EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI: 10.20885/EKSAKTA.vol1.iss1.art9

Research Article

Clustering Provinces in Indonesia based on Community Welfare Indicators

Sekti Kartika Dini[1,*], **Achmad Fauzan**[1]

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia
[*] Corresponding author: sektidini@uii.ac.id

Received: 5 January 2020; Accepted: 13 February 2020; Published: 15 February 2020

Abstract: The Preamble of the 1945 Constitution of the Republic of Indonesia explicitly states that the main task of the government of the Republic of Indonesia is to advance general prosperity, to develop the nation's intellectual life, and to realize social justice for all Indonesian people. Social inequality is a problem that is still faced by Indonesian people today. To solve the problem required supporting data analysis as a basis for policy formulation. This research was conducted with the aim of clustering provinces in Indonesia based on community welfare indicators using K-Means cluster analysis. K-Means cluster analysis is chosen based on the variance value (0.101), which is smaller than the variance value in the average linkage cluster analysis (0.152). Based on data analysis, provinces in Indonesia are clustered into three where the first cluster consists of 21 provinces, the second cluster consists of 3 provinces, and the third cluster consists of 10 provinces. Each cluster has different characteristics that can be of concern to the parties concerned to overcome the social welfare gap. Besides, in order cluster results are more easily understood, visualization of results is added with a Geographic Information System (GIS) using Indonesian maps accompanied by differences in color gradations for each cluster.

Keywords: average linkage clustering, clustering, GIS, K-Means, public welfare

Introduction

The Preamble of the 1945 Constitution of the Republic of Indonesia explicitly states that the main task of the government of the Republic of Indonesia is to advance general prosperity, to develop the nation's intellectual life, and to realize social justice for all Indonesian people [1].

Meanwhile, according article 1 and 2 of Law No. 11 of 2009 concerning Social Welfare explained that social welfare is a condition of fulfilling the life needs of citizens to be able to develop themselves and be able to carry out their social functions that can be carried out by the government, regional governments, and the community in the form of social services which include: social rehabilitation, social security, social empowerment, and social protection [2]

Social inequality is a problem that is still faced by the Indonesian people, for example, unequal education and health facilities, fewer employment opportunities when compared to the workforce, unequally on population density and population growth rate, and so forth. The existence of social inequality will have an impact on the emergence of other problems. Therefore, a specific strategy is needed to overcome the problem of social inequality.

The government, as a policymaker, certainly requires supporting data as one of the bases for making decisions on target. Likewise, the problem of social inequality requires proper data analysis so that it produces meaningful information that can be used by the government as a basis for policymaking. The availability of relevant information is expected to help the government in determining regional or provincial priorities that need attention in social welfare. This is necessary so that in the future social welfare can be fairly felt by all Indonesian people.

Several studies related to social welfare in the regions based on social welfare indicators have been carried out. [3] research to clustering regencies and cities in Jawa Tengah based on indicators of people's welfare.

The welfare indicators used in the study include GDRP per capita, population density, real expenditure per capita, the number of poor people, the number of the labor force, life expectancy, and the average length of schooling. The study resulted in grouping in 35 regencies and cities in Central Java into three groups. The first group consists of 28 regencies and cities with relatively low welfare characteristics. The second group consists of 2 regencies and cities that have characteristics of a level of welfare that is relatively moderate or better than the first group.

Meanwhile, the third group is a group with characteristics of a better level of welfare than the other two groups. [4] researching to cluster districts and cities in Jawa Barat based on community indicators using the K-Means method. This study uses community indicators consisting of population density, labor force, population growth rate, the average per capita expenditure, life expectancy, and the average length of schooling. Based on the analysis, 27 districts and cities in West Java are grouped into two groups with the characteristics of the first group consisting of eight groups with population density, labor force, population growth rate, average per capita expenditure, life expectancy, and length of school have more values higher than the second group consisting of 19 city and districts.

a review of the application of K-Means cluster analysis and hierarchical cluster analysis to air pollution analysis for 1980-2019 [5]. An interesting presentation of analysis results has an important role so that information can be easily understood so that related parties can apply it. One way to present the results of data analysis involving regions is to use a Geographic Information System (GIS). Geographic Information Systems (GIS) are computer-based tools that can be used to collect, store, manipulate, and display spatial information [6]. GIS can be used effectively to support decision making in various fields such as social, economic, health, education, and so on [7].

