

Biomass Prediction of Mangrove Forests in Langkat Regency Using Random Forest Regression Method

Ramalan Biomas Hutan Bakau di Daerah Langkat Menggunakan Kaedah *Random Forest Regression*

Rahmat Hidayat¹, Freza Riana², Sahid Agustian Hudjimartsu³

^{1,2,3} Informatics Engineering, Faculty of Technology & Science, Universitas Ibn Khaldun Bogor Jl. Sholeh Iskandar, RT.01/RW.10, Kedungbadak, Kec.

Email: ¹chimet25@gmail.com, ²freza@ft.uika-bogor.ac.id, ³shudjimartsu@uika-bogor.ac.id

ABSTRACT

Mangrove forests are important ecosystems that play a role in maintaining the stability of coastal ecosystems, absorbing carbon dioxide (CO₂) more effectively than terrestrial forests, and protecting coastal areas from abrasion and the impact of natural disasters. This research was conducted in Lubuk Kertang Village, Langkat Regency, North Sumatra, with an area of 3026 km². The main objective of this study is to predict the biomass of mangrove trees using the Random Forest Regression (RFR) method by using unmanned aerial vehicles (UAVs) to obtain Biomass on the Ground (AGB). Model validation was carried out using Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), R-squared (R²). The results of the analysis showed that the RFR model gave excellent results with an R² value of 0.9697, RMSE of 2.55, and MAE of 1.82. The distribution of biomass in the study area showed significant variation, with the average tree biomass of 21,136 Kg/tree, the lowest biomass of 6,588 Kg/tree, and the highest biomass of 46,700 Kg/tree.

Keywords: Biomass; UAV; Random Forest Regression; AGB; Mangrove

INTRODUCTION

Mangroves are a type of tree vegetation that is able to develop in the transition area between the sea and land. Mangrove ecosystems have distinctive characteristics, one of which is the influence of tidal waves. Mangrove habitats are generally located in the confluence area between the estuary and the river. In addition, mangrove roots are often seen sticking upwards to get oxygen. (Apriliyani et al., 2020). With a strong and extensive root system, mangrove forests play a role in protecting coastlines from abrasion and are even able to withstand the impact of tsunami waves. In addition, mangrove forests also function as an important element in reducing the impact of climate change and global warming. The ability of mangrove forests to absorb carbon dioxide (CO₂) and release oxygen (O₂) is more effective than other terrestrial forests (Hidayah, Rachman, and As-Syakur, 2023).

This research was conducted in Lubuk Kertang Village with an area of 30.26 km² or (3026 Ha) with 33.70% of the total area of West Brandan District (West Brandan in Figure, 2023) (Function et al., 2024). The benefits of biomass, such as ash from burning can be used as a natural fertilizer that adds



nutrients to the soil and supports organic farming, to know biomass we must collect field data to estimate the amount of biomass, in the form of height data, diameter and number of trees. Therefore, remote sensing is a viable data for estimating Above Ground Biomass (AGB) on a large scale (Navarro et al, 2020).

Unmanned Aerial Vehicles (UAVs) are often used for various purposes, such as monitoring, shooting, mapping, unmanned aerial vehicle UAVs that will make it easier and faster to accurately detect biomass above the ground (AGB) (Harapan et al., 2021).

Research related to biomass estimation has been carried out with various approaches and data sources. One of the studies by (Evita Fitri. 2023) compares the Linear Regression and the Random Forest Regression methods in the context of house price prediction. The results showed that Linear Regression produced the smallest RMSE value of 0.433, while Random Forest Regression produced an RMSE of 0.440, although more complex, demonstrating the superiority of Linear Regression in the context of data that are linear and simple structured. This study shows that the performance of the model is highly dependent on the characteristics of the data used.

Meanwhile, in the specific context of biomass estimation, (Nurhidayat et al. 2023) used the Multiple Linear Regression method based on the Forest Canopy Density (FCD) and Canopy Height (CH) parameters, with a correlation coefficient (r) value of 0.664 and a determination coefficient (R^2) of 0.4779. The model was validated using an RMSE of 46.5, which suggests that although the linear regression approach is capable of delivering fairly good results, there is still room for improvement in accuracy especially when compared to non-linear machine learning methods.

