JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 4, No 01, March 2025
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

Comparison of Poverty Clustering Results based on Distance Measurement with the Complete Linkage Method in Indonesia

Fira Anggraini[1], Devni Prima Sari[1]*
[1] Departement of Mathematics, Universitas Negeri Padang, Indonesia
* Corresponding Author: devniprimasari@fmipa.unp.ac.id

Article Info
Article History:
Received: 01 03 2025
Revised: 17 04 2025
Accepted: 23 04 2025
Available Online: 24 04 2025

Key Words:
Cluster
Distance
Complete linkage
Poverty
Standard deviation

Abstract
Every year, population growth in Indonesia increases and has the potential to trigger poverty. Poverty indicators include the number of poor people, per capita expenditure, human development index, average years of schooling, and unemployment. The clustering of regions is necessary for the government to be more effective in development. One of the methods used is cluster analysis, a statistical technique that groups objects based on similar characteristics. This research compares the results of clustering poverty in Indonesia's Regency/City in 2023 using the complete linkage method, which is based on the farthest distance. The distances analyzed include Euclidean, Square Euclidean, Manhattan, and Minkowski, resulting in two clusters at each distance. Minkowski proved to be the best distance with the smallest standard deviation ratio, which was 1.518 for cluster 1 and 2.225 for cluster 2, compared to the other distances. These results show that the Minkowski method is superior in clustering poverty areas in Indonesia.

1. INTRODUCTION

Indonesia is one of the most populous countries in the world. According to the Central Bureau of Statistics (BPS), Indonesia's total population reached 281.6 million in June 2024. The increase in population growth rate that takes place consistently every year is projected by BPS to reach 315 million people in 2035, with an annual growth rate of around 1,11%. This rapid increase in population can trigger various problems. Based on Malthus' theory, high population growth in a country risk triggering chronic poverty. And explains that the population tends to increase with a measuring series pattern [1].

Indonesia has a wealth of potential resources, such as mining (e.g. coal), minerals, plantations, livestock, fisheries, tourism, and other sectors. However, much of this potential has not been optimally utilized. As a result, there is still a large proportion of the community living in conditions of poverty. Poverty can be defined as a situation in which a person or group is unable to meet basic needs, such as food, clothing, shelter, education, and access to health services, which are considered basic needs according to certain standards [2].

There are various indicators of poverty in Indonesia, namely, the poor population, adjusted per capita expenditure, human development index, average years of schooling, and unemployment rate are important indicators. Therefore, grouping the characteristics of each region is necessary so that the government can more easily achieve success in regional development in Indonesia. To understand the condition of individuals and groups, grouping is needed based on the poverty rate in each district/city in Indonesia. One approach that can be applied to group data is cluster analysis. Cluster analysis is a method that aims to group certain objects into groups that have similarities or similar characteristics. Objects in one group usually have similarities, while objects in different groups show striking differences. [3].

In cluster analysis, there are various methods of clustering, such as single linkage, complete linkage, average linkage, and centroid linkage [4]. This research applies the complete linkage method, by grouping objects based on the farthest distance between groups. The complete linkage method is based on the farthest distance. This method measures the similarity or dissimilarity of objects based on distance, such as Euclidean, square Euclidean, Minkowski, and Manhattan. The evaluation is done by comparing clustering results based on distance measurements to determine the most effective method.

This study aims to identify the comparison of poverty clustering results based on distance measurement and to determine the best distance used after comparing the results of poverty clustering based on distance measurement using the complete linkage method in Indonesia. The selection of the method with the best clustering quality is done by considering the lowest value of the ratio of the average standard deviation within clusters (Sw) and the highest value of the standard deviation between clusters (Sb) [5].

