Above Ground Biomass Estimation of a Cocoa Plantation using Machine Learning

Sabrina Sankar, Marvin Lewis and Patrick Hosein
The University of the West Indies
St. Augustine, Trinidad
sabrina.sankar@my.uwi.edu, lewismarvin766@hotmail.com, patrick.hosein@sta.uwi.edu

***Abstract*—The rapid increase in carbon dioxide in the atmosphere and its associated effects on climate change and global warming has raised the importance of monitoring carbon sequestration levels. Estimating above ground biomass (AGB) is one way of monitoring carbon sequestration in forested areas. Quantifying above ground biomass using direct methods is costly, time-consuming and, in many cases, impractical. However, remote sensing technologies such as LiDAR (Light Detection And Ranging) captures three dimensional information which can be used to perform this estimation. In particular, LiDAR can be used to estimate the diameter of a tree at breast height (DBH) and from this we can estimate its AGB. For this research we used LiDAR data, along with various Machine Learning (ML) algorithms (Multiple Linear Regression, Random Forest, Support Vector Regression and Regression Tree) to estimate DBH of cocoa trees. Various feature selection methods were used to select the most significant features for our model. The best performing algorithm was Random Forest which achieved an R^2 value of 0.83 and Root Mean Square Estimate (RMSE) value of 0.062. This algorithm then estimated an AGB value of 28.75 ± 2.34 Mg/ha (Megagram per hectare). We compared this result with that obtained from locally-developed allometric equations for the same cocoa plot. The comparison proved our estimate to be 14.7% lower than the allometric equation. The results demonstrated that using ML with LiDAR measurements for AGB estimation is quite promising.**

***Index Terms*—Above Ground Biomass, AGB, LiDAR, Machine learning, Random Forest, Regression Tree, Regression, Forest Variable Estimation, Climate Change, Carbon Sequestration**

I. Introduction

Within recent decades, climate change and its effects have grown in importance. Without immediate intervention to reduce global greenhouse gas emissions, climate models estimate an increase of over 2°C in global temperature from pre-industrialized levels [1]. Carbon dioxide is the main greenhouse gas that is produced because of human activities, comprising an estimated 79% of total U.S. greenhouse gas emissions in 2020 [2]. Due to human activity since the Industrial Age, the proportion of carbon dioxide in the atmosphere has increased by 45% from 280 to 412 ppm (parts per million) in 2019 [3]. The effects of further increases in atmospheric carbon dioxide and global temperature can lead to serious consequences for the environment and human life. As such, new protocols and strategies have been put in place to manage and reduce carbon emissions.

The occurrence of carbon fluxes, which is the transfer of carbon from one pool to another [4], within the last few hundred years, have led to carbon being locked away in continental crusts and oceans. The extraction of fossil fuels and other resources due to mining and deforestation, have resulted in the release of the stored carbon into the atmosphere.

Atmospheric carbon can be managed in several ways, such as reducing fossil fuel consumption from industrial activities, utilizing alternate energy technologies like renewable energy resources and increasing carbon sequestration throough existing carbon pools. The latter is the process of increasing the carbon content of a carbon pool other than the atmosphere [4]. One of the more sustainable and cost-effective methods of sequestration can be achieved by using forests and vegetation as a carbon pool. In forest systems, carbon is stored in the vegetation, including the root system, dead wood, litter, and the soil, by absorbing carbon from the environment. The AGB in the vegetation consists of stems, branches, and foliage. It usually comprises between 35-65% of the total dry weight. This provides a measure of how much carbon is stored by a tree. Variation in biomass per tree is due to differing tree species and climate conditions.

Cocoa, a woody plant, is one of the commodities grown in plantations that span many hectares and has been described in some cocoa producing countries, such as Indonesia, as being strategically placed to increase carbon sequestration efforts [5], [6]. The increasing cost and demand for cocoa provides financial incentives for the development of cocoa plantations. These plantations contain shade trees that protect the young cocoa trees and thus, are classed as a cocoa-based agroforestry system [7]. These systems, once sustainably maintained, possess significant economic and carbon sequestration potential [8].

Carbon dioxide absorbed by cocoa plants under optimal conditions have been observed to be 80 Mg/ha/year while releasing 63 Mg/ha/year. When fruit production and other processes are taken into account, the yearly net uptake is 73 Mg/ha/year [5], [9]. In Indonesia, carbon stock analysis for cocoa plantations have reported carbon sequestration levels ranging from 25.52 Mg/ha to 33.19 Mg/ha [5] whilst in Central America, levels were estimated at around 49 Mg/ha [6].

Estimating the biomass of trees is therefore very important when monitoring carbon sequestration levels. However, exact methods of biomass calculation involve the use of destructive methods to determine the dry weight of the entire plant. Allometric equations are a non-destructive alternative that have

2375-9259/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDMW58026.2022.00147

been used to estimate tree biomass based on its relationship with the different tree components.

Gathering information for allometric modelling is time consuming and expensive especially in areas where forest inventories do not exist. LiDAR is a remote sensing technology, which allows data collection in a quick and efficient manner. The best use of this technology is its application collecting data in remote areas that are too difficult or costly to navigate. This type of technology not only allows for accurate data to be obtained but also reduces the need for destructive methods.

