BAREKENG: Journal of Mathematics and Its Applications
December 2023 Volume 17 Issue 4 Page 1915–1926
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss4pp1915-1926

IMPROVING ACCURACY OF PREDICTION INTERVALS OF HOUSEHOLD INCOME USING QUANTILE REGRESSION FOREST AND SELECTION OF EXPLANATORY VARIABLES

Asrirawan[1], Khairil Anwar Notodiputro[2*], Bagus Sartono[3]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, University of West Sulawesi
St. Baharuddin Lopa, Majene, 91412, Indonesia

[1,2,3]Department of Statistics and Data Science, Faculty of Mathematics and Natural Sciences, IPB University
St. Raya Dramaga, Kampus IPB Dramaga Bogor, 16680, West Java, Indonesia

*Corresponding author's e-mail: *__khairil@apps.ipb.ac.id__*

ABSTRACT

Article History:

Received: 19th May 2023
Revised: 22nd August 2023
Accepted: 19th September 2023

Keywords:

Household Income;
Prediction Interval;
Quantile Regression Forest;
Random Forest.

Quantile regression forest (QRF) is a non-parametric method for estimating the distribution function of response by using the random forest algorithm and constructing conditional quantile prediction intervals. However, if the explanatory factors (covariates) are highly correlated, the quantile regression forest's performance will decrease, resulting in low accuracy of prediction intervals for the outcome variable. The selection of explanatory variables in quantile regression forest is investigated and addressed in this paper, using several selection scenarios that consist of the full model, forward selection, LASSO, ridge regression, and random forest to improve the accuracy of household income data prediction. This data was obtained from National Labour Force Survey in 2021. The results indicate that the random forest method outperforms other methods for explanatory selection utilizing RMSE metrics. With regard to the criteria of average coverage value just above the 95% target and statistical test results, the RF-QRF and Forward-QRF methods outperform the QRF, LASSO-QRF, and Ridge-QRF methods for constructing prediction intervals.

How to cite this article:

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. Math. & App.*, vol. 17, iss. 4, pp. 1915-1926, December, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

In recent years, ensemble-based prediction algorithms have been widely developed to examine the effect of high-dimensional data. The introduction of quantile regression by Koenker [1] and Koenker and Hallock [2] enables a flexible investigation of covariate impacts on the conditional tail distribution that cannot be solved with mean regression. Furthermore, suppose the normality assumption on the outcome distribution is not met. Predicting variance and making prediction intervals for mean regression predictions becomes challenging, necessitating a more flexible quantile approach to the distribution [3], [4]. However, if the proper conditional quantile function (TCQ) is not a linear combination of the covariates or if the connection between predictor variables and the TCQ is non-linear, quantile regression predictions become less reliable [5], [6]. The Quantile Regression Forest (QRF) model evolved to address this shortcoming. Meinshausen [7] introduced the QRF model, a non-parametric ensemble approach based on decision trees. Using the framework of a random forest, this method is used to estimate the conditional quantile distribution function and provide prediction intervals for the outcome variables. The purpose of random forest algorithm is to generate a distribution function by modifying a large number of decision trees. Furthermore, because it employs decision trees as its underlying model, random forest can capture non-linear correlations between predictor variables and responses.

The estimation of quantiles depends on predictor variables in quantile regression forests. However, since predictor variables are highly correlated, it may result in uncertainty in the quantile estimate. This is due to the quantile regression forest model allowing interactions between predictor variables [7], [8]. Consequently, the prediction intervals will decrease. To address it, we can use variable selection techniques [9]–[11]. This approaches can be used to remove uninformative (irrelevant) or highly correlated predictor variables, improving model interpretation and making quantile estimation more accurate, as demonstrated by Meinshausen [7] and Youngjae and Chang [12]. Meinshausen [7] examined the goodness of prediction and prediction intervals by simulating different predictor variables. A loss function was used to assess prediction accuracy, and it was discovered that irrelevant variables tend to increase the loss function for each conditional quantile, implying that prediction accuracy diminishes. Also, Nguyen et al. [13] demonstrated that when employed with high-dimensional data, the performance of random forest may decrease due to increasingly complex interactions between predictor variables, making quantile regression forest predictions less accurate. The results may be biased due to the decision tree's random selection of variables, as uninformative variables are more likely to be preferred [14].

In this paper, we propose several techniques for variable selection (full model, LASSO [15], forward regression [16], ridge regression [17], and random forest [18], [19]) to improve the accuracy of prediction intervals for household income quantiles in Bogor. The variables chosen in the previous step will be utilized in the forest quantile regression model (QRF) models since we will have QRF, RF-QRF, Forward-QRF, LASSO-QRF, and Ridge-QRF models. The average of RMSE and coverage will deliver to compare them with different quantiles. Preliminary identification, such as normality and outlier assumptions, are displayed to guarantee that household income data is appropriate for modeling using forest quantile regression. It will also determine whether or not there is a relationship between predictor variables. To resolve these covariate models, we propose some previous research. According to Pramika [20] and Putri [21], education, age, occupation, and family size are all characteristics that influence household earnings. Participation in training or courses, working hours, job search length, internet access availability, health insurance (BPJS), and the pre work card are other predictor variables for the current research. National Labour Force Survey data was used.

2. RESEARCH METHODS

2.1 Data Description

The data on the total number of household incomes in Bogor Regency, West Java, used as the outcome variable in this study, was obtained from the National Labor Force Survey, 2021 [22]. West Java had 2,985 household heads or roughly 5.13% of total respondents. Only 1,565 households (52.4%) held jobs (employed status) for the previous week, while others were unemployed. However, just 1,498 working families were chosen for quantile regression forest modeling. The family size, age, last education level, training type (course/training), employment status, duration of job search, working hours, availability of internet services, availability of health insurance services (BPJS), and ownership status of the PRAKERJA card were the predictor variables used. Table 1 contains the specifics.

