OIL PALM MAPPING USING SUPPORT VECTOR MACHINE WITH LANDSAT ETM+ DATA

NOONI ISAAC KWESI

May, 2012

SUPERVISORS:
Dr. I. van Duren, ITC, The Netherlands
Dr. A. Duker, KNUST, GHANA
Mr. L. Addae-Wireko, KNUST, GHANA

OIL PALM MAPPING USING SUPPORT VECTOR MACHINE WITH LANDSAT ETM+ DATA

NOONI ISAAC KWESI

Enschede, The Netherlands, May, 2012

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente and the Faculty of Renewable Natural Resources of the Kwame Nkrumah University of Science & Technology in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialisation: Natural Resource Management

SUPERVISORS:
Dr. I. van Duren, ITC, The Netherlands
Dr. A. Duker, KNUST, GHANA
Mr. L. Addae-Wireko, KNUST, GHANA

THESIS ASSESSMENT BOARD:
Dr. Yousif, Hussin (Chair), ITC, The Netherlands
Mr. B. Kumi-Boateng (External examiner), University of Mines & Technology, Ghana
Dr. E. M. Osei Jnr (Internal Examiner), KNUST, Ghana

ABSTRACT

Oil palm is cultivated extensively in the humid tropical land. It the most productive oil seed in the world because the economic importance of oil palm is in two distinct products; the palm oil and kernel oil. Historically, oil palm is native to West African coast and the palm oil is mainly used for cooking. Oil palm expansion and production in Ghana within the last 2 decades were due to factors such as commodity price, market availability and government intervention. Juaben oil mills located in Ejisu-Juaben district is one of the oldest mills in the country established during the post-independence era. Since its privatisation in 1992, the supply of adequate fresh fruits bunches has been a challenge due to demand. So the Ghanaian government with assistance separately from World Bank and Africa Development fund in 1997 and 2004 respectively launched oil palm plantation initiatives to boost palm oil production, improve employment opportunities while at the same time control rural-urban migration. However, the cultivation of oil palm has raised issues of environmental sustainability. To assess sustainability of palm oil production and oil palm expansion, the roundtable for sustainable palm oil has defined principles and criteria. Several of these criteria link to land use and land cover. Yet, there is insufficient guidance from roundtable for sustainable palm oilon how to map and quantify oil palm related land cover changes. So there is a need to develop a methodology to map oil palm related land cover changes at the local level. The study objective seeks to map oil palm related land cover of a section from northern portion of Ejisu-Juaben district in the Ashanti Region of Ghana using support vector machine (SVM) with Landsat ETM+. The districtlies within Longitude 6° 15' N and 7° 00' N and Latitude 1° 15' W and 1 ° 45' W and is characterised by both agricultural and socio-economic activities. The Landsat ETM+ data acquired in 2010 was used for processing and image classification. Field data were acquired in October 2011 through stratified random sampling. A total of 343 samples were collected for classification and accuracy assessment. The classification was carried out using MLC and SVM based on best three band combination from the image. The SVM and MLC performance evaluation was done using overall accuracy assessment and kappa statistics procedure. The results of separability analysis showed that ETM+ data provides spectral discrimination of land cover types found in the study area. The best three bands that provided the optimum spectral separability based on Bhattacharyya distance are 4, 5, and 3.The result showed that band 4, band 5 and band 3 provided best spectral separability. The overall accuracy result of the SVM classification was 78.29% (kappa statistic = 0.73). The RBF parameter setting in SVM was an important variable in the classification process, because it helped control the number of support vector used in the classification. The overall accuracy for MLC was 71.7% (kappa statistics = 0.65). The results indicated that SVM can improve the classification of oil palm mapping.The estimated area covered by oil palm was 904.95 ha and 993.78 ha for MLC and SVM respectively. SVM and MLC varied in their ability to map and quantify oil palm. SVM is more accurate than MLC. SVM is suitable method for identifying and mapping oil palm.

Key words: **support vector machine, maximum likelihood classifier, spectral separability, oil palm**

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Almighty God for His utmost guidance and protection. I am deeply grateful to my supervisors; Dr. Iris van Duren, Dr. A. Duker and Mr L. Addea-Wireko for their insightful comments, guidance and supervision. I am thankful to Dr. Michael Weir for his guidance during the field work and Dr. David Rossiter for his practical tutoring of R statistical language. My sincere appreciation goes to Dr. Valentyn Tolpekin for his patience, guidance and willingness to share his knowledge on support vector machine and taking time off his busy schedules to reply to my emails. Dr. Pal Mahesh (Department of Civil Engineering National Institute of Technology, India) for the suggestions I received through email correspondence. I am thankful to the Dutch government for sponsoring me during my stay and study in the Netherlands.

I also want to extend my gratitude to Mr Fynn (outgrower manager), Mr Ofori (Extension Officer), and Mr Kyei (Extension Officer) of Juaben Oil Palm Outgrowers Cooperative Society (JOPOCOS) for their time, energy and above all allowing the research team into their plantation. Thanks also go to the Samuel (driver), Mr Frimpong (Field assistant) for their devotion during the field work period. To my colleagues especially Abel Chemura and Enock Mutanga for their cooperation and support during and after the field work.

And thanks to Daniel Tutu Benefoh (Environmental Protection Agency, Accra), Divine Aboadoh (Environmental Protection Agency, Ho, Volta Region) and Emmanuel Boakye (Working Group on Forest Certification (FSC-Ghana), Dr. K. Forkuo, Dr. E.M. Osei Jnr for their suggestions and insightful comments. To my colleagues on the GISNATUREM programme; Lillian Lucy Lartey, Kofi Loh and Isaac Amoafo-Addo, I value your friendship and thank you very for the pieces of advice. To Franz Alex Gaisie-Essilfie, Eric Attah, George Asamoah, I wish them all the best on the programme and the sky should be their limit.

Finally, I thank my family especially my dear mum and two brothers, Richard Nooni and Alexander Nooni for their prayers and support. Not forgetting Rebecca Naa Aku Adamah for your prayers, love and care.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF PLATES

LIST OF EQUATIONS

LIST OF ACRONYMS

ADf: African Development fund

ECW: Enhanced Compressed Wavelength

ERDAS: Earth Resource Data Analysis System

EU: European Union

FAO: Food & Agriculture Organisation

FFB: Fresh Fruits Branches (FFB)

GDP: Gross Domestic Product

GIS: Geographic Information System

GNIWG: Ghana National Interpretation Working Group

GOPDC: Ghana Oil Palm Development Cooperation

GPS: Global Positioning System

GSS: Ghana Statistical Service

Ha: Hectare

ISODATA: Iterative self-Organising Data Analysis

JOPOCOS: Juaben Oil Palm Outgrowers Cooperative Society

MLC: Maximum Likelihood Classification

RBF: Radial Basis Function

RMSE: Root Mean Square Error

RSPO: Roundtable on Sustainable Palm Oil

SVM: Support Vector Machine

1. GENERAL INTRODUCTION

1.1 Background

Oil palm (*Elaeis guineensis*) is a perennial crop, which is cultivated extensively in the humid tropical land. It is one of the most productive oil seed in the world (Figure 1) and becoming an increasingly important agricultural product for tropical countries around the world (Butler *et al.*, 2009) because the economic importance of oil palm is in two distinct products; the palm oil and kernel oil. Historically, oil palm is native to West African coast and originated from this region (FAO, 2005b). Traditionally, palm oil is mainly used for cooking in Western Africa (Thenkabail *et al.*, 2004). Although oil palm is regarded as an African crop, it is now found and grown in countries with similar tropical climate. Tropical forest areas are ideal, because rainfall is plentiful, temperatures and humidity are high. Oil palm is now an important crop for countries in the Far East and the Americas where the climatic conditions favours its growth (FAO, 2005b).

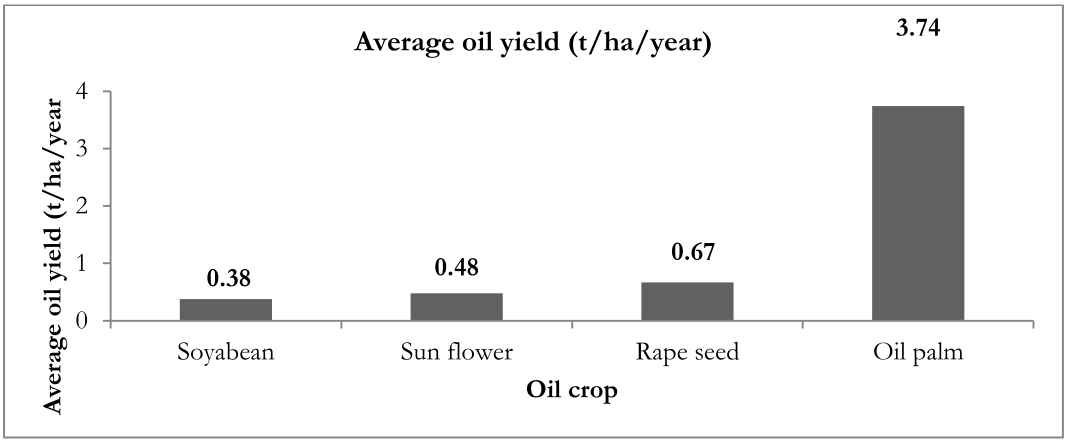


Figure 1: Annual yield of oil crop for the year 2007

Source: Oil world (2007)

Global demand for production of palm oil has increased (Figure 2) over the last 20 years. The demand may be attributed to high consumption of palm oil as a result of high population growth, cosmetic and bio-fuel industry (Koh & Ghazoul, 2008). Asia, particularly Indonesia and Malaysia are estimated to be the world's top producers of palm oil accounting for 87% percentage of global production (Huguenin *et al.*, 2007). For instance, Malaysia due to its large oil palm plantation, has utilised palm oil in the production of biodiesel for buses and cars (Yusoff, 2006). In Brazil, biodiesel from oil palm (Da Coata, 2004) is used to generate electricity (Coelho *et al.*, 2005).

In Africa, the case is not different as oil palm plantations can be traced back to pre-colonial days in Western Africa. In Cameroon, oil plantations were promoted and established by the Germans and further developed under the Franco-British regimes before becoming stated-owned after independence (Carrere, 2010). In Ghana, oil palm plantations were grown during the pre-colonial era initially along the Ghanaian coast before spreading to forest zone of the country. These oil palm plantations and mills later become state-owned after independence (Gyasi, 1992).

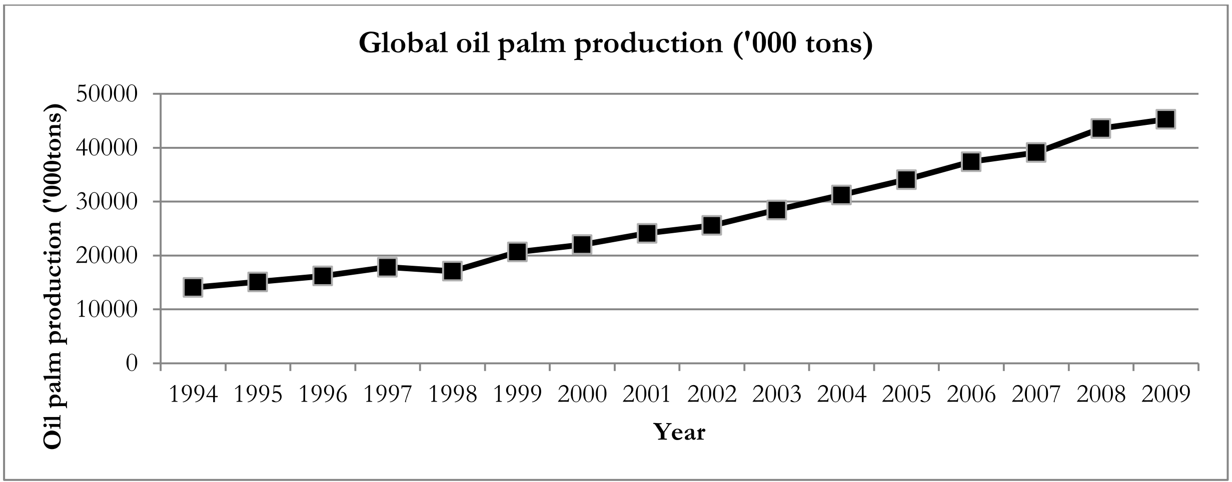


Figure 2: Global Palm oil production ('000tons) from 1994-2009

Source: World oil (2012)

Oil palm expansion and production in Ghana within the last 2 decades were as a result of factors such as price of palm oil, market availability due to existence of mills to process the palm fruits, government intervention as a means of generating employment (Carrere, 2010; World Bank-IFC, 2008). Juaben oil mills located in Ejisu-Juaben district is one of the oldest mills in the country established during the post-independence era. Since its privatisation in 1992, the supply of adequate fresh fruits bunches (FFB) has been a challenge due to demand (RSPO, 2011). So in 1997, the World Bank initiated oil palm growing project targeting smallholder plantings as a strategy to generate employment and reduce poverty in the district (Carrere, 2010; Gyasi, 2003). Additionally, in 2004, the Government of Ghana with assistance from Africa Development fund (ADf) undertook similar project called presidential special initiative (PSI) in oil palm growing areas including the Ejisu-Juaben district (Carrere, 2010). In both initiatives, free seedlings and extension services were offered to prospective farmers. Currently there are over 630 registered smallholder oil palm plantings and one large holder plantation in the district that provide raw palm fruits to the Juaben oil mills (Personal communication).

The definition of smallholder according to RSPO is farmers growing oil palm under land area of 40 hectares or less. Oil palm plantations occupying more than 40 hectares are group under medium to large holder plantation (RSPO, 2007). According to Ghana Poverty Reduction Strategy II (GPRS II, 2006) document, these initiatives have provided employment to the youth thereby controlling rural-urban migration and improved the standard of living in these oil palm growing communities.

However, cultivation of oil palm has raised issues of sustainability (Tan *et al.*, 2009) as it brings about environmental problems such as deforestation, degradation, reduction in biodiversity loss (Koh, 2008; Koh & Ghazoul, 2010). To assess sustainability of palm oil production and oil palm expansion, the Roundtable for Sustainable Palm Oil (RSPO) has defined principles and criteria (RSPO, 2005; Tan *et al.*, 2009). These principles and criteria aims at adopting proactive and multi-stakeholder approaches towards achieving certification of sustainable oil palm production (RSPO, 2007). The policy stems from the belief that expansion of oil palm, production and marketing of palm oil at the global market can be done in a clear and transparent manner without significantly compromising ecological and socio-economic sustainability (RSPO, 2007; Tan, 2007; Tan *et al.*, 2009). Several of these criteria link to land use and land cover (RSPO, 2007). In this regard, RSPO outlined in principle 7 and criteria 5 & 7 to tackle environmental challenges associated with oil palm expansion. Specifically elaborated in Principle 7 is development of new plantings and criterion 7.3 is new plantings should not replace tropical rain forest or high conservation areas (RSPO, 2007, 2009). Nonetheless, assessment of these criteria requires spatial and temporal information of oil palm related land cover changes and adopting remote sensing based approach (Laurance *et al.*, 2010) becomes a reliable option since field based approach has shown to be costly in terms of time and coverage (Janssen & van der Wel, 1994). At the moment there is insufficient guidance from RSOP on how to map and quantify oil palm related land cover changes for certification, especially for smallholder oil palm plantings in Ghana (RSPO, 2009, 2011). Remote sensing is viewed as the tool for obtaining such oil palm related cover information (McMorrow, 1995; Thenkabail *et al.*, 2004). Several research studies have applied satellite images and different methods in identifying and quantifying oil palm cover (Wahid *et al.*, 2005; Zhang *et al.*, 2009; Zhang & Zhu, 2011), however, the different classification methods employed namely object oriented classification (Wahid *et al.*, 2005), spectral angular mapper (Kamaruzaman & Mubeena, 2009), linear regression modelling (McMorrow, 1995, multiple regression modelling (Ibrahim, 2000), empirical regression modelling (Thenkabail *et al.*, 2004) targeted at mapping age related oil palm cover where mainly conducted in large holder oil palm plantations (Kamaruzaman & Mubeena, 2009; McMorrow, 2005; Thenkabail *et al.*, 2004; Wahid *et al.*, 2005) and its extension to include smallholder oil palm planting has rarely not been investigated (RSPO, 2011). Also, most of the classification methods applied are sophisticated and requires special knowledge or skill to use. Studies conducted by Wahid *et al*, (2005) and Ibrahim, (2000) mapped age related oil palm map used object oriented classification with Landsat TM and Landsat ETM+ images respectively but the research were focussed in large holder plantations.

Furthermore, studies that used high resolution images (Kamaruzaman & Mubeena, 2009; Thenkabail *et al.*, 2004) to produce age related oil palm mapping has shown to be costly when extending to larger areas (Kamaruzaman & Mubeena, 2009; Thenkabail *et al.*, 2004). Currently, in Ghana, smallholders cultivate nearly 88% of the total area under production whilst the large holder estates cultivate less than 12% of the total area (GOPDC, 2011). This means that smallholder oil palm cultivation is viewed as a lucrative venture (Butler & Laurance, 2009). Therefore, developing a methodology that will focus on smallholder plantings will improve on the present methods of mapping oil palm related cover. The method may also useful for RSPO certification of smallholder oil palm plantings (RSPO, 2011) as stipulated in criteria 5 and 7 vis-à-vis environmental assessment and integrity (RSPO, 2009, 2011). Thus, contributing to studies, it is important to develop a methodology for mapping oil palm especially smallholder plantings using medium resolution images such as Landsat ETM+ in a heterogeneous environment. This methodology should consider different age variability of smallholder planting and the complexities involved in separating the other cover types bordering the smallholder plantings in an occurring a heterogeneous environment.

Accordingly, other supervised classification methods such as maximum likelihood classifier, neural networks, and decision trees have widely been used to obtain land cover information with relatively high classification accuracies. This is because the softwares employed are readily available, easy to use and relatively cheaper than for example, the e-cognition software used for object oriented classification (Foody & Mathur, 2004; Han *et al.*, 2002; Huang *et al.*, 2002; Pal & Mathur, 2003) where affordability and use may be a challenge to resource managers in developing countries where periodic monitoring and evaluation of natural resources are essential.