Trinidad and Tobago is a country that is well-known for the high quality cocoa it produces. Cocoa production has plummeted within the last several decades from 30,000 tons to around 500 tons [10]. In recent years, the government has expressed great interest in revitalizing the cocoa industry in the country due to the increasing global demand for high quality cocoa [11]. There is limited data on tree canopy and density for local cocoa plantation, resulting in inaccurate and outdated land use information. The revitalization and establishment of cocoa plantations can play an important role by reducing global carbon footprint and increasing the economic gain associated with a rise in cocoa production. This study therefore seeks to estimate the biomass of a cocoa plantation using LiDAR technology and small-scale tree sampling to aid in carbon monitoring efforts throughout the twin island country. Our field sampled data was comparatively small for typical machine learning applications. As a consequence, various methods were used to avoid over-fitting.

II. Related Work and Contributions

Many studies have been conducted to estimate the AGB with different types of climate conditions and tree species. These studies typically create equations to quantify the biomass using allometric modelling with metrics derived from LiDAR. Others have used remote sensing techniques to estimate forest variables from which computations regarding tree biomass can be achieved. In recent years, machine learning algorithms have been applied with LiDAR data to estimate the AGB of many forested systems.

Traditional allometric modelling, undertaken by Patenaude et al., [12] utilized LiDAR to quantify the AGB of the temperate woodland in the Monks Wood Nature Reserve. This was achieved by developing an equation to calculate the biomass. In-depth field data collection was required and involved grouping trees of similar species, composition and structure from which a grid of 10, 20×20 m plots were created from the 10 groupings. From this, one sample plot from each grouping was chosen where all the trees with a diameter at breast height (DBH) greater than 7 cm were counted and measured. The total foliage, ground vegetation and litter carbon content were calculated using species specific allometric equations.

LiDAR data was used to generate a DTM raster using the triangular irregular network (TIN) based on a Delaunay triangulation [12]. The DTM was subtracted from the canopy elevation layer to generate a canopy height model (CHM). The results showed that the CHM consistently underestimated the tree heights from the reference data. Canopy density and structure as well as the configuration of the LiDAR were identified as possible causes for the underestimation. The equation to estimate the biomass utilized height metric percentiles derived from LiDAR as the predicting variable. This model obtained a R^2 of 0.74 on the field level and 0.85 on the plot level. This study showed that good results can be obtained without applying complex machine learning algorithms with the data. A similar approach was done in [13] to estimate the AGB of a subtropical forest in Hong Kong. Varying the plot size used for biomass estimation in this case significantly affected the performance of the models. A plot size of 10 m^2 provided R^2 results as high as 0.864.

Most of the studies done to derive allometric equations for a particular area employ stepwise regression for AGB estimation. Within the last decade, the application of ML algorithms to model forest data has increased, with better performing models being created. Gao et al. [14] compared the performance of five ML algorithms on optical and radar data obtained from Landsat Thematic Mapper (TM) and ALOS PALSAR (Advanced Land Observing Satellite-1 Phased Array type L-band Synthetic Aperature Radar) respectively. AGB sample plots, Landsat imagery, ALOS PALSAR L-band data, digital elevation model data and classified images of forest types were processed and variables extracted. Feature selection was done using Stepwise Regression and Random Forest from which biomass estimation models were developed using Multiple Linear Regression, Random Forest, Artificial Neural Networks, Support Vector Regression and k-Nearest Neighbours. Overestimation and underestimation of small and large AGB values were observed respectively. The Multiple Linear Regression performed best within an AGB range of 40–120 Mg/ha, whilst performance of the other ML algorithms was limited [14]. Artificial Neural Networks were deemed the best performing model within the study and Random Forest and k-Nearest Neighbours performed poorly.

Torre-Tojal et al. [15] used LiDAR data in northern Spain to estimate the AGB using Random Forest models. Allometric equations were applied to calculate the AGB based on the geographical proximity and tree species. LiDAR data was pre-processed to coincide with the ground measurement circular plots of 25 m radius in which individual digital terrain models (DTMs) and digital surface models (DSMs) were extracted to generate the canopy height model (CHMs). FUSION/LDV and PostGIS software were used to obtain 65 variables for biomass estimation. The Gini Importance was used as the feature selection method to assess and reduce the number of variables. Hyper-parameter tuning on 6 hyper-parameters of the Random Forest algorithm was done to optimize the performance of the models created. Model assessment and validation was done using 2-fold cross validation with the R^2 and RMSE performance metrics. Two of the models created performed well with R^2 values 0.726 and 0.708. Overestimation of the biomass using the two models were between 16% and 18% when compared to the ground truth biomass calculations.

Biomass estimations are most frequently conducted at plot level as it reduces the number of resources that would be needed. In [16] the authors applied image segmentation on LiDAR data to estimate the biomass of forested areas in Arkansas and Texas using three ML algorithms. The image segmentation technique groups individual pixels of similar attributes, [16], as well as incorporating spatial correlation. The study involved deriving height metrics from the LiDAR data, applying image segmentation to identify homogeneous forest units, building regression models based on the image segment data [16] and comparing the performance of the models with in-situ biomass measurements. Model performance of Regression Tree, Random Forest and Support Vector Regression were assessed and validated using a 10-fold cross validation. The training and validation accuracies obtained at both locations were extremely high, with adjusted R^2 values reaching 0.99 and 0.902 respectively. However, the performance was mainly based on the neighbourhood size used for image segmentation. Support Vector Regression and Random Forest were shown to produce the best performing models for this study.

