Determining soil fertility using principal component regression analysis of oil palm plantation in West Sulawesi, Indonesia

U S Pasaribu¹, N Nurhayati², N F F Ilmi¹, and K N Sari¹

¹Statistics Research Division, Faculty of Mathematics and Natural Sciences, Bandung Institute of Technology, Indonesia
²Department of Mathematics, Faculty of Mathematics and Natural Sciences, Jenderal Soedirman University, Indonesia

E-mail: udjianna@math.itb.ac.id

Abstract. Palm oil is one of the most popular vegetable oils in the world besides soybean oil. Nowadays, more than 80% of the world's palm oil supply comes from Indonesia and Malaysia. As the largest supplier, Indonesia must be improved and maintained its productivity with good management, such as maintaining soil fertilization to produce the best oil palm. This activity requires analysis of the soil chemical factors and soil conditions (fertile or not) from oil palm plantations, which involves a large amount of data. Principal component regression is an appropriate statistical method to solve the problem. In this method, the coefficient of the regression model is the coefficient of the principal components (PCs). Dimension reduction is performed on the predictor variables that have a high correlation so that the PCs have an insignificant correlation. The data are obtained by measuring soil samples around the trees from oil palm plantation in West Sulawesi, Indonesia. The data consist of sixteen variables and thirty-six observations (0-20 cm) for each variable. There are three PCs that become predictors of the regression model with information absorption rates reaching 78%, i.e., some macronutrients (Potassium oxide, Potassium, Calcium, and Cation exchange capacity), soil acidity and organic properties (Carbon and Nitrogen). Furthermore, the accuracy of the estimation value in this logistic regression model reaches 90% by using stepwise backward method.

1. Introduction

Since the 19th century, palm oil has been recognized in the industry, and their use has grown rapidly. It has more advantages than other vegetable oils, such as being more durable, resistant to pressure, and relatively high temperatures, and does not smell quickly [2]. In 2018, it overtook popularity from soybean oil production with a global share of 31%. Almost 85% of global palm oil productions were supplied by Indonesia and Malaysia [13]. The increasing volume of annual palm oil exports in Indonesia has made this commodity the biggest foreign exchange earner, reaching 23 billion USD in 2017 [7]. It is export volume increased by 19.45%, with export value reaching 20.7 Billion USD in the same year [1]. In addition, FAO predicts that in the coming decade, Indonesia will dominate world vegetable oil exports [15]. Therefore, oil palm has an important role in the country's economy, so an increase in productivity must be done.
Soil is considered as the main source in providing essential plant nutrients and water reserves [9]. In oil palm cultivation, it is one of the most key environmental factors when assessing the suitability of the site for oil palm cultivation [10]. In other words, the quality of soil can be an indicator of the productivity of plantation. The essential component of soil is the topography and shape, moisture availability, physical and chemical properties. Soil chemical characteristics more important consideration in palm oil plantation because they are more easily changed compared to soil physical properties [10]. The fertilization process is one of the soil maintenance activities to prevent chemical soil conditions of the plantation.

That activity can be very expensive. It can spend more than 50% of the cost of plant maintenance [10]. Through this soil analysis, it is expected to determine the level of soil fertility and fertilization activities that can be carried out appropriately (the composition of fertilizer according to soil requirements). Regression analysis can be applied by letting the chemical properties and level of soil fertility as a predictor and a response, respectively. The model will be very effective (minimum error) if there is no correlation between the predictors. Biologically and mathematically, some of the predictors are related. Consequence, the assumptions of the regression model were not fulfilled. The statistical method that can be used to solve this problem is Principal Component Regression (PCR).

The regression coefficient in PCR is the coefficient of the new variables formed called principal components (Pc’s). Dimension reduction is performed on predictor variables that have a strong enough correlation so that the Pc’s formed has no significant correlation. Therefore, the stage for selecting Pc’s is crucial. In the past five years, research on PCR in several fields was carried out. Research in health was conducted by Wang et al. (2017) regarding Nutrition factors and physical activities with diabetes [21], Ming and Lian (2015) in economics [3], Kemalbay and Korkmazoğlu (2013) in housing loan approvals for categorical data. PCR method has also been applied to genomic data by Shen and Zhu (2009) [20]. Application in agronomy was carried out by Chang et al. in 2001 [4]. In this study, the PCR method will be applied to the chemical properties of the soil in West Sulawesi oil palm plantation, Indonesia. Sixteen chemical properties were measured from Thirty-six soil samples on one of the oil palm plantations in West Sulawesi from two different afdeling.

