Clustering Indonesian Provinces Based on Poverty Levels Utilizing the Average Linkage Method with Principal Component Analysis

Paskal Immanuel Kontoro[1], Junaidi Junaidi[1], Nurul Fiskia Gamayanti[1], Arditya Sulistya Ningsih Apusing[1*]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Tadulako University
ardityasulistya6@gmail.com

Abstract. Indonesia grapples with the pervasive issue of poverty that undermines the well-being of its citizens. Recognizing the diverse characteristics of each region in Indonesia, effective poverty alleviation policies must be tailored through a nuanced approach. Therefore, this research is crucial as it aims to employ advanced clustering techniques, specifically the Average Linkage Method with Principal Component Analysis, to discern the characteristics of poverty across Indonesian provinces. Hierarchical cluster analysis with average linkage is deemed more stable. In this cluster analysis, two assumptions must be met: the assumption of sample adequacy and multicollinearity. In cases of multicollinearity violations, Principal Component Analysis is applied for resolution. This research utilizes secondary data from the Central Statistics Agency (BPS), examining factors such as the percentage of impoverished people, poverty depth and severity indices, human development index, and average and expected length of schooling to assess poverty levels. From the research results, 2 clusters were obtained. The first cluster has low poverty levels, consisting of 31 provinces excluding Papua, West Papua and East Nusa Tenggara. Those three provinces are classified as cluster 2, with high poverty levels. This research offers vital insights for policymakers, facilitating targeted policies aligned with SDGs for effective poverty reduction strategies.

Keywords: Poverty, Clustering, Average Linkage, Principal Component Analysis.

1 Introduction

In the era of globalization, Indonesia has committed itself to achieving the Sustainable Development Goals (SDGs) set by the United Nations (UN). One of the most fundamental goals of the SDGs is to eradicate poverty in all its forms and dimensions. Despite being an archipelagic country rich in social, cultural, and economic diversity, poverty remains a critical challenge for Indonesia. According to the Central Statistics Agency (BPS) data, the poverty rate in September 2022 reached 9.57%, indicating an increase of 0.03% compared to March 2022 [1]. This rise suggests that the implemented

Y. Yuyun et al. (eds.), *Proceedings of the 5th International Seminar on Science and Technology (ISST 2023)*,
Advances in Physics Research 10,
https://doi.org/10.2991/978-94-6463-520-1_13

strategies have yet to reach the desired targets. Therefore, a profound understanding of the poverty levels in each province is essential for designing appropriate policies to support the achievement of the SDGs.

Addressing the inequality in poverty levels among provinces becomes a focal point in achieving the first SDG target in Indonesia. Academic research serves as a crucial instrument to delve into comprehensive and objective information regarding the variability of poverty across regions. Cluster analysis, particularly the hierarchical method, emerges as a relevant approach to formulating more targeted strategies and interventions. Despite having certain assumptions, such as good sample representation and the absence of multicollinearity, hierarchical cluster analysis can provide a clear visual representation of the interrelationships among clusters [2]. Violations of multicollinearity assumptions can be addressed using Principal Component Analysis (PCA), simplifying variables without losing significant information [3].

In the context of clustering, the average linkage method in hierarchical cluster analysis appears as an approach that considers minor variations in merging objects into clusters and is more stable compared to other types of hierarchical cluster analysis [4]. The factors used as the basis for clustering reflect the dimensions of poverty measurement and overall aspects of human well-being. The importance of predictor variables such as the percentage of the population in poverty, poverty depth and severity indices, the Human Development Index (HDI), average years of schooling, and expected years of schooling highlight economic, educational, and health dimensions.

This research can provide deeper insights into implementing poverty alleviation policies in Indonesia, which faces unique challenges in achieving the poverty eradication goal by 2030 [5]. Therefore, this research will utilize the Average Linkage Method with Principal Component Analysis (PCA) to cluster the provinces in Indonesia based on their poverty levels.

2 Data and Method

The Average Linkage Method is a clustering process based on the average distance between objects [6]. The Average Linkage Method is more stable than other hierarchical methods in various conditions. The algorithm within this method is as follows [7].