This research is conducted with the aim of clustering provinces in Indonesia based on community welfare indicators. Furthermore, cluster results will be visualized through the Geographic Information System to be more informative and easily understood. Related parties can use the results of this study as supporting information to overcome the problem of social inequality in Indonesia based on priority provinces.

Materials and Methods

Data

The data is used in this study are sourced from the publication of Statistics Indonesia (BPS RI), namely the public welfare indicator data for 2017 [8]. The welfare indicators used include population density, labor force, population growth rate, the average per capita expenditure, figures life expectancy, and the average length of schooling from 34 provinces in Indonesia

Data Analysis

In general, the data analysis stage used in this study includes cluster analysis and results visualization. The detailed stages of data analysis are as follows.

Cluster Analysis

Cluster analysis is part of multivariate statistical analysis that is used to group objects based on the similarity of characteristics [9]. The characteristic of cluster analysis is that objects in the same cluster will have high similarity characteristics, while objects in different clusters will have low similarity characteristics.

If there are as many as n objects and p variables, then observations with $i = 1,2,3 \cdots, n$ and $j = 1,2,3, \ldots, p$ can be illustrated as follows [4].

Table 1. Illustration of the arrangement of observations in cluster analysis

	Variable 1	Variable 2	Variable 3	...	Variable p
Object 1	X_{11}	X_{12}	X_{13}	…	X_{1p}
Object 2	X_{21}	X_{22}	X_{23}	…	X_{2p}
⋮	⋮	⋮	⋮	⋱	⋮
Object n	X_{n1}	X_{n2}	X_{n2}	…	X_{np}

In cluster analysis, distance measurement is used as a measure of similarity approach. The most commonly used measure of distance is the Euclidean distance with the following formula [9].

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{1}$$

where, d_{ij} : Euclidean distance of the i^{th} data object and the j^{th} data object, p : number of variables, x_{ik}: i^{th} data object in the k^th variable, x_{jk} : j^{th} data object in the k^{th} variable.

In general, there are two cluster methods, namely hierarchical cluster analysis and non-hierarchical cluster analysis. In hierarchical cluster analysis, clusters are formed by dividing objects iteratively using an agglomerative (button-up) or divisive (top-down) approach [10]. The hierarchical cluster analysis consists of (1) single linkage, (2) complete linkage, (3) average linkage, (4) centroid linkage, and (5) Ward linkage.

Meanwhile, analysis of non-hierarchical cluster or partitional clustering grouping objects together to form a certain number of clusters [5]. One of the non-hierarchical cluster analysis that are often used is K-Means. Research on the K-Means cluster analysis has been carried out starting from [11], [12] and [13] in various fields.

In this study, the method used average linkage cluster analysis and K-Means cluster analysis. Because both of methods use the average size in cluster formation.

Average Linkage Cluster Analysis

Average linkage cluster analysis is known as the minimum variance method, where the distance between two clusters is determined based on the average distance from each member of one cluster to other cluster members [14]. The criteria between the two clusters A and B are as follows [10].

$$d_{average}(A,B) = \frac{1}{[A][B]} \sum_{a \epsilon A} \sum_{b \epsilon B} d(a,b) \qquad (2)$$

The mapping of average linkage cluster analysis can be seen in Figure 1.

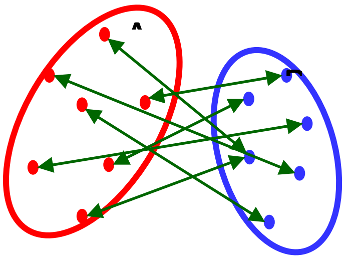

Figure 1. Mapping of average linkage cluster analysis

In general, the algorithm used in average linkage cluster analysis is as follows [15]:

a. Calculate the minimum distance of two objects
b. Combine two objects with a minimum distance into one cluster
c. Find the distance between clusters using the average value
d. Repeat steps 1-2 until all objects are joined into one cluster

The advantages of this method are (1) there is no need to determine the number of clusters beforehand, (2) produces a good visualization that joins the method, (3) uses a dendrogram for graphical representation, and (4) flexibility regarding the level of granularity. On the other hand, the disadvantages of this method are (1) not being able to make corrections when separation or merging is made, (2) lack of interpretation related to cluster description, (3) obscurity of termination criteria (end of the process), and (4) A high level of effectiveness degradation in high-dimensional space [16].