2. Principal Component Regression Method

Let \( X \) be a \( n \times p \) matrix, where \( n \) is the sample size with \( n=36 \), and \( p \) is the chemical properties of soil with \( p=16 \). Mathematically the PCR method consisted of two stages of the process and was illustrated in Figure 1 [17]:

1. \( Z = X P \) from Principal Component Analysis (PCA) process with \( P \) is a coefficient matrix from the previous variable

2. \( \bar{y} = Z \bar{b} \) as a regression model, and \( \bar{b} \) is \( n \times 1 \) parameter matrix. The parameter can be estimated by solving \( \bar{b} = (Z'Z)^{-1}Z'y \).

Through Figure 1, it can be said that the initial dimension of the data possessed is reduced first using PCA, so there are new variable matrix \( n \times k \), \( T \) with \( k < p \). Furthermore, the new variable \( T \) will be a predictor in the regression model.

![Figure 1](image-url)  
**Figure 1.** Illustration of (a) Multiple Linear Regression (MLR) and (b) PCR method. There are \( k \) independent new variables with \( k < p \) in PCR (adopted from Dunn, 2016 page. 337 [5]).
2.1 Stage 1: Principal Component Analysis

Principal Component Analysis (PCA) is a common method to select representative soil characteristics in the agronomy sector. Panishkan et al. (2012) used PCA to classify agricultural areas in Thailand based on soil properties [16]. Gonzalez and Valenzuela (2017) determine soil quality index by PCA based on the physical, chemical, and biological properties of the soil in cocoa agroforestry systems in Colombia [2]. In palm plantation, Sabrina et al. used PCA to determine major factors of the earthworm population [19]. Edokpayi et al. was also analyzing soil properties under oil palm cultivation in Nigeria by some multivariate analysis, such as PCA, cluster analysis, and discriminant analysis [6].

The essence of PCA is to convert the collection of \( p \) variables \( X \) into a system of orthogonal variables. Let \( Z \) are new orthogonal variables, so it can be written by

\[
Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\
Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\
\vdots \\
Z_p = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p
\]

Geometrically, the new variables represent a new coordinate system obtained by the rotation of the initial axis of the system, and it called by principal components (PC's) [14]. The main idea of this method is to reduce the dimension data variable by selecting the PC's that provide maximum variance [15].

Generally, the researchers are conducted data dimension reduction using the Eigen Value Decomposition (EVD) method. Ginanjar and Indratno (2015) have done in correspondence analysis using Singular Value Decomposition (SVD) to overcome the limitations of the EVD method [8]. The process of finding major components with SVD is theoretically faster than EVD. This is made possible by several reasons, including decomposition carried out without calculating the variance-covariance matrix and not sorting the eigenvalues manually. But further research is needed in algebra and statistics to be able to explain why this happens. Furthermore, PCA-SVD was carried out in this study.

Decomposition matrix by SVD can be written by

\[
X = UDV^t
\]

With \( U \) and \( V^t \) are matrixes of orthonormal eigenvector from \( XX^t \) and \( X^tX \), respectively. Matrix \( D \) is a diagonal matrix with \( D = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_p}) \), \( \lambda_1 > \lambda_2 > \ldots > \lambda_p \) and \( \lambda_i, i = 1, 2, \ldots, p \) are eigenvalues. SVD interpretation can be reviewed through the results of matrix decomposition by performing operations on the right side or left side of the equation \( XV = DU \).

| Table 1. The result of the matrix by SVD |
|-----------------|-----------------|
| (a) Left side   | (b) Right side  |
| \( X = UDV^t \) | \( XV = DU \)   |
| \( U^tX = DV^t \) | \( (XV)^t = (DV)^t \) |
| \( U^tX = Z \)  | \( V^tX^t = U^tD \) |
|                 | \( V^tX^t = Z \) |

If the left side is operated, so \( U^t \) is the basis of change from the matrix \( X \) to \( Z \). It is a random vector of rows that constructs the column of \( Z_r \). Column space means notation of possible matrix outputs. Conversely, in Table 1 (b) shows that \( V^t \) is a basis change from \( X^t \) to \( Z_r \) (right side). It is a random vector of column that constructs the row of \( Z_r \). Row space means notation of possible matrix input.