1. Determines the object corresponding to the closest distance in the distance matrix $D = \{d_{ik}\}$. After that, the obtained cluster (UV) is obtained by combining the corresponding U and V objects. Next, calculate the distance between the cluster $(UV)$and the unconnected object (W) using the following formula.

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{UV} N_W} \tag{1}$$

where:

$d_{(UV)W}$: Number of objects in the cluster (UV)
d_{ik} : Distance between objects i in the (UV) cluster dan k in the W *cluster*
N_{UV} : Number of objects in the cluster (UV)
N_W : Number of objects in the cluster (W)

2. Recalculate the new distance matrix by deleting rows and columns corresponding to clusters U and V, adding rows and columns for the distance between clusters (UV) and the remaining clusters.
3. Repeat step 2 until it forms into 1 cluster.

Principal Component Analysis (PCA) is a multivariate analysis technique that transforms correlated original variables into new uncorrelated variables by reducing the variables in such a way as to absorb most of the original data variance without eliminating the contained information [8]. The presence of correlated variables is the main principle of PCA. The principal components can be selected through three methods: choosing components with a cumulative explained variance of 80%, selecting eigenvalues greater than 1, and examining the scree plot. However, some experts recommend prioritizing the selection of principal components based on eigenvalues exceeding 1. The PCA algorithm for standardized data based on the correlation matrix is as follows [9].

1. Determining the matrix Z containing standardized data from variable X.
2. Computing the correlation matrix R using the following model.
$$R = \frac{1}{n-1} Z'Z \tag{2}$$
3. Determining the eigenvalues $\lambda_1, \lambda_2, \lambda_3, \cdots, \lambda_p$ using the following model.
$$|R - \lambda I| = 0 \tag{3}$$
4. Form the principal components (PC) based on the following model.
$$PC_i = \hat{y}_i = \hat{e}_i z = \hat{e}_{i1} z_1 + \hat{e}_{i1} z_1 + \cdots + \hat{e}_{ip} z_p \qquad i : 1,2,\cdots,p \tag{4}$$
where $\hat{e}_i$ is the eigenvector obtained from each eigenvalue λp that satisfies the equation:
$$(R - \lambda I)\hat{e}_i = 0 \tag{5}$$

The data used in this research is secondary data obtained from the Central Statistics Agency website. This research uses data on poverty levels in 2022 based on each province in Indonesia. The author analyzed six variables in this research as follows.

Table 1. Description of Variables

Variables	Description
X_1	Percentage of Poor Population
X_2	Poverty Depth Index
X_3	Poverty Severity Index
X_4	Human Development Index
X_5	Average Years of Schooling
X_6	Expected Years of Schooling

Data analysis in this research used the Average Linkage method with Principal Component Analysis (PCA) assisted by RStudio software. The following are the analysis stages carried out in this research:

1) Data collection and inputting data to RStudio.
2) Analyzing descriptive statistics.

3) Standardization of data by transforming it into z-scores.
4) Testing assumptions of sample adequacy and multicollinearity.
5) Application of the PCA method to address multicollinearity issues.
6) Determination of similarity measures using Euclidean distance.
7) Application of the average linkage algorithm.
8) Determination of the optimal cluster number using silhouette width.
9) Division of clusters based on the optimal number of cluster.
10) Obtaining cluster results from average linkage analysis with PCA.
11) Profiling groups using mean values.
12) Interpretation and conclusion.

3 Results and Discussion

a) Descriptive Statistical Analysis

Descriptive analysis was conducted to comprehensively examine the characteristics and general description of each indicator or variable utilized in the study. The following are the results of descriptive statistical analysis:

Table 2. Descriptive Statistical Analysis Results

No	Variable	Min	Max	Mean	Standard Deviation
1	Percentage of Poor Population	4,45	26,56	10,24	5,25
2	Poverty Depth Index	0,60	6,16	1,80	1,22
3	Poverty Severity Index	0,13	2,10	0,48	0,41
4	Human Development Index	61,39	81,65	71,97	3,90
5	Average Years of Schooling	7,02	11,31	8,84	0,92
6	Expected Years of Schooling	11,14	15,65	13,24	0,74

b) Data Standardization

The data used in this research has different units of measurement, so it is necessary to standardize the data before further analysis for more valid results. The results of data standardization can be seen in the Table 3 below.

