should be so chosen so that the classes not considered in training the classifier would lie below the threshold value. Pixels belonging to these classes will eventually be rejected.

#### 2.4.2 Limitations of Maximum-likelihood Classifier

MLC is a parametric supervised classifier and, therefore, assumes data of each class in each band to be normally distributed, which may not be the case generally. Hence the data has to be checked for its distribution, usually by viewing the histogram of each class in each band. If the distribution assumption is violated, then it is invalid to represent the class by normal probability function.

Apart from being normally distributed, the training data sample has to be large enough to derive statistics (like mean and covariance). The statistics can be derived from a small sample but the key thing is that they should be representative of the area under study.

The MLC classifier requires the mean and covariance for each class. At least  $(n+1)$  training samples are, therefore, needed for each class (where  $n$  is the number of bands) otherwise variance/covariance matrix will be singular (*i.e.*, its determinant will be zero and

the matrix will not be able to be inverted). This would make it impossible to derive the n-dimensional probability density function (Richards and Jia, 1998).

Typically, the training data should be 10 to 100 times the discriminating variables (*e.g.*, wavebands) (Swain and Davis, 1978). Likewise, the increase in wavebands, as those in imagine spectrometers would increase the demand of training data linearly. Clearly a very large training set is required for mapping from multispectral data sets and this runs contrary to a major goal of remote sensing, which involves extrapolation over large areas from limited ground data (Foody, 1999). Infact, according to Hughes effect, increase in dimensionality of data sets may decrease the classification accuracy in MLC.

## **2.5 Artificial Neural Networks**

The term artificial neural network refers to a network of interconnected neurons. It goes by many names such as connectionist models, parallel distributed processing models. The very development of ANN can be called biologically inspired following the idea using general organization principles found in human brains (Atkinson and Tatnall, 1997). The principles on which the brain works and used in ANN are parallel and distributed processing that is the information is not processed serially and is not stored at one fixed location.

### **2.5.1 Multi-layer Perceptrons**

In remote sensing, multi-layered feed-forward networks are most commonly encountered (Foody, 1999; Zhan *et al.*, 2003; Pal and Mather, 2004). Multi-layer



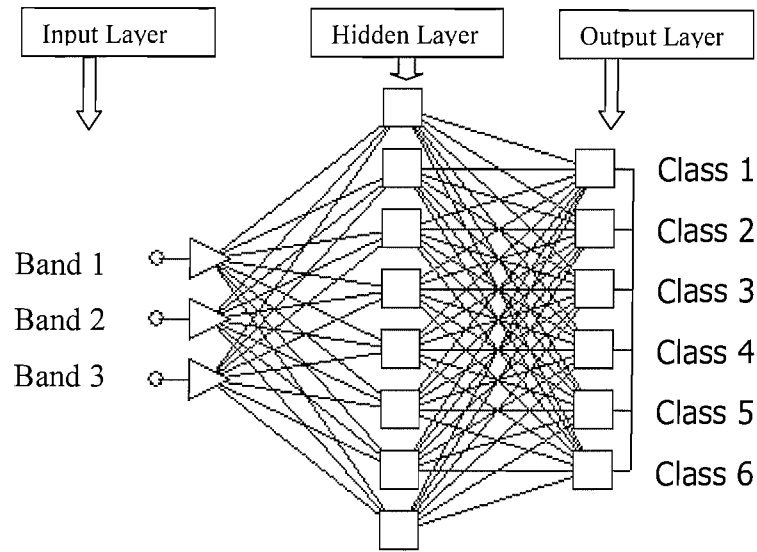


Figure 2.4: Multi-layer perceptron.

perceptrons (MLP) or feed-forward networks have been widely used for supervised image classification in remote sensing (Kanellopoulos and Wilkinson, 1997; Atkinson and Tatnall, 1997; Foody, 1999). MLP are, therefore, used in the present research and discussed in detail here.

MLP consists of a layered structure with three layers of neurons, an input layer, an output layer and a layer in between; called as the hidden layer (Figure 2.4). The input layer comprises of one unit for each discriminating variable (*e.g.*, waveband). There may be one or more hidden layers each containing units as defined by the user. The output layer consists of one unit for each class.

### 2.5.1.1 Training of Multi-layer Perceptron

In MLP, the data is processed in parallel (Figure 2.4). The data in the input layer are multiplied by the weight of the associated interconnecting channel and are summed to derive:

$$net_{pj} = \sum_i w_{ij} o_{pi} \quad (2.19)$$

where:

$w_{ij}$  is the weight from node  $i$  to node  $j$

$o_{pi}$  is the output of unit i for pattern p

This net input is then transformed by the activation function  $f_{net}$ , usually sigmoid function, such as:

$$f_{net} = \frac{1}{1 + e^{-k \cdot net}} \quad (2.20)$$

where, k is a gain parameter, usually set to 1 .

During training, the network is fed with the training data in a feed-forward fashion with weights set randomly initially for the connecting channels. At the output, the network error, which is the difference between the desired (as obtained from training) and actual network output is calculated. This error is then passed backwards through the net towards the input layer and in the process alter the weights connecting the units in the proportion of the error. The process is repeated a number of times till the output error declines to an acceptable level or had stabilized (Foody, 1995).

The network error  $E_p$  can be defined as

$$E_p = \frac{1}{2} \sum_j (t_{pj} - o_{pj})^2 \quad (2.21)$$

where:

$t_{pj}$  = Target output for pattern p on node j

$o_{pj}$  = Actual output for pattern p on node j

Each iteration computes the gradient or change in error ( $\frac{\delta E_p}{\delta w_{ij}}$ ) with respect to each

weight:

$$\frac{\delta E_p}{\delta w_{ij}} = \frac{\delta E_p}{\delta net_{pj}} \times \frac{\delta net_{pj}}{\delta w_{ij}} \quad (2.22)$$

$$\frac{\delta E_p}{\delta w_{ij}} = -\delta_{pj} \times o_{pi} \quad (2.23)$$

Where,  $\frac{\delta E_p}{\delta net_{pj}} = -\delta_{pj}$ , is the change in error as a function of change of net inputs,

$$\frac{\delta net_{pj}}{\delta w_{ij}} = o_{pi}$$

Equation 2.23 shows that decreasing the value of  $E_p$ , changes the weight proportional to  $\delta_{pj} o_{pi}$ , that is:

$$\Delta_p w_{ij} = \eta \delta_{pj} o_{pi} \quad (2.24)$$

where:

$\Delta_p w_{ij}$ , is the change for the weight, which connects, channel  $i$  to  $j$

$\eta$  is a constant, defines the learning rate

The calculation of  $\delta_{pj}$  is different for output and hidden layers, as the desired output is known from the training data for the output layer, but not for the hidden layer. The error for the output layer can be calculated as:

$$\delta_{pj} = -\frac{\delta E_p}{\delta net_{pj}} = -\frac{\delta E_p}{\delta o_{pj}} \times \frac{\delta o_{pj}}{\delta net_{pj}} \quad (2.25)$$

$$\frac{\delta o_{pj}}{\delta net_{pj}} = f'_j(net_{pj}) \quad (2.26)$$

$$\frac{\delta E_p}{\delta p_j} = t_{pj} - o_{pj} \quad (2.27)$$

$$\delta_{pj} = f'_j(net_{pj})(t_p - o_{pj}) \quad (2.28)$$

$$= k \times o_{pj}(1 - o_{pj})(t_p - o_{pj}) \quad (2.29)$$

$$\delta_{pj} = o_{pj}(1 - o_{pj})(t_p - o_{pj}) \quad (\text{for } k = 1) \quad (2.30)$$

For hidden units, where output are connected to  $k$  other units, the error is defined in proportion to the sum of the errors of all  $k$  units as modified by the weights connecting these units.

$$\delta_{kj} = -(\sum_k \delta_{pk} w_{jk}) o_{kj} (1 - o_{kj}) \quad (2.31)$$

The equation 2.31 shows that the change in error is proportional to the errors  $\delta_{pk}$  in subsequent units, so the error has to be calculated in the output unit first as given by equation 2.21 and then passed back through the hidden layers using equation 2.31 to alter the connecting weights between the units. It is this process of passing back of the error value that leads to the network being referred to as back propagation network.

During the training phase, the iterations are continued till the output error declines to an acceptable level (Kavzoglu and Mather, 2003). The training data are used to adjust the weights and thresholds to minimize the error as given by equation (2.21). This equation represents the amount by which the output of the neural network differs from the required output. This error can be represented as energy function. The energy function is related to the weights between the units and the input data.

In the error surface, each of the N weights and thresholds are taken as single dimension with N+1 being the network error. For example, with one weight, the graph (Figure 2.5) is two-dimensional, one representing weight and other, the error.

The energy surface (Figure 2.6) consists of valleys, wells and peaks. The points of minimum energy correspond to the wells and maximum are related with peaks. The aim of neural network is to minimize the error (equation 2.21) by adjusting the weights of the network so as to find the lowest point in the error surface called as global minima. The error surface in general are characterized by local minima that is the error surface has a lot of wells which are lower than the surrounding terrain but are above the global minima.

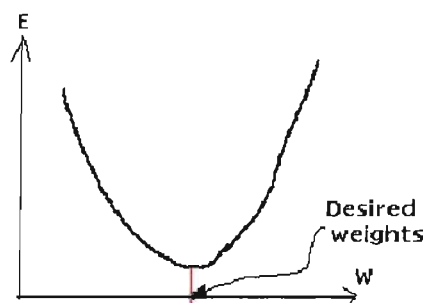


Figure 2.5: Error surface.

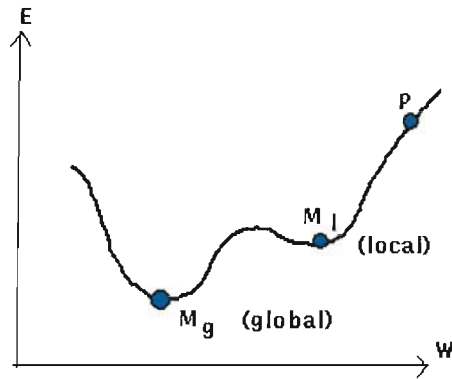


Figure 2.6: Energy surface showing local and global minima.

In case energy function is in a local minima (*i.e.*, in every direction in which the network could move, the energy is higher than at the local minima), it becomes difficult to reach global minima. This problem of reaching global minima can be minimized by altering the weights in a progressively decreased manner. A gain term  $\eta$  (equation 2.24) is therefore incorporated. Large gain means that large steps are taken across the error surface and small steps when the gain value is smaller. If the gain term is made large in the beginning and smaller later, the gradient descent will take larger steps at first thereby possibly bypassing local minima in the initial stage. The smaller steps at the later stage will help to settle in some deeper minima possibly global minima.

The changes in the weights can also be given some momentum by introducing a momentum factor into the weight adaptation equation (equation 2.31). This will produce a proportional change in the weight (*i.e.*, it will produce a large change in weights if changes are currently larger and smaller if the changes are smaller). This process is likely to skip local minima in the initial stages as the momentum term will push the changes in the direction of downward slope. The inclusion of momentum term will therefore reduce time to converge towards global minima.

The training stops when the error stops decreasing. To achieve this, the network may over-fit the training data (Figure 2.7) and may not generalize well for unseen cases. A solution to check over-fitting of training data are by making use of a validation data. The

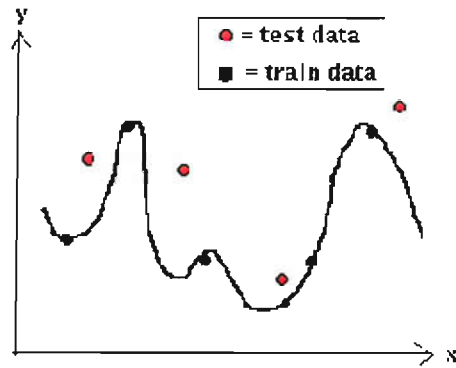


Figure 2.7: Over-fitting of training data.

validation data are like training data, which has not been used to train the network. The accuracy of networks trained by the training data are evaluated by the validation data and the network with the smallest error with respect to validation data are selected. The use of validation data to evaluate the accuracy of network can over-fit the validation data and, therefore, the accuracy of the selected network should be confirmed by evaluating its accuracy on a third independent data called as test data.

### 2.5.2 Limitations of Artificial Neural Network

Artificial neural networks operate directly on the training data without regard to any distribution assumption as in MLC. They, therefore, fall into the class of non-parametric classifiers. They, however, have all the problems associated with supervised classification. For instance, the accuracy of the classification is dependent upon the quality of the training data. Factors like number and composition of training samples, number of bands used, affect the accuracy of ANN (Foody and Arora, 1997). Error in correct identification of an individual training sample pixel may not influence statistical classifier like MLC but has considerable impact on ANN (Mather, 1999). For example one erroneous pixel may not contribute too much to the mean of the spectral distribution of the category under consideration on which the MLC is trained but will have direct effect on ANN as the classifier is trained directly on the training data. The composition of the training data especially those lying close to the decision boundary, prone to be removed by

conventional training set refinement techniques (for use in parametric classifier like MLC) play a very important role in accurate classification using ANN ( Foody, 1999). A neural network aims to minimize an overall error and, therefore, the proportion of training data of the categories under investigation is of paramount importance. For example, a network trained on unbalanced data set (unequal training data for classes under investigation), would bias its decision towards the majority class. This is so as the network may minimize overall error by classifying majority class accurately as compared to the minority class. The training data should, therefore, have equal representation for all the categories under investigation.

One of the most important limitations of ANN is the judicious selection of various parameters associated with the design. Many of these parameters are interrelated and in the absence of definite rules, trial runs or experience of the analyst is the usual recourse to develop an optimal network. Thus for instance, the number of units and hidden layers that should be used is a commonly encountered problem (Foody and Arora, 1997). Kolmogorov's theorem, however, sheds some light on the requirement of hidden layers and associated units. It states that a three-layer perceptron with  $n(2n+1)$  nodes can compute any continuous function of  $n$  variables.

The architecture of the network has to be so designed, that the network has the capacity to learn accurately during the training phase and also maintain a high generalization power. A large network is likely to learn the training data accurately but may not generalize well, but a small network on the other hand, will have difficulty in learning but may end with a greater generalization power. There exists a variety of procedure which determine the network size like pruning, growing hidden units or by cross validation. The learning and generalization power of a network is also dependent on the number of training iterations. More the iterations, more accurate is the training but lesser the generalization ability. As the objective in classification is to generalize well for unseen

cases, utmost care should be taken, not to over train the network, as it would decrease the generalization power.

Feature reduction can also be employed to reduce the network size, as only the most discriminating variables are then chosen for analysis, thereby decreasing the training time while maintaining if not possibly increasing the classification accuracy (Battiti, 1994; Lee and Landgrebe, 1997; Benediktsson and Sveinsson, 1997; Foody, 1999).

The remote sensing specialist is usually not a neural network specialist, therefore, the design of the classifier (network) should be automated as much as possible. There are software's available which automatically generate promising networks from the training data set. Such an arrangement reduces time to design optimal networks manually by repeated trial and runs, and thus compensates for the long time taken to train the network, and makes neural network a very strong candidate for classification problems.

## **2.6 Decision Trees**

A decision tree can be defined as a classification procedure that recursively partitions data into smaller subsets on the basis of tests or thresholds at each node in the tree (Figure 2.8). The tree is composed of a root node, which is at the top of the hierarchy (unlike a natural tree) contains all the input data, a set of internal nodes (splits) and a set of terminal nodes (leafs).

Decision tree classification techniques have been used successfully for a wide range of classification problems but have not found widespread use in remote sensing until recently (Safavian and Landgrebe, 1991), although these techniques have many advantages over the traditional supervised classification procedures such as the maximum-likelihood classifier. Decision trees are non-parametric and, therefore, do not make any assumptions regarding the distribution of input data. They are computationally efficient and have intuitive appeal, as the tree structure is easily interpretable (Pal and Mather, 2004). The



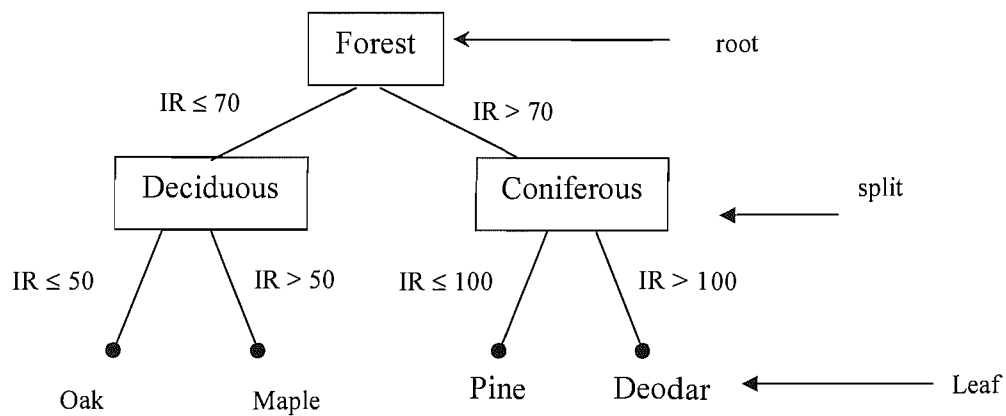


Figure 2.8: Classification of a forest using a decision tree. Each box is a node with root at the top which contains all the data (Infrared values (IR)). Splitting rules based on the value of IR values are applied at each node to make the data in the child nodes homogeneous. The forest is classified as Oak, Maple, Pine and Deodar as shown by the leaf of the decision tree.

tree structure clearly shows the features (discriminating variables, for example spectral bands) of the input data used by the tree to discriminate the categories.

### 2.6.1 Classification of Decision Trees

Decision trees can be classified as homogenous or heterogeneous based on the algorithms employed to estimate the splitting of data at the nodes. Traditionally, homogenous decision trees are used which employ only a single algorithm to estimate the split at the node. A hybrid decision tree on the other hand employs more than one algorithm. Homogenous decision trees can further be subdivided or classified as univariate and multivariate.

#### 2.6.1.1 Univariate Decision Tree

A univariate decision tree (UDT) is a type of decision tree in which the decision boundaries at each node of the tree are defined by a single feature of the input data (Swain and Hauska, 1969). The threshold/test applied at each node for splitting the data are

estimated from the training data. For continuous data, the test is of the form;  $u_i > t$ , where  $u_i$  is a feature in the data space (training data) and  $t$  is a threshold (lies in the observed range of  $u_i$ ). The threshold  $t$  can be estimated by some measure, which maximizes the dissimilarity in the descendent nodes.

### **2.6.1.2 Multivariate Decision Tree**

A multivariate decision tree (MDT) is like a UDT, except for the splitting test used at the nodes. The splitting test in MDTs is based on more than one feature of the input data, unlike UDTs which are based on only one feature.

MDTs are often more compact and can be more accurate for classification than UDT's (Brodley and Utgoff, 1995). However relative to UDT's, MDT's have many disadvantages because of their complex structure (splitting test based on more than one feature of the input data). First, unlike UDT many different algorithm/rules can be used to split the nodes (Briemmen *et al.*, 1984; Murthy *et al.*, 1994), which makes MDT more difficult to interpret than UDTs. Secondly, as the split at the nodes are governed by more than one feature, so several different algorithms are available to perform feature selection. Further, the feature selection is carried locally rather than globally that is, they choose the features to include in each test at the nodes on the basis of the data observed at a particular node rather than selecting a uniform set of features to be used for the entire tree (Friedl and Brodley, 1997).

### **2.6.1.3 Hybrid Decision Tree**

A hybrid decision tree is a decision tree, where different classification algorithms may be used in different subtrees of a larger tree (Brodley, 1995). Figure 2.9 shows an example of hybrid decision tree in which three different classification algorithms, a linear

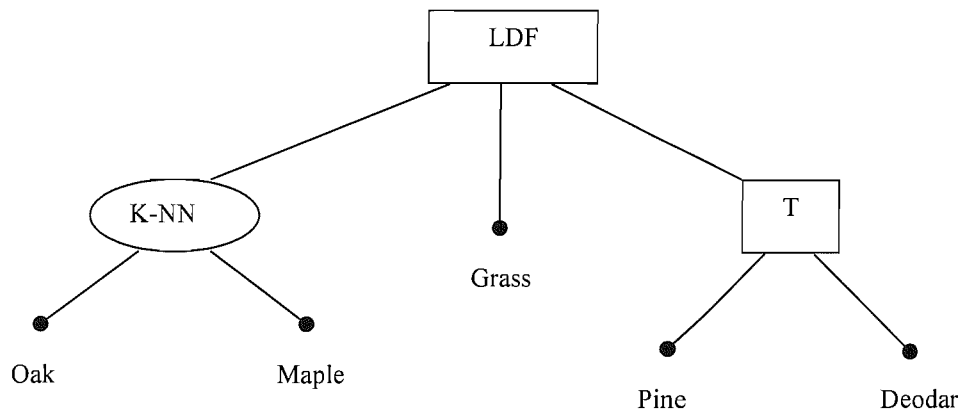


Figure 2.9: A hybrid decision tree classifier, using splitting rules as linear discriminant function (LDF) at root, k nearest neighbour on the left tree and a univariate decision tree on the right to classify four classes Oak, Maple, Grass, Pine and Deodar.

discriminant function (LDF), K nearest neighbour (K-NN) and Univariate decision tree (UDT) are used to classify a dataset.

The objective of using different classification algorithms in a hybrid decision tree is due to the fact that different algorithms can classify different data with different accuracy. This property of classification algorithms is termed “selective superiority” (Brodley, 1995). By allowing a hybrid hypothesis space, a decision tree can be adapted to the problem, thereby producing a more accurate classification result (Brodley, 1995).

The decision tree generation broadly consists of two phases; the tree design and the tree pruning.

### 2.6.2 Design of a Decision Tree

The main objectives of decision tree as any other classifier are to correctly classify training samples and generalize well to unseen cases. Additionally, the tree should be so designed that it has a simple structure and easy to update if more training data are made available (Safavian and Landgrebe, 1991).

The design of a decision tree depends on appropriate choice of the tree structure, choice of feature subsets to be used at each internal node and choice of the decision rule or strategy to be used at each internal node (Kulkarni and Kanal, 1976; Kurzynski, 1983).

Since, the number of possible tree structures, even for a moderately small number of classes can be astronomical; it is very difficult to design an optimal classifier. To make the design task easier, binary decision trees are often adopted. Binary trees are those in which each node contains two children identified as left and right child. However, the discrimination ability is not necessarily weakened by choosing a binary approach, since a general decision tree can be uniquely transformed into an equivalent binary tree (Rounds, 1980). The accuracy of classification with a decision tree depends on how the tree was designed (Safavian and Landgrebe, 1991). The various heuristic methods for designing a decision tree can roughly be divided into four categories (Safavian and Landgrebe, 1991) bottom-up-approach, top-down approach, the hybrid approach and tree growing-pruning approach.

#### **2.6.2.1 Bottom-up Approach**

In a bottom up approach, the information classes are combined until one is left with a node containing all the classes. In this approach, the design is initiated at the level of leaf, where all information classes reside and move upwards till the root is reached, where all the data resides.

Initially, pair-wise class separation are computed (Richards and Jia, 1998) using a distance metric, such as Mahlanobis distance. The two most similar classes are merged and their mean calculated. The process is continued until all the classes lie in a single group at the top of the tree that is the root. In a tree constructed this way, the more separate classes are discriminated first, near the root and more subtle ones at the later stages of the tree.

### **2.6.2.2 Top-down Approach**

The top-down approach employs splitting rules starting from the root of the tree, the classes are divided until a stopping criterion is met. It is also called as a progressive two class decision classifier (Richards and Jia, 1998). The top down method of tree design is the best known (Floriana *et al.*, 1997) amongst all the approaches.

In top down approaches, the design of a decision tree reduces to the selection of a node splitting rule, decision as to which nodes are terminal and assignment of each terminal node to a class label (Safavian and Landgrebe, 1991).

The task of class assignment is the easiest of the three tasks. The objective is to minimize the misclassification rate by assigning terminal nodes to the classes that have the highest probabilities of membership (based on the majority rule that is assign to the terminal node, the label of the class that has the most samples present).

The decision whether a node is terminal can be made by the splitting rules itself as the use of stopping rules may halt the growth of the tree too soon at some nodes and too late at the others (Brieman *et al.*, 1984). The use of the splitting rules, on the other hand, helps the tree to grow until the leafs are left with only the pure classes. This may result in very large trees, which can be pruned. Most of the research in decision tree design has concentrated in finding various splitting rules (Safavian and Landgrebe, 1991).

#### **2.6.2.2.1 Selection of Node Splitting Rules**

The splitting rule is aimed to make the data in the child nodes purer (homogeneous) (Fraser *et al.*, 2005). This can be achieved by two approaches; the first approach measures the goodness of split at the nodes while the second approach tries to minimize the impurity of the training data. Information gain and Information gain ratio, Gini-index, Towing rule and Chi-square contingency method are some of the node splitting rules. However in the present work Information gain and Information gain ratio splitting rule has been used and discussed in detail.

### 2.6.2.2.1.1 Information Gain and Information Gain Ratio

Quinlan (1986) proposed the use of information gain and information gain ratio based on the concept of entropy to represent information in data sets. Quinlan described the entropy or information content as:

$$- \sum p_j \log(p_j) \quad (2.32)$$

Where,  $p_j$  is the probability of class  $j$

For a given training set  $G$ , the probability that a case selected randomly belongs to class  $C_j$  is given by:

$$\frac{\text{freq}(C_j, G)}{T} \quad (2.33)$$

Where,  $\text{freq}(C_j, G)$  is the number of cases in  $G$  that belongs to  $C_j$ , and  $T$  denotes total cases in training data. The information (entropy) gained using equation 2.32 and 2.33 can be defined as:

$$\text{info}(G) = - \sum_{j=1}^k \frac{\text{freq}(C_j, G)}{G} \times \log_2 \frac{\text{freq}(C_j, G)}{G} \quad (2.34)$$

If a test  $X$ , partitions the set  $T$  into  $n$  outcomes, the expected information content can be found, as the weighted sum over the subtrees as:

$$\text{info}_x(G) = - \sum_{j=1}^n \frac{G_j}{G} \times \text{info}(G_j) \quad (2.35)$$

The information, therefore, gained by splitting training set  $T$  using test  $X$  can be measured by:

$$\text{Gain}(x) = \text{info}(G) - \text{info}_x(G) \quad (2.36)$$

This criterion is called as the gain criterion and is used to select a test, which maximizes the information gain. The gain criterion however has a serious drawback as it

has a strong bias in favour of tests with many outcomes. The bias can be rectified by normalization of equation 2.35 as:

$$\text{split info } o(x) = -\sum_{i=1}^n \frac{G_i}{G} \log_2 \frac{G_i}{G} \quad (2.37)$$

The split info represents the potential information generated by splitting T into n subsets, whereas the information gain measures the information relevant to classification that arises from the same division. The gain ratio gives the proportion of information generated by the split useful for classification.

$$\text{Gain ratio } (x) = \text{gain } (x) / \text{split info } (x) \quad (2.38)$$

The training data T then recursively partitions the data set, so that the gain ratio is maximized at each node of the tree. The procedure is continued until each leaf node contains data only from a single class or any further splitting yields no increase in information.

### 2.6.2.3 Hybrid Approach

The hybrid approach proposed by Kim and Landgrebe (1991) uses both bottom-up and top-down approaches sequentially. The rationale for the approach is that the bottom up procedure assists the top-down procedure in the growth of the tree. The procedure, first consider the entire data, uses a bottom up approach to come up with two clusters of classes. Then mean and covariance of both the clusters are computed, to be used by top down algorithm to generate two new clusters from each of the original clusters. If the resulting clusters contain only one class, the cluster is labelled as terminal; else the procedure is repeated till all the clusters are labelled as terminals.

### 2.6.3 Pruning of Decision Trees

A decision tree can be grown so as to have zero error on the training data. In other words, the decision tree can grow indefinitely until all the classes are separated. This process can lead to a very large decision tree (Simard *et al.*, 2000) and may over-fit to the

noise in the training data. This would result in erroneous classification of unseen cases. To overcome this problem, the tree needs to be pruned in order to generalize accurately to unseen cases (Pal and Mather, 2004).

The pruning process involves the elimination of the inefficient or weak branch (those branches of tree, the removal of which do not alter classification accuracy) of the decision tree. This results in a less complex and more interpretable tree. There are two different ways of pruning (Breiman *et al.*, 1984); either by prospectively deciding when to stop the growth of a tree referred to as pre-pruning or by reducing the size of a fully expanded tree by pruning some branches (post-pruning).

Pre-pruning methods establish stopping rules for preventing the growth of those branches that do not seem to improve the predictive accuracy of the tree (Floriana *et al.*, 1997). The problem with this approach is to specify a correct stopping rule (Breiman *et al.*, 1984) as also to understand the benefits of the splits, to take place in the pruned part of the branch. To ward off these problems, post pruning methods are adopted.

In post pruning, a tree is grown, even when it seems worthless and is then retrospectively pruned of those branches that seem superfluous with respect to predictive accuracy (Niblett, 1987).

In general, pruning methods aim to simplify decision trees that over-fits training data resulting in higher accuracies for unseen data. The benefits of pruning lured many researchers with the outcome that many pruning methods are available now.

Some methods follow the top-down approach that works from root to leaf to examine the branches to be pruned, other follows the reverse direction; bottom up approach. Furthermore, some method employ only the training set to evaluate the accuracy of a decision tree; other works on additional data set called as pruning set. The use of an independent pruning set might be problematic especially when small training samples are involved.



There are a number of post pruning methods available because of the inherent advantages over pre-pruning approach, but the following six have achieved widespread popularity (Floriana *et al.*, 1997).

1. Reduced error pruning (REP)
2. Pessimistic error pruning (PEP)
3. Minimum-error pruning (MEP)
4. Critical value pruning (CVP)
5. Cost-complexity pruning (CCP)
6. Error based pruning (EBP)

The pessimistic error pruning has been used in the present work and, therefore, discussed hereafter.

### 2.6.3.1 Pessimistic Error Pruning

The PEP method was proposed by Quinlan (1987) uses the same training data, both for growing and pruning a tree. The apparent error, that is the error rate on the training set is, therefore, biased and should not be used to select the best pruned tree. Quinlan (1987), therefore, introduced the continuity correction, considering binomial distribution that might give a more realistic error rate.

Let  $e(t)$  and  $n(t)$  represent the number of examples misclassified and total number of examples at node  $t$  respectively, then apparent error rate  $r(t)$  at node  $t$  is given by:

$$r(t) = \frac{e(t)}{n(t)} \quad (2.43)$$

Similarly, apparent rate  $r(T_t)$  for whole subtree  $T_t$  with total  $n$  leaves is:

$$r(T_t) = \frac{\sum_{i=1}^n e(i)}{\sum_{i=1}^n n(i)} \quad (2.44)$$

The corrected misclassification rate  $r'(t)$  at node  $t$  after continuity correction for binomial distribution is:

$$r'(t) = \frac{[e(t) + 1/2]}{n(t)} \quad (2.45)$$

Similarly corrected misclassification rate  $r'(T_s)$  for the whole subtree is given by:

$$\begin{aligned} r'(T_s) &= \frac{[\sum_{i=1}^n (e(i) + 1/2)]}{\sum_{i=1}^n n(i)} \\ &= \frac{\sum_{i=1}^n e(i) + n/2}{\sum_{i=1}^n n(i)} \end{aligned} \quad (2.46)$$

The numerator of equation 2.45 and 2.46 represents the number of errors at node  $t$  and for the subtree  $T_s$  respectively. The subtree is expected to make fewer errors on the training set than the parent node  $t$  when  $t$  becomes a leaf, but sometimes the reverse may happen, that is  $n'(t) \leq n(T_s)$  due to continuity correction, in which case the node  $t$  is pruned. For this reason, Quinlan (1987) weakens the condition so that:

$$e'(t) \leq e'(T_s) + SE(e'(T_s)) \quad (2.47)$$

Where, SE is the standard error for subtree  $T_s$ .

Hence, the algorithm only keeps the subtree, if the corrected figure for subtree is more than one standard error better than the figure for the node.

As the algorithm evaluates each node starting from the root of the tree and, if a branch is pruned then its descendent structure is not examined. This top-down approach, therefore, accomplishes the task very quickly.

#### 2.6.4 Limitations of Decision Tree Classification

A number of factors have to be considered in the optimal design of decision trees which includes the choice of decision tree (univariate, multivariate or hybrid), attribute selection methods to be employed to split the nodes and pruning methods to prune the tree.

Reliably few studies using decision trees have been undertaken using remote sensing data (Safavian and Landgrebe, 1991; Friedl and Brodley, 1997; De Colstrun *et al.*, 2003).

De Fries *et al.*, (1998) used decision tree classifier for global land cover classification at 8 km. spatial resolution data. They studied the effect of imbalanced training sets and found that the tree might be slightly biased towards those cover types with a large number of training pixels as compared to a tree trained with equal number of training pixels.

Pal and Mather (2002) employed decision tree on Landsat-7 ETM+ data and compared the effect of various node splitting rules and pruning methods on classification accuracy. They compared the effect of four node splitting rules, the information gain, gini index, information gain ratio and chi-square and concluded that the overall accuracy obtained is almost same, except for information gain ratio, which results in an increase of less than one per cent. The study confirms the finding of Brieman *et al.*, (1984) that classification accuracy is not affected by choice of attribute selection measure. They also compared the accuracy resulting from pruning methods REP, PEP, CVP, CCP and EBP (Esposito *et al.*, 1997) and found that accuracy ranged from 81.4 % to 82.9 %.

Friedl and Brodley (1997) applied decision trees on TM data (30 m spatial resolution) and concluded that hybrid decision tree provided higher accuracy (76 %) as compared to univariate (75.2 %) and multivariate decision tree (75.9 %). They also revealed a tendency of decision trees to penalize solutions for classes with fewer observations in the training data and suggested that this bias also depend on the overall separability in feature space relative to other classes.

## 2.7 Support Vector Machines

The support vector machines (SVM) have recently attracted the attention of the remote sensing community (Brown *et al.*, 1999; Huang *et al.*, 2002; Halldorsson *et al.*, 2003). They are gaining popularity due to many attractive features. A key feature behind the technique is to separate the classes with a decision surface that maximizes the margin between them called as the optimal separating hyperplane. A separating hyperplane refers to a plane in a multidimensional space that separates the data samples of two classes. There can be a number of separating hyperplane that can separate the classes but the optimal separating hyperplanes is expected to generalize well for unseen cases (Figure 2.10).

A key feature of this classifier is its ability to use high dimensional data without the usual recourse to feature selection to reduce the dimensionality and is, therefore, being used in very diverse fields like optical character recognition, hand written digit recognition (Vapnik 1995).

The SVM derived its name from support vectors, training data points which lies on two parallel hyperplane (Figure 2.11) and contain all the information relevant to the classification problem.

### 2.7.1 Design of Support Vector Machines

In this section, the mathematical derivation of support vector machines (SVM) is introduced in steps, first the simplest case of a linear classifier and linearly separable case is described, followed by a linear classifier and non-separable case and finally, a non-linear classifier (non linear decision surface) and non separable case, useful of all the three cases (Osuna *et al.*, 1997).

### 2.7.1.1 Linearly Separable Case

Linearly separable case is one in which the data set can be separated into its constituent classes by a linear separating surface (Figure 2.10).

Let the training data of two separable classes (Figure 2.10) with a total of  $r$  samples,  $a$  belonging to class I, and  $b$  belonging to class II ( $a + b = r$ ) be represented by:

$$(x_i, y_i), \dots, (x_r, y_r)$$

where,

$x_i \in \mathbb{R}^n$  is an  $n$  dimensional space and  $y_i \in \{1, -1\}$  are class labels (Huang *et al.*, 2002).

