BAREKENG: Journal of Mathematics and Its Applications
December 2023 Volume 17 Issue 4 Page 1865–1872
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss4pp1865-1872

CLUSTER ANALYSIS FOR DISTRICT/CITY GROUPING BASED ON VARIABLES AFFECTING POVERTY IN ACEH PROVINCE USING AVERAGE LINKAGE METHOD

Mirda Olivia[1], Nurviana[2*], Fairus[3]

[1,2,3]*Mathematics Departments, Faculty of Engineering, Samudra University*
Langsa Lama, Kota Langsa, 24411, Indonesia

*Corresponding author's e-mail: *nurviana@unsam.ac.id*

Article History:

Received: 5th April 2023
Revised: 15th August 2023
Accepted: 8th September 2023

Keywords:

Aceh;
Average Linkage;
Cluster Analysis;
Poverty.

ABSTRACT

Poverty is an inability of a person/household to meet basic needs in everyday life. Aceh is one of the provinces in Indonesia which is still faced with the problem of poverty. In March 2021 the poor population in Aceh numbered 834.24 thousand people and in September 2021 the poor population in Aceh increased by 16 thousand people, a total of 850.26 thousand people. Therefore the authors are interested in classifying and looking at the characteristics of 23 districts/cities in Aceh Province based on 5 variables that affect poverty. This study uses data from SUSENAS processed from BPS Kota Langsa in 2021. The variables used are households with the type of floor of a residential building made of soil/bamboo (X_1), households with a floor area of a residential building < 10 m2 per capita (X_2), households with residential walls made of bamboo/rumbia/wood (X_3), households with a source of drinking water from unprotected wells/springs/rivers/rainwater (X_5), and households whose head of household did not attend school/didn't finish primary school/only primary school (X_8). This study uses the average linkage method, namely the distance between two clusters is measured by the average distance between objects in each cluster. Of the 23 regencies/cities, 3 clusters were formed, namely cluster 1 with the lowest poverty rate consisting of 17 regencies/cities. Cluster 2 with the highest poverty rate consists of 2 districts/cities. Cluster 3 with a moderate poverty level consists of 4 districts/cities. The characteristics of the clusters that are formed are in clusters 1, 2 and 3 the dominant poverty level is influenced by the variable X_3, which means that there are still many households that have houses with inadequate wall types. In clusters 1 and 3 the poverty rate is not dominantly influenced by variable X_1, which means that many households have houses with proper floor types. In cluster 2 the poverty rate is not dominantly influenced by variable X_5, which means that many households consume drinking water from cleaner and more protected sources.

How to cite this article:

M. Olivia, Nurviana and Fairus., "CLUSTER ANALYSIS FOR DISTRICT/CITY GROUPING BASED ON VARIABLES AFFECTING POVERTY IN ACEH PROVINCE USING AVERAGE LINKAGE METHOD," BAREKENG: J. Math. & App., vol. 17, iss. 4, pp. 1865-1872, December, 2023.

Homepage journals: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail:*barekeng.math@yahoo.com;barekeng.journal@mail.unpatti.ac.id*
Research Articles• OpenAccess

1. INTRODUCTION

Poverty is a problem that still occurs in several countries, including Indonesia. Poverty is still a multidimensional problem so that it becomes a development priority. Poverty is one of the fundamental problems, because poverty involves meeting the most basic needs in life and poverty is a global problem because poverty is a problem faced by many countries [1]. According to BPS, poverty is an economic inability to meet basic food and non-food needs as measured from the expenditure side. In this case it can be concluded that poverty is an inability of a person/household to meet basic needs in everyday life.

Aceh is one of the provinces in Indonesia which is still faced with the problem of poverty. According to BPS [2] the number of poor people in Aceh was recorded at 850.26 thousand people (15.53%), an increase of 16 thousand people compared to the number of poor people in March 2021 which numbered 834.24 thousand people (15.33%). In this case, it means that during the period March 2021-September 2021 the percentage of poor people in Aceh has increased.

