Machine learning approach for estimating tree volume

Siti Hajar Mohd Mushar¹, Sharifah Sakinah Syed Ahmad¹, Fauziah Kasmin¹ and Nur Hajar Zamah Shari²

¹ Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, Durian Tunggal, 76100 Malacca, Malaysia
² Forestry and Environment Division, Forest Research Institute Malaysia (FRIM), 52109 Kepong, Malaysia

sakinah@utem.edu.my

Abstract. With the forestry and logging activities contributing to 5.6% of the agricultural sector in Malaysia’s 2018 GDP growth, this had thus implied the forest as having a significant role in national growth and the critical need of a precise tree volume estimation. Although regression has been the most common method used for this form of estimation, the expansion of information technology had, however, led to the use of a machine learning technique that is capable of overcoming the issues posed by the regression analysis. In this paper, the estimation of the tree volume was not only conducted via the regression method but had also involved the use of two machine learning techniques, namely the artificial neural network (ANN) and that of the epsilon-Support Vector Regression (ε-SVR). By comparing the root mean square error (RMSE) and standard deviation (SD) values from each of the volume model that had been obtained in this study, the machine learning technique was thus found to have demonstrated a better precision and accuracy level than that of the regression method.

1. Introduction

Since forest had made up about 30.6% of the world’s terrestrial surface [1], its importance as an essential component in the terrestrial ecosystem as well as for regulating the climate and in the mitigation of natural disasters had thus been heavily emphasized by Henry et al. [2], Kuyah et al. [3], Xia et al. [4] & Mugasha et al. [5] in the literature. According to the Food and Agriculture Organization of the United Nations (FAO) [6], the world’s forests are estimated to store more than 650 billion tons of carbon, of which 44% is stored in the aboveground living biomass, while the respective 11% and 45% are being stored in the deadwood and forest soil. For these reasons, the forest is not only seen as being important for reducing the amount of carbon dioxide in the atmosphere by way of the photosynthesis process but is also vital for reducing the thinning of the ozone layer and preventing the occurrences of other natural disasters. Apart from being a carbon sink, forests are also known to play a pivotal role in the national growth, where its non-timber and forest products have been found to benefit the rural communities as well as to the country as a whole [4], [7].

Being richly endowed with a tropical rainforest climate, the Department of Statistics Malaysia (DOSM) had revealed the latest Malaysia’s forestry and logging statistics as contributing to 5.6% of the agricultural sector in Malaysia’s 2018 GDP growth. Since Japan, US, India, Korea and China have been known to be the main importers of Malaysia’s timber products, a precise tree volume that is highlighted.
by JR. and Wood [8] & Shari et al. [9] would then be deemed as necessary for a better management and administration of the Malaysian forest.

Although the tree volume can be either expressed in terms of the total tree or the total area cubic volume, this paper had, however, focused on the volume estimation technique that is being used on those of individual trees. According to the forest mensuration book that was written by Jr. et al. [10], the estimation of a tree volume can be conducted from the use of a graphical and integration method, the water displacement technique or the regression analysis method. While the regression analysis had been known to be the most favoured method used for estimating the tree volume, the expansion of information technology had nevertheless led to the eventual transformation from one of a volume modelling to that of a machine learning (ML) technique.

By stating the values of the dependent variable in the forest modelling as being directly proportional to the error variations, Parresol [11] had contradicted the initial assumption of the least square regression with independent errors and that of a normally distributed mean and constant variance of $X \sim \mathcal{N}(0,1)$. Since the high variability or non-normal distribution of the data could have been due to the noise that had existed in the real tree data, Swingler [12] had therefore suggested a possible solution with the use of a machine learning (ML) method. Generally, the ML is part of an artificial intelligence system that automates and eases the construction of the model analytics process, and because it had consisted of several types, the application of the ML technique is not only limited to performing regression and classification but can also be used for clustering and anomaly detection purposes.

