Data Mining for Development Inequality

Tb Ai Munandar, Dwipa Handayani

Abstract— **Development inequality occurs in many parts of the world and is a global problem. Each country has made many efforts to reduce development inequality, especially developing countries. They are grouping the development inequality between regions to be used for future development decisions. However, the group often does not provide a complete representation. Many techniques are commonly used to classify development inequalities, such as the Williamson Index, Klassen typology, and shift-share analysis. But all three of them are not sufficiently able to visualize the proximity of inequality within and between regions. On the other hand, the development of computational science, such as data mining, can be used for data grouping tasks. This research intends to provide another alternative to grouping development inequalities using clustering tasks. The hierarchical agglomerative clustering (HAC) approach was used to group the GRDP data for 41 regions in the western part of Java Island, Indonesia. The 41 regions are spread out over three provinces: West Java, the Special Capital Region of Jakarta, and Banten. In this research, we use HAC techniques, namely single, average, complete, and ward linkage. The results showed that the single and average linkage techniques performed better than the other two. Single and average linkage performance can be seen from the silhouette coefficient of each cluster member, which is closer to +1. The cluster results can also show where each region stands in terms of development inequality. We divided cluster results into three groups to facilitate knowledge representation. Based on the average linkage technique, we have 35 regions that are members of the C1 cluster group; 3 regions are members of the C2 cluster group; and the other 3 regions are C3.**

Index Terms— **development inequality, regional economy, hierarchical agglomerative clustering, data mining, knowledge representation**

I. Introduction

Inequality of development is a global problem that is not only faced by Indonesia but also by other countries in various parts of the world. Inequality occurs because of the inequality of the development process in an area. It starts from the province and Regency to the lowest administrative area level. Various efforts continue to be made to reduce the inequality that occurs. The search for techniques, formulations, and policies related to the direction of development priorities in the future continues to be pursued. In Indonesia, development inequality is still relatively high in various provinces, especially in the western part of Java Island. To reduce the inequality level that occurs, one of the efforts that can be done is the early identification of inequality itself so that policymakers can make appropriate development decisions, for example, with a data mining approach.

Data mining has been widely used in various fields. Several data mining applications include analyzing the risk of ship collisions in marine traffic [1], characterizing forms of corporate communication based on social media [2], and even finding patterns of relationship between a feature and certain conditions (such as the relationship between reading activity and abstraction) through the amount of text and time spent reading [3]. Specifically, data mining has several tasks, including forecasting, classification, and clustering. Forecasting is intended to project future events. Several forecasting applications include predicting the demand for spare parts for inventory management needs [4] and the academic grades of college students [5]. Classification aims to group data into clear classes. Classification is widely used for various needs, such as classifying types of heart disease based on ECG time series data [6], classifying X-ray images to identify whether there are lesions so that additional diagnostic results are obtained [7], classifying types of diabetes mellitus [8], and classifying incidents in the software service distribution process [9]. In contrast to classification, clustering is intended to classify data into groups that do not yet have a clear label.

There are lots of clustering task implementations in data mining. Some of them are to solve the problem of grouping digital images for various needs. [10], [11], [12], [13], [14], post-earthquake road restoration investigations [15], [16], long-term wind speed estimates. [17], [18], modeling of the troposphere's refractive profile [19], grouping social media user behavior [20], grouping hashtags based on properties owned [21], product grouping at grocery stores [22], grouping customer behavior for marketing strategy and customer relationship maintenance [23], heart rate variability clustering [24], damage breakdown and bridge structural degradation for prediction of bridge status [25], high-dimensional data stream clustering [26], industry-standardized service modules [27], short text clustering based on semantic vectors [28], and information extraction from documents [29].

This study focuses on clustering tasks. Given the wide range of applications that clustering can solve, it is possible that the grouping of regional development inequalities can be resolved as well. In this case, inequality is grouped by extracting patterns from development data so as to form groups of development inequality between regions. Policymakers can use the clustering results to determine

Manuscript received October 13, 2022; revised March 18, 2023.

This work was supported in part by Institute for Research, Community Service and Publications, funded by Universitas Bhayangkara Jakarta Raya, INDONESIA.

Tb Ai Munandar is a lecturer and young researcher at informatics department of Universitas Bhayangkara Jakarta Raya, Indonesia (e-mail: tbaimunandar@gmail.com).

Dwipa Handayani is a lecturer in the informatics department and currently serves as vice dean for human resources at the Faculty of Computer Science at Universitas Bhayangkara Jakarta Raya, Indonesia (e-mail: dwipa.handayani@dsn.ubharajaya.ac.id).

which areas have development inequality and which have the same inequality.

The discussion of development inequality in regional economics is nothing new. Various methods are used to classify development inequalities based on a region's gross regional domestic income (GRDP). For example, using Klassen typology, Williamson Index and shift-share analysis [30], [31], [32], [33]. However, the three previously mentioned techniques have their drawbacks. First, the grouping of development inequalities is done by looking at the growth rate value and the contribution of development based on the regional GRDP value. The growth rate and development contribution depend on the value of the GRDP of the administrative region above it. Growth rates and development contributions are generally only based on two years of data (the current year and the previous year). In fact, the GRDP used to calculate the growth rate and development contribution is a time series of data that can be analyzed as a whole without being limited by the number of years of data. Second, the interpretation of the grouping results with the three approaches is incomplete. A combination of grouping results between techniques is needed to get an in-depth interpretation. Therefore, another approach is needed so that the identification of development imbalances can be carried out without having to depend on the GRDP value of other administrative regions directly and provide complete information in one execution..

II. Related Works

Research on the classification of development inequalities using data mining is not a new thing. According to the findings of the literature review, there have been many articles related to data mining for grouping regional development inequalities, such as the use of fuzzy cluster means (FCM) and hierarchical clusters to classify regional development inequality in Ukraine. The indicators used are economic activities such as industry, agriculture, construction services, and public services [34]. Another method is to use the k-means algorithm and partition around medoid (PAM) to classify development inequality in Bangladesh [35], Uttar Pradesh [36], Portugal [37], Croatia [38], European Union countries [39], West and East Germany [40], the Czech Republic [41], Ukraine [42], Pakistan [43], and even Indonesia [44]. This research indicates that the concept of data mining can be widely used, even to solve the problem of development inequality. The cluster technique is widely used to group development inequality data because it can learn from the dataset provided without having to have a cluster label first. The clustering technique studies the patterns and characteristics of the data and then groups them based on their similarity and dissimilarity.

III. Development Inequality

Inequality of development is the difference between the development achievements of a region and those of other regions as seen from the level of the economy and welfare [45]. Differences in conditions between regions are also the cause of development inequality. In general, there are two perspectives on development inequality. The first perspective views inequality as a change from one region to another regarding geographic size (vertically). At the same time, the second perspective looks at the level of social life in the community, the economy, and the conditions in an area. The second perspective is more widely used in the field to measure the existence of development inequality.