In algebra, the left and right multiplications do not provide meaningful interpretations. However, in the PCA method, we must understand very well which variables will be reduced. The process of reducing the variables to be examined in this soil analysis is the matrix \( U \).
The results of the PCA process on this data are generally expressed in a scree plot with a horizontal axis stating variables (1 to \(p\)) and the vertical axis of 1-100%, which represents the absorption of a new random variable. The percentage of variance for each PC obtained by

\[
\%PC \text{ } k\text{-th} = \frac{\lambda_k}{\sum_{j=1}^{p}\lambda_j}, k < p
\]

The selection of the PCs is based on criterion sufficient quality of representation, i.e., we Retain sufficient components to account for a specified percentage of the total variance, say, 70-80% [18].

2.2 Stage 2: logistic regression

The general form of the logistic regression model for \(k\) predictor variables is

\[
\pi(z) = \frac{\exp(\beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k)}{1 + \exp(\beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k)}
\]

With: \(\pi(z)\) = probability of success (fertile)

\(\beta_i\) = parameter regression, \(i = 1,2,\ldots,k\)

\(z_i\) = predictor (\(i\)-th principle component)

The logit function represented by

\[
g(z) = \ln \left( \frac{\pi(z)}{1 - \pi(z)} \right) = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k
\]

Significant test for the parameters will be applied with Wald test test to determine significant factor to level of soil fertilization. The hypothesis is

\(H_0: \beta_i = 0\)

\(H_1: \beta_i \neq 0\)

The Wald statistics [11] is defined as

\[
W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \sim N(0,1)
\]

with \(\hat{\beta}_i\) and \(SE(\hat{\beta}_i)\) are estimator and standard error from coefficient \(\beta_i\). The null hypothesis will be rejected if \(W > Z_\alpha\) or if \(p\)-value is greater than level of significant (\(p\)-value < \(\alpha\)). It means that the \(z_i\) influences to the soil condition with contribution \(\beta_i\). Oppositely, if it is not rejecting null hypothesis, the \(z_i\) does’nt influence to the level of soil fertilization.

The goodness of regression model can be determined by calculating the percentage of accuration. In this paper, the accuration of the model explained by

\[
\%\text{ accuracy} = \frac{\text{number of fit data}}{n} \times 100\%
\]

3. Data and Result

The data from oil palm plantation in West Sulawesi, Indonesia, was carried out in this study. There are 16 chemical properties (macronutrients) in the soil that will be defined as the random variable as follows (Table 1). Such as pH KCl 1 N, C-Organic, N, P₂O₅, K₂O, P₂O₅ Olsen P₂O₅ Bray, Ca, Mg, K, Na, H⁺, Al³⁺, C/N, Cation Exchange Capacity (CEC) and pH H₂O. The soil measurement is done through systematic sampling with a depth between 0-20 cm for each plantation. There are 36 samples for each property.
Table 2. There are Sixteen chemical properties from the soil of oil palm plantation in West Sulawesi. Its measurement is done through a systematic sampling from 36 observations with a depth between 0-20 cm for each ranch.

| No. | Chemical Property                        | Symbol          | Random variable |
|-----|-----------------------------------------|-----------------|-----------------|
| 1   | Potential acid                          | pH KCl 1 N      | X_1             |
| 2   | Carbon organic                          | C-organic       | X_2             |
| 3   | Nitrogen                                | N               | X_3             |
| 4   | The proportion of Carbon and Nitrogen   | C/N             | X_4             |
| 5   | Phosphorus pentoxide                    | P_2O_5          | X_5             |
| 6   | Potassium oxide                         | K_2O            | X_6             |
| 7   | Availability of Phosphor with Olsen method | P_2O_5 Olsen   | X_7             |
| 8   | Availability of Phosphor with Bray method | P_2O_5 Bray    | X_8             |
| 9   | Calcium                                 | Ca              | X_9             |
| 10  | Magnesium                               | Mg              | X_10            |
| 11  | Potassium                               | K               | X_11            |
| 12  | Natrium                                 | Na              | X_12            |
| 13  | Cation Exchange Capacity                | CEC             | X_13            |
| 14  | Ion hydrogen                            | H^+             | X_14            |
| 15  | Ion Aluminium                           | Al^{3+}         | X_15            |
| 16  | Active acid                             | pH H_2O         | X_16            |

The statistics descriptive (such as mean, median, minimum, maximum, first quartile, third quartile, standard deviation, variance, skewness, and kurtosis) of the data are summarized in Table 3 (a) and shows some information. This table shows that the mean of pH KCl 1 N is lower than mean pH H_2O with a value around one unit. It means the soil in that location is fertile because the value of potential acid is lower than the value of active acid.