Table 1. **Description of Response and Predictor Variables**

Variables	Type	Label
Household Income (Response)	ratio	PRT
Family size	ratio	AK
Age	ratio	USIA
Highest education level: (1) Uncompleted primary school, (2) primary school, (3) Junior high School, (4) Senior High School, (5) Vocational High School, (6) Madrasah High School, (7) Diploma I-III, (8) Diploma IV, (9) Bachelor, (10) Magister, (11) Doctoral degree	ordinal	PT
Participation in training: (1) Yes, (2) No	nominal	KP
Employment Status: (1) Self-employed, (2) Assistance from Temporary Employees, (3) Assistance from Permanent Employees, (4) laborers/staff members/official working, (5) Workers in the agricultural sector, (6) Workers in the non-agricultural sector.	nominal	SP
Job search time	ratio	LMP
Working hours	ratio	JJK
Availability of Internet services: (1) Yes, (2) No	nominal	KLI
Accessibility of Health Insurance Services (BPJS): (1) Yes, (2) No	nominal	LJK
PRAKERJA card ownership status: (1) Yes, (2) No	nominal	KPK

2.2 Random Forest

Random forest generates hundreds or even thousands of decision trees that act as independent regression functions, and the ultimate output of the RF regression is the average of all decision tree outputs. RF is an expansion of Classification and Regression Trees (CART) initiated by Breiman et al [23]. Given X as an input vector with m features and $X = \{x_1, x_2, \ldots, x_m\}$, Y as a scalar output, and S_n as training data with a total of n observations, it can be represented as in Equation (1).

$$S_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}, X \in R^m, Y \in R \tag{1}$$

The RF algorithm procedure comprises separating input data at each node to improve the splitting function parameters to suit the set S_n. The decision tree must first determine the optimal separation from all variables. The splitting process starts at the root and proceeds to each node, which applies the separation function to the new input X. This technique is continued until you reach the terminal or leaf node. Typically, the tree runs out when it comes to the maximum number of levels or when a node acquires a certain amount of observations [24], [25]. Let θ be a set of built trees and a random parameter vector. We also have a weight vector $w_i(x,\theta)$, a positive constant if observation X occurs on the tree leaves $\ell(x,\theta)$, and 0 otherwise. Equation (2) may be used to calculate the weight $w_i(x,\theta)$.

$$w_i(x,\theta) = \frac{1_{\{X_i \in R_{\ell(x,\theta)}\}}}{\{j : X_i \in R_{\ell(x,\theta)}\}} \tag{2}$$

It $w_i(x)$ is the average of $w_i(\theta)$, $w_i(x)$ may be determined as shown in Equation (3).

$$w_i(x) = \frac{1}{k}\sum_{i=1}^{k} w_i(x,\theta) \tag{3}$$

The number of trees is denoted by k. Thus, the observation Y prediction could easily be written as Equation (4).

$$\hat{\mu}(x) = \sum_{i=1}^{m} w_i(x) Y_i \tag{4}$$

2.3 Quantile Regression Forest

Quantile regression forest is a random forest generalization that remains resilient, non-linear, and non-parametric in estimating conditional quantiles [3]. Consider the τ -th quantile of Y with $X = x$

(5)

designated $q_\tau(Y \mid X = x)$with $\tau \in (0,1)$. As indicated in Equation (5), the conditional distribution function for $X = x$ and $F(y \mid X = x)$ is the probability of Y existing less than or equal to $y \in R$.

$$F(y \mid X = x) = P(Y \le y \mid X = x)$$

Quantiles are constructed using this distribution function. In general, the QRF equation can be represented as in Equation (6).

$$q_\tau(Y \mid X = x) = \inf\{y : F(y \mid X = x) \ge \tau\} \qquad (6)$$

The weighted distribution of the response variables is utilized to estimate the conditional distribution function, as shown in Equation (7).

$$\hat{F}(y \mid X = x) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \le y\}} \qquad (7)$$

The weight $w_i(x)$ may be observed in Equation (3).

In addition, quantile regression forests may deliver more robust findings against outliers compared to other regression approaches [3], [26]. This is because QRF employs the median or different quantiles as the primary statistic in decision-making at each tree node, which is less impacted by extreme values. However, outliers can still affect QRF in some circumstances. Outliers that are too far away from the majority of data points might interfere with the construction of tree nodes in QRF, resulting in erroneous or inaccurate predictions. In general, the QRF Algorithm proceeds as follows:

1. As in random forests, put k trees $T(\theta_t), t = 1, \ldots, k$. Make a note of all observations in this leaf, not just the average, for each leaf of each tree.
2. Drop x down all trees for a specified $X = x$.
3. For each tree, compute the weight of observation $w_i(x, \theta_t)$, $i \in \{1, \ldots, n\}$ as seen in Equation (2).
4. Calculate the weight $w_i(x)$ for each observation $i \in \{1, \ldots, n\}$ as an average of overusing (3).
5. By applying the weights from Step 2, compute the distribution function estimate as in (7) for every. $y \in R$
6. Estimates of the conditional quantiles are obtained by plugging instead of into (1).
7. Estimates $\hat{q}_\alpha(x)$ of the conditional quantiles $q_\alpha(x)$ are produced by substituting $\hat{F}(y \mid X = x)$ for $F(y \mid X = x)$ in (6).

2.4 Prediction Intervals

Prediction intervals are constructed utilizing conditional quantiles of QRF-predicted household revenue responses. Prediction intervals give a range of values for actual data at an acceptable level of confidence. In particular, Equation (8) constructs the prediction interval $(1-\alpha)\times 100\%$ for a given covariate (p-dimensional vector) response.

$$I(x) = [q_{\alpha/2}(Y \mid X = x), q_{1-\alpha/2}(Y \mid X = x)] \qquad (8)$$

For example, the 95% prediction range for the response Y is calculated by Equation (9).

$$I(x) = [q_{0,025}(Y \mid X = x), q_{0,975}(Y \mid X = x)] \qquad (9)$$

This suggests that for a given value x, the household income is likely to fall inside the interval. The length of the predicted interval varies X. The coverage value is used to compare the reliability of the prediction interval for family income response. The coverage value is the percentage of sample points that fall inside the prediction interval.