Furthermore, (Tien Dat Pham et al. 2020) estimated biomass using the Support Vector Regression (SVR) method based on ALOS-2 PALSAR-2 and Sentinel-2A data, and managed to achieve more accurate results than the other four models with $R^2 = 0.596$, $RMSE = 0.187$, and $MAE = 0.123$. The estimated AGB in this study was in the range of 36.22 to 230.14 Mg/ha, with an average of 87.67 Mg/ha, indicating that the integration of high-resolution remote sensing data and non-linear algorithms can provide highly precise biomass prediction results.

Compared to previous studies, this study tried to fill a methodological gap by combining UAV (Canopy Height Model) data and field measurements in the Random Forest Regression (RFR) model. In contrast to the linear regression approach used by Nurhidayat et al. or SVR by Pham et al., this study shows that Random Forest is able to handle non-linear relationships and complex interactions between variables such as diameter, height, and canopy density commonly found in mangrove ecosystems. In addition, the use of UAVs provides advantages in terms of spatial coverage and efficiency of data collection in the field.

Thus, the main contribution of this study lies in the incorporation of UAV data and non-parametric machine learning in the context of mangrove biomass estimation in Indonesia, which was previously still limited. This approach is expected to be the basis for the development of digital and spatial biomass monitoring systems in the future.

This study aims to predict biomass in mangrove trees using UAV drones using the Random Forest Regression method in Langkat Regency, North Sumatra Province.

PROBLEM STATEMENT

The formulation of the problems that arise from the background that has been explained earlier is as follows:

1. How to predict biomass using the Random Forest Regression method?
2. What is the distribution of Biomass Above the Ground (AGB) in mangrove forests in Langkat Regency?

LITERATURE REVIEWS

Mangrove Forest

Mangroves are forests affected by the tides of seawater, found around tropical beaches around the world. Mangrove trees also have adaptations through root arrangement to sustain themselves in fine mud deposits and transport oxygen from the atmosphere to the roots. Mangrove forests play an important role in protecting coastlines from abrasion and tsunami waves, as well as minimizing the impact of climate change. Most mangroves have floating seedlings that are produced annually in large quantities.

Remote Sensing

Remote sensing is a technique that is based on the use of electromagnetic waves. Technology to collect information about the earth's surface without direct contact using sensors on satellites or drones. Remote sensing is able to monitor large and relatively narrow mangrove areas, reducing costs and taking less time compared to direct-to-field measurements. The use of remote sensing that records the landscape at regular intervals can monitor the addition and subtraction of mangrove areas in a relatively short time.

Machine Learning

Machine learning is an integral part of many commercial applications and research projects today, in areas ranging from medical diagnosis and treatment to finding your friends on social networks. Many people think that machine learning can only be applied by large companies with extensive research teams. This engine will make it easy for you to solve and shorten Research Jobs.

Dataset

Datasets or commonly called datasets are made up of attributes, data, and classes, Datasets are used for various purposes of analysis, machine learning model training, and research, Datasets are an essential component of machine learning and data analysis, and a good understanding of how to collect, process, and analyze datasets is essential for building effective and accurate models.

Random Forest Regression

Random Forest Regression (RFR): The RFR model was identified on the same dataset. This technique is an extension of the Random Forest algorithm commonly used for classification and regression tasks, which operates by building multiple decision trees. RFR is a machine that can get a higher accuracy score than other machine learning algorithms because the prediction combination works more accurately than any of the constituent models. Individual decision trees in RFRs tend to learn highly irregular patterns, i.e. they adjust to their training datasets. RFR is a way to flatten multiple decision trees, trained on different parts of the same training dataset, with the goal of reducing prediction variance. RFR can assess the impact of all explanatory variables simultaneously, and automatically sort the importance of those variables in descending order, the algorithm divides the data into subsets randomly. On each subset, the decision tree is built using a randomly selected subset of data and a feature subset, each decision tree predicts the target value for new data, the final prediction is obtained by calculating the average of the predictions of all decision trees.