To reduce the increase in the number of poor people, poverty alleviation efforts such as development are carried out. There are two strategies that must be taken in efforts to reduce poverty, namely, firstly protecting families and groups of poor people through meeting their needs from various fields, secondly conducting training for them so that they have the ability to carry out efforts to prevent new poverty [3].In an effort to reduce poverty, information on the level of poverty in each district/city in Aceh Province is very much needed, bearing in mind that the geographical conditions in each district/city in Aceh are different which causes the status of the distribution of poverty to be different. Therefore we need a study that can classify districts/cities that have a similar status of poverty distribution.

Cluster analysis is one of the multivariate analyzes that is used to group objects into several clusters according to the similarity of the variables studied, so that the similarity of objects in the same cluster will be obtained compared to objects in different clusters [4]. The main objective of cluster analysis is to classify objects into relatively homogeneous groups based on a set of variables considered for research. In general, cluster analysis is divided into two methods, namely the hierarchical method and the non-hierarchical method [5]. In this study the application of cluster analysis was used to classify 23 districts/cities in Aceh Province based on indicators that affect poverty in 2021 and look at the characteristics of poverty in each of the cluster results. Clustering is done using the cluster average linkage analysis method, the choice of this method is because this method is considered to have better accuracy than other hierarchical methods [6].

2. RESEARCH METHODS

2.1 Poverty

Poverty is a state of inability to meet basic needs in everyday life. This situation is caused by the low income generated to meet the necessities of life such as clothing, boards and food. This situation can also have an adverse impact on meeting other living standards such as education and health. Poverty is understood in different ways. The main understanding includes: first, the description of material shortages, which usually includes daily food needs, clothing, food, housing, and health services [7]. There are four forms of poverty, namely absolute poverty, relative poverty, cultural poverty and structural poverty [8].

2.2 Factor Analysis

Factor analysis is a technique of reducing variables to be simpler based on the relationship between the variables studied into a number of factors. In principle, variable analysis is used to reduce data, namely the process of summarizing a number of variables into fewer and naming them as factors [9].

2.3 Cluster Assumptions

2.3.1 Sample Adequacy Test

To see the adequacy of the sample, the Keizer-Meyer-Olkin (KMO) test was carried out. To find out whether the data is representative of the existing population, the KMO value is needed [10]:

Hypothesis:

H0 : The data is feasible to be analyzed
H1 : The data is not feasible to be analyzed

Test Statistics :

$$r_{ij} = \frac{N\sum X_i X_j - (\sum X_i)(\sum X_j)}{\sqrt{[N\sum X_i^2 - (\sum X_i)^2][N\sum X_j^2 - (\sum X_j)^2]}} \quad (1)$$

$$a_{ij} = \frac{r_{yx_i} - r_{yx_j} r_{x_i x_j}}{\sqrt{1 - r_{x_i x_j}^2}\sqrt{1 - r_{yx_j}^2}} \quad (2)$$

$$KMO = \frac{\sum_{i \neq j} \sum r_{ij}^2}{\sum_{i \neq j} \sum r_{ij}^2 + \sum_{i \neq j} \sum a_{ij}^2} \quad (3)$$

Information :

r_{ij}^2 : Correlation between variables i and j
a_{ij}^2 : Partial correlation between variables i and j

The sample can be said to represent the existing population if the KMO value is > 0.5. KMO standard values can be seen in the following **Table 1**:

Table 1. Characteristics of KMO Values

KMO value	Information
0.8 – 1.0	Very worth it
0.7 – 0.8	worthy
0.6 – 0.7	Pretty decent
0.5 – 0.6	Not quite worth it
< 0.5	Not feasible

2.3.2 Bartlett's test

Testing with the Bartlett test is used to see whether there is a relationship (correlation) between variables in the multivariate case **[11]**.The Bartlett test is carried out using the following equation **[12]**:

Hypothesis:

H0 : R = I (the correlation matrix is the same as the identity matrix)
H1 : R ≠ I (correlation matrix is not the same as identity matrix)

Test Statistics :

$$Barlett = -\ln |R| \left(n - 1 - \left(\frac{2p+5}{6}\right)\right) \quad (4)$$

