Mapping and Estimating Forest Stand Volume using Machine Learning Methods and Multi-Spectral Sentinel 2 Data

Nover M. Matso

Department of Forestry, Abra State Institute of Sciences and Technology, Philippines

Heherson B. Ong

College of Environmental Management and Forestry, Isabela State University, Philippines

Emerson V. Barcellano

College of Environmental Management and Forestry, Isabela State University, Philippines

Abstract:

Sustainable forest management necessitates the mapping and estimation of forest stand attributes such as density, volume, basal area, and aboveground biomass. This study was conducted to explore the potential of geographic information systems (GIS), remote sensing, machine learning, and field inventories to estimate the forest stand volume of natural and plantation forests within watersheds in the Abra River Basin. The common machine learning regression techniques, which are random forest (RF), k-nearest neighbors (KNN), and support vector machines (SVM), were used to model and predict forest stand volume. The validation of the three machine learning methods showed that the best model to estimate and map forest stand volume is the RF algorithm ($R^2 = 0.42$, RMSE = 0.40 m$^3$/plot, MAE = 0.31 m$^3$/plot). Topographic variables such as the Digital Elevation Model (DEM) and the spectral band Near Infrared (NIR) were the most important variables in predicting forest stand volume. The estimated forest stand volume using the RF model ranged from 33 to 115 m$^3$/ha, with a mean of 59 m$^3$/ha. The results of this study revealed that forest volume can be measured using freely available satellite data and machine learning techniques.

Keywords: forest stand volume, geographic information system, machine learning, remote sensing, satellite images, sustainable forest management.

Introduction

Sustainable management of forests is an urgent need because it results in sustainable development, achievement of internationally agreed development goals, poverty eradication (Aerts and Honnay, 2011), food security, biodiversity conservation, and climate change mitigation (Food and Agriculture Organization (FAO), 2020). The necessity of updated forest resource inventories required for the assessment of forest stand characteristics is needed in support of sustainable forest management (Wulder et al., 2008). Forest resources assessment provides the foundation of forest planning and forest policy (White et al., 2016) and also describes the state and change of forest ecosystems (Gschwantner et al., 2022). Forest inventory is important for assessment and analysis (Dau and Chukwu, 2018), for monitoring (Henry et al., 2021), as a tool in

This work is licensed under a Creative Commons Attribution 4.0 International License. The license permits unrestricted use, distribution, and reproduction in any medium, on the condition that users give exact credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if they made any changes.
decision-making (Kangas and Maltamo, 2006), and to obtain information on the multifunction of forests (Fridman et al., 2014). The lack of reliable and sufficient information on forest resources can lead to poor management decisions, which in turn result in poor outcomes for ecosystem health and human well-being (Modzelewska et al., 2016; Lister et al., 2020). Traditionally, forest inventories involved complete enumeration (Kangas and Maltamo, 2006). However, this method is time-consuming, laborious, and costly (Ronoud et al., 2019).

The use of remotely sensed data has recently revolutionized the way forest resource assessments are conducted. Recent advances in GIS, remote sensing technology, artificial intelligence (AI), and the Internet of Things (IoT) enable the rapid, up-to-date, and dependable extraction of information on forest features (Fan et al., 2018; Ahmadi et al., 2020). Remote sensing can provide consistent, reproducible, and up-to-date data on multiple forest attributes to make forest assessment more efficient (Ganz et al., 2019). Remotely sensed data could be utilized instead of more costly ground observations and measurements (McRoberts and Tomppo 2007). The availability of free and up-to-date remotely sensed data has led researchers to explore its potential application in forest resource management, especially in the measurement of forest stand characteristics. In the Philippines, Truckenbrodt (2013), Divina II et al. (2015), Buitre et al. (2019), Dida et al. (2021), and Doyog et al. (2021) made use of LiDAR and Landsat data for land cover classification and assessment. Lumbres and Lee (2014), Pillodar et al. (2017), and Baloloy et al. (2018) used remotely sensed data to estimate aboveground biomass.