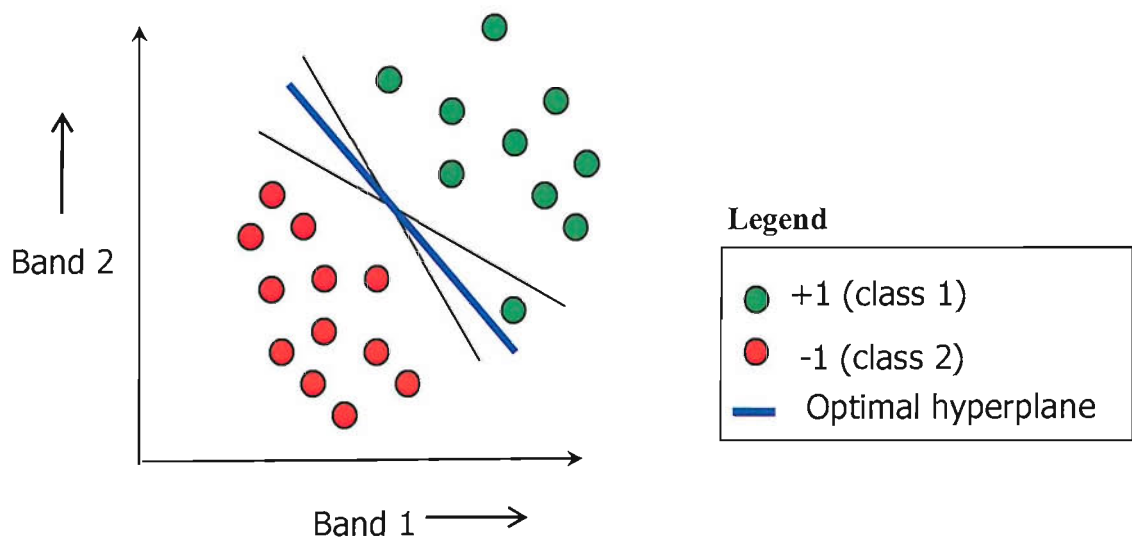


Figure 2.10: The two classes ( ● and ● ) can be separated by a number of decision surfaces (shown by black lines in between the two classes). However, there is only one decision surface called the optimal separating hyperplane (shown by dark blue line), that is expected to generalize accurately on unseen cases as compared to other decision surfaces.

The goal is to produce a classifier that works well on unseen data (*i.e.*, it generalizes well). There can be a number of hyperplanes that can separate the data but there is only one optimal separating hyperplane (Figure 2.10) which is expected to generalize well in comparison to other hyperplanes.

The goal of optimal hyperplane is that;

1. Data belonging to both the classes (Figure 2.11) should lie on its opposite side.
2. It should be so placed that the distance of the closet data points (training data) in both the classes are furthest from it (to generalize well for unseen cases).

The hyperplane can be defined by the equation:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \quad (2.55)$$

where,

$\mathbf{x}$  is a point on the hyperplane,

$\mathbf{w}$  is an  $n$  dimensional vector perpendicular to the hyperplane (determines the orientation of the plane in space),

$b$  is distance of the closet point on the hyperplane to the origin (offset of hyperplane from origin)

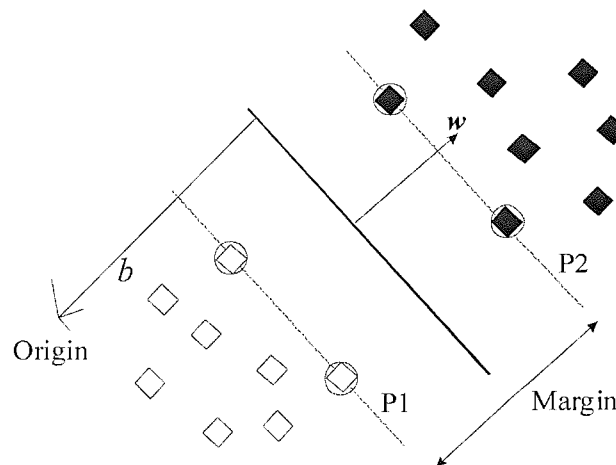


Figure 2.11: Optimal hyperplane (dark black line) with parameters  $\mathbf{w}$  and  $b$ .

Parallel planes P1 and P2 contains support vectors relevant in the formulation of optimal hyperplane.

The classifier can be defined by:

$$F_{w,b} = \text{sgn}(w \cdot x + b) \quad (2.56)$$

Let  $d_i$  be the perpendicular distance of  $x_i$  from any point  $x$  on hyperplane.

Where,  $y_i$  are class labels +1 or -1 of the two classes.

$$d_i = y_i \frac{w}{|w|} \cdot (x_i - x) \quad (2.57)$$

Substituting equation 2.55 in equation 2.57 gives

$$d_i = \frac{y_i}{|w|} (w \cdot x_i - (-b)) \quad (2.58)$$

$$d_i = \frac{y_i}{|w|} (w \cdot x_i + b) \quad (2.59)$$

The parameters  $w$  and  $b$  describing the hyperplane can be scaled by a constant without changing the hyperplane. This implies that the decision surface will remain same if both  $w$  and  $b$  are scaled by the same non-zero constant. In order to eliminate this scaling freedom so that each decision surface corresponds to a unique pair  $(w, b)$ , the following constraint is imposed.

$$y_i(w \cdot x_i + b) - 1 = 0 \quad \text{if } i \text{ is a support vector} \quad (2.60a)$$

$$> 0 \quad \text{if } i \text{ is not a support vector} \quad (2.60b)$$

or,

$$y_i(w \cdot x_i + b) - 1 \geq 0 \quad i=1,2, \dots,k \quad (2.61)$$

The set of hyperplanes that satisfy equation 2.60a are called as canonical hyperplanes. The data points that are nearest to the hyperplanes are called as support vectors and play a very important role in establishing the optimal separating hyperplane. The distance between the canonical hyperplane to the support vectors can be found, by substituting equation 2.60a into equation 2.59, as:

$$d_i = \frac{1}{|\mathbf{w}|}$$

$$d_i = |\mathbf{w}|^{-1} \quad (2.62)$$

Maximizing this distance would help in finding the optimal hyperplane that is separating hyperplane for which the distance between the two convex hulls (of the two classes of training data), measured along a line perpendicular to the hyperplane, is maximized. This distance is called as margin.

$$\max_{\mathbf{w}, b} |\mathbf{w}|^{-1} \quad (2.63)$$

under the constraints,

$$y(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0 \quad i=1,2 \dots k$$

Hence, the hyperplane that optimally separates the data are the one that minimizes:

$$\phi = \frac{1}{2} \|\mathbf{w}\|^2 \quad (2.64)$$

The quadratic optimization problem of equation 2.64 can be simplified, so as to replace the inequalities with a simpler form by transforming the problem to a dual space representation using lagrangian multipliers  $\alpha_i$  as;

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^r \alpha_i (y_i [(\mathbf{w} \cdot \mathbf{x}_i + b)] - 1) \quad (2.65)$$



The solution is basically determined by a saddle point of lagrangian, which has to be minimized with respect to  $w$  and  $b$ , and maximized with respect to  $\alpha$ , the lagrangian multipliers. The dual problem becomes:

$$\max_{\alpha_i = 1, \dots, r} \min_{w, b} L(w, b, \alpha_1, \dots, \alpha_r) \quad (2.66)$$

with constraints,

$$\alpha_i \geq 0, i = 1, \dots, r$$

$$y(w \cdot x_i + b) - 1 \geq 0, i = 1, \dots, r$$

The minimization of lagrangian (equation 2.65) with respect to  $w$  and  $b$  gives:

$$\frac{\delta L}{\delta w} = 0 \Rightarrow \sum_{i=1}^r \alpha_i y_i x_i = w \quad (2.67)$$

$$\sum_{i=1}^r \alpha_i y_i = 0 \quad (2.68)$$

Equation 2.67 shows that optimal hyperplane can be written as a linear combination of training data.

According to Karush-Kuhn-Tucker (KKT) theory, only data points that satisfy the inequality equation 2.61 can have non-zero coefficients  $\alpha_i$ . The KKT conditions play a central role in the solution of constrained optimization. The equations 2.67 and 2.68 along with the constraints of equation 2.66 are KKT conditions. As the constraints in SVM are linear (equation 2.61) and the problem is convex, the KKT conditions are necessary and sufficient for  $w, b, \alpha$  to be the solution of dual optimization problem (equation 2.65).

The pattern  $x_i$ , for which  $\alpha_i > 0$  lie exactly on the margin according to equation 2.61 and are used to establish the optimal hyperplane (equation 2.65); hence these data points are called as support vectors (Pal and Mather, 2005). All remaining examples in the training dataset are irrelevant. Their constraints are satisfied automatically and they do not

appear in the expansion of equation 2.67. The support vectors play a very important role in establishing the optimal separating hyperplane and, therefore, the classification technique is referred to as the support vector machine.

The expression obtained (equation 2.69) by substituting the expansion of equation 2.67 into the decision function given by equation 2.55, can be evaluated by the pattern to be classified and the support vectors.

$$f(x) = \text{sgn}\left(\sum_{i=1}^r \alpha_i y_i(x, x_i) + b\right) \quad (2.69)$$

### 2.7.1.2 Linearly Non-separable Case

Typically land cover classes cannot be separated by a linear separating hyperplane in feature space (Figure 2.12) because of the outliers. In such a case, a linear separating surface, as in the linearly separable case, does not exist. It is, therefore, not possible to satisfy all the constraints of equation 2.61.

$$y(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0, \quad i=1, \dots, r$$

To deal with such cases using only linear separating boundaries, Cortes and Vapnik (1995) introduced a new set of variables  $\{\xi_i\}_{i=1}^r$ ; that measures the amount of violation of the constraints.

The constraint then becomes:

$$y(\mathbf{w} \cdot \mathbf{x}_i + b) > 1 - \xi_i \quad (2.70)$$

The above constraints, in case of outliers, can always be met by making  $\xi_i$  very large, so,

a penalty term  $C \sum_{i=1}^r \xi_i$  is added to penalize solutions for which  $\xi_i$  are very large. The

constant  $C$  controls the magnitude of the penalty associated with training samples that lie on the wrong side of the decision boundary. The optimization from equation 2.64 then becomes:

$$\min\left[\frac{\mathbf{w}^2}{2} + C\sum_{i=1}^r \xi_i\right] \quad (2.71)$$

under the constraints,

$$y(\mathbf{w}\cdot\mathbf{x}_i + b) > 1 - \xi_i$$

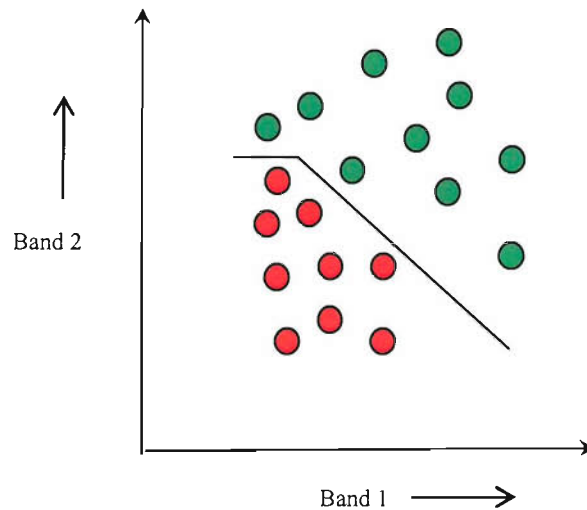


Figure 2.12: Training data cannot be separated by single linear separating hyperplane.

If the classes overlap considerably in feature space then  $C\sum_{i=1}^r \xi_i$  can be very large and the hyperplane may not generalize well.

The solution of this problem is determined by the saddle point of the lagrangian in a way similar to that described in linearly separable case. The uncertain part however is that the coefficient  $C$  has to be chosen by the analyst so as to reflect the noise in the data.

### 2.7.1.3 Decision Surfaces

In the situation, where it is not possible to define a hyperplane by linear equations on training data (*e.g.*, if the classes overlap considerably in feature space), the techniques can be extended to allow for non-linear decision surfaces (Pal and Mather, 2004). A technique introduced by Boser *et al.*, (1992), maps input data into a high dimensional space through some non-linear mapping. The transformation to a high dimensional space spreads the data in a way that facilitates the fitting of a linear hyperplane (Figure 2.13).

More precisely, one maps the input data into a high dimensional-space  $H$ , through a mapping function  $\phi$ :

$$\phi: R^n \rightarrow H \quad (2.72)$$

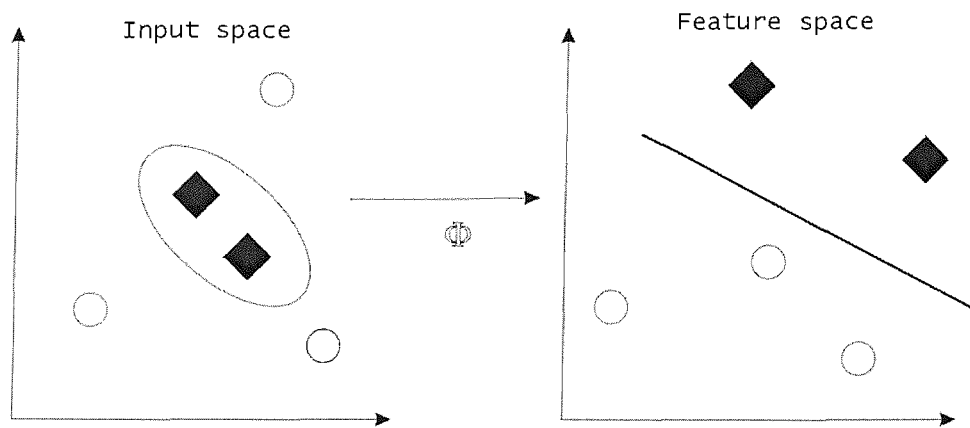


Figure 2.13: The two classes (  $\blacklozenge$  and  $\circ$  ) cannot be separated by linear decision surface but after transformation to a high dimensional feature space through a function  $\phi$ , the data can be separated by a linear decision surface (Vapnik, 1995).

An input data  $\mathbf{x}$  can be represented as  $\phi(\mathbf{x})$  in the high dimensional space  $H$ . The evaluation of the decision function given by equation 2.72 requires the computation of  $(\phi(\mathbf{x}), \phi(\mathbf{x}_i))$  in a high dimensional space. These computationally expensive calculations are reduced significantly by using a positive definite kernel (Vapnik, 1995), such that:

$$(\phi(\mathbf{x}), \phi(\mathbf{x}_i)) = k(\mathbf{x}, \mathbf{x}_i) \quad (2.73)$$

leading to decision functions of the form;

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^r \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b\right) \quad (2.74)$$

A kernel that can be used to construct a SVM must meet Mercer's condition (Huang *et al.*, 2002). The following two types of kernels meet the condition (Marcal *et al.*, 2005)

The polynomial kernels of degree  $p$ ,

$$k(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x} \cdot \mathbf{x}_i + 1)^p$$

and the radial basis functions (RBF)

$$k(\mathbf{x}, \mathbf{x}_i) = e^{-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2}$$

where  $\gamma$  is the parameter controlling the width of the Gaussian kernel

### 2.7.2 Multi-Class Support Vector Machine

Support vector machines were designed for binary classification that is one SVM can only separate two classes. Many real world problems have more than one class, one of the examples being land cover. SVMs, therefore, must be adapted to multi-class problems. Two simple ways to generalize a binary classifier to a multi-class classifier (Gualtieri and Cromp, 1998) for  $k$  classes are:

1. Train  $k$  binary classifiers, each one using training data from one of the  $k$  classes and lumping together training data of the remaining  $k-1$  classes into a mega class. In other words, the strategy is to break the  $k$  class problem, into  $k$  binary classifiers, each trained to separate one class from the rest (one-against-all approach) and then combining them by carrying multi-class classification (applying a voting scheme) according to the maximal output before applying the decision rule.
2. Construct a machine (classifier) for each pair of classes (one-against-one approach) resulting in  $n(n-1)/2$  machines. When applied to a test data, each machine gives one vote to the winning class, and the pixel is labeled with the class having most votes.

In the first option, the sizes of the two concerned classes can be disproportionately imbalanced, because one of them groups  $n-1$  classes, a mega class against a single class. A classifier may not be able to find a boundary between the two classes because the classifier probably would make least errors by labelling all data points belonging to the smaller class

with the mega class. The second option has the drawback that it requires  $n(n-1)/2$  binary classifiers to be trained for an  $n$  class problem, as compared to only  $n$  for the first option.

Multi-class classifications of remotely sensed data by SVM have to-date been based on the above approaches (Huang *et al.*, 2002; Halldorsson *et al.*, 2003; Mercier and Lennon, 2003; Gualtieri and Crompton, 1998). While both strategies to reducing the multi-class problem to a set of binary classifications enable the basic SVM to be employed, a more appropriate approach, that is less computationally demanding, is to consider all classes at one time, yielding a multi-class SVM (Hsu and Lin, 2002).

One means to achieve this, which is similar in basis to the ‘one-against-all’ approach, is by solving a single optimization problem. With this,  $n$  two class rules where the  $m^{\text{th}}$  function  $\mathbf{w}_m^T \boldsymbol{\phi}(\mathbf{x}) + b$  separates the training data vectors of class  $m$  from that of others are constructed. Hence, there are  $n$  decision functions or hyperplanes but all are obtained by solving one problem,

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \sum_{m=1}^n \mathbf{w}_m^T \mathbf{w}_m + C \sum_{i=1}^l \sum_{m \neq y_i} \xi_i^m, \quad (2.75)$$

under the constraints,

$$\mathbf{w}_{y_i}^T \boldsymbol{\phi}(\mathbf{x}_i) + b_{y_i} \geq \mathbf{w}_m^T \boldsymbol{\phi}(\mathbf{x}_i) + b_m + 2 - \xi_i^m,$$

$$\xi_i^m \geq 0, i = 1, \dots, l, m \in \{1, \dots, n\} \setminus y_i$$

The decision function is then,

$$\operatorname{argmax}_{m=1, \dots, n} (\mathbf{w}_m^T \boldsymbol{\phi}(\mathbf{x}_i) + b_m) \quad (2.76)$$

In reducing the classification to a single optimization problem this approach may also require fewer support vectors than a multi-class classification based on the combined use of many binary SVM (Hsu and Lin, 2002).

### 2.7.3 Limitations of Support Vector Machine

Support vector machines are a non-parametric supervised classification technique and, therefore, they do not assume any distribution of the training data as in conventional classifier such as the maximum-likelihood classifier. They, however, share all the problems associated with supervised classification. For instance, the accuracy of the classification is dependent upon the quality of the training data. Factors like number of training samples, number of discriminating variables affect the accuracy of support vector machines.

Support vector machines also depend upon kernel parameter choice, and class separability. The polynomial kernels, especially higher order kernels take more time than RBF kernels. Training the SVM to classify classes highly overlapping in feature space can take several times longer than training it to classify two separable classes.

Huang *et al.*, (2002) compared four classifiers MLC, DTC, ANN and SVM in land cover classification and found that SVM's were generally the most accurate and MLC the least. However the training speeds of the four classifiers were substantially different. They found that, for their data set, the MLC and DTC could be trained in a few minutes while ANN and SVM took hours and days respectively.

SVM's have been successfully used in optical character recognition, handwritten digit recognition (Vapnik, 1995), but they are relatively new to remote sensing community as compared to other classifiers like MLC, ANN, DTC (Huang *et al.*, 2002). They have however been found superior in classifying hyperspectral images acquired from air-borne visible/infrared imaging spectrometer (AVRIS) (Gualtieri and Cromp, 1998). The high dimensionality of hyperspectral data are challenging for traditional classifiers, due to the Hughes effect. The support vector machine does not suffer from this handicap and is thus suitable for use with hyperspectral data.

The use of support vector machines for data having fewer spectral bands should have a practical implication for land cover classification (Huang *et al.*, 2002), as the major

sensor systems like Thematic Mapper, Linear Image Self Scanning operate very few spectral bands as compared to hyperspectral data, and are generally employed for generating land cover maps. The performance of SVMs on data sets with very few variables should be investigated because data sets with fewer variables were not considered in previous studies (Cortes and Vapnik, 1995).

The potential of SVM reported in literature can be attributed to its ability to locate optimal separating hyperplane. The optimal separating hyperplane are expected to generalize accurately on unseen examples with fewer errors than any other separating hyperplane that might be found by the other classifiers (Huang *et al.*, 2002). The potential of SVM should be exploited in remote sensing, especially on data sets with fewer discriminating variables to generate land cover map.

## **2.8 Conclusions**

Accurate information on land cover possibly in map form is required for a plethora of applications, including land resource planning, studies of environmental change and bio-diversity conservation. Remote sensing has many advantages in generating thematic maps over conventional ground based surveys. Supervised classification methods in particular are the most popular and widely used technique for deriving thematic maps from remote sensing data. The accuracy of supervised classification, however, is often insufficient for operational applications. One of the important reasons for this is associated with the inputs in supervised classification, especially the training data.

There are a number of approaches available for supervised classification. MLC is the most popular parametric classifier but works under the assumption that the training data are normally distributed. Key distribution free methods with very different approaches are ANN, DT and SVM.

The general requirement of training data is that they should be able to characterize the classes under investigation. SVM have recently attracted the attention of the remote



sensing community. A key attraction of the SVM based approach to classification is that it seeks to fit an optimal hyperplane between the classes and since it uses only the training samples that lie at the edge of the class distributions in feature space it may require only a small training sample. Studies have shown that SVM results in higher accuracy as compared to other classifiers like ANN, DT and MLC.

This inherent property of the classifier to use only border training data provides an opportunity to review ground data collection policy. The analysis in this thesis, therefore, aims to evaluate SVM and will focus on:

- a) Relative evaluation of SVM with respect to MLC, DT and ANN by comparing the effect of training set size on classification accuracy.
- b) Assess ability to identify most useful training patterns (support vectors) in SVM classification.
- c) Intelligently reducing the requirement of training data by collecting training data of agricultural classes that acts as support vectors directly from field.
- d) Intelligently reducing the requirement of training data by relating support vectors with ancillary data.
- e) Reducing the requirement of training data if the concern is to map accurately only one class from the many land cover classes available in the study area.

# CHAPTER 3 - Relative Evaluation of Multi-Class Image Classification by SVM

## 3.1 Introduction

Supervised classification method uses training data to train the classifiers. The training data may be of unequal importance in an image classification and their importance may also depend upon the classification algorithm used. The study was undertaken with the aim to investigate the effect of training set size on classification accuracy with respect to a series of supervised classifiers. Both parametric and non-parametric classification algorithms were used. Discriminant analysis (DA) which is similar in approach to maximum-likelihood classification (MLC), the most popular parametric classifier was used as a benchmark. Key non-parametric classifiers evaluated were artificial neural network (ANN), decision tree (DT) and support vector machine (SVM).

Section 3.1.1 describes the study area and data used in the pilot study. The methodology is described in section 3.1.2 with results in 3.1.4.

### 3.1.1 Study Area and Data Used

Airborne thematic mapper (ATM) imagery with a spatial resolution of approximately 5 m acquired by a Daedalus 1268 sensor on 15 July, 1986 for a flat region of agricultural land near the village of Feltwell, U.K., were used. The data set comprised three bands (4, 6, and 9) out of possible twelve bands after feature reduction. The feature reduction was carried on the data set in an earlier study (Arora and Foody, 1997) and it was apparent that the data in all 11 bands was not required but three wavebands identified as providing the greatest level of inter class separability were selected for the analyses. The three waveband combination 4, 6 and 9 selected for analyses corresponds to electromagnetic spectrum range of 0.60-0.63  $\mu\text{m}$ , 0.69-0.75  $\mu\text{m}$  and 1.55-1.75  $\mu\text{m}$  respectively.

The ground data used, comprised a crop map produced by conventional field survey around the time of ATM data acquisition. Six agricultural classes namely; sugar beet, wheat, barley, carrot, potatoes and grass were mainly grown in the study area and were the focus of this study. The sugar beet and wheat classes were noticeably abundant in the study area.

### 3.1.1.1 Characteristics of Training Data

The training data were selected randomly from the study area. Training data comprised 100 pixels for each class.

The spectral distribution of the classes comprising the digital numbers (DN) in the three bands shows that the class sugar beet, wheat and barley were highly overlapping in the feature space (Figure 3.1). This was also confirmed in the summary description statistics of the training data (Table 3.1).

The training data were close to normal but multi-modal (Figure 3.2) and, therefore, MLC should not be used.

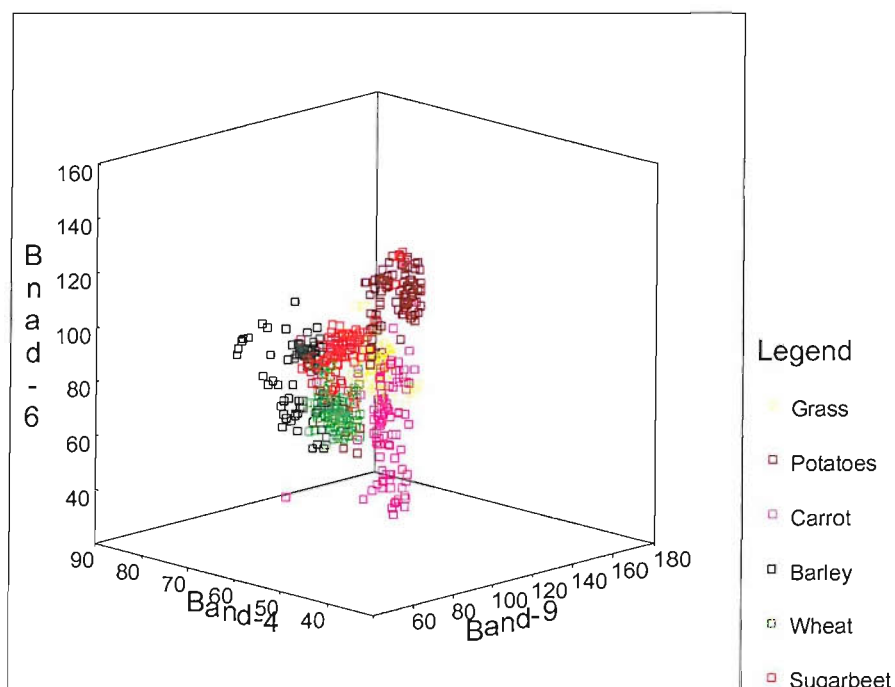


Figure 3.1: Location of training data in feature space.

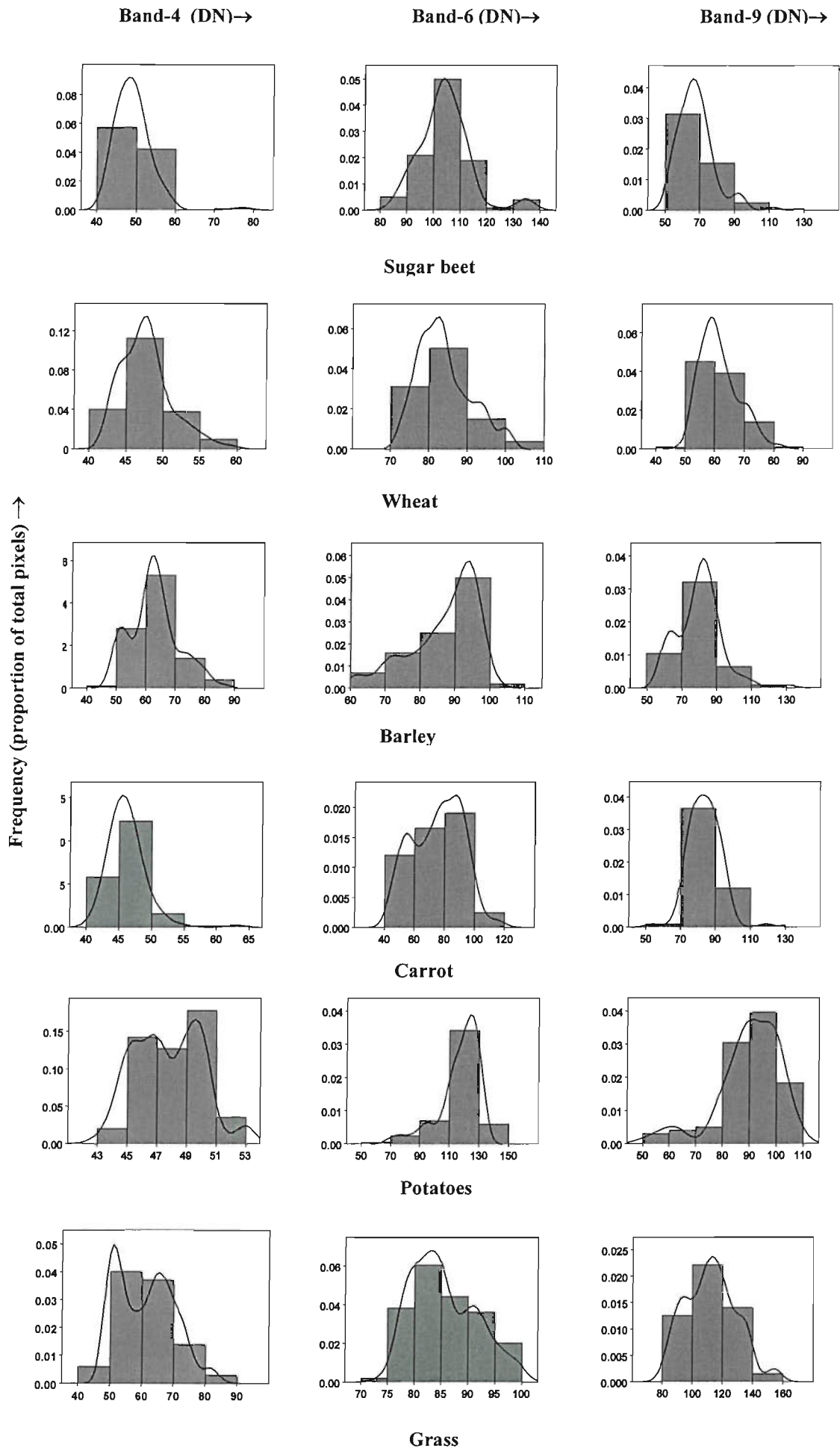


Figure 3.2: Histograms of training data for the six classes in bands 4, 6, 9. The solid lines show the smoothed histograms.

Class	Band-4 (DN)				Band-6 (DN)				Band-9 (DN)			
	Min	Max	Mean	Sd	Min	Max	Mean	Sd	Min	Max	Mean	Sd
<b>Sugar beet</b>	43	77	49.10	4.6	85	136	104.22	9.55	52	112	67.69	10.48
<b>Wheat</b>	42	59	47.71	3.39	73	101	83.88	6.82	49	82	61.43	6.48
<b>Barley</b>	49	86	62.76	8.16	61	101	86.80	9.74	60	130	80.11	12.38
<b>Carrot</b>	41	63	46.19	2.95	44	117	75.33	16.71	54	119	83.61	8.63
<b>Potato</b>	43	53	47.73	2.26	69	133	116.47	13.90	50	109	89.82	11.81
<b>Grass</b>	49	83	61.21	8.92	74	100	85.44	6.04	83	155	112.6	16.51

Table 3.1: Statistics of training data showing minimum (MIN), maximum (Max), Mean and standard deviation (sd) of digital numbers of training data of the six classes in the three bands.

### 3.1.2 Methodology

There are many factors that affect the accuracy of a classifier. This study examined the effect of training set size, classification algorithms used and testing set size on the classification accuracy. The variables studied are tabulated in Table 3.2 and detailed under section 3.1.2.1 to 3.1.2.3.

Variables	Scenarios investigated
Training set size	15 pixels per-class 30 pixels per-class 45 pixels per-class 60 pixels per-class 75 pixels per-class 90 pixels per-class 100 pixels per-class
Classification algorithms	Discriminant analysis Artificial neural network Decision tree Support vector machine
Testing set size	The classifiers were tested on testing sets comprising of two sizes (Group A and B) Group-A Comprises all the available pixels for testing Sugar beet 97 pixels Wheat 96 pixels Barley 51 pixels Carrot 31 pixels Potatoes 26 pixels Grass 17 pixels Group-B 17 pixels per class

Table 3.2: Variables studied in the study.

### **3.1.2.1 Training Set**

There can be a number of combinations of training set size that can be constructed from the available training data to evaluate the effect of training set size on classification accuracy. The training set size was defined by the number of bands used, three in the present study (*e.g.*,  $3 \times 5 = 15$  pixels/class). The training sets, therefore, comprised of 15, 30, 45, 60, 75 and 90 randomly selected pixels per-class (multiples of 3, the number of bands used) as also all the 100 pixels available for training (Table 3.2). These pixels were selected randomly from the total set of pixels available for training, which comprised of 100 pixels for each class. Since the results of a classification may be highly dependent on the specific sample of pixels selected, for each size of training set, except that using all 100 pixels available for each class, five independent samples were derived without replacement from the available training data (Table 3.3).

### **3.1.2.2 Algorithms Used**

DA is a conventional probabilistic classifier that like the maximum likelihood classifier allocates each case to the class with which it has the highest posterior probability of membership. As a basic probabilistic classifier, the discriminant analysis results provide a benchmark against which the relative accuracy of the other classifications may be assessed.

The ANN used was a basic multi-layer perceptron. The network's architecture and algorithm parameters were defined from an evaluation of several hundreds of candidate networks.

A DT can be defined as a classification procedure that recursively partitions data into smaller subsets on the basis of tests or thresholds at each node in the tree. The decision tree algorithm used the gain ratio to split nodes and the pessimistic error rate in tree pruning.

The SVM used was a one shot multi-class classifier. The advantage of multi-class support vector machine lies in the fact that all the classes gets classified in one step, unlike binary support vector machine which can work only on two classes at a time and therefore needs a number of such binary classifications (depending upon number of classes) for classifying all the classes under consideration (Gualtieri and Cromp, 1998). In reducing the classification to a single optimization problem this approach may also require fewer support vectors than a multi-class classification based on the combined use of many binary SVMs (Hsu and Lin, 2002).

Key studies reported in the remote sensing literature have used binary support vector machines only (Gualtieri and Cromp, 1998). The present study, therefore, is one of the first in the use of multi-class support vector machine in classification problem of remote sensing data.

The four classifiers described above were chosen for analysis as they differed markedly in their basis for class allocation and expected dependency on training set.

### **3.1.2.3 Testing Set**

The testing set comprised variable number of pixels to estimate the accuracy of the classification as detailed in Table 3.2. The variable number of testing pixels can be attributed to the spatial abundance of the crops in the study area. Sugar beet and wheat were noticeably abundant and therefore had greater representation in the testing set as compared to other crops.

The direct comparison of matrices can be difficult in case of variable number of testing pixels. In such instances, the error matrices can be normalized. Examination of the normalized matrices affords very convenient comparison. However, in some instances normalized values are so small that they are neglected with the notion that they do not alter the interpretation. Apart, it is also difficult to derive original matrix from normalized version and therefore should not be attempted.

An additional set comprising equal number of pixels for all the classes were, therefore, generated. Grass had the least number of pixels (17 pixels) for testing the accuracy (Table 3.2) out of all the six classes. So, a testing set comprising 17 pixels per class were generated. Since the results of a classification may be highly dependent on the specific sample of pixels selected, for each size of testing set, five independent samples were derived without replacement from the available testing data.

Two cases were analysed, hereafter referred to as case A and case B for simplicity (Table 3.3), generated from combinations of training and testing set.

In case 'A', various combination of training data size (five per size) (Table 3.3) were used and accuracy tested on testing data comprising of all the pixels acquired for testing. The combination out of the five per each training size (Table 3.3) that gave the median accuracy was selected for comparison to avoid extreme results.

In case 'B', the combination of training data was same as in case 'A', however, the testing data comprised of five different combinations of 17 pixels per class chosen randomly with replacement. Each combination of testing size was used to test the four classifiers (Table 3.2) trained by corresponding training set. For example testing set labelled as N1 was used to test the classifier trained only by corresponding training set N1 (15N1 in case of training set of 15 pixels per class).

However for comparing the effect of various variables on classification accuracy, the testing set (out of the five for any particular testing size) which gave the median overall accuracy was chosen to represent the outcome for that particular training/testing size combination to avoid extreme results *i.e.*, out of the five repetitions on testing set, the one in the middle of the overall accuracy hierarchy was selected for comparison.



Case	Training set		Testing set	
	Pixels per class	five combinations (N1 to N5)		
A	15	15N1, 15N2, 15N3, 15N4, 15N5	-all available testing data	
	30	30N1, 30N2, 30N3, 30N4, 30N5	sugar beet	97
	45	45N1, 45N2, 45N3, 45N4, 45N5	wheat	96
	60	60N1, 60N2, 60N3, 60N4, 60N5	barley	51
	75	75N1, 75N2, 75N3, 75N4, 75N5	carrot	33
	90	90N1, 90N2, 90N3, 90N4, 90N5	potato	26
	100	100 (all available training set)	grass	17
B	Same as above as in case A		17 pixels per class (five combinations N1 to N5 corresponding to five combinations of training set)	

Table 3.3: Combination of training and testing set size for analysis. The training and testing set size has been abbreviated with prefix to N as the number of pixels and suffix as the iteration number.