One of the ways of obtaining such land cover information is through the most widely used maximum likelihood classification algorithm. Maximum likelihood classifier is an example of supervised classification specifically parametric classifier (Jensen, 2005). The principle of maximum likelihood classification is based on the assumption that training data of each image band is normally distributed (Pal & Mather, 2003). But field training data is rarely normally distributed and thus pose a limitation in this type of classifier (Huang *et al.*, 2002; Pal & Mather, 2003). As a result, many advanced classification algorithms such as neural network, decision trees had emerged for land cover mapping (Foody & Mathur, 2004a; Huang *et al.*, 2002; Pal & Mather, 2003) and results show that these classifiers generally present an improved classification accuracies relative to maximum likelihood (Huang *et al.*, 2002; Pal & Mather, 2003). Despite this success, research continues to search for methods to further upgrade classification accuracies (Foody *et al.*, 2006).

In this regard, support vector machine, originally based on binary function, is viewed as one of the new ways of improving classification accuracies in remote sensing studies (Foody & Mathur, 2004a; Huang *et al.*, 2007). This is because support vector machine (SVM) has the tendency to minimise classification error by minimising the probability of misclassifying field data drawn randomly from a fixed but unknown probability distribution (Vapnik, 1995, 1998).

The support vector machine classification basically takes inputs from training data and predicts for each given inputs, which of the two classes forms the input by relating the training data set to each pixel in an image. It then operates to find a wide separating boundary between class pair by marking each pixel to belong to a class based on inputs (Foody & Mathur, 2004a; Kavzoglu & Colkesen, 2009). This is made possible through the use of a kernel function. The kernel function builds a model that assigns new classes into one class or the other. Later, test inputs can be mapped into the same space and predicted based on the side of the boundary they fall. This operation uses only pixels that lie close to the boundary called support vectors in the classification (Kavzoglu & Colkesen, 2009; Vapnik, 1995). A kernel function is used to train the classifier (Kavzoglu & Colkesen, 2009). Depending on the kernel type used, classification accuracies are improved (Huang *et al.*, 2002). But these usually comes at the expense of training time or speed as it can result in more computations (Huang *et al.*, 2002; Zhu & Blumberg, 2002). In literature, four kernel functions have been developed and reported. They are Gaussian radial basis filter (RBF), linear function, polynomial and sigma (Huang *et al.*, 2002; Zhu & Blumberg, 2002).

Although, the classification accuracy produced by support vector machine (SVM) depends on the type of kernel function used, Gaussian radial basis filter (RBF) kernel is the most widely applied kernel function in support vector machine (SVM) classification (Foody & Mathur, 2004a & 2004b). This is because the support vectors that are used in the classification are controlled by the kernel specific function parameter through cross validation (Vapnik, 1995). The significance of the support vectors in support vector machine (SVM) classification is intended to minimise confusion between classes (Huang *et al.*, 2002).

When Gaussian radial basis filter (RBF) is used, two parameters namely cost parameter (*C*) and kernel specific function parameter (γ) needs to be defined. The cost parameter (*C*) controls the penalty of wrongly placed pixels or support vectors that lie on the other side (Foody *et al.*, 2006; Hue *et al.*, 2010). The kernel specific function (γ) parameter takes care of minimising the training error (Foody *et al.*, 2006; Foody & Mathur, 2004a & 2004b).

One advantage of using support vector machine is its extension from two classes to include multiclass classification. This is done by adopting a multiclass approach (Vapnik, 1998). Several advanced approaches has been proposed and used in multiclass classification. One of approach is one-against-one (Melgani & Bruzzone, 2004; Vapnik, 1998). The use of one-against-one approach helps in building more classes and it keeps the size of training data smaller for training (Melgani & Bruzzone, 2004).

Within this context, various studies have outlined criteria for assessing performance algorithm to determine which classifier performs best (Foody *et al.*, 2006; Huang *et al.*, 2002; Jensen, 2005; Pal & Mather, 2005). For example, the use of sampling design (Jensen, 2005), sample size (Foody *et al.*, 2006; Huang *et al.*, 2002; Pal & Mather, 2005), image bands selection technique (Bruzzone & Serpico, 2000; Rahman *et al.*, 2005; Sanaeneijad *et al.*, 2009; Zhang *et al.*, 2009; Zhu & Blumberg, 2002) or separability test statistics (Kusimi, 2008) coupled with accuracy assessment and chi square statistics (Congalton, 1991) have been reported.

To determine which classifier gives high accuracy assessment for oil palm mapping, ground truth data has to be collected. Several sampling methods to collect ground truth data have been proposed: random, systematic, stratified systematic unaligned, and cluster sampling (Fitzpatrick-Lins, 1981; Jensen, 2005). Stratified random sampling approach strategy is preferred because of its reasonable approach to achieve results with high precision and reduce variation in the sampling unit (Jensen, 2005). Another consideration is the sample size used in classification (Foody & Mathur, 2004a). A guideline for choosing minimum size of samples for land cover classes have been recommended in literature (Congalton, 1991). This means that the number of samples may be adjusted based on the research study area (Jensen, 1996).

Further consideration is given to optimum bands from satellite images that provide best separation between classes of interest (Foody *et al.*, 2004a; Kusimi, 2008; Zhang *et al.*, 2009). Oil palm mapping using best bands composition from satellite images (McMorrow, 1995; Sanaeinejad *et al.*, 2009; Thenkabail *et al.*, 2004; Wahid, 1998) is ongoing. This is because maximising information from such image bands improves accuracy as well as reduces cost (Bruzzone & Serpico, 2000; Foody, 2002; Thenkabail *et al.*, 2004; Zhang *et al.*, 2009). Techniques such as principal component analysis (Huttich *et al.*, 2009), Jeffries Masuitita (Kusimi, 2008), Bhattacharyya statistical distance and Mahalanobis distance (Bruzzone & Serpico, 2000; Rahman *et al.*, 2005) have been used in literature to select image bands that provide best spectral information for classification (Zhang *et al.*, 2009). Principal component analysis is mostly used for determining best bands information however, its use alters the original image data making it difficult to relate which class pair are being distinguished (Bruzzone & Serpico, 2000; Zhang *et al.*, 2009). Applying statistical separation test using Bhattacharyya distance has become a criterion measure for band selection because it is easy to use (Bruzzone & Serpico, 2002; Zhang *et al.*, 2009). Best bands or band combinations are selected on the basis that the band that provide maximum separation between training class pair (Rahman *et al.*, 2005) will be easier to separate individual land cover class during classification (Zhang *et al.*, 2009). The optimise bands combination is then used as an input to the classifier for classification.

After classification, the final map has to be validated (Foody, 2002). The accuracy assessment measures overall accuracy and the kappa coefficients as well as individual producer and user accuracies in a form of contingency table (Congalton, 1991). The table has columns and rows that represent the reference data and classification results. Kappa statistics determines the extent of classification results (Lillesand *et al.*, 2004) and chi square statistic to test misclassified proportions in the confusion matrix has been reported as the primary criteria applied in remote sensing studies (Foody *et al.*, 2006; Huang *et al.*, 2002).

On these backgrounds, the study seeks to use separability test statistics, overall accuracy and kappa statistics to evaluate the performance of support vector machine (SVM) classifier at mapping oil palm related land cover in comparison to the most acclaimed maximum likelihood classification. The study will focus on the separability accuracies of the land cover classes involved using separability statistic test method of Bhattacharyya distance in Multi-Spec software; overall accuracy and kappa statistics with which both classifiers uses same ground truth data to estimate oil palm planting. This SVM approach targeted at oil palm mapping is new because it has not been applied in mapping smallholder oil palm related cover changes in a heterogeneous environment.

1.2 Research Objective

The study seeks to map oil palm related land cover of a section from the northern portion of Ejisu-Juaben district using support vector machine with Landsat ETM+.

The specific objectives are:

a. to evaluate the spectral separability of oil palm in relation to forest, shrub, other crops and bare
b. to analyse the performance of the support vector machine and maximum likelihood in mapping oil palm related cover using overall accuracies and Kappa statistics procedures
c. To map the spatial distribution of oil palm in the study area

1.3 Research Questions

1. Which spectral bands provide best spectral separation for mapping oil palm
2. What level of classification accuracy is attained by using
    i) support vector machine algorithm
    ii) maximum likelihood classifier
3. How well does the two classification algorithm map the spatial distribution of oil palm in the study area?

2. CONCEPTS& DEFINITION

2.1 Bhattacharyya distance

The Bhattacharyya distance is a band selection technique that uses statistical probability distribution function to measure how well two class pair are separable based on their signatures or reflectance contained in bands of satellite image data (Bruzzone & Serpico, 2000; Huttich *et al.*, 2009; Rahman *et al.*, 2005). The resulting output can be used to select the optimum subset of bands to distinguish between cover types occurring in an area. This is determined by calculating Bhattacharyya distance between two class pair by considering pixels in each band of the image. The class mean vectors and covariance matrices are estimated. It then counts the average value of the Bhattacharyya distance per each class pair and sort them based on the maximum distance or weighted interclass distance for each class (Zhang *et al.*, 2009). The results are presented in a form of table listing all possible pairwise combinations indicating the degree of similarity or difference in reflectance between land cover classes. The following mathematical illustration of Bhattacharyya distance is based on Bruzzone & Serpico, (2000). In mapping oil palm occurring in an area characterised with a heterogeneous landscape (Benefoh, 2008), in which a training data set, described by an *n* dimensional feature vector $F = (x_1, x_2, \dots x_n)$ in the feature space F, is assigned to one of c different classes $F = (\omega_1, \omega_2, \dots \omega_n)$ characterised by a priori probabilities $P(\omega_i)(i = 1,2, \dots c)$. Let $p = ({}^{x}/_{\omega_i})$ be the conditional probability density functions for the feature vector x, given the class ω_i(i=1, 2… c). Here, the criterion for selecting best bands or group of bands is based on band(s) that provides maximum average and weighted interclass distance separation for training class pair shown as:

$$\boldsymbol{B_{ave}} = \sum_{i=1}^{c} \sum_{j=1}^{c} \boldsymbol{P(\omega_i)P(\omega_j)B_{ij}} \qquad \text{Equation 1}$$

Where $\boldsymbol{B}_{ij}$, the Bhattacharyya distance between two classes, ω_i and ω_j, and may be expressed as a continuous probability functions in

$$\boldsymbol{B_{ij}} = -\ln\left\{\int \sqrt{\boldsymbol{p}\left(\frac{x}{\omega_i}\right)\boldsymbol{p}\left(\frac{x}{\omega_j}\right)\boldsymbol{dx}}\right\} \qquad \text{Equation 2}$$

$\boldsymbol{B}_{ij}$ is a measure of the average statistical distance between the conditional probability density functions related to two classes. For multivariate Gaussian distributions $\boldsymbol{B}_{ij}$ may be simplified as

$$\boldsymbol{B_{ij}} = \frac{1}{8}(\boldsymbol{m_i} - \boldsymbol{m_j})^t \left(\frac{\sum i + \sum j}{2}\right)(\boldsymbol{m_i} - \boldsymbol{m_j}) + \log\left[\frac{\left|\frac{\sum i + \sum j}{2}\right|}{\sqrt{\left|\sum i\right|\left|\sum j\right|}}\right] \qquad \text{Equation 3}$$

Where m_i and m_j and $\sum i, \sum j$ are the mean vectors and the co-variance matrices, respectively, for the classes ω_i and ω_j.

2.2 Maximum Likelihood algorithm

The maximum likelihood classifier basically develops a probability function based on inputs from a training dataset. It then considers each individual pixel in an image, compares it with known pixels and assigns unknown pixels to a class based on similarity and highest probability to belong to one of the already known classes (Jensen, 2005). Implementing maximum likelihood classifier involves the estimation of class mean variance and covariance matrices using training patterns chosen from known pixels of each particular class (Vikesh *et al.*, 2010; Cortijo & Perez de la Blanca, 1996b). The mathematical theory behind maximum likelihood expressed below follows Pal & Mather (2003).

The classifier assumes that members of each class is normal distributed in feature space and can be defined as follows: a pixel with an associated observed feature vector X is assigned to class "c_j" if

$$X\epsilon \text{"} c_j \; if \; g_j(X) > g_k(X) \; for \; all \; j \neq k, \qquad j,k = 1, \ldots, N$$ Equation 4

For multivariate Gaussian distributions $g_k(X)$ is given by: $g_k(X) = \ln\left(p(c_j)\right) - \frac{1}{2} ln \left| \sum_k \right| - \frac{1}{2}(X - M_k)^t \sum_k^{-1}(X - M_k)$ Equation 5

Where M_k and $\sum_k$ are the sample mean vector and covariance matrix of class$'k'$, and $g_k$is the g_k is a discriminating function.

2.3 Support Vector Machine

Support vector machine as explained earlier was developed based on a non-probability binary function which takes inputs from training dataset and predicts for each given inputs, which of the two classes forms the input by relating it to each pixel in the image. The known pixels of the training set are each marked to belong to one of the two classes. The support vector machine (SVM) training algorithm (i.e. kernel function) then builds a model that assigns new classes into one class or the other. This operation is carried out in feature space, where classes are separated by boundary that is wide as possible. Unseen data in the training set can be mapped into the same space and predicted to classes based on which side of the boundary they fall (Vapnik, 1995, 1998). Support Vector Machines (SVM) were first introduced as a machine learning method by Cortes and Vapnik (1995). A more detailed description of support vector machine that follows is based on Foody *et al.*, (2006) and Vapnik (1998).

Consider the training data represented by $\{x_i, y_i\}, i = 1, \ldots, r, y_i \in \{1, -1\}$ in F dimensional space. Where x_i is the observed spectral response and y_i the class label for a training case. In this instance, only an optimal hyperplane or boundary that separates the two classes in the training dataset is determined in feature space.

A hyperplane can be defined by the equation $w.x_i + b = 0$, where x is the point lying on the hyperplane, w is normal to the hyperplane; b is the bias and $\frac{|b|}{||w||}$ is the perpendicular distance from the hyperplane to the origin (see Figure 3). For linear separation, a separable hyperplane can be defined for the two classes as: $w.x_i + b \geq +1$ $(for\ y_i = +1)$ and $w.x_i + b \leq +1$ $(for\ y_i = -1)$. The two equations can be combined as

$\boldsymbol{y_i(w.x_i + b) - 1 \geq 0}$ Equation 6

The training data points found on these hyperplanes (F1 and F2) are referred to as support vectors and are central to the establishment of the optimal separating hyperplane (see Figure 3).

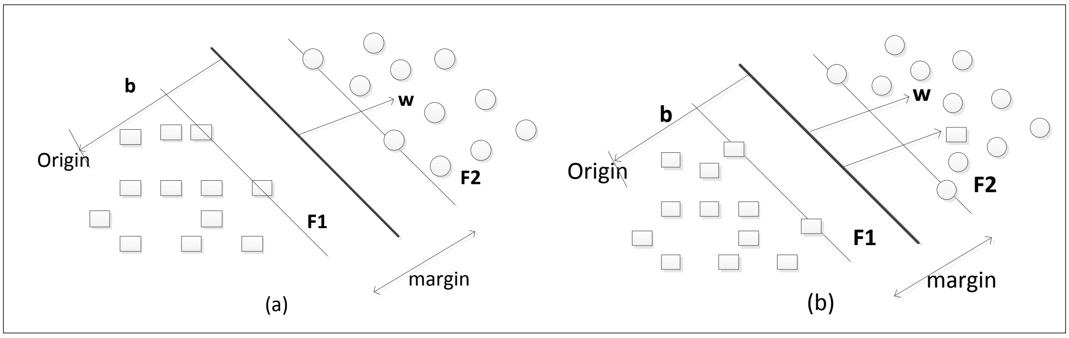


Figure 3: Basics of classification by an SVM. (a) Seperable case and (b) nonseperable case

These support vectors of the two classes lie on the two hyperplane parallel to the optimal hyperplane and are defined by $w.x_i + b = \mp 1$. The margin between these planes is $\frac{|2|}{||w||}$. The maximisation of this margin leads to the following constrained optimisation problem under the inequality constraints of equation (1).

$\boldsymbol{min\left\{\frac{1}{2}||w||^2\right\}}$ Equation 7

But in situations where the classes are not linearly separable, a slack variable, $\{\xi_i\}_{i=1}^{r}$, that indicate the distance the sample is from the optimal hyperplane to the class to which it belongs. This allows a certain amount of constraints to be introduced. The constraints then becomes,

$\boldsymbol{y(w.x_i + b) > 1 - \xi_i}$ Equation 8

The above constraints, in the case of outliers are contained in data, can always be met by making ξ_i are very large, so a penalty term, $C\sum_{i=1}^{r}\xi_i$ is added to penalise solutions for which ξ_i are very large. The constant *C* controls the magnitude of the penalty that is associated with training samples that lie on the wrong side of the decision boundary. With a low value of *C*, an inappropriately large fraction of support vectors may be derived while with a large value of *C* there is a danger of the SVM over fitting to the training data and so having low generation ability. With the addition of the penalty, the optimisation problem becomes

$$\boldsymbol{min}\left[\frac{||\boldsymbol{w}||^2}{2} + \boldsymbol{C}\sum_{i=1}^{r}\boldsymbol{\xi_i}\right] \quad \text{Equation 9}$$

If the approach is extended to allow non-linear decision surfaces, the input data are for example mapped into high dimensional space through some nonlinear mapping which has the effect of spreading the distribution of the data points in a way that facilitates the fitting of a hyperplane. This leads to decision functions of the form,

$$\boldsymbol{f(x)} = \boldsymbol{sgn}\left(\sum_{i=1}^{r} \propto_i \boldsymbol{y_i k(x, x_i)} + \boldsymbol{b}\right) \quad \text{Equation 10}$$

Where$\propto_i, i = 1, \ldots, r$ are Lagrange multipliers and $k(x, x_i)$ is a kernel function. The magnitude of $\propto_i$ is determined by the parameter *C* and lies on scale of 0-*C* (Belousov et al., 2002). The kernel used must meet Mercer's (Vapnik, 1995). Radial Basis function is one of the kernels that satisfy this condition.

$$\boldsymbol{k(x, x_i)} = \boldsymbol{e}^{-\gamma ||(x-x_i)||} \quad \text{Equation 11}$$

Where γ is the parameter controlling the width of the Gaussian kernel. The accuracy produced by SVM classifier is influenced by the magnitude of setting *C* and γparameter which can achieved through trials (cross validation). The trials are carried out until an optimal parameter setting for *C* and γ are achieved. Usually, depending on the training size the classification accuracies are improved but come at the expense of training time due to more computations (Foody & Mathur, 2004; Huang *et al.*, 2002).

