DOI: https://doi.org/10.34123/jurnalasks.v18i1.954

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Dimension Reduction of Socioeconomic Factors in Deforestation Analysis in Indonesia Using Sparse PCA

Mitha Rabiyatul Nufus[1*], Jenike Gracelya Noke[2] , Eusabius Paul Pega[3]

[1]Forest Management Study Program, Department of Forestry, Kupang State Polytechnic of Agriculture, Kupang, Indonesia; [2]Fisheries Agribusiness Study Program, Department of Fisheries and Marine Affairs, Kupang State Polytechnic of Agriculture, Kupang, Indonesia; [3]Horticultural Industrial Technology Study Program, Department of Food Crops and Horticulture, Kupang State Polytechnic of Agriculture, Kupang, Indonesia
**Corresponding Author: E-mail address: mitha.nufus@staff.politanikoe.ac.id*

ARTICLE INFO

Article history:
Received 25 March, 2026
Revised 12 June, 2026
Accepted 25 June, 2026
Published 30 June, 2026

Keywords:
Deforestation; Dimensionality Reduction; Indonesia; Socioeconomic Factors; Sparce Principal Component Analysis; Spatial Analysis

Abstract

Introduction/Main Objectives: Deforestation remains a major environmental challenge in Indonesia under diverse socio-economic conditions. This study applies Sparse Principal Component Analysis (SPCA) to identify the key socio-economic variables associated with deforestation patterns. **Background Problems:** Analyses of deforestation drivers often involve numerous correlated variables, leading to multicollinearity and making interpretation difficult. Therefore, an approach is needed to reduce data dimensionality while retaining the most relevant information. **Novelty:** This study employs SPCA to simultaneously perform dimensionality reduction and variable selection, producing a more interpretable framework for identifying socio-economic factors related to deforestation at the provincial level in Indonesia. **Research Methods:** Provincial-level socio-economic data from Statistics Indonesia were analyzed using SPCA to address multicollinearity and derive interpretable components. Spatial autocorrelation was assessed using Moran's I. **Finding/Results:** SPCA reduced the variables into two interpretable components and identified six key contributing variables while excluding three with limited influence. Moran's I values for the first (0.402) and second (0.258) sparse principal components indicated significant positive spatial clustering of provinces with similar deforestation-related characteristics. **Research Limitations:** The analysis is limited to provincial-level secondary data and may not fully capture local-scale variations or all determinants of deforestation.

1. Introduction

Deforestation is widely recognized as a major global environmental concern due to its significant implications for climate change, biodiversity loss, and ecosystem degradation. Indonesia contains one of the largest tropical forest areas in the world; however, significant forest loss has occurred over recent decades due to agricultural expansion, land-use change, and forest fires. This decline in forest cover contributes to carbon emissions, threatens biodiversity, and alters ecosystem functions, thereby affecting both environmental sustainability and climate change mitigation efforts [1]. Beyond its environmental consequences, deforestation also generates substantial social and economic implications, particularly for communities whose livelihoods depend directly on forest resources. Forests serve not only ecological functions but also provide essential goods and services that sustain local economies and support household well-being. Consequently, the promotion of sustainable forest management has become a

critical policy priority aimed at preserving environmental stability while simultaneously fostering sustainable development, especially in many developing countries. In a regional context, countries endowed with extensive tropical forest areas face significant challenges in balancing economic development with environmental conservation. Economic activities such as agricultural expansion, infrastructure development, and increasing investment frequently drive land-use changes that may accelerate deforestation processes. A growing body of literature indicates that socioeconomic factors including population growth, income levels, and the development of the agricultural sector are closely associated with the dynamics of forest cover change [2]. This finding indicates that deforestation is shaped not solely by ecological factors but also by the evolving social and economic dynamics within a region.

Indonesia, recognized as one of the countries with the largest extent of tropical forests in the world, plays a strategic role in maintaining global ecological balance. Indonesia, recognized as one of the countries with the largest tropical forest resources in the world, plays a strategic role in maintaining global ecological balance. According to the Ministry of Environment and Forestry, Indonesia's forest area was estimated at approximately 120.4 million hectares in 2023, representing a substantial proportion of the country's land area and distributed across major regions such as Kalimantan, Sumatra, and Papua. These forest ecosystems perform essential ecological functions, including acting as significant carbon sinks, regulating climate systems, and providing habitats for numerous endemic species with high ecological value [3]. However, pressure on forest areas in Indonesia continues to intensify alongside population growth, economic expansion, and the increasing demand for land to support various development activities.

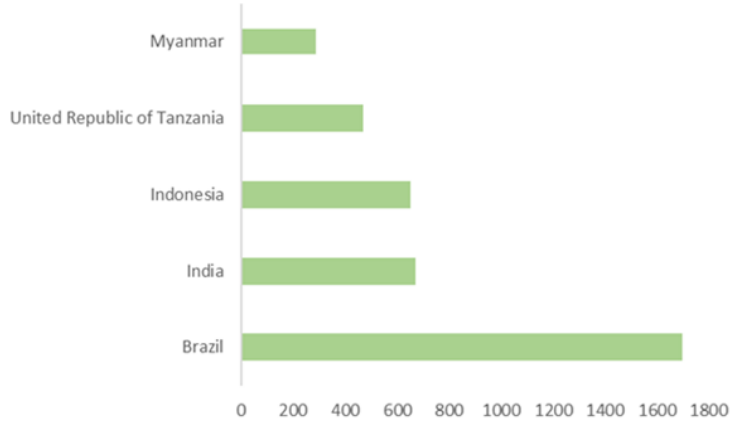


Figure 1. Deforestation rate

Based on the graphical trend, Indonesia experienced an average deforestation rate of approximately 650 thousand hectares per year (650 kha/year) during the 2015–2022 period [4]. This figure positions Indonesia among countries experiencing relatively high levels of forest loss compared with several other nations possessing extensive tropical forest areas. The elevated rate of deforestation in Indonesia is generally associated with multiple interacting drivers, including the expansion of agricultural and plantation land, population growth, and increasing development activities that require the conversion of forest areas into other land uses. These circumstances highlight the critical importance of implementing sustainable forest management practices and strengthening land-use control policies to mitigate the rate of forest cover loss in Indonesia while ensuring the long-term sustainability of forest ecosystems.

The phenomenon of deforestation in Indonesia does not occur in a linear pattern but rather exhibits fluctuations over time. Statistical records indicate that the national rate of deforestation has experienced notable variations in recent years, influenced by a combination of environmental, economic, and social factors. Socioeconomic variables such as population growth, poverty levels, and economic development are frequently associated with increasing pressure on forest resources. Population growth, for instance, tends to intensify the demand for land to support settlements, agricultural expansion, and various economic activities, thereby increasing the likelihood of forest area conversion [3]. A similar observation was reported by Njoya et al. (2024), who indicated that population growth accompanied by the expansion of economic activities tends to intensify the demand for land resources, thereby accelerating land-use change.

