Journal of Sustainability Science and Management
Volume 18 Number 12, December 2023: 95-110
eISSN: 2672-7226

MACHINE LEARNING PREDICTION OF TROPICAL FOREST ABOVE-GROUND BIOMASS ESTIMATION

NUR ILYANI MOHD ZULKIFLEE[1], NURUL AIN MOHD ZAKI[1, 5]*, TAJUL ROSLI RAZAK[3], HAMDAN OMAR[7], SHAJOERIL TAJUDIN[6], ROHAYU HARON NARASHID[1], MOHD NAZIP SURATMAN[3] AND ZULKIFLEE ABD LATIF[2,5]

[1]School of Geomatics Science and Natural Resources, College of Built Environment (CBE), Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia. [2]School of Geomatics Science and Natural Resources, College of Built Environment (CBE), Universiti Teknologi MARA, Selangor Branch, Puncak Alam Campus, 40450 Puncak Alam, Selangor, Malaysia. [3]School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Selangor Branch, Puncak Alam Campus, 40450 Puncak Alam, Selangor, Malaysia. [4]Faculty of Applied Sciences, Universiti Teknologi MARA, Selangor Branch, Puncak Alam Campus, 40450 Puncak Alam, Selangor, Malaysia. [5]Institute for Biodiversity and Sustainable Development (IBSD), Universiti Teknologi MARA, Selangor Branch, Puncak Alam Campus, 40450 Puncak Alam, Selangor, Malaysia. [6]VTS Universe Sdn Bhd, 106-2, Jalan LP 7/4, Taman Lestari Perdana, 43300 Seri Kembangan, Selangor, Malaysia. [7]Geoinformation Programme, Division of Forestry & Environment, Forest Research Institute Malaysia (FRIM), Kepong, Selangor, Malaysia.

**Corresponding author: nurulain86@uitm.edu.my*
Submitted final draft: 7 November 2023 *Accepted: 18 November 2023* http://doi.org/10.46754/jssm.2023.12.009

Abstract: Forests play a significant role as forest sources and have been commonly used to measure carbon stocks within the international carbon cycle and biomass of the forest. Land biomass is an essential element in determining the carbon and carbon balance capabilities of forest ecosystems. This study aimed to estimate forest biomass carbon stocks from the field, Airborne LiDAR, and WorldView-3 data using an Artificial Neural Network and Random Forest. In total, 245 observations and five variables including independent variables, the total height of the tree measured in field (hF), diameter at breast height (DBH), height extracted from Lidar (hL), crown projection area (CPA) and dependent variables (CS) at which based on the data used, multiple regression has been carried out to estimate the forest carbon stocks. ANN has been tested with different hidden layers by trying and error and for Random Forest, two parameters which are the number of randomly picked variables for each node of the tree (*Mtry*) and the number of trees to grow (*Ntree*), which was 500 have been used in this study. The best model obtained from both methods was used to generate the carbon stocks map prediction. This study result shows that Model 5 of the ANN algorithm obtains (RMSE = 92.248 Mg ha^{-1} and R^2 = 0.916). From this study, RF can be concluded as the best model that can be used for the estimation of biomass and carbon stocks as for this study as Model 3 of RF shows the lowest error compared to ANN (RMSE = 49.417 Mgha^{-1} and R^2 = 0.976) and the effectiveness of R as the best model for biomass estimation has been proven from the previous research.

Keywords: LiDAR, WorldView-3, Artificial Neural Network (ANN), Random Forest (RF), Machine Learning (ML).

Introduction

The tropical forest is viewed as a huge biodiversity focal point and stores roughly 40% of the Earth's general carbon stock (Metzker *et al.,* 2012). It plays a huge part in the worldwide carbon cycle, which represents 30-40% of net essential earthly yield (Clark *et al.,* 2001). Likewise, the significant effect of the tropical backwoods on the carbon cycle can be seen through the high primary production rate and wide value of pools and flux volumes. Therefore, the estimation of the biomass of the woods is significant for the storing of the carbon spending plan, monitoring the flux of carbon, and comprehension of forest ecosystems' reaction to the changing climate (Nandy *et al.,* 2019). It is important to perceive the likely job of different carbon isolation pools in limiting the collection of air CO_2, neighbourhood, provincial, and

public carbon inventories of vehicle sources and sinks, just as on solid land breaking measures in forestalling a worldwide temperature alteration (Salunkhe *et al.,* 2018). With rising awareness of the important ecological resource of the forest environment, humans have understood that biomass, amongst the most crucial parameters, and not only significantly related to wood products but also directly linked to global carbon storage and cycle (Lu, 2005).

However, modification of the land use of tropical, in particular deforestation and degradation of the forest, has contributed to 12-20% of the global emissions of greenhouse gases (GHG) throughout the last two decades (Harris *et al.,* 2012); thus, the tools for addressing climate change are reforestation, afforestation and preventing deforestation (Luong *et al.,* 2015) is needed. Assessing forest biomass and carbon conservation is not only to reduce deforestation and carbon emissions corruption but also to control practical forest areas (Hussin *et al.,* 2014). Greenwood biomass is a significant component for the planning of carbon and carbon stream observation, including a dangerous atmospheric deviation examination. Along these lines, the improvement of a dependable strategy to appraise the biomass of the woods and carbon stock becomes fundamental (Dang *et al.,* 2019). Remotely sensed data that has been combined with greenwood listing data for the evaluation of AGB and, ultimately, carbon stocks has become one of the effective solutions. Worldwide activities, including the decrease of contamination due to logging and timberland corruption (REDD) and REDD+, have been set up to push the meaning of tree biomass in carbon emission stability and energise a more prominent comprehension of carbon emission reduction (Olander *et al.,* 2008).

The United Nations Cooperation Program on REDD (UN-REDD) has proposed that greenwood assets identification frameworks ought to incorporate the utilisation of Remote Sensing (RS) stock advances for carbon stock resource evaluation, tree species monitoring, and forest degradation estimation (Kushwaha *et al.,* 2014). Satellite data of Remote Sensing is accessible in any scope of scales, from nearby to worldwide, and from any customised. Independent data types, such as an example data of optical, radar data, and LiDAR data, also exist, each of which has its advantages over the others (Kumar & Mutangga, 2017). To promote REDD+ (reduction of logging and woodland corruption, long-term forest protection, and improvement of woodland carbon stocks) methods, accurate overland forest biomass (AGB) is significant for keeping up woodland the executives and lessening a dangerous atmospheric deviation (Chen *et al.,* 2018).

The advancement of the ML technique has given established researchers a range of valuable tools to get another comprehension of the worldly and spatial varieties of different carbon streams in earthbound environments (Dou, Yang & Luo, 2018). ML methods have been widely utilised throughout the most recent twenty years and have been managing the various issues associated with carbon motion gauges (Huang & Hsieh, 2020). Location and the type of allometric equations gained from calculating the parameter of forest biometrics, such as diameter at breast height (DBH), height, crown closure, and stem density, are currently the most effective methods in obtaining the aboveground biomass of the forest (Chave *et al.,* 2014; Paul *et al.,* 2016). Nevertheless, data on the biomass is frequently out of date when it is used because of the time taken to obtain field data (Chave *et al.,* 2014).