Entropy

Information entropy is a basic concept in information theory that is related to any random variable. Information entropy can be interpreted as the average level of information, surprise, and uncertainty inherent in a variable's possible outcomes. The concept of information entropy was introduced by Claude Shannon in his 1948 paper, entropy was used to measure the degree of uncertainty or randomness in a dataset. The higher the entropy, the more uncertain or random the information contained in the data.

Gain

The gain is based on the decrease in entropy after the dataset which is then divided on each attribute, This Gain Method Builds a More Accurate Model: By selecting the most influential features (high gain), the model will focus more on the important factors that affect the biomass content.

Above-ground biomass (AGB)

Above-ground biomass (AGB) is a significant phenotypic index for evaluating photosynthesis capacity, healthy growth, and estimating crop yields. Accurate monitoring of AGBs helps improve agricultural fertilization management and optimize planting patterns, AGBs are an important component of forest ecosystems and their assessment plays an important role in understanding carbon dynamics, ecosystem health, and biodiversity conservation. By accurately measuring and monitoring AGBs, we can make informed decisions for sustainable forest management, climate change mitigation, and biodiversity conservation.

Digital Surface Model (DSM)

The Digital Surface Model (DSM) is a digital representation of the earth's surface that includes all objects or features above the surface, including man-made elements such as buildings, trees, and other structures.

Model Medan Digital (DTM)

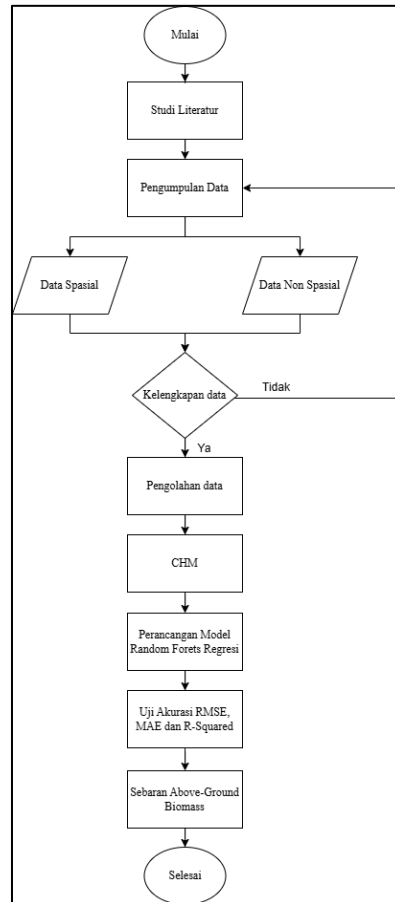
Digital Terrain Model (DTM) is a digital terrain model that contains information on the naked surface of the earth without being affected by vegetation or other man-made features.

Canopy High Model (CHM)

The Canopy Height Model (CHM) is a Digital Surface Model (DSM) that normalizes the height obtained by subtracting the Digital Terrain Model (DTM) from the DSM. On sloping grounds, points at the same height at the top of the tree appear to increase in a downward direction, based on the ground height at these points, CHM means Canopy Height Model, it is a 3D digital representation of the height of the vegetation peaks in an area. In other words, CHM describes how tall trees, shrubs, and other plants are.

METHODOLOGY

Below are the steps taken to support the research process so that the research carried out can run more structured and systematic as shown in Figure 1.

Figure 1: Research flow diagram

The following are the steps of RFR in this study:

1. The data used were 174 samples, divided into two parts with a simple 70/30 division method, namely 70% (125 samples) as training data and 30% (49 samples) as test data. In addition, to improve accuracy and avoid overfitting, the model is validated using a 5-fold cross-validation method. Avoiding overfitting this division helps ensure that the model not only memorizes patterns from training data, but can also predict well on new data. Overfitting occurs when the model matches too well with the training data so that its performance decreases on different data.
2. Measuring Model Accuracy Data testing helps assess how accurate the model's predictions are, providing an overview of real-world performance. This evaluation method helps adjust the parameters.