Apart from pointing out the ML as being a more favourable method because of its capability in automating the detection of hidden data patterns, Diamantopoulou et al. [13] had also claimed the Artificial Neural Network (ANN) as being a reliable branch of the ML forest modelling system with an extensive usage in the forestry and healthcare studies [14]–[17]. Since the ANNs are biologically inspired computer programs that had been designed to simulate the way in which the human brain processes information [18], [19], it has now been successfully applied in the studies of forestry, particularly in the biomass and volume prognosis [20]–[23], the construction of a forest growth and forest mortality model design [24], [25] and that of the tree volume estimation [21], [26]. Although the ANN model had been employed extensively in most of the volume estimation studies, many scholars had still debated on the accuracy levels of the said method with Lacerda et al. [21] asserting the ANN as producing a better volume estimate than the regression method and Tavares Júnior et al. [26] as stating otherwise.

Besides the ANN, the support vector regression (SVR) is also another dominant approach that can be used for resolving the regression issue. Initiated by Vapnik–Chervonenkis in the 1960s, the support vector machine (SVM) can be categorized into the support vector classification (SVC) as well as the support vector regression (SVR) [27]. Briefly, the SVR that was introduced by Vapnik, Steven Golowich and Alex Smola in 1992 is a prediction tool that utilises the application of machine learning theory, where the overfitting of data would be screened out by an automatic screening process as a way of ensuring the accuracy of the regression prediction [27], [28]. Although the SVR had consisted of the epsilon-Support Vector Regression ($\varepsilon$-SVR) and that of nu-Support Vector Regression ($\nu$-SVR), Diamantopoulou et al. [13] had, however, demonstrated the former as producing a better result of the two.

With this in mind, this study had therefore set out to explore and assess the performance of the tree volume estimation data from one of the Malaysia forest reserves with the use of the regression and machine learning techniques.

2. Material and methods

Apart from presenting the methodology of the tree volume estimation, this section will also introduce some background information pertaining to the regression, $\varepsilon$-SVR and the ANN models. By starting off with the data description and measurement procedure, this section had proceeded with further elaboration on the statistical techniques that were used in this dataset before concluding with the assessment results of the two methods used.
2.1. Data description
This study had adopted the selective sampling method, where the trees were chosen based on the diameter classes and species group. By using the real tree data that had been gathered from Compartment 37 in the Cherul Forest Reserve, Terengganu, the measurements were then carried out on both the standing and felled trees, but with the exclusion of the damaged, dead and dying trees. From the respective handheld Laser Criterion 400 as well as the callipers and measuring tape that was used for measuring the diameter and height of both the standing and felled trees, these trees were found to have a diameter at breast height (DBH) with less than 15 cm.

Apart from measuring the merchantable length and stump height, where it had revealed the DBH measurement as being 1.3m above the ground level, the total height of the tree was also obtained from predetermined ground height and that of the crown peak. Other than the DBH and total height, the diameter of the stem (overbark) too was measured from the stump height up to the first main branch at a 2m interval, while the Huber’s formula, which is a common method that is used in both the European and those of the tropical countries had been employed for gauging the tree volume [29], [30]. The Huber’s formula can, therefore, be stated as follows:

\[ V_i = f \times L \times (dM)^2 \]  

where:

- \( V_i \) = Log volume at i/th (m³)
- \( L \) = Length of log (m)
- \( f \) = 0.00007854 (metric units)
- \( dM \) = Diameter at the mid-length log end (cm)

From the 265 tree sample data that had been collected, the number of the tree sample data was then reduced to 241 after being subjected to the pre-processing stage. Since the aim of this study had been to produce a universal volume equation that can be applied for the dipterocarp and non-dipterocarp tree-families, the 241 tree sample data were then subdivided into these two categories with 201 forming the non-dipterocarps and the rest being the dipterocarps. Since the dipterocarps are a family of hardwood, tropical trees, they would normally cover about 57% of the Peninsular Malaysia’s lowland forest with the non-dipterocarps making up the rest of the woodlands.