In this research, an ML approach is being utilised to evaluate the aboveground biomass and carbon stock using the data LiDAR data that was collected in August 2013 and also WorldView-3 data of 9 December 2015 with spatial resolution 0.30 m (panchromatic), 1.2 m (multispectral) and super-spectral high resolution of 25 km^2. This study aims to estimate the aboveground biomass (AGB) and carbon stock from the field, Airborne LiDAR, and WorldView-3 data using a machine learning approach at which the data has been collected from Ayer Hitam Forest Reserve, Puchong in Selangor. To achieve the

aim, the objectives are to (1) identify the data of aboveground biomass and carbon stock for Ayer Hitam Forest Reserve in Selangor, (2) to develop the carbon stock estimation using a machine learning approach, and (3) to produce an aboveground biomass carbon stock map. To accomplish the objective, this software which is open-source R will be used to calculate the estimated value of biomass and carbon stock, and ArcGIS software which is remote sensing software, will be used to produce the outcome which is an aboveground biomass map and carbon stock map.

Materials and Method

Study Area

The research was conducted at 3°00′24.19" N, 101°38′25.24" in the Ayer Hitam Forest Reserve in Selangor State, Malaysia (Mohd Zaki *et al.,* 2018). The Ayer Hitam Reserve, located in Puchong, falls under the category of lowland Dipterocarp forests. It is referred to as an auxiliary upset backwood, as it has undergone a few rounds of logging and treatment since the 1930s (Syafinie & Ainuddin, 2013). The Ayer Hitam Forest Reserve was designated as a woodland-safe route in 1906 and covers 4,270 hectares. The reserve experiences a range of temperatures, with a minimum of 22.7°C and a maximum of 32.1°C. The average temperature in the area is 26.6°C (Syafinie & Ainuddin, 2013). The evolving canopy stand is approximately 20 metres above ground level. The secondary layers are 12 to 16 metres above the ground and saplings and seedlings are part of the lower canopy. Based on the map (a) shows the peninsular Malaysia, (b) shows the Selangor state, and (c) shows the study area, which was Ayer Hitam Forest Reserve, Selangor, Malaysia.

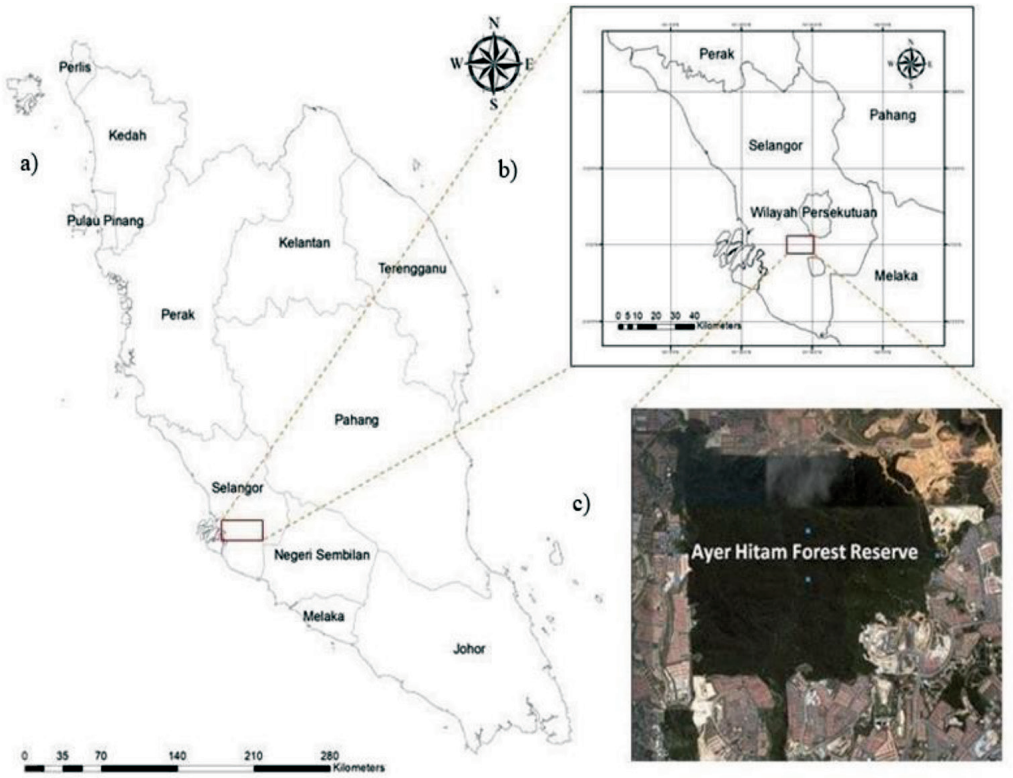

Figure 1: The location of the study area which is Ayer Hitam Forest Reserve

Research Methodology

The four phases in this research include data collection, pre-processing, OBIA processing and machine learning processing. For the first phase which is data collection, three types of data will be involved: field data, Airborne LiDAR Data, and WorldView-3 data. For field data, it involves two (2) main types of data which are diameter breast height (DBH) and height of the tree (HT). Airborne LiDAR data involved with 11 point/m2 and for WorldView-3 data are 0.3 m of multispectral and 1.2 m of panchromatic. The second phase involves the pre-processing of the data at which the calculation of AGB using Chave *et al.* (2014) equation and calculation of carbon stocks value from field data. LiDAR data is involved with the process to generate CHM and WorldView-3 data is involved with georeferencing the data with GCP. Images obtained from LiDAR and WV3 have been fused.

The third phase of this study is OBIA processing. This phase involves the process of generating height and CPA using the fusion image obtained from the processing of phase two. The fourth phase is the machine learning process. During this process, open R software will be used to estimate the data of carbon stock that have been obtained from previous calculations during phase two by using the Artificial Neural Network (ANN) algorithm and Random Forest (RF). This process involves five variables, including one independent variable (Carbon Stocks) and four dependent variables (hF, DBH, hL, and CPA). In order to carry out model fitting using the training dataset and prediction of the dependent variable using the testing dataset, the data has been split into training and testing sets. The model validation has been conducted (RMSE, MAE, MSE, R^2) and the best model with lower error was used for the production map of aboveground carbon stock using ArcGIS Software.

Parameters of the Study

This study used parameters from a previous study made by Mohd Zaki *et al.* (2018). The parameter used has been proven accurate by previous research and further study has been done using machine learning using the same parameter. The selection of the parameters from the previous studies depends on the availability of the data in the study area (Table 1).

