Neuroinformatics approach: Hierarchical cluster analysis of Indonesian provinces based on people's welfare indicators in the realm of data science and network studies

Restu Arisanti[a*], Aissa Putri Pertiwi[a], Sri Winarni[a] and Resa Septiani Pontoh[a]

[a]Department of Statistics, Padjadjaran University, Indonesia

C H R O N I C L E	A B S T R A C T
	The welfare of people has always piqued our interest, and it remains the primary goal of nations around the world in their development endeavors. To effectively drive development efforts, it is critical to understand the diverse welfare features that exist in different locations. Thus, the purpose of this statistical analysis is to classify Indonesian provinces based on a comprehensive set of People's Welfare Indicators, which includes Population Density (PD), Percentage of Poor Population (PPP), Life Expectancy Rate (LER), and Average Years of Schooling (AYS). The methodology used in this study is Hierarchical Cluster Analysis, which employs five distinctive techniques: Single Linkage, Average Linkage, Complete Linkage, Ward's Linkage, and the Centroid Method. The data for this study was obtained from reliable secondary sources, notably the official website of the Central Bureau of Statistics (BPS), and it provides insights on Indonesia's welfare picture in 2021. The average linkage approach shows as the most suitable of the five hierarchical cluster analysis methods used, with the closest cophenetic correlation to 1. The analysis reveals three distinctive clusters within the Indonesian context. Cluster 1 demonstrates a tendency toward low PWI (People's Welfare Index) status, while Cluster 2 exhibits a notably high PWI status. Cluster 3 occupies an intermediate position, characterized by moderate PWI status. These findings not only give useful classification but also act as an important reference point for the Indonesian government. They provide an in-depth insight into each province's distinct welfare features, supporting smart resource allocation and prioritizing aid distribution in regions of highest need. As a result, this research is an essential resource for creating equitable and effective policies and methods to improve people's well-being throughout Indonesia.

1. Introduction

The Indonesian Constitution, UUD 1945, expressly mandates the Republic of Indonesia's government to "promote the general welfare, educate the nation's life, and realize social justice for all Indonesian people." This emphasizes the idea that the right to a life free of poverty and the pursuit of happiness is a fundamental human right. It is the obligation of the government to make this right a reality through national development efforts. As a result, poverty alleviation remains a top development priority. In 2019, 9.22% of Indonesia's population lived below the poverty level. This percentage rose by 0.97 points the next year to 10.19%. However, by 2021, it will have dropped to 9.71%. However, a closer look reveals large differences in poverty rates between Indonesian provinces. For example, Papua province had the highest poverty rate of 27.38%, while Kalimantan Selatan province had the lowest rate of 4.56% (Statistics of Indonesia, 2022). This disparity highlights the importance of focusing attention and initiatives in certain regions. Poverty can have a negative impact on both education and labor force participation. Poverty frequently restricts access to decent education, lowering an individual's chances of finding a steady job.

The consequent lack of economic stability and employment chances can greatly impair a country's general welfare, increasing a cycle of adversity and limiting opportunities for personal and communal improvement.

To solve this, accurate identification of target groups and places is required. This is where cluster analysis comes in efficiently, since it organizes Indonesian provinces based on the People's Welfare Indicators for 2021. Cluster analysis is a powerful tool for categorizing observations or variables, particularly when the number of groupings is unknown at the beginning. Its goal is to cluster n items based on p variables with comparable features, allowing objects to be classified into one or more clusters based on shared characteristics. Finally, the study's findings are intended to assist the Indonesian government in determining the distinct characteristics of each province in terms of People's Welfare Indicators, which will help in the prioritization of assistance distribution where it is most required.

2. Methodology

In a recent study, the data used is secondary data obtained from the official website of the Central Statistics Agency (bps.go.id), which consists of 34 samples which are the number of provinces in Indonesia. And the data used is the People's Welfare Indicator in Indonesia in 2021 which consists of four variables, namely Population Density (PD), Percentage of Poor Population (PPP), Life Expectancy Rate (LER), and Average Years of Schooling (AYS).