The data from each of the group were then further divided according to its diameter class by using the general 70:30 training: testing ratio, where the data will not only be free from any form of biasness, but the testing will also be totally independent of any of the fitting procedures that had been employed in the constructed models (such as those of the regression and machine learning techniques). The summary statistics of all the species, the dipterocarps as well as the non-dipterocarps tree-families are thus shown in Table 1.

| Variable | Fitting data | Testing data |
|----------|--------------|--------------|
|          | Mean | Min. | Max. | SD | Mean | Min. | Max. | SD |
| **D (cm)** | All species (169 trees) | 52.2 | 12.8 | 98.0 | 17.0 | 53.1 | 23.6 | 129.6 | 18.0 |
|          | H (m) | 17.0 | 10.7 | 30.0 | 3.9 | 17.3 | 6.5 | 25.9 | 3.7 |
| **Dipterocarp species (28 trees)** | D (cm) | 61.7 | 25.8 | 98.0 | 20.0 | 66.6 | 32.8 | 129.6 | 26.9 |
|          | H (m) | 20.4 | 13.3 | 30.0 | 4.6 | 17.1 | 11.9 | 20.9 | 2.6 |
| **Non-dipterocarp species (141 trees)** | D (cm) | 49.3 | 23.6 | 86.9 | 14.4 | 52.66 | 12.8 | 99.4 | 17.26 |
|          | H (m) | 16.4 | 6.5 | 28.2 | 3.5 | 17.3 | 11.2 | 25.9 | 3.75 |
2.2. Regression analysis

From the extensive research that was conducted on the volume table development, it was revealed that there had been a list of commonly used regression models that were favoured by the researchers. For this reason, each of the datasets was then fitted according to the volume models that had been listed below:

1. \( V = b_0 + b_1D \)  
2. \( V = b_0 + b_1D + b_2D^2 \)  
3. \( V = b_0 + b_1D^2 \)

where:

- \( V = \) Tree volume (m³)
- \( D = \) Diameter (m)
- \( H = \) Log length (m)
- \( b_i = \) Regression coefficients

2.3. Support vector regression models (SVRs)

Under the ε-SVR algorithm, the input data would first be mapped onto an m-dimensional feature space with the use of a non-linear mapping and followed by the construction of a linear model in the feature space, where the goal would be to find a function \( f(x) \) among the pairs of the training data (input \( x_i \), target \( y_i \)) = \( (x_i, y_i) \), but not those that had demonstrated a larger \( \epsilon \) deviation from the SV. By optimizing the parameters of the ε-SVR models with the use of an “e1071” package in the R software, the error band of the function \( f(x) \) was thus found to have been between that of the \([-\epsilon, \epsilon]\) interval.

2.4. Artificial neural network (ANN)

Consisting of one or more processing units of ‘artificial neurons’ or ‘perceptrons’ [31], these ANN neurons are connected with one another by a series of weighted connections. Depending on the system that is being replicated, the perceptrons of an ANN would be arranged in layers, where each of the perceptrons from the preceding layer would have a weighted connection with those of the proceeding layer. When a system is being replicated from the ANN training process, the training data set would then be fed through the network with each of the perceptron processing the input data or input signal from either the input layer or of the preceding perceptrons before the final emission of an output signal. Under this circumstance, the weights and the structure of the network would be altered according to that of a specified training algorithm.

2.5. Statistical criteria for model evaluation

This study had adopted the Furnival’s Index instead of the other common statistical test as a way of preventing the inclusion of the dependent variable transformations and the weighted regression in the analysis phase. Since a standard error (SE) would be required in the calculation of the Furnival’s Index, its formula can thus be shown as below:

\[
FI = [f''(V)]^{-1} \times s
\]

where:

- \( FI = \) Furnival’s Index
- \( s = \) Residual standard error from the fitted regression
- \([f''(V)]^{-1} = \) Geometric mean
Although Walther and Moore (2005) had stated precision to be a consistent measure of prediction, the precision level would still be affected by the measurement or estimation techniques that were being used at a different time point of time [10] and that of the standard deviation (SD). With that, the precision can be estimated by using the following formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(V_i - \bar{V})^2}{(n - 1)}}$$

(6)

where:

- \(V_i\) = Log volume at \(i\)th (m³)
- \(\bar{V}\) = Mean volume (m³)
- \(n\) = Number of sample

Since accuracy refers to how closely the measured value of a quantity had corresponded to its true value, the mean square error (MSE) would then be regarded as the most common accuracy measure that is denoted by a small variance and the least bias in its prediction level [32]. Hence, the RMSE, which is the standard deviation of the prediction error, can thus be expressed by using the following formula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (V_i - \bar{V})^2}$$