1.1 Data Sufficiency Test

It is important to ensure that the data used is representative enough to represent the population. Therefore, the KMO (Keiser-Meyer-Olkin) test was conducted using SPSS [6]. Test statistics:

$$KMO = \frac{\sum_{i=1}^{n}\sum_{J=1}^{n} {r_{ij}}^2}{\sum_{i=1}^{n}\sum_{J=1}^{n} {r_{ij}}^2 + \sum_{i=1}^{n}\sum_{J=1}^{n} {a_{ij}}^2} \quad (1)$$

With $r_{ij} =$ correlation coefficient between variables *i* and *j*, a_{ij} = partial correlation coefficient between variables *i* and *j*

1.2 Multicollinearity Test

This step aims to determine whether there is a relationship between variables. VIF is used as a tool to detect multicollinearity in clusters that include more than two independent variables. If the VIF value exceeds 10, it indicates a significant multicollinearity problem. Test statistics:

$$VIF_J = \frac{1}{1 - R_j^2} \quad (2)$$

With R_j^2 = coefficient of determination between X_j and independent variables, j = 1, 2, ...

1.3 Data Standardization

Data needs to be standardized if it has different units. The standardization process is usually done using a Z-score. To calculate data standardization can be done with the following formula [8]:

$$Z = \frac{x_i - \bar{x}}{S_x} \quad (3)$$

With Z = nilai z-score, x_i= nilai sampel ke-I, $\bar{x}$ = nilai mean, S_x= Standar deviasi

1.4 Types of Distance Measurements

There are 4 types of distance measurements, namely Euclidean distance, square Euclidean, Manhattan, Minkowski. First Euclidean distance, the Euclidean distance measure between two objects $x' = [x_1, x_2, ..., x_p]$ and $y' = [y_1, y_2, ..., y_p]$ whose dimension is p is [9]:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - y_{jk})^2} \quad (4)$$

Second, the square Euclidean distance formula is as follows [10]:

$$d_{ij} = \sum_{k=1}^{p} \left(x_{ik} - y_{jk}\right)^2 \quad (5)$$

Third, the Manhattan distance can be expressed as follows [11]:

$$d_{ij} = \sum_{i=1}^{p} \left|x_{ik} - y_{jk}\right| \quad (6)$$

Fourth, Minkowski distance is a generalization of Euclidean and Cityblock distance. If m = 2, this distance changes to Euclidean distance, while if m = 1, this distance becomes Cityblock distance. The Minkowski distance can be expressed as follows:

$$d_{ij} = \left[\sum_{i=1}^{p} \left|x_{ik} - y_{jk}\right|^m\right]^{\frac{1}{m}} \quad (7)$$

Minkowski distance places a stronger emphasis on the differences between coordinates if the value of m > 1 [12]. With d_{ij} = distance between *i*-th and *j*-th objects, x_{ik} = observation value of the *i*-th object of the *k*-th variable, y_{jk} = observation value of the *j*-th object of the *k*-th variable, p = number of variables, m = parameters

1.5 Complete Linkage Method

In the complete linkage method stage, the first step is to choose the closest distance between the clusters in $D = \{d_{uv}\}$. Next, the objects are combined by considering the farthest distance, for example, the object is represented by cluster u and cluster v to form a combined cluster (uv). This results in the following formula [13]:

$$d_{(uv)w} = \max\left(d_{uw}, d_{vw}\right) \quad (8)$$

1.6 Determining the Best Distance

Complete cluster analysis with various distance measures or cluster methods, and then compare the results [14]. The selection of the method with the best clustering results is done by taking into account the lowest value of the average ratio of within-cluster standard deviation (Sw) and the highest value of between-cluster standard deviation (Sb) [5]. The formula for calculating the within-cluster standard deviation (Sw) is:

$$s_w = \frac{1}{c}\sum_{k=1}^{c} s_k \quad (9)$$

$$s_k = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{i(k)} - \bar{x}_{(k)}\right)^2}{n-1}} \quad (10)$$

The standard deviation between clusters can be formulated as in equation

$$s_b = \sqrt{\frac{1}{c-1}\sum_{k=1}^{p}(\bar{x}_{(k)} - \bar{x}^2)} \tag{11}$$

Selection of the best method with the formula:

$$S = \frac{s_w}{s_b} \tag{12}$$

2. METHOD

The data used is secondary data that includes poverty indicators based on districts/cities in Indonesian provinces in 2023. This data is obtained from the Central Bureau of Statistics. The observation unit in this study covers 491 districts/cities out of a total of 514 districts/cities in Indonesia. Due to regional expansion, the available and accessible data only covers 491 districts/cities. The variables applied in this study are the percentage of poor, adjusted per capita expenditure (Thousand Rupiah), human development index, average years of schooling (Years), and percentage of unemployment.