2.5 Evaluation Metrics

The root means square error (RMSE) measure is used to evaluate the accuracy of the QRF algorithm's forecast values to actual values acquired from trials. RMSE is comparable to mean absolute error (MAE), except it gives more weight to bigger fundamental values than MAE [27], [28]. A significant discrepancy between MAE and RMSE suggests the presence of variance in individual mistakes. RMSE can be defined as follows Equation (10).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} y_i - \hat{y}_i}{m}} \tag{10}$$

We also include a coverage level to measure the accuracy of prediction intervals. The coverage probability for such intervals is commonly chosen by convention or brilliant judgment. The wider the prediction interval, the greater the coverage probability, and vice versa [29].

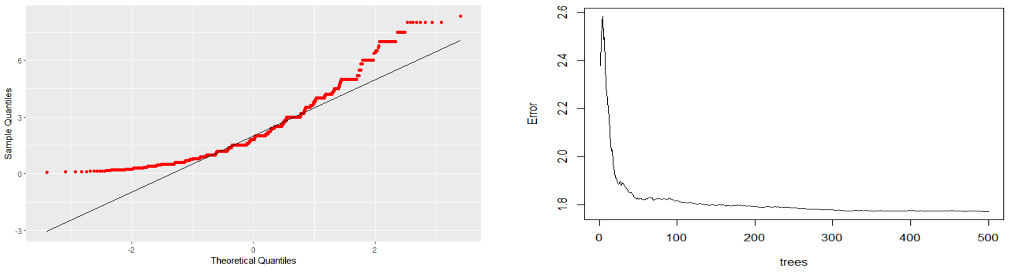
2.6 Step Analysis

This study's analytical steps were as follows:

1. Explore variable response data and predictor variables as follows:
   a. calculating correlations between mixed-scale variables based [30]
   b. plotting between TCQ and predictor variables
   c. create a boxplot to detect outliers
   d. Create a Q-Q plot to see the distribution of the response variables
2. Simulate variable selection using the full method, forward, LASSO, ridge, and random forest for the number of variables p = 10, 8, 5, and 2 with the following:
   a. Dividing the training data and test data by a ratio of 80:20
   b. Selecting the best variables for p = 10,8,5 and 2 for each method by looking at RMSE (full, forward, LASSO, ridge) and variable importance measures (random forest).
   c. Calculating the RMSE value of each variable combination
   d. Step b is repeated ten times
   e. Calculating the average of RMSE
   f. Comparing the average RMSE values through the plots
3. From step 2, determine the best combination of variables to be used in the forest quantile regression method for each forward, LASSO, ridge, and random forest method.
4. Predict and construct forest quantile regression prediction intervals with the following steps:
   a. Split the training data and test data by comparison 80:20
   b. Estimated conditional quantile predictive value ($\tau = 0{,}005; 0{,}025; 0{,}05; 0{,}5; 0{,}975; 0{,}995$) for QRF, RF-QRF, Forward-QRF, LASSO-QRF, and Ridge-QRF methods, respectively.
   c. Step b is repeated ten times
   d. Calculating the average of RMSE
   e. Plot the mean RMSE values for all quantiles
   f. Create an RMSE boxplot for the median quantiles
5. Make prediction intervals
6. Calculate coverage values
7. Create a boxplot for the average coverage value
8. Perform statistical tests on RMSE mean and coverage based on steps (4) and (5) using paired t-test.

3. RESULTS AND DISCUSSION

Data exploration investigation and the analysis of assumptions, including TCQ and outlier detection about the distribution of response data, will be provided in early research to identify whether the quantile regression forest method is able to be used to predict and improve the accuracy of prediction intervals for household income information. According to the Q-Q normal plot, the normal distribution assumption for household income data is not met. Figure 1a shows this. The red dots on the plot do not follow the diagonal line but instead create a distinct pattern, suggesting that the household income response data does not follow a normal distribution. Meanwhile, the actual conditional quantile relationship (TCQ) and predictor variables are non-linear, as illustrated in **Figure 1** (b), which is a plot of the number of trees against the error of quantile $\tau = 0{,}5$. The link between the number of trees and the error is a way to determine whether or not the relationship between the TCQ and the predictor variables is linear. **Figure 1** (b) depicts the association pattern between the predictor variable using by KP variable at the median quantile. This variable was selected because it has the highest correlation value as compared to the others. However, the plot for countless additional variables shows a similar pattern, despite the fact that the correlation value is low.



(a) (b)

Figure 1. (a) Q-Q normal of household income (PRT), (b) Median true conditional quantile (TCQ)

According to **Figure 1** (b), the relation between TCQ and the predictor variable KP is non-linear since it does not form a straight-line pattern and has a trend in error values. Furthermore, the enormous number of trees reflects the complexity of the non-linear connection in TCQ. Thus, boxplots, as illustrated in **Figure 2**, are used to ensure that there are no outliers that are too far out. As previously stated, the quantile regression forest approach is resistant to outliers but will impair prediction accuracy if outliers are too widely apart. The graphic shows that the few outlier points (in red) are still within a respectable range and will not have a substantial impact on the quantile regression forest's performance.

