DOI: 10.15849/IJASCA.240330.01

Int. J. Advance Soft Compu. Appl, Vol. 16, No. 1, March 2024
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Copyright © Al-Zaytoonah University of Jordan (ZUJ)

Modified Generalized Lasso for Variable Selection in Lag Distributed Modeling of Fresh Fruit Bunch Production from Oil Palm Plantations in Riau-Indonesia

Anang Kurnia[1], Septian Rahardiantoro[2], Sachnaz Desta Oktarina[3], Rahma Anisa[4], Nafisa Azzahra Nur Rahman[5], Dian Handayani[6]

[1,2,3,4,5]Department of Statistics, IPB University, Indonesia
e-mail: anangk@apps.ipb.ac.id
[6]Department of Statistics, Universitas Negeri Jakarta, Indonesia

Abstract

This research identified factors affecting the productivity of oil palm fresh fruit bunches (FFB) in metric tons per ha (hectare). Current research rarely includes spatial and temporal aspects, so we proposed the modified Generalized Lasso, which can be used in the lag-distributed regression by considering the adjacency of time lags and locations in the data. The modification is located in how the definition of the regression model and the penalty matrix in the Generalized Lasso, which considers the adjacencies between blocks and time lags. The method is applied on plantations in the Riau context in Indonesia. The oil palm management data used consists of 42 months of observations in 16 planting blocks spanning from 2020 to 2023. The response variable was the productivity of oil palm FFB, with predictor variables consisting of the number of rainy days, rainfall, the dosage of NPK fertilizer, and palm age. We compared our proposed method with standard Lasso. As a result, our proposed model obtained a smaller error value than the standard lasso models. It is indicated that the lag of the productivity variable and the lag of the number of rainy days influence the FFB productivity for almost all blocks.

Keywords: *fresh fruit bunches, lag distributed modeling, generalized lasso regression, spatio-temporal modeling.*

Received 19 October 2023; Accepted 13 December 2023

1 Introduction

The palm oil industry is a leading commodity that contributes to national foreign exchange in Indonesia. It is recorded that more than 13% of non-oil and gas exports and 3.5% of the national GDP contributors are generated from the oil palm plantation industry [1]. These economic benefits must of course be accompanied by sustainability initiatives to achieve the sustainable development goals by 2030. One of the ways to realize the principle of sustainability is by increasing productivity through precision farming, automation, or plantation intensification [2]. With the advance of technology, it is expected that the same optimal production of fresh fruit bunches (FFB) can be produced from relatively fewer inputs/production factors. In attempt to produce optimal FFB production, palm oil plantation companies generally conduct fruit survey (three times a year) to determine production projections. This survey activity costs a lot of money, resources, and manpower, yet the results are not necessarily 100% accurate. In order to increase the effectiveness and efficiency of harvest costs and reduce environmental risks, it is expected that forecasting analysis of FFB production can be carried out by the means of machine learning approach.

Previous machine learning approach by Firdawanti et al. [3], namely random forest lag distributed regression, was conducted to identify factors that influence FFB production projections. As a result, the projection is strongly influenced by planting age, area, plant density, temperature lag-9, and humidity lag-10. However, it is necessary to study more deeply whether other production factors such as genetic material and maintenance techniques (fertilization) have a real influence on the production system considering that the highest production costs in palm oil production are fertilizer costs (more than 50% of the total costs).

Meanwhile, it is documented that fertilizer application will help to increase oil palm productivity [4]. Aside from that, other determining factors that might affect the production of FFB are available radiation, CO2 concentration, temperature, planting material, planting density, culling, pruning, pollination, crop recovery, rainfall, soil type, topography, water conservation techniques, pest, and diseases [5], [15]. This research considers identifying factors which effected the FFB production in Riau Plantation according to the plantation block and some series of time points.

With considerable amount of variables in the model, accompanied by a time series data type with several time points bring about a wide choice of analysis methods that can be carried out. This research seeks to develop a generalized lasso method in lag distributed regression modeling which can later be used in identifying influential variables. The Generalized Lasso method is a method for estimating parameters in a regression model that utilizes the regularization approach along with the penalty matrix based on certain tuning parameters, which can be used to reduce regression coefficients with adjacent structures [6]. This approach has enormous potential to be developed in lag distributed regression analysis efforts, considering that there are many applications of the Generalized Lasso method on time series data [7]. By using the Generalized Lasso method, we aim to identify a group of lag variables that affect response. This research also attempts to determine the appropriate method for selecting the tuning parameter to use in adapting it to the penalty matrix structure which cannot be partitioned using cross validation methods in general. In this case, approximate leave-one-out cross-validation (ALOCV) [8] is proposed to be applied.

2 Literature Review

2.1 Lasso and Generalized Lasso

Suppose a response variable vector $\boldsymbol{y} \in R^{n\times 1}$, predictor variable matrix $\boldsymbol{X} \in R^{n\times p}$, regression coefficient $\boldsymbol{\beta} \in R^{p\times 1}$, and error vector $\boldsymbol{\varepsilon} \in R^{n\times 1}$, with a regression model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}. \tag{1}$$

The Lasso method was developed by Tibshirani [9] to estimate the regression coefficient $\boldsymbol{\beta}$ by adding the L_1 penalty. The estimated value of $\boldsymbol{\beta}$ can be obtained by minimizing

$$||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1, \tag{2}$$

where $||\boldsymbol{a}||_1 = \sum_i |a_i|$ and $||\boldsymbol{a}||_2 = \sqrt{\sum_i |a_i|^2}$; with $\boldsymbol{a}$ denoting an arbitrary vector, and $\lambda \geq 0$ is the tuning parameter. The λ is a value to reduce the estimated regression coefficient. If $\lambda > 0$, then the regression coefficient will decrease towards 0, which is useful for selecting predictor variables. If $\lambda = 0$, then equation (2) will be equivalent to the OLS (ordinary least square) method for estimating the regression coefficient $\boldsymbol{\beta}$.

