MSI Journal of
AI and Technology

CLASSIFICATION OF WOOD STRENGTH CLASSES BASED ON MECHANICAL PROPERTIES USING CLUSTER ANALYSIS

Muhammad Nur Aidi[1*]

[1*] Department of Statistics, IPB University

*** Correspondence: Muhammad Nur Aidi**

The authors declare that no funding was received for this work.

OPEN ACCESS

Received: 20-April-2025
Accepted: 30-April-2025
Published: 08-May-2025

This article is published by **MSI Publishers** in **MSI Journal of AI and Technology**
ISSN: XXXX-XXXX (Online)
Volume: 1, Issue: 1
(April-Jun) (2025)

ABSTRACT: This paper presents a classification process of various wood types by applying a hybrid methodology that combines unsupervised and supervised classification techniques. The unsupervised classification employs Hierarchical Cluster Analysis to group wood types based on their mechanical characteristics. The resulting clusters are then used as input for Discriminant Analysis to derive discriminant functions and validate the classification. The combined approach clearly demonstrates the effectiveness of clustering in distinguishing between groups, as shown by significantly different mean values for each mechanical variable across the clusters. This approach not only enhances classification accuracy but also provides a clearer understanding of wood strength classes, particularly for lesser-known wood species.

Keywords: *Wood classification, mechanical properties, cluster analysis, discriminant analysis, multivariate analysis, wood strength class*

1. Introduction

Wood is one of the most essential natural resources, with approximately one-third of the earth's land surface covered by forests, containing an estimated 300,000 million m³ of wood (Steinlin, 1979 as cited in Fengel & Wegener, 1995)[1]. Its usage spans construction materials to raw inputs for chemical industries. One of wood's most important characteristics is its renewability, making it a sustainable choice for long-term use.

Despite the abundance of wood types, users often prefer familiar species due to perceived reliability in strength, overlooking many others that may offer similar or even superior properties.

In Indonesia, wood strength is traditionally classified into five strength classes based on its specific gravity. However, Karnasudirdja et al. (1973)[2] argued that these classifications are tentative. New or unfamiliar wood species are typically not classified, leaving users with little guidance in selecting alternatives. Therefore, a robust classification system based on measurable physical and mechanical properties is needed.

2. Objective

This study aims to classify 50 wood species based on their physical and mechanical characteristics using multivariate statistical methods. Specifically, it applies hierarchical cluster analysis followed by discriminant analysis to determine distinct wood strength classes.

3. Theoretical Framework

3.1. Wood Characteristics

Indonesia is home to around 4,000 wood species, with about 400 having commercial potential. Of these, 259 species are officially traded and grouped into approximately 120 trade categories (Martawijaya et al., 1989)[3]. For practical applications and pricing, wood species are grouped by strength and durability. Strength classes are based on specific gravity, modulus of rupture (MOR), and compressive strength parallel to grain. Durability is closely related to specific gravity and extractive content (Seng O.D, 1990)[4]. The Indonesian classification system divides wood strength into five classes (SNI, 1999)[5]:

- Strength Class I: Very strong
- Strength Class II: Strong
- Strength Class III: Moderately strong
- Strength Class IV: Weak

- Strength Class V: Very weak

3.2. Multivariate Analysis

Multivariate analysis refers to a collection of statistical techniques used to analyze data involving multiple variables simultaneously. Data can be metric (interval or ratio) or non-metric (nominal or ordinal) (Santoso, 2005)[6]. A variate in multivariate analysis is a linear combination of measured variables with empirically determined weights: **Variate value** = $w_1X_1 + w_2X_2 + ... + w_nX_n$. Where X_i is the i-th variable, and w_i is the corresponding weight determined through the analysis.

3.3. Principal Component Analysis (PCA)

PCA and Factor Analysis aim to extract a reduced set of components or factors from a larger set of variables while retaining most of the data's variability (Wuensch, 2005)[7,8]. The main objectives of PCA are:

- Dimensionality reduction
- Data interpretation

If $\mathbf{X'} = (\mathbf{X_1, X_2, ..., X_p})$ represents a vector of p variables with mean vector μ and covariance matrix Σ or correlation matrix R, the j-th principal component Y_j is defined as: $\mathbf{Y_j = a_{1j}X_1 + a_{2j}X_2 + ... + a_{pj}X_p = a_jX}$. Where a_j is the eigenvector corresponding to the eigenvalue λ_j. If variables are measured on different scales, the correlation matrix R is used to extract the principal components. The number of principal components retained is usually determined by a cumulative variance threshold, commonly 75% or higher (Wuensch, 2005). The strength of association between the original variables and components is evaluated using correlation coefficients derived from eigenvectors and eigenvalues.