The highest mean concentration is P_2O_5 with value 107.717 and followed by K_2O with value 77.067. They are very important for an oil palm tree to grow. Na is the lowest mean concentration, with a value of 0.101. This is because Na can be harmful to the soil if the concentration is too high. C/N is a chemical factor that shows the ratio between C-Organic with Nitrogen. It also can influence the composting process in the soil. The mean of C/N is 15.139, which means soil in good condition because the organic component can be absorbed very well.

There are five chemical properties which have symmetry distribution, there are pH KCl 1 N, C-Organic, N, P_2O_5, and Mg, with the absolute value of skewness is less than 0.75. It means that the properties tend to evenly distributed. Ca and K is the essential nutrient which is needed by plant with a significant proportion. Hence they have negative skewness. Component with the highest variance is P_2O_5, with value 15635.2821. Thus it will become one of the PC in the next analysis. The skewness of pH H_2O is almost zero because of that the distribution of pH H_2O is symmetry. There is three chemical component with negative skewness value that are pH KCl 1 N, Mg, and pH H_2O.
Table 3. Descriptive statistics for the chemical component condition in the soil with 36 samples for each observation. Table (a) is a descriptive statistics before standardization, and Table (b) is a descriptive statistics after standardization.

| Variable      | Mean  | Q1    | Q2    | Q3    | Variance | Skewness | Kurtosis | Maximum | Minimum |
|---------------|-------|-------|-------|-------|----------|----------|----------|---------|---------|
| pH KCl 1 N    | 4.265 | 3.883 | 4.305 | 4.935 | 0.542    | -0.149   | -0.474   | 5.920   | 3.000   |
| C-Organic     | 9.762 | 1.368 | 1.583 | 10.235 | 204.496  | 1.245    | -0.394   | 39.203  | 0.993   |
| N             | 0.416 | 0.146 | 0.162 | 0.479 | 0.206    | 1.539    | 0.975    | 1.562   | 0.130   |
| C/N           | 15.139| 9.000 | 10.000| 14.000| 119.552  | 1.756    | 1.815    | 46.000  | 7.000   |
| P2O5          | 107.717| 41.248| 71.518| 175.986| 6827.076 | 0.959    | -0.161   | 305.313 | 22.586  |
| K2O           | 77.067| 36.280| 63.093| 130.480| 2549.539 | 0.400    | -1.357   | 166.903 | 13.429  |
| P2O5 Olsen    | 15.396| 0.000 | 4.518 | 22.468| 441.329  | 1.318    | 0.829    | 77.004  | 0.000   |
| P2O5 Bray     | 12.712| 0.000 | 6.565 | 19.243| 297.825  | 1.850    | 4.092    | 75.308  | 0.000   |
| Ca            | 24.671| 14.078| 17.335| 37.450| 181.231  | 0.148    | -1.500   | 47.670  | 0.150   |
| Mg            | 2.983 | 2.455 | 2.915 | 3.760 | 0.960    | -0.368   | 0.920    | 4.880   | 0.060   |
| K             | 0.798 | 0.310 | 0.515 | 1.225 | 0.412    | 0.898    | -0.498   | 2.310   | 0.010   |
| Na            | 0.101 | 0.050 | 0.070 | 0.100 | 0.012    | 3.406    | 13.640   | 0.600   | 0.010   |
| CEC           | 39.590| 23.640| 29.810| 49.120| 513.650  | 0.986    | 0.178    | 100.430 | 1.060   |
| H+            | 0.351 | 0.109 | 0.287 | 0.558 | 0.071    | 0.811    | -0.425   | 0.958   | 0.022   |
| Al3+          | 0.599 | 0.000 | 0.000 | 0.219 | 1.889    | 2.488    | 5.611    | 5.743   | 0.000   |
| pH H2O        | 5.422 | 4.870 | 5.390 | 6.160 | 0.522    | -0.095   | -1.439   | 6.460   | 4.230   |