The Generalized Lasso method was developed by Tibshirani and Taylor [7] by adding a penalty matrix $\boldsymbol{D} \in R^{m\times p}$ to the penalty L_1. The penalty matrix $\boldsymbol{D}$ represents the geometric structure contained in the regression coefficient $\boldsymbol{\beta}$. The estimated value of $\boldsymbol{\beta}$ in the Generalized Lasso method can be obtained by minimizing

$$||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2 + \lambda||\boldsymbol{D\beta}||_1, \tag{3}$$

In equation (3), it can be observed that if $\boldsymbol{D} = \boldsymbol{I}$, then this problem will be equivalent to the Lasso problem in equation (2).

2.2 Approximate Leave-One-Out Cross-Validation Method

One very important aspect in obtaining the appropriate estimated coefficients in the Generalized Lasso method is the selection of the optimal tuning parameter λ. Zhao and Bondell [10], applied the k-fold cross-validation (k-fold CV) method with $k = 10$ in the application of the Generalized Lasso method. However, the k-fold CV method has a fairly large bias in terms of prediction error [8] [11]. In addition, with spatio-temporal based data, partitioning the data into k-folds which damages the location and time structure is highly discouraged. This research applies approximate leave-one-out cross-validation (ALOCV), which is a method with a relatively smaller bias, with a low level of computational load [8].

Below, is the ALOCV algorithm using the Generalized Lasso method for selecting the optimal tuning parameter λ [8].

a. Estimate $\boldsymbol{\beta}$ which is the solution of equation (3),
b. Estimate $\boldsymbol{u}$ which is the solution to the dual problem of equation (3), which can be expressed as:

$$\frac{1}{2}\|\boldsymbol{\gamma} - \boldsymbol{y}\|_2^2 \; s.t. \;\; \|\boldsymbol{u}\|_\infty \leq \lambda \text{ and } \boldsymbol{X}^T\boldsymbol{\gamma} = \boldsymbol{D}^T\boldsymbol{u}. \tag{4}$$

c. Set aside rows in the penalty matrix $\boldsymbol{D}$ that belong to the set $E = \{s = 1, \dots, m \colon |\widehat{\boldsymbol{u}}_s| = \lambda\}$, to form submatrix $\boldsymbol{D}_{-E}$,
d. Form a matrix $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{B}$, where $\boldsymbol{B}$ has the null space of $\boldsymbol{D}_{-E}$,
e. Compute $\boldsymbol{H}^* = \boldsymbol{A}\boldsymbol{A}^+$, where $\boldsymbol{A}^+$ is the Moore-Penrose inverse matrix of $\boldsymbol{A}$,
f. Calculate ALOCV error as follows:

$$\frac{1}{n}\sum_{c=1}^{n}\left(\frac{y_c - \boldsymbol{x}_c^T\widehat{\boldsymbol{\theta}}}{1 - h_{cc}^*}\right)^2. \tag{5}$$

Where h_{cc}^* is the c-th diagonal of the matrix H^*.

3 Modified Generalized Lasso for Lag Distributed Model based on Adjacency Blocks

Technically, the Generalized Lasso method can be applied in many cases, depending on how we define the predictor matrix $\boldsymbol{X}$ and the penalty matrix $\boldsymbol{D}$. When $\boldsymbol{X} = \boldsymbol{I}$, the Generalized Lasso is generally applied to spatial smoothing, spatial clustering, and trend filtering. The applications of spatial smoothing and clustering could be found in [7] [12-16]. The application of trend filtering could be found in Kim et al. [17].

Moreover, the Generalized Lasso can be applied in the modeling analysis when $\boldsymbol{X} \neq \boldsymbol{I}$. Some applications for modeling are in [10] [18-19]. This study focuses on the condition when the predictor matrix $\boldsymbol{X} \neq \boldsymbol{I}$. In practice, this study modifies the Generalized Lasso by considering to the lag distributed regression model and the adjacencies of each plantation block.

Suppose the time series data $\{y_t, x_{t1}, \dots, x_{tp}\}$ for time points $t = 1,2,\dots,T$, then the lag distributed regression model with a maximum lag-K can be written as

$$y_t = \beta_0 + \sum_{j=1}^{p}\sum_{k=1}^{K} x_{(t-k),j}\beta_{kj} + \sum_{k=1}^{K} y_{(t-k)}\beta_{k(p+1)} + \epsilon_t, \tag{6}$$

where $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = \sigma^2$ [20]. In this case, this study also considers the adjacency between planting blocks. So, the regression model becomes

$$y_t = \beta_0 + \sum_{i=1}^{B}\sum_{j=1}^{p}\sum_{k=1}^{K} x_{(t-k),i,j}\beta_{kij} + \sum_{i=1}^{B}\sum_{k=1}^{K} y_{(t-k),i}\beta_{ki(p+1)} + \epsilon_t \tag{7}$$

where $i = 1,2,\dots B$ is the index for the planting block. Based on this case, modifying the generalized lasso to estimate the regression coefficients of model (7) can be expressed as a problem

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum\left(y_t - \beta_0 - \sum_{i=1}^{B}\sum_{j=1}^{p}\sum_{k=1}^{K} x_{(t-k),i,j}\beta_{kij} \right.\right. \\ \left.\left. - \sum_{i=1}^{B}\sum_{k=1}^{K} y_{(t-k),i}\beta_{ki(p+1)}\right)^2 + \lambda\|\boldsymbol{D\beta}\|_1\right\}. \tag{8}$$