3.4. Cluster Analysis

Cluster analysis belongs to the category of *interdependence techniques*, which aim to group variables or objects based on similarities among them. In this context, the treatment is applied to rows, i.e., observations (Santoso, 2005)[6]. The primary objective of cluster analysis is to classify a set of objects into groups, or *clusters*, such that objects within the same cluster are more similar to each other than to those in other clusters. A well-defined cluster exhibits high internal homogeneity and high external heterogeneity (Santoso, 2005[6]; Kaufman & Rousseeuw, 2005[9]).

The process of cluster analysis typically involves the following steps (Santoso, 2005[6]; Everitt et al., 2011[11]):

a. Measuring Similarity or Distance Between Objects:

- Correlation-based similarity: measuring correlation between pairs of objects across variables.
- Distance-based methods: commonly used is Euclidean distance, which is computed as:

$$\Delta_{jk} = \left[\sum_{i=1}^{n} \left(X_{ij} - X_{ik} \right) \right]^{1/2}$$

Where Δ_{jk} is the distance between objects j and k, and Xij, Xik are the values of variable Xi for objects j and k respectively.

- Association measures can also be used to determine similarity.

b. Forming Clusters:

There are two main approaches:

- Hierarchical Methods (e.g., agglomerative or divisive), often visualized using a *dendrogram*. The clustering begins by grouping the most similar objects and proceeds iteratively.
- Non-Hierarchical Methods, such as the K-Means algorithm, require the number of clusters to be specified in advance, and the algorithm assigns observations to the nearest cluster center iteratively.

c. Cluster Validation and Profiling:

Once clusters are formed, they must be validated to assess their reliability and distinctiveness. *Cluster profiling* involves interpreting the characteristics of each cluster. Further analysis, such as discriminant analysis, can be performed based on the profiling results (Hair et al., 2010)[11].

Hierarchical clustering methods include:

- Single linkage (nearest neighbor): merges clusters with the smallest minimum distance.
- Complete linkage (farthest neighbor): merges clusters with the smallest maximum distance.
- Average linkage: merges based on average distance between all pairs of objects.
- Ward's method: minimizes the total within-cluster variance; uses the distance between cluster centroids.

Cluster analysis assumes the following (Santoso, 2005[6]; Everitt et al., 2011[11]):

- The sample is representative of the population.
- There is minimal multicollinearity between variables; ideally, correlations should not exceed 0.5.
- Variables should be standardized (e.g., transformed into z-scores) to prevent bias from differing measurement scales.

Hierarchical clustering can be divided into *agglomerative* (bottom-up) and *divisive* (top-down) methods. The clustering result is often visualized using a dendrogram, which can be cut at the point of the largest increase in fusion distance to determine the optimal number of clusters (Dillon & Goldstein, 1984)[12].

3.4. Discriminate Analysis

Discriminate analysis is a type of *dependence technique*, involving both dependent and independent variables. Its unique feature is that the dependent variable must be categorical (e.g., group codes), while independent variables are typically metric (Santoso, 2005)[6].

The main objectives of discriminant analysis are (Santoso, 2005[6]; Hair et al., 2010[11]):

- To determine whether there are significant differences between groups on the dependent variable.
- To identify which independent variables contribute most to the discrimination between groups.
- To develop a discriminant function similar to a regression model.

- To classify observations (rows in SPSS) into their respective groups based on the discriminant function.

The general steps in discriminant analysis include:

- Defining dependent and independent variables.
- Choosing a method to build the discriminant function:
  - Simultaneous Estimation: includes all predictors at once.
  - Stepwise Estimation: variables are entered sequentially based on their discriminating power.
- Testing the significance of the discriminant function using Wilks' Lambda, Pillai's Trace, F-test, etc.
- Evaluating the classification accuracy of the model, including individual case diagnostics.
- Interpreting and validating the discriminant function.

Assumptions required for valid discriminant analysis include (Santoso, 2005[6]; Tabachnick & Fidell, 2013[13]):

- Multivariate normality: the independent variables must be normally distributed.
- Homogeneity of covariance matrices: variances across groups should be equal.
- Independence among predictors: absence of multicollinearity.
- No significant outliers among the independent variables.