K-Means Cluster Analysis

K-Means Cluster Analysis is one of the non-hierarchical cluster analysis methods used to divide existing data into one or more clusters [17]. The algorithm used in the K-Means cluster analysis is followed [18].

a. Determine k, i.e., the number of groups to be formed
b. Generates k centroid (cluster center point) randomly based on available objects as many as cluster k. Next, to calculate the next i-centroid cluster, the following formula is used:

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \qquad n = 1,2,3,\ldots \qquad (3)$$

where, v : centroid on the cluster , x_i: i^{th} object, n : the number of objects that are members of the cluster
c. Calculate the distance of each object to each centroid in each cluster using Euclidean Distance (Equation 1)

d. Group each data according to the closest distance to the centroid
e. Determine the new centroid position (C_k) by calculating the average value of objects in the same centroid

$$C_k = \left(\frac{1}{n_k}\right) \sum d_i \tag{4}$$

where, n_k : number of members in cluster k, d_i: every member in the cluster k.

The advantages of K-Means cluster analysis are (1) the time needed in computing calculations is faster and easier when implemented, (2) produces relatively good results in the Convex cluster [19], (3) the principle used is simple, (4) K-Means produces high accuracy of the size of the object, this results in this algorithm being relatively more scalable and efficient in managing large data, no effect on the order of objects [20]. Meanwhile, the disadvantages of this method are (1) it is difficult to determine the number of clusters before knowing the optimal number of clusters [21], (2) the initiation of k value is done randomly, so the grouping of clustered data obtained can also vary, this results if the initiation is not good then the grouping obtained is not optimal, (3) the use of k random fruit, makes no guarantee in finding the optimal cluster of clusters, (4) cannot handle data that has outliers [22].

The assumption that needs to be considered in cluster analysis is the absence of multicollinearity on the research variables. Multicollinearity is defined as a linear relationship between research variables. To find out whether there is multicollinearity among research variables can be based on VIF values. If the VIF value of the research variable is less than 10, then there is no symptom of multicollinearity on the research variable.

$$VIF_i = \frac{i}{1-R_i^2} \tag{5}$$

Evaluation of Clustering Results

In determining the best method in cluster analysis, it can use analysis of variance, namely variance within cluster (V_w) and variance between cluster (V_b) [23],[24]. The ideal cluster has a minimum variance within cluster that represents homogeneity within the cluster and has a maximum variance between cluster that illustrates the heterogeneity between clusters [25].

The formula for variance within cluster (V_w) is as follows [26].

$$V_w = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) V_i^2 \tag{6}$$

where, N: number of objects, k : number of clusters, n_i: number of members in cluster i^{th}, V_i^2: variance of cluster i^{th}.

The formula for variance between cluster (V_b), is presented in Equation 7

$$V_b = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{x_\iota} - \bar{x})^2 \tag{7}$$

where, k : number of clusters, n_i: number of members in cluster i^{th}, $\bar{x_\iota}$: mean of cluster $i^{th}, \bar{x}$: mean of all data.

Furthermore, to find out the variance of all clusters can be measured by comparing values variance within cluster (V_w) and variance between cluster (V_b) [26].

$$V = \frac{V_W}{V_B} \tag{8}$$

Visualization of Results

After the clustering process is obtained, visualization is given using a Geographic Information System (GIS). Making visualizations using the R program with packages GISTools, sp, OpenStreetMap, raster, and rJava [27].

Various previous studies have used GIS in research, including Bunch, that uses GIS in spatial planning and environmental management [6]. Harahap in applying GIS for zoning fishing routes in the waters of Kalimantan Barat [28] and Grace is doing research related to GIS in determining health facilities [7].

Results and Discussions

Data Description

A descriptive analysis of the research variables is presented in Table 2.

Table 2. Description of research variable data

Variables	Data Summary		
	Minimum Value	Mean	Maximum Value
Population Density (per km^2)	9	727	15624
Labor Force	330731	3766551	22391003
Population Growth Rate (%)	0.64	1.76	3.9
The Average per Capita Expenditure (Rp)	935538	1303697	1997446
Life Expectancy (Years)	64.34	69.41	74.74
The Average Length of Schooling (Years)	6.27	8.26	11.02

The six variables used in this study have different units. Therefore, to do cluster analysis, it is necessary to standardize.

