RESEARCH ARTICLE • OPEN ACCESS

A Study on Prediction Intervals Produced Using Quantile Regression Forest With and Without Variable Selection

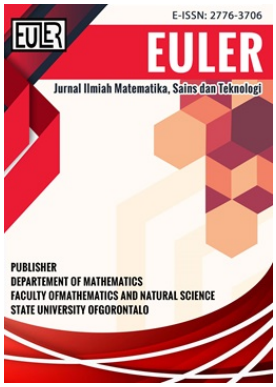
Megawati, Bagus Sartono, and Sachnaz Desta Oktarina

Volume 13, Issue 3, pp. 319–327, Dec. 2025

Received 10 September 2025, Revised 20 October 2025, Accepted 1 November 2025, Published 1 December 2025

To Cite this Article : M. Megawati, B. Sartono, and S. D. Oktarina,"A Study on Prediction Intervals Produced Using Quantile Regression Forest With and Without Variable Selection", *Euler J. Ilm. Mat. Sains dan Teknol.*, vol. 13, no. 3, pp. 319–327, 2025, *https://doi.org/10.37905/euler.v13i3.34392*

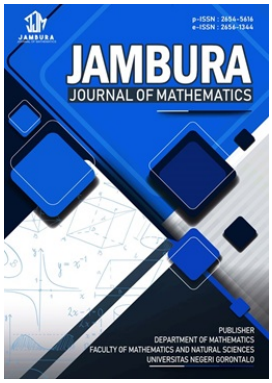
JOURNAL INFO • EULER : JURNAL ILMIAH MATEMATIKA, SAINS DAN TEKNOLOGI



Homepage	:	http://ejurnal.ung.ac.id/index.php/euler/index
Journal Abbreviation	:	Euler J. Ilm. Mat. Sains dan Teknol.
Frequency	:	Three times a year
Publication Language	:	English (preferable), Indonesia
DOI	:	https://doi.org/10.37905/euler
Online ISSN	:	2776-3706
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/euler/oai
Google Scholar ID	:	QF_r-_gAAAAJ
Email	:	euler@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

Jambura Journal of Biomathematics



Jambura Journal of Mathematics

Jambura Journal of Mathematics Education

Jambura Journal of Probability and Statistics

Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi, Volume 13, Issue 3, Pages 319–327, Dec. 2025 *https://doi.org/10.37905/euler.v13i3.34392*

Research Article

Check for updates

A Study on Prediction Intervals Produced Using Quantile Regression Forest With and Without Variable Selection

Megawati[1], **Bagus Sartono**[1,*], **Sachnaz Desta Oktarina**[1]

[1] *School of Data Science, Mathematics and Informatics, IPB University, Bogor 16680, Indonesia*

ARTICLE HISTORY
Received 10 September 2025
Revised 20 October 2025
Accepted 1 November 2025
Published 1 December 2025

KEYWORDS
Adaptive-LASSO
Oil Palm Productivity
Prediction Interval
Quantile Regression Forest
Variable Selection

ABSTRACT. *Quantile Regression Forest (QRF) is a method that utilizes the random forest algorithm to estimate the conditional distribution of response variables and form quantile prediction intervals. However, when there is a high correlation between covariates, QRF performance may decrease due to the multicollinearity effect, thereby reducing the accuracy of the prediction interval for the target variable. In linear models, multicollinearity must be addressed because it can cause large variances. This study contributes to enhancing the reliability of prediction intervals in correlated data through the integration of adaptive-LASSO with QRF. Specifically, it examines the role of variable selection by the adaptive LASSO method on the performance of the QRF prediction interval in the simulated data, and the best model obtained in the study is then applied to predict the interval in the productivity data of oil palm fresh fruit bunches. The results of the study show that variable selection is proven to produce coverage close to the target prediction interval. In addition, the QRF model with variable selection applied to the productivity data of oil palm fresh fruit bunches produces a good prediction interval.*

1. Introduction

Machine learning-based prediction algorithms are widely used for decision-making. However, the resulting predictions are often accompanied by uncertainties that need to be considered. Quantile regression (QR) addresses this issue by estimating not only central values, such as the mean or median, but also other quantiles of the response variable, including lower and upper bounds. This flexibility allows researchers to analyze how covariates influence the entire distribution of outcomes [1]. However, because QR is a linear method, its performance can be compromised when assumptions, such as the absence of multicollinearity, are violated.

Multicollinearity can destabilize QR estimates, particularly at extreme quantiles [2]. To address this limitation, Meinshausen [3] proposed the non-linear quantile regression forest (QRF), which extends random forest (RF) to estimate conditional quantiles. QRF retains all observations in each tree node, enabling accurate quantile estimation and reliable prediction intervals. Prediction intervals provide informative ranges rather than point estimates, reflecting uncertainty [4]. The width of the hose reflects the accuracy of the estimate and how likely it is that the hose can cover the actual parameters [5].

QRF estimates depend on predictor variables, high correlations among predictors can increase uncertainty. Very high correlations may lead to multicollinearity, which inflates error variance and widens prediction intervals [6, 7]. Despite this, QRF can identify the most influential variables by filtering out highly correlated predictors, for example, those with correlations above [8]. On the other hand [9], the integration of non-linear modeling with adaptive-LASSO regularization yields parameter estimates and covariance matrices that demonstrate asymptotic consistency, Asrirawan et al. [6] highlighting the importance of careful variable selection. As a nonparametric ensemble method based on decision trees, QRF's primary strength lies in its robustness to heteroskedasticity, autocorrelation, and model misspecification [10].

This study investigates the role of variable selection using the adaptive-LASSO method on the performance of QRF prediction intervals using simulated data, then the best model obtained in the study will be applied to predict the interval in the productivity data of oil palm fresh fruit bunches. Palm oil is an agricultural commodity with high economic value and plays a role in supporting food security. In recent years, palm oil has also been recognized as one of the most promising renewable energy sources [11–13]. Firdawanti et al. [14] the presence of free variables that are suspected to have an effect at different times on oil palm production or there is a lagged effect.