The penalty matrix $\boldsymbol{D}$ is defined as a combination of two aspects, the first stating the adjacency between blocks in a group of blocks, and the adjacency of the lag between predictor variables. $\boldsymbol{D}_1$ is defined as a diagonal block matrix indicating the adjacency between blocks for each predictor variable in model (8), which is expressed as:

$$\boldsymbol{D}_1 = \begin{bmatrix} \boldsymbol{D}_b & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}_b & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{D}_b \end{bmatrix}, \tag{9}$$

where the matrix $\boldsymbol{D}_b$ shows the adjacency between blocks in each block group. For example, blocks A_1, A_2, and A_3 are in one group, then their adjacencies can be expressed as three rows in the matrix $\boldsymbol{D}_b$. The first row contains the values -1 and 1 in the columns corresponding to blocks A_1 and A_2 and the remaining components contain the value 0. The second row contains the values -1 and 1 in the columns corresponding to blocks A_1 and A_3 and the remaining components contain the value 0. The third row contains the values -1 and 1 in the columns corresponding to blocks A_2 and A_3 and the remaining components contain the value 0. These three rows can be written as follows:

$$D_b = \begin{array}{c} \begin{matrix} A_1 & A_2 & A_3 & \dots \end{matrix} \\ \begin{bmatrix} -1 & 1 & 0 & \dots \\ -1 & 0 & 1 & \dots \\ 0 & -1 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array}. \tag{10}$$

In addition, $\boldsymbol{D}_2$ is defined as a diagonal block matrix containing a sequential structure for adjacent lag of predictor variables, which is expressed as

$$D_2 = \begin{bmatrix} D_l & 0 & \cdots & 0 \\ 0 & D_l & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_l \end{bmatrix}, \tag{11}$$

where $\boldsymbol{D}_l$ indicating the adjacency of lags on the same predictor variable, with each row defined as a connection between a pair of adjacent lags marked with elements -1 and 1, along with element 0 for the others [7]. As a result, the penalty matrix $\boldsymbol{D}$ can be expressed as

$$D = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}. \tag{12}$$

4 Data and Method

4.1 Data Description

This study uses secondary data, namely the productivity of oil palm FFB in 2020 to June 2023 which is observed every month in the oil palm plantations in Riau, Indonesia [21]. The smallest unit of observation in this study is the block. There were 16 blocks observed over 42 months, so there were 672 observations in this study. The predictor variables used are factors that are considered to influence oil palm FFB production [5]. The following is a list of predictor variables used:

Table 1: List of Variables

Variable	Variable Name	Type	Unit	Reference
Y	Productivity	Numeric	MT/ha	
X_1	Palm age	Numeric	year	[5] [22]
X_2	Rainyday	Numeric	days (Number of rainy days)	[5] [23]
X_3	Rainfall	Numeric	mm/day	[5] [24] [25]
X_4	Dosage of NPK fertilizer	Numeric	kg/palm/ha	[5] [26]

This study uses lag variables 1 to 12 for the Rainyday and Rainfall variables, also lag variables 1 to 12 for the Productivity variable as predictor variables in the model.

In this data there are 16 planting blocks which are divided into 7 groups as shown in Table 2. In one block group, the location of each block in it is close to each other. On the other hand, between block groups, the locations are far apart. In this data, the block group that contains the most blocks are PNDEE which contains 4 blocks.

Table 2: Planting Blocks of Data

Block Group	Block Name
PNDEC	PNDEC16a
PNDED	PNDED16a
PNDEE	PNDEE10c, PNDEE11a, PNDEE15a, PNDEE20a
PNDEF	PNDEF09c, PNDEF20a
PNDEG	PNDEG14a, PNDEG16a
PNDEH	PNDEH16a, PNDEH17a, PNDEH18a
PNDEI	PNDEI13a, PNDEI14a, PNDEI15a

4.2 Data Analysis

a. Exploration of each variable using boxplot and histogram.

At this stage, data exploration is carried out to see the distribution of productivity, palm age and NPK fertilizer dosage variables in the 16 existing blocks. Then, variables of number of rainy days and rainfall are displayed in histogram to see the distribution of the data.

b. Apply Lasso regression.

For identifying factors that affecting the productivity of FFB, firstly the Lasso [9] (according to equation (2)) is applied for estimating coefficients of the linear model:

$$y_{ti} = \beta_0 + x_{1i}\beta_{1i} + \sum_{j=2}^{3}\sum_{k=1}^{12} x_{t-k,ji}\beta_{kji} + x_{4i}\beta_4 + \sum_{k=1}^{12} y_{t-k,i}\beta_{k5i} + \epsilon_{ti} \qquad (13)$$

where $i = 1,2 \ldots 16$ which indicates indices of block. The influenced variables are identified for each block based on non-zero estimated coefficients obtained.

Next, 10-fold CV is applied to select optimum tuning parameter λ of Lasso for each block. Finally, the average CV-error for all is calculated as the goodness of fit value of the model. These Lasso modeling use R software with the "glmnet" package.

c. Apply the Generalized Lasso.

The modified Generalized Lasso is applied for estimating coefficients based on problem using our proposed model on equation (8). The $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are defined as a diagonal block matrix for the adjacency between blocks for each predictor variable and a sequential structure for adjacent lag of predictor variables, respectively.

In this step, the optimum tuning parameter λ is searched using ALOCV as described in subchapter 2.2 and then calculated the CV-error as goodness of fit value of the model. The result is displayed using a heatmap clustering for easier understanding. The Generalized Lasso modeling use R software with the "genlasso" package.

d. Interpretation of the results.

This stage compares the modeling results based on CV-error between the estimated model coefficients with Lasso for each block and the Generalized Lasso. The best model is selected which has the smallest CV error. Then, in the best model, it is interpreted what variables influence the FFB productivity variable.