(7)

where:

- \(V_i\) = Log volume at \(i\)th (m³)
- \(\bar{V}\) = Mean volume (m³)
- \(n\) = Number of sample

3. Results

The results of the regression coefficient, RMSE and SD for all the species, dipterocarps and non-dipterocarps from both the fitting and testing datasets are thus shown in Table 2.

| Model (Species) | Fitting data | Testing data |
|-----------------|--------------|--------------|
| All species (169 trees) | \(\beta_0 = -2.8658\), \(\beta_1 = 0.1120\), \(\beta_2 = 1.3550\), \(F = 1.2907\), \(RMSE = 1.9864\), \(SD = \) | \(\beta_0 = 0.2031\), \(\beta_1 = -0.0181\), \(\beta_2 = 0.0012\), \(F = 1.7553\), \(RMSE = 2.6214\), \(SD = \) |
| Dipterocarp species (28 trees) | \(\beta_0 = -0.2408\), \(\beta_1 = 0.0011\), \(\beta_2 = 1.3980\), \(F = 1.6763\), \(RMSE = 2.5367\), \(SD = \) | \(\beta_0 = 1.7472\), \(\beta_1 = -0.0842\), \(\beta_2 = 0.0019\), \(F = 4.0170\), \(RMSE = 5.9939\), \(SD = \) |
| Non-dipterocarp species (141 trees) | \(\beta_0 = -0.1076\), \(\beta_1 = -0.0010\), \(\beta_2 = 1.0610\), \(F = 1.4890\), \(RMSE = 1.8213\), \(SD = \) | \(\beta_0 = -2.3691\), \(\beta_1 = 0.0990\), \(\beta_2 = 1.0600\), \(F = 1.3980\), \(RMSE = 1.6105\), \(SD = \) |
By referring to Table 2, since the fitting of the 169 trees under the volume model 1 and 2 had recorded an equal and the lowest FI value of 1.3550, this had thus indicated a good fit of the model, where a lower value would indicate the favourability of the model and a larger value as indicating otherwise. The model fit was also verified from the volume model 3 and volume model 2 of the respective dipterocarp and non-dipterocarp dataset, where they too had shown the least corresponding values of 2.0820 and 1.0590.

Once the volume equation modelling process had been completed, model validation was then conducted on the dataset that had not been part of the regression analysis as a way of preventing the occurrence of biasness in the data. The results that were obtained from the performance of the model validation, as depicted in Table 2 were then evaluated in terms of their SD and RMSE values.

The results of the respective SD and RMSE values that had been attained by each of the datasets had not only revealed the consistency of the prediction but had also provided an account on how far the estimated volume had deviated from that of its true volume. As shown by the results of the statistical tests in the above dataset, since all of the lowest RMSE and SD values for the respective all species, dipterocarp and the non-dipterocarp were found to have occurred at volume model 1, this had thus implied the volume model as producing the most accurate and precise values.

In order to meet the objective of the study, an accuracy assessment was then performed on the volume model with the use of regression and that of the machine learning techniques. The accuracy of the assessment results that had been obtained by way of regression, ANN and SVR are thus shown in Table 3.

| Table 3. Comparison of results between the regression and the machine learning techniques. |
|---------------------------------------------------------------|
| All species | Dipterocarp | Non-dipterocarp |
| ANN | 1.2639 | 1.6750 | 0.6377 |
| SVR | 0.8076 | 2.1627 | 0.8762 |
| Regression | 1.2907 | 1.8680 | 1.3980 |

By referring to the numeric values in Table 3, the accuracy of the results that had been obtained from all species and that of the dipterocarp with the use of the regression method were found to have been between those of the ANN and SVR, while the SVR for all the species data had shown the lowest RMSE value when being compared to that of the ANN and regression methods. Since the RMSE indicates the distance between the estimated and that of the true volume, the volume estimation from the use of the regression method was found to have deviated by 1.2907 from its true volume as opposed to the dipterocarp, where its SVR had shown the highest RMSE value to be of 2.1627. While both the dipterocarp and non-dipterocarp had demonstrated the lowest RMSE values from the use of the ANN, the use of the regression method on all species and the non-dipterocarp categories had, however, shown to have resulted in the highest RMSE values in both of the datasets.