Previous studies have not systematically examined how predictor correlations affect QRF prediction intervals, nor compared QRF with and without adaptive-LASSO variable selection across multiple confidence levels. This study addresses these gaps through a simulation varying predictor correlations and evaluating coverage rates at 90%, 95%, and 99%. Therefore, this study aims to: (1) evaluate the role of variable selection on the performance of the hose estimator produced by QRF. Answered using a simulation study involving generated data, (2) implementing the QRF approach in obtaining a prediction interval for palm

* Corresponding Author.

Email : *bagusco@apps.ipb.ac.id* (B. Sartono)
Homepage : http://ejurnal.ung.ac.id/index.php/euler/index / E-ISSN : 2776-3706

Table 1. List of study variables

Code	Variable	Type	Code	Variable	Type
y	Productivity	Numerical	x_{10}	MOP Fertilizer	Numerical
x_1	Rainy days	Numerical	x_{11}	HGFB Fertilizer	Numerical
x_2	Rainfall	Numerical	x_{12}	$CuSO_4$ Fertilizer	Numerical
x_3	Sunlight duration	Numerical	x_{13}	Dolomite Fertilizer	Numerical
x_4	Air temperature	Numerical	x_{14}	Status	Categorical
x_5	Plant age	Numerical	x_{15}	Land type	Categorical
x_6	Land area	Numerical	x_{16}	Topography	Categorical
x_7	Plant density	Numerical	x_{17}	Varieties	Categorical
x_8	NPK Fertilizer	Numerical	x_{18}	Soil type	Categorical
x_9	Urea Fertilizer	Numerical			

oil FFB productivity. The best method is one that provides a coverage rate equal to the confidence interval.

2. Methods

2.1. Research Design and Tools

This computational study employed secondary data obtained from an Indonesian oil palm company, comprising 13 categorical and 5 numerical variables observed from January 2019 to September 2023. The analysis was conducted using *R Studio 4.2.2*, utilizing the *quantregForest package* for QRF modeling and the*glmnet package* for adaptive-LASSO variable selection. This study employs both simulated and empirical datasets.

2.2. Simulation Study Procedure

The simulation study serves to establish the foundational model and algorithm for subsequent analyses. The simulated data analysis will be conducted according to the following procedure:

1. Set the correlation matrix values to 0.1 (low correlation), 0.5 (moderate correlation), and 0.9 (high correlation).
2. Specify the sample size as n=70.000.
3. Generate 15 numerical predictor variables from a multivariate normal distribution $\mathbf{X} \sim \mathcal{MN}(\mathbf{0}, \Sigma)$
4. Generate 5 categorical predictor variables as follows:
   a. Generate 5 numerical variables from a multivariate normal distribution $\mathbf{X} \sim \mathcal{MN}(\mathbf{0}, \Sigma)$.
   b. Determine the first quartile Q_1 and second Q_2 of each generated variable
   c. Categorize each variable into three categories, denoted as A, B, and C, according to the following criteria: category A data value $\leq Q_1$, category B $Q_1 <$ data value $\leq Q_2$, category C data value $> Q_2$
5. Set the vector of parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{11})$. The relationship between the numerical predictor X_i and the response Y is assumed to be nonlinear in the form of a stepwise function, where each X_i is divided into 11 segments based on its decile values.
6. Generate $\varepsilon \sim \mathcal{N}(0, 1)$.
7. Generate the response model Y according to the equation:

$$Y = \sum_{i=1}^{15} \left(\sum_{j=1}^{11} \beta_{ij} (Q_{ij} \leq X_i \leq Q_{i,j+1} \right) + \sum_{i=16}^{20} \left(\sum_{j=1}^{3} \beta_{ij} (Q_{ij} = X_i \right) + \varepsilon,$$

   where: β_{ij} is the regression parameter for the i-th variable in the j-th segment, X_i is the i-th predictor variable, and Q_{ij} is the i-th predictor's j-th decile value.
8. Data Preparation for QRF modeling.
   a. Scenario 1 – Complete data without variable selection. All simulated data were used in the analysis, and 24 lags were created for the last five numerical variables.
   b. Scenario 2 – Data after variable selection using adaptive-LASSO.
      i. Adaptive-LASSO was applied to estimate the model and obtain the selected variables with adaptive weights for each predictor. Adaptive weights were calculated for each predictor
      ii. Create 24 lags for the last five numerical variables If any of the selected variables are among these five numerical variables, no lag is created for those variables.
9. QRF modelling.
   a. Build QRF prediction models using two approaches: without variable selection and with variable selection (adaptive-LASSO).
   b. Determine conditional quantile predictions $\tau =$ (0.005, 0.025, 0.05, 0.5, 0.95, 0.975, 0.995).
   c. Define prediction intervals with 90%, 95% and 99% confidence levels.
10. Calculate the coverage rate for each generated prediction interval.
11. Repeat steps 1–10 100 times to assess the consistency of the model in generating prediction intervals.
12. Evaluate the performance of the best QRF model, with and without variable selection, for subsequent prediction interval analysis on empirical data.

Note: the QRF model was constructed based on the RF parameters with ntree = 250 and nodesize values ranging from 5 to 50, selected according to the smallest RMSE without repetition.

2.3. Implementation Methodology of Oil Palm FFB Prediction Intervals

The data used in this study are secondary data with productivity of fresh fruit bunch (FFB) as the response variable and 18 predictor variables. The list of variables is shown in Table 1.

Several constraints complicate the modeling and forecasting of oil palm production. One significant challenge is the suspected variability in predictor variables, where different factors may influence production at differing time intervals (lag effects).