Using data testing, we can derive performance metrics, such as the Coefficient of determination (also known as R-squared or R^2), Root Mean Square Error (RMSE) or Mean Absolute Error (MAE), to assess the quality of the predictions. This is important for model comparisons, so you can choose the best model to use. Actual values to measure the model's performance. Some of the evaluation metrics that are often used include accuracy, and kappa, if the accuracy does not exceed 80% then it will be retested and if the accuracy value reaches the value of $>80\%$ then the model will be saved. In R studio, some packages such as random Forest, caret, can be used to build random forest models.

To ensure methodological transparency, the estimate of Above-Ground Biomass (AGB) in this study is based on the tree parameters measured in the field, namely Chest Height Diameter (DBH) and tree height (H). The biomass calculation uses allometric equations developed by Chave et al. (2005), which are: widely used in tropical forest studies:

$$AGB = 0.0673 \times (\rho \times D^2 \times H)^{0.976}$$

- AGB is biomass above ground level (kg),
- ρ is the density of wood (g/cm^3),
- D is the diameter of the tree at chest height (cm),
- H is the total height of the tree (m).

This equation has been widely validated in various tropical ecosystems and is considered appropriate for mangrove forests. For biomass prediction, this study uses the Random Forest Regression (RFR) algorithm implemented with the help of the random Forest package on the RStudio software. The parameters used in the model include:

- Number of trees (n estimators): 100
- Maximum tree depth (max depth): 10
- Separation criteria: Mean Squared Error (MSE)

FINDINGS AND DISCUSSION

Data Processing

At this stage, spatial data processing, the software used is ArcGIS to obtain CHM values and RStudio for spatial analysis.

1. Canopy High Model (CHM)

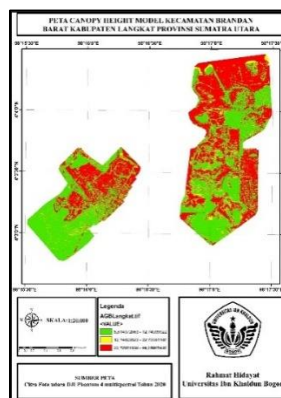
The Canopy Height Model (CHM) is obtained from the representation of tree height in the measurement area. The height of a tree is measured through the distance of the ground surface (DTM) to the canopy point or the tallest tree. The calculation of the CHM value is carried out by means of the results of the DSM – DTM stage and produces a Canopy Height Model (CHM).

Spatial Data Processing Results

Canopy Height Model (CHM)

The data obtained from the data processing of drones is CHM or Canopy Height Model which is a representation of the height of trees in the measurement area. As shown in Figure 2.

Figure 2: CHM for Regency Brandan Barat



To find out the value of CHM by using the tool function Multi Value Extract to Points software Arcgis in data that has performed the $CHM = DSM - DEM$ stage. Extracts raster cell values based on a set of

feature points and records the values in the attribute table from the output feature class. As a result of processing CHM by performing Extract Multi Values to Point in Arcgis we get a CHM value.

Eliminating Dataset Outliers

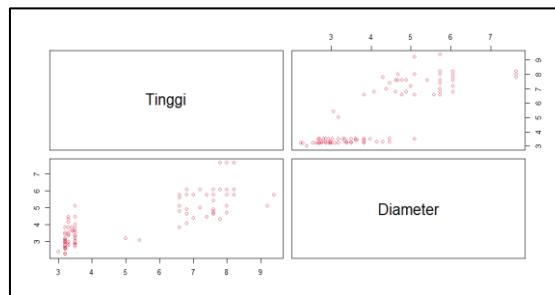
Random Forest strong enough to Outlier (because it randomly divides the data into multiple trees). However, if Outlier excessive or affect the results of the tree data model In Langkat district, the next data processing is to find the value of the regression metric, but the value obtained by R² from each data set, which is 0.1, is included in the weak category. There are three categories of groupings in the R² value, namely the strong category, the medium category, and the weak category (Hair et al., 2011). Hair et al stated that the value of R²

Table 1: Value range R²

Range of values	Level
0,75	Strong
0,50	Moderate
0,25	Weak

174 CHM value tree data in Langkat district produce R² Moderate value and there are several noise of 174 tree data points. The next stage is to eliminate the error value with Clustering which is density-based (Density-Based) from the position of data observation with the principle of grouping relatively adjacent data or DBSCAN (Spatial Grouping of Applications based on Density with Noise) using R studio software. After that export the values to excel. As shown in Figure 3.