### 3.1.3 Accuracy Assessment

Fundamental to this work is the comparison of classification accuracy statements.

The evaluation and comparison of classifications were based on the overall accuracy.

Apart from comparing the overall accuracy, the statistical significance of differences in classification accuracy was also evaluated.

The testing set, especially for case-A analysis, was same for all the analysis, as such, the statistical significance of differences in classification accuracy was evaluated based on related samples using M<sup>c</sup>Nemar test (Foody, 2004). The test is non-parametric and is based on a 2x2 dimensional error matrix (Table 3.4). The error matrix generated as a result of classification can be collapsed into the required 2x2 dimensional error matrix by focusing only into correct and incorrect class allocations.

		Classification 2 ↓		
		Allocation	Correct	Incorrect
Classification 1 →	Correct	$f_{11}$	$f_{12}$	
	Incorrect	$f_{21}$	$f_{22}$	
Σ				

Table 3.4: 2x2 error matrix to calculate the statistical significance of differences in classification accuracy based on M<sup>c</sup>Nemar test for related samples.

The M<sup>c</sup>Nemar test is given by:

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \quad (3.1)$$

For case-B, the testing data size was 17 pixels per class generated randomly from the overall testing set. There were in all five combinations of testing set for each training set size as 17N1, 17N2, 17N3, 17N4 and 17N5 (Table 3.2). In some of the cases, the testing set was same for the two classifiers to be compared; as such M<sup>c</sup>Nemar test as detailed above was evaluated. In case, where the testing set were different for the two classifiers to be compared, the statistical significance of differences between the outcomes were based on independent samples (Foody, 2004) and evaluated as;

$$Z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (3.2)$$

Where,  $x_1$  and  $x_2$  are correctly allocated cases in the two independent samples of sizes  $n_1$  and  $n_2$  respectively and  $p = \frac{x_1 + x_2}{n_1 + n_2}$ .

### 3.1.4 Results

The overall classification accuracy obtained by different classifiers were compared to understand the effect of training set size on classification accuracy. Section 3.1.4.1 to 3.1.4.4 focuses on the overall accuracy obtained by employing the four classifiers DA, ANN, DT and SVM respectively on different training set sizes. The comparison of the results of the four classifiers is followed in section 3.1.5 with conclusions in 3.1.6. However, the error matrices for each classification are given in Appendix (Tables A.1 to A.56) to derive the various indices of error matrices if desired.

### 3.1.4.1 Discriminant Analysis (DA)

#### Case A (all available testing data)

Table 3.5 shows the effect of training set size on classification accuracy of individual classes and overall accuracy on the test set for case A analysis (using all the available test set). The table shows that wheat class had a very large accuracy of 94.8 % even when the training set size was only 15 pixels per class. This shows that wheat class was spectrally discriminable from other classes.

The overall accuracy was positively related with training set size (Table 3.5) and is compatible with the results reported in literature (Huang *et al.*, 2002). The difference in accuracy between the classifications trained on the largest and smallest training set size was 2.20%, and this difference was statistically significant at 95% confidence level (Table 3.15).

Class	Training set size						
	15	30	45	60	75	90	100
Sugar beet	85.60	85.60	88.70	89.70	88.70	88.70	89.70
Wheat	94.80	93.80	93.80	94.80	93.80	94.80	93.80
Barley	84.30	86.30	84.30	88.20	88.20	88.20	88.20
Carrot	78.80	87.90	93.90	84.80	87.90	87.90	87.90
Potato	92.30	88.50	88.50	88.50	88.50	88.50	88.50
Grass	82.40	82.40	88.20	82.40	82.40	82.40	82.40
Overall accuracy	87.80	88.40	90.00	90.00	89.70	90.00	90.00

Table 3.5: Overall and class wise accuracy (%) on testing data using discriminant analysis for case A (all available testing data) analyses.

#### Case B (17 pixels per class)

It is evident from Table 3.6 that the overall accuracy was in general positively related with training set size. The difference in the overall accuracy between the classifications trained on the largest and smallest training set size was 3.90%, but the difference was not statistically significant at 95 % confidence level (Table 3.15).

Wheat class had the highest accuracy as compared to other classes in 4 out of 6 training sizes analysed. Sugar beet had a very low accuracy of 64.7 % and 76.5 % when

Class	Training set size						
	15	30	45	60	75	90	100
Sugar beet	64.70	94.10	76.50	100.00	82.40	94.10	100.00
Wheat	94.10	88.20	100.00	94.10	100.00	100.00	94.10
Barley	82.40	82.40	88.20	76.50	88.20	76.50	76.50
Carrot	94.10	82.40	94.10	94.10	82.40	82.40	94.10
Potato	88.20	88.20	88.20	88.20	82.40	88.20	82.40
Grass	82.40	82.40	82.40	82.40	82.40	82.40	82.40
Overall accuracy	84.30	86.30	88.20	89.20	86.30	87.30	88.20

Table 3.6: Overall and class wise accuracy (%) on testing data using discriminant analysis for case B (17 pixels per class in testing data) analyses.

training size was 15 pixels and 45 pixels per class respectively. However the accuracy of class grass was 82.4 % irrespective of training set size.

The comparison of the two cases, A (unequal testing data set) and B (equal testing set with 17 pixels per class) shows that overall accuracy was generally positively related with training set size. Wheat class was the most discriminable class out of all the six in both the cases for most of the training set size.

### 3.1.4.2 Artificial Neural Network

#### Case A (all available testing data)

Table 3.7 shows that there was a marginal increase in overall accuracy when training size increased from 15 to 100 pixels per class. This increase was, however, not significant at 95 % confidence level (Table 3.15). The increase in classification accuracy was in general positively related with training set size. The class grass had the highest individual class accuracy (100%) when training set size was 45 pixels/class or more.

Class	Training set size						
	15	30	45	60	75	90	100
Sugar beet	86.59	85.56	87.62	89.69	90.72	88.65	92.78
Wheat	92.70	94.79	90.62	91.66	92.70	90.62	87.50
Barley	90.19	84.31	96.07	92.15	94.11	98.03	96.07
Carrot	93.93	96.96	96.96	96.96	96.96	96.96	93.93
Potato	88.46	88.46	88.46	84.61	88.46	88.46	88.46
Grass	82.35	88.23	100.00	100.00	100.00	100.00	100.00
Overall accuracy	89.68	89.68	91.56	91.56	92.81	92.18	91.88

Table 3.7: Overall and class wise accuracy (%) on testing data using artificial neural network for case A (all available testing data) analyses.

### Case B (17 pixels per class for testing data)

There was no noticeable trend of overall accuracy related with training set size (Table 3.8). However, barley and carrot classes showed the highest accuracy amongst all the classes. The class sugar beet had a very low accuracy of 64.7 % when training set size was 15 pixels/class. The difference in the overall accuracy between the classifications trained on the largest and smallest training set size was 5.88 %, but the difference was not statistically significant at 95 % confidence level (Table 3.15).

Class	Training set size						
	15	30	45	60	75	90	100
Sugar beet	64.70	82.35	94.11	88.23	94.11	88.23	88.23
Wheat	82.35	94.11	94.11	88.23	82.35	100.00	88.23
Barley	94.11	100.00	94.11	100.00	100.00	100.00	100.00
Carrot	100.00	94.11	100.00	100.00	100.00	100.00	100.00
Potato	88.23	82.35	88.23	88.23	94.11	88.23	88.23
Grass	100.00	100.00	100.00	100.00	100.00	94.11	100.00
Overall accuracy	88.23	92.15	95.09	94.11	95.09	95.09	94.11

Table 3.8: Overall and class wise accuracy (%) on testing data using artificial neural network for case B (17 pixels per class in testing data) analyses.

The comparison of the two cases, A (unequal testing data set) and B (equal testing set with 17 pixels per class) shows that overall accuracy in both the cases increased with training set size but the increases in accuracy were not monotonic with increase in training set size. In addition, the accuracy for case 'B' with a smaller testing set had generally greater overall accuracy as compared to case 'A' with a larger testing set.

#### 3.1.4.3 Decision Tree

##### Case A (all available testing data)

Table 3.9 shows that the increase in overall accuracy was positively related with training set size. The difference in accuracy between the classifications trained on the largest and smallest training set size was 23.13 % and this difference in accuracy was significant at 95 % confidence level (Table 3.15).

Grass and carrot classes had the highest accuracy of 100 % when the training set was 100 pixels per class.

Class	Training set size						
	15	30	45	60	75	90	100
Sugar beet	87.62	81.44	82.47	90.72	93.81	85.56	91.75
Wheat	71.87	76.04	86.45	78.12	71.87	82.29	82.29
Barley	86.27	80.39	76.47	88.23	94.11	94.11	94.11
Carrot	48.48	87.87	87.87	93.93	100.00	93.93	100.00
Potato	73.07	88.46	84.61	88.46	84.61	88.46	88.46
Grass	82.35	100.00	100.00	76.47	94.11	94.11	100.00
Overall accuracy	77.18	81.87	84.37	85.94	87.19	87.50	90.31

Table 3.9: Overall and class wise accuracy (%) on testing data using decision tree algorithm for case A (all available testing data) analyses.

A very low percentage of pixels (48.48 %) of carrot could be correctly classified when the training size was 15 pixels per class, but the carrot class could be classified with 100 % accuracy when the classifier was trained with the largest training set size (100 pixels/class).

### Case B (17 pixels per class for testing data)

The examination of Table 3.10 shows that the increase in overall accuracy of classification was positively related with training set size. The difference in accuracy between the classifications trained on the largest and smallest training set size was 15.69 % and this difference was statistically significant at 95 % confidence level (Table 3.15).

Typically all the classes showed higher accuracies with larger training set. For example, classes sugar beet, carrot and grass obtained 100 % accuracy when the training size was 100 pixels per class.

Class	Training set size						
	15	30	45	60	75	90	100
Sugar beet	94.11	94.11	82.35	88.23	82.35	82.35	100.00
Wheat	82.35	64.70	82.35	88.23	70.58	82.35	88.23
Barley	70.58	94.11	82.35	94.11	94.11	88.23	88.23
Carrot	70.58	82.35	100.00	94.11	100.00	100.00	100.00
Potato	82.35	82.35	88.23	82.35	100.00	88.23	88.23
Grass	70.58	94.11	82.35	82.35	100.00	100.00	100.00
Overall accuracy	78.43	85.29	86.27	88.24	91.18	90.20	94.12

Table 3.10: Overall and class wise accuracy (%) on testing data using decision tree for case B (17 pixels per class in testing data) analyses.

The class sugar beet could be correctly classified to the tune of 94.11 % even when the training set size was of 15 pixels per class.

In both the cases, A and B overall accuracy was positively related with training set size and classes carrot and grass could be classified with 100 % accuracy. In addition, the accuracy for case ‘B’ with a smaller testing set had greater overall accuracy as compared to case ‘A’ with a larger testing set.

### 3.1.4.4 Support Vector Machine

#### Case A (all available testing data)

Table 3.11 shows that the increase in overall accuracy was positively related with training set size. The difference in accuracy between the classifications trained on the largest and smallest training set size was 6.25 % and the difference was statistically significant at 95 % confidence level (Table 3.15). Sugar beet and wheat classes had very high accuracy even when the training set size was as low as 15 pixels per class.

Typically, higher accuracies of the classes were associated with larger training set as a larger training set had more chances of including appropriate support vectors to generate the optimal boundaries between the classes.

Class	Training set size						
	15	30	45	60	75	90	100
Sugar beet	92.78	88.65	91.75	87.62	91.75	92.78	91.75
Wheat	94.79	92.70	85.41	93.75	90.62	89.58	91.66
Barley	76.47	92.15	96.07	92.15	94.11	98.03	96.07
Carrot	66.67	90.90	90.90	96.96	100.00	93.93	100.00
Potato	88.46	88.46	92.30	88.46	88.46	88.46	92.30
Grass	88.23	94.11	100.00	94.11	100.00	100.00	100.00
Overall accuracy	87.50	90.94	90.93	91.56	92.81	92.81	93.75

Table 3.11: Overall and class wise accuracy (%) on testing data using support vector machine for case A (all available testing data) analyses.

### Case B (17 pixels per class for testing data)

Table 3.12 shows that the increase in overall accuracy was generally positively related with training size. The highest overall accuracy of 96.07 % was obtained when the training set was 30 pixels per class. This can be attributed to the fact that the training data included the appropriate support vectors for all the classes, resulting in higher overall accuracy even with smaller training sets.

Class	Training set size						
	15	30	45	60	75	90	100
Sugar beet	70.58	100.00	82.35	76.47	88.23	94.11	94.11
Wheat	76.47	94.11	82.35	100.00	88.23	94.11	94.11
Barley	94.11	100.00	94.11	94.11	100.00	94.11	94.11
Carrot	94.11	100.00	88.23	100.00	94.11	100.00	100.00
Potato	88.23	94.11	100.00	94.11	88.23	82.35	88.23
Grass	82.35	88.23	100.00	94.11	100.00	100.00	100.00
Overall accuracy	84.31	96.07	91.17	93.13	93.13	94.11	95.09

Figure 3.12: Overall and class wise accuracy (%) on testing data using support vector machine for case B (17 pixels per class for testing data) analyses.

The difference in accuracy between the classifications trained on the largest and smallest training set size was 9.80 % and the difference was statistically significant at 95 % confidence level (Table 3.15).

The comparison of the two cases, A (unequal testing data set) and B (equal testing set with 17 pixels per class) shows that overall accuracy in both the cases generally increased with training set size. The results also show that higher accuracies can be obtained with smaller training set size in SVM if appropriate support vectors are available in the training data. In addition, the accuracy for case 'B' with a smaller testing set had generally greater overall accuracy as compared to case 'A' with a larger testing set.



### 3.1.5 Results and Discussion

#### Case A Analysis (all available testing data)

From the range of classifications undertaken, the highest accuracy, 93.75%, was obtained from the SVM trained with 100 cases of each class (Table 3.13). Moreover, this classification was significantly more accurate than that derived from the decision tree and discriminant analysis ( $p < 0.05$ ) (Table 3.16).

With all the four classification methods it was apparent that classification accuracy was positively related to training set size (Table 3.13). For the SVM based classifications, the difference in accuracy between the classifications trained on the largest and smallest training sets was 6.25%. Classification by the decision tree algorithm appeared to be most sensitive to training set size, with the accuracy increasing from 77.18% to 90.31% as the training set increased from containing 15 to 100 cases of each class. For all classifiers, except the artificial neural network, the difference in the accuracy of the classifications derived with the use of the largest and smallest training sets was statistically significant ( $p < 0.05$ ) (Table 3.15). At each training set size, the SVM was also relatively accurate and often the most accurate classifier, with accuracies often statistically different from those derived from the other classifiers (Table 3.16).

The effect of variation in training set size on the accuracy of the classifications by the four classifiers is compatible with results reported in the literature (Huang *et al.*, 2002). The sensitivity of the SVM classification to the nature of the sample is also evident in Table 3.13 which shows that the five SVM classifications based on a training set comprising 15 cases of each class were very varied in accuracy. Thus, while the SVM classification may be based on the information provided by a small number of training sites, forming the support vectors (Table 3.13), a large training sample may still be required to ensure that appropriate support vectors are available.

Although the four classifiers were able to classify the data very accurately, each  $>90\%$  accurate (Table 3.13) for the analyses based on the largest training set size, there

were some important differences. It was apparent, for example, that the classifiers varied in their ability to distinguish between specific classes and the accuracy with which individual classes were classified differed markedly (Appendix).

### **Case B Analysis (17 pixels per class for testing data)**

From the range of classifications undertaken, the highest accuracy, 96.07%, was obtained from the SVM trained with 30 cases of each class (Table 3.14). Moreover, this classification was significantly more accurate than that derived from the decision tree and discriminant analysis ( $p < 0.05$ ) (Table 3.17).

With all the four classification methods it was apparent that classification accuracy was in general positively related to training set size (Table 3.14). For the SVM based classifications, the difference in accuracy between the classifications trained on the largest and smallest training sets was 9.80%. Classification by the decision tree algorithm appeared to be most sensitive to training set size, with the accuracy increasing from 78.43% to 94.12% as the training set increased from containing 15 to 100 cases of each class. The difference in the accuracy of the classifications derived with the use of the largest and smallest training sets was statistically significant ( $p < 0.05$ ) for decision tree and support vector machine classifiers (Table 3.15). At each training set size, the SVM was also relatively accurate.

The sensitivity of the SVM classification to the nature of the sample is also evident in Table 3.14 which shows that the five SVM classifications based on a training set comprising 15 cases of each class were very varied in accuracy. Thus, while the SVM classification may be based on the information provided by a small number of training sites, forming the support vectors, a large training sample may still be required to ensure that appropriate support vectors are available.

Although the four classifiers were able to classify the data very accurately, each >88% accurate (Table 3.14) for the analyses based on the largest training set size, there

were some important differences. It was apparent, for example, that the classifiers varied in their ability to distinguish between specific classes and the accuracy with which individual classes were classified differed markedly (Appendix).

For both, case A and case B analysis, the results are data-specific but they do indicate the value of multi-class SVM classification. The SVM classifications were generally more accurate than ANN, DT and DA and, with the analyses constrained to a single optimization problem, requiring fewer support vectors as compared to binary analyses (Hsu and Lin, 2002). Classification accuracy was, however, a function of training set size and the potential of using small training sets in SVM based classification will require a means of intelligent training data acquisition.

### **3.1.6 Conclusions**

In this chapter classification accuracy on testing set resulting from training four different classifiers by differentially sized training set were compared. The results can be summarized as:

- SVM have considerable potential for the classification of remotely sensed data.
- It has been demonstrated here that a single multi-class SVM classification may be undertaken and used to derive very accurate classifications.
- In general, the SVM classifications were more accurate than comparable classifications derived with the use of the other classification techniques.
- The accuracy of the classifications produced from all of the classifiers for both Case A and Case B analyses were in general positively related to training set size, with the accuracy of the classifications derived from three of the classifiers increasing significantly as the training set size increased from 15 to 100 cases per-class for case A analysis (Table 3.16) and two for case B analysis (Table 3.17).

- Although a SVM classification is effectively based on a small number of training sites a large training sample may still, therefore, be required to ensure that appropriate support vectors (training data) are available.
- The sensitivity of the accuracy of the SVM classifications to training set size indicates the need for the training set to include the cases of the classes that lies on the border of the spectral distribution of the classes in feature space essentially those that faces the cases of other classes to yield appropriate support vectors.
- While a large training sample may not be required in order to estimate a statistical distribution it is, however, critical for the training sample to include useful support vectors and, unless some intelligent training data acquisition process is followed, these are more likely to be found from a large rather than small sample.

Training size	Discriminant analysis		Decision Tree		Artificial Neural Network			Support Vector Machine			
	Training	Testing	Training	Testing	Training	Testing	Parameters	Training	Testing	RBF (gamma)	No of support vectors (nsv)
15N1	85.60	87.50	94.44	81.25	87.77	89.68	MLP, 13	98.88	88.12	0.03	83
15N2	81.10	87.80	93.33	<b>77.18</b>	85.55	89.38	MLP,8	86.67	<b>87.50</b>	0.005	74
15N3	75.60	87.80	94.44	81.56	85.55	89.68	MLP,4	90.00	89.37	0.005	71
15N4	93.30	88.40	96.67	76.87	93.33	<b>89.68</b>	MLP,7	96.67	87.18	0.01	58
15N5	82.20	<b>87.80</b>	91.11	75.31	85.55	88.75	MLP,5	82.22	84.38	0.001	61
30N1	83.90	<b>88.40</b>	95.56	78.75	85.55	89.68	MLP,8	89.44	90.00	0.005	101
30N2	81.70	87.80	96.11	84.06	87.77	89.37	MLP,7	93.89	<b>90.94</b>	0.01	118
30N3	82.20	87.80	95.56	83.75	90.55	91.25	MLP,12	98.33	90.00	0.03	142
30N4	82.20	89.10	94.44	79.68	91.66	<b>89.68</b>	MLP,12	97.22	91.87	0.03	148
30N5	77.20	89.70	93.89	<b>81.87</b>	90.00	91.25	MLP,19	97.22	90.00	0.03	154
45N1	83.00	90.00	93.70	<b>84.37</b>	89.25	91.56	MLP,16	96.66	90.93	0.03	194
45N2	84.80	<b>90.00</b>	95.56	85.62	89.25	90.93	MLP,3	97.40	90.31	0.03	211
45N3	83.00	90.00	93.70	84.37	89.25	92.81	MLP,24	96.66	<b>90.93</b>	0.03	194
45N4	82.20	90.60	96.30	85.62	91.11	<b>91.56</b>	MLP,13	90.37	90.93	0.005	151
45N5	83.00	89.10	93.33	83.43	89.62	90.93	MLP,15	86.51	90.93	0.01	175
60N1	81.90	<b>90.00</b>	94.17	83.75	90.83	92.18	MLP,25	94.72	91.25	0.03	259
60N2	82.80	90.30	95.00	<b>85.94</b>	90.27	91.87	MLP,15	97.22	92.50	0.05	290
60N3	81.10	90.00	94.17	87.19	90.00	90.93	MLP,11	94.17	92.19	0.03	257
60N4	83.90	89.40	95.56	88.44	91.38	91.25	MLP,16	97.50	89.69	0.03	234
60N5	81.10	90.30	95.83	85.62	90.00	<b>91.56</b>	MLP,25	86.67	<b>91.56</b>	0.005	183
75N1	82.00	90.30	93.11	84.06	89.77	92.18	MLP,16	94.44	92.81	0.03	296
75N2	83.80	<b>89.70</b>	95.78	88.75	92.00	93.43	MLP,25	98.88	<b>92.81</b>	0.08	226
75N3	82.00	89.40	95.33	90.94	92.00	<b>92.81</b>	MLP,15	91.78	92.81	0.01	235
75N4	80.90	89.70	94.67	<b>87.19</b>	89.77	91.56	MLP,25	94.89	92.50	0.03	284
75N5	81.10	88.80	94.22	85.94	88.44	92.81	MLP,9	94.89	92.50	0.03	227
90N1	82.40	89.70	94.26	87.19	91.11	92.81	MLP,14	94.81	92.81	0.03	332
90N2	81.10	89.70	94.63	87.81	90.92	91.56	MLP,14	94.63	93.12	0.03	333
90N3	80.90	90.30	95.00	89.37	88.88	91.56	MLP,16	94.44	92.50	0.03	332
90N4	81.90	90.30	94.81	87.19	90.55	92.5	MLP,25	94.81	92.50	0.03	393
90N5	82.40	<b>90.00</b>	94.26	<b>87.5</b>	90.74	<b>92.18</b>	MLP,16	95.00	<b>92.81</b>	0.03	331
ALL(100)	81.80	<b>90.00</b>	95.17	<b>90.31</b>	91.66	<b>91.88</b>	MLP,25	93.33	<b>93.75</b>	0.02	302

Table 3.13: Overall accuracy for case A analysis. Bold values indicate the median overall accuracy obtained for the five combinations of training data and is used for comparison with accuracies obtained with different training set size combinations to avoid extreme results.

Training size	Discriminant Analysis		Decision Tree		Artificial Neural Network			Support Vector Machine			
	Training	Testing	Training	Testing	Training	Testing	Parameters	Training	Testing	Parameter RBF (gamma)	nsv
15N1	85.60	87.30	94.44	<b>78.43</b>	94.44	93.13	MLP, 13	96.67	90.19	0.010	72
15N2	81.10	83.30	93.33	72.55	88.88	88.23	MLP,8	86.67	82.35	0.005	74
15N3	75.60	87.30	94.44	80.39	85.55	92.15	MLP,4	90.00	88.23	0.005	71
15N4	93.30	<b>84.30</b>	96.67	79.41	93.33	<b>88.23</b>	MLP,7	96.67	<b>84.31</b>	0.010	58
15N5	82.20	84.33	91.11	77.45	88.88	88.23	MLP,5	93.33	84.31	0.010	72
30N1	83.90	89.20	95.56	91.20	90.55	96.07	MLP,8	99.44	98.03	0.030	144
30N2	81.70	84.30	96.11	86.27	91.66	<b>92.15</b>	MLP,7	93.89	89.21	0.010	118
30N3	82.20	<b>86.30</b>	95.56	<b>85.29</b>	91.66	93.13	MLP,12	98.33	89.21	0.030	142
30N4	82.20	84.30	94.44	79.41	92.77	89.21	MLP,12	97.22	<b>96.07</b>	0.030	148
30N5	77.20	87.30	93.89	82.35	91.11	90.19	MLP,19	97.22	96.07	0.030	154
45N1	83.00	86.30	93.70	82.35	91.17	92.59	MLP,16	96.66	<b>91.17</b>	0.030	194
45N2	84.80	89.20	95.56	<b>86.27</b>	94.44	<b>95.09</b>	MLP,3	90.37	87.25	0.005	146
45N3	83.00	84.30	93.70	86.27	90.74	92.15	MLP,24	96.66	93.13	0.030	194
45N4	82.20	93.10	96.30	94.12	91.85	96.07	MLP,13	92.96	91.17	0.010	162
45N5	83.00	<b>88.20</b>	93.33	86.27	90.55	96.07	MLP,15	94.44	94.11	0.030	207
60N1	81.90	<b>89.20</b>	94.17	91.20	92.22	93.13	MLP,25	94.72	93.13	0.030	259
60N2	82.80	90.20	95.00	85.29	92.77	96.07	MLP,15	88.06	94.11	0.005	185
60N3	81.10	89.20	94.17	<b>88.24</b>	87.50	94.11	MLP,11	99.44	<b>93.13</b>	0.100	330
60N4	83.90	88.20	95.56	91.18	94.16	<b>94.11</b>	MLP,16	97.50	92.15	0.030	234
60N5	81.10	92.20	95.83	84.31	85.55	96.07	MLP,25	86.67	93.13	0.005	183
75N1	82.00	86.30	93.11	86.27	90.44	94.11	MLP,16	91.56	93.13	0.010	229
75N2	83.80	89.20	95.78	<b>91.18</b>	89.33	95.09	MLP,25	89.77	95.09	0.005	209
75N3	82.00	<b>86.30</b>	95.33	91.18	92.00	91.17	MLP,15	95.78	<b>93.13</b>	0.030	305
75N4	80.90	90.20	94.67	91.20	89.11	95.09	MLP,25	90.67	93.13	0.010	227
75N5	81.10	86.30	94.22	89.22	91.33	<b>95.09</b>	MLP,9	94.89	93.13	0.030	284
90N1	82.40	86.30	94.26	<b>90.20</b>	93.14	<b>95.09</b>	MLP,14	94.81	95.09	0.03	332
90N2	81.10	<b>87.30</b>	94.63	89.22	86.66	94.11	MLP,14	94.63	94.11	0.03	333
90N3	80.90	88.20	95.00	93.14	89.81	96.07	MLP,16	94.44	<b>94.11</b>	0.03	332
90N4	81.90	91.20	94.81	93.34	91.66	97.05	MLP,25	91.29	96.07	0.01	254
90N5	82.40	86.30	94.26	87.25	90.74	94.11	MLP,16	91.66	92.15	0.01	264
100(17n1)	81.80	88.20	95.17	93.14	90.16	94.11	MLP,25	91.00	95.09	0.01	278
100(17n2)	81.80	87.30	95.17	<b>94.12</b>	90.50	93.13	MLP,11	94.50	<b>94.11</b>	0.03	348
100(17n3)	81.80	<b>88.20</b>	95.17	94.12	90.33	94.11	MLP, 10	94.50	94.11	0.03	348
100(17n4)	81.80	91.20	95.17	98.04	90.83	97.05	MLP, 16	94.50	96.07	0.03	348
100(17n5)	81.80	86.30	95.17	92.16	91.00	<b>94.11</b>	MLP, 15	91.00	94.11	0.01	278

Table 3.14: Overall accuracy for case B analysis. Bold values indicate the median overall accuracy obtained for the five combinations of training data and is used for comparison with accuracies obtained with different training set size combinations to avoid extreme results.

Classifiers	Z values	
	CASE A	CASE B
Discriminant analysis	<b>2.11</b>	0.81
Decision tree	<b>4.90</b>	<b>3.25</b>
Artificial neural network	1.28	1.48
Support vector machine	<b>3.24</b>	<b>2.25</b>

Table 3.15: Significance value (Z) of differences between accuracies of testing set obtained when the classifiers were trained with smallest training set size of 15 pixels and largest available size of 100 pixels per class at 95 % confidence level. Differences significant at the 95% confidence level ( $Z \geq 1.96$ ) are highlighted in bold with positive values indicating higher accuracy when training data was 100 pixels/class.

Training set size	SVM v DA	SVM v DT	SVM v ANN	ANN v DA	ANN v DT	DT v DA
15N	-0.22	<b>4.09</b>	-1.35	1.60	<b>4.65</b>	<b>-3.95</b>
30N	-0.27	<b>3.84</b>	0.85	1.07	<b>3.15</b>	<b>-2.64</b>
45N	0.00	<b>4.58</b>	-0.40	1.14	<b>4.13</b>	<b>-3.28</b>
60N	1.21	<b>2.65</b>	0.00	1.21	<b>2.65</b>	-1.85
75N	1.62	<b>3.00</b>	0.00	<b>2.13</b>	<b>2.85</b>	-1.09
90N	1.56	<b>3.40</b>	0.44	1.70	<b>3.00</b>	-1.26
100N	<b>2.27</b>	<b>2.30</b>	1.50	1.18	0.96	0.16

Table 3.16: Comparison of classification accuracy statements. The classifications derived with each method (SVM = support vector machine, DA = discriminant analysis, DT = decision tree and ANN = artificial neural network) at each size of training set for case A (all testing set), defined by the number of cases of each class, were compared using a  $M^c$ Nemar test. Differences significant at the 95% confidence level ( $Z \geq 1.96$ ) are highlighted in bold with positive values indicating that the first named classifier had the higher accuracy.

Training set size	SVM v DA	SVM v DT	SVM v ANN	ANN v DA	ANN v DT	DT v DA
15N	0.00	1.08	-1.15	1.50	1.88	-1.08
30N	<b>2.47</b>	<b>2.64</b>	1.18	1.35	1.55	-0.24
45N	0.69	1.11	-1.10	1.77	<b>2.71</b>	-0.42
60N	0.99	1.41	-0.29	1.27	1.48	-0.22
75N	1.73	0.52	0.60	<b>2.17</b>	1.10	1.11
90N	1.69	1.04	0.31	<b>1.98</b>	<b>2.12</b>	0.66
100N	1.48	0.00	0.00	1.48	0.00	1.48

Table 3.17: Comparison of classification accuracy statements. The classifications derived with each method (SVM = support vector machine, DA = discriminant analysis, DT = decision tree and ANN = artificial neural network) at each size of training set for case B (testing set comprising of 17 pixels per class), defined by the number of cases of each class, were compared using a  $M^c$ Nemar test. Differences significant at the 95% confidence level ( $Z \geq 1.96$ ) are highlighted in bold with positive values indicating that the first named classifier had the higher accuracy.

# **CHAPTER 4 - Reducing Requirement of Training Data by Relating Support Vectors with Soil Type of Training Fields**

## **4.1 Introduction**

The potential of SVM was demonstrated from a series of analyses that classified land cover from imagery of Feltwell, Norfolk (chapter 3). The inspection of the number of support vectors used in each SVM classification indicated a potential to reduce training set size without any negative impact on classification accuracy. This requires a means to intelligently identify training samples. The present study focuses on one such approach, based on the use of ancillary data on soil type to direct training site acquisition. The study was undertaken with the intent to use information on soil type to focus on regions most likely to furnish support vectors. In this way, an accurate SVM classification may be undertaken using a small training set.

### **4.1.1 Data and Methods of Classification**

Imagery acquired by SPOT HRV with a spatial resolution of 20 m on 16 June, 1986 for a flat region of an agricultural land near the village of Feltwell in Eastern England was used. The data set comprised three bands corresponding to electromagnetic spectrum range of 0.50-0.59  $\mu\text{m}$  (Green band), 0.61-0.68  $\mu\text{m}$  (Red band) and 0.79-0.89  $\mu\text{m}$  (Near-Infrared band). Near the time of the data acquisition a crop map for the test site was constructed by conventional field survey methods. This map identified the single crop type planted in the fields that were very large in comparison to the spatial resolution of the imagery.

Most of the test site had been planted to winter wheat and barley crops. The study area comprised of two types of soil, sand and peat. The sand soil comprised of humic gleyic rendzinas, brown rendzinas, typical brown sands, typical humic-sandy gley soils,



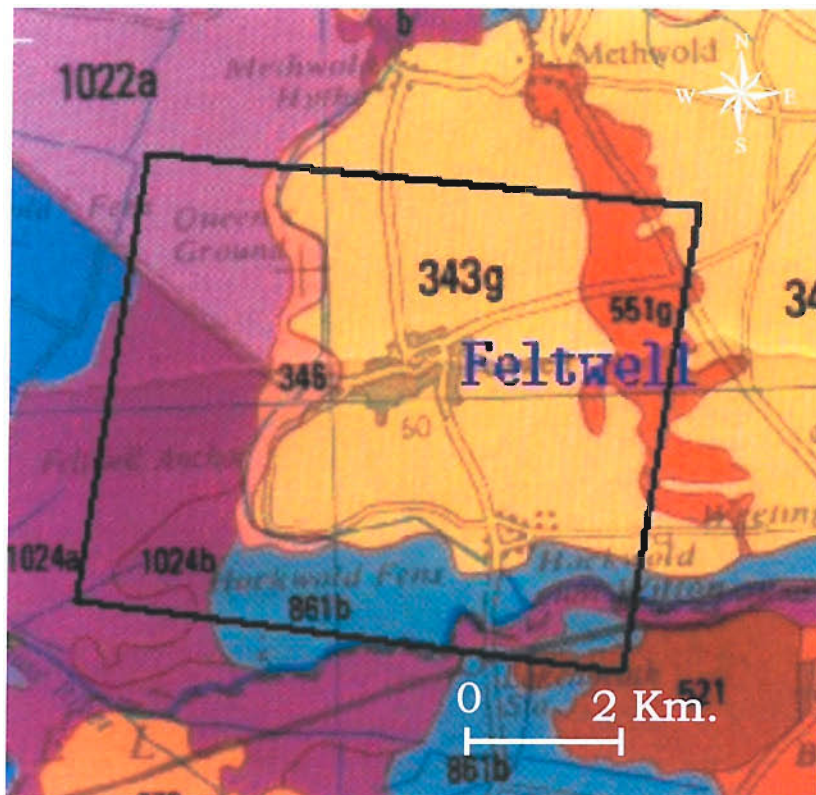


Figure 4.1: Soil map of Feltwell area. The black box represents the bounds of study area. The sand soil comprised of humic gleyic rendzinas (346), brown rendzinas (343g), typical brown sands (551g), typical humic-sandy gley soils (861b), whereas, peat soils comprised of earthy eutro-amorphous peat soil (1024a and b) and earthy eu-fibrous peat soils (1022a) (source: Soil Survey of England and Wales).

brown rendzinas whereas, peat soils comprised of earthy eutro-amorphous peat soil and earthy eu-fibrous peat soils (Figure 4.1). Winter wheat was planted in both types of soil, whereas barley was limited only to sandy soils (Figure 4.2). Focusing on just these two classes, a random sample of 75 pixels per-class was derived for each class. For winter wheat class, out of the 75 pixels for training, 40 were derived from sandy soils and 35 from peat soils.

To ensure that the basic assumptions that underlie classification, namely of pure pixels and discrete classes, were satisfied, locations in the vicinity of field boundaries were masked-out of the analyses to ensure that the sampled pixels were located within the relatively homogeneous cover of the crop planted in the large fields.

The spectral distribution of training data (Figure 4.3) shows that the cluster of training data of winter wheat class has two very distinct zones in feature space, one

populated by pixels from sandy soils and faces the other class barley, whereas the other zone comprises of pixel of winter wheat class from peat soils.

In SVM, the individual training cases vary in importance, with those lying close to the class borders most informative and helpful in determining the location of the classification hyperplanes. This property of SVM was utilized to reduce the requirement of training data.