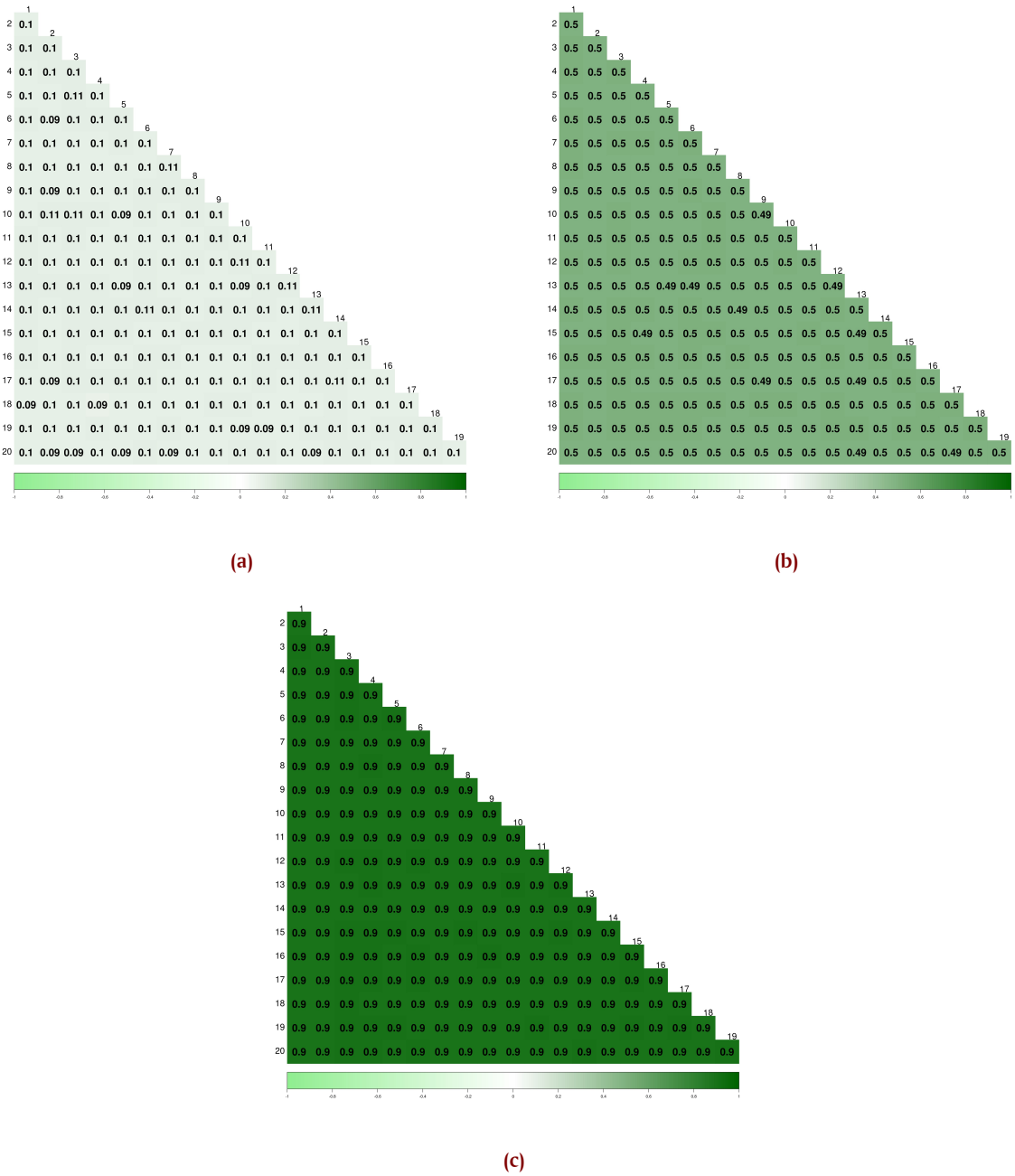


Figure 1. Correlation matrix between predictor variables in the simulation data in three scenarios, namely (a) low correlation scenario; (b) medium correlation scenario; (c) high correlation scenarios. Data were generated through simulation as described in Section 3.1

The analysis follows these steps:

1. Data standardization is performed to ensure that the variables have the same value range.
2. Data exploration is conducted to observe the overall data patterns.
3. Analysis is carried out based on the best model obtained from the simulation study.
    a. Train the QRF model (with or without variable selection), as done in Steps 8–10 of the simulation phase.
    b. Interpret the QRF model based on the prediction intervals and coverage rates to evaluate the adequacy of the desired confidence levels.

3. Results and Discussion

3.1. Simulation Study

The simulation model was constructed using data consisting of 10 scenarios, as described in the methodology section. The data were designed by considering the correlation levels among predictor variables. Each scenario represents different inter-variable relationships to assess the method's performance under varying degrees of multicollinearity. Figure 1 presents the correlation matrices for the three correlation scenarios.

