Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics
Technomathematics

Mawazo Mtafya

FOREST VARIABLE ESTIMATION BASED ON SATELLITE IMAGE DATA USING THE SPARSE BAYESIAN METHOD

Master's Thesis

Examiner: Prof. Tuomo Kauranne

Supervisor: Prof. Tuomo Kauranne
Dr. Virpi Junttila

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Computational Engineering and Technical Physics
Technomathematics

Mawazo Mtafya

Forest Variable Estimation based on Satellite Image Data using the Sparse Bayesian Method

Master's Thesis

2018

34 pages, 16 figures, and 2 table.

Examiner: Prof. Tuomo Kauranne

Keywords: Forest Variable, Bayesian Method, Satellite Image, Band Values

Forest inventory studies have been done to estimate and predict forest variable characteristics based either on field plot measurements or remote sensing data or both using different computational tools. The prediction results from those computational methods have proven to be efficient in different studies. One of the approaches that have been applied is the Sparse Bayesian method whereby the predictions have been done when both sattelite image and Laser scanning data as well as field plot measurement are considered. The method works such that regression models are formulated automatically by an algorithm that derives an optimal set of features by comparing different weighted combinations of forest stand variable values with one another. The formulated regression models estimate different forest stand variables like timber volume. The objective of the study is to estimate forest variable based on satellite image data using the sparse Bayesian method. The estimation of forest variable is done by considering different options such as direct estimation using bands values, defining those band values based on forest canopy density, biophysical indices as well as vegetation indicators like Normalized Difference

Vegetation Index (NDVI). Also, satellite image of several years around the target year are considered in the extraction of data.

PREFACE

I sincerely thank God for the love and courage that He keeps showing to me. Since I started the academic journey He has never let me down. He raise me up when I am down, He shows me the way where I could see darkness and He gives me strength in times of difficulties. I believe without him I could not have made it. Thanks to my parents and family as whole for the love they have shown me. They keep on advising, helping and encouraging me on any matter that I get concerned.

Many appreciation to my supervisor Prof. Tuomo Kauranne, for his kindness, good heart and willingness to help that He has offered to me as it is of high value. I really thank Dr. Virpi Junttila, for the cooperation and everything she has offered to me. I appreciate the quality of advice they gave to me through out this work. Without forgetting my friends and fellow students, I am very grateful for their assistance.

Lappeenranta, May 25, 2018

Mawazo Mtafya

CONTENTS

1 INTRODUCTION **7**
1.1 Background 7
1.2 Objective of the Study 9
1.3 Structure of the Thesis 9

2 BASIC TERMINOLOGY **11**
2.1 Satellite Images 11
2.2 Canopy Cover 11
2.3 Forest Canopy Density 11
2.4 Normalized Difference Vegetation Index (NDVI) 13

3 METHODOLOGY **15**
3.1 Study Area 15
3.2 Data Description 15
3.2.1 Field Plot Measurement (Verification Data Set) 16
3.2.2 Training Set 19
3.3 Sparse Bayesian Method 20
3.3.1 Linear Regression Model 21
3.3.2 Bayesian Regression Model 22
3.3.3 The Sparse Bayesian Model 23
3.3.4 Bayesian Inference 24

4 RESULTS AND DISCUSSION **25**

5 CONCLUSIONS **31**

REFERENCES **33**

ABBREVIATIONS AND SYMBOLS

AVI	Advanced Vegetation Index.
BI	Bare Soil Index.
BIO	Normalized Vegetation Difference Backgrounds
LiDAR	Light Detection and Ranging.
NAFORMA	National Forest Resources Monitoring and Assessment of Tanzania
NDVI	Normalized Difference Vegetation Index.
PFP	Private Forest Programme.
SI	Canopy Shadow Index.
TI	Thermal Index.
MNRT	Ministry of Natural Resources and Tourism
NIR	Near-infrared
OLS	Ordinary least squares (OLS
PLS	partial least squares methods
R	Red
SUR	seemingly unrelated regression

1 INTRODUCTION

1.1 Background

Forests play a significant role in the world ranging from ecological to economic importance. Ecologically they regulate climate by acting as a carbon sink for harmful atmospheric gases, they act as catchment areas for water sources and as habitat for different wildlife. Forest products including non-wood forests products are of great importance to humans as they are useful in different contexts of their lives: from household use they are sources of firewood, slash and burn farming, material for making furniture, construction materials and sources of food and income to the community, to industrial use acting as source of energy and the use of timber in sawmills. This diversity of forest resources renders the need for its estimation and prediction.

Due to great importance of forests, the study and management planning of forests should not be neglected. The advancement in science and technology has made the study and management of forest resources more efficient. The use of computation tools in prediction and estimation of forest characteristics like tree height, diameter, basal area, canopy cover and timber volume has shown to be efficient and accurate in performance [1]. Therefore, the development in science and technology has simplified the study and management planning of forest resources including timber volume.

Forest management planning and modelling mostly involves the collection of data from different forest compartments (stands) or on the individual trees by taking measurements of forest characteristics like tree height, diameter, basal area, canopy cover and timber volume. Previously, the process involved visiting the field area and taking measurements on sampled plots either on individual trees or compartment wise measurements and making inferences on forest variable characteristics using those data [2]. Due to time as well as cost constrains to visit each plot and collect information for forests characteristics of interest, field measurements approach can results in too few measurements on the features as well as too few plots which are to be taken into account. In such regard, the spatial accuracy of estimates and prediction based on the approach for the study involving large areas will be low due to poor representation of the true estimates.

The modern approach for forest modeling and management is carried out based on combined field measurement data and remote sensing data such as satellite image data and laser scanning data [3]. Field sample measurements are collected both as a verification

set and as calibration data versus remote sensing data that is used to compute the estimates. Compared to field measurements which are taken on sample plots only, remote sensing data covers the whole study area of interest and estimation of forest stand parameters is carried out over the whole target area as shown in [4]. This research states that, a mathematical modelling approach for forest characteristics is needed, as remote sensing data gives an estimate on forest characteristics based only on variables that correlate more or less with remote sensing data measurements. These models are formulated such that they connect sampled field plot measurements and remote sensing data of the same area and give an output model that estimates or predicts forest characteristics. Therefore, the approaches that combine both remote sensing data and field measurements are more likely to be cost efficient and give more accurate estimate as they covers a large area with more estimates.

Different mathematical models such as k-nearest neighbour (k-NN) and k-most similar neighbour (k-MSN) methods have been used in forest inventories to estimate forest variable characteristics ([5],[6],[7]). According to Junttila, those methods work such that the plots are classified into homogeneous strata and the forest stand variables of each plot in the given stratum is estimated as an average of the measured field plots of that stratum [4]. She states that, the drawback of those methods is that they ignore the variation of plot characteristics. Other methods like the Ordinary least squares (OLS) method, seemingly unrelated regression (SUR) and partial least squares (PLS) methods have been used in forest variable modelling using remote sensing data ([8],[1],[7]). The stand variable parameters under regression methods are predicted from the constructed regression model based on remote sensing data [1].

Junttila introduced the Sparse Bayesian regression method as an automated and adaptive method to forest stand variable selection and estimation in forest inventory as an alternative method to cross-validation approach for variables selection used in OLS and SUR [4]. Different authors have compared the performance of various methods in different studies. Naesset compared different regression methods (OLS, SUR and PLS) on forest stand parameter estimation whereby the result shows very little discrepancy on the model performance [8]. Also as the introduction of Sparse Bayesian method, Junttila et al. compared the performance accuracy of the method with other regression methods (OLS and SUR) on forest stand characteristics estimation where both methods accurately estimated the forest parameters [1]. Therefore, the results from those methods have shown efficiency in the estimation of forest stand characteristics with no significant difference.

