Random Forest Lag Distributed Regression for Forecasting on Palm Oil Production

Aulia Rizki Firdawanti[1], I Made Sumertajaya[2], Bagus Sartono[3]
{arfirdawanti@gmail.com[1], imsjaya.stk@gmail.com[2], bagusco@gmail.com}

Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia[1],
Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia[2],
Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia[3]

Abstract. Palm oil is one of the most cultivated potential commodities so it is necessary to do research to determine the determinants of production and forecasting on palm oil production. The objectives are perform data modeling dan forecasting using random forest lag distributed regression on palm oil production. This analysis combines the lag distributed regression and random forest methods. The results showed that the performances for this model are the correlation value is 0.9302, RMSE is 20.379, MAE is 14.143, and R-Square is 0.829. The 5 most important variables were quantity of palm oil, land area, palm oil age, 8th lag of wind velocity, and 1st lag of temperature. The distribution of data forecasting results are not much different from the distribution of testing data and original data.

Keywords: lag distributed regression, palm oil production, random forest regression

1 Introduction

Palm oil is one of the most cultivated potential commodities. Palm oil is an industrial plant for cooking oil, industrial oil, and fuel (biodiesel). Indonesia is the largest producer of palm oil in the world, based on data processed by the Indonesian Palm Oil Association (GAPKI IPOA) [3], in 2016, Indonesia produced 36 million tons of palm oil where 25.11 million tons were exported. The export of Indonesian palm oil increased significantly by 23% or from 25.11 million tons in 2016 to 31.05 million tons in 2017. It is estimated that this number will continue to increase along with the increasing number of palm oil plantations in Indonesia.

Palm oil production is an important part of the Indonesian economy because it contributes high foreign exchange to the country. Palm oil has an important role so it is necessary to do research to determine the determinants of production and forecasting the size of palm oil production for the next period. Analysis cannot be done easily because the factors that affect palm oil production are very large. To increase the volume of palm oil production, optimal factors or variables are needed.

This research will be focused on determining the variables that affect the increase of palm oil production then do a forecast on palm oil production for the next periods. There are several problems that make modeling and forecasting difficult. The problems are palm oil production data is a nonlinear variable, the variables used consist of numerical and categorical variables, it is assumed that there are lag effects from the independent variables used, so the method used

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290493

in this study is the random forest lag distributed regression. This analysis combines the lag distributed regression and random forest methods. This analysis is able to overcome many variables, can accommodate lag effects, see relationships that are not linear, see interactions between variables, can be used for numerical or categorical variables, and be fast in computing, but this analysis is difficult in tuning hyperparameters.

Based on the background described, this research objectives are:

1. Perform data modeling using Random Forest analysis Lag Distributed Regression.
2. Forecasting data on palm oil production

2 Materials

This research uses secondary data from 4 companies engaged in palm oil production in Borneo. The company has a total of 14 estates, 63 afdellings and 1312 blocks. Block is the smallest unit in this research, in one afdelling there are several blocks, and in one company there are several afdelling.

Table 1. Variables Used

Variables	Variables Description	Periods
Y	Production of palm oil (tons)	2016-2018
X_1	Number of rainy days	2016-2018
X_2	Rainfall	2016-2018
X_3	Temperature	2016-2018
X_4	Air Humidity	2016-2018
X_5	Wind velocity	2016-2018
X_6	Lighting intensity	2016-2018
X_7	Long of lighting	2016-2018
X_8	Treatment and maintenance for blocks	2016-2018
X_9	Treatment and maintenance when harvesting	2016-2018
X_{10}	Treatment and maintenance when fertilizing	2016-2018
X_{11}	Attack of rat pest(*Rattus tiomanicus*)	2016-2018
X_{12}	Attack of tirathaba (*Tirathaba* sp.)	2016-2018
X_{13}	Attack of caterpillar (*Setothosea asigna*)	2016-2018
X_{14}	Attack of bagworms (*Mahasena corbetti*)	2016-2018
X_{15}	Dosage of *borate*	2016-2018
X_{16}	Dosage of CPD-HIK	2016-2018
X_{17}	Dosage of CPD-25	2016-2018
X_{18}	Dosage of CPD-45	2016-2018
X_{19}	Dosage of $CuSO_4$	2016-2018
X_{20}	Dosage of MOP	2016-2018
X_{21}	Dosage of *rockphosphate*	2016-2018
X_{22}	Dosage of *super dolomite*	2016-2018
X_{23}	Dosage of urea	2016-2018
X_{24}	Topography or slope	-
X_{25}	Varieties of oil palm seedlings	-
X_{26}	Block hectares	-
X_{27}	Quantity of palm oil plant in a block	-
X_{28}	Class	-
X_{29}	Ageof the palm oil plant	-

