ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
vol. 1, No. 2, October 2020, pp. 94-102

Grouping Of Districts Based on Poverty Factors In Papua Province Uses The K-Medoids Algorithm

Afdelia Novianti [a,1,*], Irsyifa Mayzela Afnan [b,2], Rafi Ilmi Badri Utama [b,3], Edy Widodo [b,4]

[a] Department of Statistics, Universitas Islam Indonesia, Jl.Kaliurang km 14,5, Yogyakarta 55584, Indonesia
[b] Department of of Statistics, Universitas Islam Indonesia, Jl.Kaliurang km 14,5, Yogyakarta 55584, Indonesia
[1] 18611082@students.uii.ac.id*; [2] 18611181@students.uii.ac.id; [3] 18611085@students.uii.ac.id, [4]edywidodo@uii.ac.id

ARTICLE INFO

Article history
Received September 21, 2021
Revised November 30, 2021
Accepted December 7, 2021

Keywords
Poverty
K-Medoids
Clustering
Papua

ABSTRACT

Poverty is an essential issue for every country, including Indonesia. Poverty can be caused by the scarcity of basic necessities or the difficulty of accessing education and employment. In 2019 Papua Province became the province with the highest poverty percentage at 27.53%. Seeing this, the district groupings formed in describing poverty conditions in Papua Province are based on similar characteristics using the variables Percentage of Poor Population, Gross Regional Domestic Product, Open Unemployment Rate, Life Expectancy, Literacy Rate, and Population Working in the Agricultural Sector using K-medoids clustering algorithm. The results of this study indicate that the optimal number of clusters to describe poverty conditions in Papua Province is 4 clusters with a variance of 0.012, where the first cluster consists of 10 districts, the second cluster consists of 5 districts, the third cluster consists of 12 districts, and the fourth cluster consists of 2 districts.

1. Introduction

Poverty can be caused by the scarcity of basic necessities or the difficulty of accessing education and employment. The definition of the term poverty is the condition of a person or group of people, both men and women, who are unable to fulfill their basic rights to maintain and develop a dignified life. The community's fundamental rights include the need for food, health, education, decent work for humanity, housing, clean water, land, natural resources, a healthy environment, a sense of security for both men and women, and equality [1].

In mid-2020, the Head of BAPPENAS said that the poverty target in 2021 had fallen to 9.2%, where Indonesia achieved this poverty rate in 2019. In 2019 also poverty in Indonesia reached the lowest rate in the 2006-2020 period. The province with the highest poverty percentage in 2019 was Papua Province at 27.53%, and the lowest was DKI Jakarta at 3.47% [2].

Seeing this, a grouping of districts was formed to describe the conditions of poverty in Papua Province based on similar characteristics using several variables carried out in previous research. These variables are the Percentage of the Poor which describes the percentage of poor people compared to the total population, Gross Regional Domestic Product, Open Unemployment Rate, Life Expectancy, Literacy Rate, and Population Working in the Agricultural Sector in 2019 to help meet the 2021 target.

p-ISSN 2798-253X
e-ISSN 2798-3153

2. Literature Review

Research conducted by Baiq Tiswati in 2012 related to the Analysis of Factors that affect the poverty rate in Indonesia concluded that the variable Life Expectancy affects poverty in Indonesia [3]. Then in 2015, Nur Ika Septiana conducted research related to poverty analysis in Central Java Province using the spatial regression method and concluded that one of the influential variables was the Percentage of Population Working in the Agricultural Sector [4]. Furthermore, there is research by Abid Muhtarom regarding the Effect of Literacy Rate on Poverty in East Java Province for the 2008-2015 period, and it was found that AMH had an effect of 35.7% [5]. In 2018, Sri Wahyuni and Yogo Aryo Jatmiko conducted research related to Grouping Districts in Java Based on Poverty Factors with An Average Linkage Hierarchical Clustering Approach using variables of the open unemployment rate, percentage of households working in agriculture, household expenditure per capita, and the average length of schooling [6]. Next, in 2019, research conducted by Irfan Eko Saputera entitled Factors Affecting Poverty in Papua Province For the period 2006-2016 obtained the results that factors that affect poverty levels in Papua Province based on data are Government Spending, Unemployment Rate, and Inflation [7]. In the same year, I Nyoman Giri Saputra and Ni Nyoman Yuliarmi conducted a study entitled Analysis of Factors Affecting Poverty Rates in Nusa Tenggara and Papua showed Gross Regional Domestic Product, Unemployment, Investment, and Inflation had a significant effect on poverty [8].