The research method used in this research is Hierarchical Cluster Analysis using the agglomerative method which aims to group provinces in Indonesia into clusters according to the characteristics of each province based on the People's Welfare Indicators.

2.1. Cluster Analysis

Cluster analysis is one of the multivariate analyses that aims to identify and group objects based on information obtained on data that describes objects and their relationships (Johnson and Wichern, 2002). A good cluster is one that is homogeneous or similar to a high degree among members in one cluster, and heterogeneous or highly different between clusters with other clusters. Some of the benefits of cluster analysis are multiple variable data exploration, data reduction, sampling stratification, prediction of object state.

In cluster analysis, there are two assumptions that must be met. The first is that the sample must be able to represent the population. Then, the sample must not have multicollinearity between variables. However, if multicollinearity is detected, it must be less than 0.5.

2.2. Object Similarity Measurement

Before grouping data or objects for detection, first determine the size of the proximity distance between the data elements. To determine the distance between data, the Euclidean distance is used which is the root of the sum of the squared differences between the 2 vectors, namely the x and y vectors. The smaller the value of d(x,y), the more similar the two vectors are matched/compared. Vice versa, if the greater the value of d(x,y), the more different / there is no similarity between the two vectors that are matched / compared (Budi Santosa, 2007). The following function is used to measure the distance between data using Euclidean distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - y_{jk})^2} \qquad (1)$$

with,

d_{ij}: distance between object i and j

x_{ik}: value of object i in variable x_k

y_{jk}: value of object j in variable y_k

2.3. Assumption of Cluster Analysis

Multicolinearity

In detecting whether there is a perfect linear relationship between independent variables or independent variables, the Multicollinearity Test is used. One way to detect multicollinearity is to use VIF (Variance Inflation Factor). So that the Multicollinearity Test hypothesis can be stated as follows:

H_0: There is no multicollinearity

H_1: There is multicollinearity

The test statistics used are as follows:

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2}$$

with,

i : 1, 2, 3, …, p

j : 1, 2, 3, …, p

r_{ij} : observed correlation coefficient between variables i and j

a_{ij} : partial correlation coefficient between variables i and j

Test Criteria: Reject H_0 if KMO value <0.50.

The eligibility criteria can be seen through the KMO table according to Subhash Sharma which is shown as follows:

Table 1
Keiser-Meyer-Olkin (KMO)

KMO Measurement		Recommendation
	≥ 0.90	Excellent
	≥ 0.80	Good
	≥ 0.70	Medium
	≥ 0.60	Fair
	≥ 0.50	Insufficient
Under	0.50	Rejected

Independence between Variables (Bartlett Sphericity Test)

In knowing whether there is a relationship between the variables analyzed, Bartletts's test is used. If the variables X_1, X_2, ..., X_p is independent (mutually free), then the correlation matrix between variables is equal to the identity matrix. So, the Bartlett's test hypothesis can be stated as follows:

H_0: Correlation matrix = identity matrix (no correlation)

H_1: Correlation matrix ≠ identity matrix (there is a correlation)

The test statistics used are as follows:

$$\chi^2 = -\left(n - 1 - \frac{2p + 5}{6}\right) ln|R|$$

with,

$ln|R|$: determinant value of the correlation matrix

n   : the number of observations

p   : the number of variables

Test Criteria: Reject H_0 if $\chi^2_{hitung} \leq \chi^2_{\frac{1}{2}p(p-1)}$ or p-value $< \alpha$ (0.05

2.4 Hierarchical Cluster Analysis

The hierarchical cluster method is a method used to group observations in a structured manner based on similarity of properties and the number of groups that can be formed from the largest to the smallest group and vice versa. Hierarchical clustering starts with one cluster obtained from the observation of an object and ends with a cluster containing all objects or vice versa. The hierarchical cluster method is a method in cluster analysis that forms certain levels, such as in the form of a tree to perform a gradual clustering process. The results of this clustering are presented in the form of a dendrogram (Bridges Jr, 1966; Köhn & Hubert, 2014).