The estimation of forest stand volume is done by developing regression models from remotely sensed data in combination with field measurements. Individual spectral bands and spectral indices (vegetation indices) extracted from satellite images are the common input variables in developing the regression models. Modeling techniques such as parametric and non-parametric have shown the capability of estimating forest attributes such as canopy width and cover, leaf area index, basal area, height, volume, and aboveground biomass based on remotely sensed data (Zhou et al., 2020). The common machine learning techniques include Bayesian Additive Regression Trees (BART), Classification and Regression Trees (CART), KNN, RF, and SVM. Bulut et al. (2022) modeled the relationship of forest stand attributes using multiple linear regression (MLR), SVM, and deep learning (DL) using Landsat 8 and Sentinel 2 data. Their study showed that SVM had better performance in predicting density and basal area. Ahmadi et al. (2020) used four machine learning algorithms (Generalized Linear Model (GLM), BART, KNN, and SVM) for Sentinel 2 data to calculate and compare the forest stand attributes. Their study found that BART had the best performance for predicting forest volume outperforming KNN, SVM and GLM. Noorian et al. (2016) found that Quickbird data performed best with an RMSE of 2.44 m²/ha for basal area, 50.98 m³/ha for volume, and 125 n/ha for density using CART. The above-mentioned researches prove the important role of remote sensing and machine learning techniques in forest resource assessment. However, its application, particularly in estimating forest attributes such as volume, tree density, and basal area, is still limited in the Philippines, especially at the local level. This study was then conducted to explore the possibility of remotely sensed data and machine learning techniques to model and predict forest stand volume at the local level. The specific objectives were to: (1) evaluate the correlation between forest stand volume and spectral bands, indices, and topographic variables; (2) assess the accuracy of KNN, RF and SVM for modeling the volume of forest stands using R-squared ($R^2$) values, mean absolute error (MAE), and root mean square error (RMSE); and (3) determine the best variables for predicting and mapping forest stand volume.

**Materials and Methods**

**Location of the Study Area**

The study was conducted within the municipality of Licuan-Baay, Abra (Figure 1). Licuan-Baay is
one of the municipalities within the Abra River Basin with a large tract of forest cover. It is located in the northern part of the Philippines, with geographic coordinates of 17° 35' 08.60" N and 120° 32' 33.50" E. The area is 30,567.70 ha. The forest cover is around 19,774.10 ha. The study area is characterized by having different forest stands. Most part of the study area is vegetated by natural forests while some part are forest plantations. Based on the Modified Coronas Classification, the municipality is under climatic type II, which is characterized by two seasons: dry from November to April and wet from July to November. The mean annual temperature is 24.0 °C, while the mean annual rainfall is 3,012 mm.

**Remote Sensing Data**

This study utilized data from Sentinel 2A, a satellite operated by the European Space Agency under the Copernicus Program. The Sentinel 2A satellite uses a multi-spectral instrument (MSI). The Sentinel satellite captures images with 13 spectral bands ranging from 0.443 μm to 2.190 μm. The 10 m resolution product of the Sentinel 2 is one of the highest among free-available satellite products (Abdi, 2020).

![Figure 1. Study Site: (a) Geographical Location of the Abra River basin, (b) Geographical Location of the Study Area (Municipality of Licuan-Baay) (c) Location of Sample Plots](image)

**Field Measurement of Forest Stand Attributes**

A total of 155 sample plots measuring 10 x 10 m were established for the collection of data on height and diameter at breast height (DBH). The volume of individual trees was calculated based on the DBH and height of the individual trees. Only trees with a DBH of more than 10 cm were considered for this study. Total and merchantable height were measured using a digital laser range finder, while DBH was measured using a diameter tape. Table 1 shows the descriptive statistics of the sample plots.
Table 1. Descriptive Statistics of the Measured Forest Stand Volume

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Volume (m³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.66</td>
</tr>
<tr>
<td>Min</td>
<td>0.03</td>
</tr>
<tr>
<td>Max</td>
<td>1.89</td>
</tr>
<tr>
<td>Sd</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Note: Values are at plot level (100 m²)

Correlation Analysis

A correlation analysis was performed to establish the relationship between the vegetation indices and forest attributes using the Pearson correlation coefficient. The Pearson product-moment correlation coefficient (r) measures how closely two quantitative variables are related, which is either a positive or negative correlation. Instead of calculating the ranks of the variables, the coefficient calculates the strength of the "linear" correlations between the raw data from both variables. Since this coefficient lacks dimensions, there are no boundaries about the data that need to be considered while using this formula for analysis. The Pearson product-moment correlation coefficient is computed by the following formula:

\[ r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2-(\sum X)^2][n\sum Y^2-(\sum Y)^2]}} \]  (1)

