P-ISSN: 2580-1244
E-ISSN: 2580-2305

Classifying regencies and cities on human development index dimensions: Application of K-Means cluster analysis

Nurhasanah Nurhasanah[1*], Nany Salwa[2], Lyra Ornila[3], Amiruddin Hasan[4], Martahadi Mardhani[5]

[1,2,3]Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala (USK), Indonesia
[4]Faculty of Teacher Training and Education, Universitas Syiah Kuala (USK), Indonesia
[5]Faculty of Economics, Universitas Samudra (UNSAM), Indonesia
[*]Corresponding author: nurhasanah@unsyiah.ac.id

ABSTRACT

The Human Development Index (HDI) is a measurement that analyzes a region's development in improving human development. The government's development plan aims to create a successful and peaceful life. The unbalanced development in every regency and city in Indonesia is a typical issue during the development process. It may also be shown that the HDI level changes across regencies and cities in Indonesia. This research aims to identify Indonesian regencies and cities based on HDI indices. K-Means clustering algorithm is the clustering method adopted. The results of the analysis formed 4 clusters. The first cluster consisted of 20 regencies with a low average HDI indicator. The second cluster consisted of 148 regencies and cities with an average HDI indicator is medium. The third cluster consisted of 88 regencies and cities with an average HDI indicator. The fourth cluster consists of 258 regencies and cities with high HDI indicators.

Keywords: Human Development Index, K-Means Cluster Analysis, Regencies and Cities.

INTRODUCTION

Development is one effort made by the government to create a prosperous and peaceful society. Successful development in a region can be measured using the HDI. Indonesia, HDI is calculated as the geometric average of life expectancy index, education index, and gross national income (GNI) index. The life expectancy index of life expectancy at birth, education indices consisting of expected years of schooling and mean years of schooling, and GNI index are GNI per capita indicators.

According to (BPS 2016), in 2015, the regencies and cities with the highest HDI score in Indonesia were the city of Yogyakarta with the HDI value of 84.56, while the regencies and cities with the lowest HDI score are Nduga Regency with the HDI score of 25.47. There is a considerable distance between the highest and lowest HDI values. This condition illustrates the unevenness of human development in Indonesia, and also, there is a thigh gap. Several ways can overcome development gaps; one of them is by knowing the characteristics and factors that affect the difference in the value of HDI.

Research to determine the factors that affect the HDI has been done by many researchers before. One research on factors affecting HDI conducted by Putra (2015) states that the variables affecting HDI in East Java are infant mortality rate, high school enrollment rate, and number illiterate. Trianggara (2015) researched HDI modeling using a fixed-effect spatial panel. The results stated that the variables affecting HDI are per capita GRDP. In addition, according to (BPS 2011), factors affecting HDI include school enrollment rates, poverty rates, and the percentage of birth attendants with medical personnel.

Based on the above description, the K-Means cluster analysis method identifies the characteristics of regencies and cities in Indonesia based on factors affecting the HDI. K-means cluster analysis is one cluster method that aims to categorize objects based on the similarity of characteristics possessed by the object. K-Means cluster analysis can see the characteristics of regencies and cities in Indonesia through the formed groupings. According to some previous studies, regencies and cities grouping are based on HDI factors. After this first part is followed by section 2 literature review, section 3 methods, section 4 results and discussion, and section 5 are conclusions.

LITERATURE REVIEW

Research on cluster analysis has been widely discussed in grouping various cases worldwide. For example, Chan et al. (2008) identified the health profile of a group of Hong Kong Chinese and Liu (2011) who investigated the typology of the fiscal decentralization system. We used K-Means cluster analysis (Macqueen, 1967). The application of K-Means cluster analysis has been carried out in various studies. Among them, Castro et al. (2018) which identifies logistics and city categorization by HDI and taxes in São Paulo; Pejić Bach et al. (2018) examine the economic cost impact of violence on 119 countries worldwide; and Sharkh & Gough (2010), who grouped a large number of developing countries based on their level of welfare and stability.

