Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 2 No. 1 2021, 8 - 15
DOI: *https://doi.org/10.22487/27765660.2021.v2.i1.15675*
eISSN: 2776-5660

K-MEANS CLUSTERING FOR GROUPING INDONESIA UNDERDEVELOPED REGIONS IN 2020 BASED ON POVERTY INDICATORS

Resti Wahyuni[1*]
[1]Badan Pusat Statistik Provinsi Sulawesi Tengah, Indonesia

e-mail: resti.wahyuni@bps.go.id

ABSTRACT

Poverty is still a problem in Indonesia, especially in underdeveloped areas. Underdeveloped areas are areas where the region and its people are less developed than other regions on a national scale. The classification of disadvantaged areas is determined by the president in the Presidential Regulation of the Republic of Indonesia Number 63 of 2020 concerning the Determination of Underdeveloped Regions of 2020-2024. Various policies need to be set by the government to overcome poverty in underdeveloped areas. Program planning strategies may be different for each region. Therefore, in order to achieve an optimal implementation of poverty alleviation programs, it is necessary to group the districts covered in underdeveloped areas in Indonesia based on poverty indicators. The data used is macro data from the characteristics of each region in disadvantaged areas obtained from regional publications in the figures for each district. From the results of the analysis of k means clustering formed three groups with different characteristics in each cluster. In cluster one, the focus of government policies is on employment and sanitation aspects, cluster two is on health, education, and employment aspects, cluster three is on all aspects because cluster three is the area with the highest percentage of poor people compared to the other two clusters. The high percentage of poor people is also followed by other poor aspects.

Keywords: *Poverty, Grouping, Underdeveloped Areas, K-Means*

Cite: Wahyuni, R. (2021). K-Means Clustering for Grouping Indonesia Underdeveloped Regions in 2020 Based on Poverty Indicators. *Parameter: Journal of Statistics, 2*(1), 8-15.

[*]Received: Oct 2021; Reviewed: Dec 2021; Published: Dec 2021

INTRODUCTION

Currently, poverty is still a problem for all countries in the world. Therefore, the main agenda of the SDGs is to eliminate poverty in the world by 2030. In the Outcome Document Transforming Our World: The 2030 Agenda for Sustainable Development, it is stated that the main goal of development is to end poverty in all its forms everywhere. In the RPJPN 2005-2025, the problem of poverty is seen in a multidimensional framework, therefore poverty is not only related to the size of income, but involves several things, including: (i) the vulnerability and vulnerability of people or communities to become poor; (ii) regarding the presence/absence of the fulfillment of the basic rights of citizens and the presence/absence of differences in the treatment of a person or group of people in living a life with dignity.

To measure poverty, the Central Statistics Agency uses the concept of the ability to meet basic needs (basic needs approach). With this approach, poverty is seen as an economic inability to meet basic food and non-food needs as measured from the expenditure side. The number of poor people in Indonesia in March 2020 reached 26.42 million people. Compared to March 2019, the number of poor people in Indonesia increased by 1.28 million people. The percentage of poor people in March 2020 was 9.78 percent, an increase of 0.37 percent from March 2019.

In Indonesia itself, there are still many areas with the percentage of poor people above the national figure, especially in disadvantaged areas. Underdeveloped areas are district whose regions and communities are less developed than other regions on a national scale. The classification of disadvantaged areas is determined by the president in the Presidential Regulation of the Republic of Indonesia Number 63 of 2020 concerning the Determination of Underdeveloped Regions of 2020-2024. The percentage of poor people in underdeveloped areas reached 41.76 percent, namely Deiyai Regency, Papua Province. This figure is considered very high and is still very far from the SDGs goal of eradicating poverty in the world.

Poverty in society is caused by low human capital, such as education, training, or the ability to build (Dowling and Valenzuela, 2010). According to Prastyo (2010), in terms of the source of the cause, poverty can be caused by two kinds. First, cultural poverty is poverty which refers to the attitude of a person or society caused by lifestyle. Second, structural poverty is poverty caused by an unequal community structure, either because of differences in ownership, income capabilities and unequal employment opportunities or because of the uneven distribution of development results.