Biomass components and forest variables can also be estimated using LiDAR technologies. He et al. [17] estimated foliage, fruit, crown, stem, branch and AGB of a coniferous forest in western China using predictors extracted from LiDAR metrics. LiDAR metrics derived from CHM were then used to model the respective biomass components. The biomass equations used were a function of the DBH and total height for each respective component. The field measurements were used to compute the biomass of each of the components which were used to train Stepwise Regression models. Results from the regression produced adjusted R^2 values ranging from 0.749 to 0.356. Lower adjusted R^2 values were obtained from the biomass modelling of the smaller components such as foliage and fruit. Adjusted R^2 values for the total above ground biomass was 0.727. Zhao et al. [18] also utilized LiDAR data to extract forest stand parameters in the Dayekou forest, China using the Sparse Bayesian Regression model instead of typical regression techniques to offset the large amount of plot data usually required [19]. LiDAR data was used to extract 76 predictor variables based on height percentiles for the first and last pulse returns and individual mean heights and ratios of first and last pulses. The leave-one-out cross validation method was used to evaluate the performance of the models on predicting the mean height, average DBH, basal area and stand volume. The respective R^2 values obtained for each were 0.744, 0.720, 0.562, and 0.696.

Järnstedt et al. [20] compared the use of a photogrammetric surface model and ALS (airborne laser scanning) LiDAR data to estimate the forest variables diameter, mean height, basal area and volume of growing stock in Southern Finland. A high-resolution DSM was generated from the radiometric resolution of the aerial imagery combined with the TIN algorithm. Metrics extracted from both methods included the minimum, maximum, average and mode height values as well as the proportion of values from increasing percentile heights [20]. To reduce the dimensionality of the data, forward stepwise regression was done. The k-nearest neighbours estimation method along with the leave-one-out cross validation technique was used for variable estimation. The accuracy of the models was measured by the RMSE%, with comparable results being obtained from the photogrammetric model and the LiDAR model.

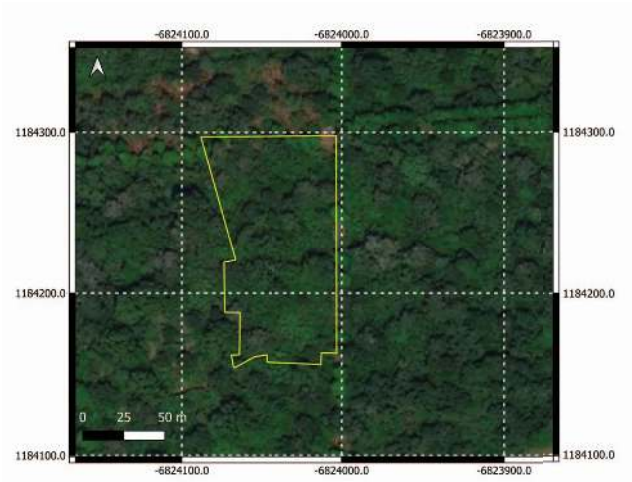

Fig. 1. Satellite view of study area, coordinate reference system.: WGS 1984, UTM projection Zone 20N

In Otsu City, Japan, Iizuka et al. [21] utilized unmanned drone imagery to collect aerial photos of the region. Photogrammetric processing was applied to generate DTMs and DSMs and extract the CHM. Laser technology was used to collect DBH and total height measurements for 51 individual samples to correlate with the aerial data. Regression analyses were used to estimate the tree height and DBH. The predictors, canopy area and canopy width, were found to predict DBH relatively well with R^2 values of 0.792 and 0.779 respectively. Tree height estimation performed poorly with a R^2 of 0.208 mainly due to density of the forested area and the inability to detect smaller trees from an aerial view.

Our contribution is the application of Machine Learning techniques using LiDAR data for Cocoa Plantations in a Small Island Developing State. Cocoa grows best in hilly terrain where it is difficult to take measurements manually. We believe that our results can be useful to other countries that seek to estimate carbon sequestration of their cocoa plantations.

III. Data and Method

The study area lies within a 100-acre region which makes up the International Cocoa Genebank (Figure 1) located in the La Chaguaramus Estate, Centeno, Trinidad. The Genebank consists of a large variety of some of the world's most diverse public collections of cocoa plant material. The different species of cocoa available at the genebank include *Theobroma cacao*, *Theobroma grandiflorum*, *Theobroma speciosum*, *Theobroma microcarpum* and *Theobroma obovatum* [22].

A. Field Data

Field data collection was conducted in Field 5B – Section F which spanned an area 0.9836 hectares. Field measurement data for 52 trees were established in the field at an accuracy of 1.5 m using Global Navigation Satellite Survey equipment. Measurements included data on the location, girth and height, of the trees. Within the area of interest, 957 cocoa and 37 shade trees were detected from the field survey.

B. LiDAR Data and Processing

The airborne LiDAR data was collected in 2014 and has now been curated by the Lands and Surveys Division. The extent of the LiDAR coverage is a 2 km^2 area spanning 685000, 687000, 1169500, 1170500 (xmin, xmax, ymin, ymax). The coordinate reference system used was that of WGS 1984 with a UTM projection, Zone 20N. The entire area consisted of 11.18 million points with a density of 5.6 points/m^2 and a pulse density of 2.7 pulses/m^2. The data was broken up into 8 blocks, each with an area of 250 m^2. Section F (Figure 1), the study area, contained 77.5 thousand points with a point density of 7.88 points/m^2 and a pulse density of 3.44 pulses/m^2.