5 Results and Discussions

5.1 Data Exploration

It is necessary to explore the data distribution as well as to compare them between the observed blocks for a better understanding before the modeling process. Figure 1 shows that there are slight differences among the productivity on each block, in terms of its median. Some blocks of the highest, which exceeds 2 MT/Ha, are PNDED16a, PNDEF20a, PNDEG14a, and PNDEG16a. In contrast, block of PNDEF09c and PNDEI13a tend to yield the lowest productivity. However, blocks with lower productivity tend to have smaller variances. Overall, the average of productivity is 1.875 MT/ha with the standard deviation 0.531 MT/ha. In addition, some outliers are observed in several blocks. The planting block group with the highest productivity is the PNDEG group. In the PNDEE block group, productivity is quite diverse, with the lowest productivity found in PNDEE11a. However, the other blocks in the PNDEE block group have almost the same range of productivity values.

Furthermore, Figure 2 demonstrates that based on its palm age, the blocks could be categorized into four groups: the lowest (median of 12 years), the middle lower age (median of 13 years), the middle upper age (median of 14 years), and the highest age (median of 15 years). There two blocks having the highest age, PNDEG14a and PNDEG16a. Meanwhile, the blocks with the youngest palm ages are in blocks PNDEE10c, PNDEF09c, and PNDEI15a. In the PNDEE block group, there is diversity in palm age. The lowest palm age is in PNDEE10c, followed by PNDEE11a and PNDEE20a, and PNDEE15a has the highest age range. Overall, the average of palm age is 13.66 years old with standard deviation 1.39 years old.

In general, the palm age of all blocks tend to be identically distributed, that is right skewed with no outliers.

In terms of the NPK dosage, each block tends to have different distribution (Figure 3). It can be portrayed that most of them are left skewed. Moreover, it can be seen that the dosage tends to differentiate blocks into two groups: lower dosage (median value of about 4 kg/palm/ha), and the higher dosage (median value of about 6 to 7 kg/palm/ha). The highest dose is in the PNDED16a block, while the lowest dose is in the PNDEH and PNDEI block groups. The PNDEH and PNDEI block groups have a relatively lower NPK dose range compared to the other block groups. On average, the NPK dose for all blocks reached 3.892 kg/palm/ha with a standard deviation of 2.713 kg/palm/ha.

Figure 4 displays the histogram of the variable number of rainy days (Figure 4(a)) and the rainfall variable (Figure 4 (b)). In both histograms, it can be seen that the distribution is skewed to the right. There is an extreme value for the rainfall variable, namely 577 mm/day. This is unsurprising because it is quite common to observe the large extreme value of rainfall. Based on this data, the average value of the number of rainy days reached 16.48 days with a standard deviation value of 3.60 days. The rainfall variable has an average of 218.86 mm/day with a standard deviation of 90.04 mm/day.