According to Junttila et al. [1], the Sparse Bayesian method is used in modelling forest

compartment characteristics from remote sensing data with an automatic selection of features used in the model. They state that regression models are formulated automatically by an algorithm that derives an optimal set of features by comparing different weighted combinations of forest stand variable values with one another. The method involves remote sensing data (satellite image and airborne laser scanning data) and field plot measurements where it establishes a correlation between satellite band values and the measured quantities. Thus, the regression models are formulated to estimate or predict different forest stand variables like basal area, canopy cover, tree height and diameter based on remote sensing measurements.

The data used in forest inventory studies combines remote sensing data (satellite data, laser scanner data and aerial photography data) with field plot measurements as the verification set. The estimation of timber volume, canopy cover and ages of trees (study of forest inventories) has been done based on LiDAR data using the Sparse Bayesian approach. However, there hasn't been studies based on satellite image data only and sample plots using Sparse Bayesian methods. The estimation based on LiDAR data has been shown to work well [1], but it is not known how well it works based on satellite image data directly without LiDAR data. Therefore, the aim of the study is to assess how well satellite image data can predict timber volume of a given plantation forest using the Sparse Bayesian method.

1.2 Objective of the Study

The aim of the study is to test how well the Sparse Bayesian regression methods predicts the biomass in planted forest areas using satellite image data.

The specific objective is to use an advanced Bayesian method to predict the timber volume based on canopy cover, plantation species and plantation age using satellite image data.

1.3 Structure of the Thesis

The following chapters are structured as follows: Chapter 2 is about description and review of various variable (terms) used in the study. We discuss satellite images as well as extraction of satellite image data. We describe canopy cover and its significance to forest modelling. Moreover, one vegetation indicator as well as different biophysical indices

used to assess forest condition are discussed in this chapter. In Chapter 3 we describe the study area, the data set used in the study featuring both field plot measurements and remote sensing data specifically satellite image data. We discuss the methodology starting with the regression model, then Bayesian regression models and Sparse Bayesian model which is the focus of this study.

Chapter 4 provides results from different approaches considered for the analysis of our data. In Chapter 5 we give conclusions.

2 BASIC TERMINOLOGY

2.1 Satellite Images

Satellite images are images of the Earth's surface taken by satellite to be used for different purposes. Among many uses, data are extracted from satellite images using some tools like QGIS that are widely used in forest inventories studies. Compared to other sources of data for forest modelling like field measurements, satellite images cover large areas and therefore they are widely used in forest inventories that cover large areas. Also, as multiple images of the same area are taken, Satellite images are more likely to be cloud-free images [4]. Moreover, they contain different band values each representing an image with a certain wavelength, ranging from ultra violet light (UV), visible light (RGB) to infrared (IR) [4]. Based on the specific analysis we considered a set of satellite images for the years 2015, 2014, 2013 and 2012 collected by United States Geological Survey (USGS).

2.2 Canopy Cover

Canopy cover is the fraction/proportion of the ground area obscured by vegetation when viewed vertically from above. Canopy cover defines the percentage of ground overshadowed by vertical projection of tree crowns onto the ground [9]. Its measurement plays a significant role in many forest inventories. The estimation of canopy cover can be applied in different ecological studies as well as in estimation of leaf area index (LAI) that quantifies the photosynthesizing leaf area per unit ground area ([10], [11]). According to Bolduc et al [12] and Maltamo et al [13], another important role of canopy cover is estimation of timber volume based on remote sensing data using field measurements as verification set.

2.3 Forest Canopy Density

The dynamics in forests have largely increased due to human as well as climate change influence. Humans are continuously exploiting forests so as to meet some of their needs: economical needs such as harvesting forest to get timber and poles as well as forest products like fruits as source of food to people. Forest canopy density biophysical indices

such as advanced vegetation index (AVI), bare soil index (BI), canopy shadow index (SI) and thermal index (TI) can be used to study the condition of forest [14]. The indices have characteristics which show their relationship to forest condition. The features of indices to forest condition are such that: SI increases as the density of forest increases, BI increases as the degree of bare soil exposure to ground increases, TI increases as the quantity of forest increases and AVI responds to all vegetation types like forest and grassland [15]. A graphical representation of relationship between indices characteristics and forest condition is shown in Figure 1.

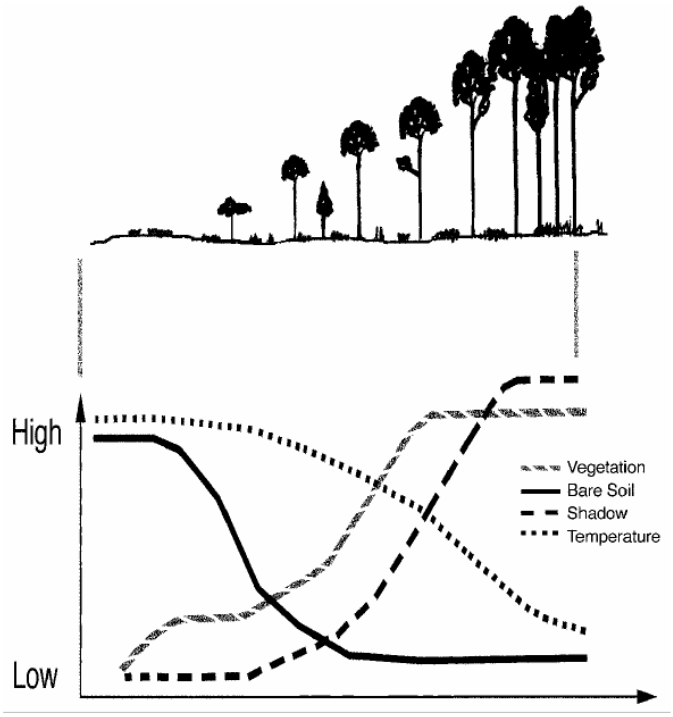

Figure 1. Biophysical indices characteristics relationship to forest condition [14].

To assess the vegetation status of forests, three indices i.e. AVI, BI and SI are calculated from landsat satellite image band values based on the following formulas [14]:

Advanced Vegetation Index (AVI):

$$\begin{aligned} AVI &= 0, \text{ if } B4 < B3 \text{ after normalization} \\ AVI &= ((B4+1)(256-B3)(B4-B3))^{1/3} \end{aligned} \tag{1}$$

whereby:

B3 ⇒ Band three
B4 ⇒ Band four

Bare Soil Index:

$$\begin{aligned} BI &= BIO \times 100 + 100 \\ BIO &= (B5 + B3) - \frac{B4 + B1}{B5 + B3} + (B4 + B1) \end{aligned} \tag{2}$$

whereby:
BI ⇒ Bare soil index
BIO ⇒ Normalized Vegetation Difference Backgrounds
B1 ⇒ Band one
B3 ⇒ Band three
B4 ⇒ Band four
B5 ⇒ Band five

Shadow index (SI):

$$SI = \sqrt[3]{(256 - B1)(256 - B2)(256 - B3)} \tag{3}$$

whereby:
B1 ⇒ Band one
B3 ⇒ Band two
B4 ⇒ Band four

2.4 Normalized Difference Vegetation Index (NDVI)

Vegetation indicators have been applied in measuring the extent and biological characteristics of plants using the spectral reflection (red and infra-red wavelength) of plants [16]. Normalized Difference Vegetation Index (NDVI) as one of vegetation indicators [16], eliminates variations in the sun illumination angle, topographical effects and some other atmospheric elements like haze [17]. The NDVI is calculated as:

$$NDVI = \frac{NIR - R}{NIR + R} \tag{4}$$

whereby:
NIR $\Rightarrow$ Near-infrared (B5)
R $\Rightarrow$ Red (B4)