Table 3. Data Standardization

Province	X_1	X_2	X_3	X_4	X_5	X_6
Aceh	0,84	0,56	0,33	0,21	0,65	1,52
North Sumatera	-0,35	-0,36	-0,33	0,19	0,94	0,09
West Sumatera	-0,82	-0,82	-0,76	0,33	0,37	1,16
⋮	⋮	⋮	⋮	⋮	⋮	⋮
West Papua	2,11	2,47	2,72	-1,56	-1,08	-0,04
Papua	3,11	3,57	3,93	-2,71	-1,97	-2,84

c) Cluster Analysis Assumption Test

1. Sample Adequacy Test

Table 4. Kaiser Meyer Olkin (KMO) Result

Variable	X_1	X_2	X_3	X_4	X_5	X_6	Overall
KMO	0,57	0,58	0,58	0,80	0,80	0,77	0,64

Table 4 shows that the overall KMO value is 0,64. Based on the KMO values above, KMO> 0,5, it can be concluded that the sample used in this research is representative, which means the sample represents the population, so it is appropriate to carry out cluster analysis.

2. Multicollinearity Test

Table 5. Bartlett's Test of Sphericity

Chi-Square	Df	P-value
45,111	15	$2{,}2 \times 10^{-16}$

Based on Table 5, the results of the multicollinearity test show p-value = $2{,}2 \times 10^{-16}$, which means that with a significance level of 5%, the p-value ($2{,}2 \times 10^{-16}$) < α(0,05). It indicates that there are symptoms of multicollinearity in the variables X_1, X_2, X_3, X_4, X_5, X_6. Therefore, we can proceed to principal component analysis (PCA) testing to overcome multicollinearity violations.

d) Principal Component Analysis

Table 6. Kolmogorov-Smirnov Test

Component	1	2	3	4	5	6
Eigen Values	4,03	1,26	0,50	0,17	0,04	0,00

Based on the principal component analysis results in Table 6, the number of PCs formed was two PCs, namely PC1 and PC2 because only these components have eigen values > 1. The distance is measured after obtaining the formed PC for cluster analysis.

e) Distance Measures

Table 7. Estimated Value of Dispersion

Province	Aceh	North Sumatera	⋯	Banten	⋯	Papua
Aceh	0	2,04	⋯	2,76	⋯	7,58
North Sumatera	2,04	0	⋯	0,75	⋯	8,31
West Sumatera	2,38	0,72	⋯	1,02	⋯	9,02
Riau	2,48	0,45	⋯	0,34	⋯	8,50
Jambi	2,73	0,94	⋯	0,50	⋯	7,98
South Sumatera	2,78	1,88	⋯	1,78	⋯	6,71

Province	Aceh	North Sumatera	⋯	Banten	⋯	Papua
Bengkulu	1,05	1,67	⋯	2,23	⋯	6,94
Lampung	2,79	1,84	⋯	1,73	⋯	6,76
Bangka Belitung Islands	4,22	2,39	⋯	1,68	⋯	8,35
Riau Islands	2,71	1,00	⋯	1,09	⋯	9,30
DKI Jakarta	3,53	2,37	⋯	2,52	⋯	10,64
⋮	⋮	⋮	⋱	⋮	⋮	⋮
North Kalimantan	2,82	0,82	⋯	0,08	⋯	8,44
North Sulawesi	2,57	0,53	⋯	0,30	⋯	8,57
Central Sulawesi	1,66	1,72	⋯	2,09	⋯	6,63
South Sulawesi	2,12	0,44	⋯	0,70	⋯	7,95
Southeast Sulawesi	1,34	0,78	⋯	1,43	⋯	7,73
Gorontalo	2,46	2,86	⋯	3,12	⋯	5,47
West Sulawesi	2,96	2,49	⋯	2,47	⋯	6,01
Maluku	0,59	2,47	⋯	3,15	⋯	7,07
North Maluku	2,40	0,38	⋯	0,37	⋯	8,38
West Papua	4,35	5,47	⋯	5,82	⋯	3,36
Papua	7,58	8,31	⋯	8,46	⋯	0