Ground classification of the LiDAR point cloud was done using the Cloth Simulation Function (CSF) developed by Zhang et al. [23]. DTMs are used to model the ground elevation, slope and other hydrologic and environmental features and can be generated from in-situ measurements using Global Positioning Systems [24]. The invert distance weighting (IDW) method, which is one of the main methods used in spatial interpolation [25], was used to generate the DTM. To predict the value of an unsampled point, the weighted average of the known values within the defined neighbourhood [25] was used. The resolution was set to 1 square metre to create the model. The parameters, k and p representing the number of k-nearest neighbours and power respectively, were set to 10 and 2.

Next the point cloud was normalized by subtracting the DTM raster elevation from all the non-ground returns that were previously classified so that the elevations would be relative to the surface that represents the ground. With the normalized point cloud and elevations all being relative to the ground, a CHM representing the surface of the canopy was created using the point-to-raster algorithm [26]. Here the pixel size was set to 1 where Triangular Irregular Network processing was used to remove the occurrence of empty pixels and reduce edge artifacts.

To estimate the AGB, individual trees must be detected. This was done by applying a local maximum filter (LMF) to distinguish between individual treetops. Application of the LMF was done on the CHM raster using a fixed window size, of 2.5 m. This allowed for the evaluation of the neighbourhood points within a 1.25 m radius circle to determine the local highest point which would represent the treetop.

Due to the limited information available for the trees, individual tree segmentation was also done on the CHM raster using the *Dalponte2016* algorithm developed by Dalponte and Coomes [27]. This technique involves identifying individual tree crowns from airborne LiDAR data using a region-growing algorithm [27] which has the ability to identify individual tree species. The crown area definition and metrics were extracted for model creation. This was merged with the individual tree dataset containing (a) the height distribution, percentile and cumulative percentage variables, (b) the intensity distribution percentile and cumulative percentage variables and (c) the return metrics of the LiDAR data, to create the final dataset of 57 variables. Field measurements in Section F sampled 52 trees. These were correlated with the individual tree metrics based on height and location.

C. Feature Selection

Selecting the most relevant variables from the raw data plays a significant role in the accuracy of the model especially when there are many variables in the dataset. Feature selection reduces the number of variables by choosing those that are not correlated or biased. This not only increases the model performance, but also decreases the computational load. We investigated three feature selection methods, Boruta, Gini Importance and Stepwise Regression. They utilize both machine learning and statistical models to determine which variables have the best predictive power.

Boruta uses a wrapper approach built around the Random Forest machine learning algorithm [28]. Two iterations of the algorithm were able to reduce the number of independent variables by classifying them into two categories: important and unimportant.

Gini Importance, also known as Mean Decrease in Impurity (MDI), can be implemented in the Random Forest algorithm as a feature selection method. The feature importance for each attribute is calculated as the sum over the number of splits (across all trees) that include that feature, proportionally to the number of samples it splits [29]. Here, the pre-processed dataset was partitioned into 70% training and 30% testing sets ($n = 52$). The training partition was used with the Random Forest algorithm to create a model where each of the attribute's importance was evaluated. The attributes with values over 0.01 were used for model building.

Stepwise Regression is used when selecting explanatory variables for a Multiple Linear Regression model by adding and deleting variables based on their computed statistical values. A Linear Regression model using the pre-processed data was first created. Then, the Akaike Information Criterion (AIC) was used to determine the most significant variables.

D. Model Description and Validation

The datasets derived from each of the three feature selection methods as well as the full dataset, were used with multiple ML algorithms to determine which would perform best at predicting values for the DBH. The ML models included Multiple Linear Regression, Random Forest, Support Vector Regression and Regression Tree. The models were trained on 70% of the dataset and tested on 30% ($n = 52$).

We used a 10-Fold Cross Validation in our analysis, splitting the dataset into 10 equal subsets. This approach uses one subset at a time for testing while training is performed on

the remaining 9 subsets. The error for the test set is computed for each run and the average over the 10 runs is computed and reported. Errors were only computed for the test set samples (which were distinct from the corresponding training set samples for that run) and hence training set errors were not included in our results. Hence any over-fitting would potentially result in poor reported results. A more detailed explanation of why K-Fold Cross Validation detects over-fitting can be found in [30].

E. Above Ground Biomass Estimation

Allometric equations used in biomass estimation are useful since it allows the biomass for large areas such as forests and plantations to be estimated without having to use expensive, time-consuming or destructive methods. Allometric equations use metrics derived from the DBH and the height of sample trees to estimate the dry weight volume or biomass of the tree [31].

To estimate the AGB, the two best performing models were used to predict DBH values for the 957 trees detected from the LiDAR data. This comprised of the two main types of trees, cocoa and shade trees, providing metrics for both. Shade trees are significantly taller than cocoa tress and hence the trees could be categorized by height. Trees with a height greater than 9 m were classed as shade trees whilst those lower were classed as cocoa. Next, individual tree allometric equations were used to calculate the AGB for each of the cocoa and shade trees which were identified as part of the *Erythrina* species.

The allometric equation for cocoa trees was taken from Mustari et al. [5] and is given by $BK = 0.1208D^{1.98}$ where BK is the dry weight (kg/tree) and D is the DBH in centimeters. The allometric equation for the shade trees was taken from Rojas-García et al. [32] and is given by $AGB = 0.0433D^{2.3929}$.