The increasing correlation among variables is indicated by the darker shades in the correlation matrix. Furthermore, the relationship between variables X and Y is designed in a non-linear form to reflect the complexity that may be encountered in empir-

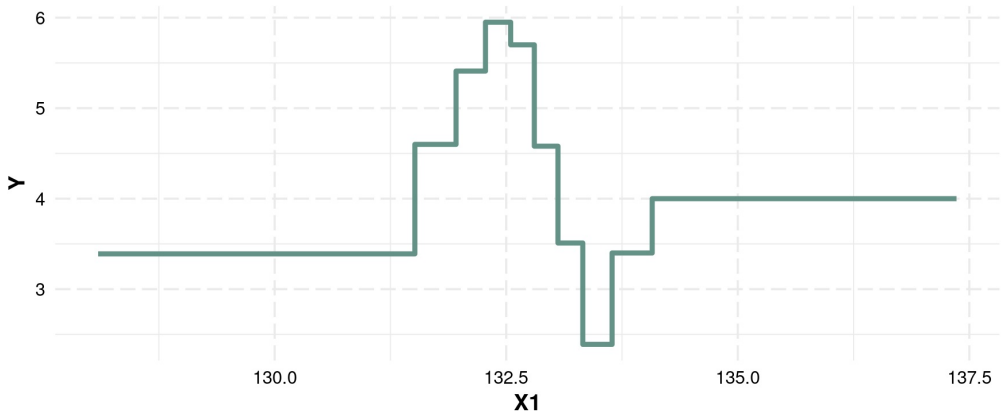


Figure 2. Piecewise relationship patterns between the response variable and the predictor variable X_1. Data were simulated to reflect non-linear relationships commonly found in empirical datasets

Table 2. Average QRF coverage rate (90% prediction interval) in low, medium, high correlation scenarios, with and without variable selection

Correlation	Coverage rate without selection (%)	Coverage rate with selection (%)	p-value
0.1	94.5	92.8	0.00
0.5	93.8	91.9	0.00
0.9	92.2	90.1	0.00

Significant at 5% significance level (Wilcoxon signed-rank test)

Table 3. The average QRF coverage rate (95% prediction interval) in low, medium, high, and contrast tests between and without variable selection

Correlation	Coverage rate without selection (%)	Coverage rate with selection (%)	p-value
0.1	96.6	95.5	0.00
0.5	96.2	94.9	0.00
0.9	94.9	93.3	0.00

Significant at 5% significance level (Wilcoxon signed-rank test)

ical data. All numerical variables are simulated to follow a piecewise relationship pattern, as illustrated in Figure 2.

Figure 2 exhibits a piecewise regression or nonlinear pattern. This pattern arises because, during the model construction process, the data are divided into several categories based on specific deciles. Prior to implementing the QRF method, the model was developed using the parameters of the RF model by setting a fixed value of ntree (in this simulation study, ntree = 250) and varying the nodesize. The optimal parameter was subsequently determined by identifying the nodesize value that yielded the smallest root mean square error (RMSE). Figure 3 shows the results of the RMSE obtained for both scenarios (with and without variable selection) at each correlation.

Optimization of RF parameters shows that at a correlation of 0.1, the smallest RMSE without variable selection was found at nodesize 26, while with variable selection, the smallest RMSE was obtained at nodesize 7. At a correlation of 0.5, the smallest RMSE without variable selection occurred at nodesize 22, whereas with variable selection, the smallest RMSE was found at nodesize 5. Meanwhile, at a correlation of 0.9, the smallest RMSE without variable selection was achieved at nodesize 8, while with variable selection, the smallest RMSE occurred at nodesize 10.

The obtained optimal parameters were applied to the QRF model to analyze prediction intervals involving 20 predictor variables. In the simulation without variable selection, 5 predictor variables were designed to have 24 lag effects for each correlation level (0.1, 0.5, and 0.9), specifically predictors 11 to 15, resulting in a total of 135 variables. The complete set of variables includes 10 numerical variables, 5 categorical variables, and 120 lag effect variables. On the other hand, in the simulation with variable selection using the adaptive-LASSO method, the number of consistently selected variables at each correlation level was observed.

Under the scenario with a correlation of 0.1, 14 variables were consistently selected in 98 repetitions, while 12 and 15 variables were selected only once. At a correlation of 0.5, 9 variables were selected in 99 repetitions, whereas 10 other variables appeared only once. At a correlation of 0.9, the number of selected variables was more varied, with combinations of 12, 11, 10, and 13 variables selected in 40, 37, 13, and 10 repetitions, respectively. The QRF prediction interval analysis also considered the lag effects on the five selected predictor variables. Variables that were not selected were excluded from lag formation.

The determination of the best approach between variable selection and no variable selection—was conducted through hypothesis testing on the average coverage rate to determine