Estimating Aboveground Biomass and Carbon Stocks

Aboveground biomass data has been calculated using an allometric equation according to Chave *et al.* (2014) [1]. The data used included ρ is the wood density (g cm^{-3}), DBH in cm, and the total height of the tree (h) in m (Mohd Zaki *et al.,* 2018). This calculation was used for AGB computation because it functioned well over a wide range of forest types and bioclimatic states (Chave *et al.,* 2014).

$$AGBest = 0.0673(\rho(DBH)^2h)0.976 \quad (1)$$

The carbon value was calculated or transformed by implementing an aspect of 0.47 which represents 47% of the dry biomass concluded to be carbon for all parts of the tree as the default value that has been recommended by IPCC (IPCC, 2006; Mohd Zaki *et al.,* 2018).

Table 1: List of the parameters used in the study

Author & Year	Parameter Used
(Mohd Zaki *et al.,* 2018)	• hF = total height of tree measured in the field • DBH = diameter at breast height • hL = height extracted from Lidar • CPA = crown projection area

Machine Learning Processing

Data Splitting

The training set was used to discover the link between dependent and independent variables, while the test set was used to evaluate the model's performance. 70% of the dataset has been used as the training set and 30% as the testing set. Random sampling, sample() function was used in the data training and testing to perform random sampling. Also, the set.seed() function with starting point 12345 has been used to produce the same random sample every time and keep consistency. Note that the number 12345 in this paper demonstrates that the seed number we choose is the starting point for generating a sequence of random numbers. Given a seed number of 12345, we will obtain the same results. To build training and test data sets, an index variable has been used when fitting a neural network. Different between ANN and RF processes before proceeding to model fitting, data normalisation needs to be carried out when using the ANN algorithm. If the data is not normalised, the predicted value will frequently be the same across all observations, regardless of the input values. Min-max normalisation was selected to standardise the data utilised in this investigation.

Model Fitting

A neural Network was fitted to the data using the "neuralnet" library for analysis. Using the training data, a neural network was formed. The dependent variable is "regressed" against the other independent variables using "neuralnet". Given that the impression of the independent factors on the dependent variable (dividend) is anticipated to be non-linear, the linear output variable is set to FALSE. The number of hidden layers in a neural network is not a precise science. In fact, without any hidden layers, accuracy is likely to be higher in some cases. As a result, trial and error are crucial in this process. Two types of hidden layers were used to construct the optimal model for this investigation: one hidden layer (1, 2, 3) and two hidden layers (c(2,2), c(5,1), c(5,2). The hidden layer selected has been tested to achieve the lowest error.

Random Forest was fitted to the data using the "randomForest" library for analysis. In this stage, the *Ntree* value is generated by the computer system that will give the best *Ntree* value to be used for further process. The plotting of the OOB error for the *Ntree* value that has been used was necessary to show which *Ntree* obtain can generate the best *Mtry* value. During the fit model, the *Mtry* was calculated using the formula used by López-Serrano *et al.* (2020) in the previous research ($m = \sqrt{P}$, $m = P/3$, $m = P$), where P represents the number of independent variables. *Mtry* = 1, *Mtry* = 2, and *Mtry* = 4 have been used and the validation of each model test with different *Mtry* will be compared to obtain the best model to be used for carbon stocks prediction.

The randomForest package includes two indices for measuring variable significance, which is the percentage increase in mean square error (%IncMSE) that has been tabulated from permuting OOB data, and the total decrease in node impurities from splitting on the variable (IncNodePurity), which is calculated from splitting on the variable (Liaw & Wiener, 2018; Li *et al.,* 2020; Nguyen & Kappas, 2020). Inflated %IncMSE values imply a better significant predictor. According to Strobl *et al.* (2007), the IncNodePurity approach is biased and should not be used. As a result, in this study, we exclusively employ the percentage IncMSE metric to determine the relevance of factors.

Prediction of Carbon Stocks

Prediction is the process of predicting the independent variables using a testing dataset. The prediction has been carried out based on the fit model process before to show that the prediction has been made using the targeted model and for this study, the ANN and RF model has been chosen. The prediction process was important to generate the predicted value of the dependent variables and also the predicted model was necessary to be used for the model validation process to be carried out.

Model Validation

RF and ANN performances were evaluated in the training and validation phases. Accordingly, the techniques were used to predict the carbon stock in the data set intended for model validation (Dantas *et al.,* 2021). Evaluating the model accuracy is an essential part of the process of creating machine learning models to describe how well the model is performing in its predictions. The MSE, MAE, RMSE, and R-squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis (Han *et al.,* 2019).

a. MAE (Mean Absolute Error)

Represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \quad (2)$$

b. MSE (Mean Squared Error)

Represents the difference between the original and predicted values extracted by squaring the average difference over the data set.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (3)$$

c. RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2} \quad (4)$$

Represents the error rate by the square root of MSE.

d. R-squared (Coefficient of Determination)

Represents the coefficient of how well the values fit compared to the original values. The values from 0 to 1 are interpreted as percentages. The higher the value is, the better the model.

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad (5)$$

Map Production

For map production, the best model that obtains less error was used for the prediction of carbon stocks from each model. From the result obtained, Model 5 of ANN shows the lowest error obtained (RMSE = 92.248 Mg/ ha and R^2 = 0.916) compared to another model. From the RF process, the lowest error was from Model 3 that have to obtain the best parameter result, *Mtry* = 4 with the accuracy of RMSE = 49.417 Mg/ha and R^2 = 0.976. Map of carbon stocks (kg/tree) using ANN and RF predicted results have been made using ArcGIS software.

Table 2: Descriptive statistic of data used

Variable and Unit	N	Min	Max	Mean	SD
Height from Lidar (hL)	245	10.851	37.822	20.696	5.149
Height from field (hF)	245	10.000	37.000	20.210	5.167
Diameter at breast height(DBH)	245	10.000	113.000	28.496	14.264
Crown projection area	245	7.514	214.283	30.679	22.177
Above-ground biomass (AGB)	245	32.000	17167	761.094	1383.355
Carbon Stocks (in kg)	245	15.000	8068	357.686	650.179

(N = number of trees; Min = minimum; Max = maximum; and SD = standard deviation)

Results and Analysis

Descriptive Statistics

The descriptive analysis in this study has been conducted using Excel of the observed data for this study which are the height of field (hF), diameter breast height (DBH), the height of Lidar (hL), canopy height model (CPA), above-ground biomass (AGB) and carbon stock.

The data in this study were divided into two dependent variables (aboveground biomass and carbon stock) and independent variables (hF, DBH, hL, and CPA). All the variable has 245 data. From the result mentioned above, the mean value of AGB was 761.094 with a standard deviation of 1,383.355 while the minimum and maximum of AGB were 17167 and 761.094, respectively. For the carbon stock, the maximum and minimum values were 8068 and 15.000 (Mean = 357.686, SD = 650.179). For the independent variable, the maximum and minimum height from the field were 37.000 and 10.000 (Mean = 20.210, SD = 5.167). For the height of Lidar (hL), the maximum and minimum values were 37.822 and 10.851 (Mean = 20.696, SD = 5.149). Other than that, the DBH maximum and minimum values were 113.000 and 10.000 (Mean = 28.496, 14.264). Lastly, the maximum and minimum values of CPA were 214.283 and 7.514 (Mean = 30.679, 22.177).