3. Method

3.1. Data and Data Sources

The data used in this study are secondary data with the variables namely Percentage of Poor Population (X1), Life Expectancy (X2), Literacy Rate (X3), Population Working in the Agricultural Sector (X4), Gross Regional Domestic Product (X5), and Open Unemployment Rate (X6) of Papua Province in 2019 which was obtained from www.bps.com.

3.2. Research Stage

The flowchart of the research to be carried out is as follows:

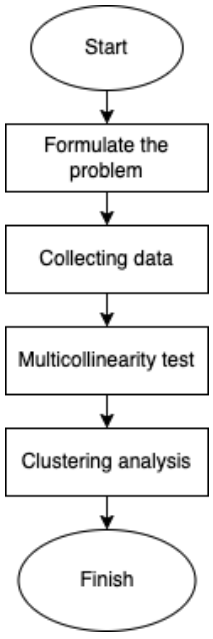

Fig. 1. Research stages

The research method used in this study is the K-Medoids method. The first stage is to formulate the problem, then collect and clean the data. The data will then be analyzed using descriptive statistical analysis and clustering analysis. In K-Medoids analysis, the steps involved are multicollinearity tests and clustering analysis to determine the best cluste to conclude

3.3. Poverty

Poverty is the inability of individuals to meet minimal basic needs to live a decent life. Poverty is a condition that falls below the standard value line of minimum needs, both for food and non-food, called the poverty line or poverty threshold. The poverty line is the amount of rupiah required by each individual to pay for the equivalent of 2100 kilos of calories per person per day and non-food needs consisting of housing, clothing, health, education, transportation, and various other goods and services [9]. A population is said to be poor when characterized by low levels of education, work productivity, income, health and nutrition, and well-being of its life, which embody its circle of helplessness. Limited human resources can cause poverty , both formal and non-formal education pathways, that ultimately lead to consequences for the low level of informal education. [10].

3.4. Descriptive statistics

Descriptive statistics serve to describe the objects studied through sample or population data as is, without conducting analysis and making conclusions that apply to the public. [9]. Descriptive statistics are methods related to collecting and presenting data to provide useful information. [11].

3.5. Cluster Assumptions

Multicollinearity tests are also required to determine the existence of linear relationships between 2 or more variables.The absence of multicollinearity in the analysis of the assumption cluster must be met [12]. Symptoms of multicollinearity can be detected in several ways, such as the following:

- Calculates the coefficient of a simple correlation between fellow free variables; if there is a simple correlation coefficient that reaches or exceeds 0.8, then there is multicollinearity.

  $$r=\frac{n\sum XY-\sum X\sum Y}{\sqrt{(n\sum X^2-(\sum X)^2)(n\sum Y^2-(\sum Y)^2)}} \quad (1)$$

  where:
  X = independent variable
  Y = dependent variable
  n = number of samples
- Calculates the tolerance value or VIF; if the tolerance value is less than 0.1 or the VIF value exceeds 10, it indicates multicollinearity between variables. Large Variance Inflation Factor (VIF) values indicate high multicollinearity between variables [13], and the formula for calculating VIF is:

  $$VIF=\frac{1}{1-R_i^2} \quad (2)$$

  where R_i^2 is the coefficient of determination.