Where:
- \( n \) = number of sample size
- \( \sum X \) = Summation of X
- \( \sum Y \) = Summation of Y
- \( \sum XY \) = Summation of the product of X and Y variables
- \( \sum X^2 \) = Summation of the squared value of X
- \( \sum Y^2 \) = Summation of the squared value of Y

Spectral Bands, Vegetation Indices, and Topographic Feature Extraction

Spectral Bands and Indices Extraction

The blue, green, red, and NIR bands of Sentinel 2A were used in this study. These bands were also used as inputs for extracting the common vegetation indices. Vegetation indices were extracted following the formula for each index shown in Table 2. The indices were extracted using QGIS 3.24 software.

Table 2. Vegetation Indices as Predictor Variables for Developing Regression Models

<table>
<thead>
<tr>
<th>Vegetation Index</th>
<th>Formula</th>
<th>Description</th>
<th>Reference/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI</td>
<td>( \frac{NIR - RED}{NIR + RED} )</td>
<td>Normalized Difference Vegetation Index</td>
<td>Ahmadi et al., 2020</td>
</tr>
<tr>
<td>SAVI</td>
<td>( \frac{NIR - RED \cdot 1.5}{NIR + RED + 0.05} )</td>
<td>Soil Adjusted Vegetation Index</td>
<td>Dos Reis et al., 2018</td>
</tr>
<tr>
<td>GCI</td>
<td>( \frac{NIR}{GREEN - 1} )</td>
<td>Green Light Chlorophyll Vegetation Index</td>
<td>Wang et al., 2022</td>
</tr>
<tr>
<td>IPVI</td>
<td>( \frac{NIR}{NIR + RED} )</td>
<td>Infrared Percentage Vegetation Index</td>
<td>Ahmadi et al., 2020</td>
</tr>
<tr>
<td>TVI</td>
<td>( \sqrt{\frac{NIR - RED + 0.05}{NIR + RED}} )</td>
<td>Transformed Vegetation Index</td>
<td>Vivar-vivar et al., 2022</td>
</tr>
<tr>
<td>GNDVI</td>
<td>( \frac{NIR - GREEN}{NIR + GREEN} )</td>
<td>Green Normalized Difference Vegetation Index</td>
<td>Liu et al., 2022;</td>
</tr>
<tr>
<td>DVI</td>
<td>( \frac{NIR - RED}{NIR + RED} )</td>
<td>Difference Vegetation Index</td>
<td>Xue and Su 2017</td>
</tr>
</tbody>
</table>
### Topographic Features

For the topographic variable, the Advanced Land Observing Satellite (ALOS) Phased Array L-band Synthetic Aperture Radar (PALSAR) DEM of the study area was used in this study. The DEM has a 12.5 m spatial resolution. The slope feature of the study area was derived from the DEM. The DEM and slope were then resampled to 10 m based on Sentinel 2A data.

### Variable Importance Selection

In this study, the Recursive Feature Elimination (RFE) technique was used to identify the most important variables. The RFE is a recursive method that begins with all of the dataset's features and then repeatedly reduces the least important characteristics until the required number of features is achieved. The primary rationale underlying RFE is that the most relevant variable will have the biggest impact on the target variable, making it more valuable for forecasting the target.

### Machine Learning Methods

The geographical coordinates of the field inventory plots were used to extract the values of the predictor variables. The field inventory data, spectral bands, and spectral indices values were used in developing the regression models. The total data sets were divided into a training data set (75%), and a test data set (25%). Therefore, 116 data sets were used for training, while the remaining 39 data sets were used for testing. Three machine learning regression techniques, such as k-NN, RF, and SVM, were used in this study.

**k-Nearest Neighbor (k-NN).** The KNN is a non-parametric machine learning technique, both for regression and classification, that may be used for a broad variety of nonlinear variables. (Ahmadi et al., 2020). In un-inventoried areas, k-NN imputation is commonly employed to estimate the properties of a forest inventory (Falkowski et al., 2010).

Finding the k closest reference samples for each target unit in the feature space specified by predictor variables is the task of applying the kNN approach (Fu et al., 2019). Following that, the target unit is allocated to the average of each response variable's values found in these k-nearest samples (Cosenza et al., 2021). In regression, the predicted response of a new sample is usually the mean of the k-neighbor responses (Hawrylo et al., 2018).