There is no strict rule for the ideal sample size, but a general guideline suggests a minimum of five cases per independent variable. For example, if six independent variables are used, at least 30 cases are recommended (Santoso, 2005)[6].

4. Methodology

4.1. Data Source

The data used in this study are secondary data, derived from a portion of a research report by the Center for Research and Development of Forest Product Technology, Ministry of Forestry, as cited in Purnamasari (2002)[14]. The dataset comprises mechanical properties of various wood species commonly found in Indonesia.

Observed Variables

Eight mechanical and physical properties of wood were considered as variables for analysis:

- **X1:** Modulus of Elasticity (MOE) (kg/cm^2)
- **X2:** Modulus of Rupture (MOR) (kg/cm^2)
- **X3:** Maximum compressive strength parallel to grain (kg/cm^2)
- **X4:** Radial shear strength (kg/cm^2)
- **X5:** Tangential shear strength (kg/cm^2)
- **X6:** End hardness (kg/cm^2)
- **X7:** Side hardness (kg/cm^2)
- **X8:** Specific gravity (kg/dm^3)

These variables are selected based on their relevance to the classification of wood strength as recommended by various forestry standards (Skaar 1988; Bowyer et al., 2003)[15,16].

4.2. Analytical Procedure

The data analysis follows these steps:

1. **Exploratory Correlation Analysis**

   A correlation matrix was constructed to assess the degree of multicollinearity among variables. Due to the existence of strong correlations and different measurement units, a **Principal Component Analysis (PCA)** was deemed necessary using the correlation matrix instead of the covariance matrix.

2. **Standardization**
   Variables were standardized into **Z-scores** to normalize differences in scale across measurements, as recommended for PCA when variables are measured in different units (Jolliffe & Cadima, 2016)[17].

3. **Principal Component Analysis (PCA)**

PCA was employed to reduce dimensionality and to identify components that capture the majority of the variance among variables. The **eigenvalue >1 criterion** (Kaiser's rule) was applied to select principal components (Kaiser, 1960)[18].

4. **Hierarchical Cluster Analysis**

Hierarchical clustering was conducted using the **Euclidean distance** and **average linkage** (UPGMA method). A **dendrogram** was constructed to visualize groupings, with cluster cutting determined at the point of greatest inter-cluster distance gain, ensuring meaningful separation between groups.

5. **Cluster Characterization**

Each resulting cluster was characterized based on the mean scores of the selected principal components to interpret the mechanical profiles of wood groups.

6. **Discriminant Analysis**

A **Linear Discriminant Analysis (LDA)** was performed to evaluate the accuracy of cluster membership assignments and to develop classification functions for wood strength groupings. Discriminant functions were interpreted based on their standardized coefficients.

5. Result and Discussion

5.1. Correlation Among Observed Variables

The correlation analysis revealed that all pairs of variables exhibited significant positive correlations. The highest correlation ($r = 0.978$) was found between wood hardness at the end (X6) and side hardness (X7), indicating that wood hardness is relatively uniform across different orientations. Modulus of Rupture (MOR, X2) was strongly associated with parallel compression strength (X3) as well as with both types of hardness (X6 and X7), suggesting these properties are critical indicators of overall wood strength. Radial shear (X4) and tangential shear (X5) were also highly correlated ($r = 0.835$), indicating similar response patterns under shear stress.

Conversely, the lowest correlations were observed between modulus of elasticity (MOE, X1) and radial shear (X4) ($r = 0.159$), and between MOE and tangential shear (X5) ($r = 0.231$). This suggests that stiffness (as represented by MOE) does not necessarily align with the material's shear response—an observation consistent with prior studies on anisotropic mechanical behavior of wood (Tsoumis, 1991[19]; Kollmann & Côté, 1968[20]).

5.2. Principal Component Analysis (PCA)

Prior to performing PCA, all variables were standardized using Z-scores due to differing measurement units. As illustrated in **Figure 1** (Scree Plot), only the first two principal components (PCs) had eigenvalues greater than 1. A sharp drop was observed between the first and second components, with subsequent components contributing marginally to total variance.