For the present study, from the above discussed property of SVM, the support vectors for winter wheat class was expected to be mainly derived from sandy soils as training data of winter wheat class from sandy soils faces the other class, barley in feature space. In such a scenario, no training data would be required for winter wheat class from peat soils and that reduces the requirement for training data.

Two SVM analyses were, therefore, undertaken to understand the effect of removing training data from peat soils for winter wheat class.

1. SVM trained for the two classes of interest (barley & winter wheat) for training data derived from both type of soils (sand and peat) and tested on an independent test set.

2. SVM trained for the two classes of interest (barley & winter wheat) for training data derived only from sandy soils and tested on an independent test set as above.

The SVM approach was used with a RBF kernel to classify the data. There are two parameters  $C$  (section 2.7.1.2) and  $\gamma$  (section 2.7.1.3) in SVM to be optimised before training the classifier. A 5n cross validation approach (using a random sample of one fifth of the training set for validation purposes) was used to fine tune the two parameters.

Accuracy was assessed using a further, independent, random sample of 40 pixels/class. The testing set for winter wheat class comprised of 20 pixels each from sand and peat soils. This testing set was used in the evaluation of the accuracy of the two classification analyses undertaken.

The comparison of classification accuracy statements for both the above analyses was undertaken in a statistically rigorous fashion. Here, the statistical significance of differences in the accuracy of classifications on testing set derived using different training

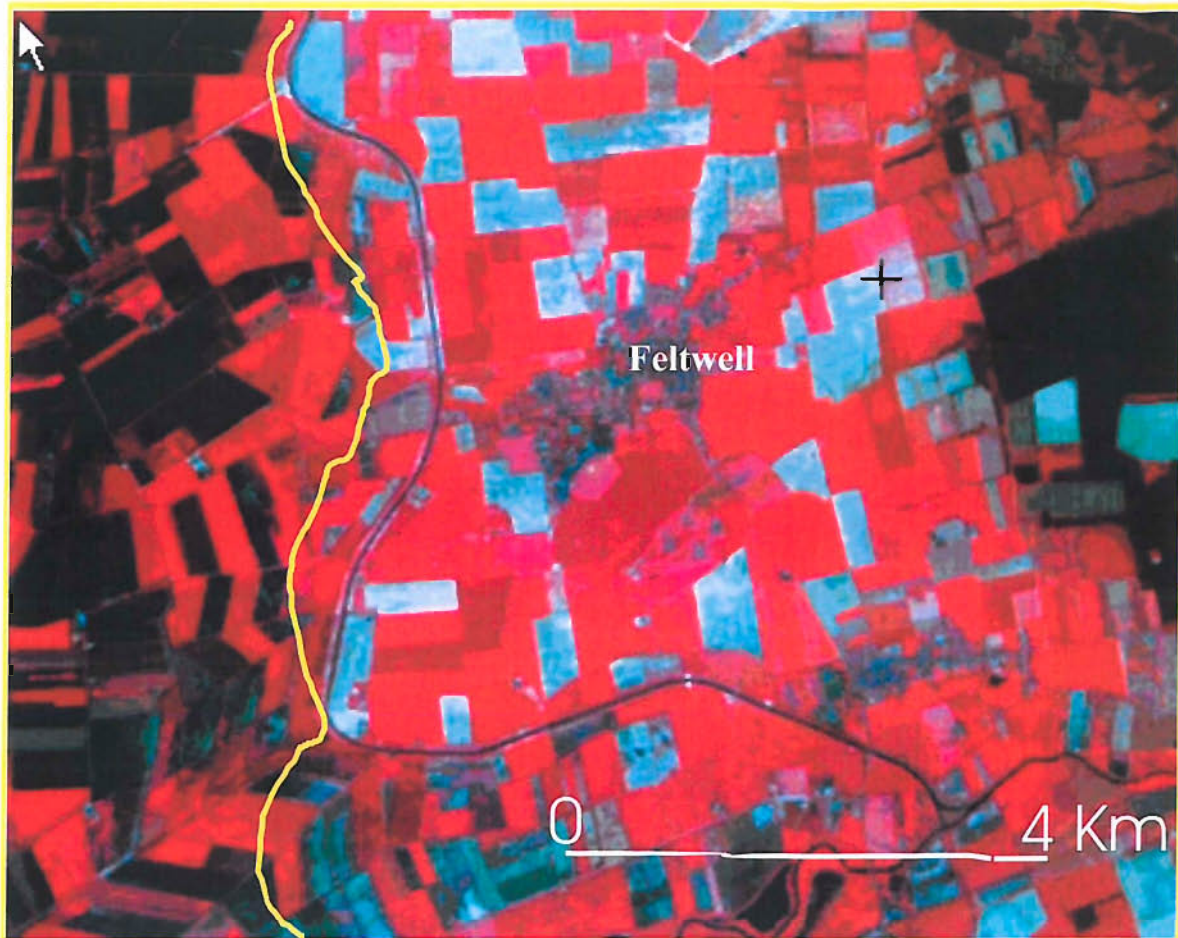


Figure 4.2: SPOT HRV FCC of study area demarcated into sandy and peat soils by the yellow line (Data, courtesy Natural Environment Research Council (NERC)). The study area was dominated by winter wheat and barley crops. The area west of yellow line is covered by peat soils and that to its east by sandy soils.

data was assessed using a  $M^c$ Nemar test, without correction for continuity, for related samples (Foody, 2004) given by:

$$Z = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \quad (4.1)$$

where,  $f_{12}$  and  $f_{21}$  represent the off-diagonal entries in the matrix

#### 4.1.2 Results and Discussion

The optimal values for parameters  $C$  and  $\gamma$  deduced from 5n cross validation (Table 4.1) for training set derived from both type of soils were  $2^4$  and  $2^{-8}$  respectively. These parameter settings were used to train the SVM classifier. The support vectors generated as a result are given in Table 4.2.

$\gamma$	$C=2^{-2}$	$C=2^{-1}$	$C=2^0$	$C=2^1$	$C=2^2$	$C=2^3$	$C=2^4$	$C=2^5$	$C=2^6$
$2^{-12}$	79.33	82.00	84.66	90.667	91.33	90.67	89.33	90.00	91.33
$2^{-10}$	82.00	84.00	91.33	92.667	92.00	90.67	91.33	92.00	92.00
$2^{-9}$	85.33	91.33	91.33	90.66	91.33	91.33	91.33	92.00	92.00
$2^{-8}$	90.00	89.33	90.00	89.33	90.00	90.67	<b>92.67</b>	91.33	92.00
$2^{-7}$	89.33	90.00	90.00	91.33	91.33	91.33	92.00	90.66	90.67
$2^{-6}$	90.00	90.00	90.67	91.33	91.33	92.00	90.67	90.00	90.67
$2^{-5}$	90.00	90.00	90.67	92.00	92.00	91.33	90.67	91.33	90.00
$2^{-4}$	89.33	90.00	91.33	91.00	91.33	90.67	91.33	86.67	86.00
$2^{-3}$	89.33	90.67	90.67	90.67	91.33	90.67	88.67	89.33	89.33
$2^{-2}$	88.66	89.33	89.33	89.33	88.66	88.00	88.00	88.00	88.00
$2^{-1}$	72.00	88.67	87.33	87.33	87.33	87.33	87.33	87.33	87.33
$2^0$	46.00	73.33	86.66	86.66	86.67	86.67	86.67	86.67	86.67
$2^1$	43.33	52.66	74.00	74.66	74.67	74.66	74.67	74.67	74.67
$2^2$	43.33	50.00	64.00	66.66	66.67	66.67	66.67	66.67	66.67

**Table 4.1:** Results of 5n cross-validation on training data for optimal selection of parameters  $C$  and  $\gamma$ . The value in the bold gives the highest accuracy obtained on training data with parameters  $C=2^4$  and  $\gamma=2^{-8}$  respectively.

The relation between  $\alpha$  and  $C$  is given by the equation:

$$0 \leq \alpha \leq C \quad (4.2)$$

The values in first column in Table 4.2 represent  $\alpha$  values and ranges from 0 to 16 as the optimised value for  $C$  parameter used was 16 ( $2^4$ ). The overall distribution of support vectors and their  $\alpha$  values are summarized in Table 4.3.

SV ( $\alpha$ )	B3	B2	B1	Crop	Soil
16	106	37	60	Barley	sand
4.494650187	109	37	60	Barley	sand
16	103	36	59	Barley	sand
8.04803743	96	37	59	Barley	sand
16	118	36	57	Barley	sand
16	111	36	57	Barley	sand
16	119	35	59	Barley	sand
6.554303967	125	34	58	Barley	sand
16	121	36	58	Barley	sand
11.87109497	118	36	59	Barley	sand
16	116	36	58	Barley	sand
16	110	35	56	Barley	sand
16	119	34	58	Barley	sand
16	120	35	57	Barley	sand
9.744216682	118	36	59	Barley	sand
7.529921847	116	37	59	Barley	sand
16	108	34	55	Barley	sand
16	111	35	58	Barley	sand
16	117	36	59	Barley	sand
4.324733828	101	33	56	winter wheat	peat
16	119	35	57	winter wheat	sand
16	118	35	57	winter wheat	sand
16	111	38	59	winter wheat	sand
16	106	37	59	winter wheat	sand
4.437130987	112	35	57	winter wheat	sand
16	117	36	58	winter wheat	sand
6.532216624	112	35	57	winter wheat	sand
16	121	35	56	winter wheat	sand
16	115	35	57	winter wheat	sand
16	118	35	58	winter wheat	sand
16	122	34	57	winter wheat	sand
16	119	35	59	winter wheat	sand
16	103	35	57	winter wheat	sand
16	111	35	57	winter wheat	sand
16	115	35	58	winter wheat	sand
16	102	36	57	winter wheat	sand
16	115	36	58	winter wheat	sand

Table 4.2: Support vectors when training data comprised of pixels from both type of soils with parameter settings of  $C$  and  $\gamma$  of  $2^4$  and  $2^{-8}$  respectively deduced from 5n cross validation. The first column shows the  $\alpha$  values of each support vector followed by spectral values of the support vector (training data) in the three bands under B3, B2 and B1.

Class	Total number of support vectors (number of support vectors with maximum $\alpha$ value of 16).
Barley (sand)	19 (13)
Winter wheat (sand)	17 (15)
Winter wheat (peat)	1 (0)

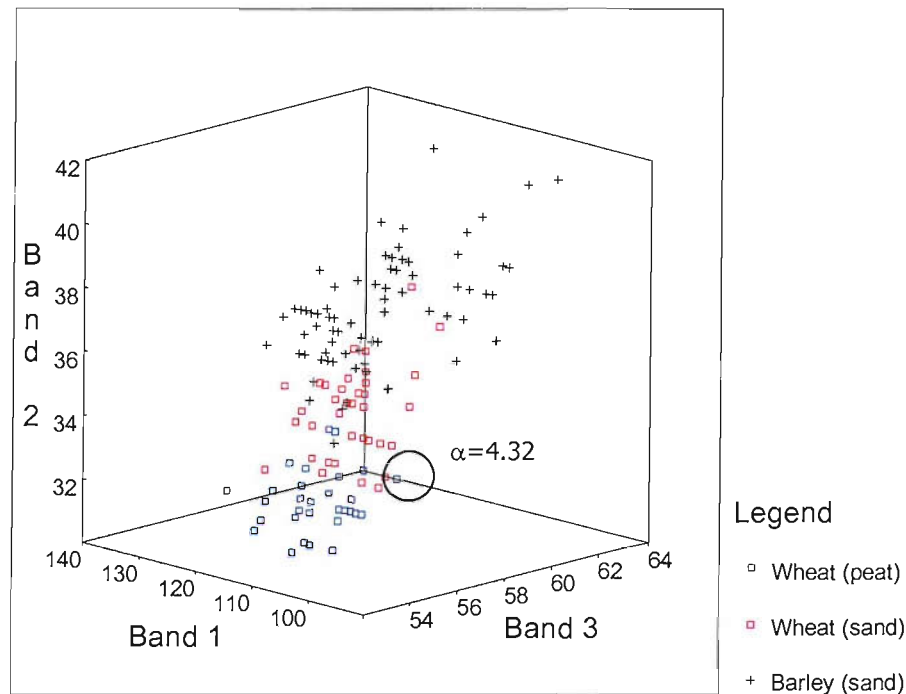
Table 4.3: Support vectors with  $\alpha$  values.

Table 4.3 shows that support vectors (which are central to the establishment of the decision surface (classifier)) in case of winter wheat class was mainly derived from sandy soils. This can be attributed to the fact that the training data for winter wheat class from sandy soils faced the other class barley in feature space. There was only one support vector for winter wheat class from peat soil (Table 4.2 and Table 4.3) and it can be considered of little consequence as its  $\alpha$  value was very low to the tune of 4.32 (Table 4.2 and Figure 4.2) as compared to very high values (maximum possible) of 16 for majority of support vectors (15 out of 17) for winter wheat class from sandy soils (Table 4.3).

The meager contribution of  $\alpha$  value of 4.32 by the lone support vector of winter wheat class from peat soil was expected to hardly affect the decision surface (hyperplane) given by the equation:

$$f(x) = \text{sgn}\left(\sum_{i=1}^r \alpha_i y_i k(x, x_i) + b\right) \quad (4.3)$$

It can, therefore, be presumed that removing the lone support vector of winter wheat class from peat soils (in other words removing training data of winter wheat class from peat soils as data other than support vectors are redundant and do not play any role in the establishment of the classifier) will hardly/marginally effect the outcome on the testing data as the classifier (equation 4.3) is hardly altered.



**Figure 4.3:** The distribution of training data of winter wheat and barley class in feature space. The lone support vector of winter wheat class from peat soil is encircled and its  $\alpha$  value of 4.32 highlighted.

In light of above, two SVM machines were trained, one trained with training data of both classes from both type of soils discussed above (section 4.1.1) and other trained with training data of both classes only from sandy soils (75 pixels for barley class and 40 pixels of winter wheat class from sandy soils leaving the 35 pixels of winter wheat class from peat soils). Both the classifiers were tested on an independent test set resulting in same outcome (Table 4.4).

Actual ↓

	Barley	Winter wheat	Total
Classified →	Barley	1	40
	Winter wheat	35	40
	Total	44	80

Table 4.4: Confusion matrix of testing set for both the analysis (classifier trained with or without wheat pixels from peat soils).

The comparison of classification accuracy statements for both the above analyses was undertaken in a statistically rigorous fashion. Here, the statistical significance of differences in the accuracy of classifications on testing set derived using two different training data was assessed using a M<sup>c</sup>Nemar test resulting in Z value of 0.

## 4.2 Conclusions

In this chapter, the support vectors of an SVM classification were related with soil type of training sites. The results can be summarized as:

- The analyses shows that support vectors of class winter wheat were mainly drawn from sandy soils implying that only training data of winter wheat from sandy soils are relevant in establishing the decision surface between winter wheat and barley classes using SVM classifier.
- The analyses shows that training data for winter wheat class from peat soil are not required as the accuracy on the testing set remains same with or without training data of winter wheat class from peat soils.
- The study shows that with information on soil type, training sample acquisition can be focused to regions most likely to furnish support vectors.
- An accurate SVM classification may be undertaken using a small training set if support vectors resulting from SVM classification can be identified with ground properties such as soil type of training fields.



# **CHAPTER 5 - Intelligently Reducing Training Requirements of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification**

## **5.1 Introduction**

The analysis detailed in chapters 3 and 4 demonstrated the potential of SVM as a classifier. SVM classifications were more accurate than classifications derived with other classification techniques. In addition, SVM used only a fraction of training data as compared to other classifiers. Although SVM classifications are effectively based on a small number of training data, the support vectors, a large training set may still be required to ensure that appropriate training data are included. Indeed, the nature of the training data can have a larger impact on classification accuracy than the classifiers used (Hixson *et. al.*, 1980, Campbell 2002).

The design of training stage is often guided by the classical statistical view of the classification process generally considering a probabilistic algorithm such as the MLC. This type of classifier requires a complete description of each class in feature space. To achieve this, a large training set, spread over the entire study area is required. Inappropriate placement or too few pixels in training site produces statistics which may not be able to characterize the land cover classes. The requirement for large sample sizes is, therefore, not unusual and penalties for collecting less in some cases can be severe (Curran and Williamson, 1986). Conventional training data acquisition schemes, therefore, aim to capture a large training set spread all over the study area to obtain representative samples of the classes. For data to be representative, a number of training data acquisition schemes has been suggested in literature (*e.g.* random, conventional and systematic sampling techniques) (section 2.3.3.1.1.4).

It is costly in terms of time and finance to acquire large, representative (spread all over the study area) training samples (Buchheim and Lillesand, 1989; Jackson and Landgrebe, 2001). Much research has, therefore, focused on ways to reduce the requirement of training data. This includes methods to reduce the dimensionality of the data sets to be classified to avoid Hughes phenomenon (for finite training samples accuracy first increases with increase in dimensionality and then decreases) (section 2.2.1), designing efficient training sampling scheme by incorporating the spatial dependence of the classes (section 2.3.3.1.1.4).

An alternative procedure is to recognize that individual training samples vary in value and this importance depends on the classifier used. Thus by focusing only on the most informative training samples, an accurate classifier may be defined at the cost of only a small training set. With SVM as a classifier, only the training samples that lie at the edge of the class distributions in feature space (support vectors) are relevant in the establishment of the decision surface. Data other than support vectors can effectively be discarded without compromising the accuracy of the classification (section 2.7.1.1).

The objective in classification is to extract the maximum information from the remotely sensed data and if possible with a small number of training sets to make the classification process economical. Given that SVM classifiers are more accurate than other classifiers (section 3.1.4.4), they should be adopted increasingly and, critically, the design of their training stage constructed around their nature.

SVM is expected to generalize more accurately as compared to other classifiers even if trained with a small training set that provides appropriate support vectors. The requirement of small training set translates to less expenditure on acquiring training data from field both in time and monetary terms. Thus the key property of SVM to use a small training set provides an opportunity to review the ground data collection policy with the intent to reduce the requirement of training data.

The study discussed in this chapter aims to evaluate classification accuracy with special regard to SVM classifier and will focus on:

1. An assessment of the ability to intelligently identify most useful training samples, the support vectors for SVM classification directly from field.
2. Comparing classification accuracy resulting from SVM classification with a suite of other classifiers namely DA, DT and ANN to appreciate the generalibility of SVM.
3. Identifying support vectors resulting from an SVM classification with ground attributes to help focus training acquisition process for future analysis to a limited area.
4. Evaluating financial implication of training a SVM with a small intelligently collected training set with a large training set collected under conventional training data acquisition scheme.
5. Evaluating reduced requirement of training data to accurately classify only one class from the many land cover classes available.

## **5.2 Study Area**

The study area comprises the south-western districts of the Punjab state of India. It includes the districts of Bathinda, Faridkot, Muktsar, Moga and parts of Ludhiana (Figure 5.1) covering an area of 11037 km<sup>2</sup>. The area falls under sub-tropical, semi-arid climate and the region has, therefore, marked extremes of climate. It is influenced by westerly winds in summer raising temperatures as high as 44 °C in May and June and in winters, the north easterly winds reduces temp to 2 -3 °C. The area is basically flat with alluvial soil and dotted with a number of sand dunes (notably in the southern part of study area) locally called as tibbas. The area is criss-crossed by a number of canals. The Sirhind canal and Rajasthan feeder are major canals of the area.

The area is predominantly under agriculture with kharif or monsoon season (May to October) followed by rabi season (October to May). The major crops grown in the kharif season are cotton and rice followed by wheat during rabi season. Besides this, citrus and grape orchards are also found in the area.

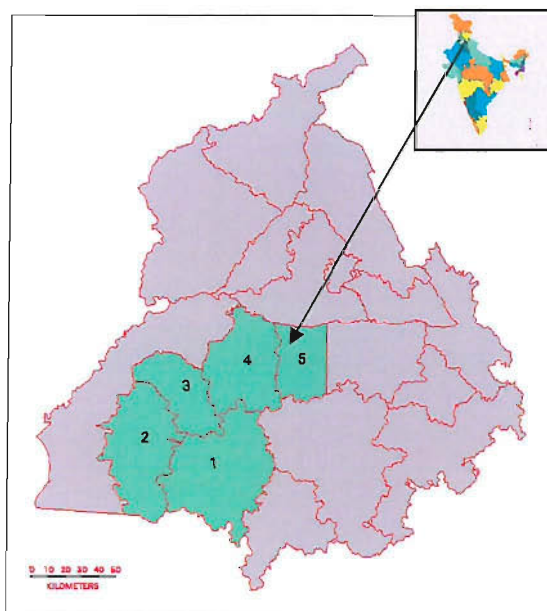


Figure 5.1: Study area shows the districts (1. Bathinda 2. Muktsar 3. Faridkot 4. Moga 5. Part of Ludhiana district) of Punjab state.

The southern part, especially the Muktsar district (Figure 5.1), is affected with the problem of waterlogging which in turn affects agricultural productivity.

### 5.2.1 Description of Study Area

The study area is agrarian with 84% of land under agriculture (Director of Land Records, Punjab, 2004). The state witnessed a Green revolution in 1970s with the introduction of the high yield varieties (HYVs) of wheat developed by Nobel laureate Normal Borlaug (Shiva, 1991). The Green revolution in Punjab transformed India from a position of a begging bowl to a position of a bread basket (Shiva, 1991).

The native varieties of wheat sown by farmers before Green revolution tend to lodge or fall over when subjected to intensive fertilizer to supplement organic manure. The varieties introduced under Green revolution were shorter with stiff stems and were able to

Page 105 missing

2. The topography of the area is very flat with slopes ranging approximately from  $1^{\circ}$  to  $1.5^{\circ}$ . As a result water due to rains or water applied to agricultural fields stagnates in area. The percolation of accumulated water increases water table depth and subsequently results in waterlogging.
3. Less with drawl of ground water for irrigation due to its poor quality for agriculture.
4. Construction of roads, railway lines and canals obstructing the natural gradient of flow of water.
5. Increase in area cultivated for water intensive crops such as rice (Table 5.1). The appropriate water supplied to rice fields from canals seeps into the ground, thereby, increasing the ground water level and resulting in waterlogging at places.



Figure 5.2: An unlined canal.

Year	Area under crop ('000 Km <sup>2</sup> )	
	Rice	Cotton
1960-1961	2.27	4.47
1970-1971	3.90	3.97
1980-1981	11.83	6.49
1990-1991	20.15	7.01
1999-2000	26.04	4.77
2000-2001	26.12	4.73
2001-2002	24.89	6.07
2002-2003	25.30	4.50

Table 5.1: Comparative area under rice and cotton in Punjab state (Source: Director of Land Records, Punjab, 2004).



Figure 5.3: Salt affected land due to waterlogging.

The study area especially, the Muktsar district (Figure 5.1) was severely affected by waterlogging in 1995 and 1997 (Singh, 1998). This problem has continued since (Figure 5.3), though anti-waterlogging measures like construction of drains along with medium-depth tube-wells have been installed (The Tribune newspaper, 19 July, 1999).

Waterlogging adversely affects the growth of plants as (a) humid conditions are conducive for the growth of insects, pests and pathogens which attacks the crops (b) waterlogging of the root zone results in oxygen deficiency, leading to a halt in root growth



and metabolism, death of the roots and eventual wilting of the crops (Figure 5.4) (c) soil in some areas gets affected with salinity affecting the growth of the crops (Figure 5.3).

The growth of cotton crop in particular has been the hardest hit of all the crops in the area as it is sensitive to water (needs less water) and farmers at time have to uproot the whole crop wilted as a consequence of waterlogging (The Tribune newspaper, July 2003; The Indian Express newspaper, Oct 29, 1998). The damage to cotton crop is due to attack by pest, lack of fruitification due to excessive humidity (*i.e.* dry conditions are required after flowering



Figure 5.4: Wilted rice as a result of waterlogging.

stage for fruitification to take place (Figure 5.5))

(<http://www.onlypunjab.com/latest/fullstory-newsID-1445.html>).

As a result of waterlogging, cotton has often been replaced by rice. The diversion from cotton to rice requires more canal water as quality of ground water in the study area is not suitable for agriculture and rice is a water intensive crop. As a consequence, more water is being added into the area through canals to provide appropriate water for rice crop and as there is no with drawl of ground water (because of the poor quality for agriculture)



creating a water imbalance in the study area. This is reflected by the increased problem of high water table and waterlogging afflicting the study area.

The diversion from cotton cropping to rice brought immediate benefits to the farmers because of secured rice crop being water intensive. However, the increase in area under rice (Table 5.1), a water intensive crop, aggravated the problem of waterlogging detrimental to agriculture in the long run. The farmers are, therefore, being lured back into cotton cropping. This decision to shift back from rice to cotton cropping is also aided by the reduced demand for rice as the state of Punjab is surplus in rice production like other parts of the country. As a consequence, the surplus rice is rotting in the store houses of the state (B. K. Chum, July24, 2002, [http://economictimes.Indiatimes.com/cms.dll/articleshort/art\\_id=16897102](http://economictimes.Indiatimes.com/cms.dll/articleshort/art_id=16897102)). Thus the rice crop is no longer a lucrative crop as the demand has decreased in the markets and is also one of the reasons of waterlogging in the area. However, the domestic demand of cotton of the country is met on many occasions by imports ([www.Indiaone.stop.com/cotton/cotton.html](http://www.Indiaone.stop.com/cotton/cotton.html)). In addition, the crop insurance for cotton (<http://www.tribuneindia.com/1999/agro.html>) introduced from October 1999 by the government and contract farming by textile mills (<http://www.thehindubusinessline.com>, 03 March, 2004), which compensates for any loss of the crop are added incentives for growing cotton crop (<http://www.onlypunjab.com/latest/fullstory-newsID-1445.html>).



Figure 5.5: The cotton crop is not watered after flowering stage. The exposed soil is dry as farmers do not water cotton crop after flowers appear to avoid attack by insects and pathogens.

Cotton is a major input to the textile industry and, therefore, plays an important role in India's agrarian and industrial economy. The pre-harvest area and production of cotton is given by the Directorate of Economics and Statistics (DES), Ministry of Agriculture and cooperation, apart from a number of trade organizations namely Cotton Corporation India and The North India Cotton Association (Charanjit Ahuja, <http://www.expressindia.com//daily//19980310/06955704.html>). The estimates provided by these organisations vary greatly which, therefore, does not give a clear picture of demand and supply. This makes it imperative to have a more scientific approach which can provide pre-harvest estimates of cotton crop needed especially by the textile mills and the government so that procurement, storage or import/export strategies for cotton can be planned in advance of the harvest of the cotton crop.

Remote sensing technology can provide timely and accurate estimates of area cultivated under the cotton crop before its harvest. Remote sensing based procedures for

pre-harvest acreage estimation have already been developed for important cereals like wheat and rice under CAPE project in India.

The estimation of area under the cotton crop prior to its harvest (pre-harvest acreage) is undertaken by the “Crop Acreage and Production Estimation” (CAPE) project by Department of Space, Government of India using remote sensing technology (Navalgund *et. al.*, 1991). This project is on a continuous basis over the years since its inception in early 1990s. The results provided notably in terms of area under the cotton crop are important because of the fluctuation in area sown under the crop over years (Table 5.1) due to waterlogging problem. The CAPE project, however, has certain drawbacks which can be appreciated after understanding the methodology it uses.

#### **5.2.1.1 Methodology of the CAPE Project**

The project uses the district as the unit of study (*i.e.*, provides pre-harvest acreage and production at district level). As the area under the study was very large (average area per district is 3000 km<sup>2</sup>) (Figure 5.1), a considerable effort is needed to collect ground data and also in subsequent analysis. It, therefore, becomes difficult to classify the whole area (complete enumeration) in short time to provide pre-harvest estimates of area under the crop.

The alternative to complete enumeration, therefore, followed in the project is based on sampling technique. The remotely sensed imagery is chosen for around late September, when cotton is at its maximum vegetative phase (onset of flowering) (Navalgund *et. al.*, 1991). The period of late September is based on the premise that the cotton crop is expected to be most spectrally separable from the other crops and, therefore, could be classified accurately. The process includes overlaying a grid of 5 X 5 Km size on a false colour composite (FCC) of IRS LISS-III data. This results in a number of squares (segments) of 5 X 5 Km. size for each district. These segments are further classified as A, B, C based on cotton crop proportion deduced from FCC ( $A > 50\%$ ,  $25\% < B < 50\%$  and

C < 25 % of the area of crop under the segment). This stratification of the area into A, B and C type is usually undertaken once in 4-5 years. For sampling, 15 per cent of these segments from each category A, B, C are randomly selected (Figure 5.6) for further analysis (classification). Thus, the grid is used to ensure that training data extracted have wide coverage, a requirement of the conventional training data acquisition scheme.

The training data for all the land cover classes present in the selected segments are collected by visiting the field around the most vegetative phase of the cotton crop (second week of September). Training signatures are generated for all the classes and the selected segments are then classified using MLC. The classified output gives area under cotton in the selected segments. The area under cotton for each district in the study area is then computed by extrapolating the results of the selected segments (15 per cent of study area) with respect to the total segments available for each type of all segments (A, B and C) available for the district given by the equation:

$$\text{Area of cotton in the district} = \left\langle \frac{n_a \times N_a + n_b \times N_b + n_c \times N_c}{N_a + N_b + N_c} \right\rangle \times \langle A\_D \rangle$$

where,

$n_a, n_b, n_c$  = Mean crop (cotton) proportion of the selected A, B and C type segments deduced from classification.

$N_a, N_b, N_c$  = Number of A, B and C type segments in the district.

$A\_D$  = Area of the district under study.

Thus the net result comprises of a classified output of the selected segments (15 per cent of the study area) and a numerical value indicating the area under the cotton crop in the district.



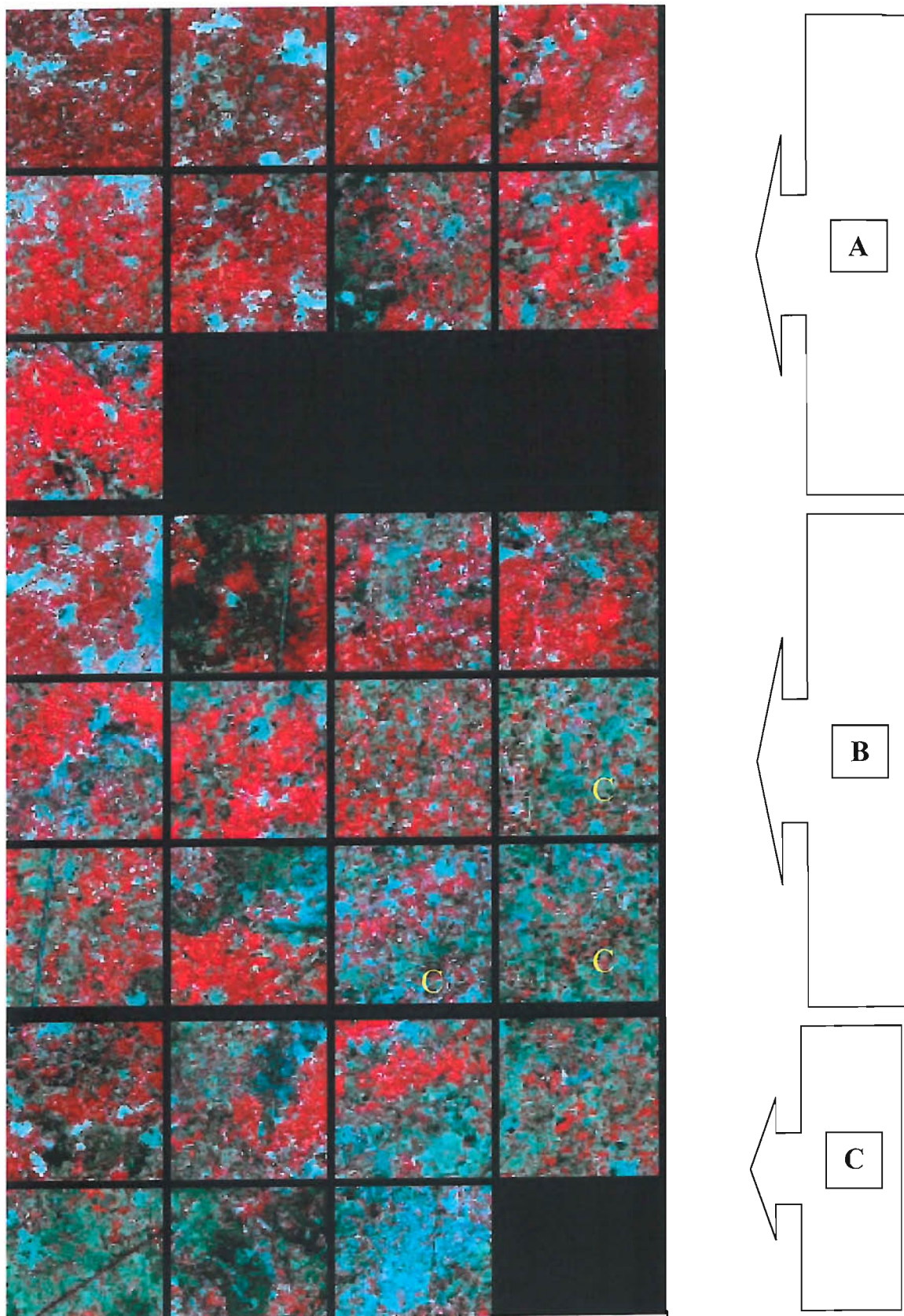


Figure 5.6: FCC of raw IRS-1D satellite data (date of acquisition 16-09-2002) of selected segments A, B and C of the Muktsar district for cotton area estimation under CAPE project. These segments constitute 15 per cent of Muktsar district in area. The figure also shows that some of the B type segments (marked as C in the FCC) definitely belong to C type. The problem arises because the segments are selected once in 4 -5 years and the area under cotton is fluctuating.

### **5.2.1.1.1 Drawbacks of the CAPE project**

The CAPE project has certain drawbacks as regards training data requirements and classifier used as enumerated below.

1. The training data are collected from blocks of pixels in the selected segments and are, therefore, affected by the problem of spatial auto correlation (section 2.3.3.1.1.4). The block of pixels results in a very large training set which does not provide any additional information as compared to training data comprising of single pixels spread throughout the study area.
2. There appears to be no apparent advantage of segment approach over the one based on individual pixels for training distributed randomly throughout the study area as far as the field visit is concerned. The team has to traverse the whole study area to reach from one segment to another which are spread all over the study area just as for collecting training data for individual pixels (section 2.3.3.1.1.4).
3. The objective of Department of space in CAPE project is to get an accurate map and get it cheaply. However, training data collected are very large (though spatially auto correlated). The objective can be met if the requirement of training data can be reduced to make the classification process economical.
4. The project uses MLC for classification. One of the objectives of classification is to produce a classifier that generalizes accurately on unseen cases. Studies have shown that SVMs are generally more accurate than other classifiers (section 2.7.3).
5. The project produces classified output of only the selected segments and not the whole study area. As such no thematic map (crop map) is produced. Only area under cotton is given as numerical value like other statistical departments involved in acreage estimation (Table 5.1). Thus remote sensing is not fully exploited as it can give area under cotton not only in statistical terms but also in spatial context.

6. The categorization of segments as A, B, C is undertaken once in 4-5 years which does not hold good for areas where the area under crops are changing over years. Cotton in the present study is one such example (Table 5.1).

The present study, therefore, aims to solve the draw backs of CAPE project in conjunction with the broader issues of supervised classification specifically relating to the use of reduced training set without compromising the accuracy of the classification as detailed under section 5.1.

### **5.3 Data**

Indian Remote Sensing Satellite (IRS-1D) with a spatial resolution of approximately 24 m acquired by LISS-III sensor, date of pass 22<sup>nd</sup> September, 2003 path/row of 93/49 and 93/50 encompassing the study area were used. The Red (0.62-0.68  $\mu\text{m}$ ), near-infrared (NIR)(0.77-0.86  $\mu\text{m}$ ) and middle-infrared (MIR) (1.55-1.75  $\mu\text{m}$ ) bands were used. These bands are tailored for agricultural crop discrimination.

The ground data were collected by visiting the field from 15<sup>th</sup> to 21<sup>st</sup> September, 2003, near the time of satellite sensor data acquisition. Three agricultural classes namely: cotton, rice (basmati), rice (local) dominated the study area and were the focus of this study. Built-up land and sand which were also abundant were included in the study.