The stages taken in this research are: (1) Collecting the data to be analyzed. (2) Carry out the data feasibility test using the KMO (Keyser Meyer Olkin) test. (3) Running the multicollinearity assumption test to find out if there is a relationship between variables. (4) Perform data standardization if there are unit differences between data using Z-Score. (5) Performing distance measurement calculations. (6) Comparing distance measurement calculations. (7) Forming district/city clusters in Indonesia using the complete linkage method. (8) Determining the total clusters and cluster members formed. (10) Interpreting the results of the clusters formed. (11) Determining the best distance to use. (12) Drawing conclusions.

3. RESULTS AND DISCUSSION

3.1 District/City Grouping in Indonesia

The clustering of regencies/cities in Indonesia was conducted using the hierarchical cluster analysis method. The process is:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.803
Bartlett's Test of Sphericity	Approx. Chi-Square	1494.090
	df	10
	Sig.	.000

Figure 1. KMO

In Figure 1 with a KMO value of 0.803 which is ≥ 0.5, the conclusion is accepted. This indicates that the available data can represent the population well. Therefore, the data meets the criteria for cluster analysis and can be used in further analysis.

Table 1. Multicollinearity Test

Variable	N	Min	Max	Mean	Sd
Percentage of Poor	491	2.27	41.42	11.66	7.20
Per Capita Expenditure (Thousand Rupiah)	491	4597	24975	11034.83	2791.75
Human Development Index	491	35.19	88.28	71.21	6.41
Average Years of Schooling (Years)	491	1.92	13.04	8.62	1.62
Percentage of Unemployment	491	0.41	35.83	4.67	2.99

In Table 1, it can be concluded that the tolerance value > 0.10 and the VIF value < 10.00, which indicates the absence of multicollinearity.

3.2 Hierarchical Cluster Analysis on Poverty Indicators

Table 2. Data Standardization

Regency	Z Score				
	X_1	X_2	X_3	X_4	X_5
Aceh Barat	0.86	-0.34	0.26	0.84	0.47
Aceh Barat Daya	0.52	-0.71	-0.49	0.09	-0.20
Aceh Besar	0.24	-0.26	0.52	1.08	1.17
Aceh Jaya	0.11	-0.21	-0.05	0.07	-0.60
Aceh Selatan	0.06	-0.83	-0.43	0.18	0.02
Aceh Singkil	1.04	-0.59	-0.17	0.05	0.72
Aceh Tamiang	0.12	-0.71	0.02	0.38	0.85
Aceh Tengah	0.38	0.10	0.53	0.79	-0.08
Aceh Tenggara	0.11	-0.88	-0.03	0.91	0.11
Aceh Timur	0.24	-0.57	-0.30	-0.09	1.12
Aceh Utara	0.69	-0.74	-0.05	0.14	0.80
Banda Aceh	-0.64	2.32	0.51	2.73	1.12
Bener Meriah	0.92	0.29	0.38	0.93	-0.74
Bireuen	0.06	-0.46	-0.36	0.43	-0.18
Gayo Lues	0.99	-0.58	2.41	-0.12	-0.68

Table 2 pieces of data standardization table. The data standardization process is needed if there are different units in the data to be analyzed. In this study, because there are different units in the data, standardization needs to be done.