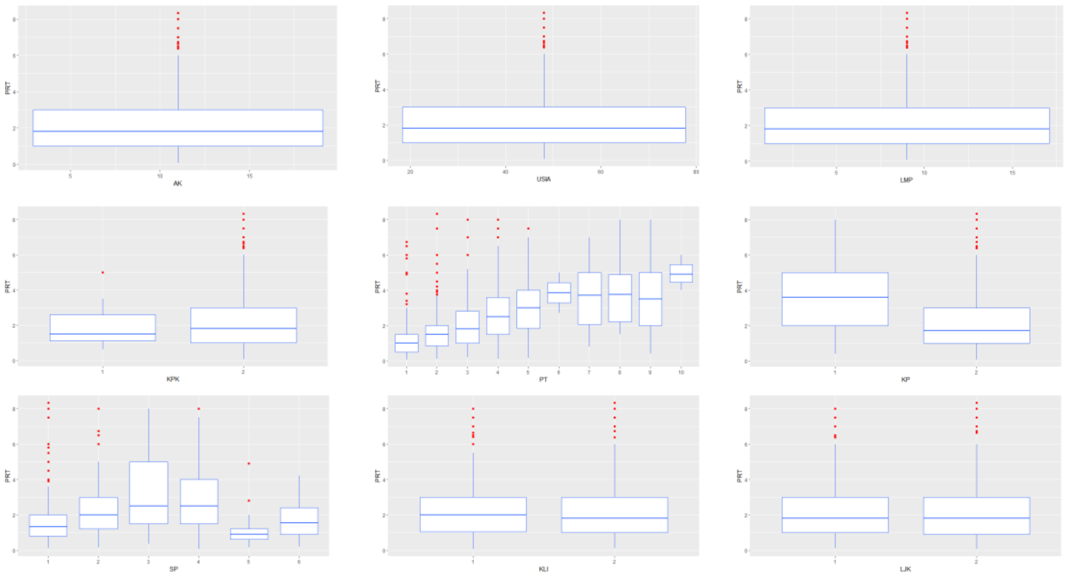


Figure 2. Categorical and continuous explanatory variables

The next stage is to investigate the correlation between predictor variables after checking the distribution analysis, TCQ, and outliers of household income data. It has the potential to improve the quantile regression forest approach's prediction effectiveness. **Figure 3** depicts the study's level of association among the predictor factors (mixed-scaled type) and also between these variables and household income. To allow the calculation of correlations between any form of mixed variable, we use the idea of semi-parametric latent Gaussian copula approaches proposed by [30]. The higher the negative correlation between variables, the more blue-black the hue, and the stronger the positive correlation between variables, the more yellowish-green the color. The findings imply that the correlation between variables is more diminutive than r=|0.6|. With r=-0.508, the variable of training involvement shows a high negative connection with household earnings. Some factors, for example, age and highest education level (r=-0.492), household income and highest education level (r=0.427), and household income and employment status (r=0.407), have correlation values greater than r=|0.4|. It also indicates that among the predictor variables a fairly significant correlation. This also holds for the relationship between predictors and responses.

As a result, variable selection must be made, starting with a simulated study of the number of variables using various selection methods such as forward selection, full model, LASSO, ridge, and random forest. **Figure 4** displays the simulation results for household income data using these selection approaches with the number of predictor variable combinations of p = (10, 8, 5, 2) and ten repetitions. The RMSE value is used to evaluate the approaches. In general, variable selection results show that the number and mix of variables utilized affect the decrease or rise in the RMSE value. **Figure 4** further indicates that the random forest approach has a lower average RMSE value than other methods for each value of p and repetition available.

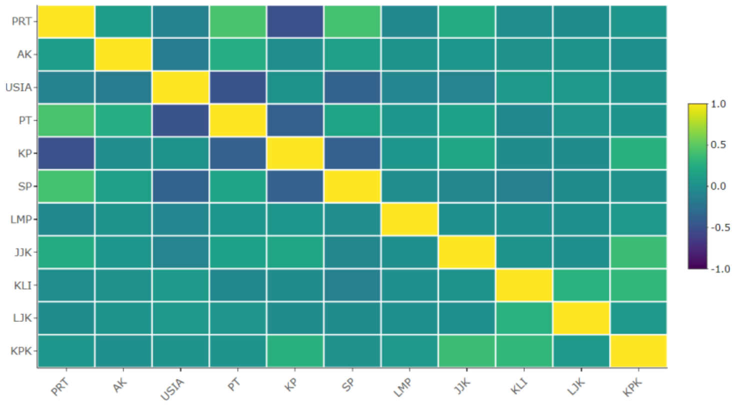
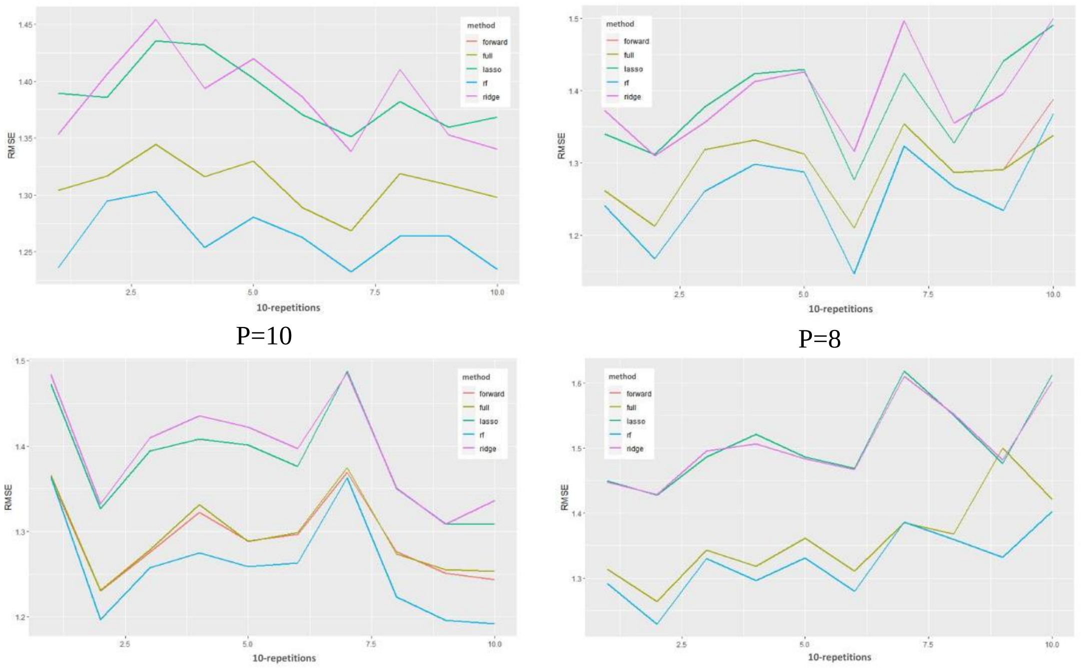


Figure 3. Correlation Of Mixed-Scaled Types Of Explanatory And Outcome Variables

Meanwhile, the average RMSE of the LASSO and ridge regression models is nearly the same, but it is still extremely high when compared to the average RMSE of the full model and forward model.