In the opinion of Sharp, et. al (2000) tries to identify the causes of poverty from an economic point of view. First, viewed from micro factors, poverty arises due to the unequal pattern of resource ownership which results in an unequal distribution of income. The poor have only limited resources and their quality is still quite low. Second, poverty arises due to differences in the quality of human resources. The low quality of human resources can be interpreted as low productivity, which in turn leads to low wages. The low quality of human resources is due to low levels of education, disadvantaged fate, discrimination, or heredity. Third, poverty arises due to differences in access to capital.

Various programs have been launched by the government, both central and regional, to tackle poverty problems ranging from education, food, health, expansion of employment opportunities, assistance for agricultural facilities and infrastructure, business credit assistance for the poor and so on. The number of programs that have been launched, of course, needs to be supported by a careful planning strategy. Program planning strategies may be different for each region. Therefore, in order to achieve an optimal implementation of poverty alleviation programs, it is necessary to group the districts covered in underdeveloped areas in Indonesia based on poverty indicators.

MATERIALS AND METHODS

1. Literature Review

The Central Bureau of Statistics uses the concept of the ability to meet basic needs (basic needs approach). With this approach, poverty is seen as an economic inability to meet basic food and non-food needs as measured from the expenditure side. So the poor are people who have an average monthly per capita expenditure below the poverty line. The Poverty Line (GK) is the sum of the Food Poverty Line (GKM) and the Non-Food Poverty Line (GKNM). The Food Poverty Line (GKM) is the value of spending on minimum food needs which is equivalent to 2100 kilocalories per capita per day. The

commodity package for basic food needs is represented by 52 types of commodities (grains, tubers, fish, meat, eggs and milk, vegetables, nuts, fruits, oils and fats, etc.).

One of the measuring tools that can be used to measure the level of poverty experienced by a person or group of people is the poverty indicator used by BAPPENAS (Harniaty, 2010). The poverty indicators in question are:

a. Food limitation is a measure that looks at the adequacy of food and the quality of the food consumed.
b. Limited access to health is a measure to see the limited access to health and the low quality of health services.
c. The limitation of education is an indicator of measuring the quality of available education, the high cost of education, the limited educational facilities, and the low opportunity to obtain education.
d. Limited access to work is used to measure limited employment and business opportunities.
e. Limited access to housing and sanitation services.
f. Limited access to clean water is used to measure the difficulty of obtaining clean water, limited use of water sources, and low quality of water sources.
g. Limited access to land, this indicator is used for land ownership structure, and land tenure, uncertainty of ownership and land tenure.
h. There is no guarantee of a sense of security, this indicator is related to the lack of security in living life.
i. Limited access to participation. This indicator is measured by low participation, and involvement in decision-making.
j. The magnitude of the population burden, this indicator is related to the size of the family's dependents and the magnitude of the pressure of life (Harniaty, 2010).

Based on the theory above, the poverty indicators used in this study include:

Table 1. Poverty Indicators Used

Poverty Indicators	Definition	Scale
Average daily consumption per capita	Total calorie consumption of all food commodities consumed by the population in one area divided by the total population	Ratio
Number of PUSKESMAS per 100.000 population	Number of PUSKESMAS in an area per 100,000 population	Ratio
Percentage of population 15 years of age and over with a junior high school diploma or below	Number of population aged 15 years and over who completed junior high school education and below which is indicated by the diploma/STTB owned divided by population aged 15 years and over	Ratio
Open Unemployment Rate (TPT)	The ratio of the number of unemployed to the number of the labor force	Ratio
Percentage of informal labours	Percentage of workers with main employment status which includes self-employment, trying to be assisted by temporary workers, casual workers in agriculture and non-agriculture, and family/unpaid workers	Ratio
Percentage of asphalt road length	Percentage of road length in an area with a type of surface covered by asphalt	Ratio
Percentage of households with own defecation facilities	Number of households with own defecation facilities divided by all households	Ratio
Percentage of households with protected cooking/washing water sources	Number of households with cooking/washing water sources divided by all households	Ratio
Percentage of households with own building ownership	Number of households with own building ownership divided by all households	Ratio
Number of dependents	The ratio of the non-productive age population (under 15 years and over 65 years) to the productive age population (between 15 to 64 years) multiplied by 100	Ratio

Based on the Presidential Regulation of the Republic of Indonesia Number 63 of 2020 concerning the Determination of Underdeveloped Regions for 2020-2024, there are 62 districts that are included in

underdeveloped areas. When viewed from the distribution, these underdeveloped areas are mostly located in the eastern part of Indonesia.