**Random Forest (RF).** Because of its versatility and ease of use, RF is one of the most used algorithms (Donges, 2021). Regression analysis using random forests can effectively depict intricate correlations between several variables. Research has demonstrated that RF can be used to integrate spectrum data into regression investigations, sometimes producing better results than conventional regression techniques (Dos Reis et al., 2018). For data-based predictions, such as forest attribute estimation, RF regression is frequently utilized (Obata et al., 2021). The ability of an RF model to assess a variable's importance—that is, the degree to which each feature variable contributes to the model's prediction—is one of its main benefits (Obata et al., 2021).

**Support Vector Machine (SVM).** The SVMs work under the assumption that every input set will have a distinct relationship to the response variable and that rules that can be used to predict the response variable from new input sets can be found by grouping and relating these predictors to each other (Dos Reis et al., 2018). SVM uses a statistical learning process to successfully handle dataset complexity and noise (Ahmadi et al., 2020). According to Hawrylo et al. (2018), the significance of support vector machines (SVM) in regression modeling lies in their ability to exclude observations whose residuals fall under a user-defined threshold from contributing to the regression fit, while data whose residuals exceed the threshold contribute linearly.
Model Evaluation

The three machine learning algorithms were evaluated using $R^2$, MAE, and RMSE. The RMSE defines the error between the actual and anticipated values and assesses how reliable the predictions inside the models are (Ahmadi et al., 2021). In general, a larger $R^2$ value implies a better fitting effect of the model, and a smaller RMSE shows higher estimation accuracy. The following equations were used for the model evaluation:

$$ R = \frac{\sum(y-\hat{y})^2}{\sum(y-y)^2} $$  \hspace{1cm} (2)

where $y$ is the value for observation, $\hat{y}$ is the predicted value of $y$ and $\bar{y}$ is the mean value of $y$.

$$ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(O_i-P_i)^2}{N}} $$  \hspace{1cm} (3)

$$ \text{MAE} = \frac{\sum_{i=1}^{n}|O_i-P_i|}{N} $$  \hspace{1cm} (4)

where $O_i$ represents the observed values, $P_i$ represents the predicted values and $N$ is the total number of samples.

Results

Correlation of Forest Stand Volume to the Predictor Variables

The Pearson correlation coefficient was used to associate the values of forest stand volume, spectral bands, spectral indices, and topographic variables, as shown in Table 3 and Figure 2. The correlation analysis was done to provide descriptive information on the strength of each variable in relation to forest stand volume. The correlation analysis showed that there is a positive and significant correlation of volume with DEM and a positive but not-significant correlation with slope. In addition, the correlation analysis showed that there is a negative and significant correlation between volume and the rest of the independent variables.

Table 3. Correlation Analysis between Forest Stand Volume and Predictor Variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>1.00</td>
</tr>
<tr>
<td>Blue</td>
<td>-0.34**</td>
</tr>
<tr>
<td>Green</td>
<td>-0.29**</td>
</tr>
<tr>
<td>Red</td>
<td>-0.27**</td>
</tr>
<tr>
<td>NIR</td>
<td>-0.49**</td>
</tr>
<tr>
<td>SAVI</td>
<td>-0.31**</td>
</tr>
<tr>
<td>OSAVI</td>
<td>-0.31**</td>
</tr>
<tr>
<td>RVI</td>
<td>-0.33**</td>
</tr>
<tr>
<td>NDVI</td>
<td>-0.31**</td>
</tr>
<tr>
<td>GNDVI</td>
<td>-0.35**</td>
</tr>
<tr>
<td>DVI</td>
<td>-0.44**</td>
</tr>
<tr>
<td>TVI</td>
<td>-0.29**</td>
</tr>
<tr>
<td>GCI</td>
<td>-0.34**</td>
</tr>
<tr>
<td>IPVI</td>
<td>-0.29**</td>
</tr>
<tr>
<td>DEM</td>
<td>0.32**</td>
</tr>
<tr>
<td>Slope</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Note: **. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Source: The Author

Figure 2. Pearson’s Correlation Coefficient Heatmap
Performance of KNN, RF and SVM

The validations of the three machine learning models are presented in Table 4. The result showed that the RF algorithm had the highest $R^2$, followed by KNN and SVM (0.32). The RMSE was calculated between the observed and predicted forest volumes. The RF algorithm achieved the best RMSE (0.40 m$^3$/plot), followed by KNN and SVM (0.44 m$^3$/plot). The difference between observed and predicted values was also compared using MAE. The RF algorithm achieved the lowest MAE (0.31 m$^3$/plot), followed by SVM (0.34 m$^3$/plot) and KNN (0.37 m$^3$/plot). Consistently, the RF algorithm showed the best performance in terms of $R^2$, RMSE, and MAE.