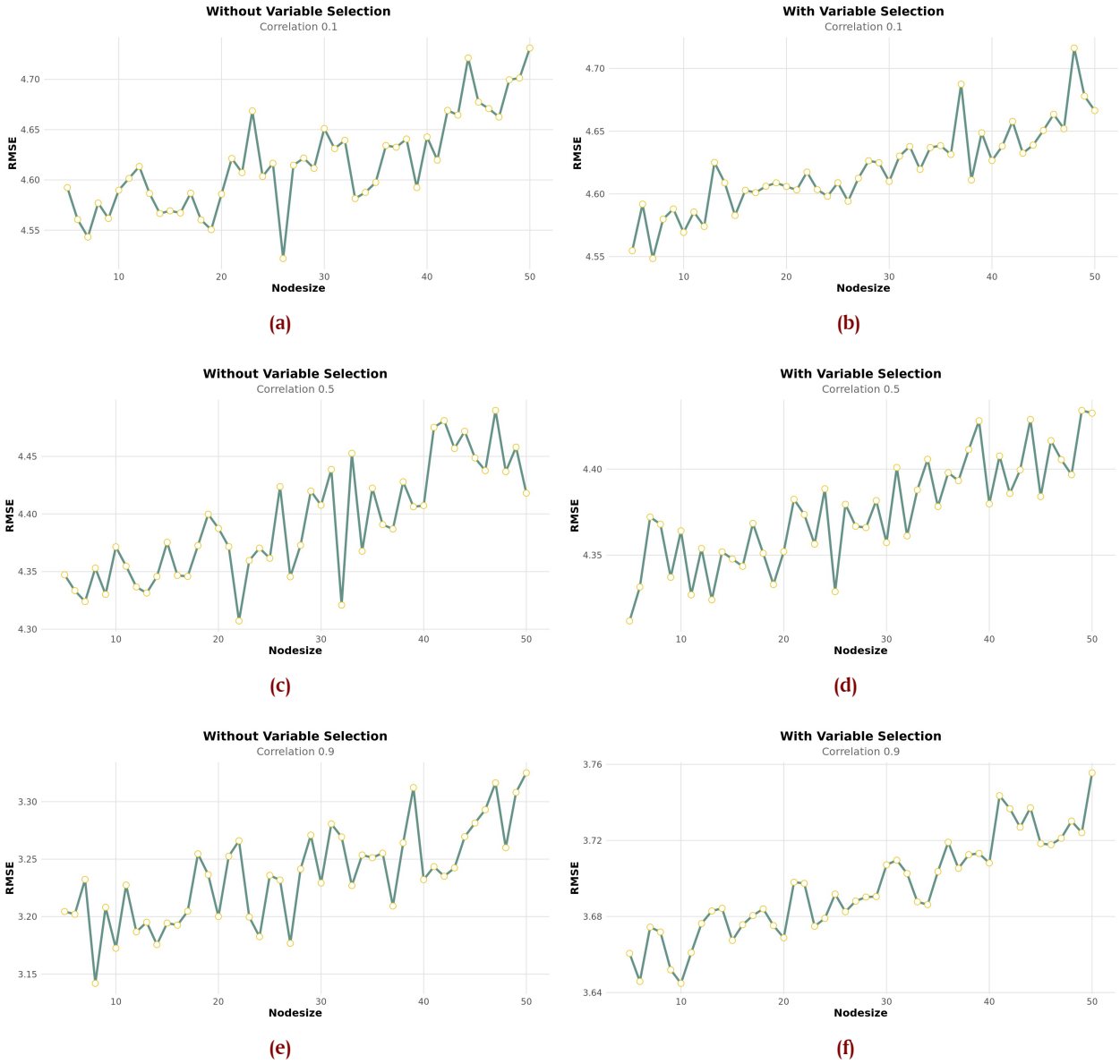


Figure 3. RMSE values based on the parameters ntree = 250 and nodesize = 5 to 50. (a) low correlation scenario without variable selection; (b) low correlation scenario with variable selection; (c) medium correlation scenario without variable selection; (d) medium correlation scenario with variable selection; (e) high correlation scenario without variable selection; (f) high correlation scenario with variable selection

Table 4. Average QRF coverage rate (99% prediction interval) in low, medium, high correlation scenarios, with and without variable selection

Correlation	Coverage rate without selection (%)	Coverage rate with selection (%)	p-value
0.1	97.5	96.6	0.00
0.5	97.1	96.0	0.00
0.9	96.0	94.8	0.00

Significant at 5% significance level (Wilcoxon signed-rank test)

whether there was a significant difference between the two methods. The hypotheses tested were as follows:
H_0: there is no significant difference in coverage rate between variable selection and no variable selection.
H_1: there is a significant difference in coverage rate between variable selection and no variable selection.

The decision was made with a 95% confidence level. If the p-value $> \alpha(\alpha = 0.05)$, then H_0 is not rejected, meaning there is no significant difference between the average coverage rates of the selection and no-selection methods. Conversely, if the p-value $\leq \alpha$, then H_0 is rejected, indicating a significant difference

between the two methods in producing coverage rates. The evaluation of the QRF model in generating 90%, 95%, and 99% prediction intervals is presented in Table 2–Table 4.

Table 2–Table 4 show that the average coverage rate decreases as the correlation increases, both with and without variable selection. For the 90% and 95% prediction intervals, this decline aligns with expectations as it approaches the target coverage rate. However, for the 99% prediction interval, the decrease moves further away from the desired target coverage rate. The probability values at all prediction interval levels are 0.00, meaning the p-value $< 5\%$. The application of the best model to empirical data heavily depends on the correlation between variables, making correlation exploration a crucial step.

3.2. Empirical Data Analysis

The empirical study was conducted using palm oil FFB productivity data obtained from a private palm oil company in Indonesia. The empirical data shares similarities with the simulation outline: it is predominantly numerical rather than categorical, exhibits a non-linear pattern, and contains variables with lag effects. The approach for selecting the optimal model for the empirical data involved examining inter-variable correlations and aligning these with the outcomes from the six previously obtained simulation scenarios.

3.2.1. Data Exploration

FFB (Fresh Fruit Bunch) productivity data of oil palm was observed monthly across 1.169 distinct blocks from January 2019 to September 2023, with a total land area of 33,367.45 hectares. The palm oil FFB data to be used contains mixed-type variables, encompassing both numerical and categorical data.

The response variable in this study is palm oil production, observed from January 2019 to September 2023. Figure 4 displays a boxplot of the response variable, namely monthly FFB productivity (tonnes/ha). Based on the figure, it is evident that the empirical data exhibits a fluctuating pattern.

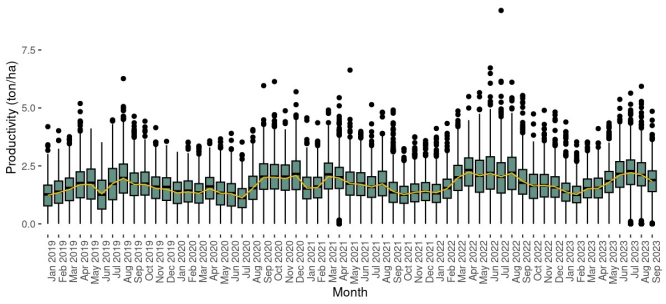


Figure 4. Oil palm FFB productivity over a 57-month period

Figure 4 shows that the productivity of palm oil FFB increased from the beginning to the middle of the year. This increase is likely influenced by optimal growth conditions after the recovery phase at the end of the previous year. After reaching its peak, productivity tends to fluctuate. This means the highly varied age of the plants also indicates that some may have surpassed their peak productive period, as illustrated in Figure 5.