Multicollinearity Assumption Check

Before a cluster analysis is performed, it is necessary to check whether there are multicollinearity symptoms in the research variable based on the VIF value (see **Equation 4**). VIF values for each variable can be seen in Table 3.

Table 3. VIF value for the research variable

Variables	VIF Value
Population Density (per km^2)	2.381
Labor Force	1.868
Population Growth Rate (%)	1.763
The Average per Capita Expenditure (Rp)	2.377
Life Expectancy (Years)	1.866
The Average Length of Schooling (Years)	1.976

Based on **Table 3**, it is known that the VIF value for all study variables is less than 10. Thus, it can be concluded that there is no multicollinearity among the research variables used so that that cluster analysis can be carried out.

Cluster Analysis

a. The Cluster Members

In cluster analysis using average linkage obtained three clusters consisting of 30 provinces in the first cluster, 1 province in the second cluster, and 3 provinces in third cluster. Meanwhile, the K-Means cluster analysis produced 3 clusters with members of 21 provinces in the first cluster, 3 provinces in the second cluster, and 10 provinces in the third cluster. Members of each cluster can be seen in Table 4.

Table 4. The member of each cluster

Cluster	The member of cluster based on Average Linkage Clustering Analysis	The member of cluster based on K-Means Cluster Analysis
1	Aceh, North Sumatera, West Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Lampung, Bangka Belitung, Riau Islands, D.I. Yogyakarta, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Borneo, Central Borneo, South Borneo, East Borneo, North Borneo, North Sulawesi, North Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Papua.	Aceh, North Sumatera, West Sumatera, Jambi, South Sumatera, Bengkulu, Lampung, West Nusa Tenggara, East Nusa Tenggara, West Borneo, Central Borneo, South Borneo, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi Maluku, North Maluku, West Papua, Papua.
2	DKI Jakarta	West Java, Central Java, East Java, Riau, Bangka Belitung, Riau Islands, DKI Jakarta, D.I. Yogyakarta.
3	West Java, Central Java, East Java	Banten, Bali, East Borneo, North Borneo, North Sulawesi.

b. Evaluation of Clustering Results

Based on Equation 8, the variance obtained for each cluster analysis, as follows:

- Average linkage cluster analysis

$$V = \frac{V_w}{V_B} = 0.152$$

- K-Means cluster analysis

$$V = \frac{V_w}{V_B} = 0.101$$

Thus, it is known that the K-Means cluster analysis variance value is smaller than the average linkage cluster analysis variance value so that the clustering provinces in Indonesia based on the community welfare indicator used in this study is K-Means cluster analysis.

c. Cluster Characteristics

The characteristics of each cluster are described in Table 5.

Table 5. The average value of the variables in each cluster

Variables	Cluster 1	Cluster 2	Cluster 3
Population Density (per km^2)	96	1074	1948
Labor Force	2094333	20446443	2274241
Population Growth Rate (%)	1.72	0.98	2.07
The Average per Capita Expenditure (Rp)	1223834	1100754	1532292
Life Expectancy (Years)	67.94	72.45	71.59
The Average Length of Schooling (Years)	7.97	7.58	9.07

Cluster 1

Based on Table 5, it is known that the first cluster is the province with the population density, labor force, and life expectancy, which the lowest value compared to the other clusters. Meanwhile, the rate of population growth, the average expenditure per capita, and the average length of school with a higher value than the provinces in the second cluster. Provinces in this cluster can be provinces that should receive special attention from related parties in terms of community welfare.

Cluster 2

The province in the second cluster is the province with the population density, labor force, and life expectancy, which the highest value compared to other clusters. On the other hand, in terms of population growth rate, the average expenditure per capita and the average length of schooling of the provinces in this cluster have the lowest value compared to the provinces in other clusters. The province in this cluster is centered on the Jawa island, which needs to get special attention on several aspects of social welfare by related parties.

Cluster 3

The third cluster consists of provinces with the highest values of population growth rate, per capita expenditure and an average length of school compared to other clusters. Meanwhile, for population density, the number of the labor force, and the life expectancy of the province in this cluster is still in the medium when compared to the provinces in other clusters.

Visualization of Cluster Results

Based on the research method, after obtaining the cluster results, visualization of clustering is done using the R program in Figure 2.