In empirical studies, the analysis of the drivers of deforestation frequently involves numerous explanatory variables that exhibit substantial intercorrelation. Socio-economic indicators such as

population size, the Human Development Index, poverty rate, and other development-related measures often display complex relationships with one another. This condition can lead to multicollinearity problems and make it difficult to determine which variables exert the most significant influence on deforestation dynamics. To address this issue, a statistical approach capable of reducing the dimensionality of the data while retaining essential information is required. One commonly applied technique for this purpose is Principal Component Analysis (PCA), which reduces the number of variables by transforming the original variables into a smaller set of uncorrelated components. Through this transformation, PCA captures most of the variability contained in the dataset using a limited number of principal components. Despite its advantages, conventional PCA has limitations in terms of interpretability. Each principal component is typically formed as a linear combination of nearly all original variables, which complicates the interpretation of the resulting components. To overcome this limitation, the Sparse Principal Component Analysis (Sparse PCA) method was developed. Sparse PCA introduces sparsity constraints that allow principal components to be constructed using only a subset of variables, thereby producing components with fewer non-zero loadings. As a result, the derived components become easier to interpret and provide more informative insights for identifying the key socio-economic factors associated with deforestation. Through this approach, a set of principal components is expected to be identified that can represent the underlying structure of relationships among socio-economic variables in a more concise and informative manner. By reducing the dimensionality of the data, the analysis is anticipated to capture the essential patterns of association among variables while minimizing redundancy. Numerous studies have examined the socioeconomic factors associated with deforestation using conventional statistical methods and Principal Component Analysis (PCA). Although PCA is effective in reducing data dimensionality and addressing multicollinearity, the resulting components are often difficult to interpret because most variables contribute to each component simultaneously. As a result, identifying the key socioeconomic factors underlying deforestation remains challenging. Furthermore, the application of Sparse Principal Component Analysis (SPCA), which produces a more parsimonious and interpretable component structure, has received limited attention in studies of deforestation in Indonesia. Therefore, this study applies SPCA to reduce the dimensionality of socioeconomic variables related to deforestation in Indonesia and to identify the most influential factors through a more interpretable component structure. The findings are expected to provide a clearer understanding of the socioeconomic dimensions associated with deforestation and support the development of more targeted forest management policies.

2. Material and Methods

2.1. Data Sources

This study utilizes secondary data at the provincial level covering all provinces in Indonesia. The data were obtained from BPS–Statistics Indonesia. The dataset consists of deforestation-related and socioeconomic indicators, including the Human Development Index, population density, population growth rate, total population, poverty rate, Gross Regional Domestic Product (GRDP) per capita, mean years of schooling, agricultural land area, and the number of forest fires. The analysis includes all provinces in Indonesia. The dataset represents provincial-level indicators in Indonesia that reflect social, economic, and environmental conditions associated with the dynamics of forest cover change. The use of data from a national statistical agency is essential because such datasets are collected through standardized procedures and rigorous verification processes, ensuring a high level of reliability for empirical analysis. Furthermore, regional statistical data are widely employed in environmental research to examine the relationship between socioeconomic factors and land-use change, particularly in studies addressing deforestation and environmental degradation [5]. The variables employed in this study consist of one dependent variable and several independent variables representing socio-economic indicators and pressures on forest resources. These variables were selected because, from a conceptual standpoint, they are frequently utilized in studies examining the relationship between socio-economic dynamics and changes in forest cover. Such indicators are particularly relevant in the context of developing countries, where development activities and socio-economic transformation often exert significant pressure on forested landscapes [5]. The inclusion of multiple socio-economic indicators in the analysis may lead to a high degree of intercorrelation among variables. To address this issue, the present study employs a dimensionality reduction approach using the Sparse Principal Component Analysis (Sparse PCA) method. This technique enables the transformation of a set of correlated variables into a smaller number of more interpretable principal components while enforcing sparsity in the component loadings. Consequently, the method facilitates a clearer identification of the key socio-economic factors associated with deforestation in Indonesia.

Table 1. Research variables

Variable	Description	Unit	Source	Year
Y	Deforestation Rate	Hectares per Year	Statistics Indonesia (BPS)	2023
X1	Human Development Index	Index (0-100)	Statistics Indonesia (BPS)	2023
X2	Population Density	People/km²	Statistics Indonesia (BPS)	2023
X3	Population Growth Rate	Percent per Year	Statistics Indonesia (BPS)	2023
X4	Total Population	Thousand People	Statistics Indonesia (BPS)	2023
X5	Poverty Rate	Percentage	Statistics Indonesia (BPS)	2023
X6	Gross Regional Domestic Product (GRDP) per Capita	Million Rupiah/Person/Year	Statistics Indonesia (BPS)	2023
X7	Mean Years of Schooling	Years	Statistics Indonesia (BPS)	2023
X8	Agricultural Land Area	Thousand Hectares	Statistics Indonesia (BPS)	2023
X9	Number of Forest Fires	Number of incidents	Statistics Indonesia (BPS)	2023

The study considered several socioeconomic indicators as predictor variables (X), including the Human Development Index, population density, poverty rate, GRDP per capita, educational attainment, agricultural land area, and forest fire incidence. These variables were included in the Sparse Principal Component Analysis to identify the underlying socioeconomic dimensions associated with deforestation. The deforestation rate (Y) was included only for descriptive and interpretative purposes. It was not used in the construction of the sparse principal components, as SPCA is an unsupervised technique that operates solely on the predictor variables. Instead, the deforestation rate was presented to provide context and facilitate the interpretation of the extracted socioeconomic patterns.

2.2 Correlation

Correlation is a statistical measure employed to quantify both the strength and direction of a linear association between two variables. Within the context of multivariate analysis, it serves as an initial step for examining the underlying relationships among variables prior to conducting more advanced modelling procedures. Among the available measures, the Pearson correlation coefficient is the most widely utilized for assessing linear dependence. Mathematically, the Pearson correlation coefficient is expressed as follows,

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left(n\sum X^2 - \left(\sum X\right)^2\right)\left(n\sum Y^2 - \left(\sum Y\right)^2\right)}} \quad (1)$$

where r denotes the correlation coefficient, n represents the number of observations, X and Y correspond to the variables under analysis. The correlation coefficient ranges between $-1 \leq r \leq 1$. A positive value ($r > 0$) indicates a direct relationship between variables, whereas a negative value ($r < 0$) reflects an inverse relationship. Furthermore, as the absolute value of the coefficient ($|r|$) approaches 1, the strength of the association becomes increasingly strong. In the context of multivariate analysis, strong correlations among independent variables may serve as an initial indication of multicollinearity. Several studies suggest that absolute correlation coefficients exceeding approximately 0.7–0.8 reflect a substantial linear association, which may imply overlapping information and potential redundancy among variables [6]. Correlation analysis is often complemented by graphical approaches, such as scatterplot matrices, to facilitate a more comprehensive examination of both linear and nonlinear relationships among variables.