In general, regional inequality can be grouped using various approaches such as the Klassen typology, shift-share analysis, and the Williamson index. Inequality in development follows its own pattern and differs depending on the approach used. The Klassen technique classifies regional development inequality into four quadrants. The first quadrant is an area that has a tendency to progress and grow rapidly. The second quadrant represents a regional pattern with rapid development. The third quadrant is an advanced but depressed area, while the fourth quadrant shows a pattern of relatively underdeveloped areas [46]. Shift-share divides the region based on three components, namely, regional share, proportional shift, and differential shift [47]. The pattern of inequality based on the Williamson Index divides regions into four groups. The first group is the region with very high inequality, with the Williamson Index (IW) value > 1, and the second is the region with high inequality, with the IW value between 0.7 and 1. Meanwhile, the moderate inequality region group has an IW value between 0.4 and 0.69, and the region with low inequality has a range of IW values below 0.39 [48].

IV. Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering (HAC) is one of the unsupervised approaches in machine learning. This technique groups data by combining two single clusters based on their similarity. A single cluster itself is a single data object from existing datasets. Unlike other cluster techniques, HAC does not require determining the number of specific clusters in the clustering process. However, the resulting cluster output can be divided into several large cluster groups to facilitate interpretation. The stages of grouping data using HAC are as follows:

1. Prepare the dataset
2. Calculate the distance matrix, for example, using Euclidean Distance as in equation (1).

$$d_{euc}(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^2\right)^{\frac{1}{2}} \quad (1)$$

3. Merge two single clusters that have something in common. Combining two single clusters can be done with several techniques, such as single, average, complete, and ward linkage. The following are the respective HAC technical equations:

- Single linkage – Equation (2):

$$f = \min\big(d(x,y)\big) \quad (2)$$

- Complete linkage – Equation (3):

$$f = \max\big(d(x,y)\big) \quad (3)$$

- Average linkage – Equation (4):

$$f = \text{average}\big(d(x,y)\big) \quad (4)$$

- Ward linkage – Equation (5):

$$f = ESS(XY) - [ESS(X) + ESS(Y)] \quad (5)$$

ESS is Error Sum Square obtained from equation (6).

$$ESS(X) = \sum_{i=1}^{Nx} \left| x_i - \frac{1}{Nx} \sum_{j=1}^{Nx} x_j \right|^2 \quad (6)$$

4. Update the distance matrix
5. Repeat steps 2 and 3 until the remaining number of clusters is 1.
6. Visualize the cluster in the form of a dendrogram.

V. Research Methods and Data

A. Research Method

This study uses the stages of the concept of data mining to solve problems. The clustering method used is hierarchical agglomerative clustering (HAC). The research begins with identifying problems, especially those related to the grouping of development inequalities. After the problem is identified, the next step is to understand and prepare the research data needs, including determining the data's variables. In the third stage, the research carries out the mining process. In the mining process, the cluster model is built with HAC. Four HAC techniques are used, namely single, complete, average, and ward linkage. After the cluster model is built, it is evaluated by looking at the silhouette coefficients to check whether the data points have been properly clustered. At this stage, the results of the fourth cluster of HAC techniques are compared based on the silhouette coefficient. The HAC cluster technique with the most appropriate grouping accuracy was used for the interpretation stage. The last stage is the interpretation of the results of the grouping that was carried out.

B. Data

The research data used is the gross regional domestic income (GRDP) of 41 regencies and municipalities in the western part of Java Island (i.e., Banten, Jakarta Capital Special Region, and West Java Provinces). The data used is GRDP data for the period 2010–2021. The data was obtained from the Central Statistics Agency of Indonesia. Data collection is done by downloading directly from the website of the central government statistics agency, both at the provincial and regional/municipal levels. In addition, provincial or regency data books in figures are also used as references in this study. This is done to clarify the data obtained from the official website of the Central Statistics Agency if there is a discrepancy.

C. Analysis Tools

Visual programming applications are used in this research, one of which is Orange Data Mining (ODM). To be able to group data using ODM, the first step is to build a workflow by adding several analysis components needed in the work area. ODM can handle various types of datasets, such as *.csv, *.xlsx, and *.txt. The data was converted into a *.csv file before use. In ODM, components such as file, data table, distance, and hierarchy cluster are used to group the DPRB data for 41 regions on West Java Island.

VI. Results

The presentation of research results follows the stages of data mining. In general, the data mining stage consists of four stages. Three stages are presented in the research results sub-chapter. While one stage in the form of interpretation is presented in the discussion chapter.

A. Data Mining Problem Statement

Identification of problems in the concept of data mining is the first step that must be done. At this stage, it produces information related to the many development inequalities in various parts of Indonesia, especially in the western part of Java Island. Inequality does not only occur between provinces but also between provinces. Even between regions of different provinces. The occurrence of development inequality often makes it difficult for policymakers to determine which areas should be prioritized and which proportion of priorities should be reduced. Inequality in development between regions is also often not well identified. So that if it is not handled seriously, it can affect the development policy-making process in the future. A comprehensive approach is needed to be able to categorize development inequalities as a whole and to see indications of the proximity of inequalities between one region and another. One approach that can accommodate this problem is to perform data mining using hierarchical agglomerative clustering techniques.

B. Data Understanding and Preparation

The data understanding stage in this study was conducted to determine the needs and the data collection process based on the results of problem identification at the problem statement stage. In the case of grouping development inequality, many variables can be used. However, in various works of literature, it is stated that GRDP data is the most used. Therefore, the data collected at this stage is GRDP data for constant prices from three provinces in West Java Island: Banten, DKI Jakarta, and West Java Provinces. Table 1 shows some GRDP data for several regions used in this study.

Table 1.
Some of the GRDP Data for Several Regions Used in Research

County Name	2010	2011	…	2015
Lebak Regency	12,572,538	13,325,629	…	16,733,238
Pandeglang Regency	12,279,542	12,984,403	…	15,974,129
Serang Regency	33,841,000	35,905,370	…	44,454,580
Tangerang Regency	18,549,119	62,022,491	…	78,093,560
Cilegon City	44,676,529	47,633,318	…	59,982,732
Serang City	12,549,572	13,595,691	…	17,808,478
Tangerang City	66,921,378	71,864,142	…	90,807,569
South Tangerang City	30,525,315	33,214,823	…	45,485,614
Kepulauan Seribu Regency	3,584,570	3,737,994	…	3,807,773
City of West Jakarta	182,020,885	205,951,118	…	328,883,065
City of Central Jakarta	258,419,705	276,997,678	…	355,092,532
City of South Jakarta	241,225,134	258,049,207	…	329,155,038

Data processing and data transformation are performed during the data preparation phase. At the stage of data processing, data integration and reduction are carried out. In order to facilitate the subsequent phase of work, 41 GRDP data files are merged into a single file via data integration. Due to the incompleteness of the collected data, the reduction stage is performed to reduce the research data by

year. There will be a decrease in Pandeglang Regency data. For example, GRDP data for Cianjur Regency includes information from 2008 to 2021. However, because most regions and municipalities lacked GRDP data from 2008, it was decided to use GRDP data from 2010 to 2021. Other areas necessitate a choice.