Poverty grouping is one way to identify the characteristics of the level of people's welfare in each region so that policies and development strategies are targeted and effective. Cluster analysis is useful in some pattern-analytical exploration, clustering, and decision making (Hafiludien & Istiawan, 2018). Several studies related to cluster analysis include research conducted by Kurniawan and Fatulloh (2017) with the title "Clustering of Social Conditions in Batam, Indonesia Using K-Means Algorithm and Geographic Information System" which groups sub-districts in Batam City based on social indicators such as data poverty, divorce data, percentage of population with the highest high school education, population distribution by age, and morbidity. Based on these indicators, 4 groups were formed with different characteristics in each group. This is similar to the research conducted by Anuraga, Cabral, and Febrianti (2018) which grouped districts and cities in East Java based on poverty indicators. There are eight poverty indicators used and resulted in four groups. In a study conducted by Chusna and Rumiati (2020) entitled "Application of the K-means and Fuzzy C-Means Methods for Grouping Junior High Schools (SMP) in Indonesia based on the Indonesian National Education Standards (SNP) which explains up to testing the grouping accuracy using icdrate (internal cluster dispersion) rate. In his research, it is stated that the K-means clustering method is the best method because it has a smaller icdrate value than the fuzzy c-means method.

2. Research Method

Group analysis is a grouping of data carried out by two kinds of methods, namely hierarchical methods and non-hierarchical methods. The hierarchical method is a grouping carried out in a hierarchical or tiered manner from n, (n-1) to 1 group. The non-hierarchical method is done by first determining the desired number of clusters. One of the well-known non-hierarchical procedures is the K-Means method (Santoso, 2004).

The K-Means algorithm is an algorithm that requires k input parameters and divides a set of n objects into k clusters so that the level of similarity between members in one cluster is high while the level of similarity with members in other clusters is very low. The resemblance of members to the cluster is measured by the proximity of the object to the mean value in the cluster or referred to as the cluster centroid. The K-Means method is the simplest and most common clustering method. This is because K-Means has the ability to group large amounts of data with relatively fast and efficient computation time. Here is the algorithm of the K-Means method:

a. Enter the data to be clustered.
b. Determine the number of clusters.
c. Take any data as much as the number of clusters at random as the center of the cluster (centroid).
d. Calculate the distance between the data and the center of the cluster, using the equation:

$$D(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + \cdots + (x_{ki} - x_{kj})^2} \quad (1)$$

where:
$D(i,j)$ = the distance of data i to cluster center j,
x_{ki} = data i at attribute j,
x_{ji} = center point j, at attribute k

e. Recalculate cluster center with new cluster membership
f. If the cluster center does not change then the cluster process has been completed, if not then repeat step d until the cluster center does not change anymore.

Determination of the best cluster method in this study is to look at the value of the icd rate (internal cluster dispersion) rate. The smaller the ICD rate, the better the grouping results. The icd rate value is the dispersion rate in the cluster, this value can be written with the equation below:

$$icd\ rate = 1 - \frac{\text{SST} - \text{SSW}}{SST} = 1 - \frac{\text{SSB}}{SST} \quad (2)$$

where:

$$SSB = \sum_{j=1}^{c} \sum_{k=1}^{p} (x_{ijk} - \bar{x}_j)^2 \quad (3)$$

Description:

SST = The total sum of the squares of the distances to the overall mean
SSW = Total sum of the squares of the sample distance to the group mean
SSB = *Sum Square Between*
c = Number of variables
p = Number of clusters
x_{ijk} = Sample i at variable j cluster k
$\bar{x}_j$ = The average of all samples on variable j

RESULT AND DISCUSSION

Based on data processing, in underdeveloped areas, there are only two regencies whose percentage of poor people is above the national percentage of poor people, namely Sula Islands Regency and Taliabu Island, both of which are in North Maluku Province. Meanwhile, the area with the highest percentage of poor people is in Dieyai Regency, Papua, which is 41,76 percent. The average percentage of poor people in underdeveloped areas is still very high and far from the national figure of 26,22 percent. Other poverty indicators in underdeveloped areas will be illustrated in the following table:

Table 2. Descriptive Results of the Poverty Indicators Used

Poverty Indiators	Minimum	Maximum	Average	Standard Deviation
Percentage of poor people	7,3	41,76	26,22	8,29
Average daily consumption per capita	1446,17	2840,58	1948,54	274,48
Number of PUSKESMAS per 100.000 population	2,22	32,29	12,17	6,56
Percentage of population 15 years of age and over with a junior high school diploma or below	48,8	89,65	69,61	9,48
Open Unemployment Rate (TPT)	0,21	8,58	3,49	1,79
Percentage of informal labours	7,96	100	76,63	16,01
Percentage of asphalt road length	0	100	49,69	34,50
Percentage of households with own defecation facilities	2,83	97,64	63,34	20,52
Percentage of households with protected cooking/washing water sources	0	88,1	39,48	28,63
Percentage of households with own building ownership	59,51	100	88,55	8,89
Number of dependents	35,1	74,08	53,31	10,42

Before performing the K-Means Clustering analysis, it is necessary to first determine the number of clusters that will be formed. In this study, the determination of the number of clusters that will be formed uses the elbow method and the gap statistic method. From the results of the calculation, 3 clusters were formed as shown in the image below. The picture on the left is the formation of clusters using the elbow method, the fracture that resembles an elbow is at number 3 which indicates that the number of clusters that will be formed is 3 clusters. It is the same as using the gap-statistical method, which together form 3 clusters.

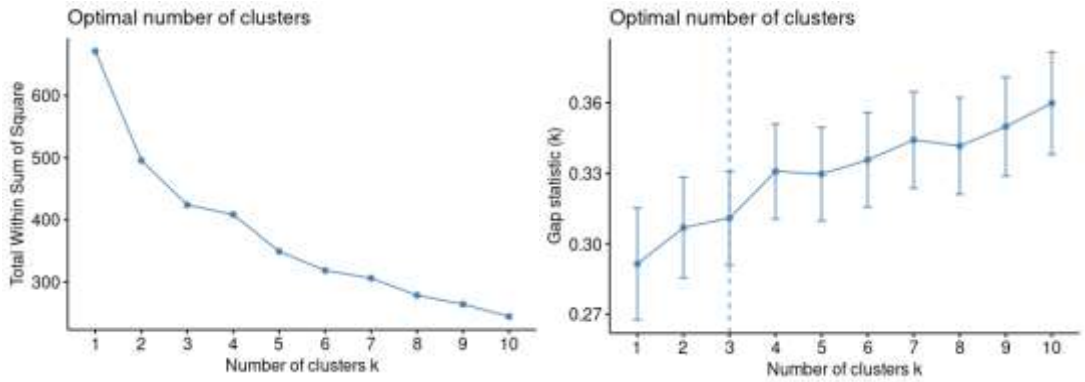

Figure 1. Optimal Cluster Number Calculation Results Using Elbow Method and Gap-Statistic

After determining the number of clusters to be formed, the next step is to determine the members of each cluster. Figure 2 is a grouping of disadvantaged areas based on poverty indicators which are divided into three clusters. In Figure 1 it can be seen that there are 6 members of cluster one, cluster two there are 38 members, and cluster three there are 18 members.