Machine Learning Model

Model Validation of Multiple Regression Model

Four error measurements, namely the coefficient of determination (R^2), the root mean square error (RMSE), the mean square error (MSE), and the mean absolute error (MAE), were used

Table 3: Model validation of 1 hidden layer

No.	Model Candidates	One Hidden Layer	MAE	MSE	RMSE	R^2	Data Sources
1	hF +DBH + hL + CPA	1	76.120	13359.31	115.582	0.868	Field, LiDAR, WV3
2	hF +DBH + hL + CPA	2	99.875	18327.71	135.380	0.907	Field, LiDAR, WV3
3	hF +DBH + hL + CPA	3	97.189	18708.56	136.779	0.815	Field, LiDAR, WV3

hF = total height of tree measured in field; DBH = diameter at breast height; hL = height extracted from lidar; CPA = crown projection area; WV3 = WorldView-3.

Table 4: Model validation of 2 hidden layers

No.	Model Candidates	Two Hidden Layer	MAE	MSE	RMSE	R^2	Data Sources
4	hF +DBH + hL + CPA	c(2,2)	66.091	9475.341	97.341	0.906	Field, LiDAR, WV3
5	hF +DBH + hL + CPA	c(5,1)	63.021	8509.641	92.248	0.916	Field, LiDAR, WV3
6	hF +DBH + hL + CPA	c(5,2)	64.304	8861.855	94.137	0.912	Field, LiDAR, WV3

hF = total height of tree measured in field; DBH = diameter at breast height; hL = height extracted from lidar; CPA = crown projection area; WV3 = WorldView-3.

to evaluate the performance of the model. In general, a higher R^2 value and lower RMSE values indicate a better estimation performance of the model (Li, Li, & Liue, 2020; López-Serrano *et al.,* 2020).

Model Validation of Artificial Neural Network

Tables 3 and 4 show the model validation or accurate assessment of the model using the Artificial Neural Network Algorithm in Rstudio software. Prediction model accuracy was assessed using standard validation indices such as MAE, MSE, RMSE, and R^2 (John *et al.,* 2020; López- Serrano *et al.,* 2020). Two types of hidden layers have been tested to obtain the accuracy value to estimate the carbon stocks which was by using 1 hidden layer and 2 hidden layers.

Based on the table, Model 5 shows the lowest accuracy (RMSE = 92.248 Mg ha^{-1} and R^2 = 0.916) obtained by using 2 hidden layers (c(5,1)) followed by Model 6 with (RMSE = 94.137 Mg ha^{-1} and R^2 = 0.912). The less accurate result based on the table was Model 3 with (RMSE = 136.779 Mg ha^{-1} and R^2 = 0.815). From the table, all the accuracy assessments of the model that used 1 hidden layer showed higher error compared to when using 2 hidden layers. We can conclude that using 2 hidden layers for prediction was better than using 1 hidden layer (Thomas *et al.,* 2017).

Model Validation of Random Forest

Table 5 shows the results of the three models that have been processed using a random forest algorithm in Rstudio software. The model differs in accuracy assessment based on the Mtry value calculated using a formula tested of (m = √P), m = P/3, m = P at which the P is the number of independent variables (López-Serrano *et al.,* 2020).

Based on the table, Model 3 with a *Mtry* value of 4, shows the best accuracy assessment (RMSE = 49.417 Mg ha^{-1} and R^2 = 0.976), followed by Model 2 (RMSE = 56.426 Mg ha^{-1} and R^2 = 0.968) and lastly Model 1 (RMSE = 67.431 Mg ha^{-1} and R^2 = 0.955). All the models presented are slightly less different as the entire model was trained using 4 variables which were low in variable number, so they do not show over or underfitting in the results (Nguyen & Kappas, 2020).

Evaluation of the Best Value of ANN and RF Model

In this study, two machine learning approaches (ANN and RF) were used to estimate the

Table 5: Model validation of random forest model based on *Mtry* value

No.	Model Candidates	*Mtry*	*Ntree*	MAE	MSE	RMSE	R^2	Data Sources
1	hF +DBH + hL + CPA	1	500	38.899	4546.896	67.431	0.955	Field, LiDAR, WV3
2	hF +DBH + hL + CPA	2	500	31.012	3183.853	56.426	0.968	Field, LiDAR, WV3
3	hF +DBH + hL + CPA	4	500	27.907	2442.020	49.417	0.976	Field, LiDAR, WV3

hF = total height of tree measured in field; DBH = diameter at breast height; hL = height extracted from lidar; CPA = crown projection area; WV3 = WorldView-3.

Table 6: Best model validation of ANN and RF

No.	Algorithm	R^2	MAE	RMSE
5	ANN	0.916	63.021	92.248
3	RF	0.976	27.907	49.417

biomass carbon stocks of Ayer Hitam Forest Reserve by integrating field data with multiple satellite data, which were LiDAR data and WorldView-3 data. Three error measurements, namely the coefficient of determination (R^2), the root mean square error (RMSE) and the parentage root mean square (RMSE%) were used to evaluate the performance of the model. In general, a higher R^2 value and lower RMSE values indicate a better estimation performance of the model (Li, Li, & Liue, 2020, López-Serrano *et al.,* 2020).

The validation result indicates that Model 3 based on the Random Forest (RF) algorithm provides the most accurate assessment with an R2 value of 0.976 and a lower RMSE of 49.417 Mg ha^{-1}. In contrast, Model 5 based on Artificial Neural Network (ANN) algorithm produces an R2 value of 0.916 with the highest RMSE of 92.245 Mg ha^{-1}. These two best models were further used to estimate the carbon stocks of the study area for map production.

The performance of the model that used the RF and ANN methods was further analysed. Scatter plots of measured forest carbon stocks against predicted data based on field data and multiple imagery data LiDAR and WorldView-3 data used were generated in Rstudio based on the different predictions used. For model ANN, the plotting is generated based on the hidden layer used, while for the RF model, the plotting is generated according to the Mtry value (shown in Figures 2 and 3). Based on the figure, the distribution of scatter points is concentrated near 1:1. At which the $Radj^2$ values are ranged (0.768 – 0.927) for the ANN model and (0.966 – 0.994) for the RF model.

Plotting Graph of the Model

For this study, a regression scatter plot has been generated according to the model used. Running a regression model in machine learning will not be manually calculated instead based on the simple scripts model run in ML. Regression is a parametric technique used to predict continuous (dependent) variables given a set of independent variables (Saraswat, 2016). Two types of scatter plots have been generated for this study.

(i) Observed versus predicted graph

This type of graph is a common and simple approach to evaluating models. The plotting plot is a scatter plot of predicted on the y-axis and observed values on the x-axis.