2.4 Multiclass Support Vector Machines

As stated earlier, support vector was originally designed to handle binary (two class) classification; however, it has been modified and extended to deal with multiclass classification. This can be achieved using two common approaches: one-against-all and one-against-one approaches (Vapnik, 2008). The principle as well as the strength and limitation of the two approaches are well explained by Melgari & Bruzzone (2004). Since, the land cover classification mostly involve more than two classes, researchers adopt one-against-one class because the approach makes building of classes easier and flexible (Burges, 1998; Melgani & Bruzzone, 2004).

3. MATERIALS AND METHODS

3.1 Study Area:

The Ejisu-Juaben district is located in the central part of the Ashanti Region and it lies within Longitude 6° 5' N to 7° 00' N and Latitude 1° 15' W to 1 ° 45' W. The district stretches over an area of about 637.2 km². The study was conducted in Bomfa, Apemso, Kote, Apraku, Juaben, and Ejisu farming communities with favourable agro-climatic conditions; located within the northern portion of Ejisu-Juaben district as shown in Figure 4 and Figure 5.

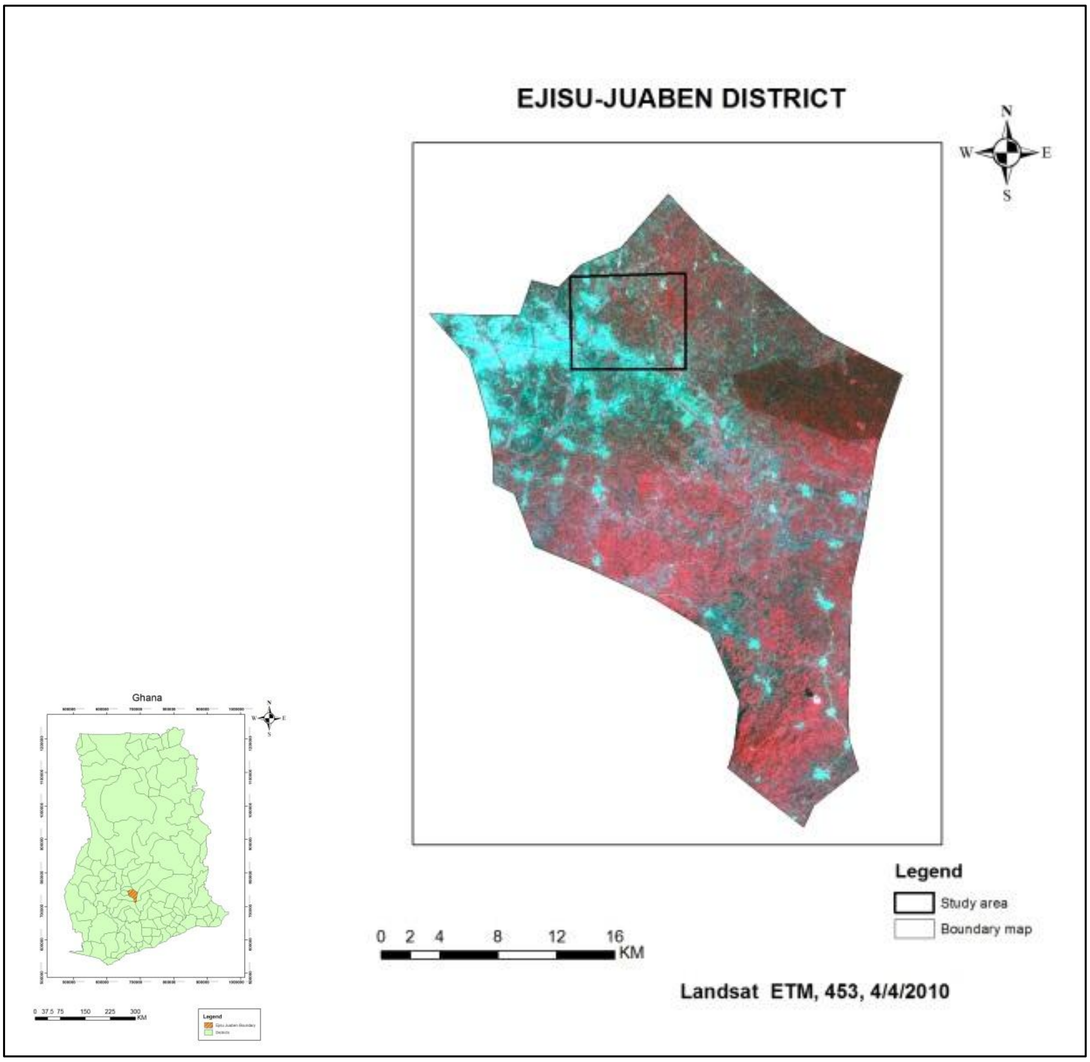


Figure 4: District map of Ghana showing false colour composite of Landsat ETM+ 2010 and the location of Ejisu-Juaben district and study area

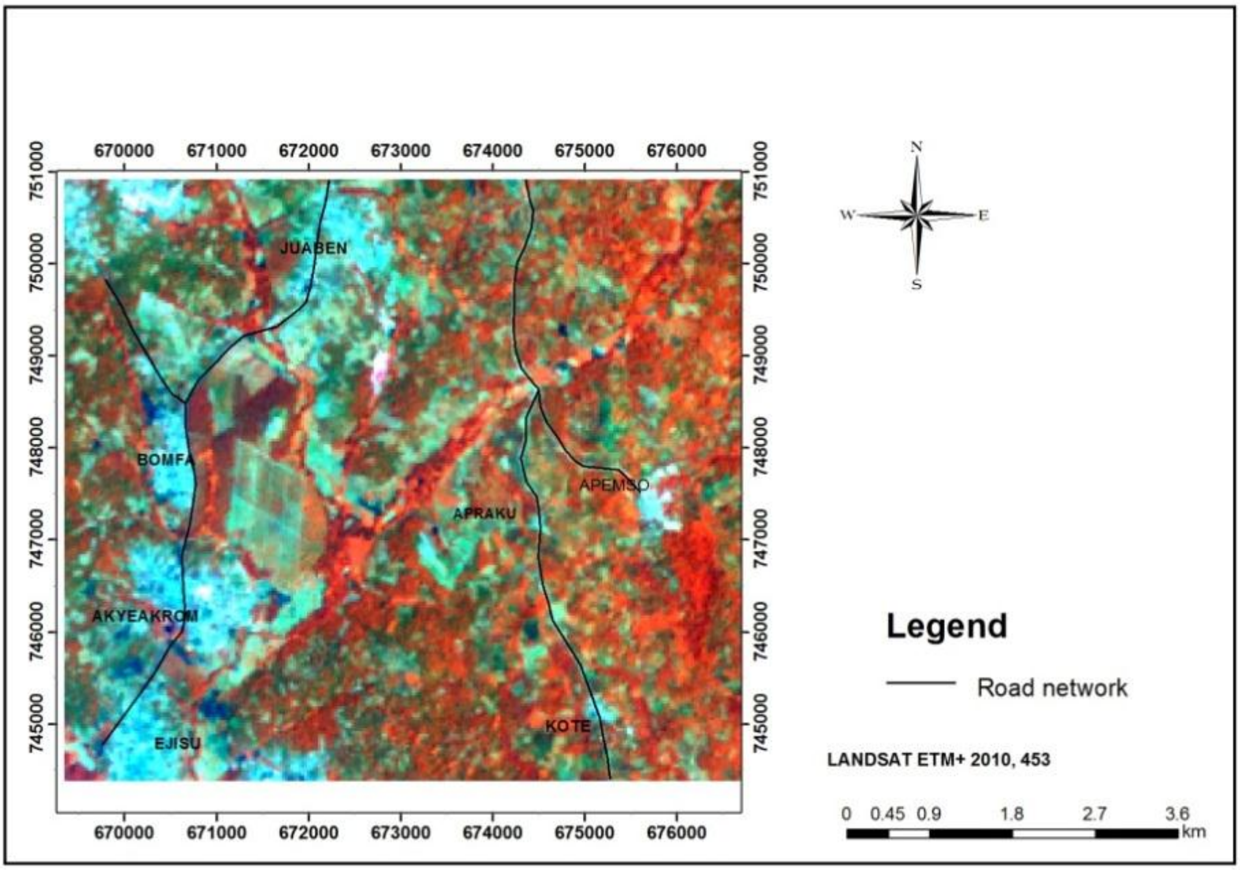


Figure 5: False colour composite of Landsat ETM+ 2010 showing the road network and locations of communities in the study area

The district experiences tropical rainfall and wet semi-equatorial climate. It is characterised by double maxima rainfall lasting from March to July and again from September to November. The mean annual rainfall is 1200mm. Temperatures range between 20°C in August and 32°C in March. Relative humidity is fairly moderate but quite high during rainy seasons and early mornings. The fair distribution of temperature and rainfall patterns enhances the cultivation of many food and cash crops (such as cocoa and oil palm) throughout the district thus making it a food sufficiency district in Ghana. The Ejisu-Juaben district falls within the forest dissected plateau terrain region. It rises from about 240 metres to 300 metres above sea level. The soils types in the district offer vast opportunities for the cultivation of traditional and non-traditional cash crops and other staple food stuff.

3.2 Materials

3.2.1 Data

Landsat ETM+ image (04/02/2010, Level 1 B with path/ row 194/55) with less than 10% cloud cover was obtained for the study. The image was selected and downloaded from ITC database. The data was chosen based on the following considerations; cost, percentage of cloud cover and image availability. A boundary shapefile of the Ejisu-Juaben district was used in the creation of the image of the study area (Figure 6). The shapefile was used to clip the Landsat ETM+ image to obtain an image of the Ejisu-Juaben district. A topographic map of scale 1: 25000 and road maps were acquired and used during the field work for navigation and collection of ground control points for geo-referencing, classification and assessment of classified map. Other data used in the research were secondary ground truth data of field points collected in 2007 in the study area by Benefoh (2008).

3.2.2 Software& Instrument

ENVI 4.7, ERDAS imagine, MultiSpec& R statistical softwares were used for image processing, image classification and accuracy assessment. GIS operations were undertaken in ArcGIS 10. The R statistical software is programming language software used for statistical analysis.

Also, IPAQ and Global Positioning system (GPS) instrucment was used for field navigation and collection of ground truth data. Garmin GPS was also used as a backup for collection of ground truth data. Digital camera was used for taking pictures of sample points.

3.3 Methods

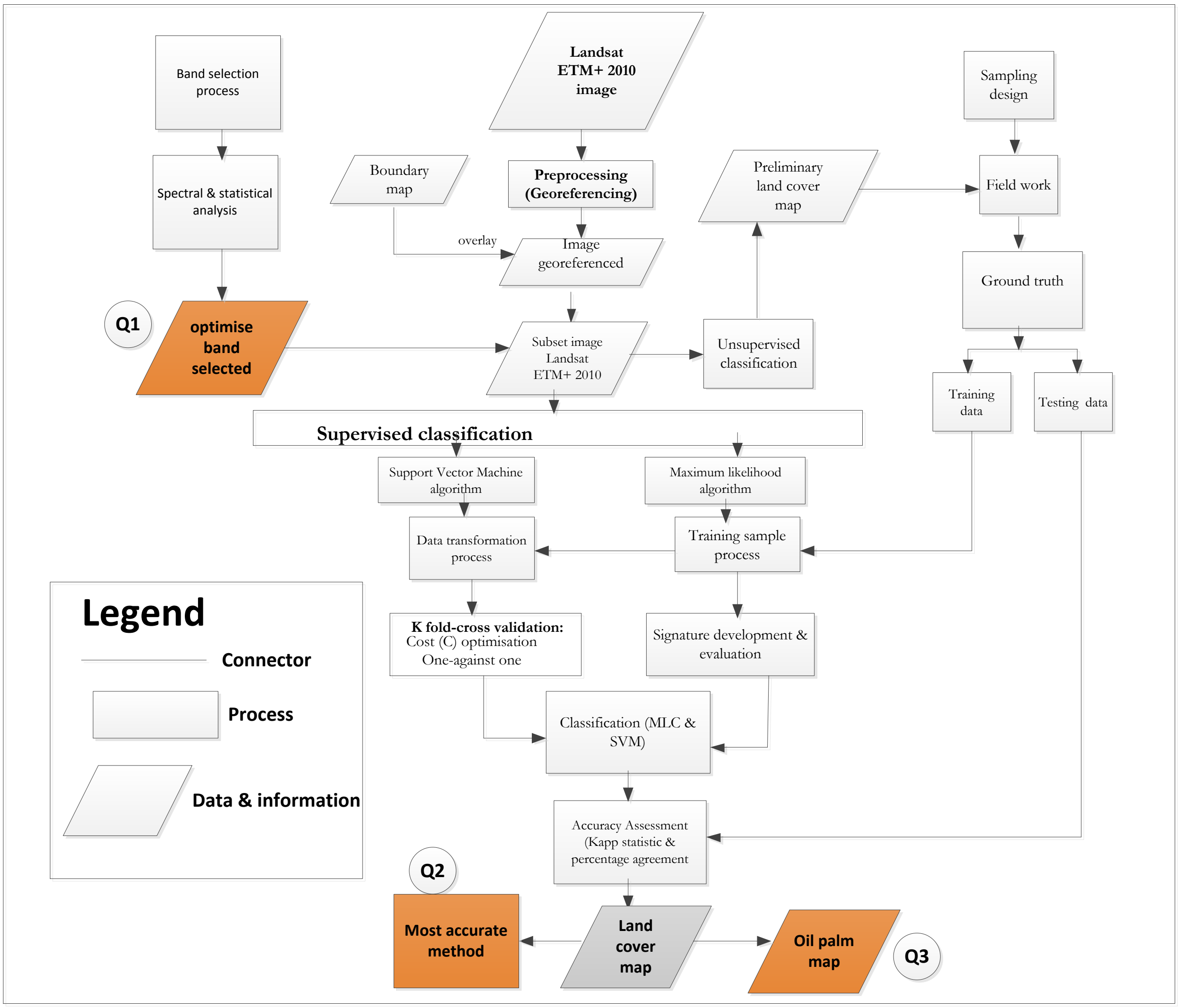


Figure 6: Methodology flow chart

3.3.1 Data pre-processing

The Landsat ETM+ 2010 image was transformed to conform to local coordinate that is system-Universal Transverse Mercator, and Legion datum map projection system. The image was geo-referenced with 25 ground control points of recognisable roads intersections in ERDAS IMAGINE 2010. A first order polynomial was used for geo-referencing and resulted in RMS error 0.27 less than 0.5 pixels. This result is considered as reasonable, because of the spatial resolution of Landsat ETM+ image used which is widely accepted in literature (Jensen, 1996). The geo-referencing was carried out to correct for geometric distortion due to Earth's rotation and other imaging conditions (Jensen, 1996).

An unsupervised classification was performed on the Landsat ETM+ image using Iterative self-Organising Data Analysis (ISODATA) classifier to produce preliminary land cover map (Khan *et al.*, 2010). The justification for adopting unsupervised classification here was due to the heterogeneity of different land use/land cover types in the area (Lillesand *et al.*, 2004). Land cover class names were chosen to match with definition used in the study area. So, the research analyst was responsible for merging and labeling spectral classes into meaningful classes. The class identification and validation was done using secondary field points collected by Benefoh (2008) before undertaking the field work. The unsupervised classification resulted in forty (40) spectral classes because relatively large clusters would be time consuming for cluster labeling and high computational demand (Lillesand *et al.*, 2004). This was grouped into five (5) major land cover types namely forest, agriculture, shrub, built-up areas and bare soil. The use of field points introduced aspects of supervised classification. A 3x3 majority filter was applied to smoothen out the "salt and pepper" appearance in the classified map (Lillesand *et al.*, 2004).

The preliminary land cover map was used with appropriate sampling design in the field to collect ground truth data. The study used stratified random sampling to collect field data. The reason is that it reduces variations within the strata and increases precision in strata (Hush *et al.*, 2003). A tool in ArcGIS 10 was used to generate random points on the classified map. The image was compressed into Enhanced Compressed Wavelength (ECW) and uploaded onto an Hp214 iPAQ for navigation during the field work.

3.3.2 Fieldwork

The field work was carried out from September to October 2011 using iPAQ GPS & Garmin 12, printed hard copy map, recording sheet & digital camera. An extra GPS device was also used as a backup to record coordinate points at the same time. This was to avoid sudden failure of device as well as confirming the values from the other GPS reading. The purpose of field work was to observe the study area and collect ground truth data for land cover mapping and accuracy assessment.

In the field, species dominance was helpful in assigning the sites to cover classes. A cover type was considered as forest, when trees crown cover is more than 10% of the ground, covering a land area of more than 0.5 ha and trees height of above 5m. A forest in the study area include both open and closed forest (FAO, 2000, 2005a). Built-up/bare cover was referred to cover of buildings, untarred or bare roads, soil, sand, or rocks surfaces.

Additionally, grass cover included all forms of grasses, ranging from creeping species up to tall elephant grass. Bush fallow included land which have been logged or farmed in the past and now left to recover with trees height less than 5m tall (Benefoh, 2008). The covers grass and bush fallow for the purpose of this study were later regrouped and called shrub. This grouping is supported by Cowardin *et al.*, (1979).