Information :

|R| : The determinant value of the correlation matrix
n : The number of observations
p : The number of variables

Reject H0 if the p-value means the variables are correlated with each other, so the data is feasible to analyze.$\leq \alpha$

2.4 Cluster Analysis With The Average Linkage Method

Cluster analysis or group analysis is a data analysis technique that aims to classify individuals or objects into several groups that have different characteristics between groups, so that individuals or objects that are in one group will be relatively homogeneous **[13]**.Cluster analysis using the average linkage method is a hierarchical cluster analysis method that is often used. In this method the distance between two clusters is measured by the average distance between an object in one cluster and an object in another cluster **[14]**.

$$d(uv)w = \frac{\sum_i \sum_k d_{ik}}{N_{uv} N_W} \quad (5)$$

Information :

d_{ik}: the distance between the i-th object in the cluster (UV) and the k-th object in the cluster to W
N_{uv}: Number of objects in the cluster (UV)
N_W: Number of objects in cluster W

2.5 Selection of Distance Measurement Methods

The similarity between two objects is indicated by the distance between the two objects. The smaller the value of the distance between the two objects, the greater the similarity between the two objects **[15].**The Manhattan distance is used if the observed variables are correlated or not independent **[16].**In this study the distance used is the Manhattan distance due to the correlation between the research variables.

Manhattan distance can be formulated as follows:

$$d_{i,j} = \sum_{k=1}^{p} |x_{ik} - x_{jk}| \quad (6)$$

Information :

$d_{i,j}$: the distance between object i and the k-th object
x_{ik}: the value of object i in the k-th variable
x_{jk}: the value of object j in the k-th variable
p.s: the number of observed variables

3. RESULTS AND DISCUSSION

3.1 Cluster Assumptions

In cluster analysis, there are two assumptions that must be met. In this study, after carrying out the KMO test and Bartlett test using 8 variables, it can be seen that there is a correlation between the variables. Thus a factor analysis will be carried out to reduce the variables by looking at the MSA value of each variable. The SPSS output results show that there are 3 variables with an MSA value of <0.5, so these variables must be eliminated and repeated cluster assumption tests are carried out.

The following are the results of the KMO and Bartlett tests for the variables X_1, X_2, X_3, X_5 and X_8:

Table 2. KMO Test and Bartlett Test Variables X_1, X_2, X_3, X_5 and X_8

Kaiser-Meyer-Olkin Measures of Sampling Adequacy		**0.875**
Bartlett's Test of Sphericity	*approx. Chi-Square*	116.454
	Df	10
	Sig.	0.000

In **Table 2** it can be seen that the KMO test values for variables X_1, X_2, X_3, X_5 and X_8 are 0.875 and greater than 0.5, which means that H0 is accepted or the total data of 23 districts/cities in Aceh Province is feasible for analysis. Bartlett test results show a significance level of 0.000 and less than 0.05, which means that H0 is rejected or there is a relationship (correlation) between the study variables. Because the two assumptions are met, the next analysis process can be carried out, namely looking at the MSA value after the variables X_4, X_6 and X_7 are eliminated.

The following are the results of the MSA test for variables X_1, X_2, X_3, X_5 and X_8 :

Table 3. MSA Value 5 Variables

Variable	MSA value
X_1	0.852
X_2	0.856
X_3	0.883
X_5	0.958
X_8	0.858

In **Table 3** above it can be seen that the MSA value of all variables is greater than 0.5, so only 5 variables are suitable for further analysis of the 8 variables.

Table 4. Number Of Variants Of Each Factor

Main Component	Root Traits (Eigen Value)	The Percentage Of Diversity	Cumulative Percentage Of Diversity
1	**4.242**	**84.835**	**84.835**
2	**0.391**	**7.826**	**92.661**
3	**0.171**	**3.429**	**96.090**
4	**0.122**	**2.444**	**98.534**
5	**0.073**	**1.466**	**100.000**

Table 4 shows that there is 1 main component that has a characteristic root (eigen value) greater than 1, thus the factor formed is 1 factor.