IV. Numerical Results

Tree detection and segmentation done on the CHM identified 957 trees out of the 994 trees within the area. This resulted in a 96.28% detection rate. The total metrics obtained from the LiDAR data included 57 variables consisting of the crown area shown previously and the 56 standard metrics associated with elevation (variables starting with z), intensity (variables starting with i), and returns (variables starting with p). Each feature selection method produced its own set of variables deemed significant. From the 57 variables, Boruta reduced the number to 7, Gini Importance reduced the number to 12 and Stepwise Regression reduced the number to 8. These are provided in Table I.

Four ML algorithms, with 10-fold cross validation, were tested with and without each of the feature selection methods described previously. These algorithms, Multiple Linear Regression, Random Forest, Support Vector Regression and Regression Tree, were used to predict the values for the DBH of the trees. Hyperparameter tuning was done on the Random Forest, Support Vector Regression and Regression Tree ML

TABLE I
SUMMARY OF FEATURES SELECTED FOR EACH METHOD

Feature Selection	Variables
Boruta	zmax, zsd, zq80, zq85, zq90, zq95, zpcum1
Gini Importance	zsd, zmax, zkurt, zq90, zq95, zq80, zpcum1, zpcum6, zpcum7, zpcum8, zpcum9, imean
Stepwise Regression	zmax, p2th, p4th, imean, zpcum6, zpcum4, p1th, zq35

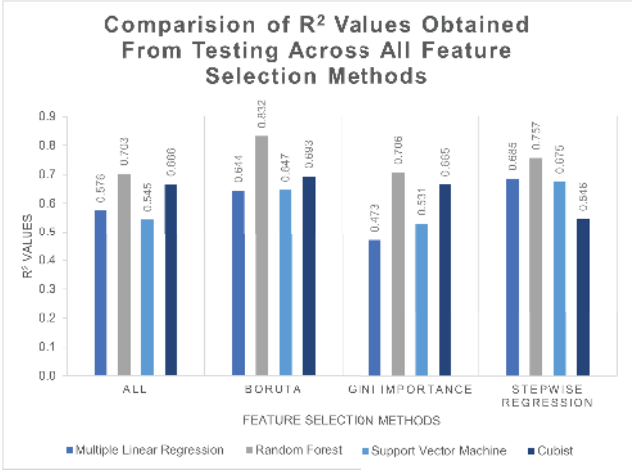

Fig. 2. Comparison of R^2 values for test samples for each model

algorithms. We computed two performance metrics, the R^2 metric and the Root Mean Square Error (RMSE) metric both with 10-fold cross-validation.

These metrics are defined as follows. If we assume N test samples and denote the set of true values by y_i and the corresponding predicted values (using the training set) by p_i then

$$\mathsf{RMSE} \equiv \left(\frac{1}{N} \sum_{i=1}^{N} (y_i - p_i)^2 \right)^{\frac{1}{2}} \tag{1}$$

and

$$\mathsf{R}^2 \equiv 1 - \frac{\sum_{i=1}^{N} (y_i - p_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{2}$$

where $\bar{y}$ represents the average over all target values of the training samples. The R^2 values for the test samples (with 10-fold cross validation) are provided in Figure 2. The RMSE values are provided in Table II.

Of all the feature selection methods used, the variables selected from Boruta performed best with testing based on the results of the ML models applied. Each of the algorithms used obtained relatively low RMSE values when compared to the same models using a different set of features. The two best performing models were Random Forest using Boruta feature selection (highest performing on the testing data) and the Random Forest using stepwise regression (second lowest

TABLE II
RMSE VALUES FOR TESTING SAMPLES FOR EACH MODEL

Algorithm	All	Boruta	Gini Importance	Stepwise Regression
Multiple Linear Regression	1.242	0.080	0.095	0.072
Random Forest	0.075	0.062	0.072	0.066
Support Vector Machine	0.087	0.076	0.087	0.074
Regression Tree	0.074	0.072	0.075	0.086

TABLE III
BIOMASS ESTIMATES USING THE TWO BEST PERFORMING MODELS

Feature Selection	Algorithm	AGB (Mg/ha)	% difference
Boruta	Random Forest	28.75 $\pm$2.34	-14.69
Stepwise Regression	Random Forest	28.10 $\pm$2.27	-16.62

RMSE). These results were then used to predict the DBH values for each of the 957 trees detected. The respective allometric equations were then applied to the cocoa and shade trees to obtain the AGB for each. The generic forestry allometric equation was applied to the data from the 994 trees identified in the satellite survey of the area to compare the accuracy of the estimations. This is given by the formula: $\mathsf{VOB} = \mathsf{WD} \times \mathsf{BEF}$, where VOB is the volume from buttress to crown, WD is the volume-weighted average wood density and BEF is the biomass expansion factor set at 1.0989. The total above ground biomass obtained using each of the models are provided in Table III.

A raster image with a resolution of 10 square metres (Figure 3) was used to display the estimated AGB values from the best model, the Random Forest using Boruta variables. The individual tree values were summed for each 10 square metre pixel.