The response variable in this research is the production of palm oil plants. The data was from January 2016 until December 2018 for each month. There are 29 independent variables with 12 lagged time of each variables that has a time periods. A list of variables along with information about the periods are shown in **Table 1**.

Topographic variables or slopes consist of wavy and flat slopes. Varieties of oil palm seedlings consist of Lonsum, Marihat, Socfin, Topaz, BLRS, and Dummy Mass. Variable of pest attack on plants, they are rat (*Rattus tiomanicus*), tirathaba (*Tirathaba* sp.),caterpillar (*Setothosea asigna*), and bagworms (*Mahasena corbetti*). Treatmentand maintenance variables include evaluation of block maintenance, evaluation when harvesting, and evaluation when fertilizing. SPH or standard per hectare is the number of principal oil palm planted in hectares. Palm oil has a planting year between 1996 and 2015, so that in 2019 the age of the plant will be 4 to 23 years. The types of fertilizers used are borate, CPD-HIK, CPD-25, CPD-45, CuSO4, MOP, rock phosphate, super dolomite, and urea.

3 Methods

Time series analysis is an analysis of a set of data in a past period of time that is useful for knowing or predicting future conditions. This is based on the fact that human behavior is much influenced by conditions or previous times so that in this case the time factor is very important role[4]. Regression models using time series data not only use the effect of changing the independent variables on non-independent variables in the same time period and during the same observation period, but also using the previous time period. The time needed for X variables in influencing non-free variables X is called time difference or lag [7].

In regression analysis involving time series data, if the regression model includes not only the current but also the lagged (past) values of the explanatory variables (the X's), it is called a distributed-lag model. The lag distribution model is a regression model that involves present and lagged data from explanatory variables [4].

[1] CART (Classification and Regression Tree) is a nonparametric statistical methodology developed for the topics of classification and regression analysis, both for categorical and continuous variable responses. This method is developed based on decision tree rules. Decision tree is formed using binary recursive partitioning algorithm. CART classifies response variables based on their explanatory variables into two blocks called child node. CART's goal is to look at the relationship between the response variables and explanatory variables through grouping based on the explanatory variables used so that an accurate group of data is obtained to characterize a classification. CART produces a classification tree if the response variable is categorical, and produces a regression tree if the response variable is continuous.

The random forest method is a development of the CART method, namely by applying the bootsrap aggregating method (bagging and random subsetting). The random forest produces many trees so that they are formed to resemble a forest, then an analysis is performed on the tree. The random forest algorithm in the data consisting of n observations and p independent variables is as follows [2]:

1. The bootstrap stage, takes a random sample of size n with recovery in the data cluster.
2. Random subsetting stages are carried out by building trees without pruning until they reach the maximum size.

3. Repeat steps 1 and 2 for k times, so that a forest consists of k trees.
4. Conducting joint estimates based on the *k* tree. The response of an observation is predicted by combining the results of the prediction of the tree. The random forest method in the estimation regression tree is based on the average.

Select a stable tree based on the MSE value generated.The selection of a stable tree based on random forest in the regression tree can be determined based on the MSE value where$\text{MSE}= \sum_{t=1}^{n}\left(y_i-\hat{y}_i\right)^2$ with $\hat{y}_i$ is the average observation prediction i and n is the number of observation.