<table>
<thead>
<tr>
<th>Validation Metrics</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KNN</td>
<td>RF</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.53</td>
<td>0.72</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.31</td>
<td>0.18</td>
</tr>
<tr>
<td>MAE</td>
<td>0.24</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Figure 3 illustrates the scatter plots of the predicted and observed forest stand volumes. Scatter plots help to visually evaluate the performance of the models. The smaller the deviation and scattered data from the regression line, the fewer random mistakes are included in regression modeling predictions.

A visual examination of the scatterplots reveals that RF exhibits the highest correlation (best fit) for volume prediction. Although evaluating forest stand attributes is difficult, the findings of this investigation were satisfactory in terms of $R^2$, MAE, and RMSE.

Assessment of Spectral Bands, Indices and Topographic Data for Modeling Accuracy

Figure 4 shows the importance of the different variables used in modeling. Eight (Blue, Green, NDVI, SAVI, OSAVI, IPVI, RVI, and TVI) of the 15 variables were eliminated by the RFE technique. The result showed that DEM and the NIR band were the most important predictors for forest stand volume, with a relative importance of 32% and 31%, respectively. The DVI showed considerable importance, which is 11%. Figure 4 also highlights the least significant contribution of the other variables, which is less than 10%.
Table 5 shows the predicted statistics of the different forest stand attributes. Figure 5 shows the mapping of the forest stand volume by the three machine learning techniques. The estimated forest stand volume using the RF model ranged from 33 to 115 m³/ha with a mean of 59 m³/ha, 24 to 121 m³/ha for the KNN model with a mean of 68 m³/ha, and 4 to 112 m³/ha for the SVM with a mean of 57 m³/ha. It is observed that the KNN and SVM models have a similar pattern in forest volume estimation. Generally, higher forest stand volumes are within the western portion of the study area.

Table 5. Statistics of Predicted and Observed Forest Stand Volume (m³/ha)

<table>
<thead>
<tr>
<th>FSA Statistics</th>
<th>KNN</th>
<th>RF</th>
<th>SVM</th>
<th>Field Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume (m³)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>68</td>
<td>59</td>
<td>57</td>
<td>66</td>
</tr>
<tr>
<td>Min</td>
<td>24</td>
<td>33</td>
<td>4</td>
<td>03</td>
</tr>
<tr>
<td>Max</td>
<td>121</td>
<td>115</td>
<td>112</td>
<td>189</td>
</tr>
<tr>
<td>Std</td>
<td>24</td>
<td>23</td>
<td>24</td>
<td>47</td>
</tr>
</tbody>
</table>

Discussion

The correlation analysis showed that volume was negatively correlated with the spectral bands and spectral indices but showed a positive correlation with the topographic variables. The negative trend observed between forest stand volume and spectral bands and indices is consistent with other studies. For instance, the studies of Fang et al. (2023) and Chrysafis et al. (2017) showed that growing stock volume is negatively correlated with the blue, green, red, and NIR bands and also with vegetation indices such as NDVI, SAVI, and DVI. The positive correlation of forest volume to the topographic variables (DEM and slope) is similar to the findings of Zhou and Feng (2023).