V. DISCUSSION

The aim of the study was to combine LiDAR data with small sample field measurements to get an approximation of the tree variable, DBH, which in turn, would be used in AGB estimation. The area of interest constitutes a small section of cocoa and shade trees that form part of the 0.4 km^2 cocoa estate. Point classification showed that much of the vegetation observed from the LiDAR data were assigned to either low or medium vegetation with a few points within the area of interest being classed as high vegetation. This corresponded with the field survey data which identified 957 cocoa trees and 37 shade trees of which the latter was of a significantly greater height than that of the cocoa.

For CHM generation the original Point-to-Raster CHM produced some empty pixels, which was one of the drawbacks of this algorithm. Further processing to fill the output was

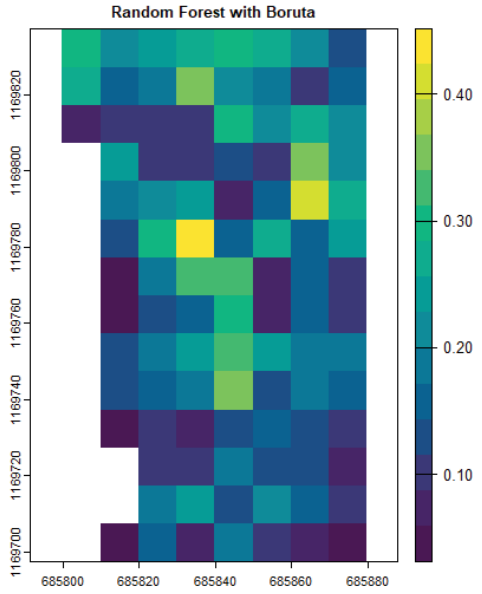

Fig. 3. 10 m $\times$ 10 m raster image showing spatial distribution of estimated AGB: Values are in Mg/ha

done using TIN to get a better model. However, some edge effects were still seen using this method likely due to the irregular shape of the area clipping. Tree density within the area was relatively high, with approximately 1000 trees per hectare. The high density proved problematic when detecting and segmenting individual trees within the point cloud and caused issues when correlating the field measurements to the detected treetops. The number of trees detected by processing LiDAR data did not exactly match with the numbers obtained from the field survey (96.28% detection) and such would result in some underestimation of the total AGB values as was observed. Edge effects, from the clipping of the area of interest, caused points from trees to be cut off leading to some of the segmented trees not being representative of the ones on the ground. Inaccuracies in modelling the data may also result from this. The sample data had sparse observations for trees greater than 7 metres and, as such, training on this height range would be limited.

The four ML algorithms, Multiple Linear Regression, Random Forest, Support Vector Regression and Regression Tree were applied with the three feature selection methods as well as with all 57 variables to predict the DBH for all the trees. The RMSE and R^2 values on a 10-fold cross validation were used as the main metrics to assess the performance of the models. The 10-fold cross validation was used because of the small training set available and reduced the likelihood of overfitting. This method divided the training data into 10 equal subsamples in which for each of the 10 iterations, one of the subsamples would be used for testing while the other 9 are used for training. The sum of the errors for each of the 10 test

sets was found and the average computed. In this approach only the error for the test sets was computed. These are distinct from the corresponding training set for that iteration and such any overfitting of the model would be reflected in the results obtained.

Figure 4 shows the distribution of the errors obtained for each feature selection method with each ML algorithm. No outliers were detected in any of the MLg algorithms used in the Boruta and Stepwise Regression feature selection models. The models that utilized the variables chosen from the Gini Importance feature selection method was seen to perform the worst overall with large errors. Here, the distribution of the errors in the Multiple Linear Regression model was large, whilst in the other models, outliers were detected. The errors were very large in the Multiple Linear Regression model using all the variables as well, with some predicted values obtaining an error as large as +3.6 m. Taking a closer look at both feature selection methods, Gini Importance and all features, it was observed that these used a large number of independent variables to generate the models. This large number of variables may have resulted in over complicated models that overfit the random noise in the training data and so performed poorly on the testing data. Random Forest and Regression Tree ML algorithms still performed relatively well, and this was likely due to each of these having their own feature selection process within the algorithm that would choose the best features for the model.

The R^2 value for the Random Forest using Boruta was 0.832 with an RSME of 0.062 m. This meant that 83.2% of the variation in the DBH was explained by the model and that the average size of the error when estimating DBH was around 0.062 m. Similarly, for the Random Forest using Stepwise Regression, R^2 was 0.757 and RSME was 0.066 m. The distribution of the errors (Figure 4) in these models were seen to be small when compared to the other models. The RMSE values (Table II) and the high R^2 values obtained from the Random Forest models from the Boruta and Stepwise Regression feature selection models, provide sufficient evidence that these were the best performing models from the study.

The predicted DBH values were used with generalized allometric equations based on the climate and area of the forest. The AGB estimates from allometric modelling done in similar latitudinal areas were found to be between 25.52 and 54 Mg/ha. The AGB estimate computed using the generic allometric equation was found to be 33.7 Mg/ha. Estimations from the models created in the study were: 28.10 Mg/ha and 28.75 Mg/ha using the Random Forest algorithm with Stepwise Regression and Boruta variables resulting in a -16.62% and -14.69% difference respectively. The Random Forest model using Boruta variables performed best, with the smallest percentage difference from the AGB estimate. These results fall within the range of previous studies done using Random Forest to estimate AGB such as in [15]. This study produced two models that resulted in differences of 16.13% and 18.44% from the AGB value obtained from ground truth