(ii) A predicted versus residuals graph

This type of plot is a scatter plot of residuals on the y-axis and predicted value on the x-axis. A residual value is a measure of how much the regression line vertically misses a data point. A residual plot is typically used to find problems with regression. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data. The plot is used to detect non-linearity, unequal error variances, and outliers.

Plotting Graph Observed Versus Predicted

Figure 2 and Figure 3 show the graph of observed carbon stocks versus predicted carbon stocks of ANN and RF. All the scatter plot ANN algorithms (Model 1, Model 2, Model 3, Model 4, Model 5, Model 6) and RF algorithm (Model 1, Model 2, Model 3) present a situation where all observations except the outlier fall around a straight-line statistical relationship. All the model shows a very strong tendency for observed and predicted to both rise above their means or fall below their means at the same time.

The straight line is designed to come as close as possible to all the data points. The trend line has a positive slope, which shows a positive relationship between observed and predicted carbon stocks. The points in the graph were tightly clustered about the trend line due to the strength of the relationship between the observed and predicted R^2 range for ANN (0.868-0.916) and R^2 range for RF (0.955-0.976) at which the R^2 value obtain from all the models near to 1.

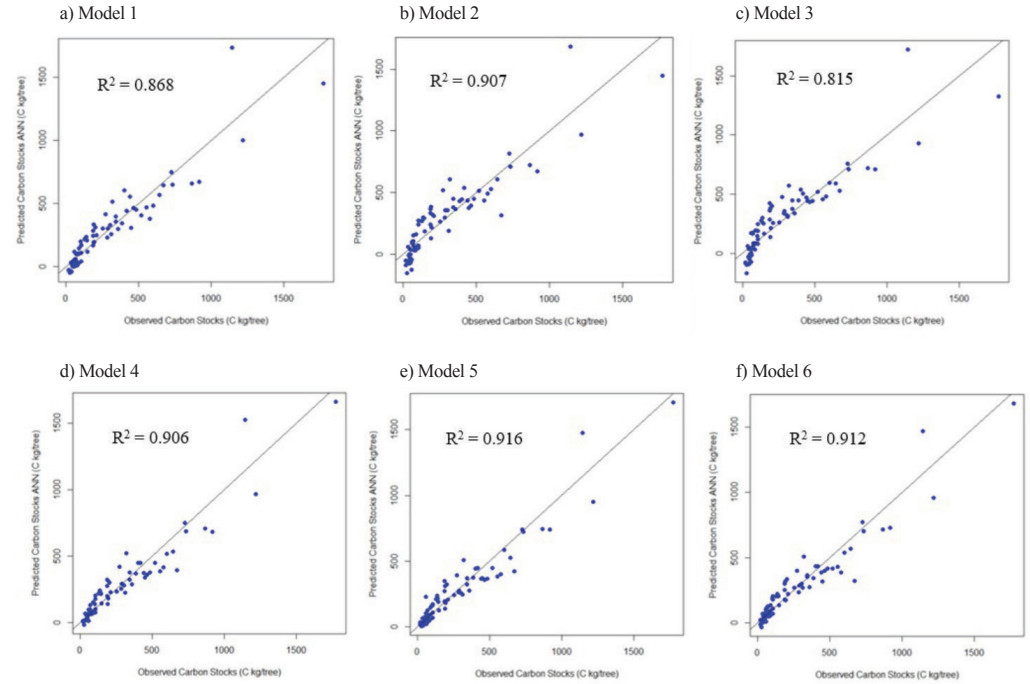

Figure 2: Observed carbon stocks versus predicted carbon stocks of ANN for validation data set (n = 74)

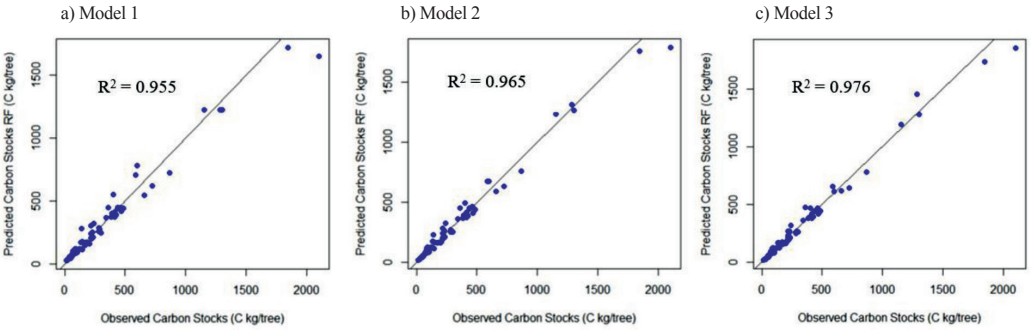

Figure 3: Observed carbon stocks versus predicted carbon stocks of RF for validation data (n = 74)

Plotting Graph Predicted Versus Residuals

Whether a linear regression function is appropriate for the data being analysed can be studied from the residual plot against the predicted variable. The nonlinearity of the regression function can be studied from a scatter plot but is not as effective as a residual plot. Figure 4 and Figure 5 show the scatter plot of the data and the fitted regression line for a study of the relationship between predicted carbon stocks and the residuals of the estimation (observed carbon stocks – predicted carbon stocks) of ANN and RF.

All the model (Model 1, Model 2, Model 3, Model 4, Model 5 and Model 6) of ANN and model (Model 1, Model 2, Model 3) of RF shows the residuals plot that was randomly dispersed around the horizontal axis at which shows that the linear regression model was appropriate for this data. The scatter plot shows that both models of ANN and RF contain outliers. Outliers are