Table 3. Euclidean Distance Matrix

d	1	2	3	4	...	491
1	0.000	1.348	1.004	1.552	...	4.575
2	1.348	0.000	2.034	0.886	...	3.597
3	1.004	2.034	0.000	2.115	...	5.381
4	1.552	0.886	2.115	0.000	...	4.188
...	...	...	...	...	...	...
491	4.575	3.597	5.381	4.188	3.876	0.000

Table 3 pieces of the Euclidean distance matrix table. It can be seen that the closest distance is between Southwest Aceh (D2) and South Aceh (D5) 0.535, while the farthest distance is between Aceh Besar (D3) and Aceh Jaya (D4) 2.115.

Table 4. Square Euclidean Distance Matrix

d	1	2	3	4	...	491
1	0.000	2.874	1.896	3.035	...	9.967
2	2.874	0.000	4.095	1.784	...	7.768
3	1.896	4.095	0.000	3.526	...	11.863
...	...	...	...	...	...	...
491	9.9671	7.768	11.86	8.7126	...	0.000

Table 4 pieces of square Euclidean distance matrix table. It can be seen that the smallest distance is found between Aceh Barat Daya (D2) and Aceh Selatan (D5), with a distance of 0.286. In contrast, the largest distance is found between Aceh Besar (D3) and Aceh Jaya (D4), which reaches 4.472. In general, the smaller the distance value, the greater the similarity in characteristics between the data being compared. Conversely, the larger the distance, the lower the similarity in characteristics.

Table 5. Manhattan Distance Matrix

d	1	2	3	4	...	491
1	0.000	1.818	1.007	2.410	...	20.933
2	1.818	0.000	4.136	0.786	...	12.941
3	1.007	4.136	0.000	4.472	...	28.960
4	2.410	0.786	4.472	0.000	...	17.536
...	...	...	...	...	...	...
491	20.933	12.941	28.960	17.536	...	0.000

Table 5 pieces of the Manhattan distance matrix table. The shortest distance was found between Aceh Barat Daya (D2) and Aceh Selatan (D5), with a value of 0.942. Meanwhile, the farthest distance was recorded between Aceh Barat Daya (D2) and Aceh Besar (D3), which amounted to 4.095. In general, the smaller the distance value between two data, the more similar their characteristics are, and conversely, the larger the distance, the lower the similarity of characteristics between the data being compared.

Table 6. Minkowski Distance Matrix

d	1	2	3	4	...	491
1	0.000	0.408	0.206	0.708	...	6.002
2	0.408	0.000	1.547	0.118	...	3.520
3	0.206	1.547	0.000	2.238	...	9.352
4	0.708	0.118	2.238	0.000	...	5.441
...	...	...	...	...	...	...
491	15.388	7.537	24.153	12.395	...	0.862

Table 6 pieces of the Minkowski distance matrix table. The closest distance was identified between Aceh Barat Daya (D2) and Aceh Selatan (D5) with a value of 0.037, while the farthest distance was found between Aceh Besar (D3) and Aceh Jaya (D4) which reached 2.238. The smaller the distance between data, the less similarity between them, and the larger the distance, the higher the level of similarity of characteristics between data.

3.3 Number of Cluster Members

Where for determining the number of clusters there are no definite rules, but there are several considerations as guidelines as follows [3]:

a. Theories, concepts, models, or practical considerations, which can provide direction to determine how many clusters.

b. In hierarchical clustering, distance can be used as a criterion.
c. The relative number of cluster members can also be taken into consideration.
Therefore in this cluster, 2 clusters were selected to form each cluster member.

3.4 Characteristics of Each Group

Table 7. Characteristics of each group

variable	Euclid		Euclid Square		Manhattan		Minkowski	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
X1	13.012	6.153	13.012	6.153	13.121	6.882	11.682	7.547
X2	10102.231	14822.918	10102.231	14822.918	9971.556	14471.164	11003.318	16161.000
X3	69.264	79.138	69.264	79.138	68.964	78.457	71.183	76.283
X4	8.079	10.814	8.079	10.814	8.010	10.570	8.605	10.913
X5	4.072	7.119	4.072	7.119	3.940	3.266	4.525	28.922

After determining the number and composition of each cluster, the next step is to perform profiling to obtain the characteristics of each cluster. The average calculation for each cluster can be found in Table 7.