3 METHODOLOGY

3.1 Study Area

The study area considered in this thesis is in Tanzania. The country is located in East Africa bordered with Uganda and Kenya to the north, with Malawi, Mozambique and Zambia to the south, and with Democratic republic of Congo, Burundi and Rwanda to the west. The country's administrative Zones include: Southern Highland Zone, Central Zone, Coastal Zone, Northern Highland Zone, Lake Zone and Southern Zones. The total forest area in the country is approximately 48.1 million ha covering 55% of the whole land area of Tanzania mainland [18]. The country's estimated plantation forest ranges between 200 000 ha to 550 000 hectares composing of four significant industrial plantation species such as pines (pinus patula, P. elliottii and P. caribaea), cypress, eucalyptus and teak with pines followed by eucalyptus as dominant species in government and private owned plantations ([19], [18], [20]).

Based on vegetation type classification by zones under National Forest Resources Monitoring and Assessment of Tanzania (NAFORMA) report [18] and Private Forest Programme (PFP) study [21], Southern Highland zone is having large (more than twice) fraction of plantation forests as compared to other zones. It is a prominent zone for forest industry location in the country with largest industrial scale plantation forests like Sao Hill industries (SHI) and Tanzania Wattle Company (TANWAT) owning tens of thousands of hectares of plantation forests [21]. According to Penttilä [22] in PFP [21] and Ngaga [19], there are also significant number of smallholders of plantation forest in the zone. Therefore, being the prominent zone for plantation forest, Southern highland zone is considered as the case study in this thesis focusing on three pilot areas as shown in Figure 2.

3.2 Data Description

Sparse Bayesian Regression works such that it establish the correlation between satellite band values and the measured quantities. In our study, we generate the dataset of plantations from satellite images (band values) and then the validation is done on the field plot measurements.

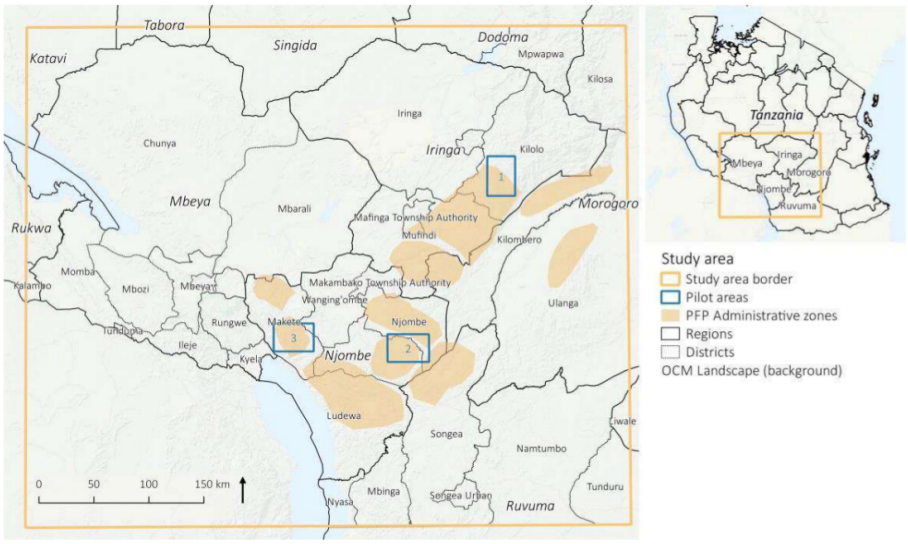

Figure 2. Study area ([21])

3.2.1 Field Plot Measurement (Verification Data Set)

The data set used as verification set in our study is from Forest Plantation Mapping programme of the Southern Highlands of Tanzania [21] based on three pilot areas as shown in Figure 2. The aim of the programme was to carry out plantation mapping of Southern Highlands with specific objectives of identifying the coverage, the number of plantation forests and its proportion of ownership in the zone. The expert team from Food and Agriculture Organization of the United Nations (FAO) and the University of Turku (UTU) were involved in data collection.

The plantation mapping study was carried out first over the whole Southern Highland zones of Tanzania at the resolution of $30m$ and then at higher spatial resolution $(10m)$ on three small pilot areas [21]. Data of various plantation attributes like plantation species, age and density classes as well as canopy cover were collected. The dominant plantation species which were considered include Pines, Eucalyptus and Wattle, with addition of two classes "eucalyptus or wattle" and "mixed or other" due to difficulties in distinguishing the planted species [21]. The plantation age was binned into three classes: recently planted $(0-3$years), growing $(3-8$ years old) and mature class $(>8$ years).

The field data for three variables based on PFP report [21] was as follows: Canopy cover measurements were binned into three classes i.e. "sparse", "intermediate" and "dense" in which for our study we re-scaled the variable to be continuous using class median value; six plantation species are considered in PFP study, but in our case only two species (Pines and Eucalyptus) are considered for prediction; and the three plantation age classes are

converted to be a single variable using the median age of each class. However, there is no measurement for canopy cover in pilot area 3. Therefore, canopy cover estimation for pilot area 3 is not.

The focus of analysis in our study is based on plantation plots from the three pilot areas lying within the plantation mask given in Figure 3. The total number of plantation plots from three pilot areas is 2239 where 819, 657 and 763 are plots for pilot area 1, 2 and 3 respectively. By using QGIS all plantation plots from three pilot areas are attached to a plantation mask (Figure 4). The corresponding plantation plots that lie within plantation areas are selected. The selected plantation plots that lie within plantation mask are 394 with 135, 70 and 189 for pilot areas 1, 2 and 3 respectively. Figure 5 shows the selected plantation plots that lie within plantation areas.

Figure 3. Plantation mask of Southern highlands

From the plantation plots selected, as the prediction on forest variables is based on two species (Pines and Eucalyptus), only plantation plots for those species are considered in our analysis. The number of plots reduces to 139 with 53, 29 and 57 for pilot areas 1, 2 and 3 respectively. The age classes for those plantation plots are converted to be a continuous variable by using the median age of each class. Also, we rescale the plantation age such that it correspond to tree age by the time the satellite image we consider was taken. This is done as follows:

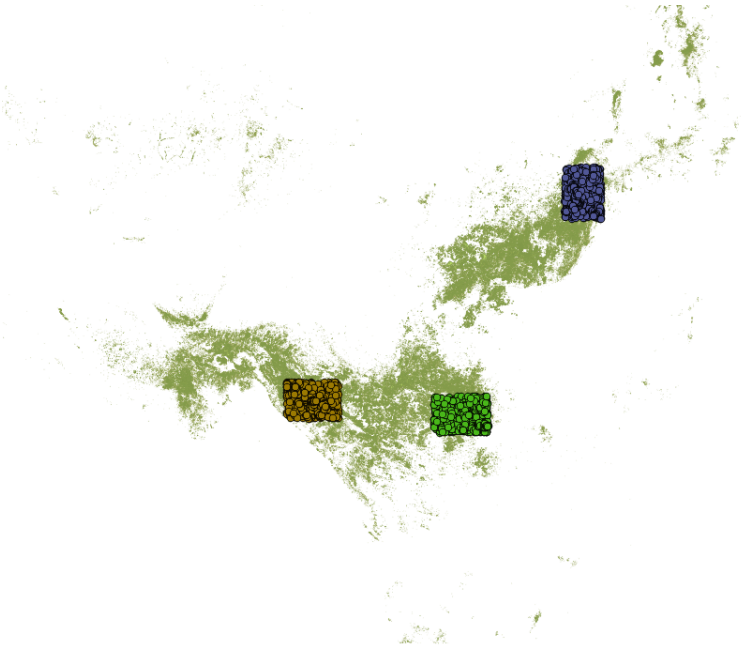

Figure 4. Pilot areas attached to plantation mask

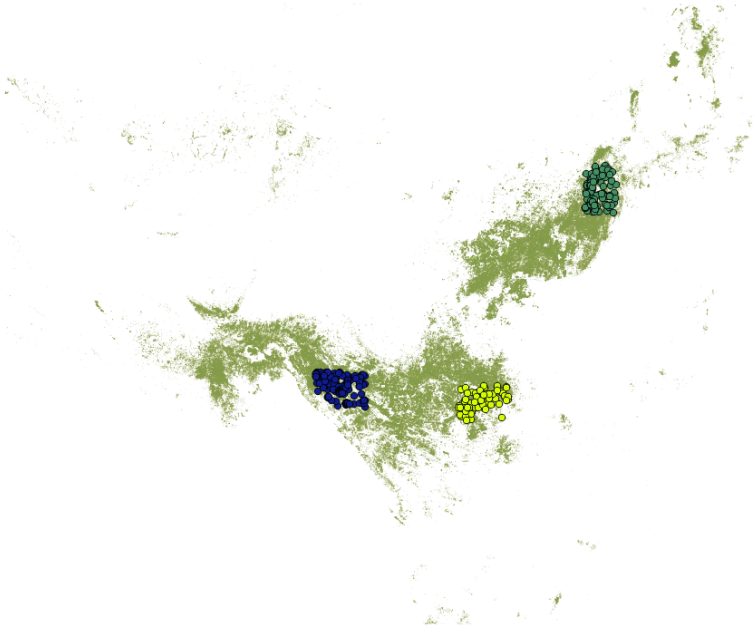

Figure 5. Selected plantation plots that fall within plantation mask

$$Age = M + (SI - IY) \tag{5}$$

where M represent median of age class, SI implies the year when the satellite image we are using was taken and IY is the imagery year which was considered in PFP. Moreover, canopy cover is converted to be a continuous variable with the scale of $0 - 100\%$ using median class value. Figure 6 and 7 are histograms for the ages and canopy cover of plantation plots.

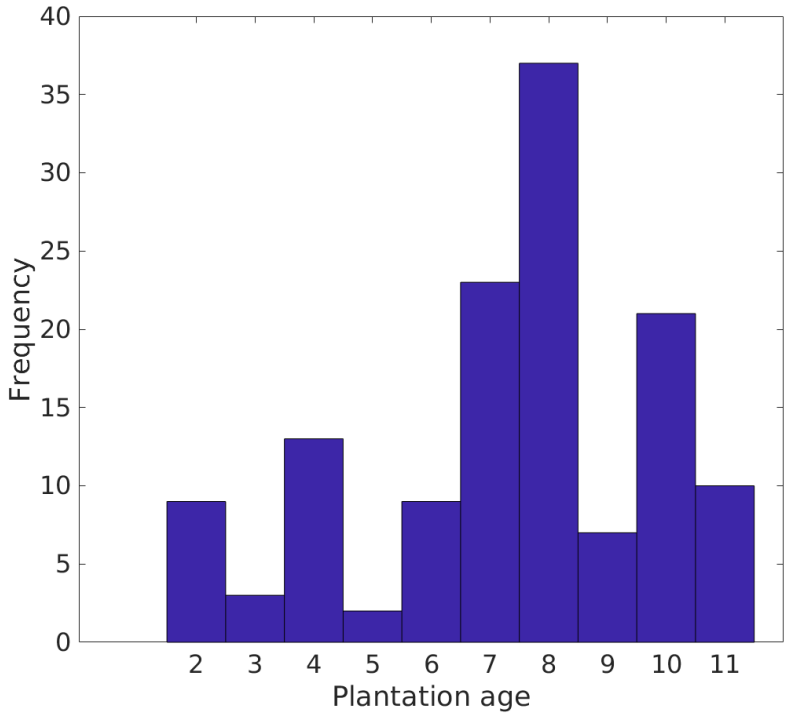

Figure 6. Hstogram plot for the ages of plantation plots

3.2.2 Training Set

Landsat 8 satellite images of different years (2012-2013) collected by United States Geological Survey (USGS) for three pilot areas are downloaded. The corresponding satellite band values for each plot are extracted using Quantum Geographic Information System (QGIS). A set of seven bands out of 11 band values are used in our study. To obtain the band values corresponding to plantation plots over the three pilot areas, pilot area plantation plots were placed over the plantation mask in QGIS. Moreover, other measurement

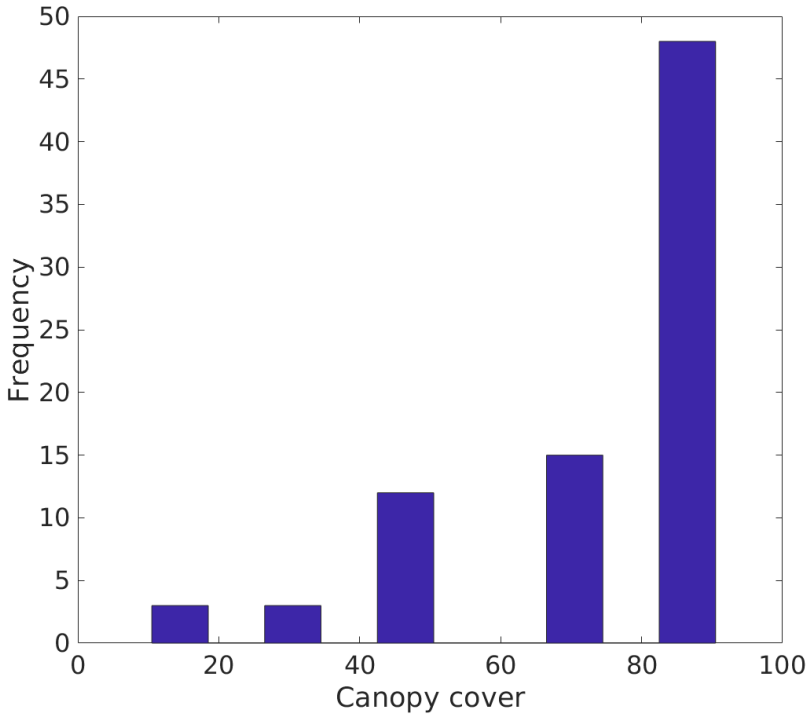

Figure 7. Hstogram plot for the canopy cover of plantation plots

variables such as NDVI, AVI, BI and SI are calculated using those band values. Satellite band values are expected to correlate with the attributes of pilot study measurements. However, the calculated correlation coeffient (R^2) for each band values, biophysical indices (AVI, BI and SI) and NDVI against pilot study measurements: age, species and canopy cover is very low as shown in Table 1 and 2.

Table 1. Calculated R^2 values between each band values and plantation age, species and canopy cover.