| Variable      | Mean  | Q1    | Q2    | Q3    | Variance | Skewness | Kurtosis | Maximum | Minimum |
|---------------|-------|-------|-------|-------|----------|----------|----------|---------|---------|
| pH KCl 1 N    | 0.000 | -0.520| 0.054 | 0.910 | 1.000    | -0.149   | -0.474   | 2.247   | -1.719  |
| C-Organic     | 0.000 | -0.587| -0.572| 0.033 | 1.000    | 1.245    | -0.394   | 2.059   | -0.613  |
| N             | 0.000 | -0.596| -0.560| 0.138 | 1.000    | 1.539    | 0.975    | 2.526   | -0.629  |
| C/N           | 0.000 | -0.561| -0.470| -0.104| 1.000    | 1.756    | 1.815    | 2.822   | -0.744  |
| P2O5          | 0.000 | -0.804| -0.438| 0.826 | 1.000    | 0.959    | -0.161   | 2.391   | -1.030  |
| K2O           | 0.000 | -0.808| -0.277| 1.058 | 1.000    | 0.400    | -1.357   | 1.779   | -1.260  |
| P2O5 Olsen    | 0.000 | -0.733| -0.518| 0.337 | 1.000    | 1.318    | 0.829    | 2.933   | -0.733  |
| P2O5 Bray     | 0.000 | -0.737| -0.356| 0.378 | 1.000    | 1.850    | 4.092    | 3.627   | -0.737  |
| Ca            | 0.000 | -0.787| -0.545| 0.949 | 1.000    | 0.148    | -1.500   | 1.708   | -1.821  |
| Mg            | 0.000 | -0.539| -0.069| 0.793 | 1.000    | -0.368   | 0.920    | 1.936   | -2.983  |
| K             | 0.000 | -0.760| -0.440| 0.666 | 1.000    | 0.898    | -0.498   | 2.357   | -1.227  |
| Na            | 0.000 | -0.470| -0.284| -0.005| 1.000    | 3.406    | 13.640   | 4.648   | -0.843  |
| CEC           | 0.000 | -0.704| -0.432| 0.420 | 1.000    | 0.986    | 0.178    | 2.684   | -1.700  |
| H+            | 0.000 | -0.905| -0.241| 0.776 | 1.000    | 0.811    | -0.425   | 2.271   | -1.230  |
| Al3+          | 0.000 | -0.435| -0.435| -0.276| 1.000    | 2.488    | 5.611    | 3.743   | -0.435  |
| pH H2O        | 0.000 | -0.764| -0.044| 1.021 | 1.000    | -0.095   | -1.439   | 1.437   | -1.650  |

Because chemical components data do not have the same units, for example, P2O5 (ppm) and CEC (me/100g), we standardize before perform any mathematical analysis. Table 3(b) summarize the statistics descriptive of chemical component data after standardization. It can be seen that the Q2 value of pH KCl 1 N, Mg, and pH H2O ranging from -0.1 to 0.1, but the other components have a value from -0.6 to -0.1. Standardization makes all of the chemical components have mean and variance 0 and 1, respectively.

A scatter diagram is generally drawn and calculate the correlation for each pair of elements to simplify the interpretation. In the case above, we get 98 for each picture and the correlation value. It presented by a matrix with a size of 16 × 16, as shown in Figure 2. We can see several variables that have fairly close linear correlations, such as pH KCl 1 N - pH H2O, Ca-pH levels of 1 N KCl, organic C / N - C, C / N - N, N - CEC, P2O5 - K2O, N-K2O, and Organic C - K2O. Statistically, the measurement
of the linearity of several pairs in Figure 2 can be used Pearson correlation. If the value is close to 1, then the linearity relationship between the two elements (random variable) is equally strong. Conversely, if it approaches -1, the linearity relationship is inversely proportional.

For example, strong correlation values (more than 0.650) are taken, namely 1 K - pH H2O KCl pH levels, 1 N KCl Ca - pH levels, organic C / N - C, C / N - N, and N - CEC. Each correlation value is 0.8763; 0.6717; 0.7106; 0.8404. Mathematically, the pH information of KCl 1 N is also represented by the pH H2O information because it has a correlation value of 0.8763. This can be done as a whole (not pair by pair) by PCA. The result of the PCA process from soil data was shown by Figure 3.

![Figure 2](image1.png)  
**Figure 2.** Scatter diagram for each pair of measured random variables (abiotic) in the soil characteristic data. Look at the pictures in row 9 for Ca and column 1 for pH KCl 1 N, row 4 for C / N and column 2 for organic C, and row 7 for Olsen and column 5 for P2O5. The upper triangle matrix shows the scatter diagram, and the lower triangle shows the correlation value between the measured variables.