Figure 5 shows that palm oil FFB (Fresh Fruit Bunch) productivity becomes evident at 3 years of age and continues to increase with plantation age. Upon reaching 16 years, most productivity

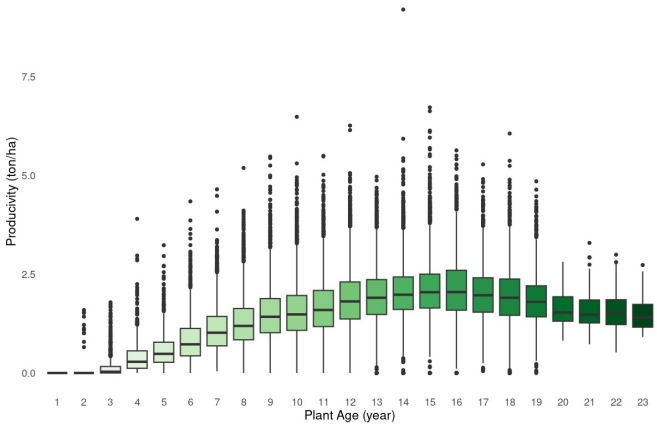


Figure 5. Oil palm FFB productivity by plant age

levels tend not to show significant further increases. According to the findings of [15] The age range of 15-25 years represents a declining yield phase as the frond production rate and number of fruit bunches begin to decrease. This condition can be caused by various factors, one of which is the rainfall factor as in Figure 6.

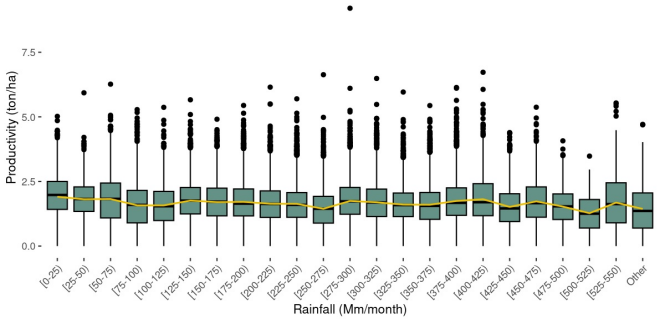


Figure 6. Relationship between FFB productivity and rainfall level

Climate change, particularly unpredictable rainfall patterns, causes a significant decline in palm oil yields. Consequently, supply is disrupted, and crude palm oil prices fluctuate [16]. Figure 6 reveals that the relationship between rainfall and oil palm FFB productivity is non-linear. Productivity does not consistently increase with higher rainfall but instead exhibits fluctuating patterns across different categories. This suggests that rainfall has an indirect influence and likely interacts with other factors in affecting productivity. The next step involves analysing the correlation among numerical predictor variables to enhance the performance of the QRF model. Based on Figure 7, the highest correlation between predictor variables is recorded at 0.7, while most variables show relatively low correlations. The NPK fertilizer variable has a moderately strong negative correlation with HGFB fertilizer, with a correlation coefficient of -0.4. Only a few variables exhibit correlations above 0.5, whereas the majority demonstrate weak relationships with one another.

These findings show that palm oil FFB data have a relatively low correlation between predictor variables. Under these conditions, the simulation results show that variable selection is important to improve the performance of the prediction model. The findings of this simulation are consistent with those reported by

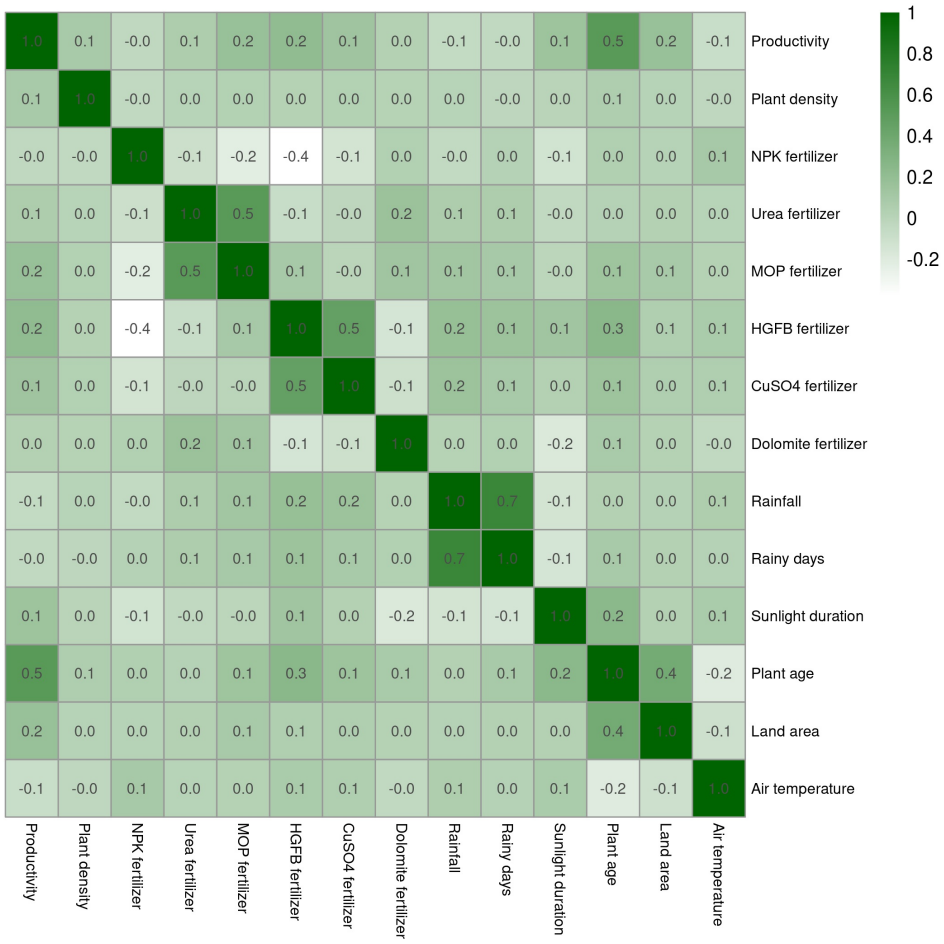


Figure 7. Correlation matrix among the predictor variables

[6], indicating that even when the data exhibit low correlations among predictor variables, the QRF method still produces narrower coverage rates compared to models without variable selection.