Based on Table 7, the characteristics of each cluster can be seen through the average in each cluster, with the following interpretations: (1) At the Euclid, Euclid Square, Manhattan, and Minkowski distances, the average poverty rate (X_1) for cluster 1 is higher than cluster 2. (3) At the Euclid, Euclid Square, Manhattan, and Minkowski distances, the average per capita expenditure (X_2), average human development index (X_3), and the average length of schooling (X_4) are lower for cluster 1 than cluster 2. (4) At the Euclid, Euclid Square, and Minkowski distances, the average unemployment rate (X_5) is lower for cluster 1 than cluster 2. However, at the Manhattan distance, the average unemployment rate (X_5) for Cluster 1 is higher than Cluster 2.

3.5 Determination of the Best Distance

The results of calculations involving the standard deviation within clusters, the standard deviation between clusters, and the comparison ratio between the two are presented in Table 8.

Table 8. Best Distance Determination

	Euclid		Square Euclid		Manhattan		Minkowski	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Mean	2226.199		2226.199		2226.199		2226.199	
Clustering Means	2039.331	2985.228	2039.331	2985.228	2013.118	2914.820	2219.863	3256.933
s_k	4124.970	6028.704	4124.970	6028.704	4069.048	5885.527	4563.791	6690.206
s_w	2062.485	3014.352	2062.485	3014.352	2034.524	2942.763	2281.895	3345.103
s_b	1346.143		1346.143		1284.444		1503.602	
S	1.532	2.239	1.532	2.239	1.584	2.291	1.518	2.225

Based on Table 8, it can be seen that the standard deviation ratio for various distance methods, namely Euclid, Euclid Square, Manhattan, and Minkowski. The Minkowski distance shows the smallest standard deviation ratio, which is 1.518 for cluster 1 and 2.225 for cluster 2. This indicates that clustering poverty based on districts/municipalities in Indonesia using the Minkowski distance provides more precise and quality clustering results compared to other distance methods.

Based on the results of the research that has been carried out, namely the comparison of Euclid, Euclid Square, Manhattan, and Minkowski distances in hierarchical cluster analysis in grouping Districts/Cities in Indonesia, the results explained are: (1) The Euclid and Euclid Square distances have the same cluster results, namely in cluster 1 with 394 Districts/Cities and cluster 2 with 97 Districts/Cities. However, it is different from the cluster results on the Manhattan and Minkowski distances, for the Manhattan distance with cluster 1 members of as many as 375

regencies/cities and cluster 2 with 116 members, and for the Minkowski distance with cluster 1 members as many as 488 regencies/cities and cluster 2 as many as 3 regencies/cities. (2) At the Euclid, Euclid Square, Manhattan, and Minkowski distances, the average poverty rate (X_1) for cluster 1 is higher than cluster 2. (3) At the Euclid, Euclid Square, Manhattan, and Minkowski distances, the average per capita expenditure (X_2), average human development index (X_3), and average length of schooling (X_4) are lower for cluster 1 than cluster 2. (4) At the Euclid, Euclid Square, and Minkowski distances, the average unemployment rate (X_5) is lower for cluster 1 than cluster 2. However, at the Manhattan distance, the average unemployment rate (X_5) for cluster 1 is higher than cluster 2.

4. CONCLUSION

Based on the results and analysis conducted in this study, it can be concluded that first, the comparison of the results of poverty clustering in Indonesia using the complete linkage method with various types of distances (Euclid, Euclid Square, Manhattan, and Minkowski) shows different results, namely: (1) When using the Euclid and Euclid Square distances, the results show that there are 394 districts/municipalities in cluster 1. Cluster 2 has 97 districts/municipalities. (2) Using Manhattan distance, cluster 1 includes 375 districts/municipalities and the second cluster has 116 districts/municipalities. (3) With the Minkowski distance, the first cluster includes 488 districts/municipalities, while the second cluster consists of only 3 districts/municipalities.