2.3 Multicollinearity

Multicollinearity refers to a situation in which two or more independent variables in a model exhibit strong linear associations. This phenomenon is frequently encountered in social, economic, and environmental datasets, where indicators are often inherently interrelated. From a conceptual standpoint, multicollinearity inflates the variance of regression parameter estimators, leading to unstable coefficient estimates and reduced interpretability. As a consequence, the overall reliability of the model may decline, particularly in terms of statistical inference and the assessment of variable significance [7]. One of the most widely applied approaches for diagnosing multicollinearity is the *Variance Inflation Factor* (VIF). This metric quantifies the extent to which the variance of an estimated regression coefficient is amplified due to linear dependencies among the independent variables. In other words, VIF reflects how strongly a predictor is explained by the remaining predictors in the model.

Mathematically, VIF is expressed as:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2}$$

where R_i^2 represents the coefficient of determination obtained by regressing the *i*-th independent variable on the remaining independent variables in the model. The interpretation of Variance Inflation Factor (VIF) values can be outlined as; a VIF equal to 1 indicates the absence of correlation among explanatory variables. Values ranging between 1 and 5 suggest a moderate degree of association that is generally acceptable within regression models. When the VIF exceeds 5, it signals the presence of potential multicollinearity, while values greater than 10 are commonly regarded as evidence of severe multicollinearity, which may substantially affect the stability and reliability of parameter estimates.

A high VIF value indicates that a given variable is substantially explained by the remaining predictors in the model, leading to an inflation of the variance of its estimated regression coefficient. From a mathematical standpoint, this increase in variance directly affects the standard error, which can be expressed as follows,

$$SE\left(\beta_i\right) = SE_{OLS(\beta_i)}\sqrt{VIF_i} \tag{3}$$

Accordingly, higher VIF values are associated with increased uncertainty in parameter estimation, reflecting greater instability in the estimated coefficients.

2.4 Sparse PCA

Principal Component Analysis (PCA) is a widely employed multivariate statistical technique designed to reduce the dimensionality of high-dimensional datasets. The method operates by transforming a set of potentially correlated variables into a smaller number of new variables, known as principal components, which are mutually orthogonal and capture the largest possible proportion of variance in the original data. Through this transformation, PCA enables the extraction of the most informative structure of the dataset while minimizing redundancy among variables. Nevertheless, a notable limitation of classical PCA lies in the fact that each principal component is typically expressed as a linear combination of all original variables. Consequently, the resulting components may involve contributions from many variables simultaneously, which often complicates the interpretability of the components in empirical applications [8]. To address these limitations, the Sparse Principal Component Analysis method was developed as an extension of Principal Component Analysis. This approach introduces the concept of *sparsity* in the coefficient vectors of the principal components. In practice, Sparse PCA produces principal components that involve only a limited subset of variables with non-zero loadings, while the remaining variables are assigned coefficients equal to zero. Consequently, this technique not only performs dimensionality reduction but also implicitly carries out variable selection. Such a property improves the interpretability of the resulting components and facilitates a clearer understanding of the relationships among variables within the analyzed dataset [9].

Conceptually, Sparse Principal Component Analysis (Sparse PCA) aims to derive principal components that not only explain the variability of the data, as in classical Principal Component Analysis (PCA), but also exhibit a sparse coefficient structure. This approach introduces additional constraints or penalty terms on the component loadings, encouraging many coefficients to shrink toward zero. As a result, the resulting components are constructed from only a limited subset of variables that contribute most strongly to the underlying data structure. This sparsity property enhances interpretability by allowing the principal components to be associated with a smaller [10]. The Sparse Principal Component

Analysis (Sparse PCA) approach has been widely applied in the analysis of high-dimensional datasets, including genomic, economic, and socio-economic data, due to its capability to simplify complex variable structures while preserving the essential information contained in the data. By imposing sparsity constraints on the component loadings, Sparse PCA produces principal components that involve only a limited subset of variables, thereby enhancing interpretability. In addition to functioning as a dimensionality reduction technique, Sparse PCA can also serve as a variable selection method in multivariate analysis, as it effectively identifies the most relevant variables that contribute to the underlying data structure [9].

Consider a data matrix $X \in R^{(}n \times p)$where (n) represents the number of observations and (p) denotes the number of variables. In the classical Principal Component Analysis (PCA) framework, the first principal component is obtained by maximizing the variance of the projected data, which can be formulated as follows,

$$\max_{\omega} \omega^T \sum \omega \tag{4}$$

With the constraint $\|\omega\|_2 = 1$, where ω denotes the loading vector of the principal component and $\sum$ represents the covariance matrix of the observed data. In the framework of Sparse Principal Component Analysis (Sparse PCA), this formulation is modified by incorporating a sparsity constraint on the loading vector. The objective is to produce principal components that not only retain the ability to explain the variability of the data, as in classical PCA, but also contain a limited number of non-zero loadings to enhance interpretability. One commonly adopted formulation is presented in Equation (1). In addition to the constraint $\|\omega\|_2 = 1$, a sparsity restriction is imposed in the form $\|\omega\|_1 \leq c$. Alternatively, the same constraint can be expressed through a penalty function approach.

$$\max_{\omega} \left(\omega^T \sum \omega - \lambda \|\omega\|_1 \right) \tag{5}$$

where,

ω : loading vector of the principal component
$\sum$: covariance matrix of the data
$\|\omega\|_1$: L_1norm controlling the level of sparsity
λ : penalty parameter that determines the degree of sparsity

The incorporation of an L_1 penalty encourages sparsity in the loading vectors by shrinking a number of coefficients exactly to zero. As a consequence, only a subset of variables contributes to the construction of the principal components. Through this mechanism, Sparse Principal Component Analysis (Sparse PCA) serves not only as a dimensionality reduction technique but also as an effective variable selection approach within multivariate analysis [10]. The approach proposed by Hui Zou reformulates principal component analysis (PCA) as a penalized regression problem. Within this framework, the objective of Sparse PCA is not only to capture the maximum variance in the data as in classical Principal Component Analysis but also to impose sparsity on the loading vectors through appropriate penalty terms. Consequently, the objective function of Sparse PCA can be expressed as follows [11].

$$\min_{A,B} \left\| \mathbf{X} - \mathbf{XBA}^T \right\|^2 + \lambda_1 \sum_j \left| b_j \right| + \lambda_2 \sum_j b_j^2 \tag{6}$$

With the constraint applied $A^T A = I$.

Where,

X : the standardized data matrix
A : the matrix representing the principal component scores
B : the matrix of principal component loadings
λ_1 : the L_1 regularization parameter (lasso) that promotes sparsity in the model
λ_2 : the L_2 regularization parameter (ridge) that ensures the stability of parameter estimation

The combination of L_1 and L_2 penalties is commonly referred to as the elastic net penalty. The L_1 penalty promotes sparsity by shrinking some coefficients exactly to zero, thereby producing a simpler and more interpretable loading structure. In contrast, the L_2 penalty contributes to stabilizing the

estimation process by mitigating the effects of multicollinearity among variables. Consequently, the elastic net framework enables the extraction of principal components that are both parsimonious and robust in the presence of correlated predictors.