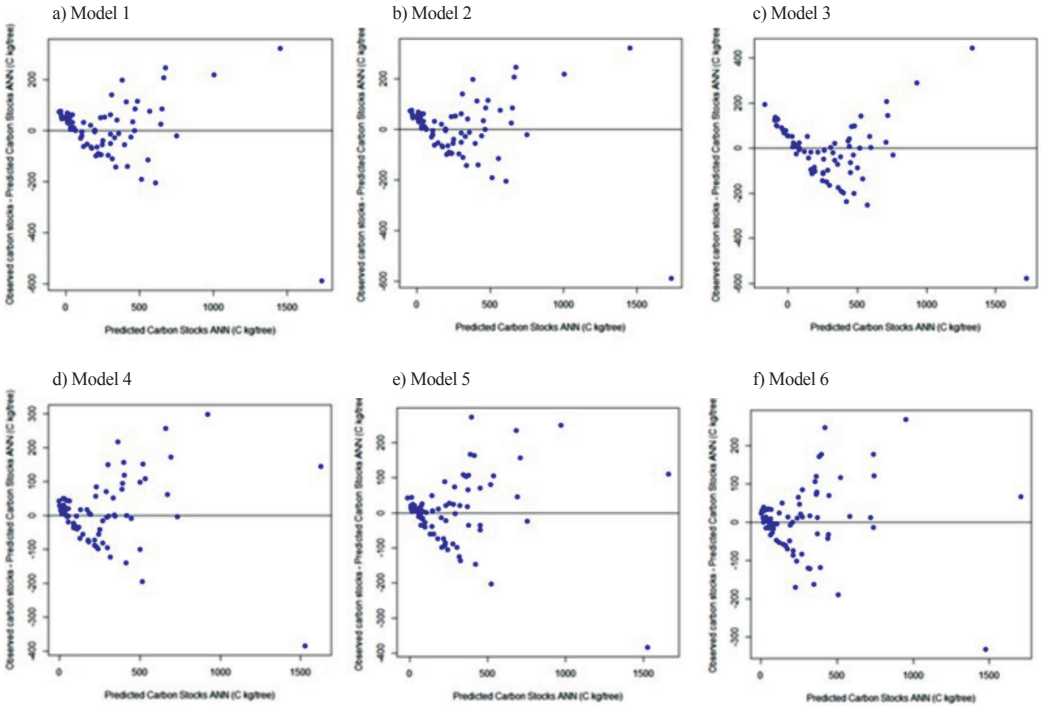

Figure 4: Predicted carbon stocks versus the residuals of the estimation (observed carbon stocks – predicted carbon stocks) of ANN for validation data (n = 74)

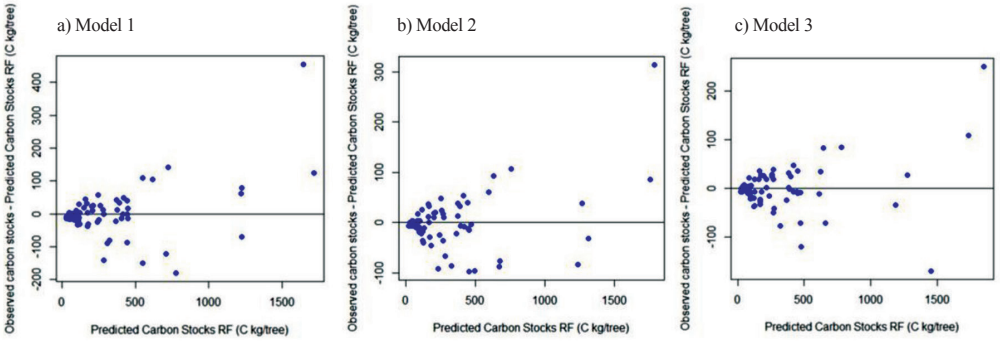

Figure 5: Predicted carbon stocks versus the residuals of the estimation (observed carbon stocks – predicted carbon stocks) of RF for validation data (n = 74)

extreme observations. Residual outliers can be identified from residual plots against X or Y.

The Comparison of Models

Likewise, earlier studies have used these two algorithms to predict forest biomass and attain fine accuracies, while the results of the model's contrast are discrete compared with this study. Exploration regarding the effectiveness of two machine learning models in estimating the carbon stocks of Ayer Hitam Forest Reserve has shown acceptable accuracy.

In this study, the Random Forest model performed best with a higher R^2 = 0.976 and lower RMSE = 49.417 Mg ha^{-1} compared to the Artificial Neural Network with accuracy (R^2 = 0.916 and RMSE = 92.248 Mg ha^{-1}), which was similar to Chen *et al.* (2018), who found

that the accuracy of the RF model was the best (R^2 = 4.43 and RMSE = 0.999 Mg ha^{-1}). Aside from that, Cao *et al.* (2018) discovered that RF was the best (R^2 = 0.9, RMSE = 13.4 Mg ha^{-1}), followed by ANN. Nguyen and Kappas (2020) conducted a previous study utilising RF to estimate aboveground biomass and obtained a value (R^2 = 0.74 and RMSE = 61.24 Mg ha^{-1}).

According to Geng *et al.* (2021), the RF model beat the ANN models, and the RF model was suggested in prior research because of its resilience and accuracy. Furthermore, Han *et al.* (2019) demonstrated that the RF model can resist overfitting and address high-dimensional data (Geng *et al.,* 2021). However, Gao *et al.* (2018) found that ANN represents (RMSE = 27.6 Mg ha^{-1}) better than RF in a relative comparison of algorithms for forest AGB forecast using ALOS PALSAR and Landsat data. According to Vahedi (2016), ANN is being used instead of traditional procedures to forecast AGB in natural forest ecosystems. Nandy *et al.* (2017) estimated biomass using ANN and obtained (R^2 = 0.74 and RMSE = 93.41 Mg ha^{-1}).

In this investigation, the RF models had the highest accuracies when compared to the ANN models. This could be related to the trivial sample sizes used for investigation, as well as the invariable random placement of samples in the research area, which is comparable to the study conducted by Geng *et al.* (2021). Even though the error value of ANN in this study is rather high, it is still lower than in earlier research by Nandy *et al.* (2017) and is still deemed the best for biomass estimation. For this investigation, only 245 observations were used, and five variables were employed, which may have limited the processing of machine learning to take place. In comparison to prior studies, the accuracy estimate for biomass carbon stocks is acceptable for future research purposes.

Findings of the Study

This study used a multiple regression method to estimate the biomass carbon stocks at which the value of the variable (hF, DBH, hL, and CPA) was fitted in the model together to predict the independent variable that might influence the increasing value of the validation model. In addition, this study only focuses on one type for every model used as an example. Only an ANN feed-forward backdrop has been used for this study which cannot show the bigger difference in error value to estimate better carbon stocks using the ANN model. Compared to previous research that has been carried out for the same study purpose, Ercanli *et al.* (2016) used four types of ANN which were ANN based on the feed-forward backdrop, based on the Elman backdrop, based on Layer Recurrent, and based on NARX and the accuracy obtained was lower in error and high value of the coefficient of determination. The method from the previous study is different from this study that only focuses on using only one type of ANN using all the variables (hF, DBH, hL, and CPA) to generate a model, and based on the model, different parameters of every model used in this study been try and error to find the best model for prediction of the carbon stocks.

In this study, set.seed( ) value has been used in a certain stage of the process at which its function is to generate a sequence of the random number instead of data generated without sequence. This is to make sure that we get the same result when running the scripts according to the sequences. Not setting the set.seed( ) will lead to the prediction that obtains differences in the arrangement of data or overfitting model when using more than one model for the same purpose and it can be considered as the not valid comparison between the model used, especially the research that includes the objective to make a comparison between the model used. From previous research regarding biomass estimation, there is still no researcher that states the use of the set seed to generate sequences of random numbers used for the prediction that fixes in accuracy assessment obtained however the use of the set.seed is quite popular in other field research as an example of nutrition rating prediction (Hou, 2018).

For this study, splitting data based on sample estimation to split training and testing data have been done to train the model used and to predict

the independent variable using test set data and validate the data using error formula (RMSE, MAE, MSE, R^2) instead of using 10-fold cross-validation method. 10-fold cross-validation is a method to assess estimating models by splitting the actual data into a training set to train the model and test set. The effectiveness of 10-fold cross-validation can be seen in a previous study by Nguyen and Kappas (2020) and Dang *et al.* (2019), in which the validation value obtained shows lower error. The difference in the method used for the model validation might lead to the less accurate model validation for this study.