DOI: 10.15849/IJASCA.240330.01

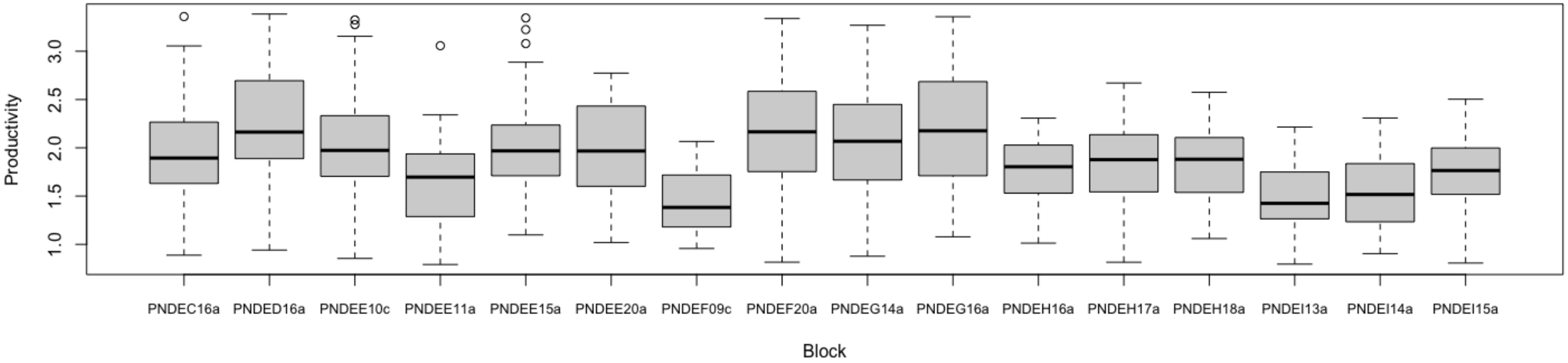


Fig 1: Boxplots of productivity (MT/ha) for each block

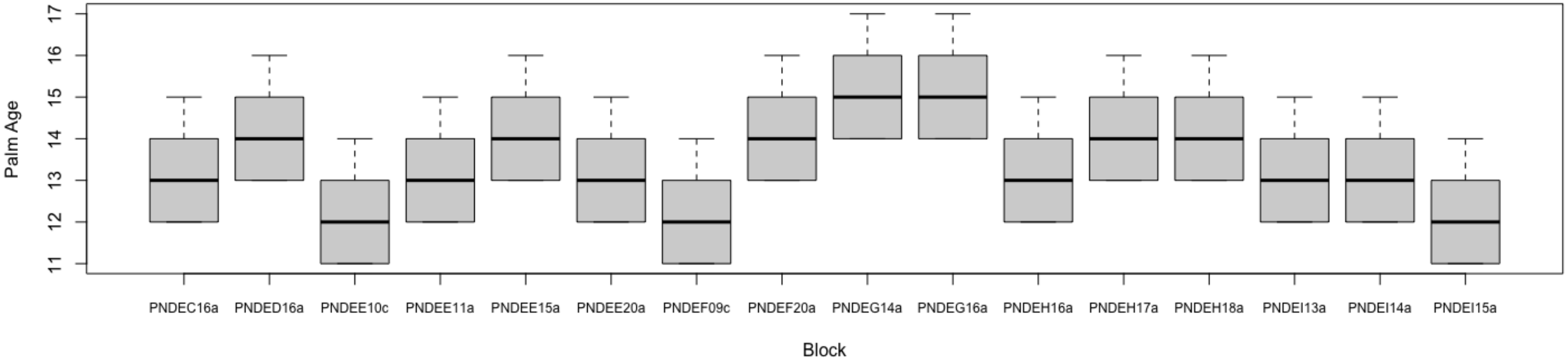


Fig 2: Boxplots of palm age (year) for each block

Received 19 October 2023; Accepted 13 December 2023

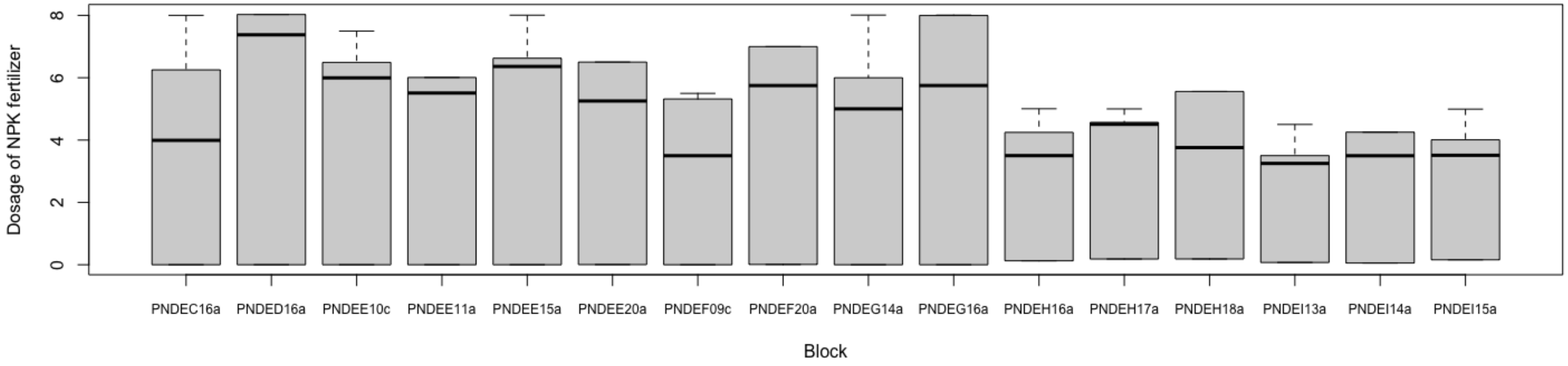
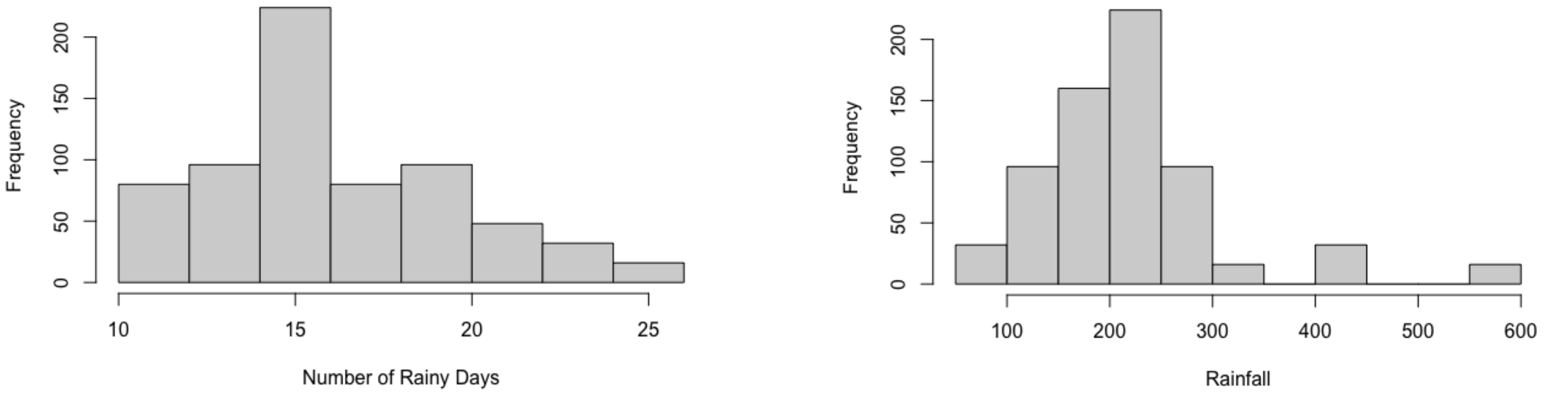


Fig 3: Boxplots of NPK dosage (kg/palm/ha) for each block



(a) (b)

Fig 4: Histogram of (a) number of rainy days (day) and (b) rainfall (mm/day)

DOI: 10.15849/IJASCA.240330.01

5.2 Result of Lasso Regression for Each Block

Modeling was conducted by using variables derived from lag-1 to lag-12 from the variables of the number of rainy days and the average rainfall. Also, the modeling was carried out using the lag-1 to lag-12 variables of the productivity variable as predictor variables according to model (13).