	B1	B2	B3	B4	B5	B6	B7
Age	-0.0557	-0.0554	-0.0626	-0.0615	-0.1713	-0.1295	-0.0961
Species	-0.0284	-0.0257	-0.0323	-0.0352	0.0628	-0.0934	-0.0813
Canopy	-0.036	-0.0622	-0.0628	-0.0790	-0.0709	-0.1937	-0.1726

3.3 Sparse Bayesian Method

The Sparse Bayesian method is used in modelling forest compartment characteristics from remote sensing data with an automatic selection of features used in the model [1]. The regression models are formulated automatically by an algorithm that derives an optimal set

Table 2. Calculated R^2 values for NDVI and biophysical indices against plantation age, species and canopy cover.

	NDVI	AVI	BI	SI
Age	-0.1701	0.0550	-0.1001	0.0583
Species	0.1223	0.0672	-0.0019	0.0291
Canopy	0.0198	-0.2298	-0.0848	0.0770

of features by comparing different weighted combinations of forest stand variable values with one another. The method involves remote sensing data and field plot measurements whereby regression models are formulated to estimate or predict different forest stand variables like basal area, canopy cover, tree height and diameter based on remote sensing measurements.

3.3.1 Linear Regression Model

These are empirical models which are used to study the relationship between variables such that, for the given input variable $\mathbf{X}$ and output variable $\mathbf{Y}$, we wish to learn the functional mapping between $\mathbf{X}$ and $\mathbf{Y}$ using some parametrized function $f(\mathbf{X}; \boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$ is the vector of model parameters. The mathematical model to study the dependency between the input vector $\mathbf{X}$ and the output variable $\mathbf{y}$ can be expressed as:

$$\mathbf{y} = f(\mathbf{X}; \boldsymbol{\beta}) + \boldsymbol{\epsilon}, \tag{6}$$

Where $\mathbf{y}$ represents measurements, $f(\mathbf{X}; \boldsymbol{\beta})$ is the function that relates the control variable $\mathbf{X}$ and unknown weights $\boldsymbol{\beta}$ to the measurement and $\boldsymbol{\epsilon}$ is measurement error assumed to follow the normal distribution with mean zero and variance σ^2. Given the model defined in the form of Equation (6), the objective is to find weights $\boldsymbol{\beta}$ such that the function $f(\mathbf{X}; \boldsymbol{\beta})$ makes good predictions for new data or it gives a good functional mapping between the variables $\mathbf{X}$ and $\mathbf{Y}$. A classical approach applied to estimate parameters $\boldsymbol{\beta}$ is the least squares method which minimizes measurement error. The method works such that we look for the weighted parameter β that minimizes the sum of squared differences between the measurements and the model written as,

$$SS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i, \beta))^2 \tag{7}$$

The least square estimator of β that minimizes the sum of squared differences between the measurements and the model is algebraic and expressed as:

$$\beta = (X^T X)^{-1} X^T y \tag{8}$$

3.3.2 Bayesian Regression Model

The models are used to study the likelihood of variable Y given measurements of variable X through conditional probability $p(Y|X)$. This can be done through probability theory, specifically Bayes' theorem which provides us with a "logic of uncertainty" (Jaynes [23] in Tipping [24]) by parametrizing the given conditional probability model such that:

$$p(Y|X) = f(x; \beta), \tag{9}$$

where β represents a vector of model parameters.

The parameters β under the least squares approach are estimated based on measurements y. In the Bayesian setting β is treated as a random variable the same way as X and Y and the goal is to find the posterior distribution of β which gives the probability density for values of β given measurements y [24]. Therefore, the conditional probability for measurement Y given X becomes $P(Y|X, B)$, and the dependency of the probability Y and X on the parameter variable β is made explicit with inference of posterior distribution of parameters β from Bayes' rule [24] defined as:

$$\pi(\beta|Y) = \frac{l(Y|\beta)p(\beta)}{\int l(Y|\beta)p(\beta)d\beta} \tag{10}$$

where, $\pi(\beta|Y)$ denotes posterior distribution, $l(Y|\beta)$ is a likelihood function which gives the probability of observing response Y given parameter β and it contains measurement errors. $p(\beta)$ is a prior distribution containing existing information for parameter β and the integral in the denominator is a normalization function.

3.3.3 The Sparse Bayesian Model

The model formulation used in the analysis of forest stand parameters for this study is based on the Sparse Bayesian model by Junttila [1]. Each forest parameter such as canopy cover, plantation species and age is estimated with a different model from Landsat 8 satellite image data for the given study area of interest. A linear model for each stand parameter or target response vector with the given matrix of proposal variable x is defined as:

$$\mathbf{y} = \boldsymbol{\beta} X + \boldsymbol{\varepsilon}, \tag{11}$$

whereby, $\mathbf{y}_{k,P}$ represents the measurements of plots P denoted as $\mathbf{y} = [y_{m,1}, \ldots, y_{m,P}]^T$, $\mathbf{x}_i = [1, x_{i1}, x_{i2}, \ldots, x_{iM}]^T$ represent the design matrix for M proposal variables with dimension $P \times M + 1$, a column of $1_{P,1}$ denote an intercept of the model . The vector $\boldsymbol{\beta}$ stands for the parameters of the model denoted as $\boldsymbol{\beta} = [\beta_{m,1}, \ldots, \beta_{m,M}]^T$ and vector $\boldsymbol{\epsilon}$ stands for measurement errors of the model expressed as $\boldsymbol{\epsilon} = [\epsilon_{m,1} \ldots, \epsilon_{m,P}]^T$.

The linear regression model given in Equation (11) for Sparse Bayesian regression is stated in probabilistic function form allowing variable selection to be based on the Bayesian approach [4]. Assuming independence in the measurements, the likelihood function for the data set to be modelled based on Equation (6) can be written as:

$$p(y|\beta, \sigma^2) = \prod_{p=1}^{P} N(X_p\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{P/2}} e^{-\left(\frac{||y - X\beta||^2}{2\sigma^2}\right)} \tag{12}$$

where $X_P\beta$ is the mean of the likelihood estimates for y_P and σ^2 is the measurement error variance. As in conventional regression models, the error term ϵ in the model (11) is assumed to be normally distributed with mean zero and variance σ^2, in probabilistic notation, $p(\epsilon) = \prod_{p=1}^{P} N(0, \sigma^2)$.

A good prediction for β, in a regression setting is obtained by minimizing the error term i.e. $\sum_{p=1}^{P} \epsilon_P^2$, while, in a Bayesian setting, this is achieved by calculating the maximum likelihood estimate for β which maximizes the probability function $p(y|\beta, \sigma^2)$ in Equation (12). When too many parameters are to be estimated from the model, it can result on over-fitting [25]. To control over-fitting of the model, the prior distribution is defined with a penalty term α which regularizes the values that β might take such that:

$$p(\beta|\alpha) = \prod_{m=1}^{M+1} N(0, \alpha_m^{-1}), \tag{13}$$

where the hyperparameter variable $\alpha = [\alpha_1, \ldots, \alpha_{M+1}]$ controls the magnitude of weights the parameters β are associated with.