Agglomerative Method

Agglomerative is a clustering method that starts with individual objects and considers the number of clusters equal to the number of objects. The most similar objects will first join to form a cluster, and so on to form one cluster while still taking into account the proximity distance between objects (Orlóci, 1967). This method has several clustering procedures, namely:

    a.   Single Linkage

The closest distance technique (single linkage clustering) (Sharma & Batra, 2019) is a method that groups two objects that have the closest distance. For example, combining two corresponding objects U and V and forming a cluster (UV). The distance between (UV) and other clusters is W, so W can be calculated by:

$$d_{(uv)w} = min\{d_{uw,}d_{vw}\}$$

where, d_{uw} is the distance between the nearest neighbors of cluster U and W, and d_{vw} is the distance between the nearest neighbors of cluster V and W.

b.   Complete Linkage

The longest distance technique (complete linkage) (Sharma & Batra, 2019) is the opposite of single linkage, which is a method that groups two objects that have the farthest distance, then the distance between objects gets closer. For example, combine two objects that correspond to U and V and form a cluster (UV). The distance between (UV) and other clusters is W, so W can be calculated by:

$$d_{(uv)w} = max\{d_{uw,}d_{vw}\}$$

where, $d_{uw,}$ is the distance between the nearest neighbors of cluster U and W, and d_{vw} is the distance between the nearest neighbors of cluster V and W.

c.   Average Linkage

The average linkage technique (Emmendorfer & de Paula Canuto, 2021) is almost the same as Single Linkage or Complete Linkage, but the criterion used is the average distance of all individuals in a cluster with the distance of all individuals in another cluster. Average linkage calculates the distance between two clusters which is referred to as the average distance where the distance is calculated in each cluster. The average linkage procedure starts by defining the matrix $D = \{d_{ik}\}$ to obtain the closest objects, for example U and V. Then these objects are merged into a form called D=$\{d_{ik}\}$. Then these objects are merged into a cluster (UV). Furthermore, the distance between (UV) and other clusters, (W).

$$d_{(uv)w} = \frac{\Sigma_i \Sigma_k d_{ik}}{N_{UV} N_W}$$

where:

d_{ik}: Distance between object i in cluster (UV) and object k in cluster W.

N_{UV}: Number of items in UV cluster

N_W: Number of items in W cluster

d.   Ward's Method

Ward's method is also called the sum of squares method because the distance between two clusters in this method is the total sum of the squares of two clusters on each variable (Rencher, 2002). This method uses a variance analysis approach to calculate the distance between clusters. The Ward method is an agglomerative clustering method to obtain groups that have the smallest possible internal variance. This method uses complete calculation and maximizes homogeneity within a group.

Ward's method is calculated based on the following equation:

$$SSE = \sum_{i=1}^{n}(X_i - \hat{X})^2$$

where:

X_i: Column vector whose entry is the first object value with, i=1, 2, ..., N.

$\hat{X}$: Column vector whose entry is the average object value in the cluster

n: The number of objects in the formed cluster

Divisive Method

The divisive method (Roux, 2018) or divisive analysis method is a clustering process based on the similarity of average values between objects. If an object has the largest average value equation, then the object will separate and turn into a splinter group. In this divisive method, the calculation is seen from the difference or difference between the average value equation and the value of the matrix element that has become a splinter group. If the value difference between the average value equation and the value of the splinter group matrix element is negative, then the calculation stops so that a new matrix must be created to get another cluster. This calculation continues until all objects are separated.

2.5 Method Selection

One measure that can be used to select the best method is the cophenetic correlation coefficient. The cophenetic correlation coefficient is the correlation coefficient between the original elements of the dissimilarity matrix and the elements of the matrix generated by the dendrogram (cophenetic matrix) (Silva, 2013). The cophenetic correlation coefficient compares the observed distance between samples and the predicted distance from the clustering process (Carvalho, 2019). The cophenetic correlation coefficient can be defined as a measure of the correlation between the cophenetic distance of two time series data objects and the original distance matrix (Kumar, 2016).