Based on the results of the factor analysis above, the variables that can be used in cluster analysis are variables X_1 (Households with a floor area of <10 m2), X_2 (Households with a type of residential building floor made of soil/bamboo), X_3 (Households with type of shelter made of bamboo/thatch/wood), X_5 (Households with a source of drinking water from unprotected wells/springs/rivers/rainwater), and X_8 (Households whose head of household does not go to school/does not finish elementary school/only SD).

3.2 Selection of Distance Measurement Methods

In this study the distance used is the Manhattan distance due to the correlation between the research variables. Following are the results of the research variable output with the Manhattan distance:

Table 5. Manhattan Distance Matrix

	Simeulue	Aceh Singkil	Aceh South	…	Subulussalam
Simeulue	0.000	1.417	1.977	…	1.711
Aceh Singkil	1.417	0.000	2.435	…	1.093
South Aceh	1.977	2.435	0.000	…	3.169
Southeast Aceh	3.158	3.841	3.186	…	4.870
East Aceh	12.374	11.385	11.394	…	11.523
Central Aceh	4.237	3.290	2.303	…	4.237
West Aceh	1.521	1.875	1.436	…	2.473
Aceh Besar	3.672	3.835	2.692	…	4.864
⋮	⋮	⋮	⋮	⋮	⋮
Subulussalam	1.711	1.093	3.169	…	0.00

The following is an example of a calculation using the Manhattan distance formula. For example, we calculated the similarity between Simeulue District and Aceh Singkil District (Objects 1 and 2).

$d_{i,j} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$

$d_{1,2}$ = |x11 – y21| + |x12 – y22| + |x13 – y23| + |x14 – y24| + |x15 – y25|

$d_{1,2}$ = |-0.45035 – (-0.58833)| + |-0.68061 – (-0.5165)| + | -0.35218 – (-0.42803)| + |-0.17705 – 0.65887| + |0.76365 – 0.56006|

$d_{1,2}$ = 0.13798 + 0.16404 + 0.07585 + 0.83592 + 0.20359

$d_{1,2}$ = 1.41738

Calculation of the similarity between objects 1 and 2 with a manhattan distance of 1.417. Then the similarities between Simeulue Regency and South Aceh Regency (Objects 1 and 3):

$d_{i,j} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$

$d_{1,3}$ = |x11 – y31| + |x12 – y32| + |x13 – y33| + |x14 – y34| + |x15 – y35|

$d_{1,3}$ = |-0.45035 – (-0.48657)| + |-0.68061 – (-0.30557)| + | -0.35218 – (-0.58766)| + |-0.17705 – (-0.40362)| + |0.76365 – 0.33958|

$d_{1,3}$ = 0.03622 + 0.37504 + 0.23548 + 0.22657 + 1.10323

$d_{1,3}$ = 1.97654

Calculation of the similarity between objects 1 and 3 with a manhattan distance of 1.977. The example of calculating the distance above shows that Simeulue Regency has characteristics that are more

similar to Aceh Singkil than to South Aceh. This is because the value of the distance between Simeulue and Aceh Singkil districts is smaller than the value of the distance between Simeulue and South Aceh districts, which is 1.417.

3.3 Grouping with the Average Linkage Method

The average linkage method will classify 23 districts/cities in Aceh Province based on the average distance between all members in one cluster and all other cluster members. Grouping begins by looking at the shortest distance between two objects using the Manhattan distance measure that has been obtained. The clustering results with the average linkage method are in the form of a dendrogram in **Figure 1** below:

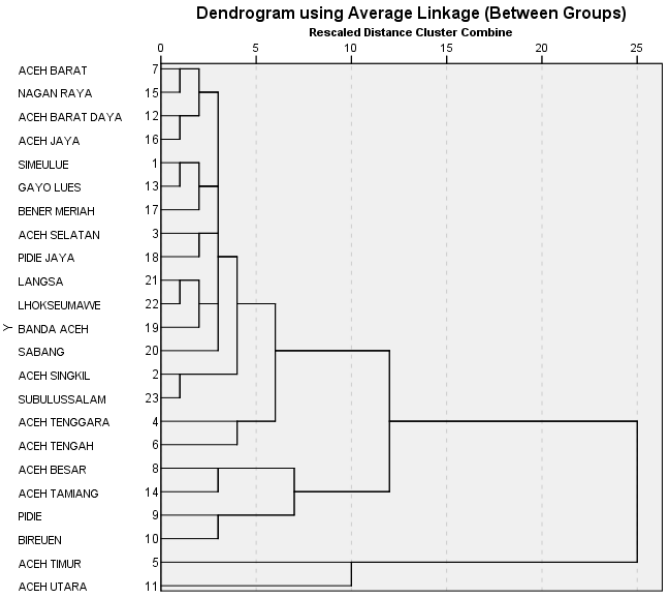

Figure 1. Average Linkage Dendrogram

The dendrogram is read from left to right where the vertical lines indicate the clusters that are merged together, while the lines on the scale show the cluster distances that are combined.

3.4 Determine The Number Of Clusters And Their Members

Details of the number of clusters with members formed can be seen in the SPSS cluster membership output table using the average linkage method. From the **Table 6** it can be concluded that the members of each cluster are:

Table 6. Cluster Members with the Average Linkage Method

Clusters	Member
1	Simeulue, Aceh Singkil, South Aceh, Southeast Aceh, Central Aceh, West Aceh, Southwest Aceh, Gayo Lues, Nagan Raya, Aceh Jaya, Bener Meriah, Pidie Jaya, Banda Aceh, Sabang, Langsa, Lhokseumawe, Subulussalam
2	East Aceh, North Aceh
3	Aceh Besar, Pidie, Bireuen, Aceh Tamiang

3.5 Cluster Interpretation

At this stage will provide specific characteristics in each cluster that is formed. Determination of the characteristics of the cluster can be seen from the centroid value (average) in each cluster. The following is a table of average values in each cluster:

Table 7. Cluster Centroid (Mean) Value

Variable	Average		
	*Clusters*1	*Clusters*2	*Clusters*3
X_1	818	9.751	2.776
X_2	6.836	32.478	16.818
X_3	14.231	67.051	44.583
X_5	1.268	7.046	3.478
X_8	10.286	40.695	29.596

*Clusters*1 has a high value on variable X_3 and has the lowest value on variable X_1
*Clusters*2 has a high value on variable X_3 and has the lowest value on variable X_5
*Clusters*3 has a high value on variable X_3 and has the lowest value on variable X_1

4. CONCLUSIONS

Based on the results of the data analysis that has been done, two conclusions are obtained. First, from the results of cluster analysis using the average linkage method, 3 clusters were formed from 23 districts/cities in Aceh Province. Cluster 1 with the lowest poverty rate consisting of 17 Regencies/Cities. Cluster 2 with the highest poverty rate consisting of 2 districts/cities. Cluster 3 with a moderate poverty level consists of 4 districts/cities. Second, cluster characteristics in terms of dominant and non-dominant variables affect the poverty rate. In clusters 1, 2 and 3 the dominant poverty rate is influenced by variable X_3, which means that there are still many households that have houses with inadequate wall types. In clusters 1 and 3 the poverty rate is not dominantly influenced by variable X_1, which means that there are already many households that have a house with a proper floor type. In cluster 2 the poverty rate is not dominantly influenced by variable X_5, which means that many households consume drinking water from cleaner and more protected sources.

REFERENCES

[1] Y. Yacoub, "Pengaruh Tingkat Pengangguran terhadap Tingkat Kemiskinan Kabupaten / Kota di Provinsi Kalimantan Barat," vol. 8, pp. 176–185, 2012.

[2] BPS Provinsi Aceh, "Profil Kemiskinan di Provinsi Aceh September 2013," no. 4, pp. 1–5, 2013.

[3] D. V. Ferezagia, "Analisis Tingkat Kemiskinan di Indonesia," *J. Sos. Hum. Terap.*, vol. 1, no. 1, pp. 1–6, 2018.