Furthermore, agricultural crop class referred to annual and cash crops such as Cocoa (*Theobroma cacao*), Citrus (*Citrus sinensis*), Cassava (*Manihof esculentus*), Oil palm (*Elaeis guineensis)*, Plantain/Banana (*Musa species*), and Maize (*Zea mays*) grown in the study area. Cassava, plantain/banana and maize are the main annual crops while cocoa, citrus, oil palm are cash crops which are grown in varying densities (FAO, 2005a). Mixed cropping is predominately practiced in the study area for especially annuals crops such as the following crops: Cassava, plantain, banana, maize. Cocoa farms are initially intercropped with plantain/banana crops for the purpose of providing shade to the young cocoa plants from intense sunlight but are later removed when cocoa reaches nearly full canopy cover. Citrus and oil palm farms are not intercropped for the purpose of avoiding nutrient competition with other agricultural crops and boosting yield. Oil palm data was recorded as a separate land cover type from agricultural crops and ground truth data were collected from both large holderholder plantations and smallholders plantings. Field data relating to oil palm ages were collected from oil palm fields in both smallholder and large holder plantation that were 5 years and above (i.e. 5, 8, 12 and 20 years old). These fields were accessible to the researcher. For smallholders, no data for oil palm fields less than 5 years were gathered within the scope of the image data used. So 5 years old or less oil palms field data were collected mainly in large plantations and observation showed that 5 or less years oil palm on the field visited do not have full canopy. This observation was important because oil palm forms full canopy cover between 5^{th} to 6^{th} years under favourable nutrient and soil condition. Full canopy minimises or eliminates background reflectance from soil or undergrowth (Wahid *et al.*, 2005).

During the same period, areas covered under raffia palm and oil palm farms with intercropping were noted. For areas predominantly under raffia palms found along water ways or streams and marshy areas, the point of location and geographical coordinate were taken. However, most of the points were inaccessible due to swampy condition. There, coordinates were estimated based on distance and compass direction.

Last but not least, field observations on the management practice in oil palm fields such as spacing (9m x 9m and growing of undergrowth) were made (see Pictures of field work in Appendices) and noted. The following considerations were taken into account during the field work; the picking of training areas near the boundary of land cover types was avoided. Extreme care was taken to prevent selection of training areas within and near road side. The reason here was to minimise mislabeling of land cover which may be caused by inherent error in the GPS. Field data were entered in MS-Excel and projected in Arc Map to show their distribution. A total of 343 field points were collected and shown in Table 1. The field data was grouped into training data and test data and subsequently the same field dataset used for class separation, land cover classification and accuracy assessment based on Bhattacharyya distance, SVM and MLC for comparison.

Table 1: Ground truth data collected in portion of Ejisu-Juaben district for training & testing of the Landsat ETM+ image

Land cover type	Training data	Testing data	Total
Forest	42	30	72
Oil palm	43	32	75
Crops	36	32	68
Shrub	32	28	60
Built-up/Bare	39	29	68

3.3.3 Bands selection procedure

To determine the measure of separability of oil palm class between the forest, shrub, other crops and builtup class occurring in the area, Bhattacharyya statistical distance measure (Rahman *et al.*, 2005) was used. A Bhattacharyya distance is a measure of statistical distance between signatures (reflectance) for all possible combinations based on the six non-thermal bands of Landsat ETM+ 2010 data (Zhang *et al.*, 2009).

In MultiSpec software, the Landsat ETM+ 2010 data and the training data were imported into the software to assess which band and band combination provided best interclass separation using Bhattacharyya distance between class pairs (Huttich *et al.*, 2009). A weighted factor value of 10 was assigned and 20 iterations were performed for all possible band combination. The optimum three band combinations were determined for bands with maximum interclass separation distance. The spectral separability is indicated for each value: good (**), fair (*) and poor ( ) (Jensen, 1996). The highest possible value is the maximum Bhattacharyya weighted interclass distance value that is considered to have good spectral separability (Zhang *et al.*, 2009). The second highest value is considered fair and the rest of the values that followed were considered as poor.

General and summary statistics (mean, standard deviation, minimum and maximum) values were obtained and analysed for all six non-thermal bands of the image. The optimum three band of combination was used as input data for maximum likelihood and support vector machine classifications. Results were presented in the form of descriptive statistical analyses such as bar graphs and tables.

3.3.4 Maximum likelihood algorithm implementation

The goal of maximum likelihood classification in this part is to produce a land cover and oil palm map. During the training process, the sample points were used to derive the signatures. The sample points were selected using random sampling in R statistical software. Randomisation was adopted to minimise the effect of spatial auto-correlation (Campbell, 2002). The sample points selected for the signature derivation process was performed only for the training samples. The minimum number of pixels required to derived a signature is the number of bands plus 1 (N+1) in ERDAS IMAGINE. This was done in order to estimate the mean vector and covariance matrix for an N-dimensional normal distribution, which is a necessary condition for the matrix to be positive.

Signatures were evaluated by examining the signature alarm and signature mean plot in ERDAS IMAGINE version 10. The best three band combinations of Landsat ETM+ data obtained from bands separation analysis were subsequently used as input to this algorithm.

The signature mean plot is a plot of the mean values of pixels comprising the area of interest (AOI) and the input bands. The signature alarm used its own pattern recognition ability, making it possible to highlight the estimated pixels in the viewer for the signature that belong to the specific class. The signature alarm was considered suitable for evaluation of the collected samples.

The conventional accuracy assessment procedure and presentation using error matrix (Congalton, 1991) was implemented. Two widely used accuracy measures; the overall accuracies and the kappa coefficient were used in this study (Congalton, 1991; Huang *et al.*, 2007). The overall accuracy has the advantage of being directly interpretable as the proportion of pixels classified correctly (Jensen, 1996) while the kappa coefficient allows for a statistical test of significance of the difference between two algorithms (Congalton, 1991).

3.3.5 Support Vector Machine implementation

The study implemented the support vector machine (SVM) on R statistical software (version R 2.13.2) available at http://www.r-project.org. The following packages were downloaded and installed: ***MASS***, ***mvtnorm***, ***kernlab***, ***gstat***, ***lattice***, ***rgdal***, ***GEOmap*** and ***mgcv.*** The packages are available on CRAN mirror in R software at http://cran.r-project.org/web/packages/e1071 for multiclass SVM. Description of packages used can be found in Appendix 8.4.

The best three band combination from Landsat ETM+ 2010 were used as image input data. Here, the training data were projected from input into two dimensional feature spaces in R programming language. The detailed procedure implemented for the support vector machine classification is outlined in Figure 7.

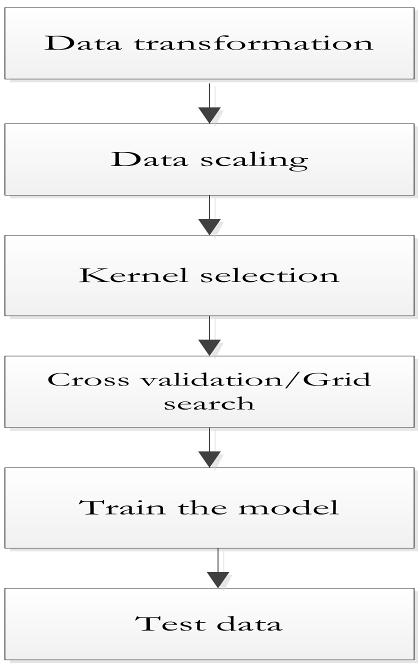


Figure 7: Procedure used in support vector machine classification of Landsat ETM+ 2010 image

Data transformation and scaling

To prepare the Landsat ETM+2010 data for classification, the image data was converted into TIFF/GeoTIFF format using ENVI 4.7 and used as input data. The training and testing data were also transformed and saved as ASCII text separately in ENVI 4.7. The training and testing data were named as separate classes; forest, oil palm, crops, shrub and built-up/bare and subsequently imported with its appropriate coordinate system into R software for data scaling. The image and the field data were transformed into format that is readable in R software as shown in Figure 8.

Following work done by Hsu *et al* (2010), Foody and Mathur (2004a), the data was rescaled from the input space to make a floating point number in the range [1,-1]. The data scaling was performed using a default function *svm* ( ) scales in package ***e1071*** available on CRAN mirror. The reason for conducting data scaling process was to improve classification results. The training and testing data were scaled using same scaling factors. The justification to compute the data scaling using the same scaling factors was to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another reason is to avoid numerical difficulties as kernel values usually depend on the inner product of the feature vectors (Hsu *et al.*, 2010).

Kernel selection

The next stage was to select the appropriate kernel function types. The kernel function seeks optimal way of choosing suitable parameter values for classification (Kavzoglu, 2009; Kavzoglu & Colkesen, 2009). Radial basis function (RBF) was chosen for the support vector machine. The parameters that need to be determined were the kernel-specific parameter (γ) as shown in Equation (11) and the cost parameter or penalty term (C) as shown in Equation (9). The aim here was to identify best parameter pair (C, γ) to train the image and subsequently classify the image. The basic reason for adopting Radial basis function (RBF) was that it has fewer numerical difficulties (Hsu *et al.*, 2010). Various parameter pair (C, γ) values were randomly selected. The parameter pair (C, γ) values were tried, and the performance was re-evaluated until all the chosen parameter pair (C, γ) has been evaluated.

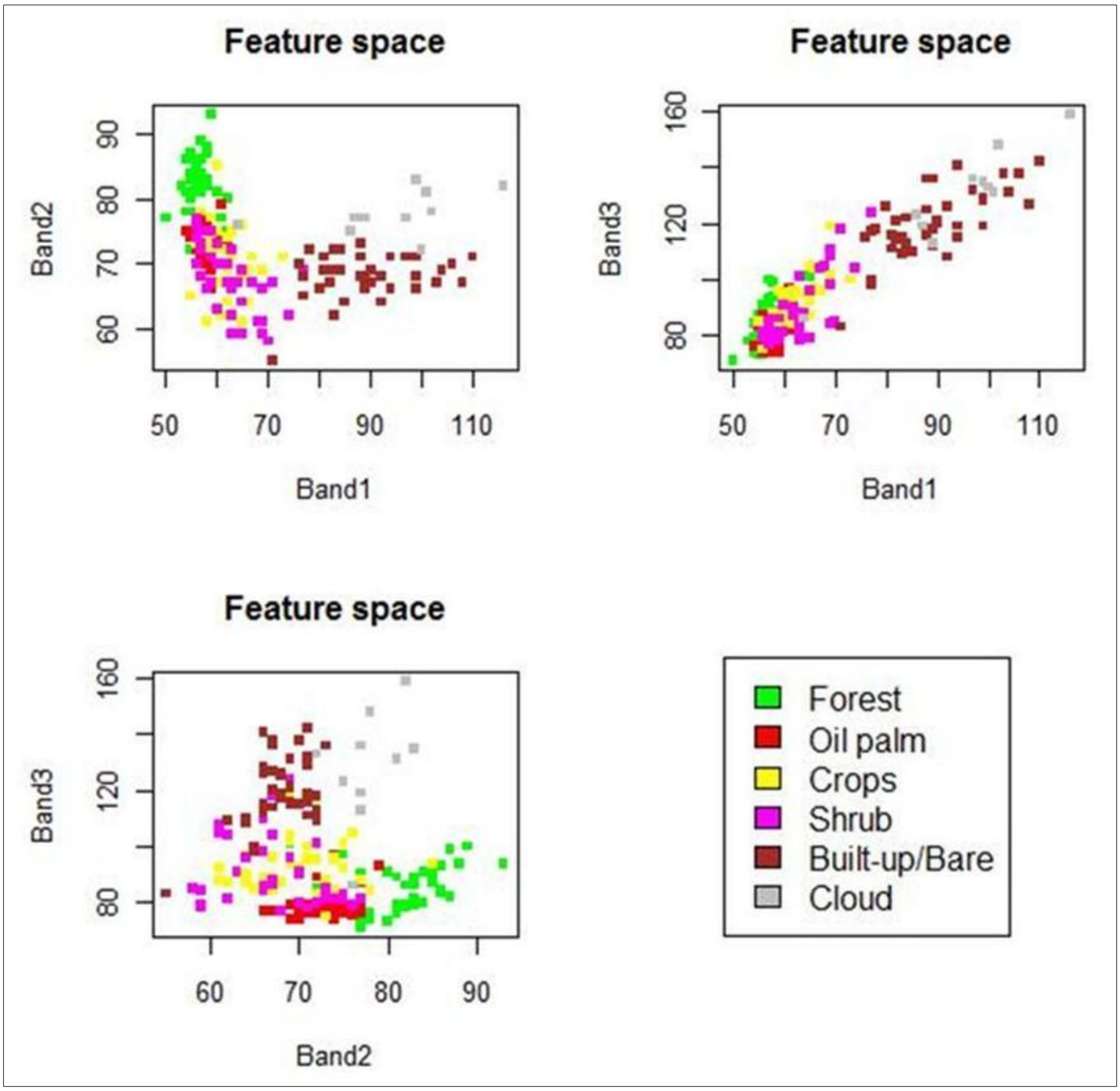


Figure 8: Distribution of training data set in dimensional feature space

Grid search and Cross validation

In this instance, a grid search algorithm was implemented to estimate cost parameter (*C)* and kernel specific function (γ) parameter values using a 10 fold cross validation. The grid search algorithm simply searched within the feature space for the best parameter pair through a *k* fold cross validation process (Foody *et al.*, 2006). The justifications for adopting cross validation was because conducting a complete grid search may be time consuming due to its high computational demand (Hsu *et al.*, 2010) and prevents over-fitting of the model (Foody & Mathur, 2004a).

For cost parameter, the range was fixed at 1 which is the default value. The reasons were that, if the cost parameter (*C*) was large, a high penalty cost for nonseparable points may store many support vectors and cause over-fitting.

Similarly if the cost parameter (*C*)was too small, the model may lead to under-fitting (Foody *et al.*, 2006). The kernel function (γ) parameters were selected from the range of 1-10. The performance was re-evaluated for all the chosen parameters.The range value of γ parameter was plotted against its corresponding accuracy assessment for testing data after cross validation (Figure 9). It was observed that for RBF kernels, the sigma (γ) parameter increases the accuracy from 1.0-3.0. Little or no trend of improvement was observed when the kernel function increased beyond 7.0.

Training the model

Generally, the cost parameter (*C*) fixed at 1 and kernel function (γ) parameter selected from the range of 1-10 produced between 150-184 support vectors through 10-fold cross validation. The best parameter pair (*C*, γ) was obtained at $C = 1$ and $\gamma = 3$. This parameter pair (*C*, γ) produced 152 support vectors yielding a misclassification error of 20.67% (Figure 9). This parameter pair (*C*, γ) generated the highest classification accuracy, therefore was selected and then applied for training the whole training data again in order to produce the final classification. The testing data was used to validate the classification from the support vector machine.

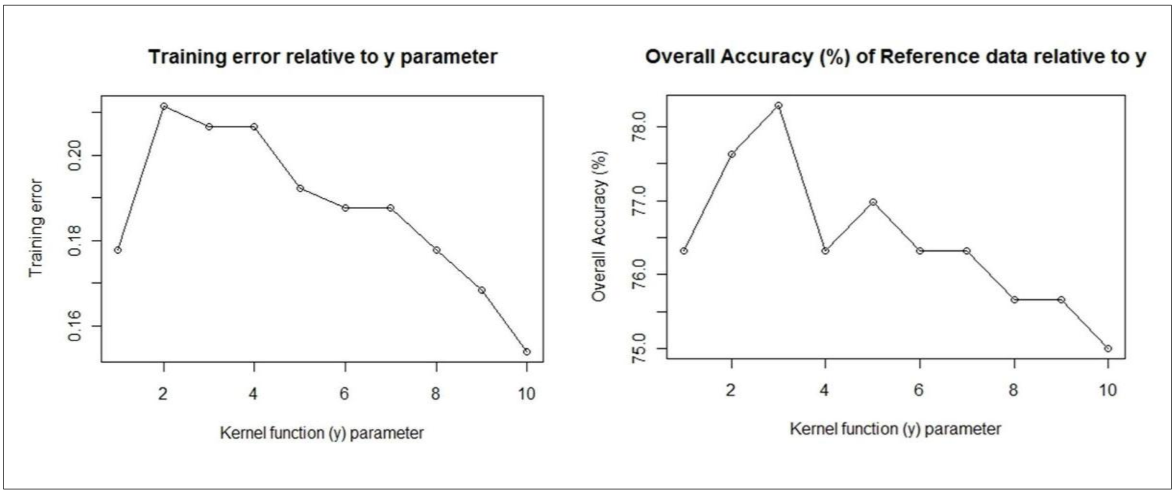


Figure 9: Relationship of training error and overall accuracy with kernel function parameter

Testing the model (Accuracy assessment)

The overall accuracy assessment and kappa statistics were derived from the contingency table or error matrix. Producer accuracy was determined by dividing the total number of correctly classified pixels for a class by the total number of reference data for that particular class. The user accuracy was also obtained by dividing the number of correct accurate sites by the total number accurate assessment sites that were classified in that group. The overall accuracy assessment is the sum of the number of samples correctly labeled for each class in the test set divided by the total number of samples in the test (Huang *et al.*, 2002; Zhu & Blumberg, 2002).

From the perspective of allocations associated with oil palm class in the confusion matrix, statistical significance of difference between classifications accuracies was evaluated using chi-square statistics and confidence limit. A chi-square was carried out to test the strength between the correctly and wrongly classified data in the confusion matrix table. The test for significance difference between misclassified proportions was based on the ***null hypothesis*** that **no significant difference exist between misclassified proportions among the classes**. The ***alternative hypothesis*** that **misclassified proportion differs between classes**. Test of significance of correlation coefficient was based on n-1 degrees of freedom and the level of confidence set at 0.05. The letter 'n' denotes the number of classes used in the study. If the calculated chi-square value is equal to or greater than the table value, reject the null hypothesis and accept the alternative hypothesis. The chi square formula used is shown below:

$$X^2 = \frac{(observed\ frequency - expeccted\ frequency)^2}{expected\ frequency}$$ ……………………………..…………Equation 12

Where observed frequency refers to the classified totals found in the confusion matrix and the expected frequency refers to the correctly classified points. The degree of freedom is 4 (i.e. 5-1) and P <0.05.