![Figure 3](image2.png)  
**Figure 3.** Scree plot (a) for absorbing variances of each new random variables (PCs). The absorption of the first, second, and third new random variables are 40.597%, 25.212%, and 12.589%, respectively. The cumulative percentage of information absorption can be seen on the plot (b).
EC has the biggest influence, i.e., −0.3494, followed by Ca, which is 0.3279. The chemical factor with the smallest influence is H+ by 0.0014. In the PC analysis, the weight of each chemical factor in the first component (Z1) can be written by the third component (Z3):

\[ Z1 = 0.2599 \text{ pH KCl 1N}, \ldots, Z3 = \text{pH H2O} \]

and the coefficient value of −0.4662, then followed by Olsen KCl 1 N pH of 0.0040. Chemical factors with little influence are CEC and proportion C/N of 0.0310 and 0.0427. From the PC3, the b factor value is 0.0606 and 0.0247.

Table 4. There is the variance, absorption of proportion data (prop.), and cumulative of prop (cum) from soil data. If only one major component is selected, the information absorbed is only 44.34% (***). Whereas, the information absorption reaches 68% (**) if two main components are selected, and so on.

| PC | var | prop. (%) | Cum. (%) |
|----|-----|-----------|----------|
| 1  | 7.0945 | 44.3409 | **44.3409** |
| 2  | 3.8091 | 23.8069 | **68.1478** |
| 3  | 1.7313 | 10.8207 | **78.9685** |
| 4  | 0.8922 | 5.5760  | 84.5445  |
| 5  | 0.6550 | 4.0936  | 88.6381  |
| 6  | 0.5085 | 3.1779  | 91.8160  |
| 7  | 0.4171 | 2.6070  | 94.4230  |
| 8  | 0.3301 | 2.0632  | 96.4861  |

Table 5: The weight of each chemical factor in the first component (Z1), the second component (Z2), and the third component (Z3). There are some factors that have big influence for each PC’s (marked by *).

| New variable (Pricipal Component) | Z1   | Z2   | Z3   |
|----------------------------------|------|------|------|
| pH KCl 1 N                       | -0.2045 | -0.4129 | 0.0606 |
| C-Organic                        | -0.2870 | 0.1495 | *0.3960 |
| N                                | -0.2664 | 0.1103 | *0.3518 |
| C/N                              | -0.2599 | 0.0427 | 0.1390 |
| P2O5                             | 0.2621 | -0.1041 | *0.4025 |
| K2O                              | *0.3350 | -0.1187 | 0.2072 |
| P2O5 Olsen                       | 0.1148 | -0.4143 | 0.2082 |
| P2O5 Bray                        | 0.1625 | 0.3185 | 0.1249 |

| New variable (Pricipal Component) | Z1    | Z2    | Z3    |
|----------------------------------|-------|-------|-------|
| Ca                               | *-0.3430 | -0.1180 | -0.0332 |
| Mg                               | 0.2765 | -0.1159 | 0.2891 |
| K                                | *0.3279 | -0.0774 | 0.2256 |
| Na                               | -0.2267 | 0.0955 | 0.1724 |
| CEC                              | *-0.3494 | 0.0310 | 0.1589 |
| H+                               | 0.0014 | 0.3523 | *0.3990 |
| Al3+                             | 0.1655 | 0.3321 | -0.2525 |
| pH H2O                           | -0.1197 | *-0.4662 | -0.0115 |

Let \( X_1 = \text{pH KCl 1N}, \ldots, X_6 = \text{pH H2O} \) dan \( Z_i = PC_i \), \( Z_2 = PC_2 \), dan \( Z_3 = PC_3 \), so the equations of the result can be written by

\[
Z_1 = -0.2045 \text{ pH KCl 1N} - 0.2870 \text{ C-Organic} - 0.2664 \text{ N} - 0.2599 \text{ C/N} + 0.2621 \text{ P2O5} + 0.3350 \text{ K2O} + 0.1148 \text{ P2O5 Olsen} + 0.1625 \text{ P2O5 Bray} - 0.3430 \text{ Ca} + 0.2765 \text{ Mg} + 0.3279 \text{ K} - 0.2267 \text{ Na} - 0.3494 \text{ CEC} + 0.0014 \text{ H}^+ + 0.1655 \text{ Al}^{3+} - 0.1197 \text{ pH H2O}
\]
Those three PCs are the regressor in the next stage, such as logistic regression. The regression model of the PCs is

\[
\hat{y}(z) = \frac{\exp(0.78 + 0.17z_1 + 0.16z_2 + 3.65z_3)}{1 + \exp(0.78 + 0.17z_1 + 0.16z_2 + 3.65z_3)}
\]

4. Discussion

The result of the PCA process also can be explained by biplot. Figure 4 shows the biplot of the first and second PC with 68.14% of cumulative variance. On these two axes, there is a percentage of data represented by the component. The red arrow is the original variable from the data. The arrow projection on each axis shows the influence of the variable on related components. The number at the green dot represents the observation number.