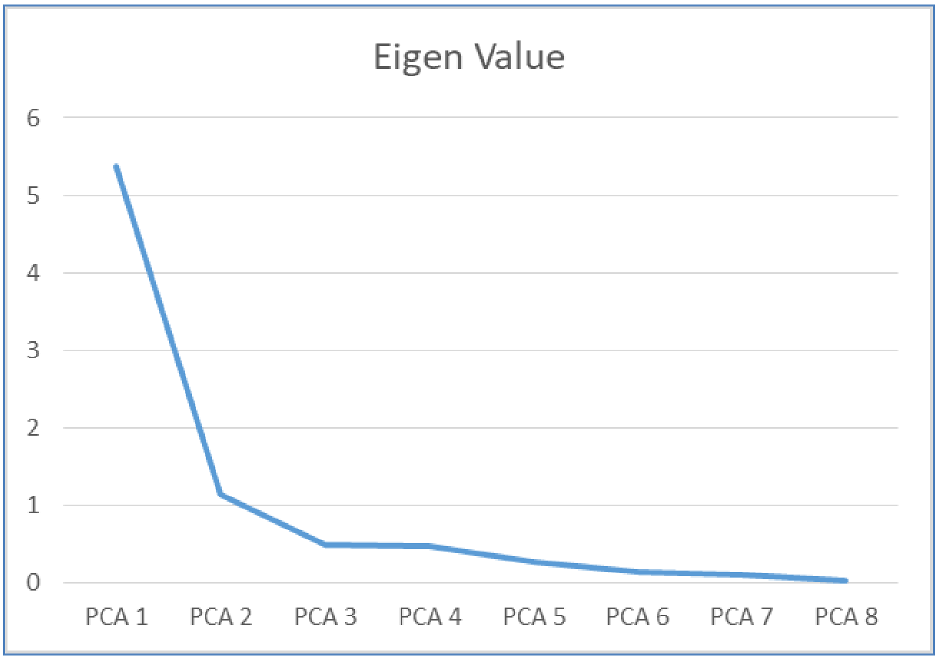

Figure 1. Scree Plot of All Principal Components

Table 1. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.363	67.034	67.034	5.363	67.034	67.034
2	1.144	14.304	81.338	1.144	14.304	81.338
3	0.497	6.215	87.553			
4	0.472	5.902	93.455			

5	0.263	3.292	96.747			
6	0.141	1.794	98.511			
7	0.099	1.237	99.748			
8	0.020	0.252	100.000			

The cumulative variance explained by the first two components was 81.34%, which is generally considered sufficient for dimensionality reduction in multivariate analysis (Jolliffe & Cadima, 2016)[17]. Hence, the original eight variables (X1–X8) can be effectively reduced to two components while retaining most of the original information.

Table 2. Component Matrix

	PCA	
	1	2
X6	0.928	-0.045
X7	0.921	-0.060
X2	0.911	-0.148
X3	0.891	-0.190
X8	0.801	-0.154
X5	0.743	0.574
X4	0.687	0.640
X1	0.603	-0.564

Based on the component matrix presented in Table 2, the corresponding mathematical expressions for the principal components are as follows:

$\mathbf{PC_1} = 0.928X_6 + 0.921X_7 + 0.911X_2 + 0.891X_3 + 0.801X_8 + 0.743X_5 + 0.687X_4 + 0.603X_1$

$\mathbf{PC_2} = -0.045X_6 - 0.060X_7 - 0.148X_2 - 0.190X_3 - 0.154X_8 + 0.574X_5 + 0.640X_4 - 0.564X_1$

The component matrix indicates that all variables within the first principal component exhibit positive correlations with the mechanical properties of wood. This suggests that increases in the values of variables X_1 through X_8 are associated with improved wood strength. In contrast, the second principal component displays strong positive correlations with radial shear strength (X_4) and tangential shear strength (X_5), with loading values of 0.640 and 0.574, respectively. This relationship implies that wood samples with high strength scores also tend to possess superior shear strength characteristics.

PC1 is characterized by uniformly positive and high loadings across all mechanical properties, suggesting it represents an overall "wood strength" dimension. **PC2** shows high positive loadings for shear strength variables (X4 and X5), suggesting it captures the wood's resistance to shear forces.

5.3. Cluster Analysis

A hierarchical cluster analysis was performed using the average linkage method and Euclidean distances. The resulting dendrogram suggested a meaningful cluster cut-off at a distance of 15, producing three clusters, as detailed in Table 3.