From the confusion matrix, the confidence interval (CI) was determined for the accuracies based on correctly classified sites and the reference and classified sample size. The formula for estimating the confidence interval is as follows:

The estimate for one proportion at 95% confidence level:

$$CI\ @\ 95\% = p \mp Z * \sigma_p$$ ………………………………………………………..Equation 13

Where CI is the confidence interval, p is the proportion in the sample, z depends on the level of confidence desired (which is 1.96 at 95%), andσ, the standard error of a proportion is equal to:

$$[p] = \frac{x+2}{N+4}$$ ……………………………………………………………………………Equation 14

The term$'x'$ is number of correctly classified sample sites and $'N'$ is the total number classified samples in the confusion matrix.

$$[\sigma_p] = \sqrt{\frac{p(1-p)}{N+4}}$$ ……………………………………………………………………Equation 15

4. RESULTS

4.1 Spectral separability assessment

The spectral mean values for the Bhattacharyya distance test are presented in Figure 10. The individual bands with the maximum Bhattacharyya mean distance as shown in Figure 10 indicates that bands 2 & 5 and band 4 &7 show similar spectral mean values. Band 3 show the maximum Bhattacharyya statistical spectral mean distance of 8.55.

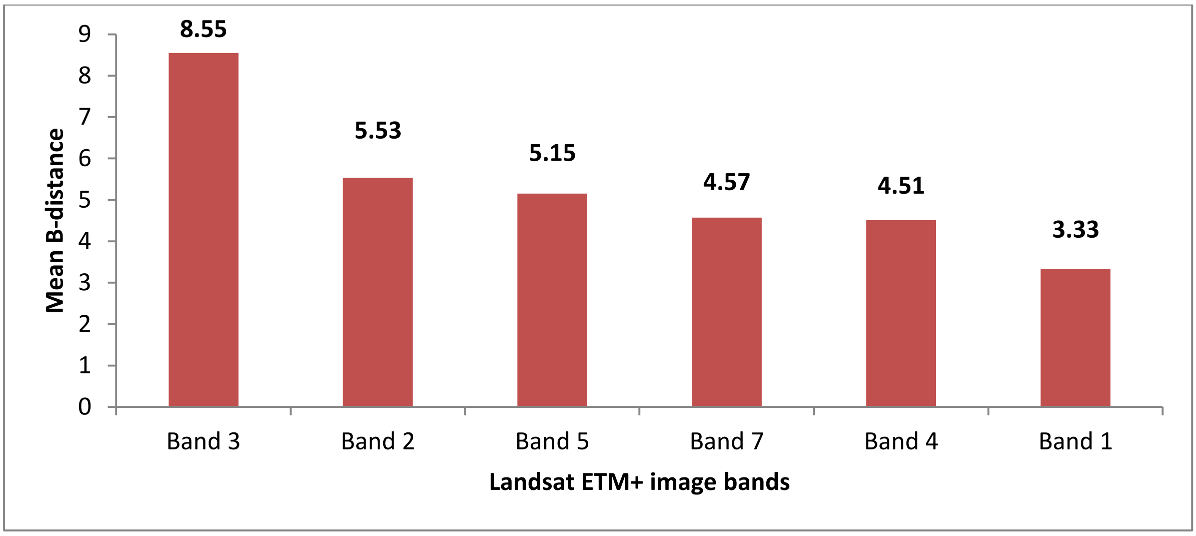


Figure 10: Bhattacharyya statistical mean distance measure for six non-thermal Landsat EMT+ bands

The result presented in Table 2 show the various Bhattacharyya weighted interclass distance for each individual band at separating land cover class pair. The spectral separability was assessed on a per-band basis and the results highlight largely on oil palm cover pair with the other classes.

The results suggest that band 3 (**15.9**) showed good spectral separability. This was established between oil palm versus other crops. This was followed by band 5 (**11.6**) with fair separability. Band7 (**6.7**), band 2 (**6.08**), band 1 (**3.26**) and band 4 (**2.81**) established poor separability and this was observed between oil palm versus other crops. Band 3 showed maximum Bhattacharyya weighted interclass distance for separating oil palm versus built-up/bare (**24.2**), followed by band 5 (**16**) with fair spectral separability. The rest of the bands provided poor separability for oil palm versus builtup/bare class.

Also, band4 (**7.6**) showed good spectral separability for oil palm versus forest followed by band 3 (**1.95**) with fair separation. The following bands; band 1 (**0.61**), band 5 (**0.38**), band 2 (**0.38**) and band 7 (**0.05**) observed poor separation according to Bhattacharyya weighted interclass distance. For separating oil palm versus shrub, bands 4 (**1.24**) showed good separation followed by band 3 (**1.12**), with fair separation. The following; band 7 (**1.09**), band 5 (**0.84**), band 2 (**0.64**) and band 1 observed poor separation for separating between oil palm versus shrub.

Generally, the results in Table 2 show that band 3 offered best spectral separability between oil palm versus other crops, oil palm versus builtup/bare (**24.2**). Band 4 offered the best spectral separation between oil palm versus forest (**7.6**) and oil palm versus shrub (**1.24**). Band 5 offered fair spectral separability for separating oil palm from the following cover types; other crops, builtup/bare and shrub.

Table 2: Bhattacharyya statistical distance measure for six non-thermal Landsat ETM+ 2010 bands and class pair.

	Weighted Interclass distance Measures (using weight factor of 10)									
Individual Band	F-O	F-C	F-S	F-B	O-C	O-S	O-B	C-S	C-B	S-B
Band 3	**1.95***	10	0.62	19.5**	**15.9****	**1.12***	**24.2****	1.36*	4.5*	6.18*
Band 2	0.38	5.36	0.51	15.2*	6.08	0.64	7.86	1.62**	2.36	7.02**
Band 5	0.41	17.30**	1.3**	10.0	**11.6***	0.85	**16***	0.71	0.39	0.96
Band 7	0.05	6.31	1.04*	12.5	6.7	1.09	12.9	0.48	1.97	2.51
Band 4	**7.6****	13.8*	0.76	1.5	2.81	**1.24****	8.5	0.92	7.43**	0.59
Band 1	0.61	1.68	0.03	7.64	3.26	0.48	10.9	1.24	1.41	6.04

Note: F stands for Forest, O-Oil palm, S-Shrub, C-Other crops, B-Built-up/Bare. Superscript ** indicates good separability and superscript * indicates fair separability.

Table 3 show Bhattacharyya statistical weighted interclass separation distances for the best combination of three bands. Band of combination 3, 4, 5 showed good spectral separability between oil palm versus other crops and oil palm versus builtup/bare class. Spectral separability between oil palm versus shrub and oil palm versus forest was fairly established. Poor spectral separation between other crops versus shrubs class was shown for this band combination. The poor separability between other crops versus shrub class in Table 3 suggests spectral overlap.

Table 3: Bhattacharyya statistical distance measure between 3 band pair of Landsat ETM+ 2010 and class pair

	Weighted Interclass distance Measures (using weight factor of 10)									
Bands	F-O	F-C	F-S	F-B	O-C	O-S	O-B	C-S	C-B	S-B
3,4,5	9.2	27.7	17.3	43.4	**24.1****	**2.2***	**54.0****	9.4	26.1	13.3
3,4,7,	9.1	25.1	13.5	36.1	21.4	**2.4****	47.6	11.1	23.5	11.8
2,3,4	**11.7****	27.8	4.9	37.1	19.6	2.1	**51.3***	9.6	22.7	8.8
1,3,4	**9.4***	25.1	4.5	33.1	**22.4***	2.0	46.7	10.0	25.1	8.8

Note: F stands for Forest, O-Oil palm, S-Shrub, C-Other crops, B-Builtup/Bare

The result for band of combination 3, 4 and 7 showed good spectral separability for separating between oil palm versus shrub and other agricultural crops versus shrub. The same band of combination showed fair spectral separability between oil palm versus other agricultural crops, forest versus shrubs, forest versus built-up/bare and shrub versus built-up/bare. Then poor separability was observed between oil palm versus forest, oil palm versus built-up/bare, forest versus other agricultural crops and other agricultural crops versus built-up/bare class.

Band of combination 2, 3 and 4 showed good spectral separability between oil palm versus forest and forest versus other agricultural crops. Fair spectral separability between oil palm versus other agricultural crops, oil palm versus built-up/bare, forest versus shrub, forest versus built-up/bare and shrub versus builtup were observed. However, poor separability between oil palm versus other agricultural crops, oil palm versus shrub, and oil palm versus builtup/bare class were also established for band of combination 2, 3 and 4.

In the same light, the results from band of combinations 1, 3 and 4 showed good spectral separability between oil palm versus forest and fair spectral separability between oil palm versus other agricultural crops, other agricultural crops versus shrub, other agricultural crops versus built-up/bare and shrub versus built-up/bare. Poor spectral separability was established between oil palm versus shrub, oil palm versus builtup/bare, forest versus other agricultural crops, forest versus shrub and forest versus built-up/bare respectively.

.

4.2 Accuracy assessment

Table 5 and Table 7 shows the accuracy assessment results of MLC & SVM respectively. The overall accuracy achieved for MLC was 71.71% with kappa coefficient of 0.65 (Table 5). From the kappa coefficient, it implied that 65% of the classification agreed with the reference data. The overall accuracy achieved SVM was 78.29% and kappa coefficient was 0.73 as presented in Table 6. From the kappa coefficient, it implied that 73% of the classification agreed with the reference data. The kappa coefficient value for MLC (0.65) was lower compared to the kappa value for SVM (0.73).

From Table 5 and Table 7, the results show that oil palm was mapped with higher user accuracies and producer accuracies than the overall classification accuracy. Also, the results show individual kappa statistics of above 80% in both classifiers for oil palm class. Figure 11 and Figure 12 shows the spatial distribution of the cover types in Ejisu-Juaben district based on MLC and SVM respectively.

The confusion matrix results for MLC and SVM are presented in Table 4 and Table 6 respectively. The results show that in both classifiers, the main confusion occurred between shrubs versus other crops; other crops versus forest. However, shrub versus other crops exhibited the lowest accuracy in relation to commission and omission error in both classifiers. For allocations associated with oil palm class, the only difference from the confusion matrix in Table 6 for SVM was that the commission error occurred with forest (1 case), shrub (2 cases) and other crops (3 cases). This translated into slightly lower user accuracy (84.21%) than user accuracy in MLC (85.71%). In SVM classification, oil palm and built-up/bare produced no error of omission and commission. This suggested that in the SVM classification, built-up was spectrally distinct from oil palm. This was reflected in producer accuracy of 100%. However, there was more confusion in MLC (Table 4) than in SVM (Table 6). The results of SVM were chosen over MLC for further statistical analysis because SVM performed better than MLC in terms of overall accuracy.

The chi- square statistical test of significance difference between misclassified proportions in Table 6 showed the calculated chi square (X^2) value of 3.51 was less than the tabulated value of 9.49 at 4 degrees of freedom and ($P < 0.05$). This result implied that there was no evidence to reject the null hypothesis. The decision was that the misclassified proportions did not vary between classes but rather the cover classes in Table 7 share the same overall accuracy of 78.29%.

The results from the confidence interval are presented in Table 9. The overall accuracy of SVM was within confidence interval of 0.71-0.84. This result showed that with repeated sampling, 95% of sampling would provide overall classification accuracies between 71% and 84%. In the case of oil palm cover, the producer and user accuracies had a standard of 0.04 and 0.06 respectively (Table 9). Producer accuracy of oil palm cover was within the confidence interval of 0.86-1.0. User accuracy of oil palm class was within the confidence interval of 0.68-0.92 as summarised in Table 9.

Table 4: Error matrix for MLC classification

Classification	Forest	Oil Palm	Other Crops	Shrub	Builtup/Bare	Total	Error of commission %
Forest	**25**	1	1	0	0	27	7.71
Oil palm	1	**30**	2	1	1	35	14.29
Other Crops	1	0	**15**	12	1	29	48.28
Shrub	3	1	10	**12**	1	27	55.56
Builtup/Bare	0	0	4	3	**27**	34	25.93
Total	**30**	**32**	**32**	**28**	**30**	**152**	
Error of omission, %	16.67	6.26	53.13	40.00	10.00		

Table 5: Accuracy assessment for MLC classification

Landcover class	Reference totals	Classified totals	Number correct	Producer accuracy, %	User accuracy, %	Kappa
Forest	30	27	25	83.33	92.59	0.91
Oil palm	32	35	30	93.75	85.71	0.82
Other crops	32	28	15	46.88	53.15	0.41
Shrub	28	26	12	42.86	46.15	0.34
Built-up/Bare	30	36	27	90.00	79.41	0.74
Totals	152	152	109			
Overall accuracies, % = 71.71%				**Overall Kappa Statistics = 0.65**		

Table 6: Error matrix for SVM classification

Classification	Forest	Oil palm	Other crops	Shrub	Built-up/Bare	Total	Error of commission %
Forest	**26**	0	1	0	0	27	3.70
Oil palm	1	**32**	3	2	0	38	15.79
Other crop	3	0	**27**	19	0	49	44.90
Shrub	0	0	0	**4**	0	4	0
Built-up/Bare	0	0	1	3	**30**	34	13.33
Total	**30**	**32**	**32**	**28**	**30**	**152**	
Error of omission, %	13.3	0	15.63	14.28	0		

Table 7: Accuracy assessment for SVM classification

Landcover class	Reference totals	Classified totals	Number correct	Producer accuracy, %	User accuracy, %
Forest	30	27	26	86.67	96.30
Oil palm	32	38	32	100	84.21
Other crops	32	49	27	84.38	55.10
Shrub	28	4	4	14.29	100
Built-up/Bare	30	34	30	100	88.24
Totals	152	152	119		
Overall accuracies, % = 78.29			**Overall Kappa Statistics = 0.73**		

Table 8: Observed versus expected values for chi square estimation

Class	Correct	Expected	Wrong	Expected	**Total**
Forest	26	*21.14*	1	*5.86*	**27**
Oil palm	32	*29.75*	6	*8.25*	**38**
Other crops	27	*38.36*	22	*10.64*	**49**
Shrubs	4	*3.13*	0	*0.84*	**4**
Builtup/bare	30	*26.62*	4	*7.38*	**34**
Total	**119**		**33**		**152**

Table 9: Estimated confidence interval for producer, user and overall accuracies (SVM)

Accuracy	Class cover	P	se	95% CI
Producer	Oil palm	0.94	0.04	0.86-1.0 (86%-100)
User	Oil palm	0.8	0.06	0.68-0.92 (68%-92%
Overall	All classes	0.78	0.03	0.71-0.84 (71%-84%)

Note: se = standard error, p = proportion in the sample, CI = confidence interval

.

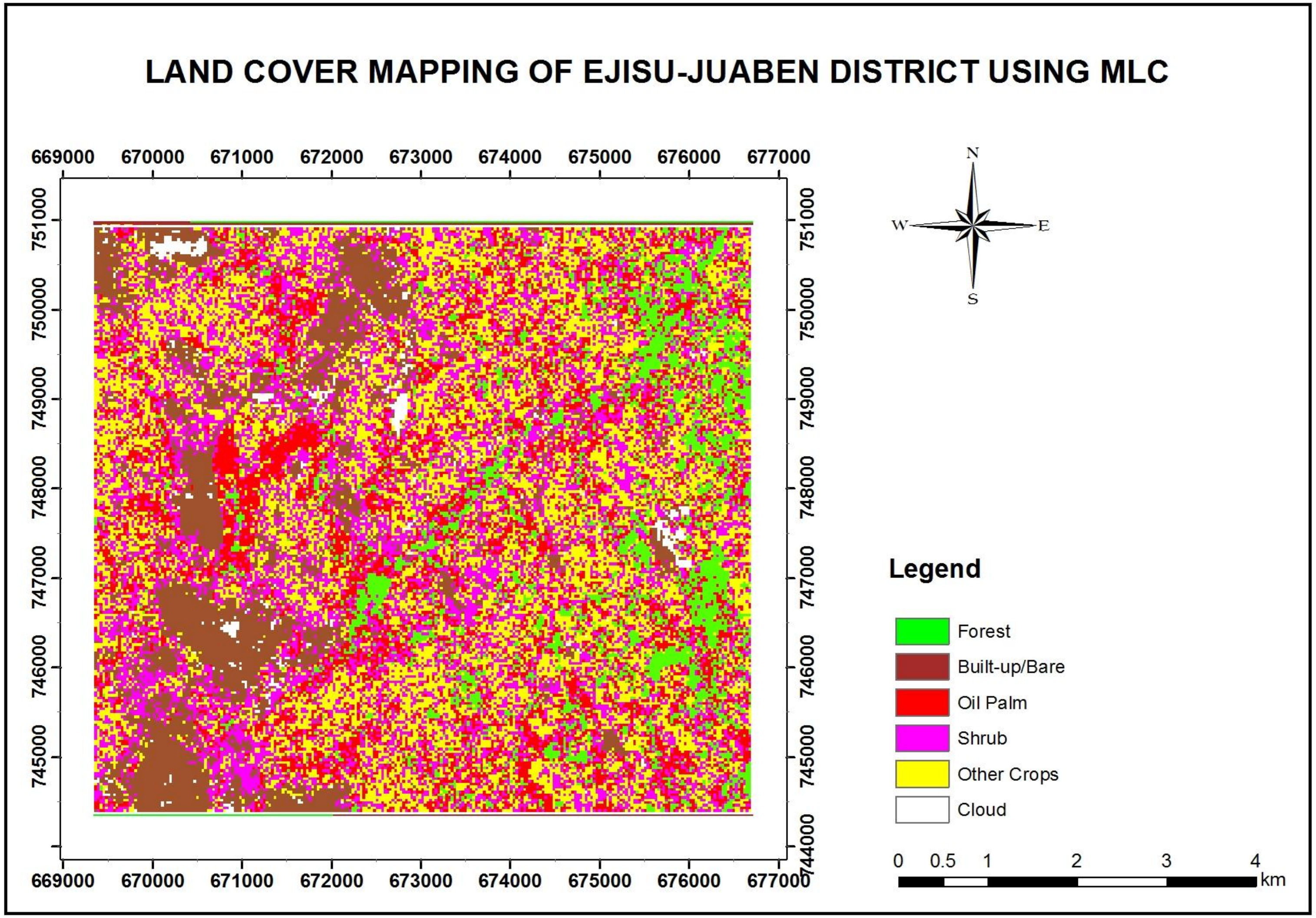


Figure 11: Maximum likelihood classified land cover map of Ejisu-Juaben district (2010)