2.5 Step of Analysis

The analytical procedure in this study was carried out through a series of systematic stages to ensure the reliability and interpretability of the results. The analysis began with the collection of secondary data representing socio-economic factors associated with deforestation across Indonesia. The dataset was subsequently cleaned and standardized to improve data quality and ensure comparability among variables. Descriptive statistical analysis and a correlation matrix were then employed to provide an initial overview of the data distribution and to explore the relationships among the socio-economic variables. To identify potential multicollinearity, the Variance Inflation Factor (VIF) was calculated for all explanatory variables. The suitability of the dataset for dimension reduction was further evaluated using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity.

Following these preliminary assessments, all variables were standardized prior to dimensionality reduction. Principal Component Analysis (PCA) was first performed to investigate the latent structure of the socio-economic variables and to summarize the original information into a smaller number of components. The appropriate number of principal components was determined by considering eigenvalues, the cumulative proportion of explained variance, and the scree plot. To improve component interpretability while retaining the most relevant variables, Sparse Principal Component Analysis (SPCA) was subsequently applied. The resulting loading matrix was examined to identify the socio-economic variables that contributed most strongly to each sparse component. Component scores and loading patterns were then visualized to facilitate the interpretation of observation clusters and variable contributions. Finally, the spatial distribution of the SPCA scores was mapped and analyzed to reveal regional differences in socio-economic characteristics related to deforestation across Indonesia. The overall findings were subsequently synthesized to draw conclusions and discuss their implications. All statistical analyses and graphical visualizations were performed using R software (version 4.5.3) with the support of several packages, including stats, car, psych, FactoMineR, factoextra, and elasticnet.

3. Results and Discussion

3.1 Descriptive Statistics

The distribution of provinces with the highest values for each research variable related to socio-economic conditions and pressures on forest resources is presented in Figure 2. In general, the distribution of maximum values across several indicators appears to be concentrated in specific provinces, indicating substantial disparities in development characteristics among regions in Indonesia. The highest values for the Human Development Index (X1), population density (X2), gross regional domestic product (GRDP) per capita (X6), and average years of schooling (X7) are observed in Special Capital Region of Jakarta. This pattern reflects the relatively advanced level of socio-economic development in the country's principal metropolitan area. In contrast, variables related to land-use pressure, such as agricultural land area (X8) and the number of forest fires (X9), reach their maximum values in provinces including Riau and Central Kalimantan, regions widely recognized for intensive land expansion activities and higher susceptibility to forest fire events. Previous research highlights that agricultural expansion and fire events are among the major drivers of forest loss in tropical regions, particularly in Southeast Asia [12]. Furthermore, the largest population size (X4) is recorded in West Java, highlighting the considerable demographic pressure experienced in this province. These patterns collectively suggest that regional development dynamics, demographic intensity, and land-use transformation play a crucial role in shaping environmental pressures and the potential risk of deforestation across Indonesia.

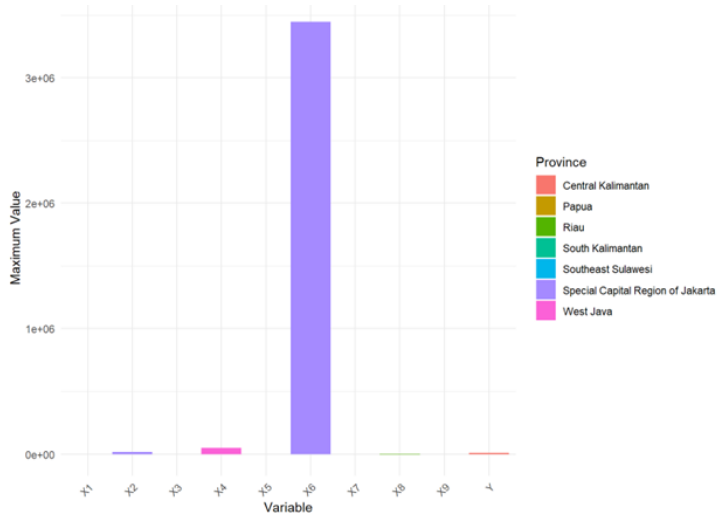


Figure 2. Descriptive statistics of each variable

3.2 Correlation

The correlation matrix and scatterplot visualization reveal varying degrees of association among variables (X1–X9), including both positive and negative relationships. Several variable pairs exhibit relatively strong and statistically significant correlations, indicating the presence of underlying structural relationships within the dataset. Notably, X1 shows a strong positive association with X7 and negative correlations with X3 and X5, while X4 and X6 demonstrate a strong linear relationship. The presence of relatively high correlation coefficients suggests potential multicollinearity among predictors, which may affect the stability and interpretability of regression-based models. Previous studies have highlighted that multicollinearity can distort parameter estimation and reduce model reliability [13]. Therefore, the application of Sparse PCA is justified, as it enables dimensionality reduction while maintaining interpretability through sparse loading structures.

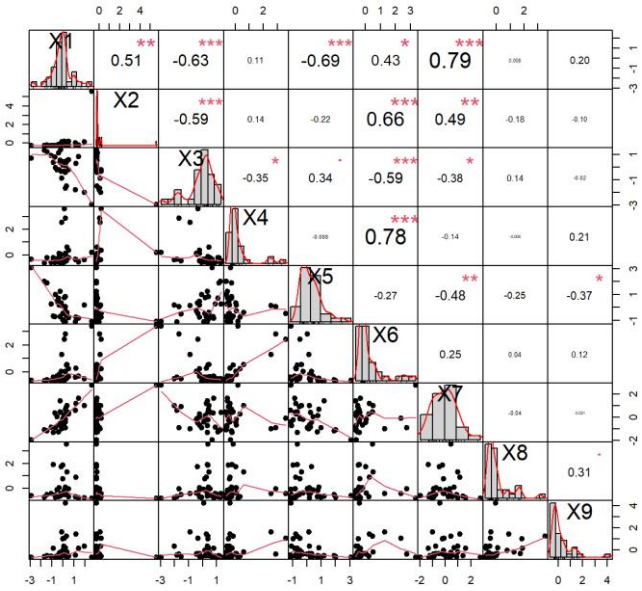


Figure 3. Matrix correlation of each variables

3.3 Multicollinearity

The Variance Inflation Factor (VIF) values for each variable, along with their corresponding visual representation, are presented as follows.