Figure 3: Graphics Dbscan



The following is the result of CHM value data in West Brandan, Langkat Regency, because the dataset obtained from the plot is small, so the dataset from the plot will be processed.

Table 2: CHM values after DBSCAN process

No	Height (m)	Diameter (cm)
1	8,2	7,64
2	7,8	4,30
3	6,8	4,62
4	7,0	5,73
5	7,2	4,99
6	7,2	6,05
.	8,0	6,05
.	8,0	4,67

173	5	3,1
174	6,8	5,7

Biomass Value of Mangrove Trees

The biomass value of Mangrove trees in West Brandan, Langkat Regency will be carried out to produce Soil Biomass (AGB) distribution.

Table 3: Biomass Values

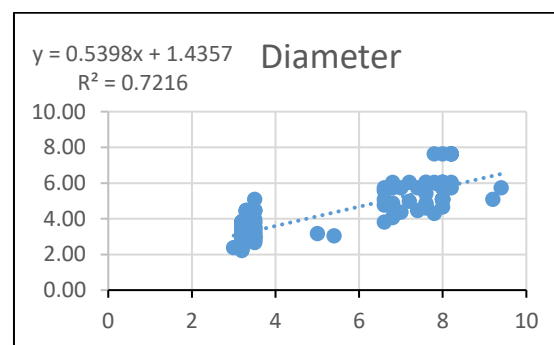
No	Tree height (m)	Tree diameter (cm)	Biomass (Kg/tree)
1	8,2	7,6	39,241
2	7,8	4,3	41,444
3	6,8	4,6	45,734
4	7	5,7	42,100
5	7,2	5,0	46,413
6	7,2	6,1	44,395
.	8	6,1	44,395
.	8	4,7	39,322
173	5	3,1	6,88
174	6,8	5,7	39,530

Random Forest Regression

Metric Regression

Regression refers to a predictive modeling problem that involves predicting numerical values. The Random Forest Regression algorithm will then be used for modeling and mapping the biomass distribution or AGB, a model that integrates height and diameter obtaining the highest accuracy ($R^2 = 0.7216$). As shown in Figure 4.

Figure 4: Metric Regression



Biomass Estimation Model with Random Forest Regression

In total, there were 174 tree samples, with 70% or 125% of the training data from the sample data, 30% or 49 from the sample data. Random Forest Regression Its implementation depends on the number of trees in order to produce a good response

Validation

The R-squared value (R^2) is used to assess how much influence a particular independent variable has on a dependent variable. The R^2 result obtained is 0.9697201. Root Mean Square Error (RMSE), Root Mean Square Error is the result of the square root of the Mean Square Error. The RMSE result obtained was 2.550351. Average Absolute Error (MAE), Grades Average Absolute Error The obtained is 1.823065.

Above-Ground Biomass (AGB)

The results of the measurement of the AGB plot of mangrove vegetation in the field showed varying values for each mangrove tree and biomass content in the plot with an average biomass of 21,136 Kg/Tree, the smallest number of 6,588 Kg/Tree and the largest number of 46,700 Kg/Tree.

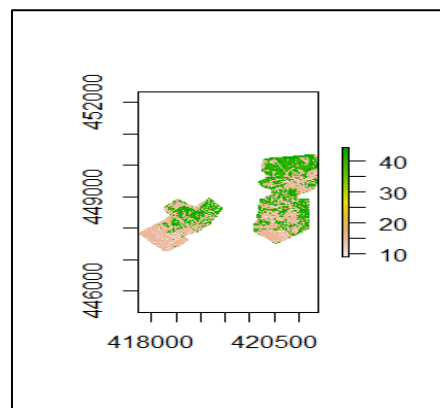
The figure below shows the results of the biomass prediction, with a value of 40 = indicating that the biomass prediction is high. As shown in Figure 4.