[4] M. Goreti, Y. Novia N, and S. Wahyuningsih, "Perbandingan Hasil Analisis Cluster dengan Menggunakan Metode Single Linkage dan Metode C-Means (Studi Kasus: Data Tingkat Kualitas Udara Ambien pada Perusahaan Perkebunan di Kabupaten Kutai Barat Tahun 2014)," *J. EKSPONENSIAL*, vol. 7, no. 1, pp. 9–16, 2016.

[5] A. N. Fathia, R. Rahmawati, and Tarno, "Analisis Klaster Kecamatan Di Kabupaten Semarang Berdasarkan Potensi Desa Menggunakan Metode Ward Dan Single Linkage," *Gaussian*, vol. 5, no. 4, pp. 801–810, 2016, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[6] D. U. Muis, "Perbandingan Analisis Cluster Hierarki Aglomeratif Dengan Menggunakan Metode Single Linkage, Complete Linkage dan Average Linkage (Studi Kasus : Indikator Kemiskinan Ditinjau dari Sektor Perumahan dan Lingkungan di Kabupaten Gunung Kidul Tahun 2015," pp. 1–14, 2017.

[7] Y. Masruroh, "Infrastruktur jalan terhadap Perekonomian Kota Malang," *J. Ekon. dan Bisnis*, vol. 11, no. 9, pp. 112–130, 2019.

[8] E. H. Jacobus, P. . Kindangen, and E. N. Walewangko, "Analisis Faktor-Faktor Yang Mempengaruhi Kemiskinan Rumah Tangga Di Sulawesi Utara," *J. Pembang. Ekon. Dan Keuang. Drh.*, vol. 19, no. 7, pp. 86–103, 2019, doi: 10.35794/jpekd.19900.19.7.2018.

[9] E. Verdian, "Analisis Faktor yang Merupakan Intensi Perpindahan Merek Transportasi Online di Surabaya," *Agora*, vol. 7, no. 1, pp. 1–8, 2019.

[10] W. Alwi and M. Hasrul, "Analisis Klaster Untuk Pengelompokkan Kabupaten/Kota Di Provinsi Sulawesi Selatan Berdasarkan Indikator Kesejahteraan Rakyat," *J. MSA (Mat. dan Stat. serta Apl.)*, vol. 6, no. 1, p. 35, 2018, doi: 10.24252/msa.v6i1.4782.

[11] Q. Nafisah and N. E. Chandra, "Analisis Cluster Average Linkage Berdasarkan Faktor-Faktor Kemiskinan di Provinsi Jawa Timur," *Zeta - Math J.*, vol. 3, no. 2, pp. 31–36, 2017, doi: 10.31102/zeta.2017.3.2.31-36.

[12] S. Machfudhoh and N. Wahyuningsih, "Analisis Cluster Kabupaten / Kota Berdasarkan Pertumbuhan Ekonomi Jawa Timur," *Sains dan Seni Pomits*, vol. 2, no. 1, pp. 1–8, 2013, [Online]. Available: http://digilib.its.ac.id/public/ITS-paper-37597-1210100028-paper.pdf

[13] M. W. Talakua, Z. A. Leleury, and A. W. Talluta, "Analisis Cluster Dengan Menggunakan Metode Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014," *J. Ilmu Mat. dan Terap.*, vol. 11, no. 2, pp. 119–128, 2017.

[14] C. Suhaeni, A. Kurnia, and R. Ristiyanti, "Perbandingan Hasil Pengelompokan menggunakan Analisis Cluster Berhirarki, K-Means Cluster, dan Cluster Ensemble (Studi Kasus Data Indikator Pelayanan Kesehatan Ibu Hamil)," *J. Media*

Infotama, vol. 14, no. 1, 2018, doi: 10.37676/jmi.v14i1.469.

[15] U. Putriana, Y. Setyawan, and Noeryanti, "Metode Cluster Analysis Untuk Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Variabel Yang Mempengaruhi Kemiskinan Pada Tahun 2013," *J. Stat. Ind. dan Komputasi*, vol. 1, no. 1, pp. 38–52, 2016.

[16] C. E. Mongi, "Penggunaan Analisis Two Step Clustering untuk Data Campuran," *d'CARTESIAN*, vol. 4, no. 1, p. 9, 2015, doi: 10.35799/dc.4.1.2015.7251.