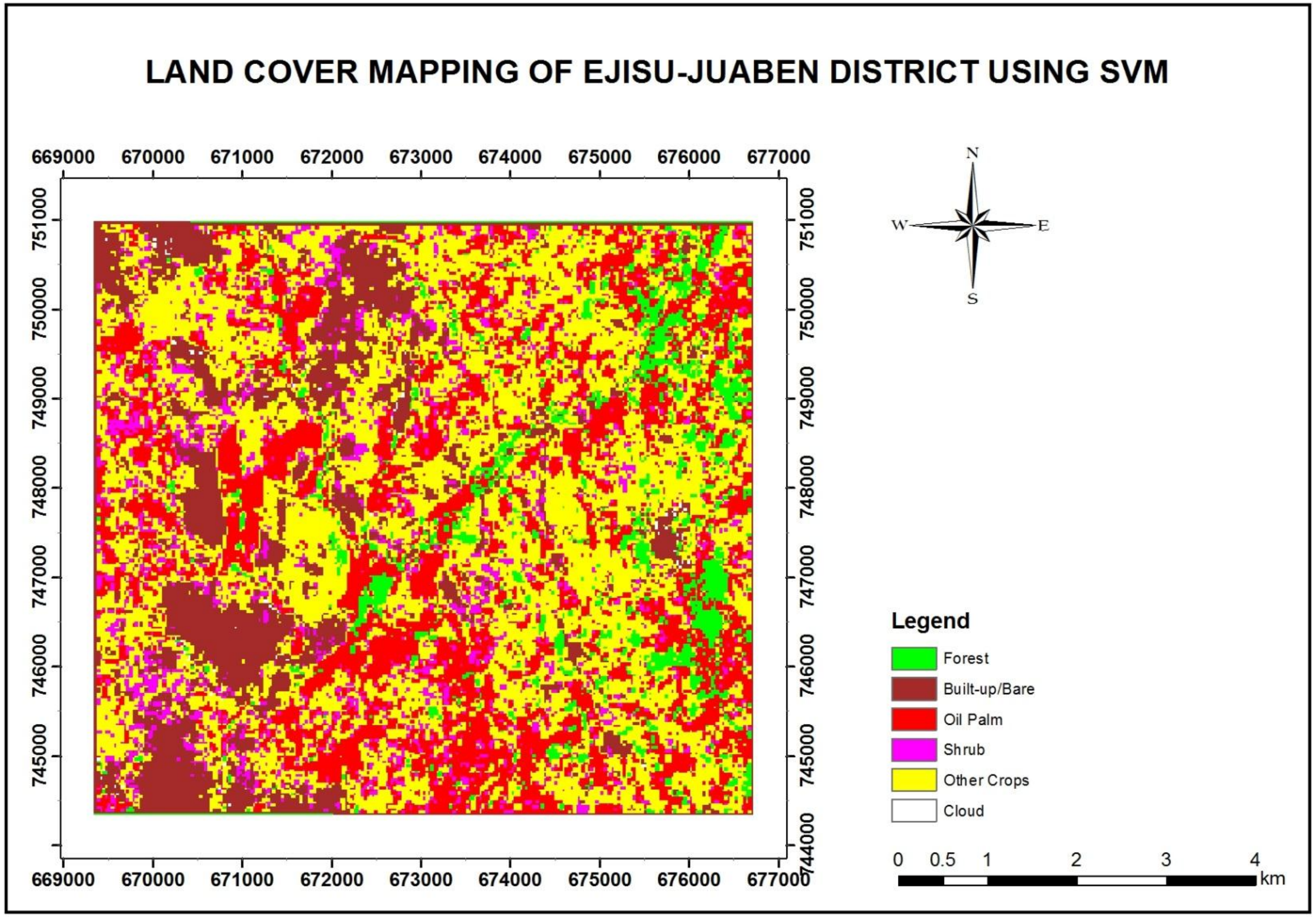


Figure 12: Support vector machine classified land cover map of Ejisu-Juaben district (2010)

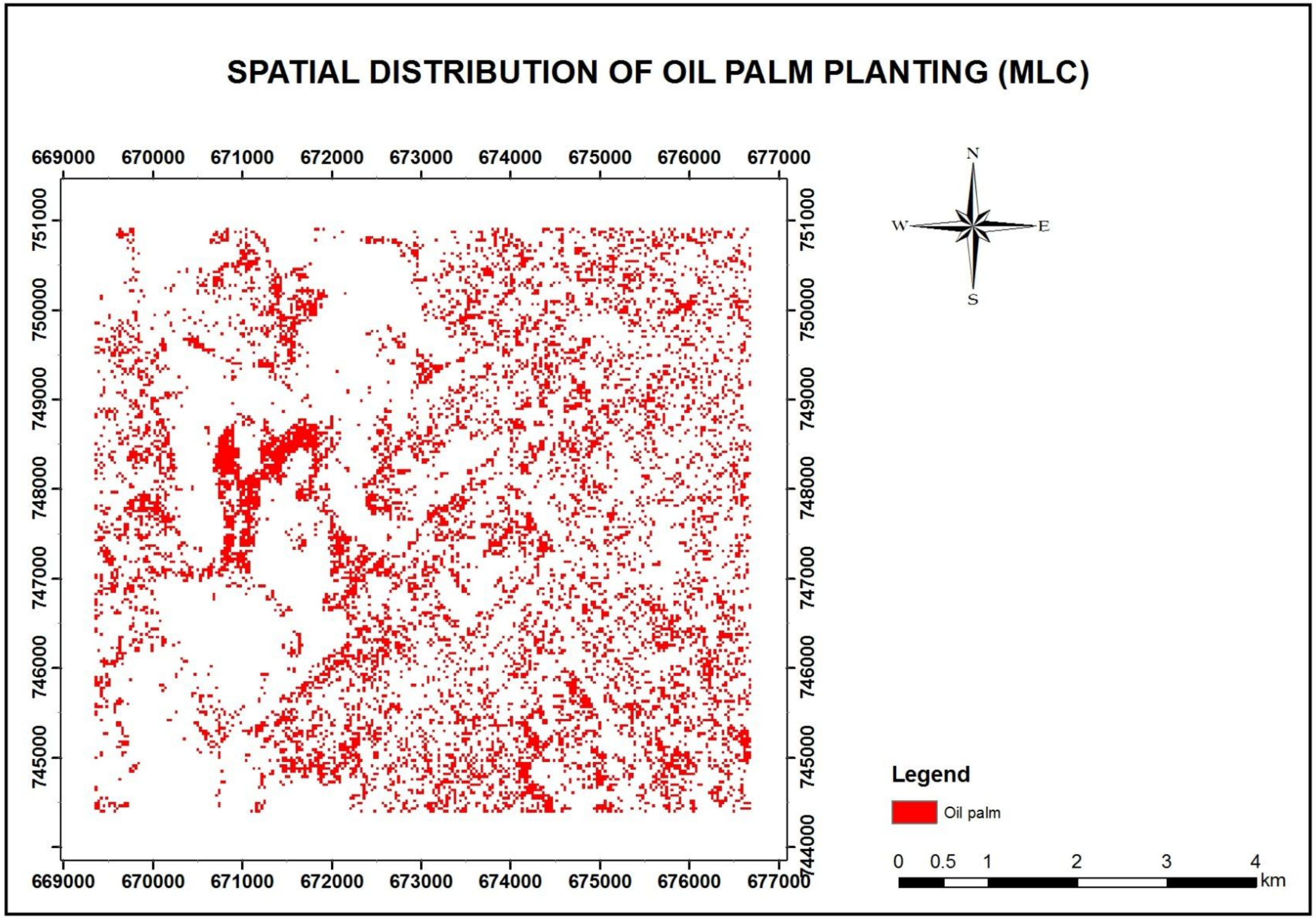


Figure 13: Spatial distribution of oil palm plantation in Ejisu-Juaben district (MLC)

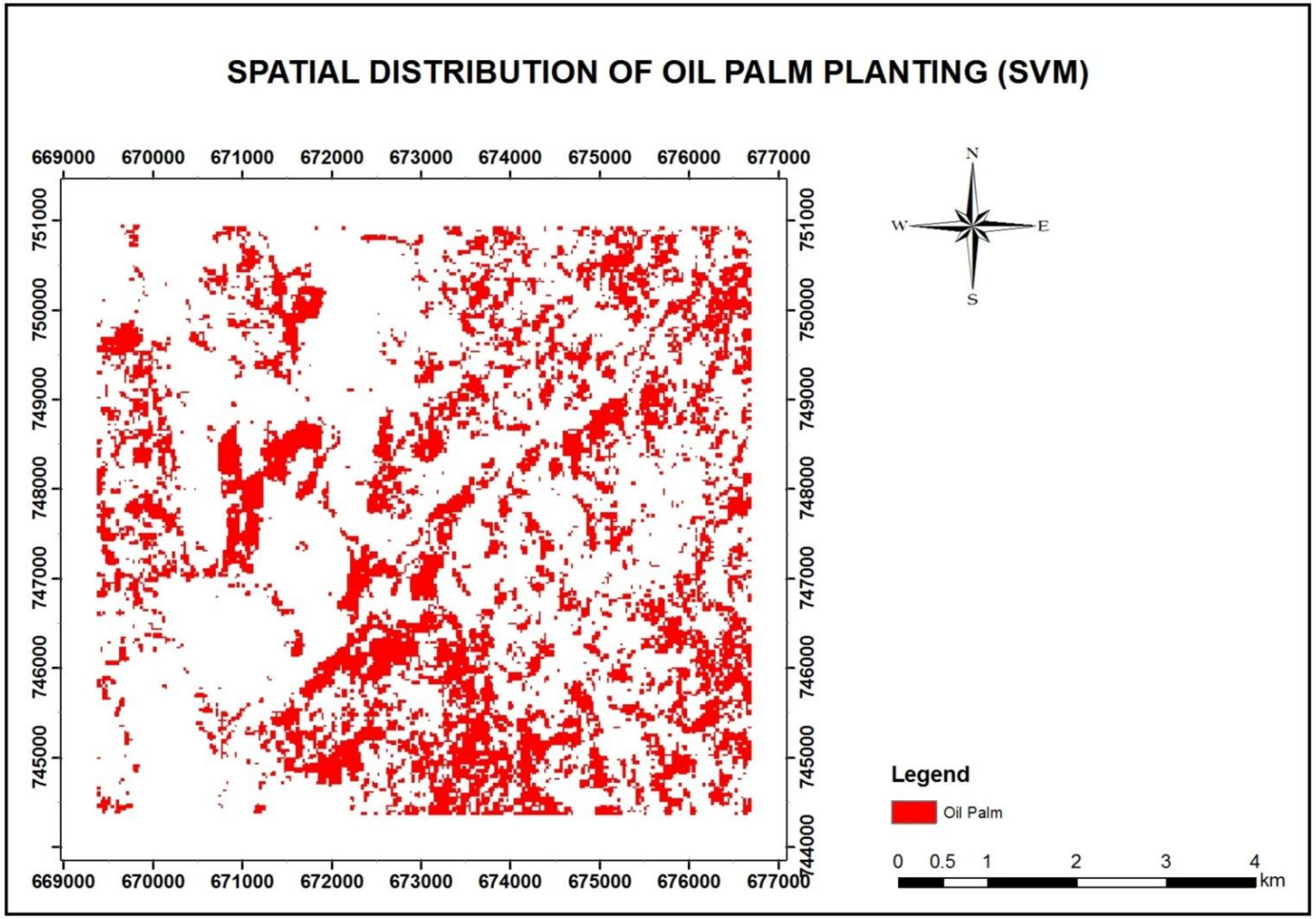


Figure 14: Spatial distribution of oil palm plantation in Ejisu-Juaben district (SVM)

4.3 Spatial distribution of land cover types

Figure 15 shows spatially, the area coverage for the various cover types for MLC and SVM. The major land cover was 'other crops' covering 1435.95 ha (29.35%) and 1461.71 ha (30.74%) based on MLC and SVM respectively. The large area covered by other crops in the study area may be linked to the extensive agricultural activities in the district and by extension labeled the 'food basket' according to FAO (2005a). Shrub recorded 1275.39 ha (26.08%) and 976.08 ha (20.22%) based on MLC and SVM.

Also, Oil palm covered 904.95 ha (18.50%) and 993.78 ha (20.54%) based on MLC and SVM. This indicates that the spatial coverage and distribution of oil palm do not vary much between both classifiers. The area estimated by both classifiers includes both areas under raffia palm and smallholder farms intercropped. This shows that oil palm and raffia palm are spectrally similar.

Forest cover was 465 ha (9.51%) and 703.47 (14.62%); built-up/bare covered 766.04 ha (15.66%) and 661.927 ha (25.32%); cloud cover on the image covered 43.69 ha (0.90%) and 28.89 ha (0.59%) based on MLC and SVM respectively.

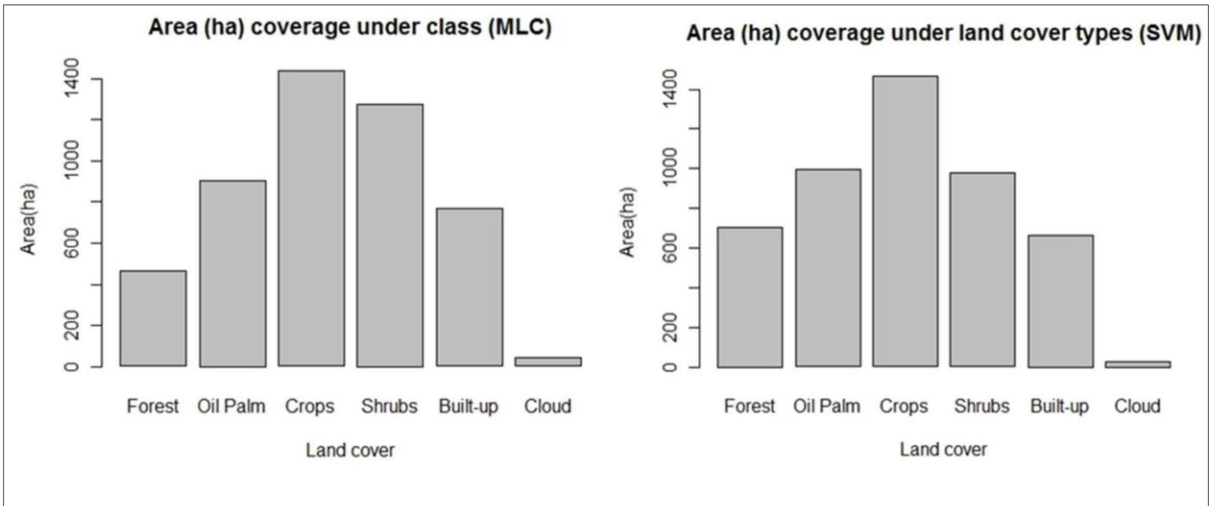


Figure 15: Estimated area covered by land cover types based on MLC & SVM

5. DISCUSSION

5.1 Spectral separability analysis

The results discussed here highlights on spectral separability established between oil palm versus the following cover types; forest, other agricultural crops, shrub and builtup/bare.

The presence of the visible wavelength (i.e. band 3, red) in the separability test of Bhattacharyya distance is remarkable (Table 2). The separation of oil palm between the following cover types; forest, shrubs, 'other crops' and built-up/bare revealed in the results highlights the usefulness of the visible spectrum for distinguishing between vegetation types. This was evident from the separability result where good separability was observed between oil palm versus other crops and oil palm versus builtup class and fair separability between oil palm versus forest. Strong chlorophyll absorption region and strong reflectance for vegetation, bare soils and urban areas in the wavelength region (0.63-0.69 Άμm) explains this result. Oil palm unique reflectance can be linked to the sunlight and minerals composition (i.e. iron, magnesium, nitrogen) purposely introduced in oil palm cultivation to boost yield which are necessary for chlorophyll production (Corley & Tinker, 2003). This result agrees with the results found by Gong *et al.* (1997), who found that the visible bands are essential for discrimination between vegetation types and bare soils.

The near-infrared band (band 4) is represented in the Bhattacharyya distance test with good spectral separability between oil palm versus forest and between oil palm versus shrub, however presented poor separability of oil palm versus 'other crops'. This is not a surprising incident, because the near infrared region is known for its information content in water and vegetation studies; however this cannot explain the poor separability between oil palm versus shrub. In the near-infrared band 4 region, presence of water absorbs well in this wavelength and vegetation has high reflectance due to moisture content.

Although, past studies (Dehnavi *et al.*, 2010; Ibrahim *et al.*, 2000) show that vegetation types reflect well in the two infrared wavelengths of Landsat ETM+ image (i.e. band 5 and band 7) because the strong water absorption and strong reflectance in the leaf provide subtle difference for vegetation mapping. However, the results found an unexpected poor spectral separability among all cover class pair especially between oil palm versus forest, oil palm versus shrub and oil palm versus other crops for band 7 (Table 2). Different reasons may cause this problem. For example, ETM+ acquisition date (February-dry season) (Benefoh, 2008; Thenkabail *et al.*, 2004) may produce difficulties due to the seasonal variations (Jensen, 1996).

Jensen (1996) and Munyati (2000) explain effect of seasonal variations on plants reflectance. This is because different moisture content may change cover classes reflectance and separation. Drier condition may cause spectral overlap among land cover classes due lower moisture content (Munyati, 2000).