Conclusion

In conclusion, all the objectives of this study were successfully achieved. All the results obtained answer all the objectives of this study. Several parameters have been used to conduct

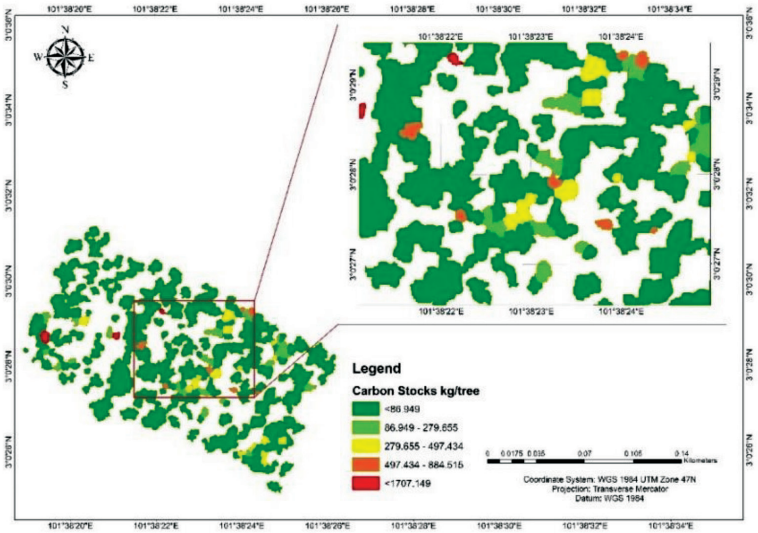

Figure 6: Predicted carbon stocks map using ANN model

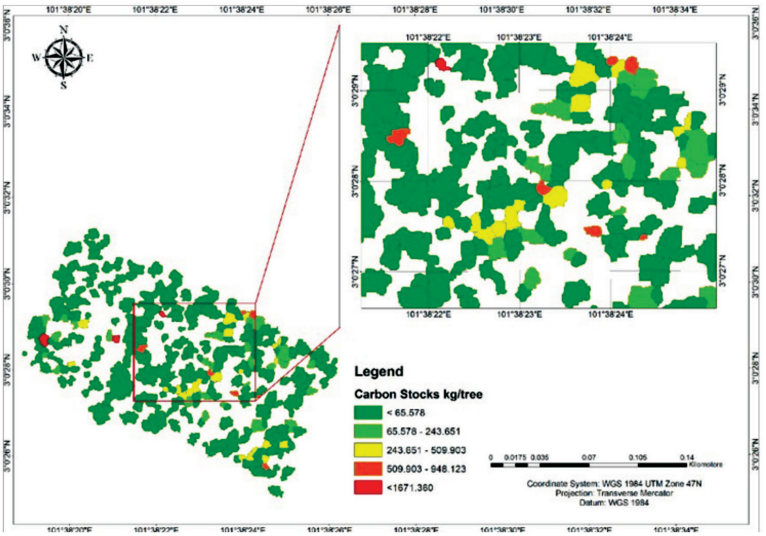

Figure 7: Predicted carbon stocks using the RF model

this research from a previous study made by Mohd Zaki *et al.* (2018). The first objective is to compose the aboveground biomass estimation for Ayer Hitam Forest Reserve in Selangor based on the previous equation Chave *et al.* (2014). All the data used have been carried out with descriptive statistics to describe the basic features of the data in the study. For objective two at which to construct the carbon stock estimation using an Artificial Neural Network (ANN) and Random Forest (RF). The result obtained from the processing shows that Model 5 of ANN (RMSE = 0.916 Mg ha^{-1}, R^2 = 92.248) and Model 3 of RF (RMSE = 0.976 Mg ha^{-1}, R^2 = 49.417) show an accurate result compared to another model. The best model from ANN and RF have been used for further process, which for prediction of the aboveground carbon stocks and the value has been used to produce a map as the final result for objective three, to produce a map of Ayer Hitam Forest Reserve based on carbon stock estimation. From this research, RF shows the best algorithm for the estimation of aboveground carbon stocks and the effectiveness of the RF algorithm has been proven by previous research (Chen *et al.,* 2018; Cao *et al.,* 2018; Nguyen & Kappas, 2020; Mohd Zaki *et al.,* 2022).

Acknowledgements

The authors would like to express their gratitude to Universiti Teknologi MARA for their invaluable support and the funding of industrial grants 100-TNCPI/PRI 16/6/2 (070/2022). Thank you to Forest Research Institute Malaysia (FRIM) for providing the remote sensing data and access to the study area.

References

Cao, L., Pan, J., Li, R., Li, J., & Li, Z. (2018). Integrating airborne LiDAR and optical data to estimate forest aboveground biomass in arid and semi-arid regions of China. Remote Sensing, 10(4), 532.

Chave, J., Réjou–Méchain, M., Búrquez, A., Chidumayo, E., Colgan, M. S., Delitti, W. B., & Vieilledent, G. (2014). Improved allometric models to estimate the aboveground biomass of tropical trees. *Global Change Biology*, *20*(10), 3177-3190.

Chen, L., Ren, C., Zhang, B., Wang, Z., & Xi, Y. (2018). Estimation of forest above-ground biomass by geographically weighted regression and machine learning with sentinel imagery. *Forests*, *9*(10), 582.

Clark, D. A., Brown, S., Kicklighter, D. W., Chambers, J. Q., Thomlinson, J. R., & Ni, J. (2001). Measuring net primary production in forests: Concepts and field methods. *Ecological Applications*, *11*(2), 356.

Dang, A. T. N., Nandy, S., Srinet, R., Luong, N. V., Ghosh, S., & Senthil Kumar, A. (2019). Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. *Ecological Informatics*, *50*(February 2019), 24–32.

Dou, X., Yang, Y., & Luo, J. (2018). Estimating forest carbon fluxes using machine learning techniques based on eddy covariance measurements. *Sustainability* (Switzerland), *10*(1), 1– 26.

Ercanli, İ., Günlü, A., Şenyurt, M., Bolat, F., & Kahriman, A. (2016). Artificial neural network for predicting stand carbon stock from remote sensing data for even-aged scots pine (*Pinus sylvestris* L.) Stands in the taşköprü-çiftlik forests. *Forest Engineering and Technologies FETEC 2016*, *170*.

Gao, Y., Lu, D., Li, G., Wang, G., Chen, Q., Liu, L., & Li, D. (2018). Comparative analysis of modeling algorithms for forest aboveground biomass estimation in a subtropical region. *Remote Sensing*, *10*(4), 627.

Geng, L., Che, T., Ma, M., Tan, J., & Wang, H. (2021). Corn biomass estimation by integrating remote sensing and long-term observation data based on machine learning techniques. *Remote Sensing*, *13*(12), 2352.