Random forest lag distributed regression is an analysis that combines the lag distributed regression and random forest methods. This analysis is able to overcome many variables, can accommodate lag effects, see relationships that are not linear, see interactions between variables, can be used for numerical or categorical variables, and be fast in computing, but this analysis are difficult in tuning hyperparameters.

There are several steps taken in analyzing data in this research. The steps are as follows:

1. Data Exploration

   Data exploration inthis research to see the whole data, such as data distribution, missing value, whether the data is missing, there is no data or missing data is a value of 0, detection of outliers which if there are outliers, will be deleted on the data, and correlation between variables.
2. Modeling using random forest lag distributed regression with the following steps:
   a. Make 12 lags for data that have a period of time.
   b. Divide data into training data and testing data (validation) into 50% by blocks.
   c. Modeling data using random forest.
   d. Evaluate the model by looking at the performance of the model.
   e. Select the important variables for each variable.This important predictor is selected based on the most important lag of each predictor for lag formed variables and variables that are not formed lag will be re-selected.
   f. Remodeling random forest models using selected important variables.
   g. Evaluate the second model by looking at the performance of the model.
3. Forecasting oil palm production data using the model obtained at the last stage of modeling.

4Results and Discussion

The initial stages of data analysis are exploration of the data. Exploration is carried out on dependent variable (production of palm oil). Based on **Figure 1**, it can be seen that production of palm oil has a non-linear distribution.

Fig. 1. Line Plot of Palm Oil Production.

Varieties of oil palm seedlings consist of Lonsum, Marihat, Socfin, Topaz, BLRS, and Dummy Mass. **Table 2** shows the varieties of oil palm seedlings with planting year. Based in **Table 2**, it can be seen that the youngest planting year of oil palm is in 2015 with the BLRS type of seeds and the oldest planting year of oil palm is in 1996 with the type of Marihat seedlings.

Table 2. Varieties of oil palm seedlings with planting year.

BLRS	Dummy Mass	Lonsum	Marihat	Socfin	Topaz
2007	2003	2003	1996	1999	2013
2008	2009	2007	1997	2000	2014
2009	2010	2008	1998	2001	
2010	2011	2011	1999	2003	
2011		2013	2000	2004	
2013		2014	2001	2005	
2014			2003	2006	
2015			2004	2007	
			2007	2013	
			2008	2014	
			2010		
			2012		
			2013		
			2014		

Topographic or slopes consist of wavy and flat slopes.**Figure 2** shows the percentage of slope types owned by the four companies. Based on **Figure 2**, it can be seen that 75% consists of flat slopes and another 25% is wavy slopes.

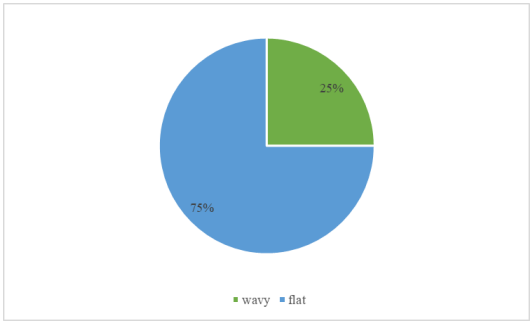

Fig. 2. Topographic.

The analysis will continue by looking at the missing value. **Table 3** shows the percentage of missing valuefor each variable.Based on **Table 3**, it shows that there are 10 variables that have a lot of missing value. This missing valuecan be overcome by using imputation techniques. The imputation technique is performed on the variables of wind velocity, air humidity, temperature, lighting intensity, long of lighting, number of rainy day, rainfall, treatment and maintenance for blocks, when harvesting, and when fertilizing. This imputation

technique aims to replace lost data with predicted data using tree regression. The stages begin by dividing the data into missing and non-existent data, making predictions about the missing data and replacing the missing data by using predictive data that has been obtained.