P=10

P=8

P=5 P=2

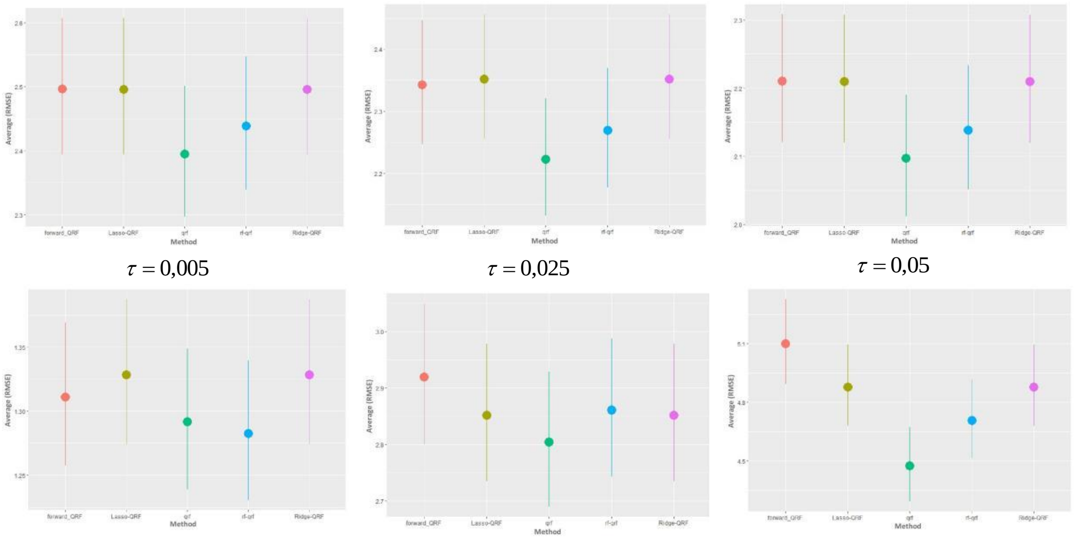
Figure 4. RMSE values from 10 replications

The variable selection results and the variables utilized for quantile estimation in the quantile regression forest approach can be shown in Table 2. The variable combinations for each method are the best combinations based on AIC (full and forward), RMSE (LASSO and ridge), and importance variables (random forest).

Table 2. Explanatory Selection

Method	Variables	Total
Forward	PT, KP, SP, JJK	4
Full	AK, USIA, PT, KP, SP, LMP, JJK, KLI, LJK, KPK	10
LASSO	PT, KP, SP, LMP, KLI, KPK	6
RF	AK, USIA, KP, SP, LMP, JJK, KLI, LJK	8
Ridge	PT, KP, SP, LMP, KLI, KPK	6

Following selecting the predictor variables, the prediction values and RMSE values of the household income response prediction based on the quantiles $\tau = 0,005; 0,025; 0,05; 0,5; 0,95; 0,995$ displayed in Figure 5 for all approaches are computed. In addition, the explanatory variables utilized in the models (QRF, RF-QRF, Forward-QRF, LASSO-QRF, and Ridge-QRF) are compared in Figure 5, utilizing each variable based on Table 2. Furthermore, the QRF model employs overall explanatory factors per the full model's suggestion.



$\tau = 0,005$ $\tau = 0,025$ $\tau = 0,05$

$\tau = 0,5$ $\tau = 0,95$ $\tau = 0,995$

Figure 5. The average of RMSE for quantiles $\tau = 0,005; 0,025; 0,05; 0,5; 0,95; 0,995$

Figure 5 clarifies that the QRF method has lower RMSE values than others for various quantile points proposed, followed by the RF-QRF method. However, the RF-QRF method has the lowest RMSE value at the median quantile point $\tau = 0,5$. The performance of RF-QRF will improve as the quantile approaches the center. Figure 6 demonstrates the comparison of RMSE values between models at the median quantile.

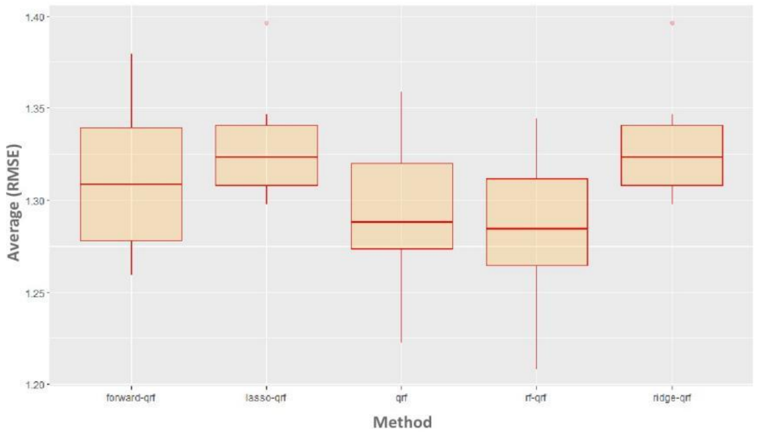


Figure 6. RMSE for median quantile regression forest

The median quantile is frequently employed in quantile regression because it has the most weight in generating the projected value and generates a more comprehensive model. Therefore, it is believed to give more valuable information for household income. The following step is to create prediction intervals and assess the method's performance using the average coverage value from 10 repetitions with a target coverage value of 95%.