3.3.4 Bayesian Inference

The inference based on the Bayesian setting is done by computing the posterior distribution of all unknown parameters for the given data. The model parameters are estimated by calculating its posterior distribution conditioned on the data through Bayes' rule defined as:

$$p(\beta|y, \alpha, \sigma^2) = \frac{p(y|\beta, \sigma^2)p(\beta|\alpha)}{p(y|\alpha, \sigma^2)}, \tag{14}$$

where $p(y|\alpha, \sigma^2) = \int p(y|\beta, \sigma^2)p(\beta|\alpha)d\beta$. The posterior distribution is also Gaussian, as its likelihood function is Gaussian combined with a Gaussian prior and a linear model such that:

$$p(\beta|y, \alpha, \sigma^2) = N(\beta|\mu, \Sigma), \tag{15}$$

with mean and covariance calculated analytically from the exponents of Gaussian distribution defined as,

$$\begin{aligned} \Sigma &= (\phi X^T X + A)^{-1} \\ \mu &= \Sigma \phi X^T y \end{aligned} \quad . \tag{16}$$

4 RESULTS AND DISCUSSION

The model is fitted such that forest parameters (canopy cover, plantation species and age) are predicted using Landsat 8 satellite image band values. The band values for each pilot study area 1, 2 and 3 are obtained from Satellite image of 2015 using QGIS. The prediction as well as the goodness of fit of a model for forest variables is made based on each pilot study area and then for combined pilot areas. The result shows that satellite band values do not seem to predict our variables of interest very well as shown in Figure 8, 9 and 10 for each pilot areas and Figure 11 when three pilot study area are combined. Canopy cover estimatimation for pilot area 3 and the combined case is not estimated since there was no measurements for it in pilot study area 3. Moreover, to test the overall fit of the models to the data, the coefficient of determination (R^2) is calculated for each model fitted and the obtained values are very low i.e. < 0.25 showing low fit of the model to the data.

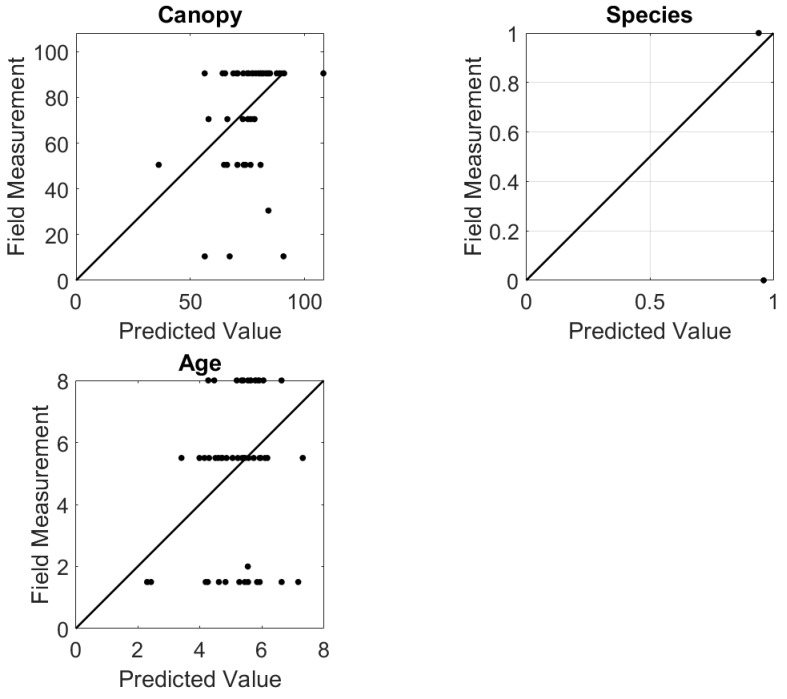

Figure 8. The scattergram plot for forest variable estimation based on pilot area 1.

Based on PFP study, satellite images of different years mostly of 2012-2015 are used in the interpretation of forest variables measurements [21]. The difference in satellite imagery year under PFP study and the one we use to obtain band values may have an effect on the results. The next step as of previous where satellite image of 2015 only was

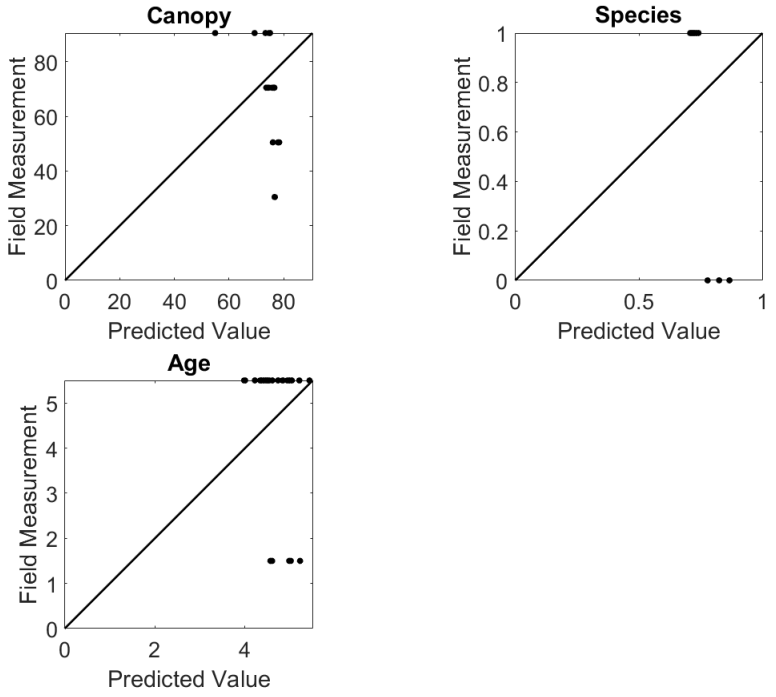

Figure 9. The scattergram plot for forest variable prediction based on pilot area 2.

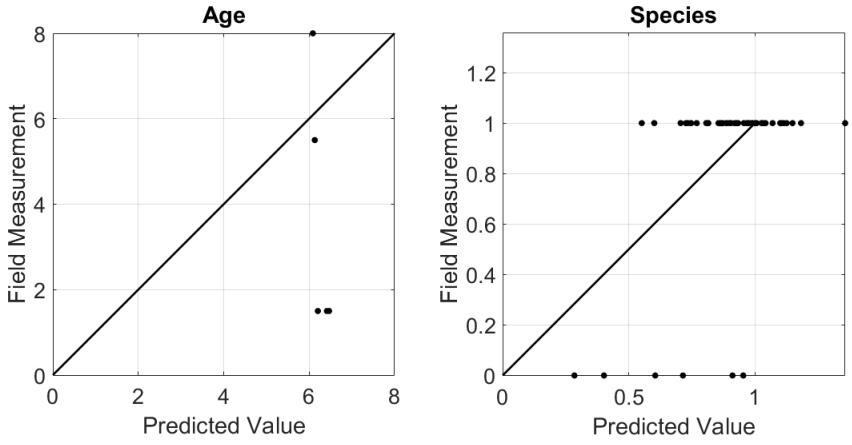

Figure 10. Forest variable estimation scattergram plot for pilot area 3.

used to obtain band values, is to obtain the band values from satellite image corresponding to the imagery year of PFP study. That is, when imagery year was of 2013 in PFP study, we obtain the corresponding band values of 2013 for the prediction of our variable of

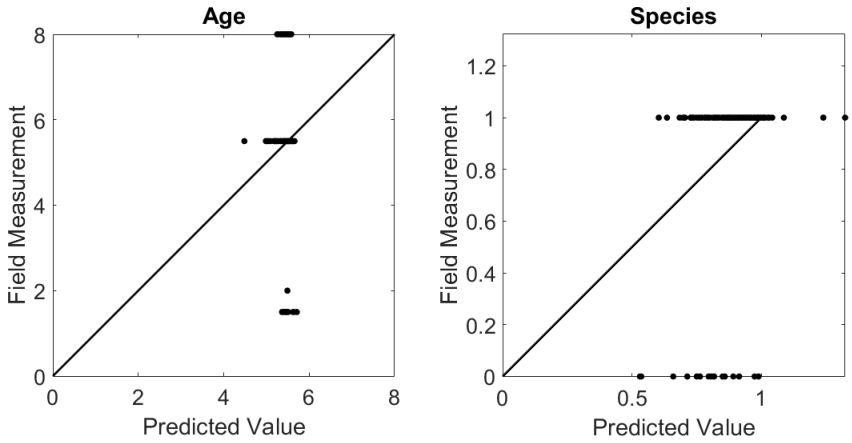

Figure 11. The scattergram plot for forest variable prediction when pilot areas are combined.

interest. Also, in case the imagery year was of 2015 and 2014 the satellite band values for year 2015 are obtained as there might be no big changes in the plantation, only if harvesting has taken place. Using QGIS, the band values for the given imagery year are obtained and prediction as well as overall goodness of fit for the model is made. The results in both cases seem not to improve as shown in Figures 12 and 13 as well as R^2 values are still low i.e. < 0.2.