3.2.2. Variable Selection using adaptive-LASSO and Lagged Transformation

Variable selection involved 13 numerical variables and 5 categorical variables, which were transformed into dummy variables. If the selected dummy categories exceeded 50%, the corresponding variable was included in the model analysis. From the selection of 18 variables using the adaptive-LASSO method, 7 variables were selected, and 2 additional selected variables were expanded into 24 lags. As a result, the total number of variables used in the QRF analysis was 55 variables.

Lag variable construction was applied to predictor variables that have time-dependent effects, such as rainfall, number of rainy days, temperature, and duration of sunlight exposure, with lags ranging from 1 to 24. In addition to weather factors, fertilization also has an indirect effect on oil palm FFB production. The types of fertilizers used include NPK, Urea, $HGFB$, $CuSO_4$, and Dolomite. The available data for all types of fertilizers are annual, so the lag variable constructed for each type of fertilizer was limited to lag 24. The lag added for each fertilizer represents a cumulative effect over two years; therefore, lags 1 to 24, as used for other variables, could not be applied. The selected variables and the lag variable construction are presented in Table 5.

Table 5. Selected variables and lag variable list

Selected	Variables
9	Plant density, Plant age, Soil type order, MOP fertilizer HGFB fertilizer, CuSO4 fertilizer, Rainfall, Rainy days
Lag	**Variables**
1 - 24	Rainfall, Rainy days
24	MOP fertilizer, HGFB fertilizer, CuSO4 fertilizer

Variables that exhibit time-dependent effects but were not selected will not be included in lag formation. Additionally, there are three categorical variables for which more than 50% of their dummy variables were not selected.

3.2.3. Variable Selection using adaptive-LASSO and Lagged Transformation

This analysis aims to compare model performance in generating accurate and reliable prediction intervals. Quantile predictions are computed at multiple τ values to obtain the range of predicted values. Subsequently, prediction intervals are constructed at 90%, 95%, and 99% confidence levels, with coverage rates calculated for each interval as part of model evaluation. Figure 8 presents the coverage rate results for productivity prediction intervals of the Socfindo variety.

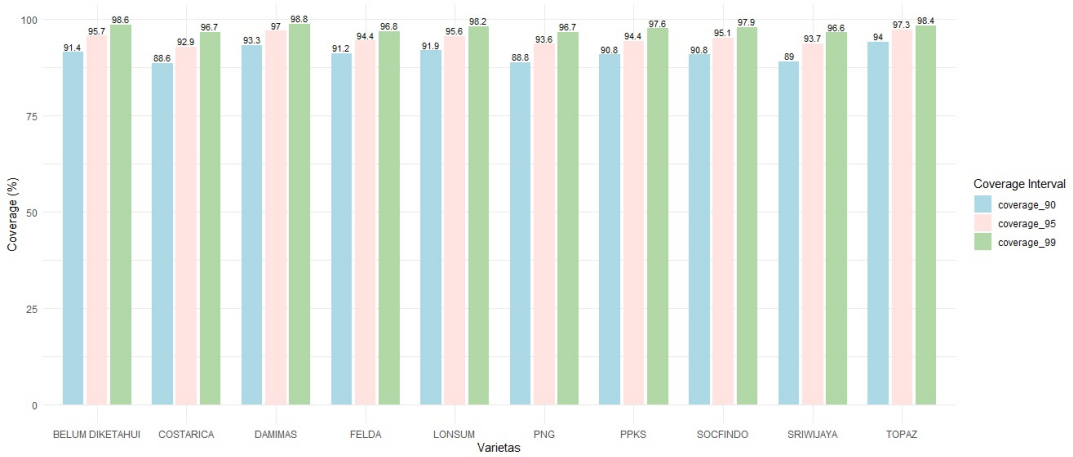


Figure 8. Overage rate of productivity prediction interval of Socfindo variety

Figure 9. Prediction intervals of FFB productivity for 250 observations randomly selected from Socfindo variety: (a) 90% prediction intervals have coverage rate of 90.8%; (b) 95% prediction intervals have coverage rate of 95.1%; (c) 99% prediction intervals have coverage rate of 97.9%

Figure 8 demonstrates that the prediction intervals generated by QRF exhibit good coverage rates across nearly all varieties. This indicates that the model reliably produces accurate prediction intervals for different types of oil palm cultivars. Among the varieties, Socfindo and PPKS show particularly strong coverage rate results compared to others, while Topaz and Damimas perform less optimally, yielding higher coverage rates than the expected prediction intervals. Figure 9 provides a detailed plot of the socfindo variety across different prediction interval levels.

The green points indicate locations that fall within the prediction interval, while the red points indicate locations outside the interval. Overall, as the confidence level increases, the prediction interval becomes wider, which in turn increases the coverage rate. The model tends to provide more conservative predictions at higher confidence levels, leading to a higher probability that the actual values fall within the prediction interval. Figure 9 which shows the 95% prediction interval, demonstrates the best performance as its coverage rate precisely meets the target. At the 90% confidence level, the coverage rate is slightly higher than expected, indicating a more conservative model. At the 99% confidence level, although the coverage rate is considerably high, the excessively wide prediction interval may reduce the interpretability and practical usefulness of the predictions.