According to Saracli et al. (2013), the formula for calculating the cophenetic correlation coefficient is as follows:

$$r_{Coph} = \frac{\Sigma_{i<k}(d_{ik} - \underline{d})(d_{c_{ik}} - \underline{d_c})}{\sqrt{\left[\Sigma_{i<k}(d_{ik} - \underline{d})^2\right]\left[\Sigma_{i<k}(d_{c_{ik}} - \underline{d_c})^2\right]}}$$

where:

r_{Coph}: cophenetic correlation coefficient

d_{ik}: original distance between objects i and k (euclidean distance)

$\underline{d}$: average of the original distances between objects i and k

$d_{c_{ik}}$: cophenetic distance of object i and k

$\underline{d_c}$: average cophenetic distance of object i and k

3. Result and Discussion

3.1 Descriptive Statistics

The following table calculates the descriptive statistics of the People's Welfare Index (PWI) in Indonesia by province.

Table 2
Descriptive Statistics

Variable	Mean	Median	Max	Min	Standard Deviation
PD	744.26	103.50	15978.00 (DKI Jakarta)	9.00 (Kalimantan Utara)	2721.058
PPP	10.43	8.51	27.38 (Papua)	4.56 (South Kalimantan Selatan)	5.411687
LER	70.20	70.11	75.08 (DI Yogyakarta)	65.29 (Sulawesi Barat)	2.513481
AYS	8.716	8.755	11.170 (DKI Jakarta)	6.760 (Papua)	0.928391

3.2 Assumptions Test

Multicollinearity test

Multicollinearity Test

Hypothesis

H_0: There is no multicollinearity.

H_1: There is multicollinearity.

Significance Level: $\alpha = 0.05$

Test Statistic: $VIF_i = \frac{i}{1-R^2}$

Test Criteria: Reject H_0 if the VIF value> 10, and accept in other cases.

Results:

Table 3
VIF Value

X_1	X_2	X_3	X_4
1.296363	1.669312	1.553479	1.627417

Based on the output results above, the 4th variable of the Indonesian People's Welfare Indicator has a variance inflation factor (VIF) value of less than 10, so H_0 is accepted. That is, with a significance level of 5%, the opinion supports that the data is identified as non-multicollinearity or there is no multicollinearity, so the assumption is met.

Keiser-Meyer-Olkin Test

Hypothesis

H_0: The amount of data is sufficient.

H_1: The amount of data is not enough.

Significance Level: $\alpha = 0.05$

Test Statistic: $KMO = \frac{\Sigma_{i=1}^p \Sigma_{j=1}^p r_{ij}^2}{\Sigma_{i=1}^p \Sigma_{j=1}^p r_{ij}^2 + \Sigma_{i=1}^p \Sigma_{j=1}^p a_{ij}^2}$

Test Criteria: Reject H_0 if KMO value <0.50, and accept in other cases.

Result: KMO = 0.6693627

Based on the output results above, the KMO result = 0.6693627 is obtained so that H_0 is accepted. So, with a significant level of 5%, it can be concluded that the sample used is sufficient for further analysis, so the assumption is met.

Bartlett's Test

Hypothesis

H_0 : Correlation matrix = identity matrix (no correlation)

H_1: Correlation matrix = identity matrix (there is a correlation)

Significance Level: $\alpha = 0.05$

Test statistic: $\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right)ln|R|$

Test Criteria: Reject H_0 if $\chi^2{}_{hitung} \leq \chi^2{}_{\frac{1}{2}p(p-1)}$ or p-value $< \alpha$

Result of Bartlett K-squared = 1200.4

Df = 3

p-value = < 2.2e-16

Based on the above output results, the p-value = 2.2×10^{-16} is obtained so that H_0 is rejected. So, with a significant level of 5% it can be concluded that there is a correlation between the independent variables, so the assumption is met.

3.3 Cluster Analysis by Hierarchical Method

In the process of choosing a cluster algorithm, the hierarchical method used consists of 4 methods, namely single linkage, average linkage, complete linkage, and ward's linkage. The clustering process was carried out with the help of R software. In this study, the grouping of provinces in Indonesia was divided into 3 clusters. This is in accordance with the HDI category based on BPS which consists of "low", "medium", and "high" categories.