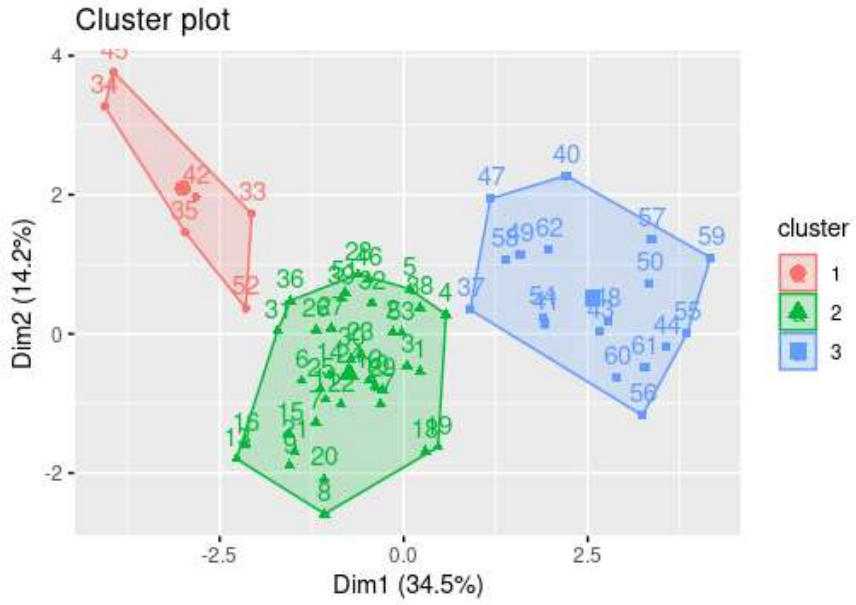

Figure 2. Results of Grouping Underdeveloped Regions Based on Poverty Indicators

Below are the details of the members of each cluster that have been formed using k means clustering

Table 3. Members of Each Cluster Calculation Results of K-Means Clustering

Cluster	Members
I	Teluk Wondama District, Teluk Bintuni, Sorong Selatan, Nabire, Boven Digoel, Waropen,
II	Nias District, Nias Selatan, Nias Utara, Nias Barat, Kepulauan Mentawai, Musi Rawas Utara, Pesisir Barat, Lombok Utara, Sumba Barat, Sumba Timur, Kupang, Timor Tengah Selatan, Belu, Alor, Lembata, Rote Ndao, Sumba Tengah, Sumba Barat Daya, Manggarai Timur, Sabu Raijua, Malaka, Donggala, Sigi, Tojo Una una, Kepulauan Tanimbar, Kepulauan Aru, Seram Bagian Barat, Seram Bagian Timur, Maluku Barat Daya, Buru Selatan, Kepulauan Sula, Pulau Taliabu, Sorong, Maybrat, Manokwari Selatan, Mappi, Keerom, Supiori.
III	Tambrauw District, Pegunungan Arfak, Jayawijaya, Paniai, Puncak Jaya, Asmat, Yahukimo, Pegunungan Bintang, Tolikara, Mamberamo Raya, Nduga, Lanny Jaya, Mamberamo Tengah, Yalimo, Puncak, Dogiyai, Intan Jaya, Deiyai

From the table above, it can be seen that the regions that are members of cluster one are spread out only in the provinces of Papua and West Papua, while regions that are members of cluster two are spread throughout Indonesia. Even in the third cluster, the territory is only spread out in Papua Province. Furthermore, the analysis is carried out by looking at the characteristics of each cluster through its centroid value. Table 4 describes the characteristics of each cluster and its centroid value.

Table 4. Centroid of Each Characteristic of Each Cluster

Cluster	**1**	**2**	**3**
Percentage of poor people	25.28	22.45	34.47
Average daily consumption per capita	2041.06	1970.58	1871.16
Number of PUSKESMAS per 100.000 population	23.52	11.37	10.07
Percentage of population 15 years of age and over with a junior high school diploma or below	53.99	67.63	78.98
Open Unemployment Rate (TPT)	6.03	3.87	1.84
Percentage of informal labours	51.15	73.94	90.77
Percentage of asphalt road length	51.76	65.39	15.85
Percentage of households with own defecation facilities	74.99	69.42	46.63
Percentage of households with protected cooking/washing water sources	40.35	55.15	6.08
Percentage of households with own building ownership	70.86	89.25	92.95
Number of dependents	50.89	54.42	51.75

Based on the table above, it can be seen that cluster three has the highest percentage of poor people and all members of cluster three are districts in Papua Province. This indicates that pockets of poverty are still concentrated in Papua Province. This high poverty is supported by the lowest average calorie consumption, the lowest number of PUSKESMAS per 100,000 population, the highest percentage of the population with the highest education at the maximum junior high school level, the highest percentage of the population working in the informal sector, the lowest percentage of asphalt road length, and the percentage of households with sources of income. main water protected for lowest cooking/washing. This is in line with the theory presented by Harniaty (2010) which states that limited population access to food, education, employment, infrastructure, sanitation, building ownership, and dependents will increase poverty.