3.6. Validity Index

Silhouette validity index is a statistical measure used to select the problem of determining the number of optimal clusters that can represent a brief graphic of how well each object is located in the cluster [14]. Silhouette validity index can be written with the following equation:

$$S(i)=\frac{b(i)-a(i)}{Max\{a(i),b(i)\}} \quad (3)$$

Silhouette validation index equations can be written as follows:

p-ISSN 2798-253X
e-ISSN 2798-3153

$$S(i)=\{1-\frac{a(i)}{b(i)}, \text{ if } a(i)<b(i)\ 0, \text{ if } a(i)=b(i)\ \frac{a(i)}{b(i)}, \text{ if } a(i)>b(i) \quad (4)$$

where:
a(i) = Average of distance from other observation
b(i) = Average of distance from observation i to all observation in the other cluster

The average *S(i)* of all objects in a cluster indicates how close the similarities of objects in a cluster are, indicating how precisely the objects have been grouped. The closer to 1, the better, and the closer to -1, the worse [13].

3.7. K-Methods Algorithm

The K-Medoids algorithm (Partitioning Around Medoid) was developed by Leonard Kaufman and Peter J. Rousseeuw, which is an algorithm similar to k-means because both algorithms break down datasets into groups. K-Medoids group based on their median value [15]. K-Medoids is a partition clustering method that minimizes the distance between a labeled point in a cluster and a point designated as the cluster's center. [16]. The K-Medoids algorithm is used to overcome the weakness of the k-means algorithm that is very sensitive to outliers because these objects are very far from the majority of other data. Therefore, if inserted into a data cluster of this kind can distort the mean value of the cluster [17]. The process of the K-Medoids cluster is as follows [18]:

1) Determine the number of clusters (k) to be formed.
2) Calculate the proximity of data using the Euclidean Distance method, with the following formula:

$$D_{ik}=\sqrt{\sum_k^n\left(x_{ij}-x_{kj}\right)^2} \quad (5)$$

With D_{ik} : Euclidean Distance, x_i : Data (i), x_j : Data (j), x_{ij} : Data (i) attribute (j), c_{kj} : Data (k) attribute (j)

3) Select any data on each cluster that will become new medoids. Then, c alculate the distance of each object in each cluster with the new medoids.
4) Calculate the deviation (S) by subtracting the total distance of the old with the new total distance. If S < 0, replace the object with a medoids data cluster to form a new set of medoids object k.
5) Repeat steps 3-5 until no medoids change.

3.8. Determining the Goodness of the Cluster

The average standard deviation in cluster (v_w) and standard deviation between clusters ($v_b\sigma_B$) can be used to determine the good performance of a clustering method W, b, B. The standard deviation formula in Cluster (v_w) [19]:

$$v_w=\frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)\times v_i^2 \quad (6)$$

where:
v_w = variance within cluster
k = number of cluster
N = amount of data
n_i = amount of data in a cluster
v_i^2 = variance in a cluster

Standard deviation formula between Clusters (v_b):

$$v_b=\frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\underline{x}_i-\underline{x}\right)^2 \quad (7)$$

where:
v_b = variance between cluster

p-ISSN 2798-253X
e-ISSN 2798-3153

k = number of cluster
n_i = amount of data in a cluster
$\underline{x_i}$ = average data on a cluster
$\underline{x}$ = average of data
Variance formula of a cluster:

$$v=\frac{v_w}{v_b} \quad (8)$$

where:
v = variance
v_w = variance within cluster
v_b = variance between cluster

4 Result and Discussion

4.1. Descriptive Analyze and Multicollinearity Test

Table 1. Descriptive Analyze

	X1	X2	X3	X4	X5	X6
N	29	29	29	29	29	29
Mean	29.220	64.978	89.447	68.040	3.448	3.278
Standard Deviation	10.023	3.816	10.478	27.827	6.698	3.320
Median	30.95	65.61	89.55	74.98	1.12	2.39
Min	10.35	55.12	70.19	4.32	0.55	0
Max	43.65	72.27	100	100	33.56	12.37
Quartile 1	24.81	64.91	81.76	49.70	0.66	0.71
Quartile 3	38.24	66.60	98.90	92.40	2.43	4.68
Skewness	-0.471	-0.874	-0.454	-0.557	3.383	1.180
Kurtosis	-1.097	0.502	-1.360	-1.009	11.655	0.494