Ancillary data comprised of the following information:

1. Soil map of Punjab state (1:250,000 scale, 1993)
2. Newspaper reports: The local newspaper especially “The Tribune” gives detailed articles about waterlogged areas and resulting losses to crops (Figure 5.7).
3. Status of crops: Agriculture departments located at district headquarters (Figure 5.1) and farmers at the field level provided valuable information about spatial distribution of crops with regard to their type/maturity stages.

4. Information gained from scientists of Punjab Remote Sensing Centre, Ludhiana, who had earlier visited the field to collect ground data for cotton and rice for CAPE project.
5. In addition, the existing expert knowledge of the author about the study area especially pertaining to cotton and rice crops because of the long involvement in the CAPE project (1994-2002) helped a lot.

The ground data collection was designed to fulfil the aims of the study. It, therefore, comprised of a conventional training data collection procedure (large training set collected by conventional approach) against an intelligent scheme (comprising of small training set) designed for SVM classification.

**Heavy rain damages cotton, paddy  
Farmers live in tents as houses collapse  
Chander Parkash  
Tribune News Service**

Midhu Khera (Muktsar), **July 22, 2003**

The heavy rain, which lashed this region three days ago, have left behind a trail of destruction as hundreds of houses collapsed and cotton, paddy and other crops, including vegetables, in thousands of acres have been badly affected in this and other villages, including Fatta Kera, Bhullarawala and Bhitwala, of the district.

The villages were presenting a picture of destruction. Some of the schools, civil and veterinary dispensaries and dharamshalas in these villages had also been inundated. Rain water could not be drained out from these pockets of Muktsar district despite the fact that crores of rupees had been spent during the previous SAD-BJP government on anti-water logging measures in this area.

The farmers pointed out that a drain passing from the village had not been cleaned by the authorities concerned. They added that various drains dig by the Irrigation and Drainage Department to prevent waterlogging also failed to drain out the rain water.

Figure 5.7: The newspaper report about waterlogging in the study area (source: The Tribune newspaper, July 22, 2003, Chandigarh edition)



### 5.3.1 Conventional Training Data Scheme

In the conventional training data scheme, the ground data (training and testing data) were collected based on stratified random (by class) sampling technique (Figure 5.8). For this, a land cover map of previous year (September, 2002) produced by remotely sensed data was used as a rough guide. A grid with a spacing of 500 m was overlaid on vector layers of each class resulting in a number of squares (segments). The function of the grid was to ensure that the training data were acquired from throughout the study area so as to capture the spectral variability of the classes, a requirement of the conventional training data acquisition scheme. In all 180 such segments were selected randomly for each of the five classes under study (90 each for training and testing). These selected segments were then visited on ground for locating homogeneous sites for classes around the satellite over pass time (section 5.3). From each selected segment, one pixel was selected from the homogenous sites of the classes visited. The size of 90 pixels per class for training set was decided based on the recommendation of 30 times the discriminatory bands to be used (Lillesand and Kiefer, 2004), three in this case. The large training set was intended to capture the spectral variability of the classes and thus able to describe the classes statistically, a requirement of the training data by conventional standards. The size was kept same for all the classes to avoid the effect of unbalanced training sets (section 2.6.4).

The small grid size of 500x500 m was chosen so that 180 independent samples of minor classes like sand and basmati rice could be sampled. The set of 180 data points for each class were transferred on the image and data extracted. Thus, each class comprised of 180 pixels which were divided equally into training and testing set randomly. Thus training and testing data comprised of 90 pixels per each class.

The spectral distribution of the training data (Figure 5.9) comprising the DN in the three bands shows that the class built up and sand overlapped mutually. This overlapping in feature space was also evident between rice basmati and rice local classes. The overlapping of the classes in feature space was also confirmed by the statistical description

of the training data (Table 5.2). However, agricultural and non-agricultural classes were very distinct in the feature space (Figure 5.9).

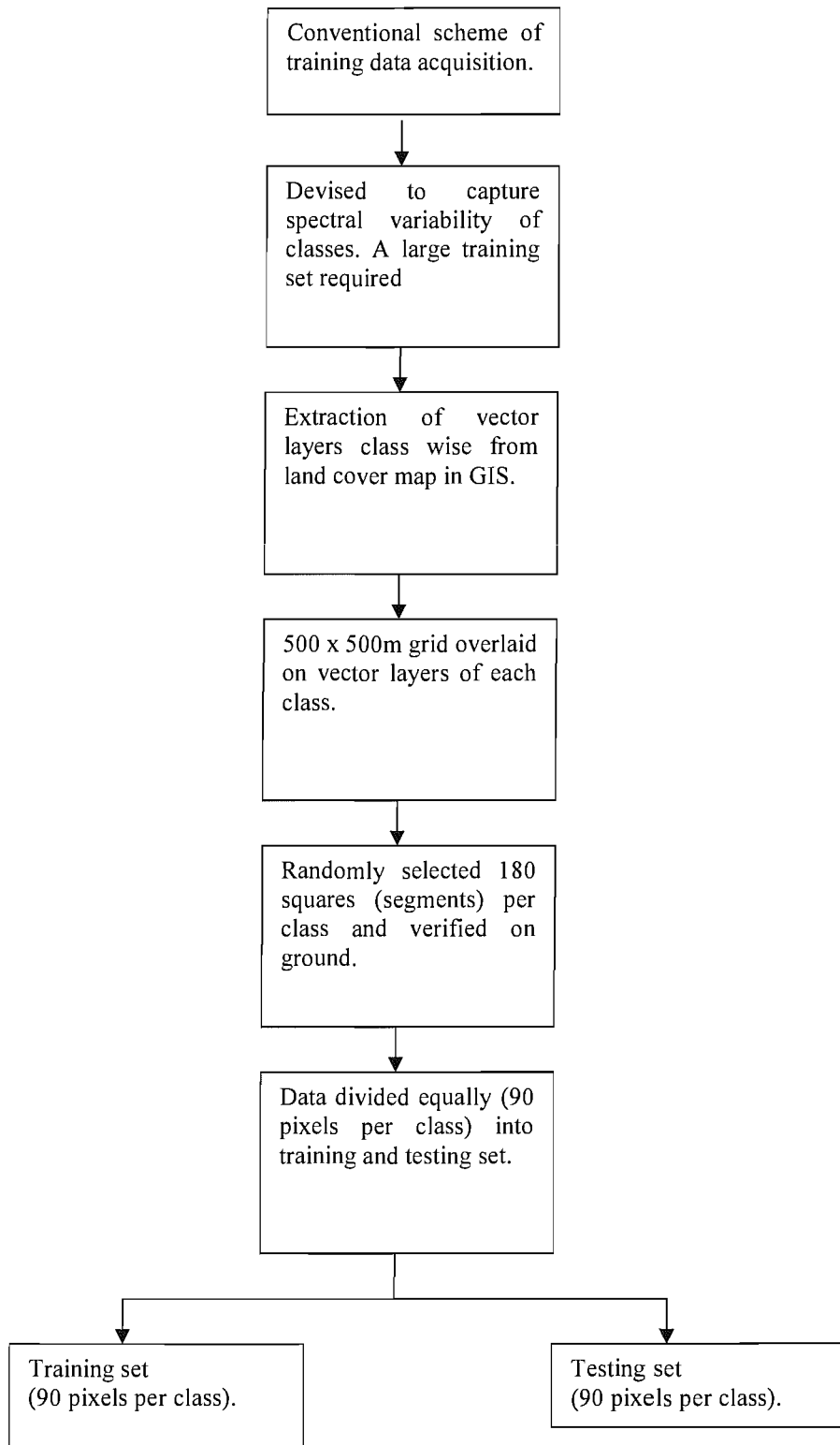


Figure 5.8: Procedure followed for training data acquisition under conventional scheme.

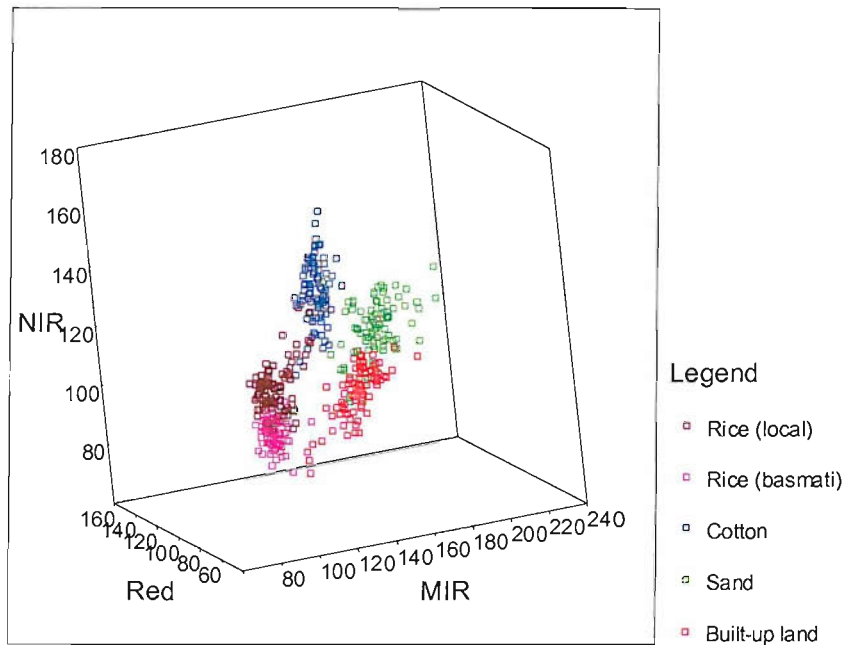


Figure 5.9: Spectral distribution of training data collected under conventional training data collection scheme.

The training data were close to normal but multi-modal (Figure 5.10) and, therefore, MLC should not be used.

Class	Red Band (DN)				NIR Band (DN)				MIR Band (DN)			
	Min	Max	Mean	Sd	Min	Max	Mean	Sd	Min	Max	Mean	Sd
<b>Built-up</b>	77	118	96.31	7.57	77	115	98.13	6.78	123	189	157.71	12.42
<b>Sand</b>	93	137	115.16	9.97	93	128	112.37	7.39	129	221	179.12	14.71
<b>Cotton</b>	48	64	52.60	3.36	119	172	144.59	10.5	103	136	118.61	5.23
<b>RiceBasmati</b>	54	67	58.14	2.85	84	115	99.42	5.14	83	114	92.70	4.81
<b>Rice Local</b>	48	67	53.49	3.87	99	139	115.26	7.64	79	115	93.73	8.59

Table 5.2: Statistics of training data showing minimum (Min), maximum (Max), Mean and standard deviation (Sd) of digital numbers of the training data of the five classes in the three bands.

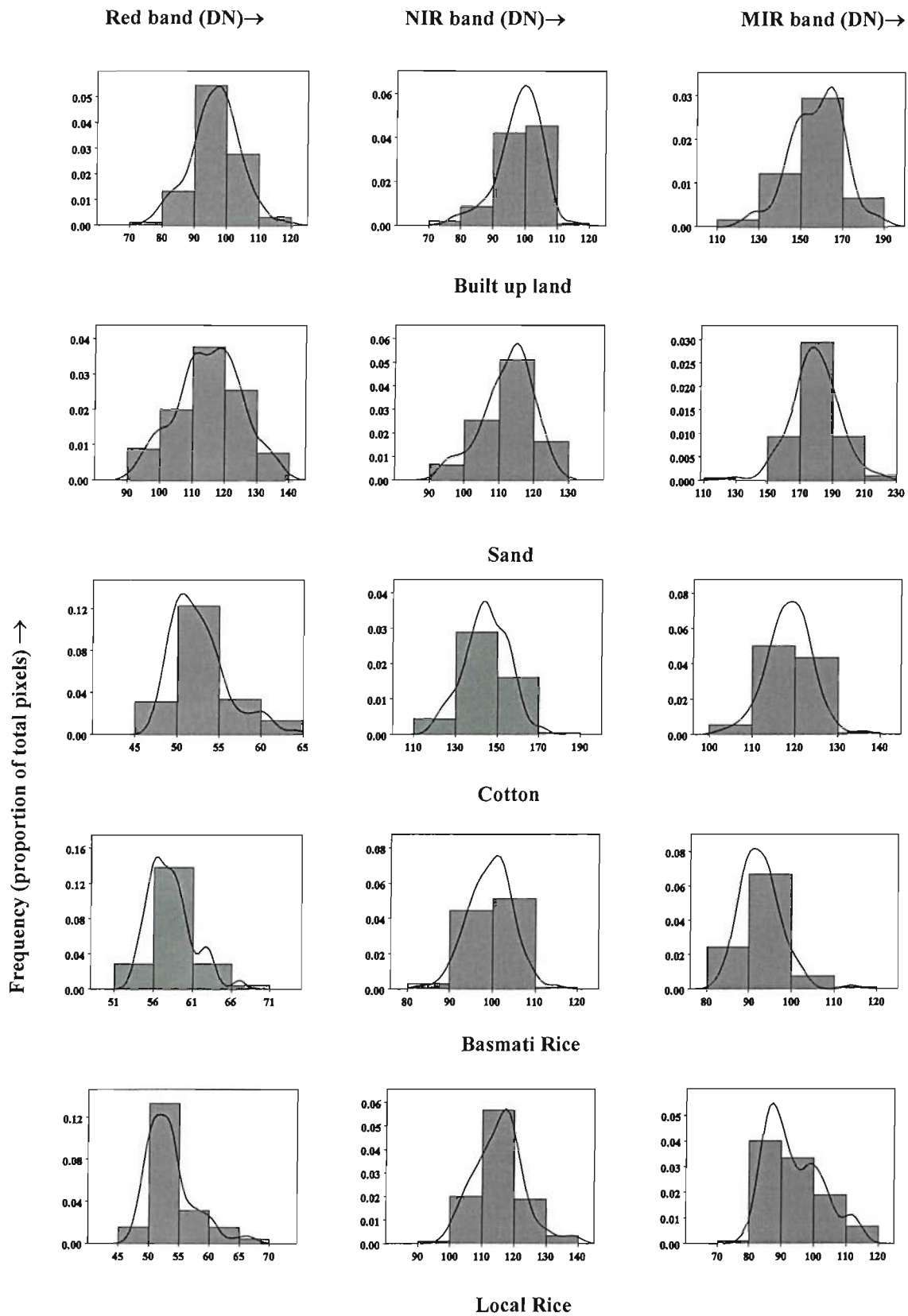


Figure 5.10: Histograms of the training data. The solid lines show the smoothed histograms.

### 5.3.2 Intelligent Training Data Scheme

The intelligent scheme for training data collection was devised to acquire training data from sites with relatively extreme spectral responses (potential border training samples) to act as support vectors for SVM classifier (Figure 5.11). This approach required understanding of the variables affecting the spectral response of the classes. These variables were well defined for agricultural crop classes. For example, in case of crops, the location in feature space is a function of many variables. These range from factors related with growth of the crop (topography, environment, management practices) to the satellite sensor characteristics (spatial and spectral resolution). The judicious combination of these factors can help to identify sites that may form useful support vectors. For example, for crops this translates to very high, very low or a combination of very high or very low values of spectral response of the bands used in feature space. For instance, a healthy crop generally has very high value in NIR band and very low in Red band, a matured crop on the other hand has comparatively low value in NIR band and high in red band (Figure 5.12). Similarly, moisture content influences MIR response, typically reducing reflectance. For example, a matured crop is dry, therefore, the spectral response in MIR band is often high but for young healthy crop, the leaves are moist, thereby reducing MIR response. The MIR response is also affected by the location of the training sites. For, example, if the training site is located near water bodies like waterlogged areas or near canal (especially if unlined), resulting in higher water-table, the MIR response reduces as compared to training site which is away from such locations and located in dry conditions. On the basis of this knowledge, one may be able to predict sites that may furnish support vectors to separate the various classes of interest.

Apart from the intrinsic properties of the crops, prior information of soil background of the crop can also be exploited to intelligently select sites to furnish the appropriate support vectors (section 4.1.2). For instance, in some cases, the feature space of a crop can be partitioned based on some soil attributes such as based on dark or light

tones. If the support vectors can be identified based on such attributes, training selection process can be limited to a smaller area (section 4.1.2). This is especially true for crops where soil is exposed to sky and contribute significantly to the spectral response of the crop. Cotton was one such crop being studied (Figure 5.5). This in turn would help future analyses in the area to focus training sample acquisition to regions most likely to furnish support vectors based on particular soil type.

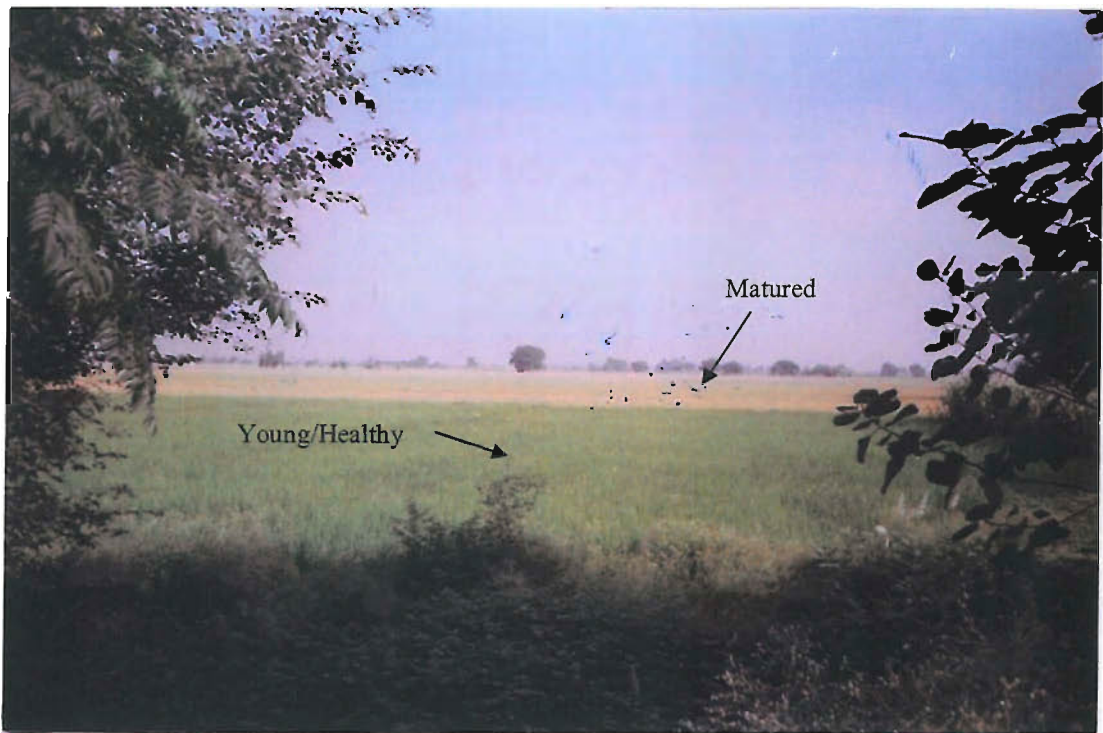


Figure 5.12: Rice fields showing matured crop in far end with nearer fields still green and healthy. This variation can be exploited to capture support vectors

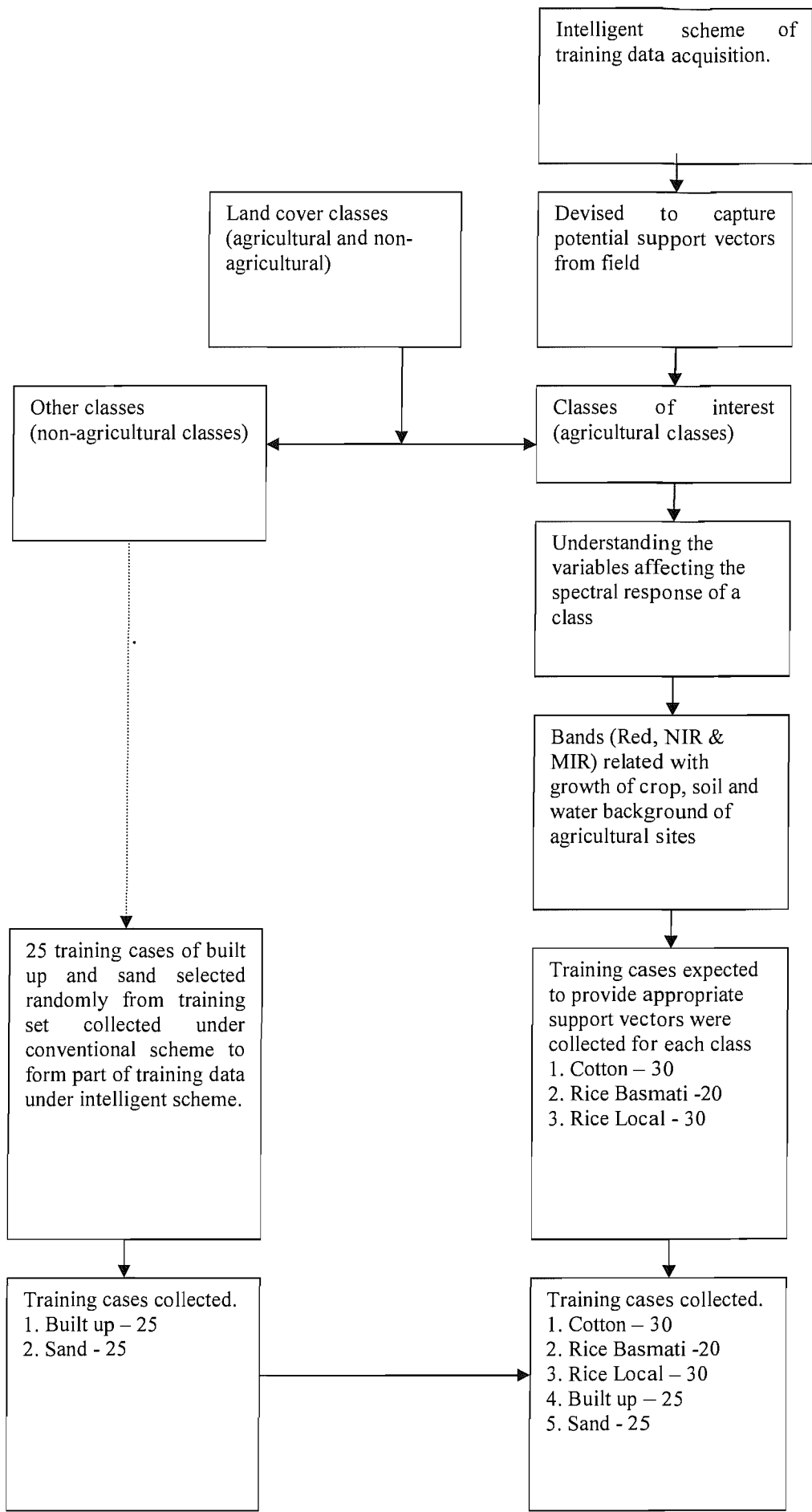


Figure 5.11: Procedure followed for training data acquisition under intelligent scheme. The scheme was tested on testing data acquired under conventional scheme.

The variables considered for collecting training data intelligently for the three crops cotton, rice basmati and rice local are enumerated in Tables 5.3, 5.4 and 5.5 respectively. The information regarding the status of crops were collected before going to field (section 5.3) and was also updated in-situ by consulting the farmers (Figure 5.13) so as to collect training data from locations that would provide appropriate support vector.

For the non-agricultural classes built-up and sand, the spectral variability was not well defined as for crops with the available spectral bands. As such, no informed guess could be made as regards the training sites that would yield support vectors, though there may be some relationship for example, for soil based on particle size. As non-agricultural classes were very distinct with respect to agricultural classes in feature space (Figure 5.9), only 25 training samples were selected randomly for each of these two classes from the training data collected under conventional scheme (Figure 5.11) to form part of training data under intelligent Scheme. The relative distribution of the classes in the feature space was expected not to result in any confusion between agricultural and non-agricultural classes in the classification process. The inclusion of training data of built up and sand classes from conventional scheme was intended to include training data of all the classes in



Figure 5.13: Farmers being consulted in field about the crop status in the area.



the intelligent scheme so that multi-class comparison of the two schemes could be made by training the classifiers with both the training schemes.

<b>General condition of crop/field from where support vectors were acquired</b>
<p>1. Generally the crop was in flowering stage throughout the study area (Figure 5.5).</p> <p>As such no variability due to growth of crop could be observed by naked eye in the field.</p>
<p>2. The soil was generally exposed to sky in the cotton fields (Figure 5.5).</p> <p>The contribution of soil to the spectral response of cotton crop comprised of (a) direct contribution (b) in the growth of the crop. So all soil types were considered to account for variability in soil type. In addition training data was also acquired from saline land (salt left on ground due to waterlogging) (Figure 5.14).</p>
<p>3. Generally the cotton fields were not watered</p> <p>The cotton fields were not watered (Figure 5.5) for frutification to take place. Training sites from near waterlogged areas or canals which have higher water table that would affect spectral response especially of water sensitive MIR band were considered.</p>

Table 5.3: Variables considered in the intelligent scheme of training data collection for cotton.

<b>General condition of crop/field from where support vectors were acquired</b>
<ol style="list-style-type: none"> <li>1. Generally the crop was young and healthy (Figure 5.15)</li> <li>2. The canopy of the crop was such that it did not permit exposure of soil to sky (Figure 5.16).</li> <li>3. The fields were generally watered as crop was young (Figure 5.15).</li> </ol> <p>There was thus no apparent variability that naked eye could notice in the field. However, training samples were collected from fields comprising all soil types and from near/away from canals that would affect the spectral response of the crop and yield support vectors.</p>

Table 5.4: Variables considered in the intelligent scheme of training data collection for rice basmati.

**General condition of crop/field from where support vectors for rice local were acquired**

1. Three different stages of maturity were noticed for the crop

Very matured-Dark yellow (Figure 5.17 a, b and c), Less matured-Light yellow (Figure 5.18 a and b), Young-Green (Figure 5.19 a and b).

The growth stages offer a defined relation with spectral response of a crop. For example, for healthy young crop, spectral response would be higher in NIR band and low in Red band but as the crop matures, there is a shift towards the red band, thereby reducing the value in NIR band and increasing in Red band as compared to young crop.

2. Training sites near water bodies especially of less (Figure 5.18a) and very matured (Figure 5.17c) varieties of local rice not watered in field would affect spectral response especially in MIR bands.

3. Different soil types (though soil has contributed only in the growth of crop and there was no direct contribution due to exposure to sky (Figure 5.21)). From saline land resulting from waterlogging (Figure 5.20)

Table 5.5: Variables considered in the intelligent scheme of training data collection for rice local.



Figure 5.14: Cotton crop in saline land. The white patches of salt due to waterlogging can be seen on the exposed soil. This was expected to increase the spectral response in all three bands.



Figure 5.15: Basmati rice was very young and green throughout the study area as such no variability could be observed in field by the naked eye.



Figure 5.16: The canopy of basmati rice does not permit soil to be exposed to sky. Thus the contribution of the soil in the spectral response of the crop could be considered only due to its contribution in the growth of the crop.



Figure 5.17a: Very matured local rice. NIR values would be low and Red higher as compared to a young healthy crop. Likewise MIR value would be higher as the crop was dry.





Figure 5.17b: Very matured local rice NIR values would be low and Red higher as compared to a young healthy crop. Likewise MIR value would be higher as the crop was dry.



Figure 5.17c: Very matured local rice adjoining a canal. Water reduces spectral response especially in MIR band.



Figure 5.18a: Matured local rice near canal. The leaves have started yellowing and grain formation has set in. Water reduces spectral response especially in MIR band.



Figure 5.18b: Matured local rice. Grain formation has taken place and yellowing of leaves has also started. The spectral values would be between young healthy and very matured local rice in similar conditions.





Figure 5.19a: Young local rice. The grain formation is there but leaves are still green. The NIR values would be high and Red very low as compared to matured crop.



Figure 5.19b: Young local rice. The grain formation is there but leaves are still green. The values in NIR would be high, low in Red and low in MIR (leaves were moist) as compared to matured crop.



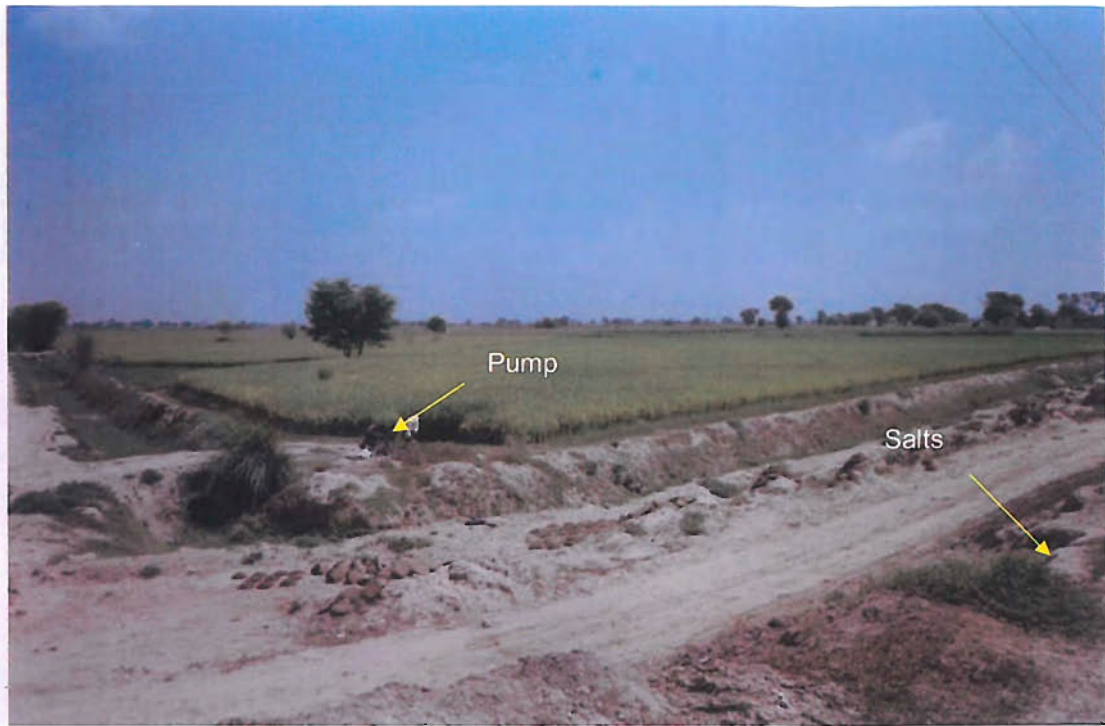


Figure 5.20: Local rice in area affected by waterlogging. The white salt can be seen on the soil (lower right corner of the photograph). The pump in the field is to drain out water due to waterlogging from the field out into surrounding drain.



Figure 5.21: Top view of local rice. The canopy does not expose soil to sky.

The spectral distribution of training data collected under intelligent scheme is given in Figure 5.22 and statistics in Table 5.6.

Class	Red Band (DN)				NIR Band (DN)				MIR Band (DN)			
	Min	Max	Mean	Sd	Min	Max	Mean	Sd	Min	Max	Mean	Sd
<b>Built-up</b>	81	104	96.16	5.5	81	109	97.56	6.76	129	189	156.68	13.21
<b>Sand</b>	101	128	116.00	6.88	99	124	113.40	6.54	159	199	178.96	8.98
<b>Cotton</b>	47	58	52.87	2.22	112	170	143.17	15.60	97	132	116.33	8.40
<b>Rice(basmati)</b>	49	65	57.95	3.93	83	110	96.70	5.69	85	103	93.55	5.41
<b>Rice (local)</b>	44	73	55.73	8.06	89	138	108.93	13.09	83	119	98.47	10.31

Table 5.6: Statistics of training data showing minimum (Min), maximum (Max), Mean and standard deviation (Sd) of digital numbers of the training data of the five classes in the three bands.

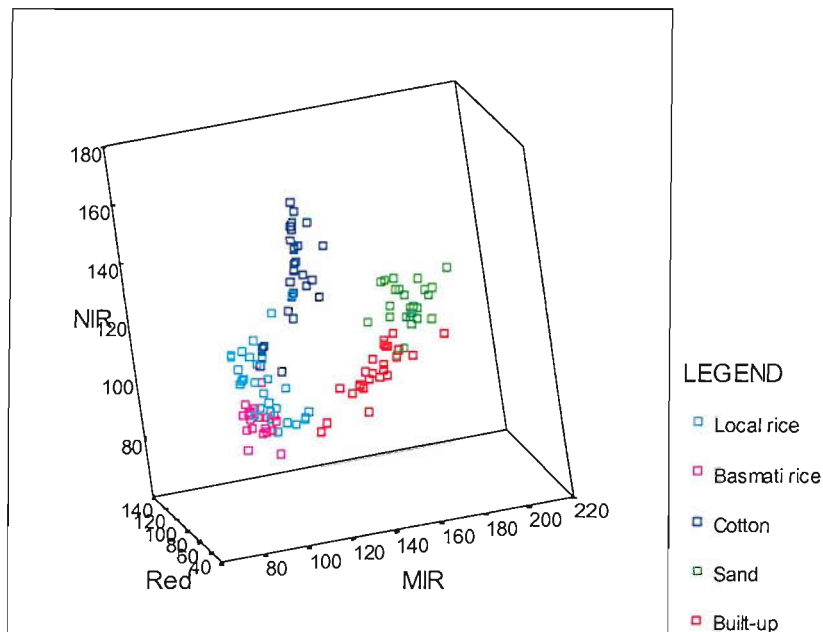


Figure 5.22: spectral distribution of training data collected under intelligent scheme.

## 5.4 Methodology of Classification

There are many factors that affect the accuracy of an image classification. This study examines the effect of training set acquired under conventional scheme and the intelligent scheme of training data acquisition along with classification algorithms used on classification accuracy (Table 5.7). Though the intelligent scheme was tailored for SVM classification, the intention of using other classifiers DA, DT and ANN in the study was to

assess the effect of intelligent scheme on the accuracy of these classifiers. In addition, the intelligent scheme was devised to acquire potential border training data and studies have shown that border training data are more important for classification by an ANN classifier (Foody, 1999), the DT uses extreme cases in the splitting rules to make data more homogeneous, as such the inclusion of other classifiers especially ANN and DT were justified to test the accuracy trained with intelligent scheme of training data acquisition.

The four classifiers were trained with both the conventional and intelligently defined training sets. The classification accuracy of the trained classifiers were tested on the same testing set which were acquired under the conventional scheme (section 5.3.1). The accuracy statements of classifications derived from both the analyses (conventional and intelligent scheme of training data acquisition) was compared in a rigorous fashion using M<sup>c</sup>Nemar test that accommodated the testing samples for the related nature in the analyses (section 3.1.3).

Variables	Scenarios investigated
Training set	<p>A) Conventional sampling scheme: training set comprised of 90 pixels per each class.</p> <p>B) Intelligent scheme: Training set comprised of variable number of cases per class as detailed below:</p> <ul style="list-style-type: none"> <li>-Built-up 25 cases</li> <li>-Sand 25 cases</li> <li>-Cotton 30 cases</li> <li>-Rice Basmati 20 cases</li> <li>-Rice Local 30 cases</li> </ul> <p>The number of training cases acquired depended upon the cases likely to provide appropriate support vectors which varied for each class.</p>
Classification algorithms	<p>Discriminant analysis</p> <p>Artificial neural network</p> <p>Decision tree</p> <p>Support vector machine</p>
Testing set	<p>Testing set comprised of 90 pixels per each class collected along with training set in conventional scheme of training data acquisition.</p>

Table 5.7: Variables considered in the study.

## **5.5 Results and Discussions**

### **5.5.1 An Assessment of Ability to Intelligently Identify Most Useful Training Samples (Support Vectors) Directly from Field**

The intelligent scheme was tailored to capture training data from sites which would provide appropriate potential border training data (potential support vectors) (Figure 5.23). This was driven by the desire that the SVM needs only training data that are located on the border of the spectral distribution of the classes in feature space. The intelligent scheme was devised for agricultural classes (section 5.3.2) as the variables affecting the spectral response are well understood for agricultural classes. Thus external knowledge of crop status, soil background and water condition of the agricultural fields were used to select suitable training sites to obtain potential border training data (Tables 5.3, 5.4 and 5.5).