After analyzing the selection of cluster algorithm methods from the four existing methods, the best method will be selected based on the cophenetic correlation value. If the cophenetic correlation value obtained from the previous analysis is close to the value equal to 1, then the method is the best method. The following table shows the results of the cophenetic correlation value of each method.

Tabel 4

Correlation of Cophenetic

| Method	Correlation of Cophenetic value
Single Linkage	0.9975311
Complete Linkage	0.9982645
Average Linkage	0.9990208
Ward's Linkage	0.9300992

Based on the table above, in this study the appropriate method to be used in the next cluster analysis step is the average linkage method with a cophenetic correlation value of 0.9990208. The following is a grouping of provinces in Indonesia using the average linkage method.

Table 5

Cluster Grouping

Cluster	Regency/City	Frequency
1	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep. Bangka Belitung, Kep. Riau, DKI Jakarta, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, dan Papua.	27
2	DKI Jakarta	1
3	Jawa Timur, Bali, Jawa Barat, Jawa Tengah, DI Yogyakarta, dan Banten	6

The dendrogram is presented in Fig. 1. Based on the dendrogram of the average linkage method above, it can be seen that the data on the number of provinces based on the People's Welfare Indicators in Indonesia in 2021 is divided into 3 clusters, where each cluster consists of provinces based on the similarity of the level of People's Welfare Indicators owned by each province in Indonesia in 2021, namely in the first cluster there is only 1 province, namely DKI Jakarta. In the second cluster there are 27 provinces, namely Central Sulawesi, Maluku, and East Kalimantan. And the rest is the third cluster, there are 6 provinces, namely East Java, Bali, West Java, Central Java, DI Yogyakarta, and Banten.
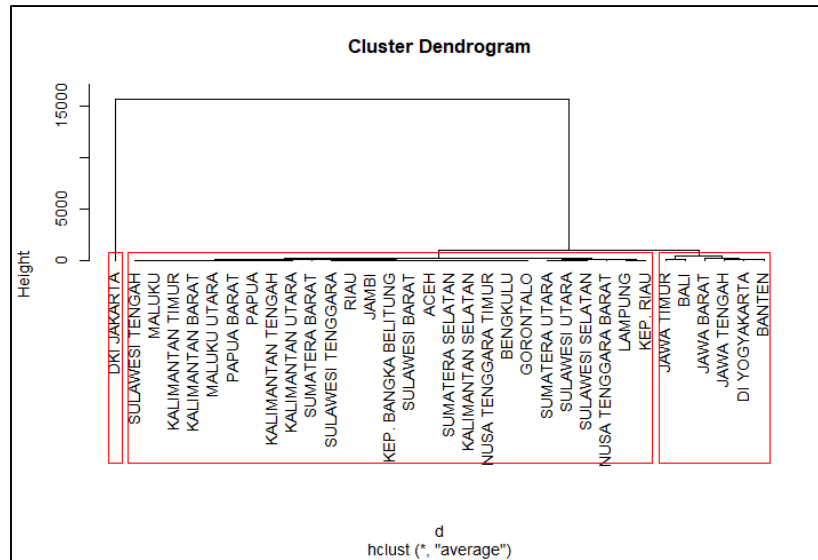
Fig. 1. Dendrogram of Average Linkage Method

3.4 Profiling and Interpretation

One way that can be done for profiling is to use the average value of objects contained in the cluster of each variable. This stage describes the profile or characteristics of each cluster.