Table 2. Multicollinearity Test with VIF values

	X1	X2	X3	X4	X5	X6
X1		1.55	3.93	9.57	1.83	4.97
X2	3.14		4.24	9.58	1.87	4.95
X3	2.77	1.48		7.29	1.80	4.99
X4	2.98	1.48	3.22		1.83	3.02
X5	2.93	1.49	4.07	9.39		4.99
X6	3.13	1.55	4.46	6.10	1.96	

The multicollinearity test aims to test whether there is a correlation between the independent variables in the regression model. The value of VIF (variance inflation factor) measures how much variance from the estimated regression coefficient will increase if there is an apparent multicollinearity problem between variables. Based on the table above. It is found that there is no VIF value > 10. so the data is free from multicollinearity, which means that the data is fulfilled for further analysis.

4.2. Analisis *Cluster* K-Medoids

4.2.1. Determining the Number of Cluster

The further analysis use the K-Medoids method. This method is a variant of the K-Means method, a non-hierarchical cluster. The number of clusters that will be formed (k) in the clustering process with the K-Medoids method uses the Silhouette method.

p-ISSN 2798-253X
e-ISSN 2798-3153

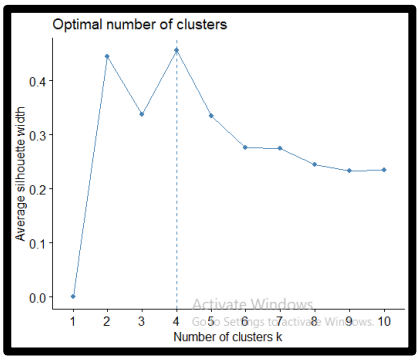

Fig. 2. Plot with silhouette method

The Silhouette method uses an average value approach to estimate the quality of the clusters formed. The higher the average value, the better. Based on the suggestion from the Silhouette method to determine the value of K seen from the highest line or see it by looking at the most optimum line. The figure 2 shows that the optimum K value is 4.

4.2.2. Result and Cluster's Profiling

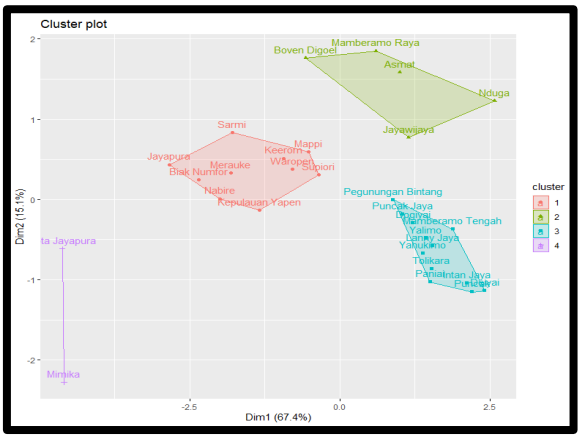

Fig. 3. Grouping visualization

Based on the image of the cluster plot, it formed with the number of clusters as many as 4 clusters. The area plot formed is based on the similarity of poverty observations. Based on the plot, the following areas are included in each cluster:

Table 3. Cluster Member

Cluster	Total	Member
1	10	Merauke. Jayapura. Nabire. Kepulauan Yapen. Biak Numfor. Mappi. Sarmi. Keerom. Waropen. Supiori.
2	5	Jayawijaya. Boven Digoel. Asmat. Mamberamo Raya. Nduga.
3	12	Paniai. Puncak Jaya. Yuhukino. Pegunungan Bintang. Tolikara. Lanny Jaya. Mamberamo Tengah. Yalimo. Puncak. Dogiyai. Intan Jaya. Deiyai.
4	2	Mimika. Kota Jayapura.

From the four clusters that have been formed the profile will be determined by looking at the average of each variable. Then, it will be categorized based on the highest to the lowest average for each variable.