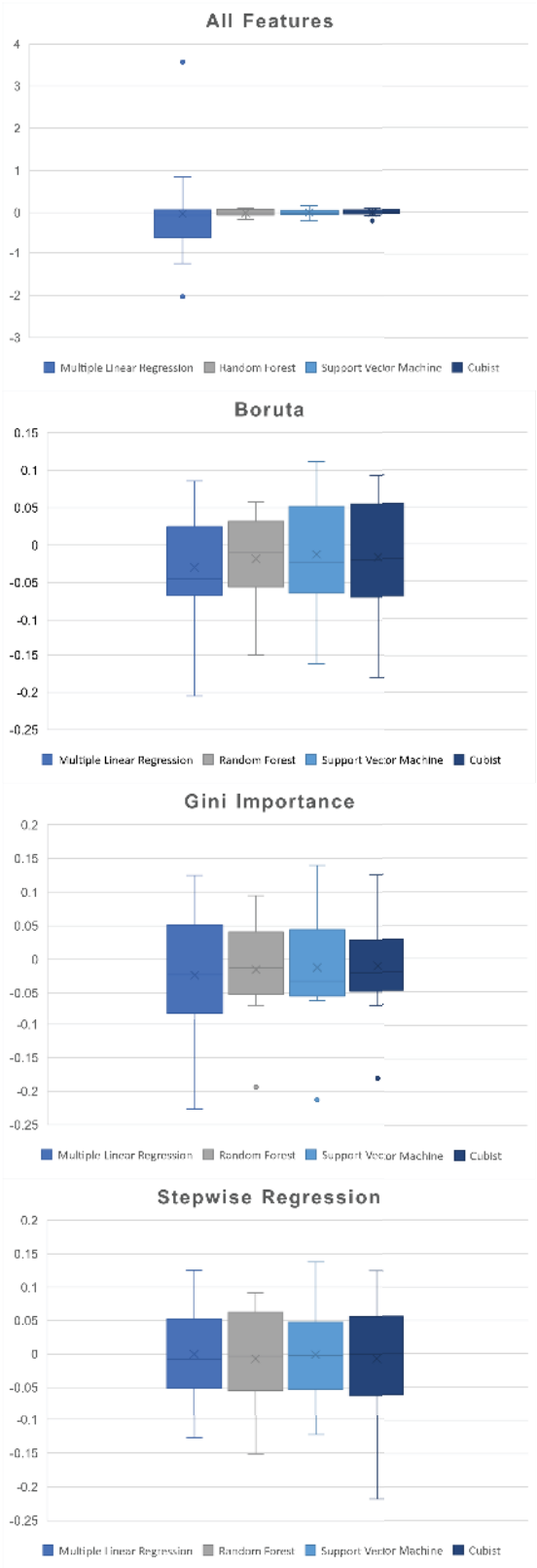

Fig. 4. Distribution of errors for each Feature Selection and Model (in metres)

measurements.

The results also show that, although there are some caveats with using Stepwise Regression techniques, model performance is still very good and comparable with other feature selection methods. Improvements to the study include testing different shapes such as a circle or rectangle to reduce edge effects for better processing of the DTM and CHM. A more in-depth look at the shade trees within the area of study would also be required to determine if tree species, other than the *Erythrina*, comprised the shade tree population. The use of higher density LiDAR would aid in increasing the detection and segmentation of individual trees as the point density of that used was relatively low at 7.88 points/m^2. Increasing the number of samples used for training, as well as, having sample tree measurements within the different height ranges that better represent the distribution of trees in the area would greatly improve the performance of the models.

VI. Conclusion

The tree variable of concern, diameter at breast height, was estimated using descriptive metrics derived from LiDAR. Due to the high density of trees within the study area, tree detection accuracy was marginally lower, however, it was able to detect approximately 96% of the trees. Feature selection methods, Boruta and Stepwise Regression were shown to improve the performance of the ML models created when compared to the models created using all of the features. The estimated biomass values from the two best performing models were close to that of the value obtained using the generic Allometric equation. The proposed approach is less costly and more efficient than traditional methods especially when large areas must be assessed. It is our intention to carry out additional experiments in order to improve upon the results obtained from this research so that we may test our methodology throughout the twin island state.

References

[1] X. Zheng, D. Streimikiene, T. Balezentis, A. Mardani, F. Cavallaro, and H. Liao, "A review of greenhouse gas emission profiles, dynamics, and climate change mitigation efforts across the key climate change players," *Journal of Cleaner Production*, vol. 234, pp. 1113–1133, 2019.

[2] U. EPA, "Overview of greenhouse gases," 2014.

[3] A. Buis, "The atmosphere: Getting a handle on carbon dioxide–climate change: Vital signs of the planet," *Nasa*, 2020.

[4] Z. Kayler, M. Janowiak, and C. Swanston, "Global carbon," *US Department of Agriculture, Forest Service, Climate Change Resource Center. https://www. fs. usda. gov/ccrc/topics/global-carbon*, 2017.

[5] K. Mustari, L. Asrul, L. Faradilla *et al.*, "Carbon stock analysis of some cocoa planting systems in south sulawesi," in *IOP Conference Series: Earth and Environmental Science*, vol. 486, no. 1. IOP Publishing, 2020, p. 012085.

[6] L. Asrul, "Cocoa agribusiness," 2013.

[7] E. Somarriba, R. Cerda, L. Orozco, M. Cifuentes, H. Dávila, T. Espin, H. Mavisoy, G. Ávila, E. Alvarado, V. Poveda *et al.*, "Carbon stocks and cocoa yields in agroforestry systems of central america," *Agriculture, ecosystems & environment*, vol. 173, pp. 46–57, 2013.