This study's data transformation takes the form of an outlier analysis-based smoothing procedure. The goal is to ascertain whether the datasets possess an abnormal data distribution. Scatter plot visualization and outlier analysis are used to determine that one data point has an abnormal distribution compared to the rest. The data point in question pertains to the city of Bandung. When data from 2010 and 2021 are compared, this anomaly becomes evident. Only Bandung City exhibits an abnormal distribution when comparing the data from these two years.

Even though it is common to get rid of outlier data, we leave it as is and check the main data source again. Repeated outlier data raises suspicions after nine comparisons of the data distribution graph. After examining the integrated and the original data source, it was determined that human error was present, specifically during data entry. The data were then modified using repeated outlier analysis. After repairs, the utilized datasets contained no anomalous data. Figure 1 illustrates the adjusted data distribution and the correlation between variables by year for each data point. The result of the data transformation is in the form of data ready to be used in the next stage, namely the data mining process. Data mining is carried out to extract hidden patterns to obtain knowledge for future development decision-making processes.

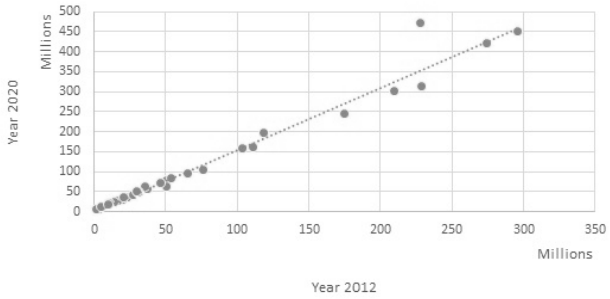

Fig. 1: Graph of Correlation of GRDP Data in 2012 and 2021

A. Data Mining Process and Model Evaluation

The goal of the data mining process is to uncover patterns in the GRDP data that belong to the 41 regions of West Java Island.This study used the hierarchical agglomerative clustering (HAC) technique to group GRDP data. Four techniques were used, and then the results were compared based on the evaluation of the cluster results according to the Silhouette coefficient. The four techniques are single, complete, average, and ward linkage. The data mining process is carried out on the GRDP data for the forty-first regions in West Java Island.

The clustering process is carried out using Orange Data Mining software. Euclidean distance is used as a distance calculation parameter for each HAC technique. Data normalization is also carried out to avoid the attribute value ranges being too far from one another. No pruning was carried out for each HAC technique to maintain the visualization of the dendrogram as a whole and to show the cluster shape of the entire area that was the research object. In addition, the dendrogram output is divided into three large cluster groups for each technique to make the interpretation process easier.

The cluster output using the single linkage technique grouped 41 areas in the western part of Java Island into three large cluster groups in the form of a dendrogram. The dendrogram shows that most areas are in the C2 cluster group, while the rest are members of the C1 and C3 cluster groups. Figure 2 shows the dendrogram for a single linkage. The C2 cluster group is mainly filled with areas from two provinces: West Java and Banten. Meanwhile, the C2 cluster group is dominated by areas in the Special Capital Region of Jakarta. The C3 cluster group contains members from the Province of West Java and the Special Capital Region of Jakarta. Overall, three regions are members of the C1 cluster group; as many as 35 regions are members of the C2 cluster group; and three regions are in C3.

The next analysis HAC is average linkage. The cluster results with this technique show that the Provinces of West Java and Banten regions dominate the C1 cluster group. Three regions in the Special Capital Region of Jakarta are members of the C2 cluster group. At the same time, the rest are members of the C3 cluster group, along with one area from West Java Province, Bekasi Regency. There are 35 regions clustered into C1, three regions into C2, and three regions into C3. The cluster results between the single technique and the average linkage are the same. The difference is the change in the cluster group label. Suppose, in a single linkage, three regions of the Special Capital Region of Jakarta, such as West Jakarta City, Central Jakarta City, and South Jakarta City, are members of the C1 cluster group. In that case, all three are still in the same cluster group, but with the C2 label. The same holds true for group C2 on a single link to group C1 on an average link. Meanwhile, the C3 cluster group remains in its original position. Figure 3 shows a cluster dendrogram using the average linkage technique.

The complete linkage technique shows different results. Although divided into three large cluster groups, there were significant changes in cluster members, especially in the C2 and C3 cluster groups. In general, the C1 cluster group is dominated by areas from the provinces of West Java and Banten. The C2 cluster group consists of three regions, with the dominant of the Special Capital Region of Jakarta Province, namely West Jakarta City, Central Jakarta City, and South Jakarta City. The total number of regions clustered into C1 is 32 regions.

The C3 cluster group is made up of six regions: Bandung City, Bogor Regency, Karawang Regency, Bekasi Regency, East Jakarta City, and North Jakarta City.For the other three regions, such as West Jakarta City, Central Jakarta City, and South Jakarta City, there was no change in cluster position, either with single, average, or complete linkage. Figure 4 shows a dendrogram with a complete linkage technique.

Changes in the positions of cluster members also have a big effect on the results of clusters that use the ward linkage method. Almost all of the areas in the Province of the Special Capital City Region of Jakarta were put into one cluster group, called C1, based on the cluster with ward linkage results. But C1 also has the Bekasi Regency, which

is part of the province of West Java. The West Java Province areas of Bandung City, Bogor Regency, and Karawang Regency make up most of the C3 cluster group. The C3 cluster group comprises 32 regions dominated by West Java and Banten Provinces. Figure 5 shows a dendrogram of cluster results with ward linkage.