Table 4. Profiling

Cluster	X1	X2	X3	X4	X5	X6	Category
1	22.74	66.891	99.091	45.967	3.096	5.547	Moderate
2	30.392	57.846	90.214	75.912	1.674	2.112	High
3	36.833	65.298	79.333	92.413	0.857	0.765	Very High
4	13.015	71.325	100	12.49	25.195	9.94	Low

Based on the table above, each K-Medoids cluster can be known as the average value and interpreted as follows:

1. 1st Cluster: There are 10 districts included in this cluster. It is categorized as moderate at this level because it has an average of all variables at ranks 3 and 2.
2. 2nd Cluster: There are 5 districts included in this cluster. It is categorized as high at this level because it has an average X2 that is 1 level higher than the cluster.
3. 3rd Cluster: There are 12 districts included in this cluster. It is categorized as very high at this level because it has a higher X4, X3, and X5 variables than all clusters.
4. 4th Cluster: There are 2 districts included in this cluster. It is categorized as low at this level compared to other clusters because it has 4 variables with the lowest average.

4.4. Validasi *Cluster*

Table 5. Validation

Validasi	X1	X2	X3	X4	X5	X6
Vw	46.63	1.40	27.65	127.9736	8.518681	3.432512
Vb	1647.53	372.72	2383.18	18482.33	1043.367	222.8279
V	0.03	0.00	0.01	0.006924	0.008165	0.015404
Var	0.012					

Based on the clustering results, the next step is to see the goodness of the cluster by looking at the value of the variance. The variance is then calculated by averaging the results of the division of Vw and Vb. The result is that the variance value is 0.012 for the K-Medoids clustering method with k=4. Because the value of the variance obtained is small, the number of clusters is good to use.

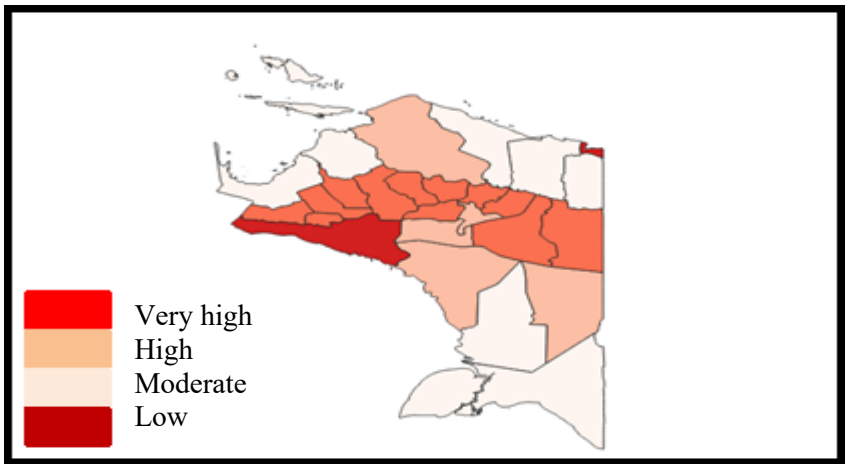

Fig. 4. Cluster visualization

Based on **Fig 4,** which is visualized using the QGis application based on the poverty level. The color orange shows areas with very high poverty rates whereas high dark cream, medium light cream, and brick red are areas with low poverty.

5. Conclusion

This research concludes that:

1. The maximum value for X1 is 43.65; X2 is 72.27; X3 and X4 are 100; X5 is 33.56; and X6 is 12.37 while the minimum value of X1 is 10.35; X2 is 55.12; X3 is 70.19; X4 is 4.32; X5 is 0.55; and X6 is 0
2. The K-Medoids method results that the optimal number of clusters to describe poverty in Papua Province is 4 clusters with a variance of 0.012, where the first cluster consists of 10 districts, the second cluster is 5 districts, the third cluster consists of 12 districts, and the fourth cluster consists of 2 districts.

References

[1] BAPPENAS, "Rencana Pembangunan Jangka Menengah Nasional 2004-2009", 2009.

[2] BAPPENAS, Badan Pembangunan Nasional, [Online]. Available: https://www.bappenas.go.id/id/berita-dan-siaran-pers/gelar-konferensi-pers-akhir-tahun-bappenas-paparkan-rkp-2021-capaian-sdgs-indonesia-dan-transformasi-ekonomi/.