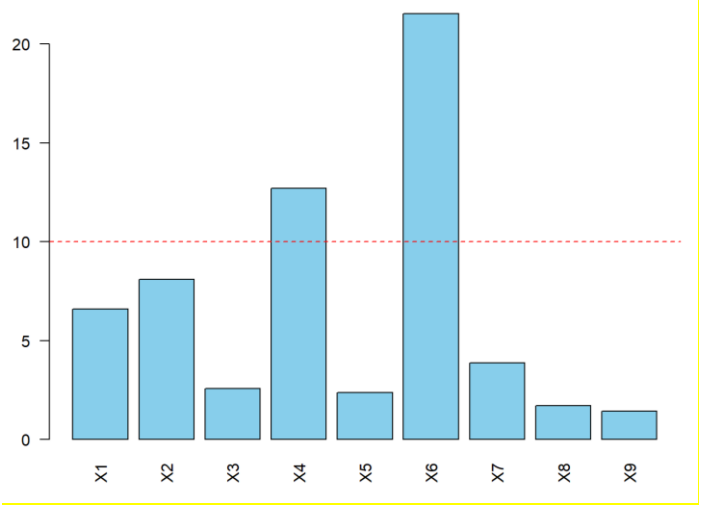


Figure 4. The visualization of VIF

The Variance Inflation Factor (VIF) analysis indicates that most variables exhibit relatively low VIF values, suggesting an acceptable level of multicollinearity. However, variables X4 and especially X6 display substantially high VIF values, exceeding common threshold levels. This suggests that these variables are highly linearly dependent on other predictors in the model. High VIF values indicate inflated variance in regression coefficients, which may lead to unstable parameter estimates and reduced interpretability. Although no universal cutoff exists, values between 5 and 10 are commonly used as indicators of potential multicollinearity [14]. Therefore, the observed values for X4 and X6 confirm the presence of strong multicollinearity. These findings justify the application of dimensionality reduction techniques such as Sparse PCA to address redundancy and improve model robustness.

3.4 Scree Plot of Sparse PCA

The scree plot is a graphical technique commonly employed in Principal Component Analysis to determine the appropriate number of principal components to retain during dimensionality reduction. This plot presents the eigenvalues, or the proportion of variance explained by each component, arranged from the largest to the smallest. Through this visualization, researchers can identify the point at which the contribution of additional components begins to decline substantially, often referred to as the *elbow*. Components appearing before this point are generally considered sufficient to capture the most relevant information contained in the dataset, while subsequent components contribute relatively little additional explanatory power.

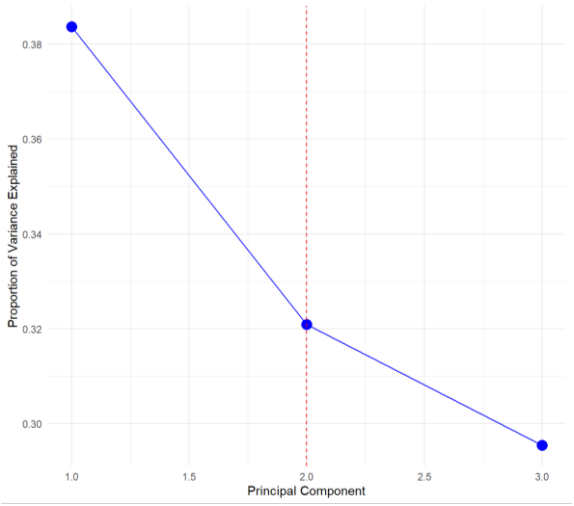


Figure 5. Scree plot sparse PCA

Consequently, the scree plot provides a clear visual basis for selecting an optimal number of components, enabling analysts to preserve essential information while reducing the complexity of multivariate data structures. For this reason, it is widely applied in statistical studies and

multidimensional data analysis as a practical tool for guiding component selection in PCA-based dimensionality reduction.

Based on the scree plot and component selection criteria, two components were retained for further interpretation. The first and second components contributed approximately 38.4% and 32.1% of the retained explained variance, respectively, indicating that these components captured the dominant structure of the socio-economic data.

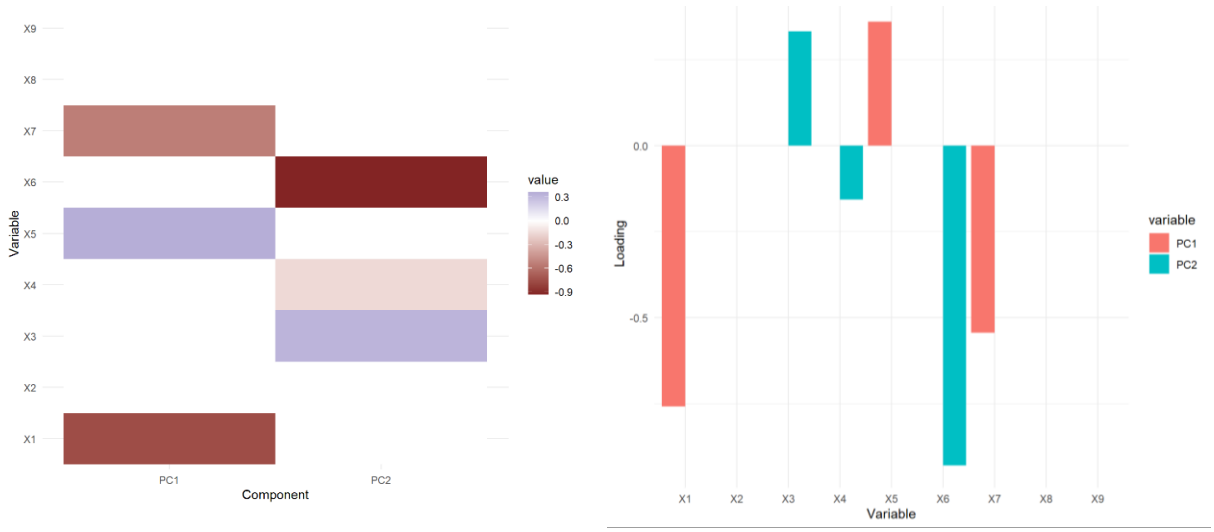
3.5 Sparse Loading

In Sparse Principal Component Analysis (Sparse PCA), the loading coefficients indicate the magnitude of each original variable's contribution to the derived principal components. A large loading value suggests that the variable is strongly associated with the corresponding component, whereas coefficients close to zero imply a relatively minor contribution. By introducing sparsity in the loading structure, Sparse PCA enhances the interpretability of the resulting components because only a limited number of variables retain substantial weights. In addition to improving interpretability, sparse loading reduces model complexity when dealing with high-dimensional datasets. Within multivariate analysis, Sparse PCA is widely applied as it preserves a substantial portion of the data variability while simultaneously identifying the most influential variables that shape the principal components [15].

Table 2. Sparse loading

Variable / PC	PC1	PC2
X1	-0.758	0.000
X2	0.000	0.000
X3	0.000	0.332
X4	0.000	-0.157
X5	0.360	0.000
X6	0.000	-0.930
X7	-0.544	0.000
X8	0.000	0.000
X9	0.000	0.000

In the first principal component (PC1), only three variables exhibit substantial loadings, namely X1 (−0.758), X5 (0.360), and X7 (−0.544). This indicates that PC1 is primarily characterized by the combined influence of these three variables, although their directions of association differ. Variables X1 and X7 are negatively associated with the first component, whereas X5 shows a positive relationship. In contrast, the remaining variables (X2, X3, X4, X6, X8, and X9) display zero loadings, indicating that they do not contribute to the formation of this component. This outcome reflects the sparsity constraint imposed by the Sparse PCA approach, which effectively eliminates variables with negligible contributions.