Table 6
Profiling

Cluster	People's Welfare Index			
	PD	**PPP**	**LER**	**AYS**
1	103.1	10.997	69.51	8.641
2	15978	4.67	73.06	11.17
3	1090.3	8.823	72.79	8.645

where

■ = high	■ = medium	□ = low

Furthermore, we can conclude that the characteristics of each cluster using the average value of the variables for each cluster, which is as follows:

a) Cluster 1 tends to have a "low" People's Welfare Index status when compared to the other two clusters. This cluster has a "low" level of Population Density (PD), Life Expectancy (LER), and Average Years of Schooling (AYS) with a "high" level of PPP (Percentage of Poor Population).

b) Cluster 2 tends to have a "high" People's Welfare Index status when compared to the other two clusters. This cluster has a "high" level of Population Density (PD), Life Expectancy (LER), and Average Years of Schooling (AYS) with PPP (Percentage of Poor Population) at a "low" level.

c) Cluster 3 tends to have a "medium" People's Welfare Index status when compared to the other two clusters. This cluster has a "medium" level of Population Density (PD), Percentage of Poor Population (PPP), Life Expectancy (LER), and Average Years of Schooling (AYS).

4. Conclusion

Based on Hierarchical Cluster Analysis, the results show that the Average Linkage method is the right method for grouping provinces in Indonesia based on the People's Welfare Indicator in 2021 with a cophenetic correlation value of 0.9990208. By using the average linkage method, the provinces in Indonesia are divided into 3 clusters where the cluster that has a "high" People's Welfare Indicator status is cluster 2, namely DKI Jakarta. While the cluster with the status of "low" People's Welfare Indicators is cluster 1 which totals 27 provinces.

From the research that has been done, the picture of the People's Welfare Indicators in Indonesia in 2021 is still uneven or it can be said that it is only superior in seven provinces. Thus, it is hoped that the Indonesian government can prioritize 27 provinces with relatively low People's Welfare Indicators, namely Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Kep. Bangka Belitung, Kep. Riau, DKI Jakarta, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, and Papua to be given appropriate treatment so that the problem of low levels of People's Welfare Indicators can be resolved.

Author Contributions

Conceptualization, Restu Arisanti; Methodology, Sri Winarni and Restu Arisanti; software, Aissa Putri; validation, Restu Arisanti; formal analysis, Restu Arisanti and Aissa Putri; data curation, Aissa Putri; writing—original draft preparation, Restu Arisanti; writing—review and editing, Resa Septiani Pontoh; visualization, Aissa Putri; supervision, Restu Arisanti. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

References

Bridges Jr, C. C. (1966). Hierarchical cluster analysis. *Psychological reports*, *18*(3), 851-854.

Carvalho, P. R., Munita, C. S., & Lapolli, A. L. (2019). Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficient. *Brazilian Journal of Radiation Sciences*, *7*(2A (Suppl.)). https://www.doi.org/10.15392/bjrs.v7i2A.668

Emmendorfer, L. R., & de Paula Canuto, A. M. (2021). A generalized average linkage criterion for Hierarchical Agglomerative Clustering. *Applied Soft Computing*, *100*, 106990.

Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis.

Köhn, H. F., & Hubert, L. J. (2014). Hierarchical cluster analysis. *Wiley StatsRef: statistics reference online*, 1-13.

Kumar, S., & Toshniwal, D. (2016). Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). *Journal of Big Data*, *3*, 1-11. https://www.doi.org/10.1186/s40537-016-0046-3

Orlóci, L. (1967). An agglomerative method for classification of plant communities. *The Journal of Ecology*, 193-206.

Rencher, A. C. (2002). *Methods of multivariate analysis*. Wiley, New York.

Roux, M. (2018). A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, *35*, 345-366.

Saraçli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of inequalities and Applications*, *2013*(1), 1-8. https://www.doi.org/10.1186/1029-242X-2013-203

Sharma, S., & Batra, N. (2019, February). Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 568-573). IEEE.

Silva, A. R. D., & Dias, C. T. D. S. (2013). A cophenetic correlation coefficient for Tocher's method. *Pesquisa Agropecuária Brasileira*, *48*, 589-596.

Statistics of Indonesia. (2019). Indikator Kesejahteraan Rakyat. Publication of Statistics of Indonesia (bps.go.id).

Statistics of Indonesia. (2022). Statistik Indonesia 2022. Publication of Statistics of Indonesia (bps.go.id).