It also shows the tendency of each variable to PC1 (horizontal axis) and PC2 (vertical axis). Variables that tend to be horizontal have a big effect on the PC1. Similarly, variables that tend to be vertical have a major effect on the PC2, as explained in Table 4, CEC and Ca have almost the same effect on PC1. This is also seen through the projection of these two chemical factors, which are almost the same length on the horizontal axis. It is also seen that the pH of H\textsubscript{2}O has the longest projection on the upright axis, which means that the pH of H\textsubscript{2}O has the greatest effect on PC2. Two chemical factors whose axes are piled up like Al\textsuperscript{3+} and P\textsubscript{2}O\textsubscript{5} Bray means they have almost the same effect on this component.

![Figure 4](image-url)

Figure 4. Biplot from the result of the PCA process (PC1 and PC2 with 68.15% of information absorption). The red arrow is the original variable from the data, and the green dots are the data that has been transformed.

The green dots look like there are groups on the left and right sides of the biplot. This indicates that the results of PCA can also be used for further analysis. In this case, the results are seen forming groups. It will be interesting if a cluster analysis is carried out.
The number of observations from 1-18 is grouping on the left side and others on the right side of biplot. The observations on the left group tend to a negative contribution of PC1, such as CEC and Ca by Table 4. Oppositely in the right group, it tends to the positive contribution of PC1, such as K and K2O. According to the raw data, the grouping of observation fits with the location of the observation. It means that the left group comes from the same block of the cultivation.

According to that, for each principal components can be called by soil macronutrients (K2O, Ca, K, CEC) for PC1, soil acidity (pH KCl and pH H2O) for PC2, and soil organic components (C-organic, N) for PC3. These PC’s are predictors in logistic regression model with level of soil fertility as the respond. The logit function for this regression model is represented by

\[ g(z) = 0.78 + 0.17z_1 + 0.16z_2 + 3.65z_3 \]

The signification test was be applied to identify the most influential predictor by hypothesis

\[ H_0 : \beta_i = 0 \]
\[ H_1 : \beta_i \neq 0 \]

The result shows as follows

| Coefficients | Statistics | p-value |
|--------------|------------|---------|
| Intercept    | 0.579      | 0.5629  |
| Z1           | 0.648      | 0.5172  |
| Z2           | 0.225      | 0.8221  |
| Z3           | 1.982      | 0.0475  |

According to table above, only \( Z_3 \) is significant to the response of the regression model (p-value < \( \alpha = 5\% \)). The regression model will rebuild based on it by stepwise backward elimination [22], so \( Z_1 \) and \( Z_2 \) are eliminated in the new model. The result shows that the intercept is not significant with \( \alpha < 10\% \) (p-value = 0.1448). Based on it, the best logistic regression model has been given by

\[ \hat{\pi}(z) = \frac{\exp(2.01z_3)}{1 + \exp(2.01z_3)} \]

In the Table 6, it shows the comparison result between before and after stepwise method in the regression model. The fitted value has been converted to binomial value by some criterion. If the value of fitted data is greater than 0.5 so it will be classified as success (equal to 1). The update model is very influence to percentage of accuracy. It can be increasing the percentage of accuracy to reaching 90%. It shows that the stepwise backward method is worked in this research.

| Indicator of Comparison | Predictors |
|-------------------------|------------|
|                         | Before     | After    |
| Model Equation          | \[ \hat{\pi}(z) = \frac{\exp(0.78 + 0.17z_1 + 0.16z_2 + 3.65z_3)}{1 + \exp(0.78 + 0.17z_1 + 0.16z_2 + 3.65z_3)} \] | \[ \hat{\pi}(z) = \frac{\exp(2.01z_3)}{1 + \exp(2.01z_3)} \] |
| Number of fitted value  | 32         | 33       |
| % Accuracy              | 88.88%     | 91.67%   |

5. Conclusion

Three PCs become predictors of a regression model with information absorption rates reaching 78%. They called by soil macronutrients (K2O, Ca, K, CEC), soil acidity (pH KCl and pH H2O), and soil organic components (C-organic, N). Other than that, the soil organic component is the most properties that influence soil fertilization from the regression model after using the stepwise backward method. It can provide accurate estimation results with an accuracy level reaches 90%. The advantage of this
research is the soil researcher can reduce half of the chemical properties to be analyzed for measuring the level of soil fertility. Furthermore, they can save costs and times for the research as well.

