International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN 2964-3139 e-ISSN (on Process)
Vol. 1, No. 1, page 66 - 70
Submitted 31/12/2022; Accepted 02/01/2023; Published 14/01/2023

K-Means Cluster Algorithm for Grouping Inequality in Regional Development

[1]Tb Ai Munandar and [2]Dwipa Handayani
[1,2] Informatics Department, Universitas Bhayangkara Jakarta Raya, INDONESIA
e-mail : [1]tbaimunandar@gmail.com, [2]dwipa.handayni@dsn.ubharajaya.ad.id

Corresponding Autor: Tb Ai Munandar

Abstract
Unsupervised learning is a subset of machine learning. Many unsupervised learning algorithms are used to solve various problems, especially the extraction of hidden data patterns. One of the problems that concerns unsupervised tasks is clustering. Clustering techniques are widely used for data grouping needs, one of which is development inequality clustering. The classification of development inequality is an important consideration in a country's regional development strategy. However, development groupings often do not pay attention to the hidden information aspects of the data, so they do not get the appropriate results. This research was conducted to provide an additional alternative in the realm of development inequality clustering, namely by classifying development data using the k-means algorithm. The data used is GRDP data for 41 regions in the western part of Java Island for the 2010–2021 range. The results show that the forty-one regions are grouped into four clusters. The first cluster (C1) contains 35 regions, the second cluster (C2) contains three regions, the third cluster (C3) contains four regions, and the fourth cluster (C4) contains three regions. Based on the cluster results, it can be seen that all members of cluster C4 are areas with the best level of development, while cluster C1 is the area with the lowest level of development. As for clusters C2 and C3, these are areas with development levels between clusters C1 and C4. The grouping results can be used by policymakers or local governments to determine the direction of future development priorities based on the cluster with the lowest level of development.

Keywords— development inequality, clustering, k-means algorithm, hidden information, development priorities.

1 Introduction

Equitable development is a global problem that Indonesia is also facing. Unequal development is synonymous with development inequality between regions, both at the provincial and regency levels and down to the level of smaller administrative areas. The government continues to seek appropriate techniques and formulations to reduce disparities in regional development, including formulating various policies related to future development directions and priorities. In the western part of Java Island in particular, disparities between regions are still quite high. Therefore, it is very necessary to identify the inequalities that occur so that policymakers can determine more targeted development policy directions. Not only that, but it is critical to understand the pattern of close inequality between regions so that dealing with unequal areas that are close together can become a special concern for policymakers. On the other hand, by looking at the closeness of patterns of inequality between regions, policy makers can find out which administrative areas have similar inequality when compared to other regions.

In the regional economic realm, there are many techniques that can be used to identify development disparities between regions. The Klassen typology, Williamson index, and shift-share analysis, for example [1], [2], [3], [4]. These three techniques, however, have flaws against one another. First, classify inequality based on the value of the growth rate and development contribution according to the GRDP value it has. The rate of growth and contribution of a region depend on the region above it. Second, to be able to produce a complete interpretation, for example, to

obtain information on areas that have high inequality along with the superior resource sectors they have, different techniques must be used. A different approach is needed so that the process of analyzing development data is not interdependent with data from other areas while at the same time being able to provide complete information without having to use different techniques. so that inequalities between regions as well as leading sectors can be presented in one complete visualization.

Research in the field of clustering techniques for solving regional economic problems is not new. Based on the literature studies conducted, many scientists use a clustering approach to analyze and identify regional development disparities not only in Indonesia but also in various parts of the world. Fuzzy cluster means and hierarchical cluster techniques are used to analyze regional inequality in Ukraine based on economic activities, such as industry, agriculture, construction services, and public services [5]. Different clustering techniques are used to classify regional developmental disparities in Bangladesh using k-means and Partition Around Medoid (PAM). In this study, five indicators of maternal and child health were used [6]. The k-means technique was used in research [7] to classify levels of infrastructure development in Uttar Pradesh and divide them into five clusters. From 2003 to 2012, researchers all over the world used cluster techniques to categorize development inequality, including in Portugal [8], Croatia [9], European Union countries [10], West and East Germany [11], the Czech Republic [12], Ukraine [13], and Pakistan [14]. In Indonesia itself, the cluster technique has been used to analyze development inequality, as was done by [15].

2 Research methods

The research to be carried out is arranged in stages for a period of one semester. The data used is data on district and city development achievements in three provinces, namely West Java, the Jakarta Capital Special Region, and Banten Province. The data was taken from the official website of the Central Statistics Agency for each of the three provinces. Broadly speaking, the research that will be carried out includes several stages, namely: (1) study of the literature and identification of problems, particularly with regard to information on current development inequality; (2) data collection from the official BPS website for each region in the three provinces; (3) clustering of inequality data using the k-means algorithm; (4) validity testing of cluster results (validation of cluster results to ensure each region is correctly clustered); and (5) interpretation of the results of inequality grouping, including comparison of the results of the k-means cluster with the raw data of the region.