30 = indicates that the biomass prediction is moderate

20 = indicates that the biomass prediction is low

10 = indicates that the biomass prediction is very low

Figure 5. Distribution of AGB Brandan Barat Langkat Regency



CONCLUSION

This study succeeded in developing an accurate biomass prediction model for mangrove forests by taking into account environmental variables and relevant vegetation characteristics, resulting in an R-square value of 0.9697201, RMSE 2.550351, and Mean Absolute Error of 1.823065. Mangrove vegetation plot (AGB) plot measurements in the field showed variations in biomass values for each tree, with an average of 21,136 kg/tree, the smallest amount was 6,588 kg/tree, and the largest amount reached 46,700 kg/tree. Based on the results of the above research, in the application of the Random Forest Regression for Biomass Estimation method on mangrove trees, this study still has many shortcomings. Therefore, the input and suggestions submitted by the author so that in the future this research can develop, namely adding samples when taking field data so that the research results are more accurate.

ACKNOWLEDGEMENT

On this occasion, the author would like to express his deepest gratitude to: My beloved parents, Dad and Mother, for their prayers, encouragement, motivation, and boundless love and support over the years. Dr. Ir. Muhammad Nanang Prayudianto, M.Sc. as the Dean of the Faculty of Engineering & Science, Ibnu Khaldun University Bogor, Fitrah Satrya Fajar Kusumah, S.Kom., M.Kom., as the Head of the Computer Engineering Study Program, Faculty of Engineering & Science, Ibnu Khaldun University Bogor, Freza Riana S.Si., M.Si. The main supervisor who has given advice and input to the

author, Sahid Agustian H, S.T, M. kom. Supervisors who have provided suggestions and input to the author, Colleagues from the Geospatial Information Technology group who have helped and motivated the author in completing this final report, friends from the Geospatial Information Technology group 2017 who have helped and encouraged the author in completing this scientific paper. The author realizes that the results of this writing are still far from perfect. Therefore, the author expects constructive criticism and suggestions to improve the quality of this final report. In conclusion, hopefully this final project report will provide benefits for the world of education.

REFERENCES

- A. Function et al. (2024). "Scientific Journal of Planning, Development, and Regional Development (JIP3W) Appropriate Email Articles." *Scientific Journal of Regional Planning, Development, and Development, Vol. –*.
- Alexander, C., Korstjens, A. H., & Hill, R. A. (2018). Influence of micro-topographic characteristics and crown parameters on tropical forest tree height estimation using LiDAR canopy height models. *International Journal of Applied Earth Observation and Geoinformation, 65*, 105–113. <https://doi.org/10.1016/j.jag.2017.10.009>
- Chave, J., Andalo, C., Brown, S., Cairns, M. A., Chambers, J. Q., Eamus, D., ... Yamakura, T. (2005). Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia, 145*(1), 87–99. <https://doi.org/10.1007/s00442-005-0100-x>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, e623. <https://doi.org/10.7717/peerj-cs.623>
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). SUPPORTS VECTOR ENGINE OPTIMIZATION USING PARTICLE SWARM OPTIMIZATION FOR MULTICLASS DATASET CLASSIFICATION. *Journal of Responsiveness, 5*(1), 120–126. No DOI.
- F., Agustina, D., Tjahjadi, M. E., & Zulhan, I. (2024). Test the accuracy of the slope gradient digital terrain model from aerial photo shooting (case study: Lowokwaru Regency, Malang City). *Journal of Geoscience, 10*(1), 1. <https://doi.org/10.12962/j25023659.v10i1.19469>
- Fitri, E. (2023). Comparative analysis of linear regression methods, random forest regression, and gradient-driven tree regression methods for home price prediction. *Journal of Applied Computation Science and Technology, 4*(1), 58–64. <https://doi.org/10.52158/jacost.v4i1.491>
- Handayani, R. N., & Tjahjadi, D. (2023). SUPPORTS VECTOR ENGINE OPTIMIZATION USING PARTICLE SWARM OPTIMIZATION FOR MULTICLASS DATASET CLASSIFICATION. *Journal of Responsiveness, 5*(1), 120–126.
- Jeon, G. (2021, Agustus 1). Information entropy algorithms for image, video, and signal processing. *Entropy, 23*(8), 926. <https://doi.org/10.3390/e23080926>
- Liu, Y., dkk. (2022). Remote sensing estimation of potato above-ground biomass using spectral and spatial features from high-definition digital camera imagery. *Computers and Electronics in Agriculture, 198*, 107089. <https://doi.org/10.1016/j.compag.2022.107089>
- Navarro, A., Young, M., Allan, B., Carnell, P., Macreadie, P., & Ierodiaconou, D. (2020). Application of unmanned aerial vehicles (UAVs) to estimate above-ground biomass in mangrove ecosystems. *Remote Sensing of Environment, 242*, 111747. <https://doi.org/10.1016/j.rse.2020.111747>
- Nurhidayat, M. Z., Aditya, T., Zannah, A. L., & Firdausia, S. (2023). Modeling above-ground biomass estimation using forest canopy height and density in secondary peat swamp forests of Kalimantan. In *IOP Conference Series: Earth and Environmental Science, 1276*(1), 012001. <https://doi.org/10.1088/1755-1315/1276/1/012001>
- Ouedraogo, I., Defourny, P., & Vanclooster, M. (2019). Application of random forest regression and its performance comparison with multiple linear regression in modeling groundwater nitrate concentration at continental scale in Africa. *Hydrogeology Journal, 27*(3), 1081–1098. <https://doi.org/10.1007/s10040-018-1900-5>