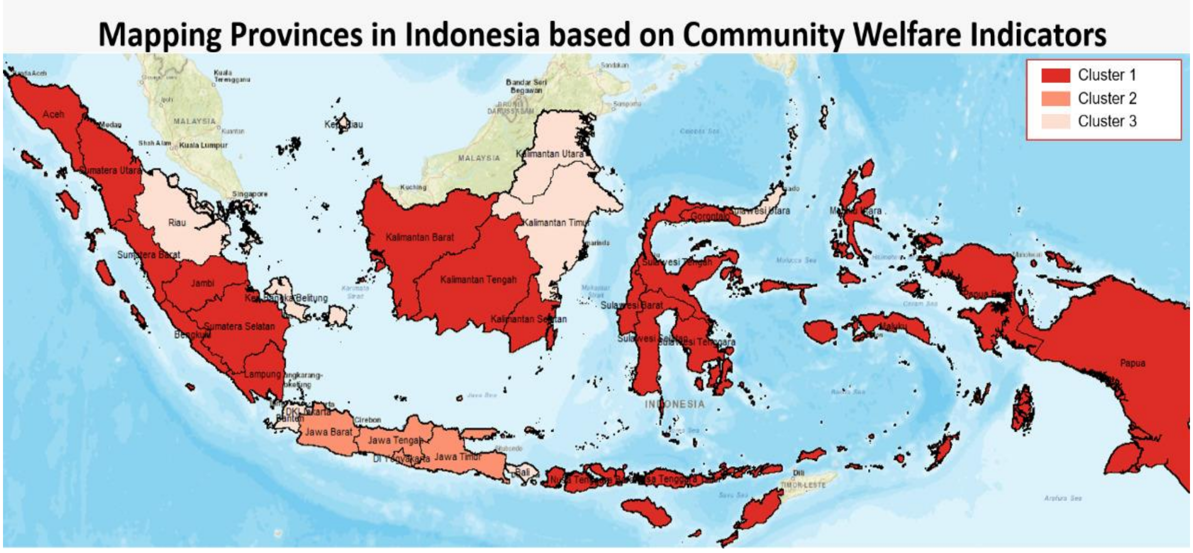

Figure 2. Visualization of cluster results

From Figure 2., the areas with colors according to the cluster results obtained are presented. The red color states cluster 1 which consists of 21 provinces (61.76%), cluster 2 is visualized with a bright red color consisting of 3 provinces (8.82%), while the other is cluster 3. When seen at a glance, in general, the provinces in Indonesia are still categorized to receive special attention by the parties concerned in terms of public welfare.

After it is known that clustering is based on community welfare indicators in each province, further research can be done related to the dominant factors of each province. Not only that, but it can also be continued with more detailed levels, namely the spatial effects of each region. Research conducted can use the Spatial Error Model (SEM), the Spatial Lag Model (SAR), the Spatial Autoregressive Moving Average (SARMA), or the Spatial Data Panel.

Conclusion

Based on the analysis, the results of clustering 34 provinces in Indonesia with indicators of community welfare consisting of six indicators obtained the best method using K-Means cluster analysis. K-Means cluster analysis was chosen because it is based on a variance value (0.101) which is smaller than the Average Linkage Clustering variance value (0.152). In addition, it is known that the welfare of the people in Indonesia is still uneven. This can be seen in the characteristics of each cluster. Each cluster has different priorities in the area of welfare, which need to be considered or improved by related parties. Provinces in the first cluster can be a priority in many aspects of welfare for the government so that welfare can be felt fairly by the community. Meanwhile, provinces in the second and third groups still need attention in several aspects of community welfare. Cluster visualization in the form of a map of Indonesia is expected to make it easier to describe the general condition of the welfare of provinces in Indonesia.

References

[1] Ministry of Foreign Affairs of the Republic of Indonesia, The Constitution of the Republic of Indonesia of 1945, 1 (2011) 1–166.

[2] Undang-Undang Republik Indonesia Nomor 11 Tahun 2009 Tentang Kesejahteraan Sosial, in: 2009.

[3] S. Yulianto, K.H. Hidayatullah, Analisis Klaster Untuk Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Indikator Kesejahteraan Rakyat, Statistika. 2 (2014) 56–63.