3 Results and Discussion

In the research conducted, the data was pre-processed before being grouped using the cluster technique. The pre-processing process is carried out in the form of data normalization for each attribute value. This is because there are several values that have quite a wide range between one attribute and another. The number of k for the k-means algorithm in this study is set to k = 4. Apart from that, another parameter that is set is the distance calculation technique. The Euclidean distance technique was used to calculate the distance between the data in this study. GRDP data for forty-one urban districts in three provinces of the western part of Java Island are clustered using the k-means algorithm. The results of the clustering show that most of the regions are grouped into cluster 1 (C1), namely 31 regions. The second group (C2) contains three regions as cluster members; the third group (C3) consists of four regions; and the fourth group (C4) consists of three regions. Figure 1 is a graph of the distribution of cluster members for the 41 regions. Meanwhile, Table 1 provides regional distribution information for each cluster.

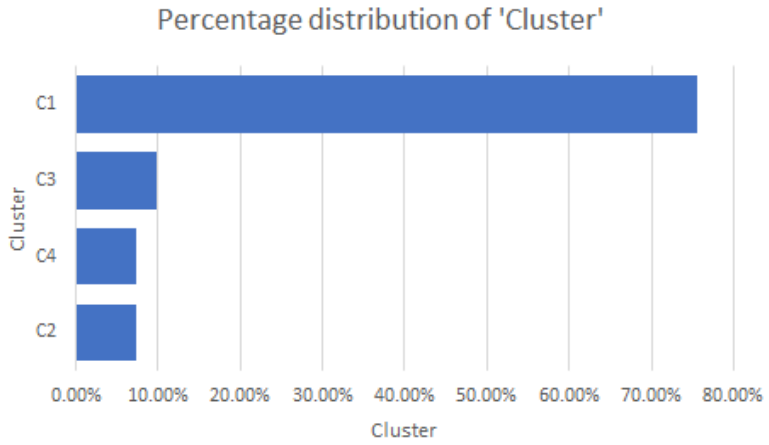

Figure 1. Percentage Distribution of Cluster Members

Cluster validity tests are conducted to determine the extent to which each member of the cluster is well clustered. The silhouette index is used to test cluster validity. The Silhouette index range with good clusters is close to 1, and the further away from 1 the cluster members are, the less properly grouped they are. Based on the calculation of the silhouette index, for the first cluster (C1), most of the areas belonging to Cluster C1 can be said to be properly grouped, except for the Tangerang Regency area. Almost all cluster members at C1 have values close to 1, while Tangerang Regency is far from 1. Thus, it can be said that Tangerang Regency is not properly grouped at C1. In cluster C2, the Silhouette Index values for each cluster member (region) are mostly close to 1, meaning that this cluster already has appropriate members. The same thing happens for clusters C3 and C4, where the Silheoutte Index value for each cluster member is close to 1. Figure 2 is a visualization for visualizing the silhouette index for each cluster.

C1 cluster members are mostly regions with an average gross regional domestic product (GRDP) ranging from 2.7 million to 76,528 million. The C2 cluster is a region with an average value of GRDP between 209,712 and 275,148 million. As for C3, it is an area with an average GRDP value between 91,505 million and 154,994 million, while C4 is an area with an average GRDP between 349,947 million and 366,594 million. According to the findings of this cluster, regions in cluster C4 have significantly higher rates of growth and development than regions in clusters C3, C2, and C1.

Table 1: Distribution of clusters per region

Name of Regions	Cluster	Silhouette	Name of Regions	Cluster	Silhouette
City of West Jakarta	C4	0.644376	Cianjur Regency	C1	0.723388
City of Central Jakarta	C4	0.684319	Cirebon Regency	C1	0.721928
City of South Jakarta	C4	0.675644	Garut Regency	C1	0.71751
Tangerang City	C3	0.578399	Indramayu Regency	C1	0.670058
Bogor Regency	C3	0.707333	Kuningan Regency	C1	0.720435
Karawang Regency	C3	0.710341	Majalengka Regency	C1	0.723846
Bandung City	C3	0.677775	Pangandaran Regency	C1	0.713893
City of East Jakarta	C2	0.683035	Purwakarta Regency	C1	0.710716
City of North Jakarta	C2	0.632583	Subang Regency	C1	0.723045
Bekasi Regency	C2	0.60133	Sukabumi Regency	C1	0.710858
Lebak Regency	C1	0.724386	Sumedang Regency	C1	0.724319
Pandeglang Regency	C1	0.724004	Tasikmalaya Regency	C1	0.724912
Serang Regency	C1	0.696858	Banjar City	C1	0.711359
Tangerang Regency	C1	0.527907	Bekasi City	C1	0.659674
Cilegon City	C1	0.636143	Bogor City	C1	0.723505
Serang City	C1	0.724767	Cimahi City	C1	0.724886
South Tangerang City	C1	0.694294	Cirebon City	C1	0.722232
Kepulauan Seribu Regency	C1	0.712742	Depok City	C1	0.71062
Kabupaten Bandung	C1	0.603146	Sukabumi City	C1	0.716201
West Bandung Regency	C1	0.723533	Tasikmalaya City	C1	0.721328
Ciamis Disctrict	C1	0.724847			