Table 3. Results of Hierarchical Cluster Analysis

Cluster	Species Number
I	3, 44, 26, 41, 46, 48, 13, 29, 32, 14, 39, 21, 28, 17, 25, 2, 11, 5, 7, 10, 15, 4, 22, 27, 12, 37, 31, 23, 38
II	6, 34, 20, 40, 43, 8, 42, 30, 35, 16, 24, 33, 39, 19, 47, 18, 50, 45, 36
III	9

This cut-off is appropriate because it maximizes dissimilarity between clusters while maintaining homogeneity within clusters. According to the wood strength classification system used by the Ministry of Forestry (Seng V, 1990[21]; Ministry of Forestry, 1992[22]), the wood species in these clusters vary significantly in strength:

- **Cluster I** includes species such as *Toona sureni* (No. 46), classified as strength class IV (low strength).
- **Cluster II** includes *Dillenia obovata* (No. 24), classified as strength class III (moderate strength).
- **Cluster III** includes *Calophyllum inophyllum* (No. 9), classified as strength class II (high strength).

These findings confirm the effectiveness of clustering in grouping wood types by mechanical performance.

5.4. Discriminate Analysis

Due to having only one member, Cluster III was excluded from the discriminant analysis. The discriminant function analysis yielded a **98% classification accuracy**, which is highly acceptable for biological and material sciences where perfect classification is rare (Hair et al., 2010)[11]. Only one misclassification occurred within Group I (Table 4).

Table 4. Classification Results Based on Discriminant Analysis

Actual Group	Predicted Group 1	Predicted Group 2	Misclassification %
1	29	0	0%
2	1	19	5%

Table 5. Linear Discriminant Functions for Each Group

Constant / Variable	Function 1	Function 2
Constant	-0.9708	-25.702
X1	-0.1791	-18.471
X2	0.4283	17.332
X3	0.3226	-0.6699
X4	0.1230	-0.3462
X5	-11.660	18.045
X6	-24.612	35.592
X7	-0.1381	-0.0351
X8	0.2364	-0.4864

The linear discriminant functions are given below:

Function 1 (Group 1):

$Y_1 = -0.9708 - 0.1791X_1 + 0.4283X_2 + 0.3226X_3 + 0.1230X_4 - 1.1660X_5 - 2.4612X_6 - 0.1381X_7 + 0.2364X_8$

Function 2 (Group 2):

$Y_2 = -2.5702 - 1.8471X_1 + 1.7332X_2 - 0.6699X_3 - 0.3462X_4 + 1.8045X_5 + 3.5592X_6 - 0.0351X_7 - 0.4864X_8$

These functions can be used to predict the group membership of a given wood sample based on its mechanical characteristics.

Dominant Variables in Each Cluster

To determine which variables are the most influential in each cluster, Principal Component Analysis (PCA) was conducted for each cluster. The results are shown in Table 6 and Table 7.

Table 6. Component Matrix for Cluster 1

Variable	Component 1
X2	0.955
X3	0.935
X7	0.883
X6	0.882
X8	0.882
X1	0.853
X4	0.769
X5	0.747

Table 7. Component Matrix for Cluster 2

Variable	Component 1	Component 2
X2	0.882	0.038
X6	0.856	0.211
X7	0.847	0.244
X1	0.717	-0.244
X3	0.697	0.343
X5	-0.641	0.453
X8	0.460	0.456
X4	-0.529	0.764

From Tables 6 and 7, it can be seen that in both clusters, the variable most strongly influencing wood strength characteristics is MOR (X2). However, the second most dominant variable differs: for Cluster 1 it is parallel compression (X3), while for Cluster 2 it is end hardness (X6).

The number of principal components that represent the total variance differs between the clusters. In Cluster 1, a single principal component explains 74.9% of the variance. In Cluster 2, two components together account for only 67.6% of the variance.

For both clusters, **MOR (X2)** consistently emerged as the most dominant variable, indicating its critical role in defining wood strength classes. However, the second most dominant variable differed: it was **parallel compression (X3)** for Cluster I, and **end hardness (X6)** for Cluster II. This variation highlights how different mechanical aspects contribute to the classification of wood strength across groups.

5.5. Mean Difference Test Between Clusters

A mean difference test was conducted to determine whether the mean values of each variable differ significantly between the two clusters. If the mean values of the variables in both clusters differ significantly, the clustering can be considered valid.