Cluster analysis is one interdependence technique that categorizes objects based on similar characteristics. The object characteristics used in the cluster analysis are called variables. Variables in cluster analysis should not strongly correlate with each other (multicollinearity). A measure of distance can measure the similarity of characteristics between objects. The measure of the distances contained in the cluster analysis is Euclidean distance, Mahalanobis distance, and City-block distance (Mattjik & Sumertajaya, 2011). In this study, the distance used is the Euclidean distance. Euclidean distance between two P and Q objects with coordinates $P = (x_1, x_2, \dots, x_p)$ and $Q = (y_1, y_2, \dots, y_p)$ and measured using Equation 1 (Johnson & Wichern, 2014).

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + \left(x_p - y_p\right)^2}$$

Cluster analysis, in general, can be approached by three methods: hierarchical method, non-hierarchical method, and method by combining hierarchical and non-hierarchical methods (Hair et al., 2010). In this research, the clustering method used is hierarchical. Grouping with hierarchical methods is faster and easier if using large samples (Mattjik & Sumertajaya, 2011). One method in the non-hierarchical grouping is the K-Means cluster method.

K-Means clustering analysis was first introduced by Macqueen (1967). The K-Means cluster method first determines the number of groups formed by partitioning *n* objects into *k* groups. The center point of the cluster is derived from the average object value in the cluster (Han et al., 2012). There are stages suggested by Macqueen (1967) in the grouping of objects:

1. Determining the number of clusters to be formed.
2. Determining the center point of the cluster.
3. Apply the objects closest to the center point of the cluster by using the Euclidean distance.
4. Re-establish the center point for each newly formed cluster.
5. Repeat step 2 until no object moves from one cluster to another (Johnson & Wichern, 2014).

METHODOLOGY

Data dan Variables

The data used in this study is the data indicator HDI 2015 in every regency and city in Indonesia. Data were obtained from the BPS website of every province in Indonesia. The amount of data used is 514 regencies and cities. The variable used consists of 10 independent variables. The definition of the variables used can be seen in Table 1.

Table 1. List of Used Variables

Variable	Explanation
X_1	Expectancy year school (year)
X_2	Mean year school (year)
X_3	Senior high school participation rate (%)
X_4	Percentage of population who have the highest diploma S1/S2/S3 (%)
X_5	Life expectancy rate (year)
X_6	Percentage of birth attendants by medical personnel (%)
X_7	Percentage of toddlers who received measles immunization toddler (%)
X_8	Per capita expenditure (Rupiah)
X_9	Percentage of poor people (%)
X_{10}	Percentage of the open unemployment rate (%)

Data Analysis Procedure

The steps taken in this research are as follows:

1. Conducting a descriptive statistical analysis by displaying the table.
2. Standardize data.
3. Conduct a multicollinearity assumption test by looking at a VIF value greater than 10.
4. They are forming four groups with the K-Means cluster method.
5. Conduct an interpretation for each group formed.

P-ISSN: 2580-1244
E-ISSN: 2580-2305

EMPIRICAL RESULTS

Descriptive Statistics

To see the description of each variable descriptive analysis is used. Descriptive statistical value for each variable used in this research can be seen in Table 2. Based on Table 2, there is a considerable distance between each variable's maximum and minimum values.

Table 2. Descriptive Statistics

Variables	Mean	Minimum	Maximum
X_1	12.38	2.17	17.01
X_2	7.86	0.64	12.38
X_3	72.65	20.82	94.28
X_4	5.71	0.00	21.23
X_5	68.71	53.60	77.46
X_6	85.86	0.11	100.00
X_7	76.63	9.63	100.00
X_8	9,399,676	3,625,000	22,425,000
X_9	12.62	1.68	45.74
X_{10}	5.62	0.34	17.26

Multicollinearity Test

One way to detect multicollinearity is to look at Variance Inflation Factor (VIF) value. VIF values for each variable can be seen in Table 3. Multicollinearity occurs when a variable has a VIF value greater than 10. Based on Table 3, it can be seen that no variable has a large VIF value of 10. So it can be said that there is no multicollinearity between these variables.