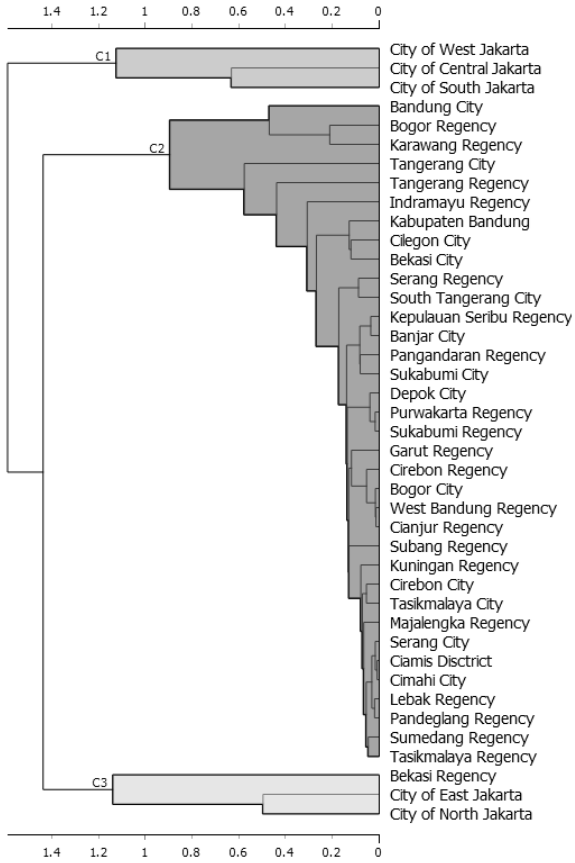

Fig. 2: Dendrogram of Single Linkage Cluster Results

After the data mining process is done, the cluster model is evaluated to see how well it works. We evaluate cluster by calculating the Silhouette coefficient for each clustering technique. Silhouette coefficient can show the relevant information according to the suitability of clustered data objects in the right cluster. Evaluation with the Silhouette coefficient is also intended to choose which HAC cluster technique is more effective in the discussion section.

Using silhouette coefficients, the cluster evaluation of the four HAC techniques shows that four cluster members are not in the right cluster for single and average linkage. As for complete linkage, eight cluster members are indicated not to be in the right cluster. On the other hand, the ward linkage shows that seven cluster members are in the wrong cluster. In Table 2, you can see a summary of what was found when the HAC cluster model was evaluated. The discrepancy between cluster members who are in a specific cluster group is seen from the silhouette coefficient. In this study, a silhouette coefficient of at least 0.60 is needed for a value to be seen as a member of a cluster. Below the range of 0.60, we think that the members of the cluster are not grouped together correctly. The range of this coefficient is set more by how tight it needs to be to figure out if an object belongs to a certain cluster or not. In the calculation of the silhouette coefficient, it can also be seen that several cluster members have a silhouette coefficient below 0 and a negative value.

Based on how well the cluster model worked, this study chose either single linkage or average linkage, which had the fewest number of clusters that didn't match up.