[8] S. Chauhan and C. R. Ritu, "Carbon sequestration in plantations. agroforestry for increased production and livelihood security," 2016.

[9] S. Abdoellah, "Co2 absorptionemission balance in cocoa plantation," *Prosiding Simposium Kakao 2008*, 2008.

[10] C. Jewell, "Breathing new life into trinidad and tobago's cocoa sector," *WIPO MAGAZINE*, 2017.

[11] "Revitalizing the Cocoa Industry - InvesTT and EU Delegation Make Strides — Trinidad and Tobago Government News — news.gov.tt," http://www.news.gov.tt/content/revitalizing-cocoa-industry-investt-and-eu-delegation-make-strides, [Accessed 19-Oct-2022].

[12] G. Patenaude, R. Hill, R. Milne, D. Gaveau, B. Briggs, and T. Dawson, "Quantifying forest above ground carbon content using lidar remote sensing," *Remote sensing of environment*, vol. 93, no. 3, pp. 368–380, 2004.

[13] E. P. Y. Chan, T. Fung, and F. K. K. Wong, "Estimating above-ground biomass of subtropical forest using airborne lidar in hong kong," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[14] Y. Gao, D. Lu, G. Li, G. Wang, Q. Chen, L. Liu, and D. Li, "Comparative analysis of modeling algorithms for forest aboveground biomass estimation in a subtropical region," *Remote Sensing*, vol. 10, no. 4, p. 627, 2018.

[15] L. Torre-Tojal, A. Bastarrika, A. Boyano, J. M. Lopez-Guede, and M. Graña, "Above-ground biomass estimation from lidar data using random forest algorithms," *Journal of Computational Science*, vol. 58, p. 101517, 2022.

[16] B. Xue, *Lidar and Machine Learning Estimation of Hardwood Forest Biomass in Mountainous and Bottomland Environments*. University of Arkansas, 2015.

[17] Q. He, E. Chen, R. An, and Y. Li, "Above-ground biomass and biomass components estimation using lidar data in a coniferous forest," *Forests*, vol. 4, no. 4, pp. 984–1002, 2013.

[18] P. Zhao, L. Gao, and T. Gao, "Extracting forest parameters based on stand automatic segmentation algorithm," *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[19] V. Junttila, M. Maltamo, and T. Kauranne, "Sparse bayesian estimation of forest stand characteristics from airborne laser scanning," *Forest Science*, vol. 54, no. 5, pp. 543–552, 2008.

[20] J. Järnstedt, A. Pekkarinen, S. Tuominen, C. Ginzler, M. Holopainen, and R. Viitala, "Forest variable estimation using a high-resolution digital surface model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 74, pp. 78–84, 2012.

[21] K. Iizuka, T. Yonehara, M. Itoh, and Y. Kosugi, "Estimating tree height and diameter at breast height (dbh) from digital surface models and orthophotos obtained with an unmanned aerial system for a japanese cypress (chamaecyparis obtusa) forest," *Remote Sensing*, vol. 10, no. 1, p. 13, 2017.

[22] A. Kennedy and V. Mooledhar, "Conservation of cocoa in field genebanks-the international cocoa genebank, trinidad," in *International Workshop on Conservation, Characterisation and Utilisation of Cocoa Genetic Resources in the 21st Century, Port-of-Spain (Trinidad and Tobago), 13-17 Sep 1992*. West Indies Univ. Cocoa Research Unit, 1993.

[23] W. Zhang, J. Qi, P. Wan, H. Wang, D. Xie, X. Wang, and G. Yan, "An easy-to-use airborne lidar data filtering method based on cloth simulation," *Remote sensing*, vol. 8, no. 6, p. 501, 2016.

[24] G. T. Raber, J. R. Jensen, S. R. Schill, and K. Schuckman, "Creation of digital terrain models using an adaptive lidar vegetation point removal process," *Photogrammetric engineering and remote sensing*, vol. 68, no. 12, pp. 1307–1314, 2002.

[25] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Computers & geosciences*, vol. 34, no. 9, pp. 1044–1055, 2008.

[26] J.-R. Roussel, D. Auty, N. C. Coops, P. Tompalski, T. R. Goodbody, A. S. Meador, J.-F. Bourdon, F. De Boissieu, and A. Achim, "lidr: An r package for analysis of airborne laser scanning (als) data," *Remote Sensing of Environment*, vol. 251, p. 112061, 2020.

[27] M. Dalponte and D. A. Coomes, "Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data," *Methods in ecology and evolution*, vol. 7, no. 10, pp. 1236–1245, 2016.

[28] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of statistical software*, vol. 36, pp. 1–13, 2010.

[29] A. Perrier, *Effective Amazon machine learning*. Packt Publishing Ltd, 2017.

[30] A. W. Moore, "Cross-validation for detecting and preventing overfitting," *School of Computer Science Carneigie Mellon University*, 2001.

[31] P. W. West, *Tree and forest measurement*. Springer, 2015.

[32] F. Rojas-García, B. H. De Jong, P. Martínez-Zurimendí, and F. Paz-Pellat, "Database of 478 allometric equations to estimate biomass for mexican trees and forests," *Annals of forest science*, vol. 72, no. 6, pp. 835–864, 2015.