Cluster validity tests are conducted to determine the extent to which each member of the cluster is well clustered. The silhouette index is used to test cluster validity. The Silhouette index range with good clusters is close to 1, and the further away from 1 the cluster members are, the less properly grouped they are. Based on the calculation of the silhouette index, for the first cluster (C1), most of the areas belonging to Cluster C1 can be said to be properly grouped, except for the Tangerang Regency area. Almost all cluster members at C1 have values close to 1, while Tangerang Regency is far from 1. Thus, it can be said that Tangerang Regency is not properly grouped at C1. In cluster C2, the Silhouette Index values for each cluster member (region) are mostly close to 1, meaning that this cluster already has appropriate members. The same thing happens for clusters C3 and C4, where

the Silheoutte Index value for each cluster member is close to 1. Figure 2 is a visualization for visualizing the silhouette index for each cluster.

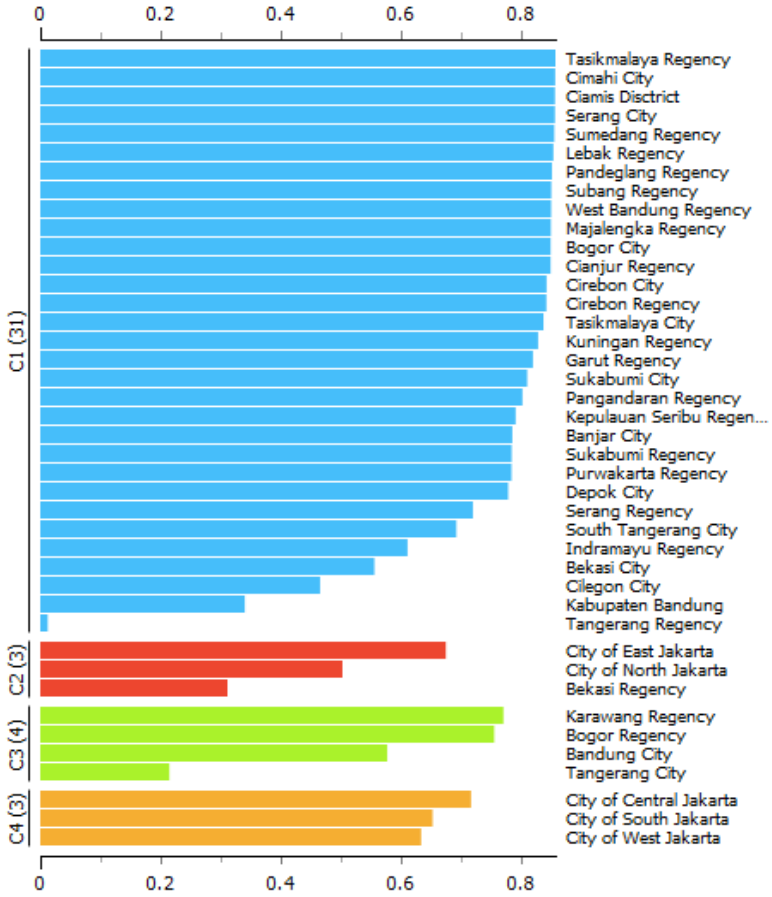

Figure 2 : Silhouette Index Distribution for Each Cluster

C1 cluster members are mostly regions with an average gross regional domestic product (GRDP) ranging from 2.7 million to 76,528 million. The C2 cluster is a region with an average value of GRDP between 209,712 and 275,148 million. As for C3, it is an area with an average GRDP value between 91,505 million and 154,994 million, while C4 is an area with an average GRDP between 349,947 million and 366,594 million. According to the findings of this cluster, regions in cluster C4 have significantly higher rates of growth and development than regions in clusters C3, C2, and C1.

4 Conclusion

Based on the results of the research that has been done, it can be concluded that the 41 regions in the western part of Java are divided into four clusters. The C4 cluster is a region with the best growth rate among other clusters. Cluster C1 is the region with the lowest growth rate compared to other regions in different clusters. This can be seen from the average GRDP value of each region based on the clusters formed. In cluster C1, there is also one area that is indicated as not clustered in the right cluster, while the rest of the Silhouette Index values are close to 1. This means that each member of the cluster has been grouped according to their respective clusters.