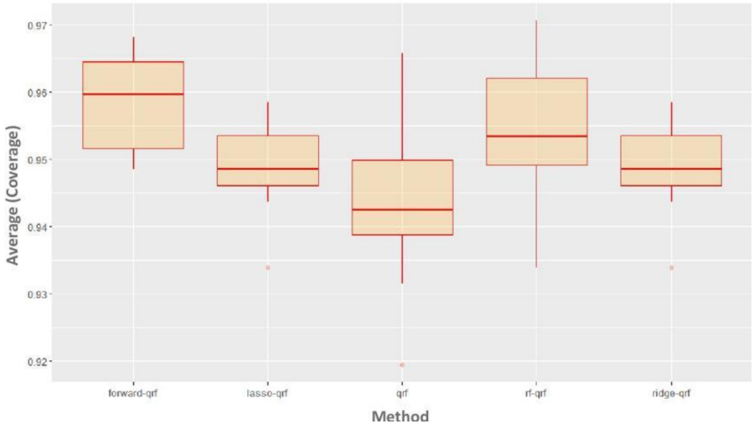
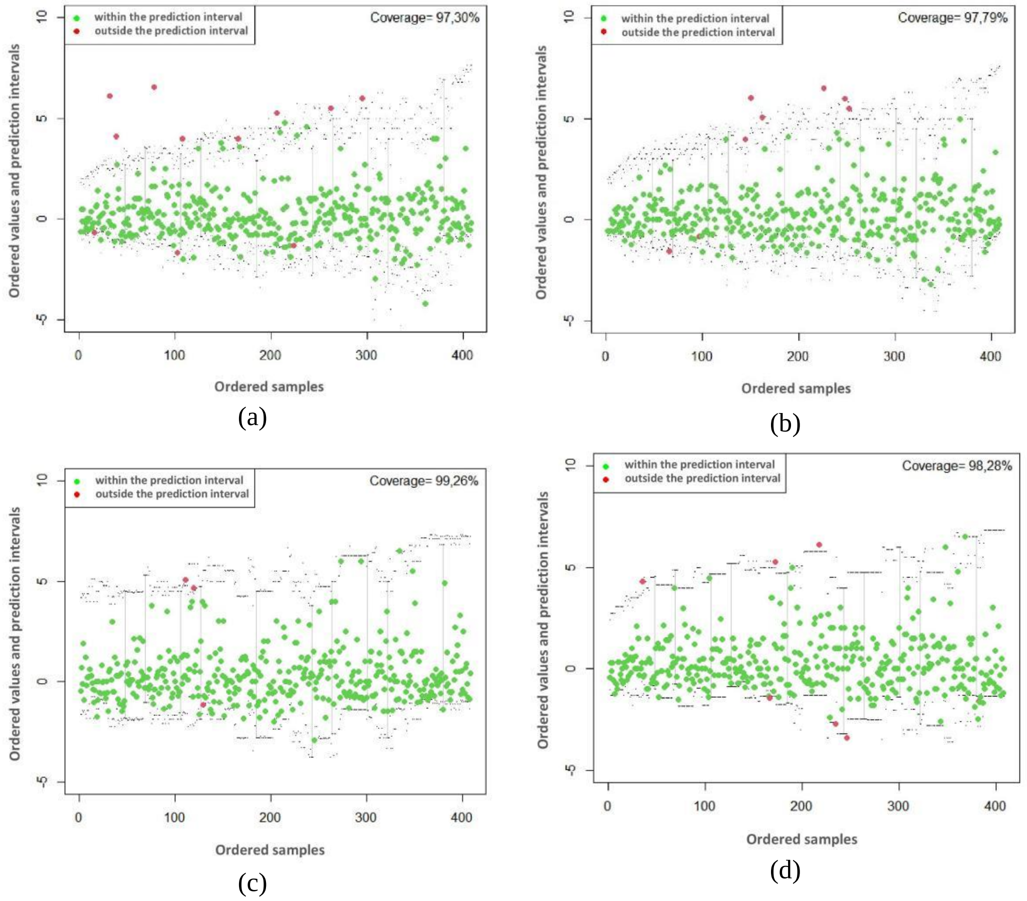


Figure 7. Boxplot Of The Average Of Coverage Quantile Regression Forest

The average coverage values of the quantile regression forest with a target coverage value of 95% and ten replications at the median quantile are presented in **Figure 7**. All approaches' average coverage levels are pretty close to the desired coverage value of 95%. However, QRF, LASSO-QRF, and Ridge-QRF have average coverage values lower than the reference target of 95%, whilst the others (Forward-QRF and RF-QRF) have values higher than 95%. Furthermore, QRF has the lowest average coverage value. This implies that the predictability of prediction intervals for family income responses, including all predictor factors, is lower than the predictability of variables generated from selection results. As a result, RF-QRF and Forward-QRF, which have average coverage values greater than 95%, may be utilized to create prediction ranges for household income. **Figure 8** and **Figure 9** provide a plot of prediction intervals $\tau = 0{,}005$ and the median quantile $\tau = 0{,}5$ for ten replications.