The training data collected under the intelligent scheme was overlaid with those collected under the conventional scheme (Figure 5.23) to visualize if the training data collected under intelligent scheme was successful in capturing border training data (potential support vectors).

The intelligent scheme was successful in capturing border training data for agricultural classes especially cotton and local rice (Figure 5.23). However, for basmati rice, there was hardly any variability that eye could notice in field and, therefore, the training data collected for basmati rice under intelligent scheme did not fully describe the border of training data as collected by conventional scheme in feature space. Further examination of Figure 5.23 shows that the element of external knowledge involved in intelligent scheme helped to capture potentially border training data especially very matured local rice (Figure 5.23), which could not be captured by the conventional scheme. This can be attributed to the reason that the very matured variety of local rice could be found only in small parts of the study area and, therefore, the training set (90 cases per

class) devised for conventional scheme was not large enough to acquire training data from very matured variety of local rice. Perhaps, a still larger training set could have acquired training data pertaining to very matured variety of local rice under conventional scheme of training data acquisition.

The distribution of training data collected under intelligent scheme over conventional scheme (Figure 5.23) shows that the intelligent scheme was successful in collecting training data from border regions of the spectral distribution of the agricultural classes in feature space. However, the success of intelligent scheme over the conventional scheme of training data acquisition needs to be confirmed by training the various classifiers, especially the SVM and testing on the same testing set for comparative analysis.

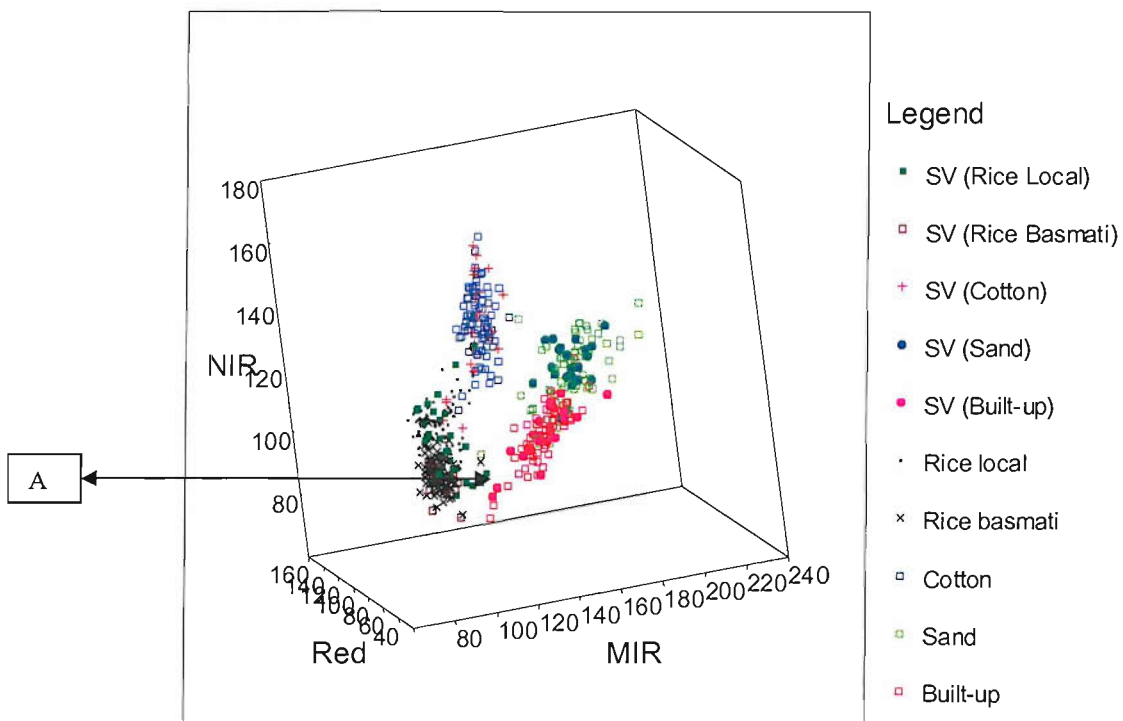


Figure 5.23: Training data of conventional scheme overlaid by that captured under intelligent scheme. The prefix SV in labels in the legend refers to training data collected by intelligent scheme. The 'A' refers to training data from site with very matured local rice collected under intelligent scheme.

## 5.5.2 Relative Accuracy

The overall classification accuracy obtained on the testing set by different classifiers trained with training data acquired under conventional scheme were compared with accuracies obtained with corresponding classifiers trained with data acquired under intelligent scheme to understand the effect of the nature of training set and classifiers used on classification accuracy. Section 5.5.2.1 to 5.5.2.4 focuses on accuracy obtained by the four classifiers DA, ANN, DT and SVM respectively, trained by training data collected by conventional and by intelligent scheme. The comparison of the results of the four classifiers is followed in section 5.5.2.5.

### 5.5.2.1 Discriminant Analysis

Table 5.8 and 5.9 shows the confusion matrix of testing set when DA was trained by training data collected by conventional and by intelligent scheme respectively.

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	82	8	0	0	0	90
Sand (S)	18	72	0	0	0	90
Cotton (C)	0	0	87	0	3	90
Rice Basmati (RB)	0	0	0	83	7	90
Rice Local (RL)	0	0	3	13	74	90
Total	100	80	90	96	84	450

Overall accuracy=88.44%

Table 5.8: Error matrix of testing set for the classification derived from the discriminant analysis (DA) trained by data acquired under conventional scheme.

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	83	7	0	0	0	90
Sand (S)	19	71	0	0	0	90
Cotton (C)	0	0	88	0	2	90
Rice Basmati (RB)	0	0	0	76	14	90
Rice Local (RL)	0	0	4	4	82	90
Total	102	78	92	80	98	450

Overall accuracy=88.88%

Table 5.9: Error matrix of testing set for the classification derived from the discriminant analysis (DA) trained by data acquired under intelligent scheme.

Comparative analysis of the two tables (Table 5.8 and Table 5.9) shows that overall accuracy obtained by DA was very similar for both the training schemes, conventional and intelligent. The differences in accuracy between the classifications trained by conventional and intelligent scheme were statistically not significant at 95 % confidence level (Table 5.16). This can be attributed to the reason that statistical parameters generated by both the training schemes (Table 5.2 and Table 5.6) were very similar. The examination of the two tables (Table 5.2 and Table 5.6) shows that the statistical parameters for the first three classes were very similar for both the training schemes resulting in very similar class accuracies for the first three classes for DA trained with either training schemes. However, for the last two classes (Rice local and Rice basmati), statistical parameters generated from the two training schemes were not similar and were reflected in their varied accuracies for the classes (local rice and basmati rice) for DA trained with conventional and intelligent schemes.

### 5.5.2.2 Decision Tree

Table 5.10 and 5.11 shows the confusion matrix of testing set when DT was trained by training data collected by conventional and by intelligent scheme respectively.

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	76	14	0	0	0	90
Sand (S)	19	71	0	0	0	90
Cotton (C)	0	0	82	0	8	90
Rice Basmati (RB)	0	0	0	78	12	90
Rice Local (RL)	0	0	2	13	75	90
Total	95	85	84	91	95	450

Overall accuracy = 84.88%

Table 5.10: Error matrix of testing set for the classification derived from the decision tree (DT) trained by data acquired under conventional scheme.

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	78	12	0	0	0	90
Sand (S)	11	79	0	0	0	90
Cotton (C)	0	0	66	0	24	90
Rice Basmati (RB)	0	0	0	66	24	90
Rice Local (RL)	0	0	5	16	69	90
Total	89	91	71	82	117	450

Overall accuracy=79.55%



Table 5.11: Error matrix of testing set for the classification derived from the decision tree (DT) trained by data acquired under intelligent scheme.

Comparative analysis of the two tables (Table 5.10 and Table 5.11) shows that overall accuracy obtained by DT trained with intelligent scheme of training data acquisition was less as compared to one trained with training data from conventional scheme. The differences in accuracy between the classifications trained by conventional and intelligent scheme were statistically significant at 95 % confidence level (Table 5.16). This can be attributed to the reason that DT is a non-parametric classifier and the node splitting rules to make child nodes purer is based on the extreme spectral values of the training data ((Figure 5.24, Table 5.2) and (Figure 5.25 and Table 5.6)) of the various classes. The small intelligently selected training data under intelligent scheme provided extreme values (for agricultural classes especially local rice, Figure 5.23) and, therefore, the classifier had more extreme values as compared to conventional scheme of training data (Table 5.6 and Table 5.2 respectively). The extreme values under intelligent scheme, therefore, provided more overlap between agricultural classes in feature space and, therefore, more confusion with DT classification.

The node splitting rules in decision tree trained with intelligent scheme, therefore, had more extreme cases (Figure 5.25) as compared to one trained with conventional scheme (Figure 5.24). This resulted in more overlap in decision rules between classes in feature space and, therefore, more confusion when trained with training data from intelligent scheme. This made DT very sensitive to the nature of training scheme used to acquire training samples. The overall accuracy for DT trained with conventional scheme of training data decreased from 84.88 per cent to 79.55 per cent when trained with intelligent scheme. This difference was very pronounced for agricultural classes.



### 5.5.2.3 Artificial Neural Networks

Table 5.12 and 5.13 shows the confusion matrix of testing set when ANN was trained by training data collected by conventional and by intelligent scheme respectively.

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	85	5	0	0	0	90
Sand (S)	13	77	0	0	0	90
Cotton (C)	0	0	85	0	5	90
Rice Basmati (RB)	0	0	0	78	12	90
Rice Local (RL)	0	0	1	8	81	90
Total	98	82	86	86	98	450

Overall accuracy = 90.22%

Table 5.12: Error matrix of testing set for the classification derived from the artificial neural network (ANN) trained by data acquired under conventional scheme.

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	81	6	0	0	3	90
Sand (S)	16	74	0	0	0	90
Cotton (C)	0	0	88	0	2	90
Rice Basmati (RB)	0	0	0	77	13	90
Rice Local (RL)	0	0	5	7	78	90
Total	97	80	93	84	96	450

Overall accuracy = 88.44%

Table 5.13: Error matrix of testing set for the classification derived from the artificial neural network (ANN) trained by data acquired under intelligent scheme.

Comparative analysis of the two tables (Table 5.12 and Table 5.13) show that overall accuracy as well as individual class wise accuracies obtained by ANN were very similar for both the training schemes, conventional and intelligent. The differences in accuracy between the classifications trained by conventional and intelligent scheme were statistically not significant at 95 % confidence level (Table 5.16). This can be attributed to the findings (Foody, 1999) that border training data are more important than core for classification undertaken by ANN classifier. This, therefore, resulted in very similar accuracy for ANN trained with intelligent scheme (training data collected with potential border cases of classes in feature space) with that collected by conventional scheme.

For agricultural classes only, for which the intelligent scheme was tailored, conventional scheme resulted in 244 pixels correct (Table 5.12), whereas the intelligent scheme provided 243 pixels correct (Table 5.13) for SVM classification. Thus the accuracy of agricultural classes were very similar for both the training schemes.

#### 5.5.2.4 Support Vector Machine

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	89	1	0	0	0	90
Sand (S)	15	75	0	0	0	90
Cotton (C)	0	0	88	0	2	90
Rice Basmati (RB)	0	0	0	82	8	90
Rice Local (RL)	0	0	3	7	80	90
Total	104	76	91	89	90	450

Overall accuracy = 92.00%

Table 5.14: Error matrix of testing set for the classification derived from the Support Vector Machine (SVM) trained by data acquired under conventional scheme.

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	88	2	0	0	0	90
Sand (S)	18	72	0	0	0	90
Cotton (C)	0	0	88	0	2	90
Rice Basmati (RB)	0	0	0	78	12	90
Rice Local (RL)	0	0	4	4	82	90
Total	106	74	92	82	96	450

Overall accuracy = 90.66%

Table 5.15: Error matrix of testing set for the classification derived from the Support Vector Machine (SVM) trained by data acquired under intelligent scheme.

Comparative analysis of the two tables (Table 5.14 and Table 5.15) show that overall accuracy as well as individual class wise accuracies obtained by SVM was very similar for both the training schemes, conventional and intelligent. The differences in accuracy between the classifications trained by conventional and intelligent scheme were statistically not significant at 95 % confidence level (Table 5.16).

The two analysis used only a fraction of the input training data, the conventional used 215 training data as support vectors from a total of 450 (47.7 %) input training set (51, 52, 28, 38, 46 support vectors for class Built-up, Sand, Cotton, Rice basmati and Rice

local respectively) whereas, the intelligent scheme used 76 training data as support vectors from a total of 130 input training set (11, 9, 11, 18, 27 support vectors for class Built-up, Sand, Cotton, Rice basmati and Rice local respectively). However, for intelligent scheme which was devised for agricultural classes 56 out of 80 (70 %) training samples were used as support vectors.

Support vectors central to the establishment of decision surfaces in SVM could be successfully captured under intelligent scheme of training data acquisition especially for cotton and local rice. However for rice basmati, there was hardly any variability that eye could notice in fields (table 5.4) and, therefore, the intelligent scheme was not very successful in capturing potential support vectors for rice basmati class (Figure 5.22). Thus, the unavailability of proper support vectors for rice basmati class resulted in confusion of rice basmati with rice local class (class facing rice basmati class in feature space) in the classification process (Table 5.15).

For agricultural classes only, for which the intelligent scheme was tailored, conventional scheme resulted in 250 pixels correct (Table 5.14), whereas the intelligent scheme provided 248 pixels correct (Table 5.15) for SVM classification. Thus the accuracy of agricultural classes were very similar for both the training schemes, though training data was reduced by more than two-thirds from 270 cases for conventional scheme to 80 pixels under the intelligent scheme.

#### **5.5.2.5 Discussion**

The results obtained for the four classifiers with focus on SVM are discussed hereafter for classifications trained for conventional scheme in next section, followed by intelligent scheme in section 5.5.2.5.2 and finally the comparative analysis of the two schemes in section 5.5.2.5.3.

#### **5.5.2.5.1 Analysis of Classifications Trained with Conventional Scheme of Training**

##### **Data Acquisition**

From the range of classifications undertaken, the highest overall accuracy of 92 % was obtained from the SVM. Moreover, this classification was significantly more accurate than that derived from DA and DT (Table 5.17) at the 95 % confidence level.

The sensitivity of the SVM classification to the nature of the training sample is also evident (Table 5.18) which shows that SVM classification are based on a fraction input training data that lie on part of the edge of class distribution in feature space. The SVM used only 215 training samples out of a possible 450 as support vectors for training (Table 5.18).

There was no confusion between agricultural and non-agricultural classes by any of the four classifiers. This is due to the reason that agricultural classes and non agricultural classes were spectrally distinct in feature space (Figure 5.9). However, SVM in general produced the most accurate classification for all the agricultural classes as compared to other classifiers.

#### **5.5.2.5.2 Analysis of Classifications Trained with Intelligent Scheme of Training**

##### **Data Acquisition**

From the range of classifications undertaken, the highest overall accuracy of 90.66 % was obtained from the SVM. Moreover, this classification was significantly more accurate than that derived from DT and ANN (Table 5.17) at the 95 % confidence level.

The sensitivity of the SVM classification to the nature of the sample is also evident (Table 5.18) which shows that SVM classification used only a fraction of the input training data. SVM used only 76 training pixels as support vectors (Table 5.18).

The accuracy obtained by ANN and SVM were very similar as expected. Studies have shown that border training data are more important in classification with ANN

(Foody, 1999) and training data captured under intelligent scheme strived to collect border training data.

DA, ANN and SVM correctly classified 88 pixels of cotton out of 90 but the SVM used only 11 pixels of cotton as support vectors (Table 5.18) as compared to all the available training data of cotton (30 pixels) by the other three classifiers.

The small intelligently collected training data under the intelligent scheme was devised for agricultural classes but for non-agricultural classes (built-up and sand) for which no educated guess could be made (may be possible with certain attributes such as particle size for sand) in the present study as regards the sites that would provide appropriate support vectors, 25 training samples for each class were included from conventional scheme. This was intended so that multi-class comparison could be undertaken for the two schemes of training data acquisition. The size of 25 training samples were chosen as the spectral distribution of the two classes (built-up and sand) (Figure 5.9) were very distinct with agricultural classes in feature space. This relative distribution of the classes was expected not to result in any confusion between agricultural and non-agricultural classes in the classification process. There was, as expected no confusion between agricultural and non-agricultural classes by any of the four classifiers used and, therefore, justifies the selection of 25 training samples each for non-agricultural classes. In addition, it can be argued that the information on built-up and sand is often available in GIS format or from earlier image analysis and can be masked out from the study area. This implies that from practical consideration mapping agricultural classes are of paramount importance. The intelligent scheme devised for agricultural classes, therefore, served its purpose.

### 5.5.2.5.3 Comparison between Classifications Trained with Conventional and Intelligent Scheme of Training Data Acquisition

SVM provided the highest accuracy for both the training schemes (Table 5.14 and Table 5.15). The SVM used only a fraction of input training data for both the training schemes. However, the intelligent scheme devised for agricultural classes was successful in its intent to capture support vectors directly from field as the SVM classification used 56 out of 80 (70 %) training samples collected under intelligent scheme for agricultural classes as support vectors as compared to 215 out of 450 (47.7 %) for all the classes used by the conventional scheme (section 5.5.3.3).

DT was most sensitive of all the four classifiers to the nature of the training data (Table 5.16). The differences in accuracy between the DT classifications trained by conventional and intelligent scheme were statistically significant at 95 % confidence level (Table 5.16).

However, there was no confusion between agricultural and non-agricultural classes by any of the four classifiers trained by either of the training acquisition schemes.

Classifiers	Z values
	Conventional ' v Intelligent scheme
Discriminant analysis	-0.447
Decision tree	<b>2.650</b>
Artificial neural network	1.290
Support vector machine	1.500

Table 5.16: Significance value (Z) of differences between accuracies of testing set obtained when the classifiers were trained with training data collected under Conventional and by Intelligent scheme of training data collection. Differences significant at the 95% confidence level ( $Z \geq 1.96$ ) are highlighted in bold with positive values indicating higher accuracy when classifier trained with training data collected under conventional scheme.

Training scheme	SVM v DA	SVM v DT	SVM v ANN	ANN v DA	ANN v DT	DT v DA
Conventional	<b>3.138</b>	<b>5.600</b>	1.7060	1.410	<b>4.110</b>	<b>-2.470</b>
Intelligent	0.125	<b>5.976</b>	<b>2.0412</b>	-0.426	<b>4.714</b>	<b>-4.817</b>

Table 5.17: Comparison of classification accuracy statements for classifications trained with conventional and intelligent scheme of training data acquisition (SVM = support vector machine, DA = discriminant analysis, DT = decision tree and ANN =artificial neural network). Differences significant at the 95% confidence level ( $Z \geq |1.96|$ ) are highlighted in bold with positive values indicating that the first named classifier had the higher accuracy.

Training scheme	DA	DT	ANN		SVM		
	Acc (%)	Acc(%)	Acc(%)	Architecture	Acc(%)	parameters	SV's
Conventional	88.44	84.88	90.22	3:8:5	92.00	C=0.25 $\gamma=0.005$	215
Intelligent	88.88	79.55	88.44	3:11:5	90.66	C=1 $\gamma=0.000625$	76

Table 5.18: Parameters used to model the classifiers. The architecture in ANN describes the input layers, the nodes in the middle layer and the output layers. The network's architecture were defined from an evaluation of several hundreds of candidate networks. The parameters for SVM were chosen with the intent to maximize accuracy on testing set.

### 5.5.3 Identifying Support Vectors with Ground Attributes of the Training Sites

The SVM analysis detailed in Chapter 4 demonstrated that the support vectors of wheat class were mainly derived from brown soils. In particular, this knowledge may allow small intelligently selected training samples to be derived from regions with particular soil type for future analysis without loss of classification accuracy. Thus in situations, when support vectors of a class can be identified with a particular variable associated with the spectral response of a class (*e.g.*, for crops, growth or background properties of training sites such as soil or water) can be exploited to direct training acquisition activities to regions most likely to furnish support vectors for future analysis.

The intelligent scheme of training data acquisition (section 5.3.2) for agricultural classes was based on ancillary information on the growth status and background conditions (soil type and water) of the training sites. The intelligent scheme used 56 out of 80 training samples of agricultural classes as support vectors (section 5.5.2.4).

If, however, support vectors resulting from SVM classification based on intelligent scheme of training data acquisition can be related with ancillary information on the growth status of crops or ground attributes (like soil or moisture status) of training sites, there is scope of further reducing the requirement of training data over and above the small intelligently acquired training data collected under the intelligent scheme for future analysis. Thus the knowledge gained about the relationship of support vectors with ancillary information from SVM classification can be exploited in case the analysis is repeated in future to focus the training data acquisition process to the regions most likely to furnish support vectors.

The analysis detailed in chapter 4 shows that the support vectors of wheat class were mainly derived from only one soil type and, therefore, provides an opportunity to acquire training data from sites with particular soil type for future analysis. It was expected based on the experience of chapter 4 that support vectors for cotton crop being studied would be related with a particular soil type. This was hypothesised as the soils in the cotton fields were exposed to sky (Figure 5.5) and were expected to contribute directly to the spectral response of the cotton crop in addition to its contribution in the growth of the cotton crop unlike the case in Feltwell study (chapter 4) where soil was not exposed to sky and contributed only in the growth of the crop. However, the examination of support vectors along with ancillary information (Table A57) suggested that in general support vectors of cotton crop were derived from training sites located near the water bodies like canals or waterlogged areas and not soil as anticipated. This can be attributed to the reason that cotton crop was generally not watered, to avoid attack by pathogens and pests for frutification to result (section 5.2.1). The training data collected from near canals or waterlogged surfaces for cotton crop resulted in higher moisture content of the soil, reducing spectral response of cotton crop especially in MIR band. These training data of cotton crop, therefore, were located between the dry cotton crop (majority condition for cotton crop) and local rice in feature space and formed support vectors for cotton crop.



In all 10 out of 11 support vectors of cotton crop (Table A57) were derived from near waterlogged or near canals. Thus there was only one exception; one of the support vectors of the cotton crop was derived from dry soil (Table A57). Moreover, this single support vector drawn from the region of dry soil had a small  $\alpha_i$  value of 0.6051 (Table A57). The contribution of the support vectors to the establishment of the optimal separating hyperplane (OSH) is directly proportional to its  $\alpha$  value as is evident from equation 2.69, with training samples for which  $\alpha_i=0$  making no contribution to the fitting of the hyperplane. Thus, training samples with  $\alpha_i=0$  carry no useful information, unlike those that lie in the border region of classes in feature space where  $\alpha_i$  tends to its maximum value (1 for cotton class in the present case as the value of parameter  $C$  is 1). Thus majority of support vectors (10 out of 11) of cotton crop were drawn from training sites with wet conditions with the highest possible  $\alpha_i$  value of 1 (Table A57).

The lone support vector of cotton crop drawn from dry soil having a small  $\alpha$  value of 0.6501 had a very meagre contribution to the fitting of the hyperplane between cotton and local rice crop. Thus removing this lone support vector of cotton crop drawn from dry soil was expected not to have a significant influence on the location /orientation of the hyperplane between cotton and local rice classes and thereby the classification accuracy. The analysis was, therefore, extended to appreciate if the support vectors of cotton crop derived only from waterlogged or near canal (*i.e.* wet conditions) provided the same accuracy as SVM trained with all training data of cotton crop (including training data drawn from dry soil (majority condition) of the cotton crop). For this, training data of cotton crop with wet ground conditions was only retained for training the SVM classifier. Repeating the SVM classification but with the training samples of cotton crop drawn from near canals or waterlogged areas (wet conditions of training sites) only resulted in the same set of class allocations (Table 5.15) as made for the testing cases as before ( $Z=0$ ). Thus, classification accuracy was maintained despite the exclusion of training samples of

cotton crop from dry soils.

The result indicates that training data for cotton needs be acquired only from near water bodies (near canals or waterlogged areas) if the analysis is repeated in future. In this way, an accurate classification may be undertaken with SVM classifier using a small training set for cotton crop derived from a small spatial area near waterlogged or from near canals.

#### **5.5.4 Financial Implication of Reducing the Requirement of Training Data**

The objective in remote sensing classification process should be to provide an accurate land cover product keeping the whole process as economical as possible. The analysis carried in the work was designed to reduce the requirement of training data under intelligent scheme of training data acquisition and was thus aimed to reduce the cost of classification process as compared to one based on a large training set collected under conventional training data acquisition scheme. Both the approaches intelligent and conventional had very different requirements especially with regard to training acquisition process. There are a number of steps involved in image classification process and efforts should be made to reduce the costs involved at each step if possible. The costs involved in classification analysis can be broken into four broad parts:

1. Set up costs (costs for acquiring hardware and software)
2. Field survey costs (vehicle costs and costs incurred on personnel involved)
3. Image acquisition costs (cost of remotely sensed data)
4. Time spent on analysis of field data and processing of imagery (costs on computer and image analyst time)

The potential to reduce the requirement of training data as detailed under the intelligent scheme has direct bearing on cost associated with field survey and also on time spent on the analysis as compared to the one carried with conventional training data scheme. The other two costs (set up costs and cost to acquire imagery) as detailed above

were, however, same for conventional and intelligent training scheme as both the analysis were carried on the same setup (hardware and software) and on the same imagery. The approximate costs incurred in the two classification analysis are tabulated in Table 5.19.

Table 5.19 shows that the intelligent scheme of training data acquisition was cheaper by Rs. 9288 (26.09 per cent) over the conventional scheme. The difference in cost for the classification process with the two schemes was mainly on account of preparatory work and distance traversed on ground to acquire training sites. The conventional scheme needed plot outs of selected segments (section 5.3.1) on transparent sheet for each class to be overlaid on the reference maps of the study area to locate the selected segments on ground so as to acquire sites for training and testing data. Thus not only was the conventional scheme costly as compared to intelligent scheme but at the same time had more paper work in the form of plot outs from plotter. The conventional scheme was based on the concept of capturing the spectral variability of the classes in feature space by the sheer large number of training sets. However, the intelligent scheme was devised to capture a small number of training set aided by external knowledge on factors like crop growth and background properties of training sites (soil type and water) and, therefore, made training data acquisition under intelligent scheme more scientific and interesting in field.

The calculation enumerated in Table 5.19 are, however, based when both training and testing data were collected together at the same time in the conventional scheme (section 5.5.1) but this difference in costs for classification undertaken with intelligent and conventional schemes of training acquisition would magnify if training and testing data had been collected independently of each other. Keeping this scenario in mind, the expenditure likely to incur if training and testing data were collected independently were recalculated. For this situation, the only difference with one tabulated in Table 5.19 would be in distance travelled for collecting training data. For efforts to acquire testing data are same for both the schemes, hence negated. The distance to be traversed to acquire only

training data under conventional and intelligent scheme would have been 1700 Km and 1040 Km respectively. The overall costs for the two schemes conventional and intelligent would have been Rs 33000 and Rs. 23390 respectively. This makes intelligent scheme cheaper by Rs 9610/- over conventional scheme or 29.12 per cent over the conventional scheme.

	Random	Intelligent
<b>Field survey</b>		
-Preparatory work	(a) Land cover map (Rs. 20,000) (b) Preparation of random segments (one man day = Rs. 600) (c) Plot outs of selected segments on 1:250000 sheets (Rs. 150 per sheet, totalling Rs 1800=00 for 12 sheets)	(a) Soil map (Rs 16000)  (b) Gathered information about growth status of crops, waterlogging through (i) news paper (ii) agriculture departments (iii) farmers (one man day Rs 600)
<b>Travelling</b>		
-In field	(a) Distance travelled 2442 Km (@Rs.3.5/Km = Rs.8547)	(a) Distance travelled 1874 Km (@Rs.3.5/Km = Rs.6559)
<b>Analysis</b>		
-Training data extraction, data formatting and SVM classification carried on personal computer (PC).	(a) 31 hours (@Rs. 150/hr for PC = Rs.4650)	(a) 21 hours (@Rs. 150/hr = Rs.3150)
<b>Total</b>	<b>Rs.35597</b>	<b>Rs.26309</b>

Table 5.19: Comparison of expenditure incurred on SVM classification process based on conventional with intelligent scheme of training data acquisition. The cost has been calculated in Indian Rupees with approximate rates prevalent in India for the work, though part of the analysis has been carried in United Kingdom too. (1 US dollars (USD) = Rs 43.40 and 1 United Kingdom Pounds (GBP) = Rs 80.26 as on 15 May 2005).

### 5.5.5 Reduced Requirement of Training Data for Classifying Accurately only One Class.

The analysis detailed in this chapter so far was directed towards the production of a standard (multi-class) land cover map. However, often the concern is to map accurately only one class from the many land cover classes available. For example, the Large Area

Crop Inventory (LACIE) project of United States Geological Survey (USGS) was concerned only with the wheat crop. The intention in such projects such as LACIE was to map accurately the class of interest without any regard to the accuracy of the other land cover classes in the area.

Conventional classifiers require training set for all the classes even if the project is focused on just one class, as in the CAPE project (section 5.2.1.1). However, training data of all the land cover classes present in the study area may not be required if the concern is to accurately map only one class from the many land cover classes available if using SVM as a classifier.

For classification by an SVM, only the training samples that are support vectors, which lie on part of the edge of the class distribution in feature space, are required; all other training samples provide no contribution in fitting the decision boundary. Thus to accurately map only one class, training data are required only from the class of interest and from classes facing the class of interest in feature space. Training data of classes not facing the class of interest in feature space do not contribute in the establishment of the SVM classifier classifying the class of interest and are, therefore, redundant. This is illustrated here with reference to the classification of cotton.

Cotton has been chosen as the class of interest in the present study (section 5.3) as the satellite sensor data used was acquired for CAPE project for accurately mapping only the cotton crop. The remotely sensed data of late September were chosen on the premise that cotton crop was at its maximum vegetative phase at that time and was thus expected to be most spectrally separable from other crops in the area.

The spectral distribution of training data in feature space acquired under intelligent scheme of training data acquisition (Figure 5.22) shows that rice basmati class did not face class cotton in feature space. Since, the focus in SVM classifier lies on the border region between the classes in feature space to establish the optimal hyperplane, it may be anticipated that for cotton crop, the training samples drawn from rice basmati class would

not contribute to the establishment of SVM classifier separating cotton class with other classes. Thus to accurately classify only one class from the many land cover classes available using an SVM classifier, training data were required from the class of interests and from classes facing the class of interest in feature space. The exclusion of training data of rice basmati class from training the SVM classifier was thus not expected to influence the accuracy of classification accuracy of cotton crop.

Two SVM classifications were trained to understand the effect of excluding training data of rice basmati class on the classification accuracy of cotton crop. First, an SVM trained with all the classes. Second, an SVM trained with training data of all the classes except rice basmati class. Both the trained SVM classifiers were tested on the same testing set as acquired under the conventional scheme (section 5.3.1) of training data acquisition that included samples from all classes including rice basmati class.

The error matrices of testing set when SVM classifier was trained with all the classes (Table 5.15) and trained with all classes except rice basmati class (Table 5.20) shows that accuracy of cotton crop classification (97.7 %) was maintained for both the cases. Thus excluding training data of rice basmati class from training the SVM classifier did not affect the accuracy of the cotton crop.

However, when the training set without basmati rice was used, all cases of rice basmati were classified as rice local for the testing set (Table 5.20). The classification of rice basmati into rice local was expected as all the cases of basmati rice were located towards the rice local side of the SVM classifier established by training data of cotton and rice local class. Thus for mapping cotton crop in future in the study area, there is no need to acquire training data of rice basmati as it does not contribute in establishing SVM classifier separating the cotton class from other classes.

The exclusion of training data of rice basmati class for accurately mapping cotton crop is contrary to the requirements of many classification projects aimed to accurately

Actual class	Predicted class					Total
	B	S	C	RB	RL	
Built-up (B)	88	2	0	0	0	90
Sand (S)	18	72	0	0	0	90
Cotton (C)	0	0	88	0	2	90
Rice Basmati (RB)	0	0	0	0	90	90
Rice Local (RL)	0	0	4	0	86	90
Total	106	74	92	0	178	450

Table 5.20: Error matrix of testing set for the classification derived from the support vector machine (SVM) trained by data acquired under intelligent scheme for all classes except rice basmati.

classify only one class from the many land cover classes, for example CAPE project where objective is to accurately map only the class of interest but training data are acquired from all the land cover classes in the study area (section 5.2.1.1).

The analysis demonstrate that for accurately mapping cotton crop only, there is no need to acquire training data from class basmati rice as the class of interest cotton did not share any boundary with basmati rice in feature space (Figure 5.22). Thus in cases where the interest is in accurately classifying only one class, the requirement of training data can be identified from the relative distribution of training data in feature space, excluding training data of classes not facing the class of focus in feature space. Thus classification accuracy of class of interest can be maintained despite the reduction in training size.

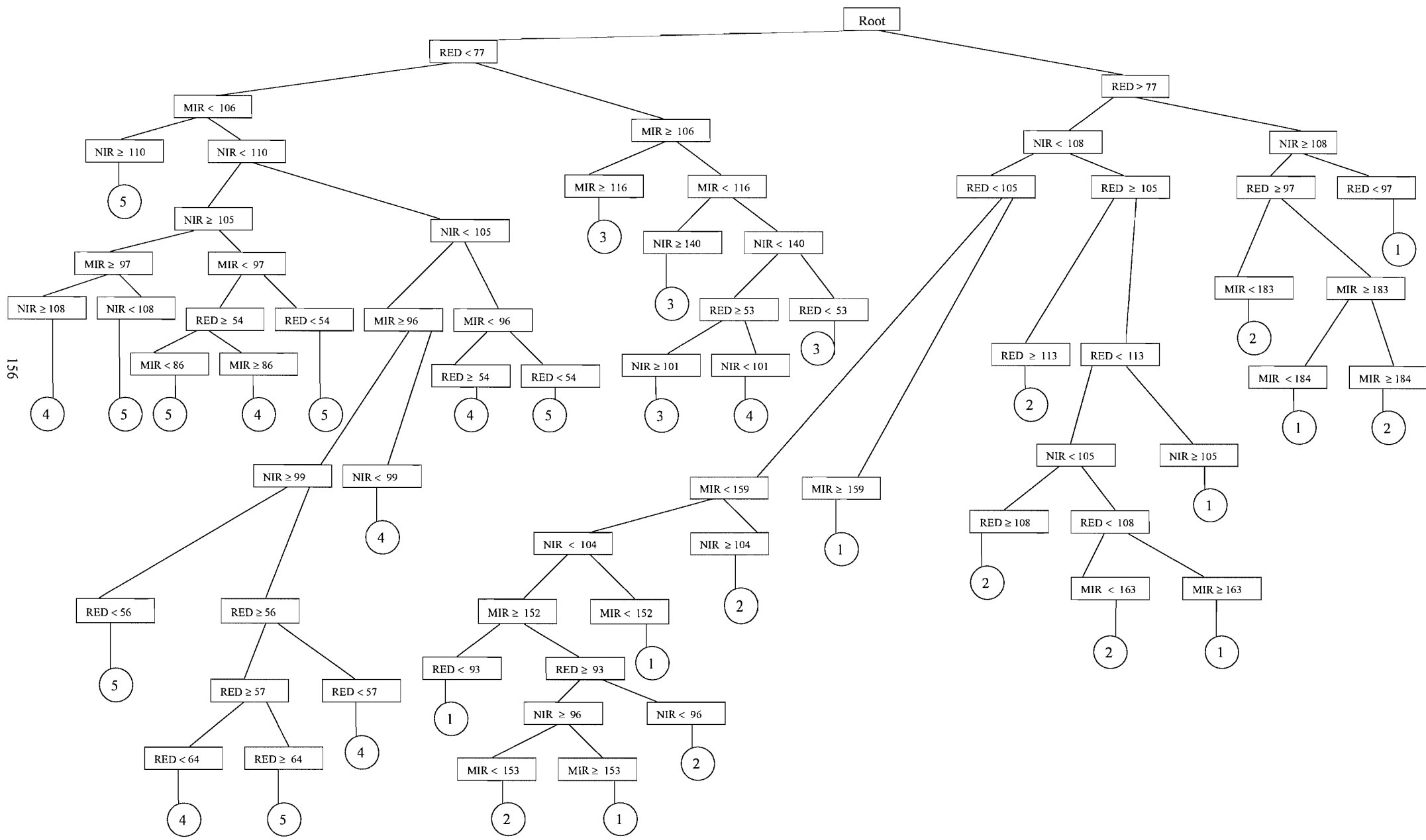


Figure 5.24: Tree structure when DT was trained by training data collected under conventional scheme. Each box is a node with root at the top which contains all the training data. Splitting rules used the values in the three input bands (Red, NIR and MIR) at nodes to make the data purer in the child nodes. For example, the left node after the root splits the data into child nodes based on values of Red band. Thus the splitting rule Red < 77 qualifies training data with values less than 77 in Red band for this branch of the tree. The terminal node (last node) circular in shape refers to classified output with numbers 1 to 5 corresponding to classes Built-up, sand, Cotton, Local rice and Pasmati rice in the study.