ACKNOWLEDGMENT

This research was supported by a grant from the Institute for Research, Community Service, and Publications, funded by Universitas Bhayangkara Jakarta Raya.

BIBLIOGRAPHY

[1] . G.O. Naibaho, J.R. Mandei, and L.R.J. Pangemanan, Analisis Ketimpangan Pembangunan dan Pertumbuhan Ekonomi Antar Wilayah Kabupaten/Kota Di Provinsi Sulawesi Utara, *Jurnal Agri-Sosio Ekonomi Unsrat*, Volume 16 Nomor 3, hal. 369 – 378, 2020, *in Bahasa*

[2] . M.J. Darmawan, dan Tukiman, Analisis Dimensi Ketimpangan Pembangunan Antar Wilayah Di Provinsi Jawa Timur Tahun 2014-2018, Jurnal Dinamika Governance: *Jurnal Ilmu Administrasi Negara*, Volume 10 Nomor 1, 2020, *in Bahasa*

[3] . R.H. Harahap, , H.B. Isyandi dan E.K. Pailis, Analis Pertumbuhan Ekonomi dan Ketimpangan Antar Kabupaten Hasil Pemekaran Wilayah Indragiri (Kabupaten Indragiri Hulu, Kabupaten Indragiri Hilir, Kabupaten Kuantan Singingi), *Pekbis Jurnal*, Vol.12, No.3, hal. 183 – 193, 2020, *in Bahasa*

[4] . Kadriwansyah, B. Semmaila, and J. Zakaria, Analisis Ketimpangan Wilayah di Provinsi Sulawesi Selatan Tahun 2014-2018, *PARADOKS: JURNAL ILMU EKONOMI* Volume 4.Nomor 1, hal. 25 – 36, 2021, *in Bahasa*

[5] . K. Gorbatiuk, O. Mantalyuk, O. Proskurovych, and O.V. Alkov, Analysis of Regional Development Disparities in Ukraine with Fuzzy Clustering Technique, *HS Web of Conferences* 65, 04008 (2019), https://doi.org/10.1051/shsconf/20196504008, 2019,

[6] . E. Raheem, J.R. Khan, and M.S. Hossain, Regional disparities in maternal and child health indicators: Cluster analysis of districts in Bangladesh, *PLoS ONE* 14(2): e0210697. https://doi.org/10.1371/journal.pone.0210697, 2019

[7] . M. Dube, S.K. Yadav, and V. Singh, Uncovering Regional Disparities in Infrastructural Development of Uttar Pradesh: An Exploratory Factor Analysis*, Journal of Reliability and Statistical Studies*, Vol. 15, Issue 1 (2022), hal. 21–36, doi: 10.13052/jrss0974-8024.1512, 2022

[8] . J.O. Soares, M.M.L. Marques, and C.M.F. Monteiro, A Multivariate Methodology To Uncover Regional Disparities: A Contribution To Improve European Union And Governmental Decisions, *European Journal of Operational Research* 145 (2003) 121–135, 2003

[9] . L.R. Bakaric, Uncovering Regional Disparities – the Use of Factor and Cluster Analysis, *Economic Trends and Economic Policy*. No. 105 , pp. 52-77, 2005,

[10] . M. Lukovics, Measuring Regional Disparities on Competitiveness Basis. *JATEPress*, Szeged, pp. 39-53, 2009

[11] . F. Kronthaler, *A Study of the Competitiveness of Regions based on a Cluster Analysis: The Example of East Germany*, Research Report of Institute for Economic Research Halle (IWH), 2003

[12] . H.V. Vydrová, and Z. Novotná, Evaluation Of Disparities In Living Standards Of Regions Of The Czech Republic, *Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis*, Volume LX 42 Number 4, 2012

[13] . O. Nosova, The Innovation Development in Ukraine: Problems and Development Perspectives, International *Journal Of Innovation And Business Strategy*, Vol. 02/August, , 2013

[14] . S. Ramzan, M.I. Khan, and F.M. Zahid F.M., Regional Development Assessment Based on Socioeconomic Factors in Pakistan Using Cluster Analysis, World Applied Sciences Journal 21 (2): 284-292, 2013

[15] . A. Widodo, and Purhadi, Perbandingan Metode Fuzzy C-Means Clustering dan Fuzzy C-Shell Clustering (Studi Kasus: Kabupaten/Kota di Pulau Jawa Berdasarkan Variabel Pembentuk Indeks Pembangunan Manusia). Tesis Magister Statistika, FMIPA-ITS, 2012, *in Bahasa*