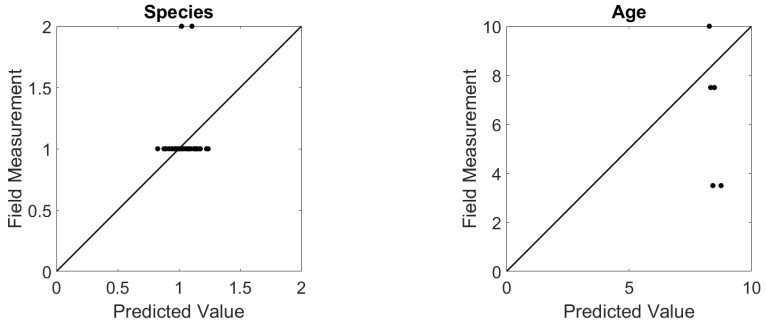

Figure 12. The plot for forest variable estiomation when satellite image of 2013 with coresponding 2013 imagery year are considered.

To improve our results, three canopy density biophysical indices: AVI, BI and SI that are defined by various ratios of band values are constructed. The model is fitted with both original band values and the new defined band values and also, the one with biophysical

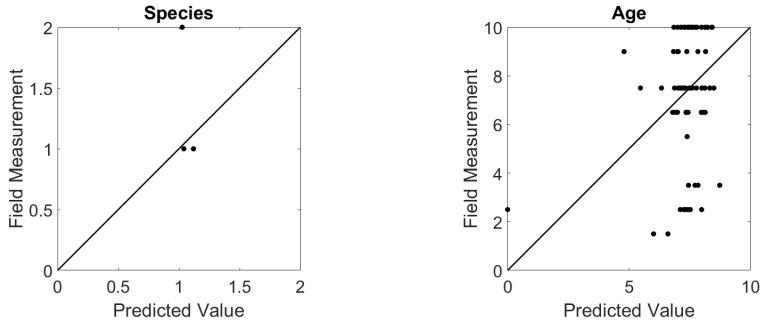

Figure 13. The plot for forest variable prediction when satellite image of 2015 with imagery year of 2015,2014 and 2013 are considered.

indices only. The overall fit of models is assessed in all cases. The results in both cases seem not to be different from the ones obtained in the previous cases: when satellite band values only and when satellite image of the same year as the imagery year under PFP study are considered. Figures 14 and 15 display a graphical representation of results when both band values and biophysical indices and when only biophysical indices are applied. Moreover, in both cases, the coefficient of determination is still low.

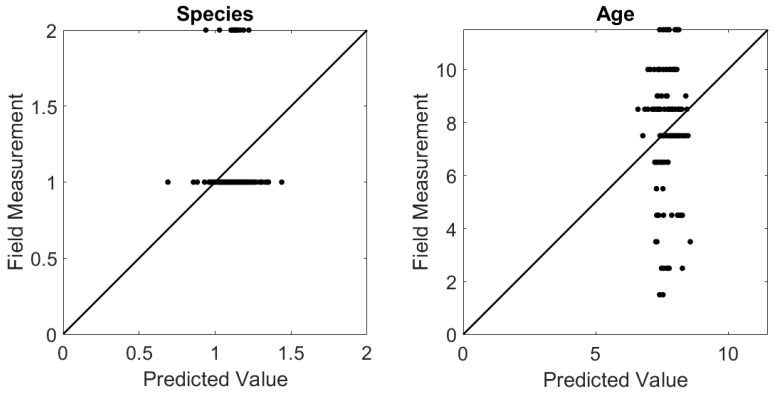

Figure 14. The scattergram plot for estimation of forest variables when both bands value and biophysical indices are applied.

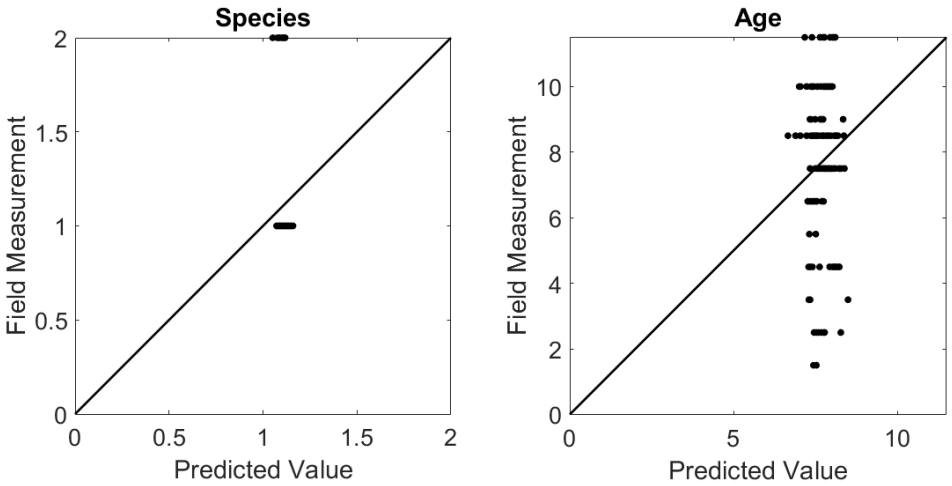

Figure 15. The scattergram plot for prediction of forest variables when only biophysical indices is considered.

From scattergrams plotted, the plantation age plots show some structure of correlation between the plantation age and NDVI. Therefore, we fit the model to estimate plantation age for all three pilot area using NDVI under the Sparse Bayesian method. Plantation plots that seem to be outliers are removed in the analysis. The resulting scattergram (Figure 16) shows very low negative correlation in the age prediction using NDVI as well as a low coefficient of determination.

The other alternative considered is to fit a normal linear regression model for age prediction using NDVI. The results remained the same as the one obtained in the previous options. The scattergram plot in Figure 16 shows partial correlation on the plantation age using NDVI.

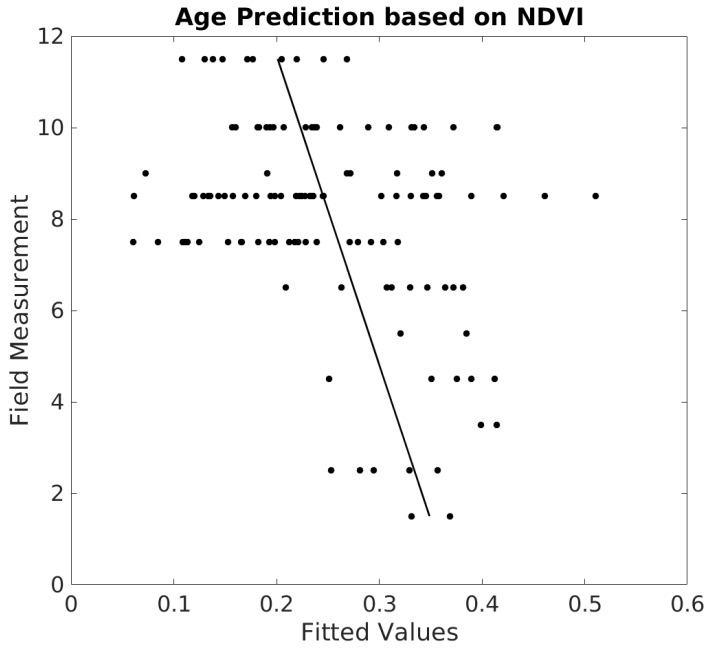

Figure 16. The scattergram for plantation age estimation using NDVI.