[3] BPS, "Penduduk Fakir Miskin Indonesia Tahun 2002," 2002.

[4] E. Saputera, "Faktor-Faktor yang Mempengaruhi Kemiskinan di Provinsi Papua Periode 2006-2016," Jurnal Ilmiah Mahasiswa Universitas Surabaya, vol. 8, no. 1, 2019.

[5] Fauziah, "Hierarchical Cluster Analysis Industri Manufaktur Besar dan Sedang Berdasarkan Status Penanaman Modal. Studi Kasus: Industri Manufaktur Besar dan Sedang di Jawa Tengah Tahun 2015," in Skripsi Jurusan Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Yogyakarta, Universitas Islam Indonesia, 2019.

[6] Febriyana, "Analisis K-Means dan K-Median pada Data Indikator Kemiskinan.," in Universitas Islam Negeri Syarif Hidayatullah, Jakarta, 2011.

[7] G. I. Nyoman, S. and N. N. Yuliarmi, "Analisis Faktor-Faktor yang Mempengaruhi Tingkat Kemiskinan di Nusa Tenggara dan Papua," Jurnal Ekonomi Pembangunan Universitas Udayana, vol. 9, no. 12, pp. 2762-2791, 2019.

[8] M. Abid, "Pengaruh Angka Melek Huruf Terhadap Kemiskinan di Provinsi Jawa Timur Periode 2008-2015," Jurnal Penelitian Ilmu Manajemen, vol. 1, no. 03, 2016.

[9] M. Bunkers, "Definition of Climate Regions in the Northern Plains Using an Objecting Cluster Modification Technique," j.Climate, pp. 130-146.

[10] N. Kaur, "K-Medoids Clustering Algorithm," International Journal of Computer Application and Technology, pp. 42-45, 2014.

[11] N. L. Anggraeni, "Teknik Clustering Dengan Algoritma K-Medoids untuk Menangani Strategi Promosi di Politeknik TEDC Bandung," Jurnal TEDC, vol. 12, no. 2, 2019.

[12] R. E. Walpole and R. H. Myers, Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan, Bandung: ITB, 1995.

[13] R. P. Silhouttes, "A Graphical Aid To The Interpretation And Validation Of ClusterAnalysis," Journal of computational and Applied Mathematics, pp. 20:53-65, 1987.

[14] S. N. Ika., "Analisis Kemiskinan di Provinsi Jawa Tengah Menggunakan Metode Regresi Spasial," in Skripsi Prodi Statistika, Yogyakarta, Universitas Islam Indonesia, 2015.

[15] S. Santoso, Menguasai Statistik Parametrik Konsep dan Aplikasi dengan SPSS, Jakarta: PT Elex Media Komputindo, 2015.

[16] S. Wahyuni and Y. A. Jatmiko, "Pengelompokan Kabupaten/Kota di Pulau Jawa Berdasarkan Faktor-Faktor Kemiskinan dengan Pendekatan Average Linkage Hierarchical Clustering," Jurnal Aplikasi Statistika dan Komputasi Statistika STIS, vol. 10, no. 1, 2018.

[17] T. Supriatna, "Birokrasi. Pemberdayaan. dan Pengentasan Kemiskinan," in Humaniora Utama Press, Bandung, 1997.

p-ISSN 2798-253X
e-ISSN 2798-3153

[18] Tiswati, "Analisis Faktor-Faktor yang Mempengaruhi Tingkat Kemiskinan di Indonesia," Jurnal Ekonomi Pembangunan, vol. 10, no. 1, Juni, 2012.

[19] Y. H. Christanto and G. Abdillah, "Penerapan Algoritma Partitioning Around Medoids (Pam) Clustering Untuk Melihat Gambaran Umum Kemampuan," Semin. Nas. Teknol. Inf. dan Komun, vol. 2015, no. sentika, pp. 444-448, 2015.

p-ISSN 2798-253X
e-ISSN 2798-3153