(a) (b)

Figure 6. (a) Heatmap loading sparse PCA (b) Visualization of sparse PCA loading

For the second principal component (PC2), the component structure is determined by three variables: X3 (0.332), X4 (−0.157), and X6 (−0.930). Among these, X6 provides the most dominant contribution to PC2, as reflected by its relatively large loading value. The other variables have zero loadings and therefore do not participate in forming the second component.

The resulting sparse loading pattern demonstrates that SPCA not only performs dimensionality reduction but also effectively highlights the most relevant variables, thereby improving the interpretability of the principal components. Such an approach is particularly advantageous in the analysis of high-dimensional multivariate data, as it enhances model interpretability while preserving essential information contained in the dataset. Overall, the results indicate that six variables (X1, X3, X4, X5, X6, and X7) make substantial contributions to the extracted sparse principal components. In contrast, variables X2 (Population Density), X8 (Agricultural Land Area), and X9 (Number of Forest Fires) exhibit negligible or zero loadings across the retained components. Consequently, these variables do not contribute significantly to the component structure identified by SPCA and were not retained as key variables in the interpretation of the resulting dimensions. This finding suggests that, within the analyzed dataset, their variability is not strongly aligned with the dominant socioeconomic patterns associated with deforestation captured by retained components. The exclusion of X2, X8, and X9 was based on their near-zero loading values in all retained sparse principal components. These findings demonstrate that Sparse PCA not only reduces the dimensionality of the dataset but also performs implicit variable selection, resulting in a more interpretable and parsimonious component structure compared with conventional PCA [8].

3.6 Score Plot of Sparse PCA

The score plot in Sparse Principal Component Analysis (Sparse PCA) is a graphical representation of two or more principal components that illustrates the position of each observation (sample data) within the component space obtained from the dimensionality reduction process. Typically, the plot is constructed using the first sparse principal component (SPC1) and the second sparse principal component (SPC2). Through this visualization, patterns, relationships, and potential groupings among observations can be more clearly identified after the original high-dimensional data have been projected into a lower-dimensional space. This representation facilitates the interpretation of data structure while preserving the most relevant variation captured by the sparse components [16].

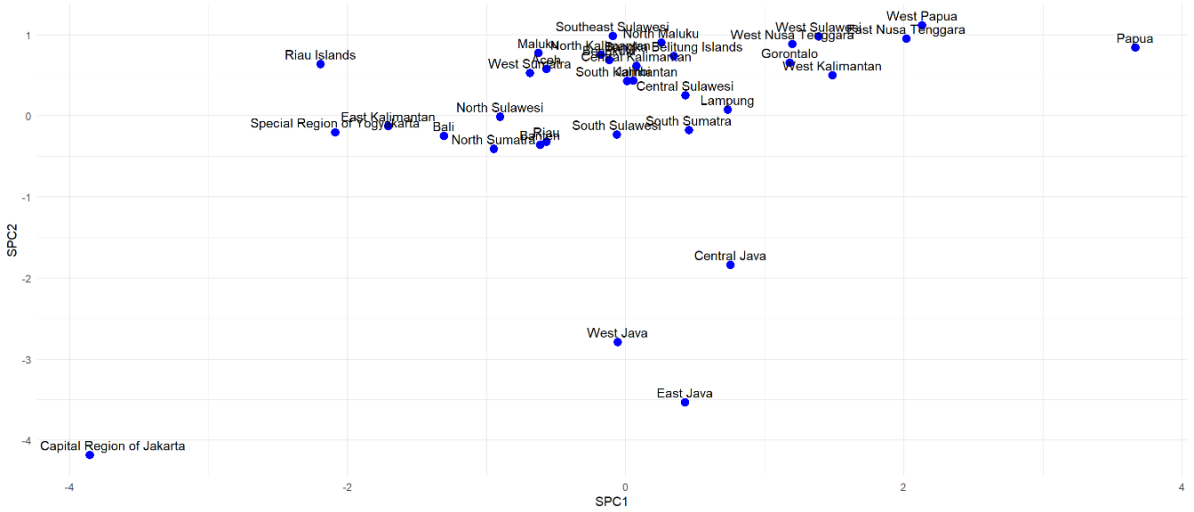


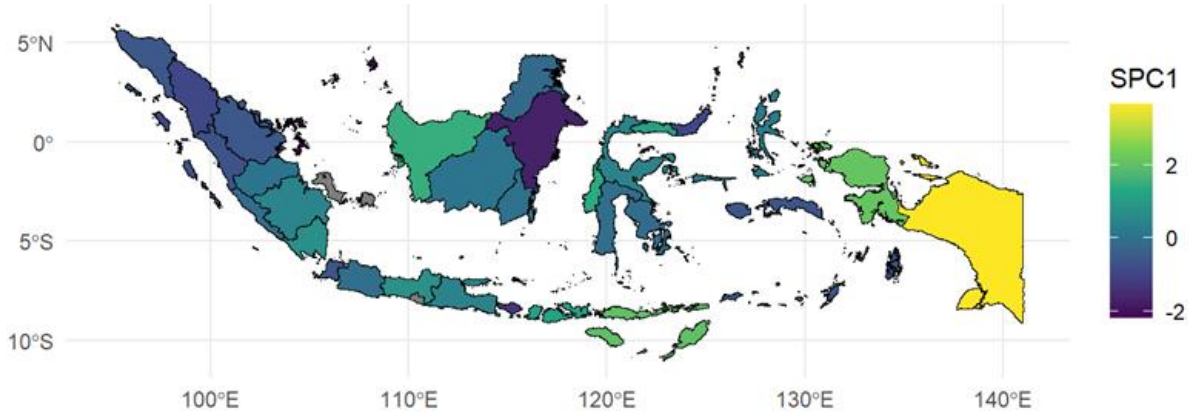
Figure 7. Sparse PCA score plot

In general, most provinces are located near the center of the coordinate system, indicating that their socio-economic characteristics are relatively similar and follow the overall national pattern. These provinces form the main cluster that represents the average condition of the analyzed variables. In contrast, several provinces appear farther from the center, indicating distinct socio-economic profiles. For example, Capital Region of Jakarta is positioned in the negative direction of both SPC1 and SPC2, reflecting a socio-economic structure that differs markedly from most other provinces due to its role as the national economic and administrative center. Some provinces on Java Island, including East Java, West Java, and Central Java, also appear slightly separated from the main cluster, particularly along the SPC2 dimension, indicating certain variations in their socio-economic indicators. Meanwhile, eastern