4. Conclusion

This study demonstrates that the application of variable selection significantly affects the performance of QRF prediction intervals, with notable differences between models with and without variable selection across all prediction interval levels (90%, 95%, and 99%). Simulation results indicate that the decision to apply variable selection depends on the correlation level among predictors, with variable selection recommended for low-correlation datasets to achieve coverage rates closer to the target.

For oil palm FFB productivity data, where most predictors exhibit low correlation, QRF was applied with variable selection, producing prediction intervals with satisfactory coverage for nearly all varieties. Specifically, Socfindo and PPKS cultivars achieved coverage rates near the expected levels, while Topaz and Damimas produced slightly wider intervals. These findings highlight that applying variable selection according to the empirical characteristics of the data can improve prediction accuracy and support more reliable decision-making in oil palm plantation management.

Author Contributions. **Megawati:** Methodology, software, resources, data curation, visualization and writing—original draft preparation. **Bagus Sartono:** Conceptualization, validation, investigation, writing—review and supervision. **Sachnaz Desta Oktarina:** writing—review and supervision and validation. All authors discussed the results and contributed to the final manuscript.

Acknowledgement. Appreciation is expressed to the Ministry of Higher Education, Science and Technology of the Republic of Indonesia.

Funding. This research was funded by Indonesian Education Scholarship (BPI), Center for Higher Education Funding and Assessment (PPAPT), and Indonesian Endowment Fund for Education (LPDP).

Conflict of interest. The authors declare that there are no conflicts of interest related to this article.

Data availability. Not applicable.

References

[1] R. Koenker and K. F. Hallock, "Quantile regression," *J. Econ. Perspect.*, vol. 15, no. 4, pp. 143–156, 2001, doi: 10.1257/jep.15.4.143.

[2] C. Davino, R. Romano, and D. Vistocco, "Handling multicollinearity in quantile regression through the use of principal component regression," *Metron*, vol. 80, no. 2, pp. 153–174, 2022, doi: 10.1007/s40300-022-00230-3.

[3] N. Meinshausen, "Quantile regression forests," *J. Mach. Learn. Res.*, vol. 7, pp. 983–999, 2006.

[4] Y. Fang, P. Xu, J. Yang, and Y. Qin, "A quantile regression forest based method to predict drug response and assess prediction reliability," *PLoS One*, vol. 13, no. 10, pp. 1–16, 2018, doi: 10.1371/journal.pone.0205155.

[5] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed., Pacific Grove: Duxbury Thomson Learning, 2002.

[6] A. Asrirawan, K. A. Notodiputro, and B. Sartono, "Improving Accuracy of Prediction Intervals of Household Income Using Quantile Regression Forest and Selection of Explanatory Variables," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 4, pp. 1915–1926, 2023, doi: 10.30598/barekengvol17iss4pp1915-1926.

[7] D. L. Shrestha and D. P. Solomatine, "Machine learning approaches for estimation of prediction interval for the model output," *Neural Networks*, vol. 19, no. 2, pp. 225–235, 2006, doi: 10.1016/j.neunet.2006.01.012.

[8] L. Hu, J. Ji, Y. Li, B. Liu, and Y. Zhang, "Quantile Regression Forests to Identify Determinants of Neighborhood Stroke Prevalence in 500 Cities in the USA," *J. Urban Health*, vol. 98, no. 2, pp. 259–270, 2021, doi: 10.1007/s11524-020-00478-y.

[9] E. S. Kravitz and R. J. Carroll, "Re-evaluating composite scores: Adaptive Lasso variable selection for non-linear models," *Stat*, vol. 8, no. 1, pp. 1–10, 2019, doi: 10.1002/sta4.251.

[10] Q. Chen, Z. Xiao, and Q. Yao, "Quantile control via random forest," *J. Econom.*, Feb. 2024, p. 105789, doi: 10.1016/j.jeconom.2024.105789.

[11] T. Bicalho, C. Bessou, and S. A. Pacca, "Land use change within EU sustainability criteria for biofuels: The case of oil palm expansion in the Brazilian Amazon," *Renew. Energy*, vol. 89, pp. 588–597, 2016, doi: 10.1016/j.renene.2015.12.017.

[12] M. J. Chin, P. E. Poh, B. T. Tey, E. S. Chan, and K. L. Chin, "Biogas from palm oil mill effluent (POME): Opportunities and challenges from Malaysia's perspective," *Renew. Sustain. Energy Rev.*, vol. 26, pp. 717–726, 2013, doi: 10.1016/j.rser.2013.06.008.

[13] S. Mekhilef, S. Siga, and R. Saidur, "A review on palm oil biodiesel as a source of renewable fuel," *Renew. Sustain. Energy Rev.*, vol. 15, no. 4, pp. 1937–1949, 2011, doi: 10.1016/j.rser.2010.12.012.

[14] A. R. Firdawanti, I. M. Sumertajaya, and B. Sartono, "Random Forest Lag Distributed Regression for Forecasting on Palm Oil Production," in *Proc. 1st Int. Conf. Statistics and Analytics (ICSA 2019)*, Bogor, Indonesia, Aug. 2–3, 2019, EAI, 2020, doi: 10.4108/eai.2-8-2019.2290493.

[15] L. S. Woittiez, M. T. van Wijk, M. Slingerland, M. van Noordwijk, and K. E. Giller, "Yield gaps in oil palm: A quantitative review of contributing factors," *Eur. J. Agron.*, vol. 83, pp. 57–77, 2017, doi: 10.1016/j.eja.2016.11.002.

[16] S. D. Oktarina, R. Nurkhoiry, and I. Pradiko, "The effect of climate change to palm oil price dynamics: A supply and demand model," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 782, no. 3, 2021, doi: 10.1088/1755-1315/782/3/032062.