Table 3. Percentage of Missing value for each variables

Variables	Variables Description	Percentage
Y	Production of palm oil (tons)	0.00
X_1	Number of rainy days	8.33
X_2	Rainfall	8.33
X_3	Temperature	16.67
X_4	Air Humidity	16.67
X_5	Wind velocity	16.67
X_6	Lighting intensity	16.67
X_7	Long of lighting	16.67
X_8	Treatment and maintenance for blocks	12.68
X_9	Treatment and maintenance when harvesting	6.46
X_{10}	Treatment and maintenance when fertilizing	24.87
X_{11}	Attack of rat pest (*Rattus tiomanicus*)	0.00
X_{12}	Attack of tirathaba (*Tirathaba* sp.)	0.00
X_{13}	Attack of caterpillar (*Setothosea asigna*)	0.00
X_{14}	Attack of bagworms (*Mahasena corbetti*)	0.00
X_{15}	Dosage of *borate*	0.00
X_{16}	Dosage of CPD-HIK	0.00
X_{17}	Dosage of CPD-25	0.00
X_{18}	Dosage of CPD-45	0.00
X_{19}	Dosage of $CuSO_4$	0.00
X_{20}	Dosage of MOP	0.00
X_{21}	Dosage of *rockphosphate*	0.00
X_{22}	Dosage of *super dolomite*	0.00
X_{23}	Dosage of urea	0.00
X_{24}	Topography or slope	0.00
X_{25}	Varieties of oil palm seedlings	0.00
X_{26}	Block hectares	0.00
X_{27}	Quantity of palm oil plant in a block	0.00
X_{28}	Class	0.00
X_{29}	Age of the palm oil plant	0.00

After the missing data is resolved, the data analysis is continued by make the lag for variables which have a time period, so that the topography, varieties of oil palm seedlings, block hectares, quantity of palm oil plant in a block, class, and age of the palm oil plant variables are not lagged. The lag used in this research reached 12 lag.

Analysis continued with modeling using random forest regression. The stages in conducting this random forest regression analysis is to split the data into training data and validation by 50% by blocks. Furthermore, modeling is done on the training data and the predicted values are released. The analysis continued with an evaluation of the model.

The evaluation carried out on the model is seen based on the correlation value between predictive value and testing data. The random forest regression analysis that was carried out obtained a correlation value of 0.9287, the value was already high so the model could be said to be good. Other evaluations are carried out by looking at the r*oot mean squared error (*RMSE) of 20.546, m*ean absolute* error (MAE) of 14.528, and R-Square of 0.826.

Then, the analysis was continued by determine at the variable important that built the model. It can be seen the 10 most important predictors that built this model were the quantity of palm oil, hectare (land area), age of palm oil plant, 1st lag of temperature, 1st lag of air humidity, 8th lag of wind velocity, 1st lag of wind velocity, 2nd lag of air humidity, 7th lag of wind velocity, and 11th lag of attack of rat pest. Of the many variables, only theimportant variables will be selected for the next modeling. This important predictor is selected based on the most important lag of each predictor for lag formed variables and variables that are not formed lag will be re-elected.

Then the next analysis is modeling on the data using only with selected important variables. Data returned is split into training data and data testing into 50% by blocks. The evaluation was carried out again on the second model and a correlation value of 0.9302 was obtained. Other evaluations are carried out again by looking at the RMSE of 20.379, MAE of 14.143, and R-Square of 0.829. The results show an improvement in the model and again seen for important predictors which built the model.

Table 4. Results of forecasting palm oil production.

Companies	Forecasting
A	33953.83
B	13620.04
C	43435.77
D	52594.26

The final analysis is to forecast the production of palm oil. Maximum data forecasting can bedone by using the minimum lag selected based on the important variables of the first model. Thus, the result for data forecasting is carried out for the next 1 month, namely for January 2019. Based on the second model, the 10 most important variable are quantity of palm oil, hectare (land area), age of palm oil plant, 8th lag of wind velocity, 1st lag of temperature, 1st lag of air humidity, 11th lag of attack of rat pest, 5th lag of rainfall, 1st lag of treatment and maintenance when fertilizer, and 11th lag of long of lighting. The analysis continued with forecasting palm oil production. **Table 4**shows the results of data forecasting carried out on each block and has been presented in the form of each companies.