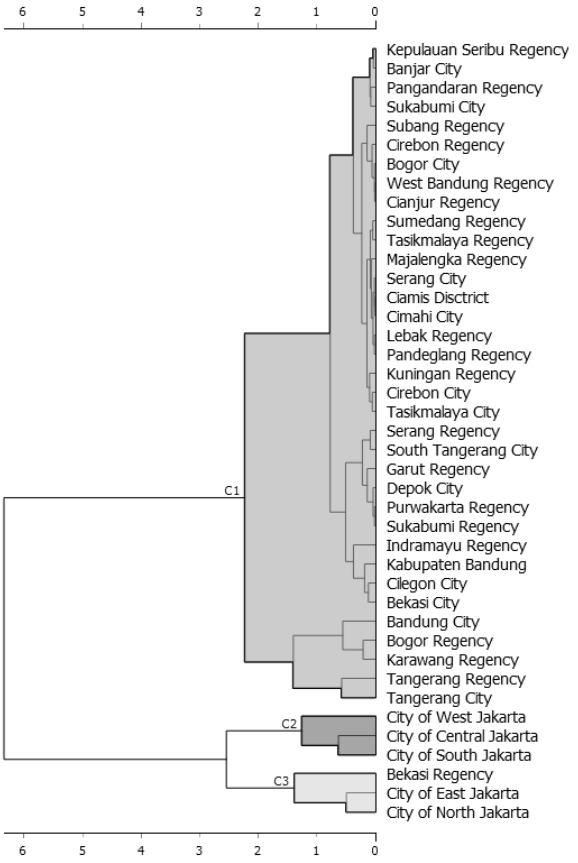

Fig. 3: Dendrogram of Average Linkage Cluster Results

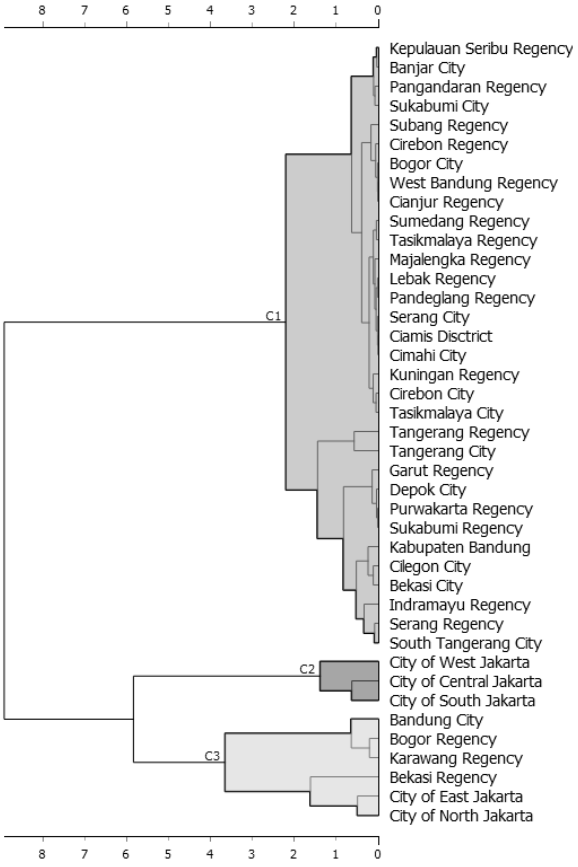

Fig. 4: Dendrogram of Complete Linkage Cluster Results

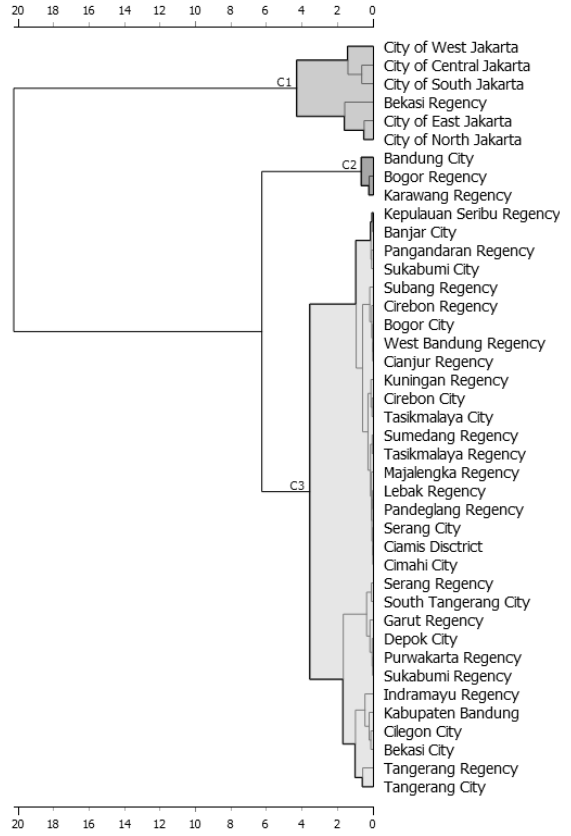

Fig. 5: Dendrogram of Ward Linkage Cluster Results

TABLE 2.
SUMMARY OF EVALUATION OF THE HAC CLUSTER MODEL

HAC method	Group Cluster	Number of Cluster Members	Number of Incompatible Cluster	Incompatible Cluster Member
Single Linkage	C1	3	0	-
	C2	35	3	Bandung City, Karawang City and Bogor Regency
	C3	3	1	City of North Jakarta
	Total	41	4	
Average Linkage	C1	35	3	Bogor Regency, Karawang Regency and Bandung City
	C2	3	0	-
	C3	3	1	City of North Jakarta
	Total	41	4	
Complete Linkage	C1	32	2	Tangerang Regency and Tangerang City
	C2	3	0	-
	C3	6	6	City of East Jakarta, City of North Jakarta, Bekasi Regency, Bogor Regency, Karawang Regency, and Bandung City
	Total	41	8	
Ward Linkage	C1	6	3	City of East Jakarta, City of North Jakarta, and Bekasi Regency
	C2	3	0	
	C3	32	4	Tangerang Regency, Cilegon City, Tangerang City, and Bandung Regency
	Total	41	7	

VII. DISCUSSION

Another essential data mining stage is representing the knowledge or insight obtained. As stated in the previous sub-chapter, this discussion focuses on cluster results using the average linkage technique. The GRDP data from 2010 to 2021 are known to be grouped into three large cluster groups based on information from data mining results using the cluster technique, particularly the average linkage technique. The first cluster group is C1 containing cluster members in the form of Lebak Regency, Pandeglang Regency, Serang Regency, Tangerang Regency, Cilegon City, Serang City, Tangerang City, South Tangerang City, Thousand Islands Regency, Bandung Regency, West Bandung Regency, Bogor Regency, Regency Ciamis, Cianjur Regency, Cirebon Regency, Garut Regency, Indramayu Regency, Karawang Regency, Kuningan Regency, Majalengka Regency, Pangandaran Regency, Purwakarta Regency, Subang Regency, Sukabumi Regency, Sumedang Regency, Tasikmalaya Regency, Bandung City, Banjar City, Bekasi City, Bogor City, Cimahi City, Cirebon City, Depok City, Sukabumi City, and Tasikmalaya City.

The cities of West Jakarta City, Central Jakarta City, and South Jakarta City are part of the second cluster group, called C2. The last group, C3, is made up of cluster members from East Jakarta City, North Jakarta City, and Bekasi Regency. Table 3 shows information about cluster groups and cluster members as well as the average value of the GRDP for 2010–2021. Table 3 also shows that all members grouped into the C1 cluster group are regions with an average GRDP between 2,708,393 and 164,994,518.76. To find out the relationship between cluster results and the average GRDP of each region, Figure 6 is then trimmed by taking a particular visualization of the C1 cluster group. The trimmed results make it easier to look at and figure out what the cluster results mean (see Figure 6).

Pangandaran Regency and Sukabumi City also have a relatively close average GRDP. Therefore, these two areas were then grouped into a new small cluster group (GKK-2). With the mechanism for calculating the average distance after being compared to other regions, the distance from the GKK-1 cluster group turned out to have a more significant closeness than other cluster groups. Thus, GKK-1 and DKK-2 were then grouped into a new group (GKK-3). Pay attention to the illustration in Figure 6, which is marked with a blue box with a dotted line.

The grouping mechanism for other regions is the same as the results of GKK-1 and GKK-2, which are grouped into a new cluster group (GKK-n) until finally forming a large hierarchical cluster group. In this study, the cluster technique can also be seen to group regions regardless of their province of origin. Like the case of the Thousand Islands Regency and the City of Banjar, each of which is an area of the Jakarta Capital Special Region and West Java, according to the GRDP value, they have identical

development achievements close to each other. Therefore, they were grouped into new groups. The Thousand Islands Regency is not grouped with other areas in the Jakarta Capital Special Region because it has a GRDP value that is significantly higher compared to other areas of the Jakarta Capital Special Region.

TABLE 3.
AVERAGE LINKAGE CLUSTER RESULTS

No.	Group Cluster	County Name	Average (GRDP)	GRDP Average Interval
1		Lebak Regency	7,169,447.71	
2		Pandeglang Regency	16,442,524.37	
3		Serang Regency	45,404,847.42	
4		Tangerang Regency	76,528,060.70	
5		Cilegon City	61,695,814.94	
6		Serang City	18,354,166.54	
7		Tangerang City	91,505,051.81	
8		South Tangerang City	47,139,810.85	
9		Kepulauan Seribu Regency	3,781,881.42	
10		Bandung Regency	66,885,467.07	
11		West Bandung Regency	26,080,560.16	
12		Bogor Regency	128,181,829.97	
13		Ciamis Regency	18,359,503.14	
14		Cianjur Regency	26,418,165.00	
15		Cirebon Regency	28,330,260.83	
16		Garut Regency	32,981,001.58	
17	C1	Indramayu Regency	52,662,379.11	2,708,393.00 - 164,994,518.67
18		Karawang Regency	136,011,496.88	
19		Kuningan Regency	12,443,740.83	
20		Majalengka Regency	18,496,330.27	
21		Pangandaran Regency	5,649,737.31	
22		Purwakarta Regency	38,608,264.86	
23		Subang Regency	23,382,450.23	
24		Sukabumi Regency	38,550,736.41	
25		Sumedang Regency	20,383,481.59	
26		Tasikmalaya Regency	20,532,809.46	
27		Bandung City	154,994,518.67	
28		Banjar City	2,708,393.00	
29		Bekasi City	56,963,930.56	
30		Bogor City	26,276,842.35	
31		Cimahi City	18,538,389.09	
32		Cirebon City	14,374,489.27	
33		Depok City	38,970,589.60	
34		Sukabumi City	7,191,757.69	
35		Tasikmalaya City	12,801,134.21	
36	C2	City of West Jakarta	342,603,546.59	366,594,421,69 - 342,603,546.59
37		City of Central Jakarta	366,594,421.69	
38		City of South Jakarta	340,947,943.77	
39	C3	City of East Jakarta	256,360,519.65	209,712,727.44 - 256,360,519.65
40		City of North Jakarta	275,148,956.75	
41		Bekasi Regency	209,712,727.44	

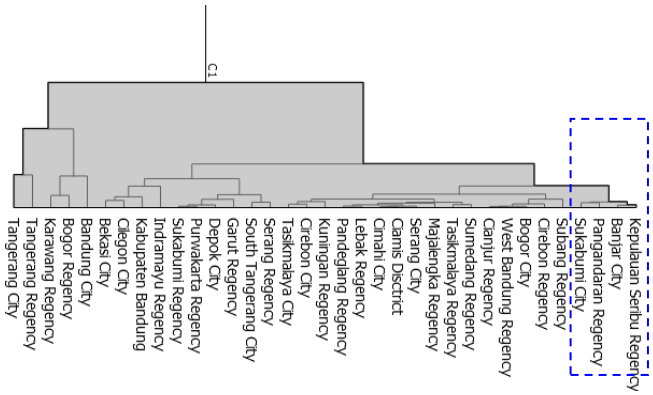

Fig. 6: Average Linkage Cluster Results for Group C1.