4. Conclusion
Although the volume model 1 of the regression analysis in all the species, dipterocarp and non-dipterocarp had demonstrated a high accuracy level, one of the essential elements that are required in an ideal volume model would be the capability of resolving the regression analysis issues. For these reasons, since the machine learning technique was not only found to be a reliable and a robust model for solving the complex environmental issues but had also produced a better output result, this research had thus selected the modelling from the machine learning technique as being a better volume model than that of the regression method.

Acknowledgement
The authors express their deepest gratitude and thanks to Universiti Teknikal Malaysia Melaka (UTeM) in supporting this research PJP/2019/FTMK(68)/S01678.
References

[1] Food and Agriculture Organization (FAO), *Global Forest Resources Assessment 2015: How are the World’s Forests Changing?*, 2nd ed. Rome, Italy: Food and Agriculture Organization of the United Nations, 2016.

[2] M. Henry et al., “Estimating Tree Biomass of Sub-Saharan African Forests: A Review of Available Allometric Equations,” *Silva Fenn.*, vol. 45, no. 3B, pp. 477–569, 2011.

[3] S. Kuyah et al., “Allometric Equations for Estimating Biomass in Agricultural Landscapes: I. Aboveground Biomass,” *Agr. Ecosyst. Environ.*, vol. 158, pp. 216–224, 2012.

[4] Z. S. Xia, W. S. Zeng, S. Zhu, and H. Z. Luo, “Construction of Tree Volume Equations for Chinese Fir Plantations in Guizhou Province, Southwestern China,” *For. Sci. Pract.*, vol. 15, no. 3, pp. 179–185, 2013.

[5] W. A. Mugasha et al., “Allometric Models for Prediction of Above- and Belowground Biomass of Trees in the Miombo Woodlands of Tanzania,” *For. Ecol Manag.*, vol. 310, pp. 87–101, 2013.

[6] Food and Agriculture Organization of the United Nations (FAO), “Global Forest Resources Assessment 2010,” 2010.

[7] Food and Agriculture Organization of the United Nations (FAO), “State of the World’s Forests,” Rome, Italy, 1997.

[8] D. W. Patterson, H. V. W. JR., and G. B. Wood, “Errors in Estimating the Volume of Butt Logs,” *For. Prod J.*, vol. 43, no. 3, pp. 41–44, 1993.

[9] N. H. Z. Shari, W. M. S. W. Ahmad, S. Musa, W. R. W. Mohd., and H. Ismail, “Development of Local Volume Table for Second Growth Forests Using Standing Tree Measurements,” *Malaysian For.*, vol. 73, no. 2, pp. 163–170, 2010.

[10] J. A. K. Jr., M. J. Ducey, T. W. Beers, and B. Husch, “Determination of Tree Volume, Weight and Biomass,” in *Forest Mensuration*, 5th ed., Wiley-Blackwell, 2016, pp. 135–183.

[11] B. R. Parresol, “Modeling Multiplicative Error Variance: An Example Predicting Tree Diameter from Stump Dimensions in Baldcypress,” *For. Sci.*, vol. 39, pp. 670–679, 1993.

[12] K. Swingler, *Applying Neural Networks: A Practical Guide*. Great Britain: Morgan Kaufman Publishers Inc., 2001.

[13] M. J. Diamantopoulou, R. Özçelik, and H. Yavuz, “Tree-bark Volume Prediction via Machine Learning: A Case Study Based Onblack Alder’s Tree-bark Production,” *Comput. Electron. Agric.*, vol. 151, pp. 431–440, 2018.

[14] D. A. Roffman et al., “Development and Validation of a Multiparameterized Artificial Neural Network for Prostate Cancer Risk Prediction and Stratification,” *JCO Clin. Cancer Informatics*, vol. 2, pp. 1–10, 2018.

[15] A. Catic, L. Gurbeta, A. Kurtovic-Kozaric, S. Mehmedbasic, and A. Badnjevic, “Application of Neural Networks for Classification of Patau, Edwards, Down, Turner and Klinefelter Syndrome based on First Trimester Maternal Serum Screening Data, Ultrasonographic Findings and Patient Demographics,” *BMC Med. Genomics*, vol. 11, no. 1, p. 19, 2018.