The situation is different in band 5 where oil palm versus shrubs; oil palm versus other crops and oil palm versus forest show fair spectral separability. The reasons here may be different. The reasons may essentially be related to the wavelength regions of band 5 (Ibrahim *et al*., 2000). Band 5 operates within the spectral wavelength of (1.55- 1.75Άμm). This wavelength range is very sensitive to moisture and useful for monitoring vegetation and soil moisture. On the other hand, for band 7 (2.08- 2.35Άμm), its application is well known in soil and geology mapping; though can be applied in vegetation mapping (Dehnavi *et al*., 2010). Also, other factors such as the separability techniques, spatial resolution, and landscape under study may play a role. The study recommends the use of other separability techniques or wet season image to confirm the results or otherwise.

Results from three band combination in the separability test show that band 1 and 2 provide the minimum Bhattacharyya separation distance in separating all the class pair (Table 2). The poor separability exhibited by band 1 and 2 in the Bhattacharyya distance test may be explained by the spectral region in which these bands operate. Band 1 (0.45-0.52Άμm, blue-green) is a short wavelength that penetrate better than other bands but its choice of application is preferred in aquatic ecosystems to monitor sediment, water depth. Band 2 (0.52-0.62Άμm, green) is similar to band 1 but not as extreme as band 1. Therefore, applying these bands to spectrally separate between vegetation and builtup/bare using Bhattacharyya testis apparently and less effective as evidenced in the results shown in Table 2.

Spectral analysis of band combination 3, 2, 1 shows poor spectral separability. The result is not surprising as band 1 and band 2 has been established to be useful for aquatic studies despite the presence of band 3. Furthermore, band 4, 3, 2 combination show similar poor separability as bands 3, 2, 1 and the inclusion of a near infrared band (band 4) makes it useful for detecting land and water boundaries rather than vegetation mapping.

Bands 4, 5, 3 contains a near infrared red (band 4), mid infrared red (band 5) and red (band 3) which can be clearly defined for different vegetation types because they exhibit variations in moisture. Band 3, 4, 7 has a near infrared red (band 4), mid infrared red (band 7) and red (band 3) similar to band 4, 5, 3. However, band 7 typically due its wavelength range is used in soil and geological studies (Dehnavi *et al*., 2010). This evidence explains why band 7 show poor spectral separability for between oil palm versus other crops, forest and shrub in (Table 2 and Table 3).

Comparing the four different band of combination based on the Bhattacharyya statistical distance, the spectral overlap shown on Table 2 and Table 3 for band of combination 2, 3 and 4; band of combination 1, 3 and 4 as well as 3, 4, and 7 do not provide an optimum spectral separability among the land cover types. However, the combination of near infrared red (band 4), mid infrared red (band 5) and red (band 3) provide spectral details for vegetation and non-vegetation mapping. With this band of combination, discrimination between oil palm and others such as forest, shrub, other crops and builtup/bare is the optimum with regards to spectral separability. This result is supported in previous studies where other techniques confirmed the use of this band combination for oil palm mapping (Liberti *et al.*, 2009; Wahid *et al.*, 2005).

5.2 Mapping oil palm with SVM.

The spectral overlap among land cover classes is known to impact on overall accuracy through both error of omission and commission during accuracy assessment. For example, poor spectral separation between different cover types is linked to low classification results (Kusimi, 2008). Kusimi, (2008) linked lower overall accuracy of land cover classification in Wassa-Amenfi district of Ghana to poor spectral separability between closed canopy versus open canopy plantation and shrub versus open mine fields in Wassa-Amenfi district of Ghana.

Thus an overall accuracy of 78.29% from SVM classification relatively lower than the recommended 85% by Campbell (2002) and kappa statistics of over 70% was not surprising. This is expected because SVM is designed to locate an optimal hyperplane for separating between classes. This optimal separating hyperplane is applied to unseen training samples with least error. Support vectors used in the SVM classification are controlled by the kernel specific function parameter (Vapnik, 1995). Since the significance of the support vectors is intended to minimise confusion between classes (Huang *et al.*, 2002), its success comes at a cost when the class separability is poor. The cost parameter controls the penalty of wrongly placed pixels or support vectors that lie on the other side by using only pixels that belong to the correct class during training. So the kernel specific function (γ) parameter which takes care of minimising the training error (Foody *et al.*, 2006; Foody & Mathur, 2004a & 2004b) achieves that through fine tuning or cross validation until an optimum support vectors are obtained at reasonably minimum misclassification error (Figure 9). This effect explains the accuracy assessment results (Foody *et al.*, 2006; Hue *et al.*, 2010).

The overall accuracy result is in agreement with other published work. Zhu and Blumberg (2002) achieved an overall accuracy of 89.90% using SVM in mapping vegetation cover from non-vegetation cover with less omission of 9.60% for 30m resolution image classification in Beer Sheva, Israel. Also, in Punjab-India, Foody *et al.*, (2006) successfully classified cotton class with a user and producer accuracies of 95.65% and 97.77% using SVM. Similarly Marcal *et al.*, (2005) obtained overall accuracy of 72.20% by using SVM on ASTER images.

Oil palm can be mapped using SVM at high user and producer accuracies of 84.21% and 100% respectively. This indicates that the pixels representing the oil palm class was well allocated and that the classification provides a clear representation of the field situation. In SVM classification, oil palm and builtup/bare produced no error of omission and relatively lower commission due to oil palm being spectrally distinct from builtup/bare. The error of commission shown in Table 6 for oil palm class is explained by the misallocations of shrub (2 cases), forest (1 case) and other crops (3 cases). This evidence explains the slightly lower producer accuracy of SVM (84.21%) relative to MLC (85.71%).

The confidence interval for the overall accuracy for SVM classification provided a range of between 71% and 84% suggesting that even when the sampling is repeated many times, there would be no significant impact on SVM classification accuracies and that the overall accuracy will be estimated within this confidence interval range (Rossiter, 2004). This evidence is revealed in the chi-statistics where the misclassified proportion showed lack of significance. This means that all the land cover classes used in the study shared the same overall accuracy.

The results further highlights on the value of SVM approach to classification when classes of interest (i.e. oil palm) are targeted. The results suggest that for oil palm class, the producer accuracy would range between the confidence interval of 86% and 100% when the sampling is repeated many times. The same applied to user accuracy with range of between 68% and 92%. This result is in agreement with past studies where Foody *et al*., (2004) and Foody *et al*., (2006) compared the level of correctly classified and wrongly classified based upon the standardised normal test statistic (McNemar test).

Thus, it can be concluded that the proportions occupied by each cover class, especially oil palm in the classified map do not significantly vary from the actual coverage in the field. This implies that SVM can be relied upon to identify and map the spatial distribution of land cover especially for oil palm cover in a heterogeneous landscape such as the one in Ejisu-juaben district although the overall accuracy was lower than the recommended level of 85%. The result show the importance of training data used in classification since a poor quality training samples can lead to the hyperplane been fitted inappropriately resulting in a low classification accuracy. Poor quality training data could pose a shortcoming to using this approach. The sampling technique used to collect the training data may explain the classification accuracy obtained. Further studies should consider other sampling techniques since Foody *et al*., (2006) employed intelligence selection of training samples to improve classification accuracies.

The SVM approach used can provide reliable information in assessing the link between oil palm expansion and deforestation (Koh, 2010), fauna and flora biodiversity loss (Koh, 2007, 2008) that has gained prominence recently. SVM approach can also be useful in identifying and mapping land cover change especially comparing images taken at two different periods for change detection studies. Additionally, SVM approach has demonstrated its potential to identify and map the distribution of oil palm in a heterogeneous landscape; thus the approach can be applied to principles 5, 6 and criteria 7.1 and 7.2 of RSPO to address issues of environmental impact assessment.

However, its application to estimate the actual area under oil palm comes with some level of uncertainty. The area covered by oil palm plantings in SVM classification is 993.78 ha. For example, the study established that the estimated area under oil palm include large holder plantations, smallholder plantings and areas occupied by wild raffia palm along the water ways and marshy areas, which were equally classified as oil palm (*Elaeis guineensis*).

This indicates that wild raffia palm and oil palm (*Elaeis guineensis*) are spectrally similar (Foody *et al.*, 2006; McMorrow, 1995). This can pose limitation to accurately use of SVM in estimating area under oil palm and this should be assessed in further studies.

Also, this calls for the need to include on-site assessment and use of ancillary data such as updated topographic map to augment the SVM approach. Figure 14 shows the spatial distribution of oil palm in the study area. Thus, the SVM approach needs refinement specifically to target discriminating of wild raffia palm, different age stand of oil palms, which may be a valuable approach for initial assessment for RSPO certification of oil palm especially criteria 7.3 of RSPO document, which states that new oil palm plantings are not expected to replace rain forest and high conservation areas.

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The general objective of the study was to map oil palm in Ejisu-Juaben district using support vector machine algorithm with Landsat ETM+ data. The study has demonstrated the use of remote sensing and GIS information for mapping oil palm related land cover in the district. The following are the study conclusions.

Objective 1 To evaluate the spectral separability of oil palm from other cover types

- Which spectral bands of Landsat 2010 image provide best oil palm separability from other cover types?
  - The best three band combination from Landsat EMT+ 2010 that can separate between oil palm from forest, other crops, shrubs and builtup/bare using Bhattacharyya distance are the near infrared (band 4), mid infrared (band 5) and red (band 3)

Objective 2: To assess the performance of the support vector machine and maximum likelihood in oil palm mapping using overall accuracies and Kappa statistics procedures

- What is the level of classification accuracy attained by using support vector machine (SVM) and maximum likelihood classification (MLC)?
  - The overall accuracy recorded for oil palm mapping of Landsat ETM+ 2010 was 71.71% for (MLC). The overall kappa statistics was 0.65 and is considered moderate agreement. Similarly, the overall accuracy of Landsat ETM+ 2010 using SVM was 78.21%. The overall kappa statistics was 0.73 and is interpreted as good agreement.
  - SVM performs better in terms of overall accuracies and kappa statistics than MLC. The accuracy assessment result for SVM is generally in agreement with past studies in which the SVM was found to be more accurate than maximum likelihood.

Objective 3: To estimate the spatial distribution of oil palm in the study area

- How well does the two classification algorithm estimate the spatial distribution of oil palm in the study area?
- The area estimation under oil palm using MLC was 904.95 ha and 993.78 ha for SVM respectively. The area estimated by MLC was lower than the area estimated by support vector machine. An area difference of 88.90 ha could be attributed to the principle by which both algorithms classified the data. However, SVM estimated accurately than MLC with overall accuracy of 78.21%.

Limitations

The non-availability of current (2011) cloud free LANDSAT ETM+ data limited the study assessment of land cover types accuracy to reflect the real cover on ground.

6.2 Recommendations

The study recommends the following:

- SVM and MLC varied in their ability to map and quantify oil palm. SVM classifier maps oil palm with a high user and producer accuracies. Therefore academic and state institution such as universities, forestry commission, and RSPO-Ghana office should consider using support vector machine as an alternate for oil palm related land cover mapping. However, the SVM classifier needs refinements.
- Studies should explore the use of other kernels types available and other remote sensing images to enhance the classification accuracy of SVM.
- The study recommends that other improved and robust techniques of training data sampling be used to improve classification. Finally, other advanced pixel based algorithms such as neural network and decision trees should be explored and then compared with support vector machine.
- Further studies to be carried out during both the dry and wet season to check the effect of spectral separability.

7. LIST OF REFERENCES

Benefoh, T. D. (2008). *Assessing the effects of land-use/cover change on ecosystem services in Ejisu-Jauben Districts, Ghana.* Unpublished MSc, University of Twente, Enschede.

Bruzzone, L., & Serpico, S. B. (2000). A technique for feature selection in multiclass problems *International Journal of Remote Sensing, 21*(3), 549-563.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and Knowlege Discovery, 2*, 121-167.

Butler, R. A., Koh, L. P., & Ghazoul, J. (2009). REDD in the red: palm oil could undermine carbon payment schemes. *Conservation Letters, 2*(2), 67-73.

Butler, R. A., & Laurance, W. L. (2009). Is oil palm the next emerging threat to the Amazon? *Tropical Conservation Science, 2*(1), 1-10.

Campbell, J. B. (Ed.). (2002). *Introduction of Remote Sensing*. London: Taylor and Francis.

Carrere, R. (2010). *Oil palm in Africa: Past present and future scenarios* (No. 15). Uruguary: World Rainforest Movement.

Coelho, S. T., da Silver, O. C., Velazquez, S., de Almeida Monteiro, M., & Silotto, E. (2005). The use of palm oil for electricity generation in the Amazon region. RIO 5-World climate & energy event.

Congalton, R. G. (1991). A review of assessing the accuracy of classification of remotely sensed data. *Remote Sensing of Environment*, *37*, 35-46.

Corley, R. H. V., & Tinker, P. B. (2003). *The Oil palm* (4th edn ed.). Oxford: Blackwell Science.

Cortes, C., & Vapnik, V. N. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Cortijo, F. J., & Perez de la Blanca. (1996b). Image Classification using non-parametric classifiers and contextual information. *International Archives of Photogrammetric Engineering & Remote Sensing, XXXI*(B3).

Cowardin, L. M., Carter, V., Golet, F. C., & LaRoe, E. T. (1979). *Classification of wetlands and deepwater habitat of the United States, .*

Da Coata, R. C. (2004). Potential for production bio-fuel in the Amazon deforested area. . *Biomass and Bioenergy, 26*, 405-415.

Dehnavi, G. A., Srikhani, R., & Nagaraju, D. (2010). Image procesing and analysis of mapping alteration zones in environmental research, East of Kurdistan, Iran. *World Applied Sciences Journal, 11*(3), 278-283.

FAO. (2000). *Land cover classification system; classification concepts and user manual*. Rome, Italy: FAO.

FAO. (2005a). *Fertilizer use by crops in Ghana* (1st edn ed.). Rome: FAO.

FAO. (2005b). Small-scale palm oil processing in Africa. *FAO Agricultural Services* Retrieved 20th August, 2011

Fitzpatrick-Lins, K. (1981). Comparison of sampling procedures and data analysis for a land-use and land-cover map. *Photogrammetric Engineering & Remote Sensing, 47*(3), 343-351.

Foody, G., & Mathur, A. (2004a). A Relative Evaluation of Multiclass image classification by Support Vector Machines. *IEEE transactions on Geoscience and Remote Sensing, 42*(6), 1335-1343.

Foody, G., Mathur, A., Sanchez-Harnandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment, 104*, 1-14.

GOPDC. (2011, 2011). Ghana Oil Palm Developement company limited. Retrieved 16 August 2011, 2011, from http://www.gopdc-ltd.com/index.cfm/page:home

Gyasi, E. A. (1992). Emergence of a new oil palm belt in Ghana. *Tijdschrift Voor Economische En Sociale Geografie, 83*(1), 39-49.

Gyasi, E. A. (2003). The environmental impact and sustainability of plantations in Sub-Saharan Africa: Ghana's experiences with oil palm plantations: UNU Press.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2010). *A Practical Guide to Support Vector Classification*. Taiwan: National Taiwan University.

Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). Assessment of the support vector machines for land cover classification. *international journal of Remote Sensing, 23*, 725-749.

Huang, C., Song, K., Kim, S., Townshend, J. R. G., Davis, P., Masek, J. G., et al. (2007). Use of a dark object concept and support vector machines to automate forest cover change analysis. *Remote Sensing of Environment, 112*(3), 970-985.

Huguenin, R. L., Karaska, M. A., van Blaricom, D., & Jensen, R. J. (1997). Subpixel Classification of Bald Cypress and Tupelo Gum Trees in Thematic Mapper Imagery. *Photogrammetric Engineering & Remote Sensing, 63*(6), 717-725.

Husch, B., Beers, T. W., & Kershaw, J. A. (2003). *Forest Measurements* (Fourth ed.). New Jersey: John Wiley & Sons.

Huttich, C., Gessner, U., Herold, M., Strohbach, J. B., Schmidt, M., Keil, M., et al. (2009). On the suitability of MODIS time series metrics to map vegetation types in dry savanna ecosystems: A case study in Kalahari of NE Namibia. *Remote sensing, 1*, 620-643.

Ibrahim, S., Zainal, A. H., & Khalid, M. (2000). *Application of optical remote sensing technology for oil palm management*. Paper presented at the ACRS 2000.

Janssen, L. F. T., & van der Wel, F. J. M. (1994). Accuracy assessment of satellite derived land-cover data: a review. *Photogrammetric Engineering & Remote Sensing, 60*, 419-426.

Jensen, R. J. (1996). *Introduction to Digital Image Processing: A remote sensing perspective, 2 edition*. Piscataway, New Jersey: Prentice Hall.

Jensen, R. J. (2005). *Introductory Digital Image Processing, 3rd edition*. New York: Prentice Hall.

Kamaruzaman, J., & Mubeena, P. (2009). Mapping of Individual oil palm trees using airborne hyperspectral sensing: An Overview. *Applied physical research, 1*(1), 15-30.

Kavzoglu, T. (2009). Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling & Software, 24*(7), 850-858.

Kavzoglu, T., & Colkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation, 11*(5), 352-359.

Khan, M. R., de Bie, C. A. J. M., van Keulen, H., Smaling, E. M. A., & Real, R. (2010). Disaggregating and mapping crop statistics using hypertemporal remote sensing. *International Journal of Applied Earth Observation and Geoinformation, 12*(1), 36-46.

Koh, L. P. (2007). Potential habitat and biodiversity losses from intensified biodiesel feedstock production. [Article]. *Conservation Biology, 21*(5), 1373-1375.

Koh, L. P. (2010). Spatially explicit scenario analysis for reconciling agricultural expansion, forest protection, and carbon conservation in Indonesia. *National Academy of sciences of the United States of America, 107*(24), 11140-11144.

Koh, L. P., & Ghazoul, J. (2008). Biofuels, biodiversity, and people: Understanding the conflicts and finding opportunities. *Biological Conservation, 141*(10), 2450-2460.

Kusimi, J. M. (2008). Assessing land use and land cover change in the Wassa West District of Ghana using remote sensing. *GeoJournal, 71*, 249-259.

Laurance, W. F., Koh, L. P., Butler, R. A., Sodhi, N. S., bradshaw, C. J. A., Neidel, J. D., et al. (2010). Improving the performance of the roundtable on sustainable palm oil for nature conservation. *Conservation Biology, 24*(2), 377-381.