The same thing is also seen in Cilegon City and Bekasi City. Two regions from two different provinces are grouped into one new cluster group. Serang City, Ciamis Regency, and Cimahi City are all grouped into the same cluster group. The results of the grouping can be used by policymakers or regional leaders to assess the regional development accomplishments that they lead.In addition, it can also be used as input regarding the focus of future development policies. The cluster results in this study also show that group C1 is a collection of areas with the lowest average GRDP value among the three other cluster groups. Meanwhile, group C2 consists of areas with the highest average GRDP value, and group C3 consists of areas with an average GRDP value in the middle (between C1 and C2).

Using the HAC technique, the results of the cluster can be used to sort development gaps between provinces based on the value of the GRDP. It can, however, also group development differences by province so that policymakers can see where inequality is not only between provinces but also within provinces. Figure 7 shows the results of regional grouping for one province, Banten Province.

Figure 7 shows that Banten Province has three big groups of clusters: C1, C2, and C3. Each cluster group represents the results of regional grouping based on the average value of the GRDP, as shown in Table 3. The results of the regional grouping in Banten Province show that Lebak Regency and Pandeglang Regency are clustered into a new cluster group (GKB-1). Notice the blue dotted box in Figure 7. The results from GKB-1 have clustered again with Serang City (GKB-2). If we look at the average GRDP of the three regions, it can be seen that all three have development achievements that are close to each other in value. Even though the data show otherwise, Lebak Regency has the smallest average GRDP value compared to Pandeglang Regency and Serang City. However, these two areas are closer to Lebak Regency, so they are clustered into the same cluster group. The same is true for other regions. The cluster results represent the GRDP values that are close to each other. In the context of this cluster visualization, the policymakers can see the position of each region under their leadership. Specifically, the grouping of development

inequalities that occur between regions.

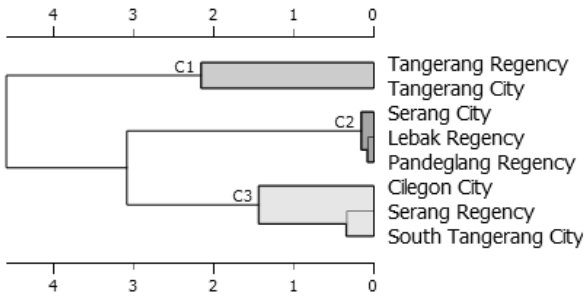

Fig. 7: Aggregated Linkage Cluster Results for Banten Province

In addition to Banten Province, this study also visualizes cluster results for two other provinces, namely West Java and the Jakarta Capital Special Region. The aim is to see the position of each region when grouped within a province. Like Banten Province, West Java Province's cluster outputs are grouped into three large cluster groups, namely C1, C2, and C3. Each cluster group represents an area based on the proximity of its average GRDP. Like Bogor Regency and Karawang Regency, both are clustered into a tiny cluster group because they have adjacent average GRDP values. Bogor Regency has an average GRDP of 159,582,650, while Karawang Regency has 166,941,492.2. These two areas are then grouped with the area that has a higher average than the two, namely the City of Bandung. Finally, after regrouping with Bekasi Regency, which has the highest average GRDP of the three, cluster group C1 was formed. The cluster members of each group are shown in Figure 8.

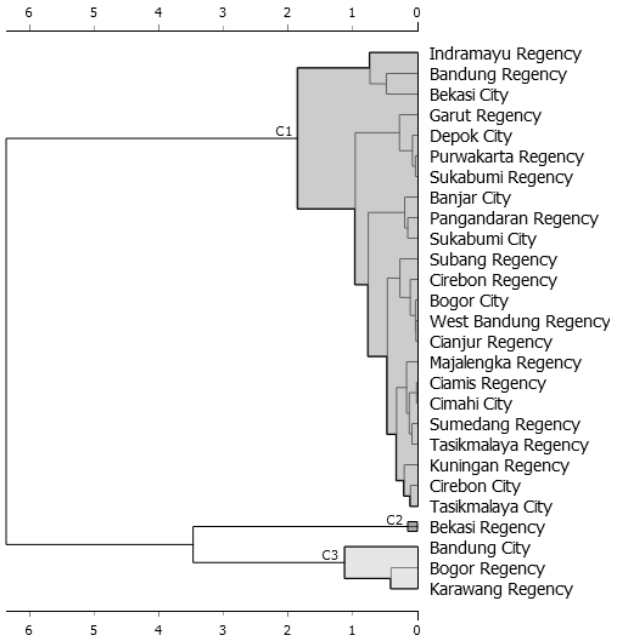

Fig. 8: Average Linkage Cluster Results for West Java Province

Among the other regions, cluster group C1 has the highest average GRDP. Group C2 is a collection of regions with the lowest average GRDP among the other groups. Meanwhile, the C3 cluster group is a collection of regions with an average GRDP value in the middle. With the results of this cluster, it can be seen that as many as 59.25% of the area in West Java Province is still in the low development achievement group, which results in inequality with other cluster groups. Only 14.8% have very high GRDP scores, namely Bandung City, Bekasi Regency, Karawang Regency, and Bogor Regency. From the facts on the ground, it is very natural that these four regions have an average GRDP value that is very different from other regions. These four regions have very supportive economic activities, supported by the presence of industry in their areas.

For areas in the Jakarta Capital Special Region Province, this study also groups them into three large cluster groups. The cluster group C1 consists of the Kepulauan Seribu Regency and C2, with members from East Jakarta City and North Jakarta City. As for the C3 cluster group, West Jakarta City, Central Jakarta City, and South Jakarta City are filled. Figure 9 shows that there is a large gap in the Jakarta Capital Special Region. The Kepulauan Seribu Regency is in a separate cluster group (C1) by itself since no other area is close to the GRDP value, which is a measure of how well development is going. Geographically, the Thousand Islands Regency is separated from the other five regions. Even if the GRDP data is checked, it is unequal compared to other cities in DKI Jakarta. The average GRDP of the Kepulauan Seribu Regency in 2010–2021 was only 3,781,881,423.