Acknowledgments
This research was supported by P3MI 2019 Institut Teknologi Bandung.

References
[1] Badan Pusat Statistik  2018 Statistik Kelapa Sawit Indonesia 2017
[2] Bank Bumi Daya 1988 Minyak kelapa sawit, suatu tinjauan produksi, pemasaran, dan prospek. (Jakarta: Divisi Penelitian dan Pengembangan)
[3] Chen M M and Ma J L 2015 Application of principal component regression analysis in economic analysis Proc. of the 2015 3rd International Conference on Management Science, Education Technology, Arts, Social Science and Economics (Paris: Atlantis Press)
[4] Chang C W, Laird D, Mausbach M J and Hurburgh Jr. C R  2001 Near-infrared reflectance spectroscopy–principal components regression analysis of soil properties. Soil Sci. Soc. Am. J. 65 480–490
[5] Dunn K 2016. Process Improvement Using Data. Lecture Notes Version 389-3b2c
[6] Edokpayi A A, Osayande P E, Ikuenobe C E, and Aghinien A E  2015 Multivariate Analysis of Soils Under Oil Palm Cultivation in Nigeria Nigerian Journal of Agriculture, Food and Environment 11(4) 97-102
[7] Gapki  2017 https://gapki.id/news/4419/sawit-sembang-devisa-300-triliun-untuk-negeri-ini-apa-maknany
[8] Ginanjar I and Indratno S W 2015 Identify the ability to purchase a home in the five major cities of Java for the Indonesian middle class using correspondence analysis. AIP Conference Proceeding 1692, 020022.
[9] Ghaemi M, Astaraei A R, Emami H, Mahalati M N, Sanaeinejad S H 2014. Determining soil indicators for soil sustainability assessment using principal component analysis of asatan quds-east of mashhad- Iran. Journal of Soil Science and Plant Nutrition 14(4)
[10] Hakim M 2007 Kelapa sawit, Teknis Agronomis dan Manajemennya (Jakarta: Lembaga Pupuk Indonesia)
[11] Hosmer D W and Lemeshow S 2000  Applied Logistics Regression  (New York: John Wiley & Son)
[12] Kemalbay G and Korkmazoğlu O K 2013 Categorical principal component logistic regression: A case study for housing loan approval Procedia-Social and Behavioral Sciences 109 730-736
[13] Ling A H  2019 Global palm oil trade-prospect and outlook. 4th Malaysian Business Forum. Kuala Lumpur, Malaysia
[14] Milewska A J, Jankowska D, Cytko D, Więsak T, Acacio B and Milewski R 2014 The use of principal component analysis and logistic regression in prediction of infertility treatment outcome Studies in Logic Grammar and Rhetoric 39(52) 7-23
[15] OECD-Food Agriculture Organization on United Nations 2017 OECD-FAO Agricultural Outlook 2018-2027: Oilseeds and Oilseeds Product
[16] Panishkhan K, Swangjang K, Sanmanee N and Sunghong D 2012 Principal component analysis for characterization in the application of some soil properties International Journal of Environmental and Ecological Engineering 6(5) 279-281
[17] Parra-Gonzales S D and Rodriguez-Valenzuela J 2017 Determination of the soil quality index by principal componen analysis in cocoa agroforestry systemin the Orinoco region Columbia Journal of Agriculture and Ecologv International 10(3) 1-8
[18] Rencher A C 2012. Methods of Multivariate Analysis (3rd ed.) (New York: John Wiley & Sons, Inc. Publication)
[19] Sabrina D T, Hanafi M M, Azwady A A N, and Mahmud T M M. 2009 Earthworm populations and cast properties in the soil of oil palm plantations *Malaysian Journal of Soil Science* **13** 29-42

[20] Shen Y F and Zhu J 2009 Power analysis of principal component regression in genetic *Association Studies Journal of Zhejiang Universiti SCIENCE B* **10** 721-730

[21] Wang K S, Liu Y, Xie X, Gong S, Xu C and Sha Z 2017 Principal component regression analysis of nutrition factors and physical activities with diabetes *J. Biom. and Biost.* **8**(4), 7 pages

[22] Walpole R E, Myers R H, Myers S L, and Le K 2012 *Probability and statistics for engineers and scientists* (Boston: Pearson Education)