Based on Table 6, it can be seen that one of the values we marked in red indicates that North Kalimantan Province and Banten Province have the closest distance with a distance of 0,08. It indicates that North Kalimantan Province and Banten Province have similar characteristics in terms of poverty levels.

f) Clustering Utilizing the Average Linkage Method

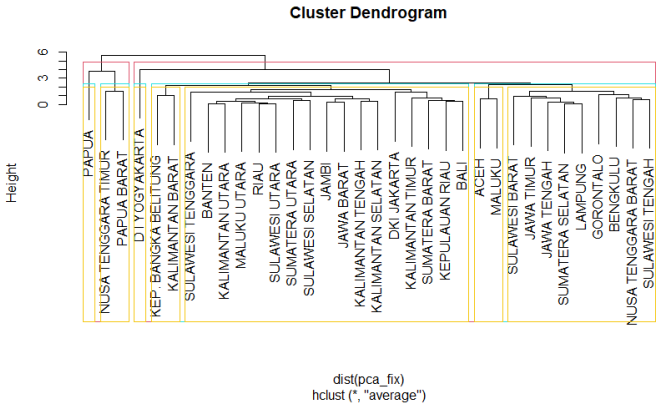

Fig. 1 Dendogram Clustering Result

The dendrogram in Figure 1 illustrates that a group is formed by combining objects that are closest in distance, resulting in a cluster containing objects with similar characteristics. Based on the clustering results using the average linkage method, it is evident that 2 clusters, 5 clusters, or even 7 clusters can be formed. Therefore, it becomes imperative to ascertain the optimal number of clusters to be established.

g) Determination of The Optimum Number of Clusters

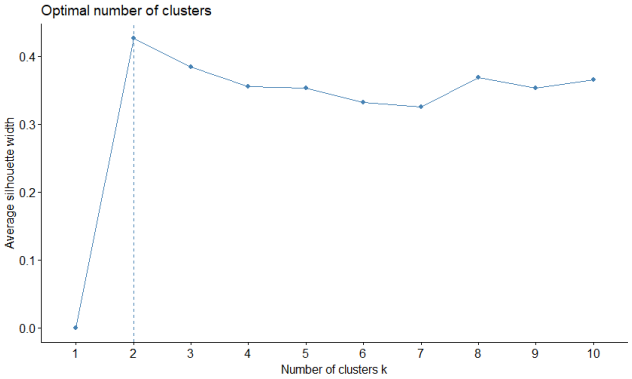

Fig. 2 The Optimum Number of Clusters with Silhouettes Method

Based on Figure 2, it can be seen that the average silhoutte width obtained has the most significant value at number 2. This indicates that the optimum number of clusters formed in this reseacrh is 2 clusters.

h) Cluster Divison with the Optimum Number of Cluster

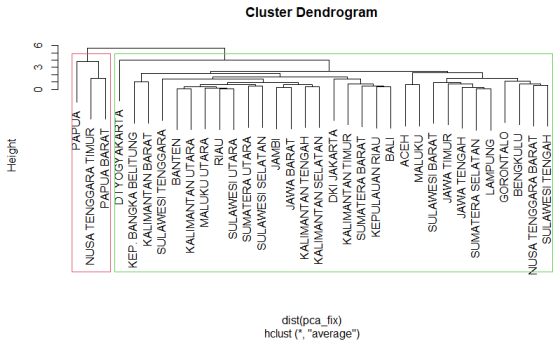

Fig. 3 Dendogram of Cut-Off Results

Based on Figure 3, it can be formed based on the cut-off results we have carried out; 2 clusters were formed. Cluster 1 consists of 31 provinces, namely Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, West Java, Central Java, D I Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, West Kalimantan, Central Kalimantan,