Table 3. VIF Values

Variable	VIF
X_1	3.333
X_2	4.049
X_3	2.005
X_4	2.252
X_5	1.724
X_6	2.242
X_7	1.611

X_8	2.742
X_9	1.688
X_{10}	1.316

K-Means Cluster Analysis

The clustering of regencies and cities in Indonesia using K-Means cluster analysis was formed into 4 clusters. Cluster 1 consists of 20 regencies, including Sampang Regency (East Java), several regencies in West Papua Province, including South Manokwari Regency, Arfak Regency, and Tambrau Regency 16 regencies in Papua Province, including Nduga Regency, Puncak, Puncak Jaya. Cluster 2 consists of 148 regencies and cities.

Most regencies in the second cluster are located in East Nusa Tenggara Province. One city is located in cluster 2; the city is Subulussalam, located in Aceh Province. Cluster 3 consists of 88 regencies and cities dominated by cities in Indonesia. Where in cluster 3, there are 77 cities and 11 regencies. The regency in cluster 3 are Aceh Besar, Natuna, Sukoharjo, Karanganyar, Sidoarjo, Sleman, Bantul, Badung, Gianyar, Malinau and Minahasa regency. Cluster 4 consists of 258 regencies and cities dominated by regencies in Indonesia. Where in cluster 4, there are 245 regencies and 13 cities. The cities that are in the cluster 4 are Tebing Tinggi, Gunungsitoli, Pagar Alam, Cilegon, Serang, Banjar, Tasikmalaya, Tegal, Pekalongan, Batu, Probolinggo, Singkawang, Tarakan, Kotamobagu, and Tidore Islands.

Cluster Interpretation

Characteristics of each cluster formed can be seen based on the average value of each cluster variable. The average value of each variable for each cluster can be seen in Table 4. Characteristics of each cluster based on Table 4 are that cluster 1 has the highest percentage of poor people (X9), and the other variable are very low. Cluster 2 has a high percentage of poor people (X9) and relatively moderate variables. Cluster 3 has the lowest percentage of poor people (X9), and other variables are very high. Cluster 4 has a relatively average percentage of poor people (X9), and other variables are relatively high.

Table 4. The Average Value of each Cluster Variable

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
X_1	8.166(4)	11.959(3)	14.077(1)	12.366(2)
X_2	3.604(4)	7.014(3)	10.294(1)	7.852(2)
X_3	49.563(4)	70.429(3)	81.961(1)	72.546(2)
X_4	1.842(4)	4.462(3)	11.422(1)	4.775(2)
X_5	63.009(4)	65.575(3)	72.018(1)	69.829(2)
X_6	29.434(4)	75.144(3)	96.453(1)	92.774(2)
X_7	33.815(4)	70.731(3)	82.518(1)	81.321(2)
X_8	5,061,590.000(4)	7,761,693.716(3)	12,957,595.796(1)	9,462,025.116(2)
X_9	35.628(1)	15.412(2)	7.381(4)	11.175(3)
X_{10}	1.758(4)	4.594(3)	7,971(1)	5,703(2)

(1)The average value of the variables is ranked 1

The value for each variable in cluster 3 indicates that this cluster is better than the other cluster. The value for each variable in cluster 4 indicates that cluster 4 is better than clusters 1 and 2 but not better than cluster 3. The value for each variable in cluster 2 indicates that cluster 2 is better than cluster 1 but not better than clusters 3 and 4.