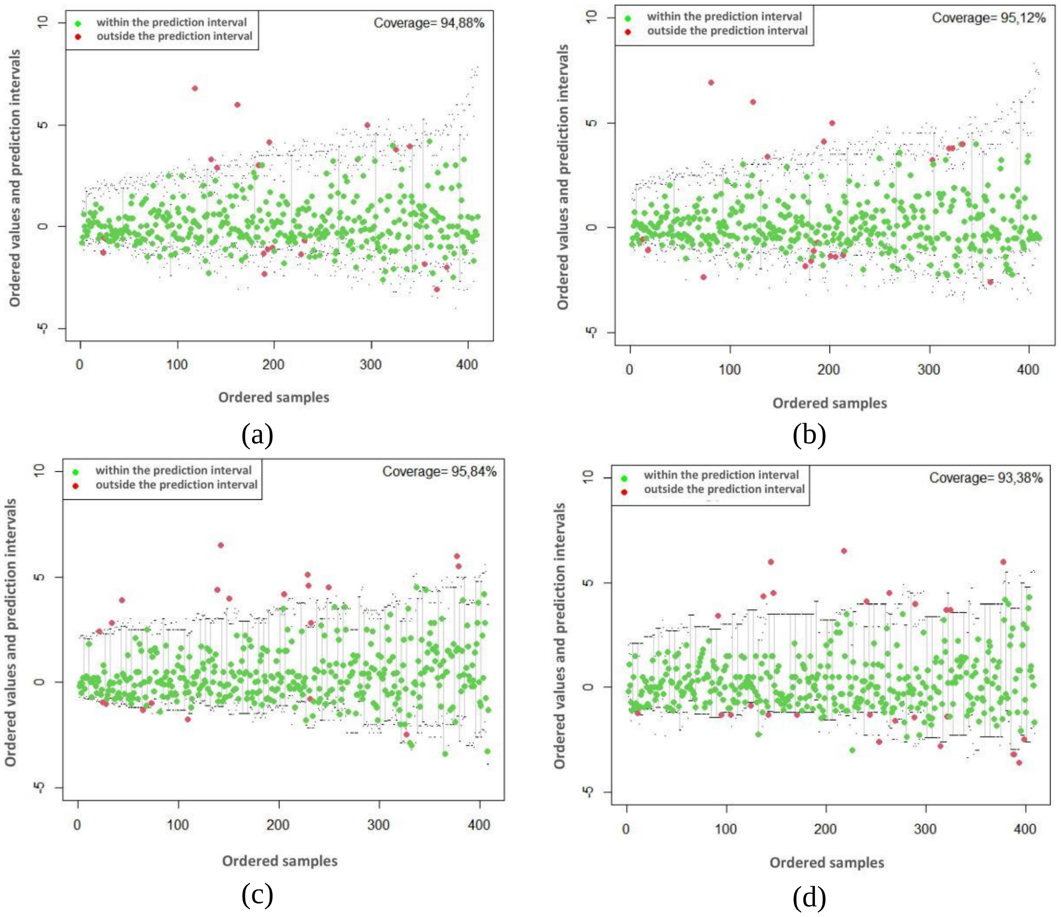
(a)

(b)

(c)

(d)

Figure 8. Prediction Interval 95% For Quantile 0,005 (A) Qrf (B) Forward-QRF (c) RF-QRF (d) LASSO-QRF



(a)

(b)

(c)

(d)

Figure 9. Prediction interval 95% for median Quantile (A) Qrf (B) Forward-QRF (c) RF-QRF (d) LASSO-QRF

The green dots represent observed sites that are within the 95% prediction interval, while the red dots represent places that are outside the interval. The coverage numbers in the quantile $\tau = 0{,}005$ are about 97%-99% (above the aim of 95%), indicating that the prediction ranges are more cautious. This suggests that the probability of the predicted value of household income fitting actual observation is pretty high. This is also evidenced by the comparatively long prediction interval, which encompasses more actual values. However, the performance of the prediction intervals diminishes from quantile $\tau = 0{,}005$ to quantile $\tau = 0{,}5$. Only RF-QRF and Forward-QRF are above the 95% interval objective, while the rest are below it. This can also be observed in the considerably lower prediction interval for the median quantile compared to the prior quantile. However, when employing the median quantile, RF-QRF is suggested since it offers accurate findings.

Table 3. Statistical Test of RMSE and Coverage

Criteria	Method	Mean difference	Lower	upper	t-value	P-value
RMSE	QRF and RF-QRF	0.0089	0.0018	0.0161	2.8368	0.0195
	QRF and Forward-QRF	0.0193	0.0078	0.03073	3.7947	0.0042
	QRF and LASSO-QRF	0.0367	0.0187	0.0547	4.6113	0.0013
	RF-QRF and Forward-QRF	0.0282	0.0161	0.0403	5.2581	0.0005
	RF-QRF and LASSO-QRF	0.0665	0.0248	0.0665	4.9565	0.0008
	Forward-QRF and LASSO-QRF	-0.0457	-0.0665	-0.0248	-4.9565	0.0008
Coverage	QRF and RF-QRF	-0.0105	-0.0154	-0.0056	-4.8138	0.0009
	QRF and Forward-QRF	0.0149	0.0038	0.0260	3.0331	0.0142
	QRF and LASSO-QRF	0.0056	-0.0046	0.0158	1.2429	0.2453
	RF-QRF and Forward-QRF	0.0044	-0.0036	0.0124	1.2431	0.2452
	RF-QRF and LASSO-QRF	-0.0049	-0.0127	0.0029	-1.4266	0.1874
	Forward-QRF and LASSO-QRF	0.0049	-0.0029	0.0127	1.4266	0.1874

A statistical analysis of the mean RMSE and coverage values was undertaken to compare the techniques, as shown in **Table 3**. The t-test findings for mean RMSE values indicate that there is a significant difference in mean RMSE values between techniques at the 95% confidence level. The RF-QRF method is superior to the others. Meanwhile, the t-test for mean coverage values reveals that significant variations in mean coverage values exist only between QRF and RF-QRF, as well as QRF and Forward-QRF. This suggests that the RF-QRF and Forward-QRF approaches outperform QRF when it comes to creating prediction intervals for household income responses.

4. CONCLUSIONS

The findings of explanatory variable selection affect predicted values and prediction intervals for the quantile regression of household income response. The random forest technique has the lowest RMSE value based on the simulated predictor variable selection procedure. The RF-QRF and Forward-QRF algorithms display an average coverage value above the given target when creating prediction intervals with a target coverage of 95%. This indicates that, when compared to other approaches, these methodologies produce more trustworthy projections of household income.

ACKNOWLEDGMENT

Special thanks to the Advanced Data Analysis (ADL) Lecturers in the Doctoral Program in the Statistics and Data Science Study Program at FMIPA IPB University for providing considerable encouragement and direction in developing this research. Thank you also to the Ministry of Education, Culture, Research, and Technology, Indonesia, for supporting my studies through the BPI scholarship.

REFERENCES

[1] R. Koenker, "Quantile Regression - book extract," *Cambridge Univ. Press*, no. February 1997, p. 198, 2005.

[2] 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.