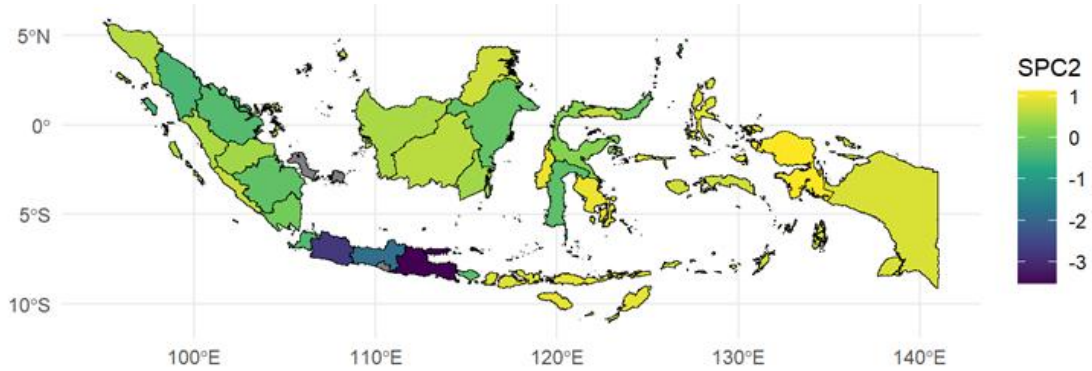
provinces such as Papua and West Papua tend to lie on the positive side of SPC1, suggesting different variable patterns compared to most provinces. Overall, the plot demonstrates that Sparse PCA effectively reduces data dimensionality while clearly illustrating similarities and differences in socio-economic characteristics among provinces. Overall, the score plot demonstrates that Sparse PCA effectively reduces data dimensionality while preserving the underlying variation among observations, thereby facilitating the visualization of similarities and differences in socio-economic characteristics among provinces [17].

3.7 Spatial Distribution of Sparse PCA

The spatial distribution in Sparse PCA refers to the mapping of principal component scores derived from Sparse Principal Component Analysis onto a geographic space. Through this approach, the dominant patterns of variation among multiple variables can be represented spatially. Such visualization facilitates the identification of regions that exhibit similar or contrasting characteristics based on the most influential combination of variables captured by the sparse components. This spatial representation enhances the interpretability of multivariate relationships and provides a clearer understanding of how underlying patterns vary across different geographic areas.



(a)



(b)

Figure 8. Spatial Distribution of (a) Sparse PCA Component 1 (b) Sparse PCA Component 2

The spatial distribution of SPC1 indicates that several provinces in eastern Indonesia, particularly Papua, exhibit relatively high component scores. In contrast, a number of provinces located in Kalimantan and parts of Sumatra show comparatively lower values. This spatial pattern suggests the presence of regional heterogeneity, which is primarily influenced by the combined effect of variables X1, X5, and X7, identified as the dominant contributors to the first sparse principal component. Such regional heterogeneity is consistent with previous studies demonstrating that socio-economic characteristics and development patterns vary considerably among Indonesian provinces, resulting in distinct spatial configurations of environmental and land-use dynamics [18]. The spatial map of SPC2

reveals a different distributional pattern. Several provinces across Sumatra, Kalimantan, and parts of eastern Indonesia display relatively high scores, while provinces on Java tend to exhibit lower values. This indicates that the second component captures an alternative dimension of variation in the dataset, driven by a distinct set of variables compared with those influencing the first component.

To verify whether the spatial patterns observed in Figure 8 reflect statistically meaningful geographic structures, Global Moran's I statistics were calculated for the Sparse PCA scores. The results, presented in Table 3, indicate significant positive spatial autocorrelation for both SPC1 and SPC2, confirming that provinces with similar component scores tend to be spatially clustered rather than randomly distributed. The spatial map of SPC2 reveals a different distributional pattern. Several provinces across Sumatra, Kalimantan, and parts of eastern Indonesia display relatively high scores, while provinces on Java tend to exhibit lower values. This indicates that the second component captures an alternative dimension of variation in the dataset, driven by a distinct set of variables compared with those influencing the first component. Differences in regional socio-economic structures and land-use characteristics have also been reported in previous spatial analyses of Indonesia, highlighting that provinces may exhibit contrasting spatial patterns depending on the underlying socio-economic drivers [19].

Table 3. Global Moran's I statistics for sparse principal components

Component	Moran’s I	p-value
SPC 1	0.40229380	3.025070e-06
SPC 2	0.25835357	1.912515e-03

To statistically assess the spatial patterns identified by Sparse PCA, Global Moran's I statistics were computed for the first two sparse principal components. As presented in Table 3, both components exhibited significant positive spatial autocorrelation. SPC1 yielded a Moran's I value of 0.402 (p = 3.025070e-06), indicating a moderate tendency for provinces with similar SPC1 scores to be geographically clustered. Likewise, SPC2 produced a Moran's I value of 0.258 (p = 1.912515e-03), suggesting a positive spatial association, although weaker than that observed for SPC1. These findings indicate that the spatial patterns displayed in Figure 8 are not randomly distributed across Indonesia. Provinces located in close geographic proximity tend to share similar characteristics represented by the sparse principal components. The stronger spatial dependence observed in SPC1 suggests that the dominant variation captured by this component is more spatially structured than the variation represented by SPC2.

Overall, the analysis demonstrates that among the nine initial variables examined, only six variables (X1, X3, X4, X5, X6, and X7) were retained because they exhibited non-zero loadings in the sparse principal components. In contrast, X2, X8, and X9 were excluded due to zero loadings, indicating that these variables contributed negligibly to the extracted component structure. This finding is consistent with the fundamental characteristic of Sparse Principal Component Analysis, which simultaneously performs dimensionality reduction and variable selection by producing sparse loading vectors. As a result, the derived components are more parsimonious and substantially easier to interpret than those obtained from conventional PCA, making SPCA particularly suitable for multivariate datasets with correlated predictors. Although the present study focuses on spatially distributed socio-economic data, the resulting sparse component structure also facilitates the interpretation of regional variation in the underlying factors associated with deforestation [20].

4. Conclusion

This study demonstrates that Sparse Principal Component Analysis (Sparse PCA) serves as an effective technique for reducing the dimensionality of socio-economic variables associated with deforestation in Indonesia, while simultaneously enhancing the interpretability of the analytical results. The findings indicate that two principal components are sufficient to represent the underlying structure of the dataset, capturing the majority of the total variance. The first component is largely characterized by the Human Development Index, poverty rate, and mean years of schooling, whereas the second component is primarily driven by population growth, total population, and GRDP per capita. Among the nine variables initially considered, only six were retained as significant contributors, while population density, agricultural land area, and the number of forest fires were excluded due to their minimal influence. This result underscores the capability of Sparse PCA not only in dimensionality reduction but also in performing implicit variable selection.