CONCLUSION

In this article, clustering 514 regencies and cities in Indonesia based on HDI dimension using the K-Means cluster analysis to form 4 clusters. Cluster 1 consists of 20 regencies, namely several regencies located in East Java Province, Papua Province, and West Papua Province. Cluster 2 consists of 147 regency and 1 city, namely Subulussalam city. Cluster 3 consists of 77 cities and 11 regencies. The regency in cluster 3 are Aceh Besar, Natuna, Sukoharjo, Karanganyar, Sidoarjo, Sleman, Bantul, Badung, Gianyar, Malinau, and Minahasa regency. Cluster 4 consists of 245 regency and 13 cities. The cities in the cluster 4 are Tebing Tinggi, Gunungsitoli, Pagar Alam, Cilegon, Serang, Banjar, Tasikmalaya, Tegal, Pekalongan, Batu, Probolinggo, Singkawang, Tarakan, Kotamobagu, and Tidore Islands.

Our findings indicate the importance of cluster analysis. Characteristics of cluster 1 have factors that influence HDI are lowest compared to other clusters. Variables in cluster 1 need to be increased, except the percentage of the poor and open unemployment rates variable. Characteristics of cluster 2 are clusters with averages of each variable better than cluster 1 but not better than clusters 3 and 4. Characteristics cluster 3 is the cluster that has the average of each variable that is the highest compared to the other group. Cluster 3 is the best.

REFERENCES

BPS. (2011). *Indeks Pembangunan Manusia 2009-2010: Keterkaitan antara IPM, IPG, dan IDG*. Badan Pusat Statistik.

BPS. (2016). *Indeks Pembangunan Manusia 2015*. Badan Pusat Statistik.

Castro, R. B., Merchán, D., Orlando JR, F. L., & Winkenbach, M. (2018). City Logistics and Clustering: Impacts of Using HDI and Taxes. In E. Taniguchi & R. G. Thompson (Eds.), *City Logistics 2: Modeling and Planning Initiatives*. ISTE Ltd and John Wiley & Sons.

Chan, M. F., Mok, E., Wong, T. K. S., Lee, R. L. T., & Fok, M. S. M. (2008). Investigating the health profiles of Hong Kong Chinese: Results of a cluster analysis. *Journal of Clinical Nursing*, *17*(7), 911–920. https://doi.org/10.1111/j.1365-2702.2007.01932.x

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis : A global edition*. Prentice Hall International.

Han, J., Pei, J., & Kamber, M. (2012). *Data Mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.

Johnson, R. A., & Wichern, D. W. (2014). *Applied Multivariate Statistical Analysis* (6th ed.). Pearson Prentice Hall.

P-ISSN: 2580-1244
E-ISSN: 2580-2305

Liu, L. C. hung. (2011). The typology of fiscal decentralization system: A cluster analysis approach. *Public Administration and Development*, *31*(5), 363–376. https://doi.org/10.1002/pad.605

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Berkeley Symposium on Mathematical Statistics and Probability*, *1*(14), 281–297.

Mattjik, A. A., & Sumertajaya, I. M. (2011). *Sidik Peubah Ganda dengan Menggunakan SAS*. IPB Press.

Pejić Bach, M., Dumičić, K., Jaković, B., Nikolić, H., & Žmuk, B. (2018). Exploring impact of economic cost of violence on internationalization: Cluster analysis approach. *International Journal of Engineering Business Management*, *10*, 1–15. https://doi.org/10.1177/1847979018771244

Putra, D. M. (2015). *Pemodelan Indeks Pembangunan Manusia (IPM) Provinsi Jawa Timur dengan menggunakan Metode Regresi Logistik Ridge*. Institut Teknologi Sepuluh Nopember.

Sharkh, M. A., & Gough, I. (2010). Global welfare regimes: A cluster analysis. *Global Social Policy*, *10*(1), 27–58. https://doi.org/10.1177/1468018109355035

Trianggara, N. (2015). *Pemodelan Indeks Pembangunan Manusia Menggunakan Spatial Panel Fixed Effect*. Universitas Diponegoro.