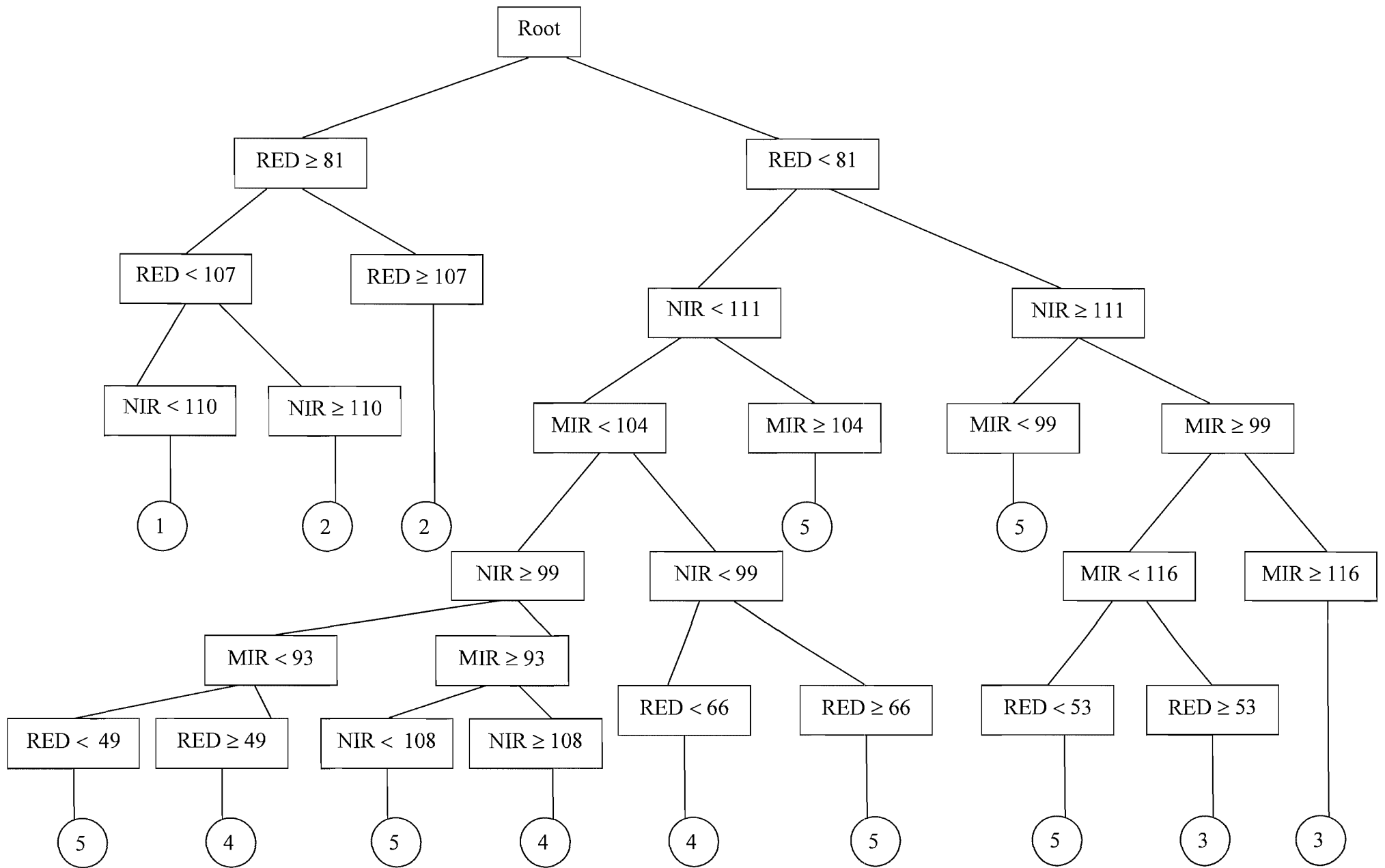


Figure 5.25: Tree structure when DT was trained by training data collected under “Intelligent” scheme.

### 5.5.6 Summary

SVM provided the highest accuracy when trained with training data acquired with the conventional (92.00 %) and intelligent training scheme (90.66 %) with respect to other classifiers DA, DT and ANN (Table 5.18). Furthermore, the accuracy of 90.66 % by SVM trained by training data from intelligent scheme was more accurate than any other classifier (DA, DT and ANN) even when the classifiers were trained with a large training data collected with the conventional scheme (Table 5.18).

However, the SVM classifiers trained by the two schemes used only a fraction of input training data (215 out of 450) for conventional scheme and (76 out of 130) for intelligent scheme as against 450 and 130 cases used by all other classifiers for training with conventional scheme and intelligent scheme respectively.

The intelligent scheme devised for agricultural classes was successful in its intent to capture support vectors directly from field as the SVM classification used 56 out of 80 (70 %) training samples collected under intelligent scheme for agricultural classes as support vectors as compared to 215 out of 450 (47.7 %) for all the classes used by the conventional scheme (section 5.5.2.4).

The intelligent scheme of training data collection was successful in capturing potential border training data for agricultural classes which provided appropriate support vectors for SVM classification. The difference in accuracy between land cover classifications trained by SVM for conventional (92.00%) and intelligent schemes (90.66 %) were not significant at 95 % confidence level (Table 5.16). For agricultural classes only, for which the intelligent scheme was tailored, the conventional scheme resulted in 250 pixels correct (Table 5.14), whereas the intelligent scheme provided 248 pixels correct (Table 5.15) for SVM classification. Thus the accuracy of agricultural classes were very similar for both the training schemes, though training data was reduced by more than two-thirds from 270 cases for conventional scheme to 80 pixels under the intelligent scheme.

The mechanism devised for intelligent scheme to acquire small intelligently selected training samples provided similar accuracy for cotton crop (97.7 %) as a large training set acquired by conventional technique for SVM classification (Tables 5.14 and 5.15). The intelligent scheme used only 30 pixels of cotton for training as compared to 90 pixels used under conventional scheme to train the SVM classifier (Table 5.7).

The small intelligently selected training data acquired under the intelligent scheme which was driven by external knowledge was able to capture cases for classes with very extreme spectral responses in feature space, for example, very matured local rice which was limited on ground to a very small area (Figure 5.23). However, the conventional scheme failed to capture cases of very matured local rice (may be possible with a still larger training set than acquired in the study). This is, therefore, one of the strengths of the intelligent scheme of training data acquisition of being able to capture sub-classes of a class that occupy very small area on ground.

The intelligent scheme which resulted in very similar accuracy for SVM trained with training data acquired by conventional scheme of training data acquisition was also very promising for DA and ANN classifier. The classification accuracy obtained with DA and ANN trained with both training schemes were very similar. The differences in accuracy between the classifications trained with conventional and intelligent schemes were not statistically significant at 95 % confidence level (Table 5.16) for DA classifiers as well as for ANN classifiers.

The similar accuracy obtained for DA classifier trained by the two schemes of training data acquisition can be attributed to the reason that statistical parameters generated by both the training schemes (Table 5.2 and Table 5.6) were very similar.

For ANN, the reason for similar accuracy when the classifier was trained by the two schemes of training data acquisition can be attributed to the findings that border training data are more important than core for classification undertaken by ANN classifier (Foody, 1999). This, therefore, resulted in very similar accuracy for ANN trained with

intelligent scheme (training data collected with potential border cases of classes in feature space) with that collected by conventional scheme.

DT classifier was most sensitive to the nature of training samples. The difference in accuracy between the classifications trained with conventional and intelligent schemes were statistically significant at 95 % confidence level (Table 5.16). This can be attributed to the reason that DT is a non-parametric classifier and the node splitting rules to make child nodes purer is based on the extreme spectral values of the training data ((Figure 5.24, Table 5.2) and (Figure 5.25 and Table 5.6) of the various classes. The small intelligently selected training data under intelligent scheme provided extreme values (for agricultural classes especially local rice, Figure 5.23) and, therefore, the classifier had more extreme values as compared to conventional scheme of training data (Table 5.6 and Table 5.2 respectively). The extreme values under intelligent scheme, therefore, provided more overlap between agricultural classes in feature space and, therefore, more confusion with DT classification.

There was, however, no confusion between agricultural and non-agricultural classes by any of the four classifiers under both the training schemes.

The SVM classifications derived with a small training set captured under intelligent scheme had very similar classification accuracy with respect to one trained with a large training set collected under conventional scheme. Thus not only did the intelligent scheme nearly maintained classification accuracy with respect to conventional scheme of training data acquisition but at the same time was less costly. The intelligent scheme of training data acquisition was cheaper by 26.09 % over the conventional scheme of training data acquisition (Table 5.19). This difference would have magnified to 29.12 % if the training and testing set would have been acquired independently of each other (section 5.5.4).

However, the cost of classification can be further reduced for future analysis by exploiting the knowledge gained once about the relationship of support vectors with ancillary information by SVM classification by focusing the training data acquisition

process to the regions most likely to furnish support vectors. For instance, the support vectors of cotton were derived from near water bodies (waterlogged or canals) (section 5.5.3). This indicates that training data for cotton need be acquired only from near water bodies (near canals or waterlogged areas) if the analysis is repeated in future. In this way, an accurate classification may be undertaken with SVM classifier using a small training set for cotton crop derived from a small spatial area near waterlogged or from near canals.

The study also demonstrated that for mapping cotton with SVM, there is no need to acquire training samples from rice basmati class. Thus in cases where the interest is in accurately classifying only one class, the requirement of training data can be identified from the relative distribution of training data in feature space, excluding training data of classes not facing the class of focus in feature space. Thus classification accuracy of class of interest can be maintained despite the reduction in training size.

The studies (section 5.5.3 and section 5.5.5) demonstrated that for mapping cotton crop in future analysis in the study area, training acquisition for cotton crop should be limited only to locations near water bodies like canals or waterlogged areas and that no training data is required from rice basmati.

The analysis confirms the results of chapters 3 and 4 that a representative training sample of each class in feature space is not required if using SVM as a classifier. The potential of SVM of using only border training samples of a class in feature space can be exploited to limit the size of training samples especially of agricultural classes using external knowledge. Thus if knowledge regarding the status of the crop (matured, young), soil and water background of the agricultural fields is known, the training data process can be directed to intelligently capture a small relevant training set from sites that would provide appropriate support vectors central to the establishment of decision surface in SVM classification. The knowledge gained about the relationship of support vectors with ancillary information in SVM classification can be exploited in future analysis by focusing the training data acquisition process from regions most likely to furnish support vectors.

### 5.5.7 Conclusions

- SVM provided the highest accuracy as compared to other conventional classifiers DA, DT and ANN for both schemes of training data acquisition, the Conventional as well as the Intelligent scheme.
- The intelligent scheme tailored for agricultural classes was successful as 70 % of the training data collected was used as support vectors in SVM classification.
- The overall accuracy obtained by SVM trained with intelligent scheme of training data acquisition was very comparable with one trained with conventional scheme. The difference in accuracy obtained by the two classifications was statistically not significant.
- The overall accuracy obtained by SVM trained with the Intelligent scheme of training data acquisition was higher as compared to that achieved by training the other classifiers DA, DT and ANN for both schemes of training data acquisition, the Conventional as well as the Intelligent scheme.
- The intelligent scheme of training data acquisition was cheaper by 26.09 % over the conventional scheme of training data acquisition.
- The support vectors of cotton class for classification trained with training data acquired under the Intelligent scheme of training data acquisition were mainly derived from near water bodies. Thus for analysis in future, training data for cotton crop may be collected mainly from near water bodies.
- The study also demonstrated that for accurately mapping cotton crop, training data of classes not facing the class of interest in feature space is not required. Thus accuracy of mapping cotton remained unaffected when training data included or excluded training data of rice basmati class.

- Essentially the analysis shows that a representative sample of each class in feature space is not required if using SVM as a classifier. The potential of SVM of using only border training data samples of a class in feature space can be exploited to reduce the requirement of training data over the conventional techniques of classification.

# CHAPTER 6 - Summary and Conclusions

## 6.1 Summary

The desire in training a supervised classifier has traditionally been to derive an accurate and complete description of the spectral response of all the classes in the study area. To achieve a complete description of each class in feature space, a large training set is typically required. Much of the literature on training data is based on the classical view of classification process with a conventional probabilistic classifier like MLC as the basis of classification.

The MLC is a parametric classifier and requires a large number of training samples acquired from across the entire study area to capture the spectral variability of the classes.

The analysis in chapter 3, designed with a large representative training set (100 cases per class) acquired with a mindset to provide a complete description of each class in feature space reinforces the effect of training set size on classification accuracy. The accuracy of the classifications produced from all the four classifiers (DA, DT, ANN and SVM) were positively related with training set size.

The increase in accuracy with training set size for the different classifiers can be attributed to different reasons based primarily on the way the classifiers allocate the cases to the various classes. The non-parametric classifiers (DT, ANN and SVM) are not based on any parametric model as DA. The classifiers differ markedly in their approach for training. For example, for training, DT is based on the extreme values of the spectral response of a class in feature space. Thus if the classes overlap considerably in feature space, there will be more confusion between the classes if using DT as a classifier. However, Foody (1999) has shown that ANN is more biased towards extreme cases of a class in feature space. The SVM classifier depends on training samples which lie on part of the edge of the class distribution in feature space, the support vectors. Data other than



support vectors are redundant and do not contribute in the establishment of the decision surface and can be discarded without compromising the accuracy of the classification. Thus the four classifiers differ in the basis of class allocations and, therefore, expected dependency on the nature of training set. For example, training set with extreme spectral response of the classes would result in higher accuracy for ANN and SVM classifier as compared to DT classifier.

The increased training set size (chapter 3) fulfilled the requirements of the four classifiers resulting in higher accuracy as the training set size increased. SVM was comparatively the most accurate classifier of all the four classifiers and provided the highest accuracy when trained with the largest training set (100 pixels/class). The increased accuracy of SVM can be attributed to the reason that a large training set has more chances of including support vectors. It is, therefore, not always necessary to have training statistics that provides a complete description of the class's especially using non-parametric classifiers. The design of training stage should, therefore, be guided by the classifiers used. For example with a SVM classifier the concern is to identify and characterize the remotely sensed data that lie near to the location of the classification hyperplane or classes in the feature space.

The potential of SVM was evident from analysis detailed in Chapters 3 and 5. In general, the SVM classifications were more accurate than comparable classifications derived with the use of other classifiers DA, DT and ANN. In addition, SVM used only a fraction of the input training data (support vectors) as compared to other classifiers and should, therefore, be increasingly used in classifying remotely sensed data. Thus for SVM, the training samples are not equally important with those lying near the edge of the class distributions in feature space and facing the distributions of other classes in feature space (support vectors) more important in the fitting of decision boundaries between the classes. The support vectors typically occupy a small discrete area of the distribution of a class in feature space and, therefore, may have something in common. The support vectors of a

class may be derived from training sites with particular ground property of training sites. For example, analysis detailed in chapter 4 demonstrated that support vectors may be related with ground property of training sites like soil. This relationship of support vectors with attributes like soil can be exploited to limit acquisition of training data from sites that provides appropriate support vectors for future analysis.

Thus if there is a prior knowledge or some ancillary information that can be used to identify/locate training sites to regions from which the most informative training samples, the support vectors can be derived, it may be possible to acquire a small intelligently selected training set that can be used to accurately classify the data. This would in turn reduce the cost of the classification process as every training data collected has a cost attached to it.

The analysis detailed in chapter 5 demonstrates that external knowledge can be employed in the training acquisition process for any current land cover classification from sites that provides the most informative training samples, the support vectors. The training acquisition scheme, in such instances, needs to be devised in advance of training acquisition process and should include the variables affecting the spectral response of the agricultural classes. The procedure detailed in chapter 5 with intelligent scheme of training data acquisition demonstrates that considering all the growth stages of the crop and background properties (water and soil) of the training sites can provide appropriate support vectors central to the establishment of SVM classifier.

The intelligent scheme devised for agricultural classes was successful in its intent to capture support vectors directly from field as the SVM classification used 56 out of 80 (70 %) training samples collected under intelligent scheme for agricultural classes as support vectors as compared to 215 out of 450 (47.7 %) for all the classes used by the conventional scheme (section 5.5.2.4). The SVM classifications derived with the intelligent scheme of training data acquisition had very similar classification accuracy (248 pixels correct) for agricultural classes with respect to one trained with a large training set

collected under conventional scheme (250 pixels correct). Thus not only did the intelligent scheme resulted in similar classification accuracy with respect to conventional scheme of training data acquisition but at the same time was less costly. The intelligent scheme of training data acquisition was cheaper by 26.09 % over the conventional scheme of training data acquisition (Table 5.19).

However, the cost of classification can be further reduced for future analysis by exploiting the knowledge gained from the SVM classification about the relationship of support vectors with ancillary information by focusing the training data acquisition process to the regions most likely to furnish support vectors. The analysis detailed in chapter 4 demonstrated that support vectors can be related with ancillary information like soil type. The analysis detailed in chapter 5 shows that the support vectors of cotton were derived from near water bodies (waterlogged or canals) (section 5.5.3). This indicates that training data for cotton needs be acquired only from near water bodies (near canals or waterlogged areas) if the analysis is to be repeated in future. In this way, an accurate classification may be undertaken with SVM classifier using a small training set for cotton crop derived from a small spatial area near waterlogged or from near canals. Thus not only can the limited training data acquisition process be tailored for any current land cover classification but can be extended for future analysis based on the relationship of support vectors derived with ground attributes like soil type and water condition of the training sites.

The study (section 5.5.5) also demonstrated that training acquisition scheme should be designed on the nature of the output desired. For example, for land cover mapping, training data is generally acquired from all the land cover classes available in the study area but if the concern is to accurately map only one of the many land cover classes in the area, training data of classes not facing the class of interest in feature space can be excluded without affecting the accuracy of the class of concern with SVM classification. This was demonstrated with cotton crop as the class of concern whereby excluding the

training data of class rice basmati did not affect classification accuracy of the cotton crop (section 5.5.5).

The hypothesis that classes which do not share their boundaries with other classes in feature space are not required for classification. This is more apparent for SVM classifications which are based on a small training set that occupies only part of border of feature space of class. This potential of SVM can be exploited to reduce the requirement of training data if the concern is to map accurately only one class.

The success of classification undertaken using a small intelligently acquired training data collected under intelligent scheme for classification task and the potential to limit training data acquisition for future analysis based on ground property of training sites like soil and water questions the very understanding prevailing in remote sensing community over many years namely;

1. Training data collection is more of an art than science.

Traditionally training data collection is based on the premise to collect a large training set spread all over the study area so as to capture the spectral variability of the classes. The external knowledge used in collecting training data under intelligent scheme from sites that provided appropriate support vectors was founded on scientific rationale. The small intelligently collected training data was based on scientific knowledge about the relation of crop status and background properties of training sites (soil and water) with spectral response. Thus scientific knowledge played a key role in training data collection.

2. Training data should be representative of the classes

For classification by SVM, only the training samples that are support vectors, which lie on part of the edge of the class distribution in feature space are required, all other training samples are redundant. Thus the accuracy of the classification was maintained despite reduction in training set size (chapter 4 and chapter 5). Thus training data need not be representative of the classes used if using SVM as a classifier.

## 6.2 Conclusions

- SVMs have considerable potential for the classification of remotely sensed data. It has been demonstrated here that a single multi-class SVM classification may be undertaken and used to derive very accurate classifications.
- In general, the SVM classifications were more accurate than comparable classifications derived with the use of the other classification techniques.
- SVM classification is effectively based on a small number of training samples, the support vectors, which are training data that lies on part of the edge of a class distribution in feature space.
- The analysis shows that a representative sample of each class in feature space is not required if using SVM as a classifier. The potential of SVM of using only border training data samples of a class in feature space, the support vectors can be exploited to reduce the requirement of training data over the conventional techniques of classification.
- The study shows that training data requirements of agricultural classes can be reduced by intelligently planning training data acquisition scheme that can provide appropriate potential border training data (potential support vectors). This was shown with the Intelligent scheme of training data acquisition (section 5.3.2), which was devised for agricultural classes as the variables affecting the spectral response are well understood for agricultural classes. Thus external knowledge of crop status, soil background and water condition of the agricultural fields were used to select suitable training sites to obtain potential border training data.
- The classification accuracy obtained by training SVM with a small intelligently selected training data acquired under the Intelligent scheme of training data acquisition was very comparable with one obtained by training SVM with a large training set collected under the conventional scheme.

- The intelligent scheme of training data acquisition was cheaper over the conventional scheme of training data acquisition essentially due to the reduced requirement of training data.
- The study shows that support vectors resulting from SVM classification may be related with ground property like soil type or water background of training sites. This knowledge may be utilized for future analysis by collecting training data from only those sites that would provide appropriate support vectors.
- The study also demonstrated that if the concern is to accurately map only one class from the many land cover classes available, training data of classes not facing the class of interest in feature space is not required. This was shown for mapping cotton crop (section 5.5.5). The accuracy of cotton class remained unaffected when training data included or excluded training data of rice basmati class
- The key conclusion of the analyses is that a complete description of each class in feature space is not required for an accurate classification. With a SVM, only training samples located near the hyperplane are required with other samples not contributing the analysis can effectively be discarded. The acquisition of training samples from beyond the border region of a class in feature space is, therefore, unnecessary and a waste of effort and resources. Thus, in situations when knowledge of a variable such as growth of crops or background soil or water status of training sites that may impact on the spectral response of a class is available, this could be used to direct training activities. In particular, this knowledge may allow small, intelligently selected, training samples to be acquired that may be used to discriminate between the classes as accurately as a much larger, unintelligently selected, sample. Knowledge may, therefore, be used to reduce training set size without loss of classification accuracy by directing the training site acquisition process to regions most likely to provide appropriate support vectors. The

requirement of reduced training set size based on scientific knowledge not only has financial implications on the classification process but makes training acquisition process very interesting in the field too. The intelligent scheme was interesting in the sense that there was more interaction as compared to the conventional scheme of training data acquisition with the farmers at the field level to locate sites that provided appropriate support vectors.

### **6.3 Future Work**

The future work relates to the implementation of the findings of the research work in the CAPE project for accurately mapping the cotton crop. This would help to reduce the requirement of training data. For accurately mapping cotton, training data for only cotton and rice local classes would be needed and there would be no need to acquire training data for rice basmati class. In addition, training data for cotton need be acquired from near water bodies like canals or near waterlogged areas. The reduced training data requirements in turn would reduce the cost on the classification process.

The analysis detailed in the research was limited to the use of only a few spectral bands. The hyperspectral remote sensing data made available now a days makes it imperative to study the feature reduction based on the training data requirements of the classifier to be used. The general approach in feature reduction uses all the training data to select the most separable bands. However, feature reduction should be based on the classifier to be used. For SVM, only training data which lie on border of the spectral distribution of a class in feature space which are potential support vectors needs be considered to select the most separable bands.

## APPENDIX

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	83	6	0	0	8	0	97
Wheat (W)	3	91	2	0	0	0	96
Barley (B)	0	8	43	0	0	0	51
Carrot (C)	0	2	0	26	5	0	33
Potato (P)	0	2	0	0	24	0	26
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	86	109	45	27	39	14	320

Overall accuracy = 87.80%

Table A.1: Error matrix for the classification derived from the DA trained with training set 5n (containing 15 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	83	7	0	0	7	0	97
Wheat (W)	3	90	2	1	0	0	96
Barley (B)	0	7	44	0	0	0	51
Carrot (C)	0	1	0	29	3	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	86	107	46	31	35	15	320

Overall accuracy =88.40%

Table A.2: Error matrix for the classification derived from the DA trained with training set 10n (containing 30 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	86	4	0	0	7	0	97
Wheat (W)	5	90	0	1	0	0	96
Barley (B)	1	7	43	0	0	0	51
Carrot (C)	0	1	0	31	1	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	1	1	15	17
<b>Total</b>	92	104	43	33	32	16	320

Overall accuracy =90.00%

Table A.3: Error matrix for the classification derived from the DA trained with training set 15n (containing 45 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	87	3	0	0	7	0	97
Wheat (W)	3	91	2	0	0	0	96
Barley (B)	0	6	45	0	0	0	51
Carrot (C)	0	2	0	28	3	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	90	104	47	29	35	15	320

Overall accuracy=90.00%

Table A.4: Error matrix for the classification derived from the DA trained with training set 20n (containing 60 cases of each class) for case A analysis.



Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	86	4	0	0	7	0	97
Wheat (W)	3	90	2	1	0	0	96
Barley (B)	0	6	45	0	0	0	51
Carrot (C)	0	1	0	29	3	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	89	103	47	31	35	15	320

Overall accuracy =89.7%

Table A.5: Error matrix for the classification derived from the DA trained with training set 25n (containing 75 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	86	4	0	0	7	0	97
Wheat (W)	3	91	2	0	0	0	96
Barley (B)	0	6	45	0	0	0	51
Carrot (C)	0	1	0	29	3	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	89	104	47	30	35	15	320

Overall accuracy =90.00%

Table A.6: Error matrix for the classification derived from the DA trained with training set 30n (containing 90 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	87	3	0	0	7	0	97
Wheat (W)	3	90	2	1	0	0	96
Barley (B)	0	6	45	0	0	0	51
Carrot (C)	0	1	0	29	3	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	90	102	45	31	35	15	320

Overall accuracy =90.00%

Table A.7: Error matrix for the classification derived from the DA trained with the largest training set (containing 100 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	11	1	0	0	5	0	17
Wheat (W)	1	16	0	0	0	0	17
Barley (B)	0	3	14	0	0	0	17
Carrot (C)	0	0	0	16	1	0	17
Potato (P)	0	2	0	0	15	0	17
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	12	19	14	17	23	14	102

Overall accuracy =84.30%

Table A.8: Error matrix for the classification derived from the DA trained with training set 5n (containing 15 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	16	0	0	0	1	0	17
Wheat (W)	1	15	1	0	0	0	17
Barley (B)	0	3	14	0	0	0	17
Carrot (C)	0	2	0	14	1	0	17
Potato (P)	1	1	0	0	15	0	17
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	18	21	15	15	19	14	102

Overall accuracy =86.30%

Table A.9: Error matrix for the classification derived from the DA trained with training set 10n (containing 30 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	13	2	0	0	2	0	17
Wheat (W)	0	17	0	0	0	0	17
Barley (B)	0	2	15	0	0	0	17
Carrot (C)	0	1	0	16	0	0	17
Potato (P)	0	1	0	0	15	1	17
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	13	23	15	17	19	15	102

Overall accuracy =88.20%

Table A.10: Error matrix for the classification derived from the DA trained with training set 15n (containing 45 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	17	0	0	0	0	0	17
Wheat (W)	1	16	0	0	0	0	17
Barley (B)	0	4	13	0	0	0	17
Carrot (C)	0	0	0	16	1	0	17
Potato (P)	0	2	0	0	15	0	17
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	18	22	13	17	18	14	102

Overall accuracy =89.20%

Table A.11: Error matrix for the classification derived from the DA trained with training set 20n (containing 60 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	14	0	0	0	3	0	17
Wheat (W)	0	17	0	0	0	0	17
Barley (B)	0	2	15	0	0	0	17
Carrot (C)	0	1	0	14	2	0	17
Potato (P)	0	2	0	0	14	1	17
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	14	22	15	15	21	15	102

Overall accuracy =86.30%

Table A.12: Error matrix for the classification derived from the DA trained with training set 25n (containing 75 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	16	0	0	0	1	0	17
Wheat (W)	0	17	0	0	0	0	17
Barley (B)	0	4	13	0	0	0	17
Carrot (C)	0	1	0	14	2	0	17
Potato (P)	0	1	0	0	15	1	17
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	16	23	13	15	19	15	102

Overall accuracy =87.30%

Table A.13: Error matrix for the classification derived from the DA trained with training set 30n (containing 90 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	17	0	0	0	0	0	17
Wheat (W)	1	16	0	0	0	0	17
Barley (B)	0	4	13	0	0	0	17
Carrot (C)	0	0	0	16	1	0	17
Potato (P)	0	2	0	0	14	1	17
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	18	22	13	17	17	15	102

Overall accuracy =88.20%

Table A.14: Error matrix for the classification derived from the DA trained with the largest training set (containing 100 cases of each class) for case B

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	85	7	0	0	5	0	97
Wheat (W)	4	69	21	0	0	2	96
Barley (B)	1	5	44	0	0	1	51
Carrot (C)	0	4	1	16	7	5	33
Potato (P)	5	2	0	0	19	0	26
Grass (G)	0	0	0	0	3	14	17
<b>Total</b>	95	87	66	16	34	22	320

Overall accuracy =77.18%

Table A.15: Error matrix for the classification derived from the DT trained with training set 5n (containing 15 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	79	4	4	0	9	1	97
Wheat (W)	12	73	10	1	0	0	96
Barley (B)	5	2	41	2	1	0	51
Carrot (C)	1	2	1	29	0	0	33
Potato (P)	1	1	0	0	23	1	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	98	82	56	32	33	19	320

Overall accuracy =81.87%

Table A.16: Error matrix for the classification derived from the DT trained with training set 10n (containing 30 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	80	6	1	8	2	0	97
Wheat (W)	3	83	7	3	0	0	96
Barley (B)	5	1	39	2	4	0	51
Carrot (C)	0	3	1	29	0	0	33
Potato (P)	1	2	0	1	22	0	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	89	95	48	43	28	17	320

Overall accuracy =84.37%

Table A.17: Error matrix for the classification derived from the DT trained with training set 15n (containing 45 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	88	3	1	4	1	0	97
Wheat (W)	12	75	7	1	1	0	96
Barley (B)	4	2	45	0	0	0	51
Carrot (C)	0	2	0	31	0	0	33
Potato (P)	1	1		1	23	0	26
Grass (G)	0	0	4	0	0	13	17
<b>Total</b>	105	83	57	37	25	13	320

Overall accuracy =85.94%

Table A.18: Error matrix for the classification derived from the DT trained with training set 20n (containing 60 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	91	1		2	2	1	97
Wheat (W)	6	69	10	3	7	1	96
Barley (B)	3	0	48	0	0	0	51
Carrot (C)	0	0	0	33	0	0	33
Potato (P)	1	1		2	22	0	26
Grass (G)	0	0	0	0	1	16	17
<b>Total</b>	101	71	58	40	32	18	320

Overall accuracy =87.19%

Table A.19: Error matrix for the classification derived from the DT trained with training set 25n (containing 75 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	83	6	2	1	4	1	97
Wheat (W)	6	79	7	2	0	2	96
Barley (B)	0	1	48	0	1	1	51
Carrot (C)	0	2	0	31	0	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	1	0	0	0	0	16	17
<b>Total</b>	90	90	57	34	28	21	320

Overall accuracy =87.50%

Table A.20: Error matrix for the classification derived from the DT trained with training set 30 (containing 90 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	89	4	1	0	2	1	97
Wheat (W)	8	79	6	1	0	2	96
Barley (B)	3	0	48	0	0	0	51
Carrot (C)	0	0	0	33	0	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	100	85	55	34	25	20	320

Overall accuracy =90.31%

Table A.21: Error matrix for the classification derived from the DT trained with the largest training set (containing 100 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	16	1	0	0	0	0	17
Wheat (W)	0	14	3	0	0	0	17
Barley (B)	2	0	12	0	3	0	17
Carrot (C)	0	1	0	12	1	3	17
Potato (P)	1	2	0	0	14	0	17
Grass (G)	0	0	3	1	1	12	17
<b>Total</b>	19	18	18	13	19	15	102

Overall accuracy =78.43%

Table A.22: Error matrix for the classification derived from the DT trained with training set 5n (containing 15 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	16	0	1	0	0	0	17
Wheat (W)	3	11	2	1	0	0	17
Barley (B)	1	0	16	0	0	0	17
Carrot (C)	2	1	0	14	0	0	17
Potato (P)	1	1	0	1	14	0	17
Grass (G)	0	0	1	0	0	16	17
<b>Total</b>	23	13	20	16	14	16	102

Overall accuracy =85.29%

Table A.23: Error matrix for the classification derived from the DT trained with training set 10n (containing 30 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	14	0	0	1	2	0	17
Wheat (W)	1	14	0	1	0	1	17
Barley (B)	1	2	14	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	1	0	1	15	0	17
Grass (G)	0	0	3	0	0	14	17
<b>Total</b>	16	17	17	20	17	15	102

Overall accuracy =86.27%

Table A.24: Error matrix for the classification derived from the DT trained with training set 15n (containing 45 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	15	1	0	1	0	0	17
Wheat (W)	1	15	0	0	1	0	17
Barley (B)	1	0	16	0	0	0	17
Carrot (C)	1	0	0	16	0	0	17
Potato (P)	0	2	0	0	14	1	17
Grass (G)	0	0	2	1	0	14	17
<b>Total</b>	18	18	18	18	15	15	102

Overall accuracy =88.24%

Table A.25: Error matrix for the classification derived from the DT trained with training set 20n (containing 60 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	14	2	0	0	1	0	17
Wheat (W)	1	12	2	1	0	1	17
Barley (B)	1	0	16	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	0	0	0	17	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	16	14	18	18	18	18	102

Overall accuracy =91.18%

Table A.26: Error matrix for the classification derived from the DT trained with training set 25n (containing 75 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	14	3	0	0	0	0	17
Wheat (W)	0	14	0	0	3	0	17
Barley (B)	0	0	15	0	2	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	2	0	0	15	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	14	19	15	17	20	17	102

Overall accuracy =90.20%

Table A.27: Error matrix for the classification derived from the DT trained with training set 30n (containing 90 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	17	0	0	0	0	0	17
Wheat (W)	0	15	2	0	0	0	17
Barley (B)	2	0	15	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	1	0	0	15	1	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	19	16	17	17	15	18	102

Overall accuracy =94.12%

Table A.28: Error matrix for the classification derived from the DT trained with the largest training set (containing 100 cases of each class) for case B analysis

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	84	4	2	0	7	0	97
Wheat (W)	3	89	3	1	0	0	96
Barley (B)	0	5	46	0	0	0	51
Carrot (C)	0	1	0	31	1	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	1	2	14	17
<b>Total</b>	87	101	51	33	33	15	320

Overall accuracy =89.68%

Table A.29: Error matrix for the classification derived from the ANN trained with training set 5n (containing 15 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	84	5	0	0	8	0	97
Wheat (W)	4	90	1	1	0	0	96
Barley (B)	1	7	43	0	0	0	51
Carrot (C)	0	1	0	31	1	0	33
Potato (P)	0	2	0	0	24	0	26
Grass (G)	0	0	0	0	2	15	17
<b>Total</b>	89	105	44	32	35	15	320

Overall accuracy =89.68%

Table A.30: Error matrix for the classification derived from the ANN trained with training set 10n (containing 30 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	85	2	1	0	8	1	97
Wheat (W)	4	87	4	1	0	0	96
Barley (B)	1	1	49	0	0	0	51
Carrot (C)	0	1	0	32	0	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	90	93	54	33	31	19	320

Overall accuracy =91.56%

Table A.31: Error matrix for the classification derived from the ANN trained with training set 15n (containing 45 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	87	3	0	0	7	0	97
Wheat (W)	4	88	3	1	0	0	91
Barley (B)	1	3	47	0	0	0	51
Carrot (C)	0	1	0	32	0	0	33
Potato (P)	0	2	0	1	22	1	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	92	97	50	34	29	18	320

Overall accuracy =91.56%

Table A.32: Error matrix for the classification derived from the ANN trained with training set 20 (containing 20 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	88	5	0	0	3	1	97
Wheat (W)	3	89	3	1	0	0	96
Barley (B)	2	1	48	0	0	0	51
Carrot (C)	0	1	0	32	0	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	93	93	51	33	23	18	320

Overall accuracy =92.81%

Table A.33: Error matrix for the classification derived from the ANN trained with training set 25n (containing 75 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	86	4	0	0	6	1	97
Wheat (W)	3	87	4	1	0	1	96
Barley (B)	0	1	50	0	0	0	51
Carrot (C)	0	1	0	32	0	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	89	95	54	33	29	20	320

Overall accuracy =92.18%

Table A.34: Error matrix for the classification derived from the ANN trained with training set 30n (containing 90 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	90	3	1	0	3	0	97
Wheat (W)	3	84	7	1	0	1	96
Barley (B)	0	2	49	0	0	0	51
Carrot (C)	0	2	0	31	0	0	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	93	93	57	32	26	19	320

Overall accuracy =91.88%

Table A.35: Error matrix for the classification derived from the ANN trained with the largest training set (containing 100 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	11	1	0	0	4	1	17
Wheat (W)	1	14	2	0	0	0	17
Barley (B)	0	0	16	0	0	1	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	2	0	0	15	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	12	17	18	17	19	19	102

Overall accuracy =88.23%

Table A.36: Error matrix for the classification derived from the ANN trained with training set 5n (containing 15 cases of each class) for case B analysis.



Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	14	0	0	1	2	0	17
Wheat (W)	0	16	0	1	0	0	17
Barley (B)	0	0	17	0	0	0	17
Carrot (C)	0	1	0	16	0	0	17
Potato (P)	0	2	0	0	14	1	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	14	19	17	18	16	18	102

Overall accuracy =92.15%

Table A.37: Error matrix for the classification derived from the ANN trained with training set 10n (containing 30 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	16	0	0	0	1	0	17
Wheat (W)	0	16	0	1	0	0	17
Barley (B)	1	0	16	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	1	0	1	15	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	17	17	16	19	16	17	102

Overall accuracy =95.09%

Table A.38: Error matrix for the classification derived from the ANN trained with training set 15n (containing 45 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	15	0	0	0	2	0	17
Wheat (W)	1	15	1	0	0	0	17
Barley (B)	0	0	17	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	1	0	0	15	1	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	16	16	18	17	17	18	102

Overall accuracy =94.11%

Table A.39: Error matrix for the classification derived from the ANN trained with training set 20n (containing 60 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	16	0	0	0	1	0	17
Wheat (W)	1	14	2	0	0	0	17
Barley (B)	0	0	17	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	1	0	0	16	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	17	15	19	17	17	17	102

Overall accuracy =95.09%

Table A.40: Error matrix for the classification derived from the ANN trained with training set 25n (containing 75 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	15	2	0	0	0	0	17
Wheat (W)	0	17	0	0	0	0	17
Barley (B)	0	0	17	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	2	0	0	15	0	17
Grass (G)	0	0	0	0	1	16	17
<b>Total</b>	15	21	17	17	16	16	102

Overall accuracy =95.09%

Table A.41: Error matrix for the classification derived from the ANN trained with training set 30n (containing 90 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	15	1	0	0	1	0	17
Wheat (W)	2	15	0	0	0	0	17
Barley (B)	0	0	17	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	2	0	0	15	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	17	18	17	17	16	17	102

Overall accuracy =94.11%

Table A.42: Error matrix for the classification derived from the ANN trained with the largest training set (containing 100 cases of each class) for case B analysis

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	90	4	2	0	1	0	97
Wheat (W)	4	91	1	0	0	0	96
Barley (B)	3	7	39	2	0	0	51
Carrot (C)	0	3	0	22	3	5	33
Potato (P)	1	2	0	0	23	0	26
Grass (G)	0	0	0	0	2	15	17
<b>Total</b>	98	107	42	44	28	20	320

Overall accuracy =87.50%

Table A.43: Error matrix for the classification derived from the SVM trained with training set 5n (containing 15 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	86	8	1	0	1	1	97
Wheat (W)	2	89	4	1	0	0	96
Barley (B)	1	3	47	0	0	0	51
Carrot (C)	1	1	0	30	0	1	33
Potato (P)	1	1	0	0	23	1	26
Grass (G)	0	0	0	0	1	16	17
<b>Total</b>	91	102	52	31	25	19	320

Overall accuracy =90.94%

Table A.44: Error matrix for the classification derived from the SVM trained with training set 10n (containing 30 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	89	4	1		1	1	97
Wheat (W)	2	82	11	1	0	0	96
Barley (B)	1	1	49	0	0	0	51
Carrot (C)	0	3	0	30	0	0	33
Potato (P)	0	2	0	0	24	0	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	93	92	61	31	25	18	320

Overall accuracy =90.93%

Table A.45: Error matrix for the classification derived from the SVM trained with training set 15n (containing 45 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	85	3	2	0	6	1	97
Wheat (W)	3	90	2	1	0	0	96
Barley (B)	1	3	47	0	0	0	51
Carrot (C)	0	0	0	32	0	1	33
Potato (P)	0	2	0	0	23	1	26
Grass (G)	0	0	0	1	0	16	17
<b>Total</b>	89	98	51	34	29	19	320

Overall accuracy =91.56%

Table A.46: Error matrix for the classification derived from the SVM trained with training set 20n (containing 60 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	90	4	0	0	2	1	97
Wheat (W)	6	83	6	1	0	0	96
Barley (B)	1	0	50	0	0	0	51
Carrot (C)	0	0	0	33	0	0	33
Potato (P)	0	1	0	0	25	0	26
Grass (G)	0	0	0	1	0	16	17
<b>Total</b>	97	88	56	35	27	17	320

Overall accuracy =92.81%

Table A.47: Error matrix for the classification derived from the SVM trained with training set 25n (containing 75 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	90	5	0	0	1	1	97
Wheat (W)	2	86	7	1	0	0	96
Barley (B)	1	0	50	0	0	0	51
Carrot (C)	0	2	0	31	0	0	33
Potato (P)	1	1	0	1	23	0	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	94	94	57	33	24	18	320

Overall accuracy =92.81%

Table A.48: Error matrix for the classification derived from the SVM trained with training set 30n (containing 90 cases of each class) for case A analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	89	6	0	0	1	1	97
Wheat (W)	2	88	5	1	0	0	96
Barley (B)	1	1	49	0	0	0	51
Carrot (C)	0	0	0	33	0	0	33
Potato (P)	0	2	0	0	24	0	26
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	92	97	54	34	25	18	320

Overall accuracy =93.75%

Table A.49: Error matrix for the classification derived from the SVM trained with the largest training set (containing 100 cases of each class) for case A analysis

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	12	1	0	0	4	0	17
Wheat (W)	4	13	0	0	0	0	17
Barley (B)	0	1	16	0	0	0	17
Carrot (C)	0	1	0	16	0	0	17
Potato (P)	1	1	0	0	15	0	17
Grass (G)	0	0	3	0	0	14	17
<b>Total</b>	17	17	19	16	19	14	102

Overall accuracy =84.31%

Table A.50: Error matrix for the classification derived from the SVM trained with training set 5n (containing 15 cases of each class) for case B analysis.

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	17	0	0	0	0	0	17
Wheat (W)	1	16	0	0	0	0	17
Barley (B)	0	0	17	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	1	0	0	16	0	17
Grass (G)	0	0	0	0	2	15	17
<b>Total</b>	18	17	17	17	18	15	102

Overall accuracy =96.07%

Table A.51: Error matrix for the classification derived from the SVM trained with training set 10n (containing 30 cases of each class) for case B analysis

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	14	2	0	1	0	0	17
Wheat (W)	0	14	2	1	0	0	17
Barley (B)	0	1	16	0	0	0	17
Carrot (C)	0	2	0	15	0	0	17
Potato (P)	0	0	0	0	17	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	14	19	18	17	17	17	102

Overall accuracy =91.17%

Table A.52: Error matrix for the classification derived from the SVM trained with training set 15n (containing 45 cases of each class) for case B analysis

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	13	3	0	0	0	1	17
Wheat (W)	0	17	0	0	0	0	17
Barley (B)	1	0	16	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	1	0	0	16	0	17
Grass (G)	0	0	0	1	0	16	17
<b>Total</b>	14	21	16	18	16	17	102

Overall accuracy =93.13%

Table A.53: Error matrix for the classification derived from the SVM trained with training set 20n (containing 60 cases of each class) for case B analysis

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	15	1	0	0	1	0	17
Wheat (W)	0	15	2	0	0	0	17
Barley (B)	0	0	17	0	0	0	17
Carrot (C)	0	2	0	16	0	1	17
Potato (P)	0	1	0	1	15	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	15	19	19	18	16	18	102

Overall accuracy =93.13%

Table A.54: Error matrix for the classification derived from the SVM trained with training set 25n (containing 75 cases of each class) for case B analysis

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	16	1	0	0	0	0	17
Wheat (W)	1	16	0	0	0	0	17
Barley (B)	0	1	16	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	1	0	2	14	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	17	19	16	19	14	17	102

Overall accuracy =94.11%

Table A.55: Error matrix for the classification derived from the SVM trained with training set 30n (containing 90 cases of each class) for case B analysis

Actual class	Predicted class						Total pixels
	(S)	(W)	(B)	(C)	(P)	(G)	
Sugar beet (S)	16	1	0	0	0	0	17
Wheat (W)	0	15	2	0	0	0	17
Barley (B)	1	1	15	0	0	0	17
Carrot (C)	0	0	0	17	0	0	17
Potato (P)	0	0	0	1	16	0	17
Grass (G)	0	0	0	0	0	17	17
<b>Total</b>	17	17	17	18	16	17	102

Overall accuracy =94.11%

Table A.56: Error matrix for the classification derived from the SVM trained with the largest training set (containing 100 cases of each class) for case B analysis

$\alpha(1)$	$\alpha(2)$	$\alpha(3)$	$\alpha(4)$	RED	NIR	MIR	Classes	Location	Soil code	Growth
0.0000	0.0000	0.0000	1.0000	96	81	129	1			
1.0000	0.0000	0.0000	0.0000	104	99	176	1			
1.0000	0.0000	0.0000	0.0000	101	106	163	1			
1.0000	0.0000	0.0000	0.0000	99	105	163	1			
1.0000	0.0000	0.0054	0.2047	95	106	189	1			
0.9903	0.0000	0.0000	0.0000	97	103	168	1			
1.0000	0.0000	0.0000	0.0000	96	109	167	1			
1.0000	0.0000	0.0000	0.0000	104	104	164	1			
1.0000	0.0000	0.0000	0.0000	102	104	165	1			
0.7776	0.0000	0.0000	0.0000	104	100	158	1			
0.0000	0.0000	0.0000	0.7893	94	84	131	1			
1.0000	0.0000	0.0000	0.0000	111	109	181	2			
1.0000	0.7849	0.5720	0.4596	107	99	169	2			
0.7110	0.0000	0.0000	0.0000	109	124	167	2			
1.0000	0.0000	0.0000	0.0000	103	122	172	2			
0.1833	0.0000	0.0000	0.0000	111	113	182	2			
1.0000	0.0000	0.0000	0.0000	110	111	159	2			
1.0000	0.3478	0.0000	0.2130	101	114	167	2			
1.0000	0.0000	0.0000	0.0000	113	110	179	2			
1.0000	0.0000	0.0000	0.0000	112	100	174	2			
0.0000	0.0000	0.0000	1.0000	54	132	113	3	Near Canal	47	
0.0000	0.0000	0.0000	0.2035	54	138	116	3	Near Canal	43	
0.0000	0.0000	0.0000	1.0000	53	130	114	3	Near Canal	46	
0.0000	0.0000	0.0000	1.0000	53	137	115	3	Waterlogged	44	
0.0000	0.0000	0.0000	1.0000	56	129	115	3	Adjoining Canal	18	
0.0348	0.0201	0.0000	0.6051	47	170	119	3	very dry	44	
0.0000	0.0000	0.0000	1.0000	53	122	99	3	Waterlogged	45	
0.0000	0.0393	1.0000	1.0000	54	116	97	3	Waterlogged	45	
0.0000	0.0000	1.0000	1.0000	58	112	107	3	Waterlogged	45	
0.0000	0.0000	1.0000	1.0000	56	118	99	3	waterlogged and near canal	43	
0.0000	0.0000	0.0000	1.0000	55	122	100	3	waterlogged along canal	43	
0.0000	0.0000	0.0000	1.0000	62	93	97	4	Along River	41	
0.0000	0.0000	0.0000	1.0000	62	92	97	4	Along River	41	
0.0000	0.0000	0.0000	1.0000	61	97	97	4	Along canal	41	
0.0000	0.0000	0.0000	1.0000	60	92	95	4	Along canal	41	
0.0000	0.0000	0.0000	1.0000	58	98	94	4	Dry (no canal)	41	
0.0000	0.0000	0.0000	1.0000	58	100	90	4	Dry (no canal)	41	
0.0000	0.0000	0.0000	1.0000	65	96	102	4	Dry (no canal)	41	
0.0000	0.0000	0.0000	1.0000	56	100	87	4	Along canal	14	
0.7038	1.0000	1.0000	1.0000	55	110	97	4	near built up	18	
0.0000	0.0000	0.0000	1.0000	55	102	92	4	dry near wasteland	18	
0.0000	0.0000	0.0000	1.0000	58	99	90	4	near road	18	
0.0000	0.0000	0.0000	1.0000	56	101	90	4	dry (near road)	19	
0.0000	0.0000	0.0000	1.0000	58	98	90	4	near built-up (dry)	14	
0.0000	0.0000	0.0000	0.2510	57	95	90	4	near built-up (dry)	19	
0.0000	0.0000	0.0000	1.0000	55	104	88	4	near built-up (dry)	16	

0.0000	0.0000	0.0000	1.0000	53	94	98	4	near road	16	
0.0000	0.0000	0.0000	1.0000	60	95	102	4	near river	41	
0.0000	0.0000	0.0000	1.0000	65	83	103	4	near river	41	
0.0000	0.0000	0.0000	1.0000	51	107	96	5	Near Canal	45	R2
0.0000	0.0000	0.0000	1.0000	48	112	92	5	Near Canal	45	R3
0.0000	0.0000	0.0000	1.0000	53	113	89	5	Near Canal	46	R2
0.0000	0.0000	1.0000	0.0000	50	112	98	5	Near Canal	46	R3
0.0000	0.0000	0.0736	1.0000	58	101	95	5	Dry (no canal)	46	R2
0.0000	0.0000	1.0000	1.0000	66	95	101	5	earlier waterlogged (now salt)	18	R2
0.0000	0.0000	0.0150	0.0000	70	94	119	5	Near Canal	16	R1
0.0000	0.0000	0.0000	1.0000	55	100	99	5	Dry (no canal)	16	R3
1.0000	0.0000	0.0000	1.0000	73	91	117	5	dry (no canal)	16	R2
0.0000	0.0000	0.0000	0.5842	67	93	116	5	dry (no canal)	16	R1
0.0000	0.0000	1.0000	0.0000	57	106	107	5		16	R3
0.0000	0.0000	0.0000	1.0000	69	91	112	5	adjoining road	16	R1
0.0000	0.0000	0.0000	1.0000	48	101	90	5		46	R3
0.0000	0.0000	0.0000	1.0000	60	94	106	5	earlier affected by waterlogged	50	R2
0.0000	0.0000	1.0000	1.0000	60	102	100	5	waterlogged	44	R2
0.6237	0.4620	1.0000	0.0000	51	138	115	5	near canal	44	R1
0.0000	0.0000	1.0000	0.0000	52	125	95	5		47	R3
0.0000	0.0000	0.0000	1.0000	50	112	86	5		46	R3
0.0000	0.0000	1.0000	0.0000	52	133	105	5		18	R3
0.0000	0.0000	1.0000	0.0000	51	120	92	5		18	R3
0.0000	0.0000	0.4770	0.0000	49	123	89	5		16	R3
0.0000	0.0000	1.0000	0.0000	50	119	97	5		16	R3
0.0000	0.0000	0.0000	1.0000	56	106	97	5		16	R3
0.4436	0.0000	0.1229	1.0000	70	89	104	5		41	R1
0.0000	0.0000	0.0000	1.0000	44	114	86	5	Adjoining canal	44	R3
0.0000	0.0000	0.0000	1.0000	62	99	103	5	Adjoining canal	43	R3
0.0000	0.0000	1.0000	0.0000	56	116	96	5	Adjoining canal and waterlogged	45	R3

Table A57: Summary of 76 support vectors resulting from SVM analysis using training data acquired under intelligent scheme of training data acquisition. The  $\alpha$  values in SVM are between a pair of classes as SVM is basically a binary classifier and, therefore, there are four  $\alpha$  values for the five classes in the analysis. The four columns follow some particular order based on class label, for example, the  $\alpha$  values for cotton class (label 3) has  $\alpha$  values in column 1, column 2, column 3 and column 4 with respect to class 1, 2, 4 and 5 respectively.







## REFERENCES

- Arora, M. K. and Foody, G. M., "Log-linear modelling for the evaluation of the variables affecting the accuracy of probabilistic, fuzzy and neural network classifications," *International Journal of Remote Sensing*, vol. 18, no. 4, pp. 785-798, 1997.
- Atkinson, P. M., Cutler, M. E. J., and Lewis, H., "Mapping sub-pixel proportional land cover with AVHRR imagery," *International Journal of Remote Sensing*, vol. 18 pp. 917-935, 1997.
- Atkinson, P. M., "Optimal Ground-Based Sampling for Remote-Sensing Investigations - Estimating the Regional Mean," *International Journal of Remote Sensing*, vol. 12, no. 3, pp. 559-567, 1991.
- Atkinson, P. M. and Tatnall, A. R. L., "Neural networks in remote sensing - Introduction," *International Journal of Remote Sensing*, vol. 18, no. 4, pp. 699-709, 1997.
- Atkinson, P. M., Foody, G. M., Curran, P. J., and Boyd, D. S., "Assessing the ground data requirements for regional scale remote sensing of tropical forest biophysical properties," *International Journal of Remote Sensing*, vol. 21, no. 13-14, pp. 2571-2587, 2000.
- Battiti, R., "Using Mutual Information for Selecting Features in Supervised Neural-Net Learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537-550, 1994.
- Benediktsson, J. A. and Sveinsson, J. R., "Feature extraction for multisource data classification with artificial neural networks," *International Journal of Remote Sensing*, vol. 18, no. 4, pp. 727-740, 1997.
- Betts, A. K., Ball, J. H., Beljaars, A. C. M., Miller, M. J., and Viterbo, P. A., "The land surface atmosphere interaction: A review based on observational and global modelling perspectives.," *Journal of Geophysical Research*, vol. 101 pp. 7209-7225, 1996.
- Boles, S. H., Xiao, X., Liu, J., Zhang, Q., Munkhtuya, S., Chen, S., and Ojima, D., "Land cover characterization of temperate east asia using multi-temporal vegetation sensor data," *Remote Sensing of Environment*, vol. 88, no. 1, pp. 157-169, 2004.
- Boser, B., Guyon, I., and Vapnik, V. N. A training algorithm for optimal margin classifiers. 144-152. 1992. Proceedings of 5 th Annual Workshop on computer Learning Theory.
- Brieman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., *Classification and regression Trees* Monterey, C.A.: Wadsworth, 1984.

Brodley, C. E. and Utgoff, P. E., "Multivariate Decision Trees," *Machine Learning*, vol. 19, no. 1, pp. 45-77, 1995.

Brown, M., Gunn, S. R., and Lewis, H. G., "Support vector machines for optimal classification and spectral unmixing," *Ecological Modelling*, vol. 120 pp. 167-179, 1999.

Buchheim, M. P. and Lillesand, T. M., "Semi-Automated Training Field Extraction and Analysis for Efficient Digital Image Classification," *Photogrammetric Engineering and Remote Sensing*, vol. 55, no. 9, pp. 1347-1355, 1989.

Campbell, J. B., *Introduction to Remote Sensing*, 3 ed. Taylor and Francis, 2002, London.

Canter, F., "Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification," *Photogrammetric Engineering and Remote Sensing*, vol. 63, no. 4, pp. 403-414, 1997.

Chavez, P. S., Jr., "An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data," *Remote Sensing of Environment*, vol. 24 pp. 459-479, 1988.

Chen, D. M. and Stow, D., "The effect of training strategies on supervised classification at different spatial resolutions," *Photogrammetric Engineering and Remote Sensing*, vol. 68, no. 11, pp. 1155-1161, 2002.

Congalton, R. G., "A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data," *Photogrammetric Engineering and Remote Sensing*, vol. 54, no. 5, pp. 593-600, 1988.

Congalton, R. G., "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sensing of Environment*, vol. 37, no. 1, pp. 35-46, 1991.

Cortes, C. and Vapnik, V., "Support-Vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

Cracknell, A. P., "Synergy in remote sensing - what's in a pixel?," *International Journal of Remote Sensing*, vol. 19, no. 11, pp. 2025-2047, 1998.

Curran, P. J. and Williamson, H. D., "The accuracy of ground data used in remote-sensing investigations," *International Journal of Remote Sensing*, vol. 6, no. 10, pp. 1637-1651, 1985.

Curran, P. J. and Williamson, H. D., "Sample-size for ground and remotely sensed data," *Remote Sensing of Environment*, vol. 20, no. 1, pp. 31-41, 1986.

De Colstoun, E. C. B., Story, M. H., Thompson, C., Comission, K., Smith, T. G., and Irons, J. R., "National park vegetation mapping using multi-temporal Landsat 7

data and a Decision tree classifier," *Remote Sensing of Environment*, vol. 85, no. 3, pp. 316-327, 2003.

DeFries, R. S. and Townshend, J. R. G., "NdvI-derived Land-cover classifications at a global scale," *International Journal of Remote Sensing*, vol. 15, no. 17, pp. 3567-3586, 1994.

DeFries, R. S., Hansen, M., Townshend, J. R. G., and Sohlberg, R., "Global land cover classifications at 8 Km spatial resolution: the use of training data derived from Landsat imagery in decision tree classifiers," *International Journal of Remote Sensing*, 1998.

Duda, T. and Canty, M., "Unsupervised classification of satellite imagery: choosing a good algorithm," *International Journal of Remote Sensing*, vol. 23, no. 11, pp. 2193-2212, 2002.

Emrahoglu, N., Yegingil, I., Pestemalci, V., Senkal, O., and Kandirmaz, H. M., "Comparison of a new algorithm with the supervised classifications," *International Journal of Remote Sensing*, vol. 24, no. 4, pp. 649-655, 2003.

Esposito, F., Malerba, D., and Semeraro, G., "A comparative analysis of methods for pruning decision trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 476-491, 1997.

Fisher, P., "The Pixel: a snare and a delusion," *International Journal of Remote Sensing*, vol. 18, no. 3, pp. 679-685, 1997.

Fitzpatrick-Lins, K., "Comparison of sampling procedures and data analysis for a Land-use and Land-cover map," *Photogrammetric Engineering and Remote Sensing*, vol. 47, no. 3, pp. 343-351, 1981.

Floriana, E., Donato, M., and Giovanni, S., "A Comparative Analysis of Methods for Pruning Decision Trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 476-491, 1997.

Foody, G. M., "Land-Cover classification by an artificial neural-network with ancillary information," *International Journal of Geographical Information Systems*, vol. 9, no. 5, pp. 527-542, 1995.

Foody, G. M. and Arora, M. K., "An evaluation of some factors affecting the accuracy of classification by an artificial neural network," *International Journal of Remote Sensing*, vol. 18, no. 4, pp. 799-810, 1997.

Foody, G. M., "The significance of border training patterns in classification by a feedforward neural network using back propagation learning," *International Journal of Remote Sensing*, vol. 20, no. 18, pp. 3549-3562, 1999.

Foody, G. M., "Status of land cover classification accuracy assessment," *Remote Sensing of Environment*, vol. 80, no. 1, pp. 185-201, 2002.

Foody, G. M., "Hard and soft classifications by a neural network with a non-exhaustively defined set of classes," *International Journal of Remote Sensing*, vol. 23, no. 18, pp. 3853-3864, 2002.

Foody, G. M., "Thematic map comparison: evaluating the statistical significance of differences in classification accuracy," *Photogrammetric Engineering and Remote Sensing*, vol. 70, no. 5, pp. 627-633, 2004.

Foody, G. M., Sargent, I. M. J., Atkinson, P. M., and Williams, J. W., "Thematic labelling from hyperspectral remotely sensed imagery: trade-offs in image properties," *International Journal of Remote Sensing*, vol. 25, no. 12, pp. 2337-2363, 2004.

Fraser, R. H., Abuelgasim, A., and Latifovic, R., "A method for detecting large-scale forest cover change using coarse spatial resolution imagery," *Remote Sensing of Environment*, vol. 95, no. 4, pp. 414-427, 2005.

Friedl, M. A. and Brodley, C. E., "Decision tree classification of land cover from remotely sensed data," *Remote Sensing of Environment*, vol. 61, no. 3, pp. 399-409, 1997.

Gallego, F. J. Crop Area Estimation in the MARS Project. Conference on ten years of MARS project, Brussels, pp.1-11, 1999.

Genderen Van, J. L. and Lock, B. F., "Testing land-use map accuracy," *Photogrammetric Engineering and Remote Sensing*, vol. 43, no. 9, pp. 1135-1137, 1977.

Gong, P. and Howarth, P. J., "An assessment of some factors influencing multispectral Land- Cover classification," *Photogrammetric Engineering and Remote Sensing*, vol. 56, no. 5, pp. 597-603, 1990.

Gualtieri, J. A. and Crompton, R. F. Support vector machines for hyperspectral remote sensing classification. 27. 1998. 27 th AIPR Workshop: Advances in Computer Assisted Recognition.

Zhan, H., Shi, P., and Chen, C., "Retrieval of oceanic chlorophyll concentration using support vector machines," *IEEE Transactions Geoscience Remote sensing*, vol. 41, no. 12, pp. 2947-2951, 2003.

Halldorsson, G. H., Benediktsson, J. A., and Sveinsson, J. R. Support vector machines in multisource classification. 2003. IEEE, IGARSS. 21-7-2003.

Hansen, M. C. and Reed, B., "A comparison of the IGBP DISCover and University of Maryland 1km global land cover products," *International Journal of Remote Sensing*, vol. 21, no. 6-7, pp. 1365-1373, 2000.

Hashemain, M. S., Abkar, A. A., and Fatemi, S. B. Study of sampling methods for accuracy assessment of classified remotely sensed data. 2005. Istanbul. 12-7-2004.

Hay, A. M., "Sampling design to test land use map accuracy," *Photogrammetric Engineering and Remote Sensing*, vol. 45 pp. 529-533, 1979.

Hixson, M., Scholz, D., and Fuhs, N., "Evaluation of several schemes for classification of remotely sensed data," *Photogrammetric Engineering and Remote Sensing*, vol. 46, no. 12, pp. 1547-1553, 1980.

Ho K.T, Hull J.J, and Srihari S.N, "Decision combination in multiple classification systems," *IEEE Transactions on Pattern Analysis and Machine Analysis*, vol. 16, no. 1, pp. 66-75, 1994.

Hord, R. M. and Brooner, W., "Land-use map accuracy criteria," *Photogrammetric Engineering and Remote Sensing*, vol. 42 pp. 671-677, 1976.

Hsu, C. W. and Lin, C. J., "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415-425, 2002.

Hsu, C. W. and Lin, C. J., "A simple decomposition method for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 291-314, 2002.

Huang, C., Davis, L. S., and Townshend, J. R. G., "An assessment of support vector machines for land cover classification," *International Journal of Remote Sensing*, vol. 23 pp. 725-749, 2002.

Huang, C., Townshend, J. R. G., Liang, S., Kalluri, S. N. V., and DeFries, R. S., "Impact of sensors point spread function on land cover characterization: assessment and deconvolution," *Remote Sensing of Environment*, vol. 80 pp. 203-212, 2002.

Jackson, Q. and Landgrebe, D. A., "An adaptive classifier design for high-dimensional data analysis with a limited training data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 12, pp. 2664-2679, 2001.

Jensen, L. L. F. and Van der Wel, F. J. M., "Accuracy assessment of satellite derived land-cover data: A review," *Photogrammetric Engineering and Remote Sensing*, vol. 60 pp. 419-426, 1994.

Kanellopoulos, I. and Wilkinson, G. G., "Strategies and best practice for neural network image classification," *International Journal of Remote Sensing*, vol. 18, no. 4, pp. 711-725, 1997.

Kavzoglu, T. and Mather, P. M., "The use of backpropagating artificial neural networks in land cover classification," *International Journal of Remote Sensing*, vol. 24, no. 23, pp. 4907-4938, 2003.

Kim, B. and Landgrebe, A., "Hierarchical classifier design in high-dimensional numerous class cases," *IEEE Transactions Geoscience Remote sensing*, vol. 29, no. 4, pp. 518-528, 1991.

Kulkarni, A. V. and Kanal, L. N. An optimization approach to hierarchical classifier design. 3. 1976. International Joint Conference Pattern Recognition.

Kurzynski, M. W., "The optimal strategy of a tree classifier," *Pattern Recognition*, vol. 16, no. 81, pp. 87, 1983.

Han, K., Champeaux, J., and Roujean, J., "A land cover classification product over France at 1 Km resolution using SPOT4/VEGETATION data," *Remote Sensing of Environment*, vol. 92, no. 1, pp. 52-66, 2004.

Wang, L., Sousa, W. P., Gong, P., and Biging, G. S., "Comparison of IKONOS and QuickBird images for mapping mangrove species on the Caribbean coast of panama," *Remote Sensing of Environment*, vol. 91, no. 3, pp. 432-440, 2004.

Lee, C. and Landgrebe, D. A., "Decision boundary feature extraction for neural networks," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 75-83, 1997.

Lillesand, T. M., Kiefer, R. W., and Chipman, J. W., *Remote Sensing and image Interpretation*, 5 ed. Wiley Text Books, 2004, NJ.

Liu, W. and Wu, E. Y., "comparison of non-linear mixture models: sub-pixel classification," *Remote Sensing of Environment*, vol. 94, no. 2, pp. 145-154, 2005.

Marcial, A. R. S., Borges, J. S., Gomes, J. A., and Da Costa, J. F. P., "Land cover update by supervised classification of segmented ASTER images," *International Journal of Remote Sensing*, vol. 26, no. 7, pp. 1347-1362, 2005.

Mather, P. M., "Land Cover Classification Revisited," in Atkinson, P. M. and Nicholas J. Tate (eds.) *Advances in Remote Sensing and GIS Analysis* John Wiley and sons Ltd., 1999, Chichester.

Mather, P. M., *Computer Processing of Remotely-Sensed images: an Introduction*, 2 ed. John Wiley and sons Ltd., 1999, Chichester.

Mcbratney, A. B. and Webster, R., "How Many Observations Are Needed for Regional Estimation of Soil Properties," *Soil Science*, vol. 135, no. 3, pp. 177-183, 1983.

Melgani, F. and Bruzzone, L., "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions Geoscience Remote sensing*, vol. 42, no. 8, pp. 1778-1790, 2004.

Mercier, G. and Lennon, M. Support Vector Machines for Hyperspectral Image Classification with Spectral-based kernels. Proceedings IEEE International

Geoscience and Remote Sensing Symposium . 2003.

Murthy, S., Salzberg, S., and Kasif, S., "A system for induction of oblique decision trees," *J.Artificial Intelligence Research*, vol. 2 pp. 1-33, 1994.

Navalgund, R. R., Parihar, J. S., Ajai, and Rao, P. P. N., "Crop Inventory using Remotely Sensed Data," *Current Science*, vol. 61, no. 3, pp. 162-171, 1991.

Niblett, T. Constructing Decision Trees in Noisy Domains. Progress in Machine Learning 87, 67-78. 1987.

Olthof, I., Butson, C., and Fraser, O., "Signature extension through space for northern landcover classification: A comparison of radiometric correction methods," *Remote Sensing of Environment*, vol. 95, no. 3, pp. 290-302, 2005.

Osuna, E. E., Freund, R., and Girosi, F. Support vector machines: Training and applications. ftp publications.ai.mit.edu . 1997.

Pal, M. and Mather, P. M., "An assessment of the effectiveness of decision tree methods for land cover classification," *Remote Sensing of Environment*, vol. 86, no. 4, pp. 554-565, 2003.

Pal, M. and Mather, P. M., "Assessment of the effectiveness of support vector machines for hyperspectral data," *Future Generation Computer Systems*, vol. 20, no. 7, pp. 1215-1225, 2004.

Pal, M. and Mather, P. M., "Support vector machines for classification in remote sensing," *International Journal of Remote Sensing*, vol. 26, no. 5, pp. 1007-1011, 2005.

Pinter, P. J., Ritchie, Jr. J. C., Hatfield, J. L., and Hart, G. F., "The agricultural research services remote sensing program: An example of interagency collaboration," *Photogrammetric Engineering and Remote Sensing*, vol. 69 pp. 615-618, 2003.

Pontius, R. G., "Quantification error versus location error in comparison of categorical maps," *Photogrammetric Engineering and Remote Sensing*, vol. 66 pp. 1011-1016, 2000.

Quinlan, J. R., "Induction of decision trees," *Machine Learning*, vol. 1, no. 81, pp. 106, 1986.

Quinlan, J. R., "Simplifying decision trees," *International Journal of Man-Machine Studies*, vol. 27, no. 3, pp. 221-234, 1987.

Richards, J. A. and Xiuping, J., *Remote sensing digital image analysis: An introduction*, 3 ed. Springer, 1998, Berlin.



Richards, J. A., "Classifier performance and map accuracy," *Remote Sensing of Environment*, vol. 57 pp. 161-166, 1996.

Rosenfield, G. H. and Fitzpatricklins, K., "A coefficient of agreement as a measure of thematic classification accuracy," *Photogrammetric Engineering and Remote Sensing*, vol. 52, no. 2, pp. 223-227, 1986.

Rounds, E., "A combined non-parametric approach to feature selection and binary decision tree design," *Pattern Recognition*, vol. 12, no. 313, pp. 317, 1980.

Safavian, S. R. and Landgrebe, D., "A Survey of Decision Tree Classifier Methodology," *IEEE Transactions on Systems Man and Cybernetics*, vol. 21, no. 3, pp. 660-674, 1991.

Schowengerdt, R. A., *Techniques for Image Processing and Classification in Remote Sensing* 1983, Academic press, London.

Shiva, V., "The Green Revolution in the Punjab," *The Ecologist*, vol. 21, no. 2, 1991.

Simard, M., Saatchi, S. S., and De Grandi, G., "The use of decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 5, pp. 2310-2321, 2000.

Singh, H. Delineation of waterlogged areas in muktsar district using remote sensing technology. L-94-AE-15-BIV, 1-24. 1998. Ludhiana, College of agricultural engineering, Punjab Agriculture University.

Smits, P. C., Dellepiane, S. G., and Schowengerdt, R. A., "Quality assessment of image classification algorithms for land cover mapping: a review and proposal for cost based approach," *International Journal of Remote Sensing*, vol. 20 pp. 1461-1486, 1999.

Steele, B. M., "Combining multiple classifiers: An application using spatial and remotely sensed information for land cover type mapping," *Remote Sensing of Environment*, vol. 74, no. 3, pp. 545-556, 2000.

Swain, P. H. and Husaka, H., "The decision tree classifier: design and potential," *IEEE Transactions Geoscience Remote sensing*, vol. 15 pp. 142-147, 1969.

Swain, P. H. and Davis, S. M., *Remote Sensing: the Quantitative Approach* McGraw-Hill Intl. Book Co, 1978, London.

Tatem, A. J., Noor, A. M., and Hay, S. I., "Defining approaches to settlement mapping for public management in Kenya using medium spatial resolution satellite imagery," *Remote Sensing of Environment*, vol. 93, no. 1, pp. 42-52, 2004.

Townshend, J. R. G. and Justice, C., "Information extraction from remotely sensed data, a user view," *International Journal of Remote Sensing*, vol. 2, no. 4, pp. 313-329, 1981.

Van Genderen, J. L. and Lock, B. F., "Testing land-use map accuracy," *Photogrammetric Engineering and Remote Sensing*, vol. 43, no. 9, pp. 1135-1137, 1977.

Vapnik, V. N., *Statistical Learning Theory* John Wiley and sons Inc., 1998.

Zhuang, X., Engel, B. A., Lonzanogarcia, D. F., Fernandez, R. N., and Johnnansen, C. J., "Optimization of training data required for neuro-classification," *International Journal of Remote Sensing*, vol. 15 pp. 3271-3277, 1994.