Furthermore, the spatial distribution and score plot analyses reveal substantial regional heterogeneity in socio-economic characteristics, particularly between western and eastern regions of

Indonesia. Overall, this study offers a more parsimonious and interpretable analytical framework for identifying key socio-economic drivers of deforestation, and provides meaningful insights to support evidence-based policy development in sustainable forest management. From a practical perspective, future studies are recommended to incorporate additional environmental and institutional variables to obtain a more comprehensive understanding of deforestation dynamics. Moreover, policymakers are encouraged to prioritize region-specific strategies that account for socio-economic disparities in order to enhance the effectiveness of sustainable forest management initiatives.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Mitha Rabiyatul Nufus: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Visualization, Manuscript Writing. **Jenike Gracelya Noke**: Investigation, Data Curation, Visualization, Manuscript Writing and Review. **Eusabius Paul Pega**: Formal Analysis, Investigation, Visualization, Manuscript Editing.

References

[1] M. Leon, G. Cornejo, M. Calderón, E. González-Carrión, and H. Florez, "Effect of Deforestation on Climate Change: A Co-Integration and Causality Approach with Time Series," *Sustainability*, vol. 14, no. 18, 2022, doi: https://doi.org/10.3390/su141811303.

[2] H. M. Njoya, K. Hounkpati, K. Adjonou, K. Kokou, S. Sieber, and K. Löhr, "Socioeconomic analysis of deforestation and economically sustainable farming systems to foster forest landscape restoration in central Togo," *Front. Sustain. Food Syst*, vol. 8, pp. 1-16, 2024, doi: https://doi.org/10.3389/fsufs.2024.1466008.

[3] M. R. Nufus, R. Silaban, E. P. Pega, J. G. Noke, E. F. Karo-Karo, "Analisis nonparametrik laju deforestasi dengan model spline truncated berbasis faktor sosial ekonomi di Indonesia," *Wahana Forestra: Jurnal Kehutanan*, vol. 20, no. 2, pp. 124-134, 2025, doi: https://doi.org/10.31849/f3xg9n24.

[4] Global Forest Watch, Retrieved from Indonesia Deforestation Rates: https://www.globalforestwatch.org/dashboards/country/IDN/?category=forest-change&lang=id&location=WyJjb3VudHJ5IiwiSUROIl0%3D&map=eyJjZW50ZXIiOnsibGF0IjotMi41Nzg0NTMwNjA1NjA0NjU2LCJsbmciOjExOC4wMTUxNTU3ODk5ODA5Mn0sInpvb20iOjIuNDMxNTQ5Njk3NjM2NDc1LCJjYW5Cb3, March. 2026.

[5] A. Tyukavina, P. Potapov, M. C. Hansen, A. H. Pickens, S. V. Stehman, S. Turubanova, D. Parker, V. Zalles, A. Lima, I. Kommareddy, X. P. Song, L. Wang, and N. Harris, "Global Trends

of Forest Loss Due to Fire From 2001 to 2019," *Front. Remote Sens*, vol. 3, pp. 1-20, 2022, doi: https://doi.org/10.3389/frsen.2022.825190.

[6] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *American Journal of Applied Mathematics and Statistics*, vol. 8, no. 2, pp. 39-42, 2020, doi: https://doi.org/10.12691/ajams-8-2-1.

[7] H. I. Dertli, D. B. Hayes, and T. G. Zorn, "Effects of multicollinearity and data granularity on regression models of stream temperature," *Journal of Hydrology*, vol. 639, pp. 1-11, 2024, doi: https://doi.org/10.1016/j.jhydrol.2024.131572.

[8] B. B. Alkan and I. Ünaldi, "Robust sparse principal component analysis: situation of full sparseness," *Journal of Applied Mathematics, Statistics and Informatics*, vol. 18, no. 1, pp. 5-20, 2022, doi: https://doi.org/10.2478/jamsi-2022-0001.

[9] A. Chowdhury, A. Bose, S. Zhou, D. P. Woodruff, and P. Drineas, "A Fast, Provably Accurate Approximation Algorithm for Sparse Principal Component Analysis Reveals Human Genetic Variation Across the World," *Annual International Conference, RECOMB*, pp. 86-106, 2022, doi: https://doi.org/10.1007/978-3-031-04749-7_6.

[10] S. Park, E. Ceulemeans, and K. Van Deun, "Acritical assessment of sparse PCA (research): why (one should acknowledgethat) weights are notloadings," *Behaviour Research Methods*, vol. 56, pp. 1413-1432, 2024, doi: https://doi.org/10.3758/s13428-023-02099-0.

[11] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics,"* vol. 15, no. 2, pp. 265-286, 2006, doi: https://doi.org/10.1198/106186006X113430.

[12] Q. Ke, S. Xu, S. Zong, X. Jiang, and S. Li, "Spatiotemporal dynamics and driving mechanisms of agricultural expansion into forests in Southeast Asia," *Journal of Environmental Management*, vol. 393, pp. 1-15, 2025, doi: https://doi.org/10.1016/j.jenvman.2025.127030.

[13] C. A. Asrat, Makkulau, and I. Yahya, "Perbandingan Metode Principal Component Analysis (PCA) dan Partial Least Square (PLS) dalam Penanganan Multikolinearitas pada Kasus Kemiskinan di Provinsi Sulawesi Tenggara Tahun 2023," *Arus Jurnal Sains dan Teknologi*, vol. 3, no. 1, pp. 68-82, 2025, doi: https://doi.org/10.57250/ajst.v3i1.1164.

[14] C.-C. Jeng, "Why a Variance Inflation Factor of 10 Is Not an Ideal Cutoff for Multicollinearity Diagnostics," *Journal of Educational Research and Development*, vol. 57, no. 2, pp.67-92, 2023, doi: https://doi.org/10.53106/199044282023105702004.

[15] F. Chen, and K. Rohe, "A New Basis for Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, vol. 33, no. 2, pp. 421-434, 2024, doi: https://doi.org/10.1080/10618600.2023.2256502.

[16] S. Widiarto, R. Fauziyah, T. P. Astari, N. L. Juliasih, S. Hadi, L. Zakaria, and I. Saputra, "Authentication of Processed Beef Sausage Products Using Chemometric Analysis Based on FTIR Spectrophotometry Data," *Jurnal Kimia Sains dan Aplikasi*, vol. 28, no. 1, pp. 39-46, 2025.

[17] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Art. no. 20150202, 2016. doi: 10.1098/rsta.2015.0202.

[18] J. Purwanto, T. Rusolono, and L. B. Prasetyo, "Spatial model of deforestation in Kalimantan from 2000 to 2013," *Jurnal Manajemen Hutan Tropika*, vol. 21, no. 3, pp. 110–118, 2015, doi: 10.7226/jtfm.21.3.110.

[19] M. F. Barri et al., Papua Bioregion: The Forest and Its People. Bogor, Indonesia: Forest Watch Indonesia, 2019. [Online]. Available: https://fwi.or.id/wp-content/uploads/2020/06/FWI-2019-Papua-Bioregion-The-Forest-and-Its-People.pdf

[20] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006, doi: 10.1198/106186006X113430