The Lasso method was applied to estimate the coefficients of the model (13) for each block. This case dismissed the adjacency between blocks in the block groups. The value of λ in the Lasso is determined using a 10-fold CV. The Table 3 displays the estimated coefficients of the variables resulting from modeling using Lasso.

Table 3: Coefficient estimates of lasso regression for each block

Block	X_1	X_2_lag7	X_2_lag8	X_2_lag9	Y_lag1	Y_lag11	Y_lag12
PNDEC16a	0.1323	0	0	0	0	0	0
PNDED16a	0.1429	0.0112	0	0	0	0	0
PNDEE10c	0.0862	0.0095	0.0044	0.0152	0	0	0.2185
PNDEE11a	0.1076	0	0	0	0	0	0
PNDEE15a	0.1281	0	0	0	0	0	0
PNDEE20a	0.1431	0	0	0	0	0	0.0182
PNDEF09c	0.0822	0	0.0067	0	0.1364	0	0.0807
PNDEF20a	0.1369	0	0	0	0	0	0
PNDEG14a	0.1136	0	0	0	0	0.0544	0.0934
PNDEG16a	0.1231	0	0	0.0040	0	0.0004	0.1013
PNDEH16a	0.1268	0	0	0	0.0205	0	0
PNDEH17a	0.1201	0	0	0	0.0458	0	0
PNDEH18a	0.1219	0	0	0	0	0	0
PNDEI13a	0.0976	0	0	0	0	0	0
PNDEI14a	0.1043	0	0	0	0	0	0
PNDEI15a	0.1141	0	0.0005	0	0.1558	0	0

Based on the Table 3, it can be obtained that the productivity of oil palm FFB in all blocks is affected by the palm age. Other influencing factors are the number of rainy days in the previous 7 to 8 months, as well as the productivity value of oil palm FFB itself in the previous 1 month and 11 to 12 months, which varies in certain plant blocks. In detail, the variable number of rainy days in the previous 7 months impacted the productivity of FFB oil palm in blocks PNDED16a and PNDEE10c. Meanwhile, the number of rainy days in the previous 8 months affected FFB oil palm productivity also in the PNDEE10c block, along with the PNDEF09c and PNDEI15a blocks. The number of rainy days in the previous 9 months also affects the PNDED16a block and the PNDEG16a block. The FFB oil palm productivity variable in the previous month had an effect on blocks PNDEF09c, PNDEH16a, PNDEH17a, and PNDEI15a. Apart from that, FFB oil palm productivity in the previous 11 months and 12 months had an effect on the PNDEG14a and PNDEG16a blocks. Furthermore, FFB oil palm productivity in the previous 12

Received 19 October 2023; Accepted 13 December 2023

months also affected blocks PNDEE10c, PNDEE20a, and PNDEF09c. As a result, these models have an average CV-error of 0.249.

5.3 Result of the Generalized Lasso in the Modified Model

Based on our modified model in (8), the Generalized lasso is carried out with a penalty matrix $\boldsymbol{D}$ with dimensions of 1164×608. The format of the predictor matrix $\boldsymbol{X}$ is also adjusted based on the embedding blocks which are arranged in diagonal blocks, so that the dimensions are 672×608. By using the ALOCV method [8], a value of $\lambda = 252.6$ is obtained with a CV-error value of 0.204 and degrees of freedom of 56. Figure 5 describes the line plot between ALOCV-error with λ.

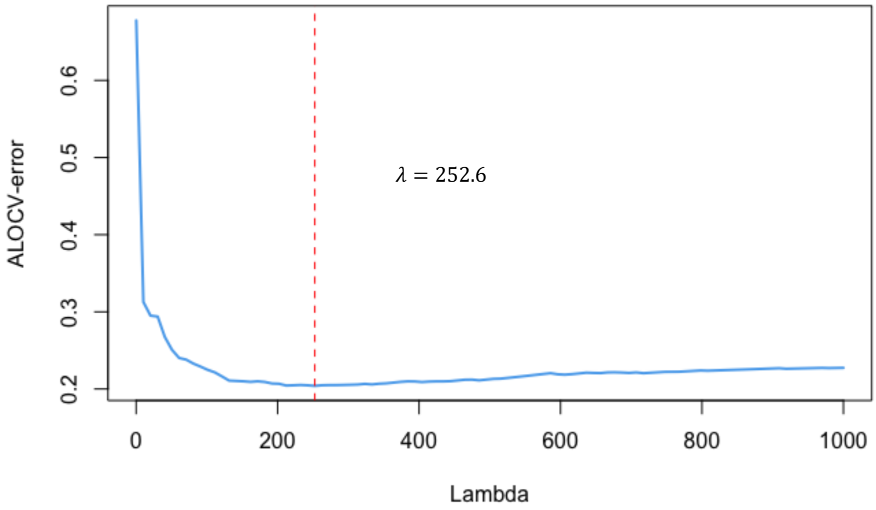


Fig 5: Line plot of ALOCV-error based on value of λ

The Generalized Lasso is applied based on the λ selected with ALOCV. The results of which are summarized in a heatmap plot with hierarchical clustering is displayed in Figure 6. As a result, there are three clusters of variables, namely Cluster 1 for variables which is a collection of variables that have a relatively strong influence which contains the productivity variable for lags 1 to 12. Cluster 2 for variables is a collection of variables that have a moderate influence in the model, which contains the variable number of rainy days from lags 1 to 12. And cluster 3 for variables contains the remaining set of variables that have no influence in the model. Apart from that, there were also three block clusters formed. Cluster 1 for blocks contains blocks with relatively the same variable influence which consists of 12 blocks, namely PNDEE11a, PNDEE20a, PNDEE10c, PNDEE15a, PNDEF20a, PNDEF09c, PNDEH18a, PNDEH17a, PNDEH16a, PNDEI15a, PNDEI13a, and PNDEI14a. Cluster 2 for blocks contains blocks that do not have influential variables, namely blocks PNDED16a and PNDEC16a. Cluster 3 for blocks contains blocks PNDEG14a and PNDEG16a which have a negative influence on several variables in them.