Liberti, M., Simoniello, T., Carone, M. T., Coppola, R., D'Emilio, M., & Macchiato, M. (2009). Mapping badland areas using LANDSAT TM/ETM satellite imagery and morphological data. *Geomorphology, 106*(3–4), 333-343.

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. (2004). *Remote sensing and image interpretation* (5th ed ed.). New York: Wiley.

McMorrow, J. M. (1995). Relation of oil palm spectral response to stand age. *International journal of Remote Sensing, 16*(16), 3203-3209.

Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE transactions on Geoscience and Remote Sensing, 42*, 1778-1790.

Munyati, C. (2000). Wetlands change detection on the Kafue Flats, Zambia, by classification of a multu-temporal remote sensing image dataset. *International journal of Remote Sensing, 21*(9), 1787-1806.

Pal, M., & Mather, P. M. (2003). *Support Vector classifiers for land cover classification*. Paper presented at the Map India Conference 2003

Rahman, M., Bhattacharya, P., & Desai, B. (2005). Similarity Searching in Image Retrieval with Statistical Distance Measures and Supervised Learning

Pattern Recognition and Data Mining. In S. Singh, M. Singh, C. Apte & P. Perner (Eds.), (Vol. 3686, pp. 315-324): Springer Berlin / Heidelberg.

Rossiter, D. G. (2004). *Technical note: Statistical methods for accuracy assessment of classified thematic maps*. Enschede: ITC.

RSPO. (2005) RSOP's principles and criteria for sustainable palm oil production. In RSOP (Series Ed.)*: Vol. Public release version 3* (pp. 1-6).

RSPO. (2007). RSPO Principles and Criteria for sustainable palm oil production, including indicators and Guidance 47.

RSPO. (2009). *RSPO principles and criteria for sustainable palm oil production; Guidance on scheme smallholders*.

RSPO. (2011). Ghana national interpretation of RSPO principles and criteria for sustainable palm oil.

Tan, K. T., Lee, K. T., Mohamed, A. R., & Bhatia, S. (2009). Palm oil: Addressing issues and towards sustainable development. *Renewable and Sustainable Energy Reviews, 13*(2), 420-427.

Thenkabail, P. S., Enclona, E. A., Ashton, M. S., Legg, C., & De Dieu, M. J. (2004). Hyperion, IKONOS, ALI, and ETM plus sensors in the study of African rainforests. *Remote Sensing of Environment, 90*(1), 23-43.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer-Verlag.

Vapnik, V. N. (1998). Statistical learning theory. *Adaptive and learning systems for signal processing, communication, and control*, 736.

Vikesh, K., Vinod, K., & Kamal, J. (2010). Development of spectral signatures and classification of sugercane using ASTER data. *International Journal of Computer Science & Communication, 1*(1), 245-251.

Wahid, B. O., Nordiana, A. A., & Tarmizi, A. M. (2005) Satellite mapping of oil palm land use. Kuala Lumpar: MPOB information series.

World Bank-IFC. (2008). *Key Sustainability issues in the Palm Oil sector: A Discussion paper for Multi-stakeholders consultations*. Washington: World Bank working group.

Yusoff, S. (2006). Renewable energy from palm oil-an innovation on effective utilization of waste. *Journal of Cleaner Production, 14*, 87-93.

Zhang, R., wenling, C., & Hui, Y. (2009). Band selection of Hyperspectral images based on Bhattacharyya distance. *WSEAS transaction on Information science and application, 6*(7), 1165-1175.

Zhu, G., & Blumberg, D. G. (2002). Classification using ASTER data and SVM algorithms;: The case study of Beer Sheva, Israel. *Remote Sensing of Environment, 80*(2), 233-240.

8. LIST OF APPENDICES

8.1 Main Functions in the e1071 Package for Training, Testing, and Visualizing

Some *e1071* package functions are very important in any classification process using SVM in R, and thus will be described here.

The first function is *svm()*, which is used to train a support vector machine. Some import parameters include:

- **data**: an optional data frame containing the variables in the model. If this option is used, the parameters x and y described below, aren't necessary;
- **x**: a data matrix, a vector, or a sparse matrix that represents the instances of the dataset and their respective properties. Rows represent the instances and columns represent the properties;
- **y**: a response vector with one label for each row (instance) of x;
- **type**: sets how svm() will work. The possible values for classification are: C, nu and one (for novelty detection);
- **kernel**: defines the kernel used in training and prediction. The options are: linear, polynomial, radial basis and sigmoid;
- **degree**: parameter needed if the kernel is polynomial (default: 3);
- **gamma**: parameter needed for all types of kernels except linear (default: 1/(data dimension));
- **coef0**: parameter needed for polynomial and sigmoid kernels (default: 0);
- **cost**: cost of constraint violation (default: 1). This is the 'C'-constant of the regularization term in the Lagrange formulation;
- **cross**:specifies the cross-validation. A k > 0 is necessary. In this case, the training data is performed to assess the quality of the model: the accuracy rate for classification;probability: logical indicating whether the model should allow for probability predictions.

8.2Bhattacharyya statistical distance measure

```
›5151
--------------------------------------------------
Display 'Landsat_data_Clip1_landsat.img' 02-27-2012 10:15:02
(MultiSpecWin32_2.20.2012)

  Output Information:

    Statistics computed from base image file and saved in
'Landsat_data_Clip1_Resample1.sta'
    Number of Bad Lines = 0
    Total Number of Pixels = 60,320

    Histogram Summary Table
    First     Last      Line     First      Last    Column
    Line      Line    Interval   Column    Column  Interval
        1       232          1            1       260        1

    Channel      Channel             Data Range         Mean      Median
       Standard
             Description
Deviation
        1            81      to       255          91.0        87
21.4
        2            61      to       255          72.5        68
24.0
        3            50      to       255          67.5        61
26.0
        4            45      to       255          74.9        72
23.6
        5            62      to       255          96.1        91
24.4
        6            32      to       255          58.0        51
29.1

0 CPU seconds for displaying image.  02-27-2012  10:15:09
  --------------------------------------------------


--------------------------------------------------
Feature Selection  02-27-2012  10:23:29 (MultiSpecWin32_2.20.2012)

  Input Parameters:

    Project = 'Untitled Project'
    Original class statistics are used.
    Base image file = 'Landsat_data_Clip1_Resample1.img'
    Algorithm used is "Bhattacharyya"
    List best 20 combinations
    Minimum value to be listed: 0
    Maximum value to be listed: 30000
List separabiliy table.
    1 contiguous channels in each channel combination group
    Step search of channel combination groups will be used

    Channels used: 1-6
```

```
    Classes used:                             Symbol
       1: Oil palm                             1
       2: Forest                               2
       3: Other crops                          3
       4: Shrub                                4
       5: Builtup/Bare                         5

  Output Information:

    There are 10 class combinations.
    There are 6 channel combination(s) for 1 group(s) of 1 contiguous
channel(s).

                 class pair symbols >    12    13    14    15    23    24
       25    34    35    45
                 weighting factor >     (10)  (10)  (10)  (10)  (10) (10)
      (10)  (10)  (10)  (10)
      Channels     Min.       Ave.     Weighted Interclass Distance
Measures
      1.     3           0.62       8.55    1.95  10.0  0.62  19.5
      15.9  1.12  24.2  1.36  4.50  6.18
      2.     2           0.38       5.53    0.38  5.36  0.51  15.2
      6.08  0.64  16.0  1.62  2.36  7.02
      3.     5           0.39       5.15    0.41  17.3  1.30  10.0
      11.6  0.85  7.86  0.71  0.39  0.96
      4.     6           0.05       4.57    0.05  6.31  1.04  12.5
      6.70  1.09  12.9  0.48  1.97  2.51
      5.     4           0.59       4.51    7.59  13.8  0.76  1.50
      2.81  1.24  8.50  0.92  7.43  0.59
      6.     1           0.03       3.33    0.61  1.68  0.03  7.64
      3.26  0.48  10.9  1.24  1.41  6.04

    There are 5 channel combination(s) for 2 group(s) of 1 contiguous
channel(s).

                 class pair symbols >    12    13    14    15    23    24
       25    34    35    45
                 weighting factor >     (10)  (10)  (10)  (10)  (10) (10)
      (10)  (10)  (10)  (10)
      Channels     Min.       Ave.     Weighted Interclass Distance
Measures
      1.   3  4    1.94      17.79    9.03  24.0  4.40  33.0  19.2
      1.94  46.6  9.10  22.5  8.06
      2.   3  5    1.27      10.88    3.40  17.6  3.82  20.1  17.2
      1.27  24.5  2.82  5.76  12.2
      3.   3  6    1.33      10.02    2.38  10.4  1.89  19.9  19.0
      1.33  25.3  4.41  5.13  10.3
      4.   1  3    0.83       9.22    2.37  11.1  0.83  19.6  18.9
      1.16  24.4  1.80  4.76  7.09
      5.   2  3    0.76       9.09    2.18  10.3  0.76  20.4  17.5
      1.13  24.5  1.90  4.73  7.28

    There are 4 channel combination(s) for 3 group(s) of 1 contiguous
channel(s).

                 class pair symbols >    12    13    14    15    23    24
       25    34    35    45
```

```
              weighting factor >     (10)  (10)  (10)  (10)  (10) (10)
     (10)  (10)  (10)  (10)
     Channels     Min.       Ave.     Weighted Interclass Distance
Measures
      1.   3  4  5       2.22       22.69    9.12  27.7  17.3  43.4
     24.1  2.22  54.0  9.42  26.1  13.3
      2.   3  4  6       2.38       20.21    9.13  25.1  13.5  36.1
     21.4  2.38  47.6  11.1  23.5  11.8
      3.   2  3  4       2.10       19.59    11.7  27.8  4.89  37.1
     19.6  2.10  51.3  9.63  22.7  8.78
      4.   1  3  4       2.00       18.74    9.40  25.1  4.51  33.1
     22.4  2.00  46.7  10.0  25.1  8.84

    There are 3 channel combination(s) for 4 group(s) of 1 contiguous
channel(s).

              class pair symbols >    12    13    14    15    23    24
      25    34    35    45
              weighting factor >     (10)  (10)  (10)  (10)  (10) (10)
     (10)  (10)  (10)  (10)
     Channels     Min.       Ave.     Weighted Interclass Distance
Measures
      1.   3  4  5  6    2.51       27.29    9.25  41.0  17.9  45.5
     42.4  2.51  59.0  14.6  27.1  13.4
      2.   2  3  4  5    2.45       26.20    12.3  36.8  20.2  51.3
     26.5  2.45  61.4  10.2  26.5  14.0
      3.   1  3  4  5    2.30       24.32    9.49  30.8  18.9  43.5
     29.8  2.30  54.1  10.6  29.8  13.5

    There are 2 channel combination(s) for 5 group(s) of 1 contiguous
channel(s).

              class pair symbols >    12    13    14    15    23    24
      25    34    35    45
              weighting factor >     (10)  (10)  (10)  (10)  (10) (10)
     (10)  (10)  (10)  (10)
     Channels          Min.       Ave.     Weighted Interclass
Distance Measures
      1.   2  3  4  5  6      2.84       31.45    13.1  42.5  21.9
     60.8  42.4  2.84  72.9  14.7  28.6  14.3
      2.   1  3  4  5  6      2.70       28.37    9.79  41.6  19.9
     46.2  42.9  2.70  60.8  15.6  30.0  13.8

    There are 1 channel combination(s) for 6 group(s) of 1 contiguous
channel(s).

              class pair symbols >    12    13    14    15    23    24
      25    34    35    45
              weighting factor >     (10)  (10)  (10)  (10)  (10) (10)
     (10)  (10)  (10)  (10)
     Channels          Min.       Ave.     Weighted Interclass
Distance Measures
      1.   1  2  3  4  5  6    2.96       33.43    13.3  44.4  22.9
     62.8  43.2  2.96  73.5  16.0  40.3  14.5


2 CPU seconds for feature selection.  02-27-2012  10:23:31
  ---------------------------------------------------
```

8.3 Pictures of field work

Plate 1: Oil palm field with puerera undergrowth & Oil palm field with soil background

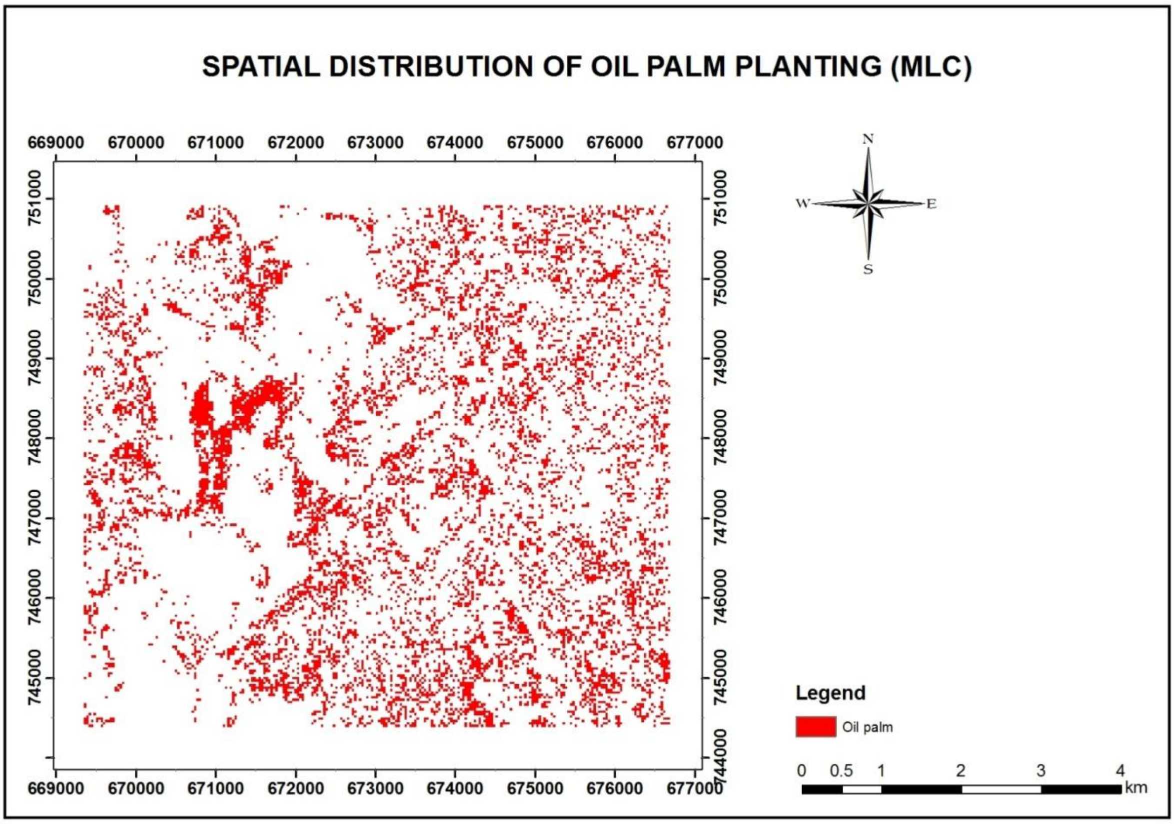
Plate 2: Field work observation made by the researcher & Field Management practices of (9mx9m) spacing

Plate 3: Mixed crops with soil background &Mixed cropping with grass underground

Plate 4: Plate 4: Harvested field observed on the field &Mixed trees species termed as shrub

8.4 Maximum likelihood algorithm

8.5 Maximum likelihood algorithm



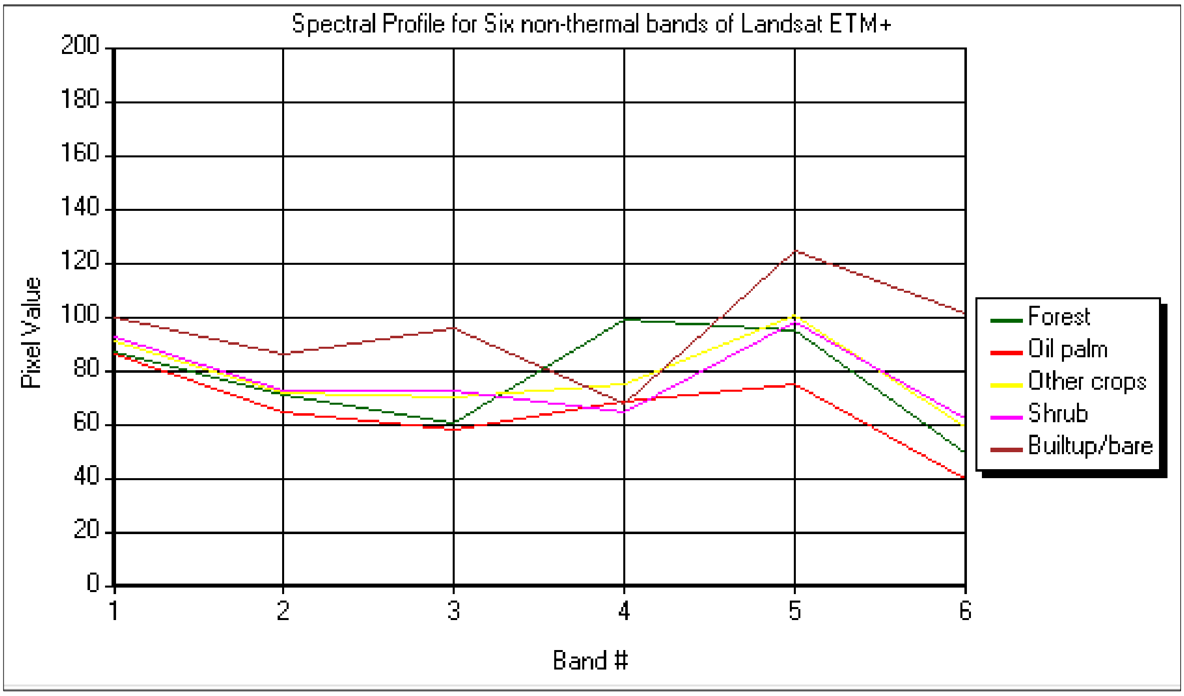


Figure 16: Spectral signatures of land cover classes (Landsat ETM+)