South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, and North Maluku, which are colored green in the box, and Cluster 2 consists of 3 provinces, which are Nusa Tenggara Timur, Papua Barat dan Papua, which is colored pink in the box.

i) Group Profiling

Table 8. Characteristics of Each Cluster Based on Mean Values

Cluster	X_1	X_2	X_3	X_4	X_5	X_6	Average of X_1, X_2, X_3	Average of X_1, X_2, X_3
1	9,04	1,51	0,37	72,70	8,97	13,31	3,64	31,66
2	22,65	4,87	1,54	64,39	7,52	12,52	9,69	28,14

In Table 8, the characteristics of each cluster are delineated, with the red color highlighting the highest values. Variables X_1, X_2, X_3 exhibit a positive association with poverty, signifying that as these variables increase, poverty levels also increase. Conversely, X_4, X_5, X_6 are negatively associated with poverty, meaning higher average values of X_4, X_5, X_6 correspond to lower poverty levels.

9 In Cluster 1, the average overall result from X_1, X_2, and X_3 is the lowest, while the average overall result from X_4, X_5, and X_6 is the highest. It suggests that Cluster 1 has a low poverty level. On the other hand, in Cluster 2, the average overall result from X_1, X_2, and X_3 is the highest, and the average overall result from X_4, X_5, and X_6 is the lowest. It indicates that Cluster 2 has a high poverty level.

4 Conclusion

Based on the results and discussion, it can be concluded that two distinct groups can be identified based on the characteristics of the predictor variables. Cluster 1, characterized by a low poverty level, encompasses 31 provinces, namely Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, and North Maluku. Meanwhile, Cluster 2, characterized by a high poverty rate, comprises 3 provinces: East Nusa Tenggara, West Papua, and Papua. Through the utilization of the Average Linkage method with Principal Component Analysis (PCA) and comprehensive predictor variables, this research offers a more nuanced understanding of the poverty dynamics in each province. This insight can serve as the foundation for developing region-specific policies to reduce poverty levels and achieve Indonesia's first Sustainable Development Goals (SDGs) target.

References

[1] [BPS] Badan Pusat Statistik. (2023). Badan Pusat Statistik Republik Indonesia. Jakarta.

[2] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning.

[3] Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer.

[4] Johnson, R. A., & Wichern, D. W. (2014). Applied Multivariate Statistical Analysis (6th ed.). Pearson.

[5] Iis, Yahya, I., Wibawa, G. N., Baharuddin, Ruslan, & Laome, L. (2022). Penggunaan Korelasi Cophenetic Untuk Pemilihan Metode Cluster Berhierarki Pada Mengelompokkan Kabupaten/Kota Berdasarkan Jenis Penyakit Di Provinsi Sulawesi Tenggara Tahun 2020. Prosiding Seminar Nasional Sains dan Terapan (SINTA) VI.

[6] Ramadani, R., & Salma, A. (2022). Metode Average Linkage Dan Ward Dalam Pengelompokan Kesejahteraan Sumatera Barat. Journal Of Mathematics UNP , Vol. 7, No. 3, pp. 11-24.

[7] Hasrul, M. (2018). Analisis Klaster Untuk Pengelompokkan Kabupaten/Kota Di Provinsi Sulawesi Selatan Berdasarkan Indikator Kesejahteraan Rakyat. Makassar: Skripsi. Fakultas Sains Dan Teknologi Universitas Islam Negeri (Uin) Alauddin.

[8] Utami, N. D. (2017). Perbandingan Hasil Pengelompokkan Antara Metode Average Linkage, Ward, Complete Linkage, Dan Single Linkage (Studi Kasus: Indikator Kesehatan Indonesia Tahun 2015). Yogyakarta: Skripsi. Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Islam Indonesia.

[9] Sriningsih, M., Hatidja, D., & Prang, J. D. (2018). Penanganan Multikolinieritas Dengan Menggunakan Analisis Regresi Komponen Utama Pada Kasus Impor Beras di Provinsi Sulut. *Jurnal Ilmiah Sains,* Vol.18 No.1, 18-24.