Han, L., Yang, G., Dai, H., Xu, B., Yang, H., Feng, H., ... & Yang, X. (2019). Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data. *Plant Methods*, *15*(1), 1–19.

Harris, N. L., Brown, S., Hagen, S. C., Saatchi, S. S., Petrova, S., Salas, W., Hansen, M. C., Potapov, P. V., & Lotsch, A. (2012). Baseline map of carbon emissions from deforestation in tropical regions. *Science*, *336*(6088), 1573–1576.

Hussin, Y. A., Gilani, H., van Leeuwen, L., Murthy, M. S. R., Shah, R., Baral, S., & Qamer, F. M. (2014). Evaluation of object-based image analysis techniques on very high-resolution satellite image for biomass estimation in a watershed of hilly forest of Nepal. *Applied Geomatics*, *6*(1), 59–68.

Huang, I., & Hsieh, C. (2020). Gap-filling of surface fluxes using machine learning algorithms in various ecosystems. *Water*, *12*(12), 1-24. https://www.mdpi.com/2073-4441/12/12/3415

Hou, J. (2018). Simple neural network for nutrition rating prediction. *Amazonaws*. https://rstudio-pubs-static.s3.amazonaws.com/390906_bbf1500b7cc4473c83b29fd14c0951e8.html

IPCC. (2006). *Guidelines for national greenhouse gas inventories* (Vol. 4). Agriculture, Forestry and Other Land Use. http://www.ipcc-nggip.iges.or.jp/public/2006gl/

John, K., Abraham Isong, I., Michael Kebonye, N., Okon Ayito, E., Chapman Agyeman, P., & Marcus Afu, S. (2020). Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land*, *9*(12), 487.

Kumar, L., & Mutanga, O. (2017). Remote sensing of above-ground biomass. *Remote Sensing*, *9*(9), 935.

Kushwaha, S. P. S., Nandy, S., & Gupta, M. (2014). Growing stock and woody biomass assessment in Asola-Bhatti Wildlife Sanctuary, Delhi, India. *Environmental Monitoring and Assessment*, *186*(9), 5911–5920.

Liaw, A. & Wiener, M. (2012). *Random forest: Breiman and cutler's random forests for classification and regression*. R Package Version 4.6-7. http://cran.r-project.org/web/packages/randomForest/

Li, Y., Li, M., Li, C., & Liu, Z. (2020). Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Scientific Reports*, *10*(1), 1–12.

López-Serrano, P. M., Cárdenas Domínguez, J. L., Corral-Rivas, J. J., Jiménez, E., López-Sánchez, C. A., & Vega-Nieva, D. J. (2020). Modeling of aboveground biomass with Landsat 8 OLI and machine learning in temperate forests. *Forests*, *11*(1), 11.

Lu, D. (2005). Aboveground biomass estimation using Landsat TM data in the Brazilian Amazon. *International Journal of Remote Sensing*, *26*(12), 2509–2525.

Luong, N. V., Tateishi, R., Hoan, N. T., & Tu, T. T. (2015). Forest change and its effect on biomass in Yok Don National Park in Central Highlands of Vietnam using ground data and geospatial techniques. *Advances in Remote Sensing*, *4*(2), 108–118.

Metzker, T., Spósito, T. C., Filho, B. S., Ahumada, J. A., & Garcia, Q. S. (2012). Tropical forest and carbon Stock's valuation: A monitoring tropical forest and carbon stock's valuation: A monitoring policy. In *Biodiversity enrichment in a diverse world*. InTechOpen.

Mohd Zaki, N. A., Latif, Z. A., & Suratman, M. N. (2018). Modelling above-ground live trees biomass and carbon stock estimation of tropical lowland Dipterocarp Forest: Integration of field-based and remotely sensed estimates. *International Journal of Remote Sensing*, *39*(8), 2312–2340.

Mohd Zaki, N. A., Asri, A. M., Zulkiflee, N. I. M., Latif, Z. A., Razak, T. R., & Suratman, M. N. (2022). Assessment of forest aboveground biomass estimation from Superview-1 satellite image using machine learning approaches. In C*oncepts and applications of remote sensing in forestry*. Springer. https://doi.org/10.1007/978-981-19-4200-6_1

Nandy, S., Singh, R., Ghosh, S., Watham, T., Kushwaha, S. P. S., Kumar, A. S., & Dadhwal, V. K. (2017). Neural network-based modelling for forest biomass assessment. *Carbon Management*, *8*(4), 305-317.

Nandy, S., Ghosh, S., Kushwaha, S. P. S., & Kumar, S. (2019). *Remote sensing of Northwest Himalayan ecosystems*. Springer.

Nguyen, D. & Kappas, M. (2020). Estimating the aboveground biomass of an evergreen broadleaf forest in Xuan Lien Nature Reserve, Thanh Hoa, Vietnam, using SPOT-6 data and the random forest algorithm. *International Journal of Forestry Research*, *2020*, 1–13. https://doi.org/10.1155/2020/4216160

Olander, L. P., Gibbs, H. K., Steininger, M., Swenson, J. J., & Murray, B. C. (2008). Reference scenarios for deforestation and forest degradation in support of REDD: A review of data and methods. *Environmental Research Letters*, *3*(2), 025011.

Paul, K. I., Roxburgh, S. H., Chave, J., England, J. R., Zerihun, A., Specht, A., Lewis, T., Bennett, L. T., Baker, T. G., Adams, M. A., Huxtable, D., Montagu, K. D., Falster, D. S., Feller, M., Sochacki, S., Ritson, P., Bastin, G., Bartle, J., Wildy, D., & Sinclair, J. (2016). Testing the generality of above-ground biomass allometry across plant functional types at the continent scale. *Global Change Biology*, *22*(6), 2106–2124.

Saraswat, M. (2016, December 6). Beginners guide to regression analysis and plot interpretations. *HackerEarth*. https://www.hackerearth.com/blog/developers/beginners-guide-regression-analysis-plot-interpretations/

Salunkhe, O., Khare, P. K., Kumari, R., & Khan, M. L. (2018). A systematic review on the aboveground biomass and carbon stocks of Indian forest ecosystems. *Ecological Processes*, *7*, 17. https://doi.org/10.1186/s13717-018-0130-z

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 1-21.

Syafinie, A. M., & Ainuddin, A. N. (2013). Biomass and carbon estimation of Eugeissona tristis. *Sains Malaysiana*, *42*(10), 1461–1466.

Thomas, A. J., Petridis, M., Walters, S. D., Gheytassi, S. M., & Morgan, R. E. (2017, August). Two hidden layers are usually better than one. In *International Conference on Engineering Applications of Neural Networks* (pp. 279-290). Springer, Cham.

Vahedi, A. A. (2016). Biomass and Bioenergy Artificial neural network application in comparison with modeling allometric equations for predicting above-ground biomass in the Hyrcanian mixed-beech forests of Iran. *Biomass and Bioenergy*, *88*, 66–76.