Figure 3 shows a box plot from the real data, forecasting data performed, and testing data. Based on the **Figure 3**, it can be seen that the distribution of data forecasting results are not much different from the distribution of testing data and original data. **Table 5** shows the comparison of the distribution produced between the real data, forecasting data, and testing data.

Production of palm oil has a non-linear distribution. The performances of model are correlation value of 0.9287, RMSE of 20.546, MAE of 14.528, and R-Square of 0.826. Model with only the important variables of the first model selected has a better performance than model with all variables. The performances of second model are correlation value of 0.9302, RMSE of 20.379, MAE of 14.143, and R-Square of 0.829. The results show an improvement in the model. Maximum data forecasting can be done by using the minimum lag selected based on the important variables of the first model. So that production data forecasting is carried out for the next 1 month.

Table 5. Comparison of the distribution of real data, forecasting data, and testing data.

Description	Minimum	Quartile 1	Quartile 2	Mean	Quartile 3	Maximum

real data	0.00	38.00	71.00	75.56	104.00	360.00
forecasting data	2.612	57.239	120.782	109.454	156.098	243.396
testing data	0.4015	39.8315	74.3958	75.0875	100.562	224.920

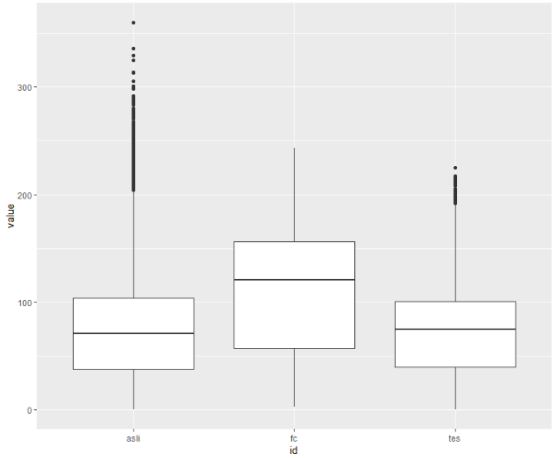

Fig. 3. Boxplot of real data, forecasting data, and testing data.

5Conclusion

Based on the research, production of palm oil has a non-linear distribution. The performances of model are correlation value of 0.9287, RMSE of 20.546, MAE of 14.528, and R-Square of 0.826. Model with only the important variables of the first model selected has a better performance than model with all variables. The performances of second model are correlation value of 0.9302, RMSE of 20.379, MAE of 14.143, and R-Square of 0.829. The results show an improvement in the model. Maximum data forecasting can be done by using the minimum lag selected based on the important variables of the first model. So that production data forecasting is carried out for the next 1 month.

References

[1] Breiman L, Friedman JH, Olshen A, Stone C J. Classification and Regression Trees. Chapman and Hall, US (1993).
[2] Breiman L, Cuttler A. Random Forest. Machine Learning. Vol. 45, pp.5-32. (2001)
[3]GAPKI (2018, January 30). Refleksi Industri Kelapa Sawit 2017 dan Prospek 2018. Retrieved from https://gapki.id/news/4140/refleksi-industri-kelapa-sawit-2017-dan-prospek-2018.
[4] Gujarati D. Basic Econometrics: 4th Edition. McGraw-Hill, US(2003).
[5] Hardon, J J,RajanaiduN, Van der Vossen. Elaeisguineensis Jacq. pp. 85-93.(2002).

[6] Murtilaksono, K, Siregar H H, Darmosarkoro W. Model neraca air di perkebunankelapasawit. JurnalPenelitianKelapaSawit. Vol. 15, pp. 21-35. (2007)
[7] Supranto J. Ekonometrika Dasar. Edisi 2.Lembaga Penerbitan Fakultas Ekonomi Universitas Indonesia,ID (1995).
[8] Verikas A, Gelzinis A, Becausekiene M. Mining data with random forest: A survey and result of new tests. Pattern Recognition. Vol 44, pp. 330-349. (2011).