[16] I. O. Ahmed, B. A. Ibraheem, and Z. A. Mustafa, “Detection of Eye Melanoma Using Artificial Neural Network,” *J. Clin. Eng.*, vol. 43, no. 1, pp. 22–28, 2018.

[17] P. F. J. Tsai et al., “Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients using Artificial Neural Network,” *J. Healthc. Eng.*, 2016.

[18] R. Özçelik, M. J. Diamantopoulou, J. R. Brooks, and H. V. Wiant, “Estimating Tree Bole Volume using Artificial Neural Network Models for Four Species in Turkey,” *J. Environ. Manag.*, vol. 91, no. 3, pp. 742–753, 2010.

[19] L. P. Reis et al., “Prognosis on the Diameter of Individual Trees on the Eastern Region of the Amazon using Artificial Neural Networks,” *For. Ecol. Manag.*, vol. 382, pp. 161–167, 2016.

[20] S. Yang, Q. Feng, T. Liang, B. Liu, W. Zhang, and H. Xie, “Modeling Grassland Above-ground Biomass based on Artificial Neural Network and Remote Sensing in the Three-River Headwaters Region,” *Remote Sens. Environ.*, vol. 204, pp. 448–455, 2018.
[21] T. H. S. Lacerda, C. D. Cabacinha, C. A. Araújo Júnior, R. D. Maia, and K. W. de S. Lacerda, “Artificial Neural Networks for Estimating Tree Volume in the Brazilian Savanna,” Cerne, vol. 23, no. 4, 2017.

[22] C. R. Sanquetta, L. R. Piva, J. Wojciechowski, A. P. Corte, and A. B. Schikowski, “Volume Estimation of Cryptomeria Japonica Logs in Southern Brazil using Artificial Intelligence Models,” South. For. a J. For. Sci., vol. 80, no. 1, pp. 29–36, 2017.

[23] A. A. Vahedi, “Artificial Neural Network Application in Comparison with Modeling Allometric Equations for Predicting Above-ground Biomass in the Hyrcanian Mixed-beech Forests of Iran,” Biomass Bioenergy, vol. 88, pp. 66–76, 2016.

[24] D. Zhang, L. Zhang, Q. Ye, and H. Ruan, “Robust Learning-based Prediction for Timber-volume of Living Trees,” Comput. Electron. Agric., vol. 136, pp. 97–110, 2017.

[25] S. J. S. da Rocha et al., “Artificial Neural Networks: Modeling Tree Survival and Mortality in the Atlantic Forest Biome in Brazil,” Sci. Total Environ., vol. 645, pp. 655–661, 2018.

[26] I. D. S. Tavares Júnior et al., “Artificial Neural Networks and Linear Regression Reduce Sample Intensity to Predict the Commercial Volume of Eucalyptus Clones,” Forests, vol. 10, no. 3, p. 268, 2019.

[27] J. Langhammer and J. Cesák, “Applicability of a Nu-Support Vector Regression Model for the Completion of Missing Data in Hydrological Time Series,” Water, vol. 8, no. 12, p. 560, 2016.

[28] D. Basak, S. Pal, and D. C. Patranabis, “Support Vector Regression,” Neural Inf. Process. Rev., vol. 11, no. 10, pp. 203–224, 2007.

[29] H. S. Grave, Forest Mensuration. New York: John Wiley & Sons, Inc., 1906.

[30] R. Ozçelik, H. V. W. JR., and J. R. Brooks, “Estimating Log Volumes of Three Tree Species in Turkey by Six Formulae,” For. Prod J., vol. 56, no. November/December, pp. 84–86, 2006.

[31] N. B. Karayiannis and A. N. Venetsanopoulos, Artificial Neural Networks. Springer, 1993.

[32] B. A. Walther and J. L. Moore, “The Concepts of Bias, Precision and Accuracy, and Their Use in Testing the Performance of Species Richness Estimators, with a Literature Review of Estimator Performance,” Ecography (Cop.), vol. 28, no. 6, pp. 815–829, 2005.