R², MAE, and RMSE are common metrics used in validating regression models. The R², or goodness of fit, indicates how well the data fit the regression model. Regression models fit data well when the difference between observed and predicted and mapping forest stand volume
predicted values is small and unbiased. Generally, higher $R^2$ values are ideal, but it also depends on the study conducted. An absolute measure of fit is provided by the RMSE, whereas the $R^2$ value offers a relative measure of fit (Ahmadi et al., 2020). In this study, the $R^2$, RMSE, and MAE were used to validate the regression models. The results showed that RF was the best algorithm to predict forest volume. Compared to the study of Ahmadi et al. (2020), which also used Sentinel 2A and machine learning algorithms, the BART algorithm is the best model to predict basal area and volume, outperforming KNN, SVM, and GLM. The study by Hu et al. (2021) using LIDAR showed that RFK, a hybrid model of the RF, had a better prediction performance for volume than the hybrid models of Artificial Neural Networks (ANN) and SVM, which are ANNK and SVMK. Bulut et al. (2023) also reported in their study that SVM had better performance than DL and MLR for predicting volume. In the study of Mohamaddi et al. (2011), which used Landsat ETM, the CART model showed significantly higher prediction accuracy than MLR models for volume. The accuracy assessment revealed that the findings of this study demonstrated varying accuracies for estimating forest stand volume using non-parametric regression techniques. Based on the $R^2$ values, RF was more accurate at estimating volume, while the other regression methods had poorer goodness of fit. All of the regression models have $R^2$ values less than 50%. Although the $R^2$ value is low, it does not always indicate that forecasts are less accurate (Ahmadi et al., 2020). Several studies also reported a lower or comparable $R^2$ compared to the result of this study. For instance, Ahmadi et al. (2020) reported an $R^2$ of 0.18 to 26 using KNN, MLR, and BART. Gomez et al. (2012) reported an $R^2$ of 0.46 using CART for density estimation. In contrast, the studies of Mohamaddi et al. (2011), Noorian et al. (2016), and Bulut et al. (2022) reported higher $R^2$ (0.49–0.96).

Spectral bands and spectral indices from different satellite sensors are commonly used as predictor variables in estimating forest stand attributes and show varying results. In addition, topographic and climatic variables are included as inputs for modeling. Some of the common satellite data used are Landsat, Sentinel, IKONOS, and QuickBird. In this study, it was found that DEM and NIR are the most important variables in predicting forest stand volume. The result of this study is similar to the findings of Ahmadi et al. (2020), where elevation data is the most important variable for stem volume. In contrast to the study of Mohammadi et al. (2011), DVI is the most important variable in predicting volume and density. Gunlu et al. (2014) estimated forest stand characteristics such as basal area, height, and volume with band reflectance values and spectral indices obtained from the IKONOS satellite image using multiple regression analysis. Their study revealed that spectral indices showed better predicting capability compared to spectral bands. In addition, their study showed that the spectral indices DVI and EVI are the best independent variables for predicting volume and basal area. Chrysafis et al. (2017) estimated volume using the spectral bands and indices of Landsat 8 and Sentinel 2 satellite images. The result of their study showed that reflectance values and vegetation indices extracted from Sentinel 2 had better performance. Bulut et al. (2022) estimated density, basal area, and volume using reflectance and vegetation indices obtained from Landsat 8 and Sentinel-2 satellite images. Using multiple regression analysis, their study showed that higher $R^2$ values were obtained with vegetation indices than reflectance values. The results of other studies show that spectral indices are better predictors than spectral bands. Modeling performance is significantly influenced by various factors, such as forest stand attributes, the time satellite images were acquired, and the characteristics of the satellite imagery. According to Bulut et al. (2020), vegetation indices are better predictors of forest stand attributes because vegetation indices are combinations of bands with different wavelengths, thus better reflecting the features of the study area. In contrast to spectral bands that measure specific wavelengths. The findings of this study, however, showed that DEM and NIR bands are
the most important variables in predicting forest stand volume using the RF algorithm.

**Conclusion**

This study investigated the potential of utilizing Sentinel 2 imagery and machine learning algorithms for estimating the forest stand volume of natural and plantation forests in the Abra River Basin. To estimate forest stand volume, regression models were built using the spectral bands of Sentinel 2, common vegetation indices, topographic variables, and field inventories. The random forest regression model is the best for predicting volume. The findings of this study also demonstrated the relevance of using topographic variables like elevation to increase model accuracy. The results of this study revealed that forest volume can be measured using freely available satellite data and machine learning techniques. The use of other machine learning methods and a larger set of data is still recommended to further improve the accuracy of the models.

**Acknowledgement**

The authors would like to extend their gratitude primarily to the Commission on Higher Education-Region 2, Philippines for funding the study and to the reviewers for providing insightful comments and suggestions.

**Conflict of Interests**

No conflict of interest.

**References**