The second highest percentage of poverty is in cluster one and members of cluster one are spread across the provinces of Papua and West Papua. In cluster one, the open unemployment rate (TPT) is the highest compared to other clusters. The highest TPT was achieved by Boven Digoel Regency with a TPT of 8,09. Meanwhile other indicators such as the percentage of asphalt roads and the percentage of households with protected main water sources for cooking/washing are also still low.

Cluster two is a cluster with a lower percentage of poor people than other clusters. Cluster two members are the most numerous and are evenly distributed in underdeveloped areas throughout Indonesia. In cluster two, health and sanitation indicators are the number of PUSKESMAS per 100,000 population and the percentage of households with protected water sources for cooking/washing is still low. Education indicators also need attention because the percentage of the population with the highest education at a maximum of junior high school is quite high, this indicates that most of the population in disadvantaged areas in cluster two only reaches junior high school and below. In addition, the indicator of the burden of dependence on the second cluster is the highest, which illustrates that the dependents of the productive age population towards the non-productive age are greater than the other two clusters.

After three clusters have been formed and members have been obtained, testing is necessary to assess how well the clusters have been formed. Testing the formed cluster using the icd rate value. The smaller the ICD rate, the better the grouping results. The icd rate value achieved is 0.208, which means that the cluster formed is quite good

CONCLUSION

From the results of the study formed three clusters, cluster one consists of 6 members, cluster two consists of 38 members, and cluster three consists of 18 members. Cluster one is the cluster with the second highest percentage of poor people. Another poverty indicator that needs attention from the government is the employment aspect, which can be seen from the highest TPT centroid value compared to the other two clusters. Cluster two is a cluster with a lower percentage of poor people than the other two clusters. Other poverty indicators that require sufficient attention from the government are aspects of education and the number of dependency burdens. With the large burden carried by the productive age in cluster two, the government must be able to maximize the potential of the productive age population. Cluster three is the cluster with the highest percentage of poor people and all members of cluster three are scattered only in Papua Province. The government should pay more attention to areas in cluster three where pockets of poverty are still concentrated in Papua.

Suggestions for further research can try other variables so that the accuracy value can be increased.

REFERENCES

BAPPENAS. (2004). *Rencana Pembangunan Jangka Panjang Nasional 2005-2025*. Jakarta: Badan Perencanaan Pembangunan Nasional

Badan Pusat Statistik. (2021). Daerah Dalam Angka 2021. Jakarta: Badan Pusat Statistik

Chusna, H. A., & Rumiati, A. T. (2021). Penerapan Metode K-Means dan Fuzzy C-Means untuk Pengelompoan Sekolah Menengah Pertama (SMP) di Indonesia Berdasarkan Standar Nasional Pendidikan (SNP). Jurnal Sains dan Seni ITS, 9(2), D216-D223.

Dowling, J. M., & Valenzuela, R. J. (2010). Economic development in Asia. Singapore: Cengage Learning.

Febrianti, A. F., Cabral, A. H., & Anuraga, G. (2018). K-Means Clustering Dengan Metode Elbow Untuk Pengelompokan Kabupaten Dan Kota Di Jawa Timur Berdasarkan Indikator Kemiskinan.

Harniaty. (2010). *Program-program Sektor Pertanian yang Berorientasi Penanggulangan Kemiskinan*. Bogor: Pusat Analisis Sosial Ekonomi dan Kebijakan Pertanian Departemen Pertanian.

Kurniawan, D. E., & Fatulloh, A. (2017). Clustering of Social Conditions in Batam, Indonesia Using K-Means Algorithm and Geographic Information System. International Journal of Earth Sciences and Engineering (IJEE), 10(05), 1076-1080.

Prastyo, A. A., & EDY YUSUF, E. Y. (2010). Analisis Faktor-Faktor Yang Mempengaruhi Tingkat Kemiskinan (Studi Kasus 35 Kabupaten/Kota Di Jawa Tengah Tahun 2003-2007 (Doctoral dissertation, UNIVERSITAS DIPONEGORO).

Republik Indonesia. (2020). *Peraturan Presiden Republik Indonesia Nomor 63 Tahun 2020 tentang Penetapan Daerah Tertinggal Tahun 2020-2024*

Sharp, Grimes, P. W., & Register, C. A. (2009). Economics of Social Issues (19th ed.). McGraw-Hill/Irwin Education.