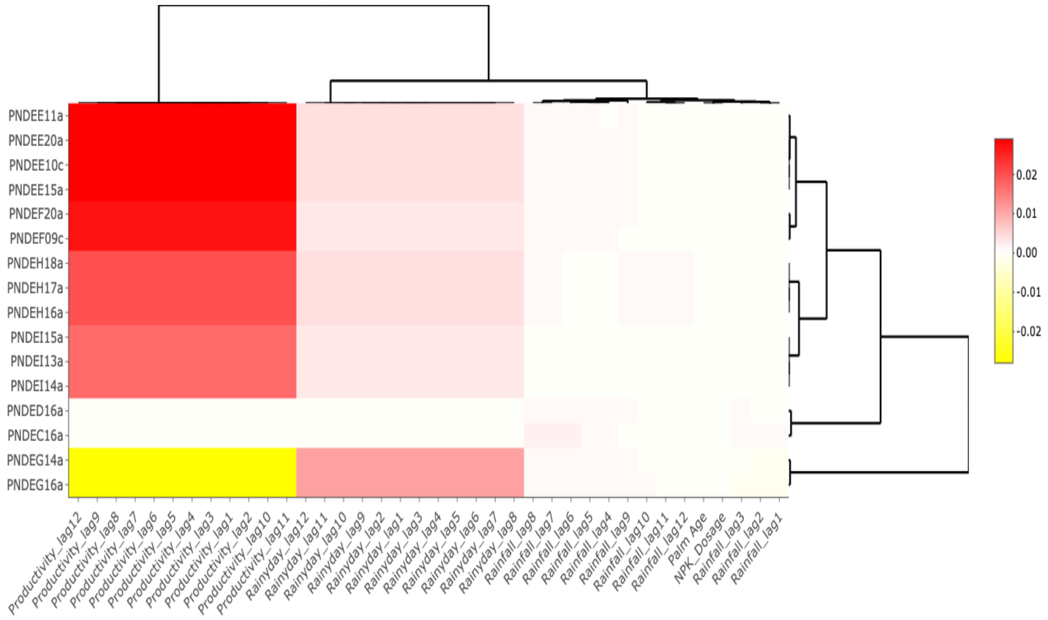


Fig 6: Heatmap clustering of generalized lasso coefficient estimates

In summary, based on these results, it can be seen that the lag variable of productivity and the lag variable for the number of rainy days influence the response variable for all blocks except for the block PNDED16a and PNDEC16a. In general, rainfall, palm age and NPK fertilizer dosage do not affect the response variable.

5.4 Interpretation of the Results

Based on the two modeling approaches used in this research, the Generalized Lasso model was chosen because it has a CV-error of 0.204 which is lower than the Lasso model (0.249). As a result, there are clusters of the same influence on the variables, and cluster into blocks with the same pattern. Based on these results, it can be seen that the variables rainfall, plant age, and NPK fertilizer dosage have no influence in the model for each block. Meanwhile, the lag of productivity variable and the lag of number of rainy days have influences in the model. In this result, the influence of the lag of productivity and the influence of the lag of the number of rainy days have a common influence in the model, therefore any lag can be selected to be used in the model.

6 Conclusion

This research developed the application of the generalized lasso in the case of distributed lag models which takes into account the adjacencies between lags on predictor variables and also the adjacencies between locations (blocks). In its application, the proposed method applied to oil palm FFB productivity data in Riau, Indonesia for 16 planting blocks, which was also compared with the Lasso regression method applied to each block. As a result, based on the CV-error values obtained, the model with our proposed Generalized Lasso was chosen as the best model because it has a smaller CV-error value. Thus, the variables that influence oil palm FFB productivity can be identified based on the planting block and the time lag of the influencing predictor variables, namely productivity at the previous time points and the variable number of rainy days along with its lag variables.

ACKNOWLEDGEMENTS
This research is supported by The Directorate General of Higher Education, Research and Technology of the Ministry of Education, Culture, Research and Technology the Republic of Indonesia based on contract number 102/E5/PG.02.00.PL/2023 for the Fundamental Research schema.

References

[1] Statistics Indonesia. (2022). Indonesia in Numbers 2022.

[2] Auernhammer, H. (2001). Precision farming-the environmental challenge. *Computers and electronics in agriculture*, 30(1-3), 31-43.

[3] Firdawanti, A R. (2019). Peramalan Produksi Kelapa Sawit Menggunakan Algoritma Random Forest Lag Distributed Regression. (Thesis. IPB University)

[4] Corley, R. H. V., & Tinker, P. B. (2008). *The oil palm*. John Wiley & Sons.

[5] Woittiez, L. S., Van Wijk, M. T., Slingerland, M., Van Noordwijk, M., & Giller, K. E. (2017). Yield gaps in oil palm: A quantitative review of contributing factors. *European Journal of Agronomy*, 83, 57-77.