[3] Q. Huang, H. Zhang, J. Chen, and M. He, "Quantile Regression Models and Their Applications: A Review," *J. Biom. Biostat.*, vol. 08, no. 03, 2017, doi: 10.4172/2155-6180.1000354.

[4] E. Waldmann, "Quantile regression: A short story on how and why," *Stat. Modelling*, vol. 18, no. 3–4, pp. 203–218, 2018, doi: 10.1177/1471082X18759142.

[5] H. Cardot and P. Sarda, "Estimation in generalized linear models for functional data via penalized likelihood," *J. Multivar. Anal.*, vol. 92, no. 1, pp. 24–41, 2005, doi: 10.1016/j.jmva.2003.08.008.

[6] K. Chen and H. G. Müller, "Conditional quantile analysis when covariates are functions, with application to growth data," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 74, no. 1, pp. 67–89, 2012, doi: 10.1111/j.1467-9868.2011.01008.x.

[7] N. Meinshausen, "Quantile Regression Forest," *J. Mach. Learn. Res.*, vol. 7, no. 2006, pp. 983–999, 2006, doi: 10.1111/j.1541-0420.2010.01521.x.

[8] M. Ardiansyah, K. A. Notodiputro, and B. Sartono, "Peningkatan Presisi Dugaan Berat Gabah Melalui Proses Seleksi Peubah Dalam Pembelajaran Mesin Statistika," in *Prosiding Seminar Nasional VARIANSI*, 2020, pp. 171–183.

[9] Y. Zhang, Q. Wang, and M. Tian, "Smoothed Quantile Regression with Factor-Augmented Regularized Variable Selection for High Correlated Data," *Mathematics*, vol. 10, no. 16, pp. 1–30, 2022, doi: 10.3390/math10162935.

[10] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Stat.*, vol. 34, no. 3, pp. 1436–1462, 2006, doi: 10.1214/009053606000000281.

[11] M. Rashighi and J. E. Harris, "Regularized Quantile Regression and Robust Feature Screening for Single Index Models," *Physiol. Behav.*, vol. 176, no. 3, pp. 139–148, 2017, doi: 10.1053/j.gastro.2016.08.014.CagY.

[12] Y. Chang, "Multi-step quantile regression tree," *J. Stat. Comput. Simul.*, vol. 84, no. 3, pp. 663–682, 2014, doi: 10.1080/00949655.2012.721886.

[13] T. T. Nguyen, H. Zhao, J. Z. Huang, T. T. Nguyen, and M. J. Li, "A new feature sampling method in random forests for predicting high-dimensional data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9078, no. May, pp. 459–470. doi: 10.1007/978-3-319-18032-8_36.

[14] T. Nguyen, J. Z. Huang, T. T. Nguyen, and I. Khan, "Bias-Corrected Quantile Regression Forests for High-Dimensional Data," in *Proceeding of the 2014 International Conference on Machine Learning and Cybernetics*, 2014, no. July, pp. 1–6. doi: 10.13140/2.1.2500.8002.

[15] R. Shrinkage, "Regression Shrinkage and Selection via the Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996, [Online]. Available: jstor.org/stable/2346178

[16] J. Wieczorek and J. Lei, "Model selection properties of forward selection and sequential cross-validation for high-dimensional regression," *Can. J. Stat.*, vol. 50, no. 2, pp. 454–470, 2022, doi: 10.1002/cjs.11635.

[17] Y. Wu, "Can't Ridge Regression Perform Variable Selection?," *Technometrics*, vol. 63, no. 2, pp. 263–271, 2021, doi: 10.1080/00401706.2020.1791254.

[18] J. M. Klusowski, "Complete Analysis of a Random Forest Model," *arXiv*, vol. 13, pp. 1063–1095, 2018.

[19] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[20] D. Pramika, "Faktor-Faktor Yang Mempengaruhi Pendapatan Rumah Tanggadi Kabupaten Empat Lawang Provinsi Sumatera Selatan," *J. Ekon. Manajemen, Bisnis, Audit. dan Akunt.*, vol. 2, no. 1, pp. 33–49, 2017.

[21] D. A. Putri and N. D. Setiawina, "Pengaruh Umur, Pendidikan, Pekerjaan Terhadap Pendapatan Rumah Tangga," *EP-Unud*, vol. 2, no. 4, pp. 173–180, 2017.

[22] BPS, "Survey Angkatan Kerja Nasional," 2021.

[23] L. Breiman, *Classification and Regression Trees*, 1st ed., no. January. New York: Taylor and Francis Group, 1984. doi: 10.1201/9781315139470.

[24] L. Breiman, "Random Forests," *Mach. Learn.*, no. 45, pp. 5–32, 2021, doi: 10.1109/ICCECE51280.2021.9342376.

[25] Y. Li *et al.*, "Random forest regression for online capacity estimation of lithium-ion batteries," *Appl. Energy*, vol. 232, no. February, pp. 197–210, 2018, doi: 10.1016/j.apenergy.2018.09.182.

[26] L. Schiesser, "Quantile Regression Forests - An R-Vignette," pp. 1–10, 2014.

[27] H. Pham, "A new criterion for model selection," *Mathematics*, vol. 7, no. 12, pp. 1–12, 2019, doi: 10.3390/MATH7121215.

[28] A. Asrirawan, S. U. Permata, and M. I. Fauzan, "Pendekatan Univariate Time Series Modelling untuk Prediksi Kuartalan Pertumbuhan Ekonomi Indonesia Pasca Vaksinasi COVID-19," *Jambura J. Math.*, vol. 4, no. 1, pp. 86–103, 2022, doi: 10.34312/jjom.v4i1.11717.

[29] J. Landon and N. D. Singpurwalla, "Choosing a coverage probability for prediction intervals," *Am. Stat.*, vol. 62, no. 2, pp. 120–124, 2008, doi: 10.1198/000313008X304062.

[30] M. Huang, C. Müller, and I. Gaynanova, "latentcor: An R Package for estimating latent correlations from mixed data types," *J. Open Source Softw.*, vol. 6, no. 65, p. 3634, 2021, doi: 10.21105/joss.03634.