Second, the comparison of the results of distance measurement with the complete linkage method in clustering poverty in Indonesia based on cluster characteristics shows the following differences: (1) Using the Euclid, Euclid Square, Manhattan, and Minkowski distances, the average poverty rate (X_1) of the first cluster is higher than that of the second cluster. (2) At Euclid, Euclid Square, Manhattan, and Minkowski distances, the average per capita expenditure (X_2), average human development index (X_3), and average length of schooling (X_4) are lower for cluster 1 than cluster 2. (3) At Euclid, Euclid Square, and Minkowski distances, the average unemployment rate (X_5) is lower for the first cluster than the second cluster. However, at Manhattan distances, the average unemployment rate (X_5) for the first cluster is greater than the second cluster.

Third, from the comparison of standard deviation ratios conducted on Euclid, Euclid Square, Manhattan, and Minkowski distances, it is known that the Minkowski distance has the lowest standard deviation ratio, which is 1.518 for cluster 1 and 2.225 for cluster 2. This shows that the Minkowski method produces a more accurate group division in clustering poverty by district/city in Indonesia.

This research focuses on comparing the results of hierarchical clustering using the complete linkage method that involves four types of distances, namely: Euclid, Euclid Square, Manhattan, and Minkowski, applied in the context of poverty level analysis. Future researchers interested in exploring distance comparisons in cluster analysis may consider using other methods and distances. With the wide selection of methods and distance types available, there is an opportunity to develop this research further, especially when applied to different fields of science.

REFERENCES

[1] M. P. Todaro and S. C. Smith, *Pembangunan Ekonomi*, edisi ke-9. Jakarta: Erlangga, 2006.
[2] Badan Pusat Statistik (BPS), *Indonesia dalam Angka 2020*. Jakarta: BPS, 2020.
[3] B. Simamora, *Analisis Multivariat Pemasaran*, edisi pertama. Jakarta: PT Gramedia Pustaka Utama, 2005.
[4] J. Supranto, *Analisis Multivariat: Arti dan Interpretasi*. Jakarta: PT Asdi Mahasatya, 2004.
[5] S. B. Purnamasari, "Pemilihan Cluster Optimum Pada Fuzzy C-Means (Studi Kasus: Pengelompokan Kabupaten/Kota di Jawa Tengah berdasarkan Indikator Indeks Pembangunan Manusia)," *Jurnal Gaussian*, vol. 3, no. 3, Universitas Diponegoro, 2014.
[6] W. Alwi and M. Hasrul, "Analisis Klaster untuk Pengelompokkan Kab/Kota di Provinsi Sulawesi Selatan Berdasarkan Indikator Kesejahteraan Rakyat," *Jurnal MSA*, vol. 6, no. 1, pp. 35–42, 2018.
[7] T. P. Ryan, *Modern Regression Methods*. New York, NY: Wiley, 1997.
[8] E. Prasetyo, *Data Mining: Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset, 2012.
[9] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, edisi ke-5. New Jersey: Prentice Hall, 2002.
[10] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*. United States: Pearson Prentice Hall, 2010.

[11] C. E. Mongi, "Penggunaan Analisis Two Step Clustering Untuk Data Campuran," *Jurnal de Cartesian (JdC)*, vol. 4, no. 1, pp. 9–19, 2015.
[12] I. Haviluddin, M. Iqbal, G. M. Putra, N. Puspitasari, H. J. Setyadi, F. A. Dwiyanto, A. P. Wibawa, and R. Alfred, "A Performance Comparison of Euclidean, Manhattan and Minkowski Distances in K-Means Clustering," dalam *2020 6th International Conference on Science in Information Technology (ICSITech)*, 2020, pp. 184–188.
[13] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. New Jersey: Pearson Prentice Hall, 2007.
[14] D. R. Ningrat, "Analisis Cluster Dengan Algoritma K-Means dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi," *Jurnal Gaussian*, vol. 5, no. 4, Universitas Diponegoro, 2016.