[6] Arnold, T. B., & Tibshirani, R. J. (2016). Efficient implementations of the generalized lasso dual path algorithm. *J Comput Graph Stat* 25(1):1–27. https://doi.org/10.1080/10618600.2015.1008638

[7] Tibshirani, R. J., & Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3). https://doi.org/10.1214/11-AOS878

[8] Rad, K. R., & Maleki, A. (2020). A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *Journal of the Royal Statistical Society Series B*. 82(4): 965-996. arXiv:1801.10243

[9] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol* 58(1):267–288

[10] Zhao, Y., & Bondell, H. (2020). Solution paths for the generalized lasso with applications to spatially varying coefficients regression. *Comput Statis Data Anal* 142:106821. https://doi.org/10.1016/j.csda.2019.106821

[11] Rad, K. R., Zhou, W., & Maleki, A. (2020). Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In: Proceedings of the 23rd international conference on artificial intelligence and statistics (AISTATS), pp 108

[12] Tibshirani, R., & Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* 9(1):18–29. https://doi.org/10.1093/biostatistics/kxm013

[13] Yang, T., Liu, J., Gong, P., Zhang, R., Shen, X., & Ye, J. (2016, August). Absolute fused lasso and its application to genome-wide association studies.

In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1955-1964).

[14] Yanti, Y., & Rahardiantoro, S. (2023). Analysis of Regency and City Pneumonia Clusters in West Java 2020. *Komputasi: Jurnal Ilmiah Ilmu Komputer dan Matematika* 20(1): 44-50. https://doi.org/10.33751/komputasi.v20i1.6412

[15] Rahardiantoro, S., & Sakamoto, W. (2022). Optimum Tuning Parameter Selection in Generalized lasso for Clustering with Spatially Varying Coefficient Models. In IOP Conference Series: Earth and Environmental Science (Vol. 950, No. 1, p. 012093). IOP Publishing. https://doi.org/10.1088/1755-1315/950/1/012093.

[16] Rahardiantoro, S., & Sakamoto, W. (2023). Spatio-temporal clustering analysis using generalized lasso with an application to reveal the spread of Covid-19 cases in Japan. Computational Statistics, 1-25. https://doi.org/10.1007/s00180-023-01331-x

[17] Kim, S.-J., Koh, K., Boyd, S., & Gorinevsky, D. 2009. l1 Trend Filtering. *SIAM Review* 51(2):339– 360. https://doi.org/10.1137/070690274

[18] Rahardiantoro, S., & Sakamoto, W. (2021, March). Clustering regions based on socio-economic factors which affected the number of COVID-19 cases in Java Island. In Journal of Physics: Conference Series (Vol. 1863, No. 1, p. 012014). IOP Publishing. https://doi.org/10.1088/1742-6596/1863/1/012014

[19] Rahardiantoro, S., & Sakamoto, W. (2022). Spatially varying coefficient modeling of numerical and categorical predictor variables in the generalized lasso. Journal of Environmental Science for Sustainable Society, 11(Supplement), PP05_p16-PP05_p19.

[20] Suo, X., & Tibshirani, T. (2016). An Ordered Lasso and Sparse Time-lagged Regression. *Technometrics*, 58(4): 415-423

[21] Bumitama Gunajaya Agro. (2023) Riau Plantation Report.

[22] Darmawan, S., Takeuchi, W., Haryati, A., AM, R. N., & Na'Aim, M. (2016, June). An investigation of age and yield of fresh fruit bunches of oil palm based on ALOS PALSAR 2. In IOP Conference Series: Earth and Environmental Science (Vol. 37, No. 1, p. 012037). IOP Publishing. doi: 10.1088/1755-1315/37/1/012037.

[23] Paterson, R. R. M., & Lima, N. (2018). Climate change affecting oil palm agronomy, and oil palm cultivation increasing climate change, require amelioration. *Ecology and Evolution* 8(1): 452–461. doi: 10.1002/ece3.3610

[24] Oktarina, S. D., Nurkhoiry, R., & Pradiko, I. (2021, June). The effect of climate change to palm oil price dynamics: a supply and demand model. In IOP Conference Series: Earth and Environmental Science (Vol. 782, No. 3, p. 032062). IOP Publishing.

[25] Abubakar, A., Ishak, M. Y., & Makmom, A. A. (2022). Nexus between climate change and oil palm production in Malaysia: a review. Environmental Monitoring and Assessment, 194(4), 262. doi: 10.1007/s10661-022-09915-8.

[26] Ummah, A. S., & Surahman, M. (2020). Acceleration and improvement of productivity by inorganic and organic fertilizer application for six-year-old mature palm oil. In IOP Conference Series: Earth and Environmental Science (Vol. 418, No. 1, p. 012047). IOP Publishing. doi: 10.1088/1755-1315/418/1/012047

Notes on contributors

Anang Kurnia is Associate Professor at the Department of Statistics, IPB University, Bogor, Indonesia. His main teaching and research interests include Data Science, Statistical Machine Learning, Statistical Inference, Generalized Linear Mixed Model, and Small Area Estimation. He has published several research articles in international journals of statistics and data science.

Septian Rahardiantoro is lecturer at the Department of Statistics, IPB University, Bogor, Indonesia. His main teaching and research interests include Data Science, Statistical Machine Learning, and Statistical Modeling in Environmental Science.

Sachnaz Desta Oktarina is lecturer at the Department of Statistics, IPB University, Bogor, Indonesia. Her research interest are Statistical Modeling and Data Science in Sustainability and Natural Resource Management.

Rahma Anisa is lecturer at the Department of Statistics, IPB University, Bogor, Indonesia. Her research interest are Spatial Statistical Modeling, Data Science, and Actuarial Study.

Nafisa Azzahra Nur Rahman is master’s student at the Department of Statistics, IPB University, Bogor, Indonesia. Her current research is the development and application of machine learning for forecasting the productivity of oil palm fresh fruit bunches in Indonesia.

Dian Handayani is Associate Professor at the Department of Statistics, Universitas Negeri Jakarta, Indonesia. Her main teaching and research interests include Statistical Inference, Generalized Linear Mixed Model, Item Response Theory, and Small Area Estimation. She has published several research articles in international journals of statistics and its applications.