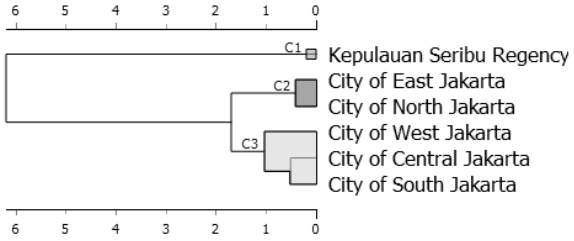

Fig. 9: Average Linkage Cluster Results for DKI Jakarta Province.

The results of this study's grouping of inequality in the Jakarta Capital Special Region Province, especially in the Kepulauan Seribu Regency, are in line with research [45] done in 2007 that showed that the Kepulauan Seribu Regency has problems with development. The development process that occurs is still not sustainable. This is seen from an ecological-economic point of view, which shows an imbalance in terms of trading activities involving fish catches. The people of Kepulauan Seribu still depend on fishing for their economy. In seven years, research conducted by [49] in 2014 said that development inequality still occurs in the Kepulauan Seribu Regency. His research report stated that development inequality in the Jakarta Capital Special Region, especially in the archipelagic region, continued to increase for the past five years starting in 2014. Several factors cause significant inequality, including geographical conditions, government policies, weak regional spatial planning, human resources, and the economy.

VIII. Conclusion

The research has grouped forty-one West Java Island areas into three large cluster groups. The forty-one areas are part of three provinces: West Java, Jakarta's Special Capital Region, and Banten. The results of the grouping indicate that there is an imbalance in development between areas that are close to each other. A comparison of cluster output to actual GRDP data also shows that the GRDP value of a region does not significantly affect the cluster group formed. The size of the distance between data points has a

significant impact on the formation of the cluster group. This study also shows that clustering tasks in data mining can overcome the limitations of tipologi Klassen, Index Williamson (IW), and Shift Share (SS) analysis. Clustering techniques can solve data time series, allowing them to fully analyze GRDP, whereas Klassen, IW, and SS cannot.

The cluster results can reveal the links of inequality between regions, both intra- and inter-provincial. Klassen, IW, and SS simply group the inequality without drawing attention to the connections. The cluster results can also show the position of inequality between one region and another, so that they can be used as material for evaluation and regional repositioning by policymakers in preparing future development plans.

REFERENCES

[1] M. Mustaffa, S. Ahmad, M.I.H.M. Nasrudin, K.A. Sekak, N.A. Aini, A.M.M. Ali, M.F.M. Taib, M.Z.A. Yahya, O.H. Hassan, A.K. Razali, and M.H.M. Jais, "Data Mining Analysis on Ships Collision Risk and Marine Traffic Characteristic of Port Klang Malaysia Waterways from Automatic Identification System (AIS) Data," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019, 13-15 March, 2019, Hong Kong, pp242-246

[2] S. Palmer, "Characterizing the Interactions of a Multinational Engineering Services Company on Twitter," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2018, 4-6 July, 2018, London, U.K., pp205-210

[3] H. Mori, R. Yamanishi, Y. Nishihara, and Junichi Fukumoto, "Relationship Between Features of Reading Behaviors and Dynamic Abstract of Novel," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2018, 14-16 March, 2018, Hong Kong, pp254-259

[4] M. Suyunova, "The Use of Demand Forecasting Techniques for the Improvement of Spare Part Management," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2018, 4-6 July, 2018, London, U.K., pp223-227

[5] SA Bogle, and KM Black, "Classifiers for Predicting Undergraduate Computer Science Performance," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2018, 4-6 July, 2018, London, U.K., pp228-231

[6] R. Rathi, N Yagnik, S. Tiwari, and C. Sharma, "Analysis of Statistical Models for Fast Time Series ECG Classifications," Engineering Letters, vol. 30, no.2, pp718-729, 2022

[7] Y. Hu, X. Zhang, J. Yang, and S. Fu, "A Hybrid Convolutional Neural Network Model Based on Different Evolution for Medical Image Classification," Engineering Letters, vol. 30, no.1, pp168-177, 2022

[8] V. B. Kumar, K. Vijayalakshmi, and M. Padmavathamma, "A Hybrid Data Mining Approach for Diabetes Prediction and Classification," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2019, 22-24 October, 2019, San Francisco, USA, pp298-303

[9] J. Messejana, R. Pereira, J.C. Ferreira, and M. Baptista, "Predictive Analysis of Incidents based on Software Deployments," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019, 3-5 July, 2019, London, U.K., pp150-155

[10] X.D. Li, D. Wei, J.S. Wang, Q.S. Guo, and L. Chen, "Colony Image Edge Detection Algorithm Based on FCM and RBI-FCM Clustering Methods," IAENG International Journal of Computer Science, vol. 48, no.2, pp356-363, 2021

[11] A. W.. Rosyadi, and N. Suciati, "Image Segmentation Using Transition Region and K-Means Clustering," IAENG International Journal of Computer Science, vol. 47, no.1, pp47-55, 2020

[12] K. Salhi, E.L. Jaara, M.T. Alaoui, and Y.T. Alaoui, "Color-Texture Image Clustering Based on Neuro-morphological Approach," IAENG International Journal of Computer Science, vol. 46, no.1, pp134-140, 2019

[13] W. Ieosanurak, and W. Klongdee, "Face Classification using Adjusted Histogram in Grayscale," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2018, 14-16 March, 2018, Hong Kong, pp276-278

[14] Arifin, S. Sumpeno, Muljono, and M. Hariadi, "A Model of Indonesian Dynamic Visemes From Facial Motion Capture Database Using A Clustering-Based Approach," IAENG International Journal of Computer Science, vol. 44, no.1, pp41-51, 2017

[15] J. Wu, N. Endo, and M. Saito, "Cluster Analysis for Investigating Road Recovery in Fukushima Prefecture Following the 2011 Tohoku Earthquake," Engineering Letters, vol. 29, no.4, pp1636-1642, 2021

[16] J. Wu, N. Endo, and M. Saito, "Cluster Analysis for Investigating Road Recovery in Iwate Prefecture Following the 2011 Tohoku Earthquake," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2021, 20-22 October, 2021, Hong Kong, pp65-69

[17] G. Chen, J. Chen, Z. Zhang, and Z. Sun, "Short-Term Wind Speed Forecasting Based on Fuzzy C-Means Clustering and Improved MEA-BP," IAENG International Journal of Computer Science, vol. 46, no.4, pp768-776, 2019

[18] G. Chen, B. Tang, Z. Zhang, Z. Zeng, and S. Li, "Short-Term Wind Speed Forecasting Based on Singular Spectrum Analysis, Fuzzy C-Means Clustering and Improved SSABP," Engineering Letters, vol. 29, no.2, pp351-364, 2021