[4] S. Soemartini, E. Supartini, Analisis K-Means Cluster Untuk Pengelompokan Kabupaten/Kota di Jawa Barat Berdasarkan Indikator Masyarakat, Prosiding. (2017) 144–154.

[5] P. Govender, V. Sivakumar, Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), Atmos. Pollut. Res. 11 (2020) 40–56.

[6] M.J. Bunch, T.V. Kumaran, R. Joseph, Using Geographic Information Systems (GIS) For Spatial

Planning and Environmental Management in India: Critical Considerations, 2012.

[7] I.U. Grace, Application of Geographic Information Systems (GIS) in the Selection of Suitable sites for health facilities establishment, Texila Int. J. Public Heal. 4 (2016) 237–243.

[8] BPS, Statistik Indonesia 2019, Badan Pusat Statistik, 2019. https://www.bps.go.id/.

[9] R. a. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis - International Edition, 2014.

[10] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O.P. Patel, A. Tiwari, M.J. Er, W. Ding, C.T. Lin, A review of clustering techniques and developments, Neurocomputing. (2017).

[11] H. Steinhaus, Sur la division del corps matériels en parties, Bull. l'Académie Pol. Del Sci. - Cl. III. IV (1956) 801–804.

[12] G.H. Ball, D.J. Hall, A clustering technique for summarizing multivariate data., Behav. Sci. 12 (1967) 153–155.

[13] J. MacQUEEN, Some Methods For Classification and Analysis of Multivariate Observations, in: FIFTH BERKELEY Symp., n.d.: 281–297.

[14] J.H. Ward, Hierarchical Grouping to Optimize an Objective Function, J. Am. Stat. Assoc. 58 (1963) 236–244.

[15] G. Abdurrahman, Clustering Data Kredit Bank Menggunakan Algoritma Agglomerative Hierarchical Clustering Average Linkage, JUSTINDO (Jurnal Sist. Dan Teknol. Inf. Indones. 4 (2019) 13.

[16] A. Bhagat, N. Kshirsagar, P. Khodke, K. Dongre, S. Ali, Penalty Parameter Selection for Hierarchical Data Stream Clustering, Procedia Comput. Sci. 79 (2016) 24–31.

[17] R.H.B. Bangun, Analisis Klaster Non-Hierarki dalam Pengelompokan Kabupaten/Kota di Sumatera Utara Berdasarkan Faktor Produksi Padi, JURNAL AGRICA 9(1) (2016) 54-61.

[18] Y. Agusta, K-Means–Penerapan, Permasalahan dan Metode Terkait, Jurnal Sistem dan Informatika 3(1) (2007) 47-60.

[19] C. Zhang, Z. Fang, An Improved K-means Clustering Algorithm, in: J. Inf. Comput. Sci. (2013) 193–199.

[20] B. Simamora, Analisis multivariat pemasaran, Gramedika Pustaka Utama, Jakarta, 2005.

[21] L. Rahmah, Perbandingan Hasil Penggerombolan K-Means, Fuzzy K-Means, Dan Two Step Clustering, J. Pendidik. Mat. 2 (2017) 39.

[22] S. Andayani, Pembentukan cluster dalam Knowledge Discovery in Database dengan Algoritma K-Means, SEMNAS Mat. Dan Pend. Mat. 2007. (2007) 1–10.

[23] S. Ray, R.H. Turi, Determination of number of clusters in k-means clustering and application in colour image segmentation, Proc. 4th Int. Conf. Adv. Pattern Recognit. Digit. Tech. (1999) 137–143.

[24] A. Khan, Jumpstart Tableau, 2016. https://doi.org/10.1007/978-1-4842-1934-8.

[25] A.R. Barakbah, K. Arai, Identifying moving variance to make automatic clustering for normal data set, in: IECI Japan Work. 2004 (2004) 26–30.

[26] A.R. Barakbah, Y. Kiyoki, A pillar algorithm for k-means optimization by distance maximization for initial centroid designation, 2009 IEEE Symp. Comput. Intell. Data Mining, CIDM 2009 - Proc. (2009) 61–68.

[27] R. Hijmans, Spatial Data in R, 2019.

[28] S.A. Harahap, I. Yanuarsyah, Aplikasi Sistem Informasi Geografis (SIG) Untuk Zonasi Jalur Penangkapan Ikan Di Perairan Kalimantan Barat, J. Akuatika. 3 (2012) 40–48.