- Patar Pasaribu, R., Kabul Pranoto, A., Sewiko, R., Afwafiah, E. (2022). Mapping of mangrove distribution with remote sensing on the coast of Karawang Regency.
- Pham, T. D., Yoshino, K., Le, N. N., & Bui, D. T. (2020). Estimating mangrove above-ground biomass using integrated gradient-boosted decision tree algorithm with Sentinel-2 and ALOS-2 PALSAR-2 data at Can Gio biosphere reserve, Vietnam. *Remote Sensing*, 12(5), 777. <https://doi.org/10.3390/rs12050777>
- Prasojo, B., & Haryatmi, E. (2021). Analysis of loan eligibility prediction using random forest method. *National Journal of Information Technology and Systems*, 7(2), 79–89. <https://doi.org/10.25077/teknosi.v7i2.2021.79-89>
- Purosojo, B., & Haryatmi, E. (2021). Analysis of loan eligibility predictions using the random forest method. *National Journal of Information Technology and Systems*, 7(2), 79–89. <https://doi.org/10.25077/teknosi.v7i2.2021.79-89>
- Raharjo, B. (2021). Machine learning. Ungaran City.
- Sari, K., Ismail, A. Y., & Hendrayana, Y. (2023). Potential biomass and carbon reserves in the mangrove forest area of Kanci Kulon Village, Astanajapura District, Cirebon Regency. *Journal of Forestry and Environment*, 5(1). <https://doi.org/10.25134/jfe.v5i1.9048>
- TS Harapan, et al. (2021). Biomass estimation on the soil of *Syzygium aromaticum* uses the structure of the Structure from Motion (SfM) perspective of UAV in the Paninggahan agroforest area, West Sumatra. *Journal of Biology UNAND*, 9(1), 39–46. <https://doi.org/10.25077/jbioua.9.1.39-46.2021>
- Wirantiko, M., Handayani, H. H., & Cahyono, A. B. (2020). A study on DTM determination using slope-based filtering and grid-filtering (Case study: Wonokromo and Lontar sub-districts, Surabaya). –, 16(1), 46–56.
- Yaqin, Y., et al. (2020). Health assessment analysis of mangrove forest in East Lampung. –. No DOI.
- Z. Hidayah, H. A. Rachman, & A. R. As-Syakur. (2023). Mapping mangrove forest condition in Madura Strait coastal area using mangrove health index approach based on Sentinel-2 satellite imagery. *Maior Geographia Indonesia*, 37(1), 84. <https://doi.org/10.22146/mgi.78136>