[19] T. Ma, H. Liu, and Y. Zhang, "A Method for Establishing Tropospheric Atmospheric Refractivity Profile Model Based on Multiquadric RBF and k-means Clustering," Engineering Letters, vol. 28, no.3, pp733-741, 2020

[20] R. Alfred, L.Y. Jie, J.H. Obit, Y. Lim, H. Haviluddin, and A. Azman, "Social Media Mining: A Genetic Based Multiobjective Clustering Approach to Topic Modelling," IAENG International Journal of Computer Science, vol. 48, no.1, pp32-42, 2021

[21] M. Rokaya, H. Turabieh, S. Al Azwari, A. Alharbi, M. Alnfiai, M. Alzahrani, W. Osaimi, and W. Alhakam, "Clustering Hashtags Based on New Hybrid Method and Power Links," IAENG International Journal of Computer Science, vol. 48, no.3, pp716-730, 2021

[22] R. R. Seva, C. Carandang, K. Lim, J.M. Khoo, A.M.J.A. Gutierrez, and J.C. Tangsoc, "Grocery Product Arrangement Using Cluster Analysis," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019, 13-15 March, 2019, Hong Kong, pp459-463

[23] J.T. Wei, S.Y. Lin, Y.Z. Yang, and H.H Wu, "Using A Combination of RFM Model and Cluster Analysis to Analyze Customers' Values of A Veterinary Hospital," IAENG International Journal of Computer Science, vol. 47, no.3, pp442-448, 2020

[24] A. Ragozin, V. Telezhkin, and P. Podkorytov, "Hierarchical Cluster-analysis of Transient Heart Rate using a Digital Spectral Analysis in the Complex Frequency Plane," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2019, 22-24 October, 2019, San Francisco, USA, pp1-3

[25] A. Guo, A. Jiang, and Z. Cheng, "A Hybrid Clustering Method for Bridge Structure Health Monitoring," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2018, 23-25 October, 2018, San Francisco, USA, pp161-165

[26] M.K. Islam, M.M. Ahmed, and K.Z. Zamli, "i-CODAS: An Improved Online Data Stream Clustering in Arbitrary Shaped Clusters," Engineering Letters, vol. 27, no.4, pp752-762, 2019

[27] L. Li, Y. Lin, X. Wang, T. Guo, J. Zhang, H. Lin, and F. Nan, "A Clustering-Classification Two-Phase Model on Service Module Partition Oriented to Customer Satisfaction," Engineering Letters, vol. 26, no.1, pp76-83, 2018

[28] K. Abdalgader, "Clustering Short Text using a Centroid-Based Lexical Clustering Algorithm," IAENG International Journal of Computer Science, vol. 44, no.4, pp523-536, 2017

[29] R. Xi, and K. Zhenxing, "Hierarchical RNN for Information Extraction from Lawsuit Documents," Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2018, 14-16 March, 2018, Hong Kong, pp266-270

[30] 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-SosioEkonomi Unsrat*, Volume 16 No 3, pp.. 369 – 378, 2020, *in Bahasa*

[31] M.J. Darmawan, and Tukiman, Analisis Dimensi Ketimpangan Pembangunan Antar Wilayah Di Provinsi Jawa Timur Tahun 2014-2018, Jurnal Dinamika Governance: Jurnal Ilmu Administrasi Negara, Volume 10 No 1, 2020, *in Bahasa*

[32] R.H. Harahap, H.B Isyandi, and 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, pp. 183 – 193, 2020, *in Bahasa*

[33] Kadriwansyah, B. Semmaila, and J. Zakaria, Analisis Ketimpangan Wilayah di Provinsi Sulawesi Selatan Tahun 2014-2018, *Paradoks: Jurnal Ilmu Ekonomi* Volume 4.No.1, pp. 25 – 36, 2021, *in Bahasa*

[34] 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

[35] 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, 2019, https://doi.org/10.1371/journal.pone.0210697

[36] 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), pp. 21–36, 2022, doi: 10.13052/jrss0974-8024.1512

[37] 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

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

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

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

[41] 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

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

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

[44] 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). *Thesis* Magister Statistika, FMIPA-ITS, 2012, *in Bahasa*

[45] S.B. Susilo, Analisis Berkelanjutan Pembangunan Pulau-Pulau Kecil : Pendekatan Model Ekologi-Ekonomi, *Jurnal Ilmu-Ilmu Perairan dan Perikanan Indonesia*, Vol. 14 No. 2, pp. 29 – 35, 2019, *in Bahasa*

[46] Hasanah, Pemetaan Sektor Unggulan di Kota Pontianak Dengan Metode Tipologi Klassen dan Location Quotient, Prosiding Seminar Nasional SATIESP 2021, p. 156 – 163, 2021

[47] S. Taniu, A.P. Yakup, and M.A. Novriansyah, Shift Share Analysis to Determine Regional Economic Performance Of Gorontalo, Gorontalo Development Review (GOLDER), Vol.3, No.2, pp. 102 – 113, 2020, Available at https://jurnal.unigo.ac.id/index.php/gdrev/article/view/1088/619

[48] V. D. Waluyaningsih, and A. H. Setiawan, "Analisis Ketimpangan Pendapatan Antarwilayah di Kawasan Kedungsepur, Barlingmascakeb, dan Subosukawonosraten Periode 2008-2017," Diponegoro Journal of Economics, vol. 9, no. 2, pp. 123-134, Jul. 2021

[49] Adinianata, A.T. Alamsyah, Mengurangi Ketimpangan Pembangunan Di Wilayah Kepulauan dan Daratan Menuju Pengembangan Pemukiman Kepulauan Mikro Yang Berkelanjutan : Studi Kasus Kabupaten Kepualaun Seribu, *Thesis* Program Studi Kajian Pengembangan Perkotaan, Universitas Indonesia, 2014, *in Bahasa*

Tb Ai Munandar is currently an assistant professor in the department of informatics, Faculty of Computer Science at Universitas Bhayangkara Jakarta Raya, Indonesia. He received his Doctorate in Computer Science at Universitas Gadjah Mada, Indonesia, in 2017. His research fields include artificial intelligence, machine learning, data mining, and implemented data science for health and regional development. He is a member of IAENG with the number 120487. In addition, he is also an active member of the Indonesian Computer Electronics and Instrumentation Support Society (IndoCEISS) for Banten Province.

Dwipa Handayani is currently an assistant professor at the department of informatics, Faculty of Computer Science, the Universitas Bhayangkara Jakarta Raya, Indonesia. She received her M.S in information management systems at Gunadarma University, Jakarta, Indonesia. Apart from being a lecturer, she is the vice dean of the faculty of computer science in the field of human resources. Her research areas are information systems, decision support, data mining, and machine learning.