5 CONCLUSIONS

Forest inventory studies have been done to estimate and predict forest variable characteristics based on both field measurements and remote sensing data using different computational tools. The results from those computational methods like regression methods and other models for forest parameters estimation have proven to be efficient in different studies. Based on those studies different authors mostly combined remote sensing data (either aerial photograph data with laser scanning data or satellite image data) and field measurements have been tested on various approaches. However, the efficiency of satellite image data alone combined with field measurements on forest stand characteristics prediction using the Sparse Bayesian approach have not been tested. In our study we focused on estimation of timber volume based on the prediction of three attributes: canopy cover, plantation species and age by Sparse Bayesian method using Satellite image data.

Different options are considered to estimate timber volume from satellite image data using the Sparse Bayesian method. The first option was to use satellite band values to estimate canopy cover, plantation age and species. To avoid the effect in the difference between imagery year and the satellite image used to extract band values, the satellite image corresponding to the imagery year is downloaded and the data for this case is used in the estimation. Also, Landsat 8 satellite band values are defined based on forest canopy density biophysical indices: AVI, BI and SI and then the prediction of timber volume is made from those indices and when band values are combined with indices. Moreover, as vegetation indicators measure the extent and biological characteristics of plants using the spectral reflection of plants, NDVI as one of the indicator is considered in the study. Irrespective of partial correlation from estimation using NDVI, generally, the results from those options considered seem not to give promising predictions of timber volume using the Sparse Bayesian method.

The dataset considered as verification set (field plot measurement) is from PFP study whereby its collection was based on high resolution visual interpretation of satellite images. Any inaccurate interpretation of satellite images like classifying natural forest as plantation forest, misclassification of canopy cover and plantation age and considering such incorrectly classified forests in the prediction of timber volume may result in incorrect results. The plantation plots considered are extracted from the plantation mask developed by PFP study which covers the whole area of southern highland zone of Tanzania. A highly accurate matching of plantation plot coordinate on the plantation mask as well as plantation mask itself to represent plantation forest would be very desirable but may be very difficult to produce. Incorrectness in either case will lead to unreliable

results.

Generally, for the results obtained based on satellite data using the Sparse Bayesian method,the prediction of forest variable seems not to be reliable. Possible reasons might be that satellite image data alone might not be good estimators of timber volume based on the three attributes considered: canopy cover, plantation age and species. Also, the data considered as the verification set as was based on visual interpretation and might not be so reliable information for plantation studies over the given area. Moreover, the plantation mask that was used in the extraction of plantation plots for the study might not have been so accurate in providing accurate information of plantation area as there might be some areas with natural forest being interpreted as plantation areas.

REFERENCES

[1] V Junttila, M Maltamo, and T Kauranne. Sparse bayesian estimation of forest stand characteristics from airborne laser scanning. *Forest Science*, 54(5):543–552, 2008.

[2] A Kangas and M Maltamo. *Forest inventory: methodology and applications*, volume 10. Springer Science & Business Media, 2006.

[3] Endre H Hansen, T Gobakken, S Solberg, A Kangas, L Ene, E Mauya, and E Næsset. Relative efficiency of als and insar for biomass estimation in a tanzanian rainforest. *Remote Sensing*, 7(8):9865–9885, 2015.

[4] V Junttila. Automated, adaptive methods for forest inventory. *Acta Universitatis Lappeenrantaensis*, 2011.

[5] P Packalén and M Maltamo. Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. *Forest Science*, 52(6):611–622, 2006.

[6] Kari T Korhonen and A Kangas. Application of nearest-neighbour regression for generalizing sample tree information. *Scandinavian Journal of Forest Research*, 12(1):97–101, 1997.

[7] A Haara and A Kangas. Comparing k nearest neighbours methods and linear regression-is there reason to select one over the other? *Mathematical and Computational Forestry & Natural Resource Sciences*, 4(1):50, 2012.

[8] E Næsset, Ole M Bollandsås, and T Gobakken. Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. *Remote Sensing of Environment*, 94(4):541–553, 2005.

[9] Almasi S Maguya et al. Use of airborne laser scanner data in demanding forest conditions. *Acta Universitatis Lappeenrantaensis*, 2015.

[10] SB Jennings, ND Brown, and D Sheil. Assessing forest canopies and understorey illumination: canopy closure, canopy cover and other measures. *Forestry: An International Journal of Forest Research*, 72(1):59–74, 1999.

[11] L Korhonen, Kari T Korhonen, M Rautiainen, and P Stenberg. Estimation of forest canopy cover: a comparison of field measurement techniques. 2006.

[12] P Bolduc, K Lowell, and G Edwards. Automated estimation of localized forest volume from large-scale aerial photographs and ancillary cartographic information in a boreal forest. *International Journal of Remote Sensing*, 20(18):3611–3624, 1999.

[13] M Maltamo, K Eerikäinen, J Pitkänen, J Hyyppä, and M Vehmas. Estimation of timber volume and stem density based on scanning laser altimetry and expected tree size distribution functions. *Remote Sensing of Environment*, 90(3):319–330, 2004.

[14] A Rikimaru, PS Roy, and S Miyatake. Tropical forest cover density mapping. *Tropical Ecology*, 43(1):39–47, 2002.

[15] MS Jamalabad. Forest canopy density monitoring using satellite images. In *Geo-Imagery Bridging Continents XXth ISPRS Congress, Istanbul, Turkey, 2004*, 2004.

[16] A Abdollahnejad, D Panagiotidis, and P Surovỳ. Forest canopy density assessment using different approaches–review. *J. for. sci*, 63(3):107–116, 2017.

[17] M Mróz, A Sobieraj, et al. Comparison of several vegetation indices calculated on the basis of a seasonal spot xs time series, and their suitability for land cover and agricultural crop identification. *Technical Sciences*, 7(7):39–66, 2004.

[18] Ministry of Natural Resources and Tourism (MNRT). *National Forest Reasources Monitoring and Assessment of Tanzania Mainland (NAFORMA). Main results*. Ministry of Natural Resources and Tourism., 2015.

[19] YM Ngaga. Forest plantations and woodlots in tanzania. In *African Forest Forum. Nairobi*, 2011.

[20] FRA and FAO. Global forest resources assessment 2015 desk reference. *Food and agriculture organization of the United Nations, Rome*, 2015.

[21] Private Forest Programme (PFP). *Forest Plantation Mapping of the Southern Highlands.* Final report. Iringa, Tanzania., 2017.

[22] J Penttilä et al. Influence of smallholder woodlot management to tree growth and tree quality. 2016.

[23] Edwin T Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

[24] Michael E Tipping. Bayesian inference: An introduction to principles and practice in machine learning. In *Advanced lectures on machine Learning*, pages 41–62. Springer, 2004.

[25] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.