Based on the individual difference tests at the 95% confidence level, the results show that the means of the two clusters differ significantly. This indicates that the two clusters have distinct characteristics (strengths). In other words, the clustering performed using hierarchical cluster analysis, as described in Section C above, is highly effective.

6. CONCLUSION

The results of the principal component analysis revealed that two principal components were sufficient to represent the total variability of all variables. Hierarchical cluster analysis identified three distinct clusters based on mechanical properties, which serve to differentiate wood strength classes. Discriminate analysis demonstrated a high level of classification accuracy, reaching 98%, with only one tree species being misclassified. Furthermore, the principal component analysis indicated that the most influential variable in both Cluster 1 and Cluster 2 is the Modulus of Rupture (MOR), denoted as variable X2. The mean difference test between the two clusters showed a statistically significant difference, confirming that the clustering approach used in the analysis was highly effective.

7. Literature References

1. Fengel, D., & Wegener, G. 1995. Wood: Chemistry, Ultrastructure, Reactions (Kayu Kimia Ultrastruktur Reaksi-Reaksi). Gadjah Mada University Press.
2. Karnasudirdja, S., B. Ginoga, and O. Rachman. 1973. Classification of Wood Strength Based on the Relationship Between Modulus of Rupture and Other Wood Strength Properties(Klasifikasi Kekuatan Kayu Berdasarkan Hubungan antara Keteguhan Lentur Patah dengan Sifat Keteguhan Kayu Lainnya) . Report No. 115, Agricultural Research and Development Agency, Bogor.

3. Martawijaya, A., Iding Kartasujana, Y.I. Mandang, S.A. Amongprawira, and K. Kadir. 1989. Atlas of Indonesian Wood. Ministry of Forestry, Forestry Research and Development Agency, Bogor.
4. Seng, O. D. 1990. Specific Gravity of Indonesian Wood Species and Understanding Wood Weight for Practical Purposes. Forest Products Research and Development Center, Ministry of Forestry, Bogor.
5. Indonesian National Standard (SNI). 1999. Wood Preservation for Housing and Buildings. Http://www.dephut.go.id/INFORMASI/SNI/pkupg.htm. Diakses 14 Juni 2007
6. Santoso, M. H., Naim, S., Suroso, S., Hayudini, M. A. A., & Shrestha, P. (2023). Influence of Work Environment and Employee Competence Analysis on Employee Performance. *Chanos Chanos*, *21*(1), 31-40.
7. Wuensch, K. L. 2005. Principal Components and Factor Analysis. East Carolina University.
8. Wuensch, K. I. 2005. Multivariate Analysis with SPSS. Department of Psychology. East Carolina University. http://core.eco.edu/psyc/wuenschk/SPSS/SPSS-MV.htm.
9. Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.
10. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis.
11. Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. In *Multivariate data analysis* (pp. 785-785).
12. Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York (NY): Wiley, 1984..
13. Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics (6. Baskı). *MA: Pearson*.
14. Purnamasari, S. 2002. Clustering of Several Wood Types Based on Mechanical Properties Using Cluster Analysis (Pengelompokan Beberapa Jenis Kayu Berdasarkan Sifat Mekanik dengan Analisis Gerombol). Bachelor's Thesis, Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB.
15. Skaar, C., & Skaar, C. (1988). Wood moisture and the environment. *wood-water relations*, 1-45.

16. Bowyer, J. L., Shmulsky, R., & Haygreen, J. G. (2007). *Forest products and wood science: an introduction* (No. Ed. 5, pp. xv+-558). Iowa: Blackwell.
17. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.
18. Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, *20*(1), 141-151.
19. Tsoumis, G. (1991). *Science and technology of wood: structure, properties, utilization* (Vol. 115). New York: Van Nostrand Reinhold.
20. Kollmann, F. F., Kuenzi, E. W., & Stamm, A. J. (2012). *Principles of wood science and technology: II wood based materials*. Springer Science & Business Media.
21. Seng Hua, L., Wei Chen, L., Antov, P., Kristak, L., & Md Tahir, P. (2022). Engineering wood products from Eucalyptus spp. *Advances in Materials Science and Engineering*, *2022*(1), 8000780.
22. Ministry of Forestry (1992). Guidelines for the Classification of Wood Strength Classes Based on Mechanical Properties. Directorate General of Forestry. (Pedoman Penggolongan Kelas Kekuatan Kayu Berdasarkan Sifat Mekanik. Direktorat Jenderal Kehutanan).