p-ISSN 2338 – 3216
e-ISSN 2528 – 1070

JURNAL STATISTIKA
UNIVERSITAS MUHAMMADIYAH SEMARANG
11(2) 2023: 12-21

Jurnal Statistika
Universitas Muhammadiyah Semarang

K-MEANS ALGORITHM FOR GROUPING PROVINCES IN INDONESIA BASED ON MACROECONOMIC AND CRIMINALITY INDICATORS

Andrea Tri Rian Dani [1*], Fachrian Bimantoro Putra [2], Meirinda Fauziyah[3], Sifriyani[4], Suyitno[5], M Fathurahman[6]

[1,2,3,4,5,6] Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Mulawarman University, Samarinda, Indonesia

***e-mail**: andreatririandani@fmipa.unmul.ac.id

Article Info:
Received: August 15, 2023
Accepted: October 21, 2023
Available Online: November 30, 2023

Keywords:
Criminality; Euclidean Distance; K-Means; Silhouette
.

Abstract: Cluster analysis is a multivariate analysis method that divides n observations into K groups ($K \leq$ n) based on their properties. One of the well-known algorithms in cluster analysis is K-Means. K-Means uses the non-hierarchical principle where at the initial initiation, it is necessary to determine the number of groups in advance. The K-Means algorithm can be used to categorize Indonesian provinces based on macroeconomic variables (such as the percentage of poor people, open unemployment rate, and Gini ratio) and crime rate. The ultimate goal of this research is of course to get optimal grouping results. The similarity measure used is Euclidean Distance. The number of groups tested K=2,3, 4,…,10, and the optimal number of groups with the highest Silhouette value was selected. Based on the results of the analysis, the optimal number of clusters is four. These four clusters have characteristics that distinguish one cluster from another.

1. INTRODUCTION

Sustainable Development Goals (SDGs) are shared commitments and agreements with the goal of achieving community welfare while preserving the environment [1], [2]. The SDGs have universal principles, are integrated with one another, and are inclusive with the intention of ensuring that no one will be left behind [3]. The SDGs are mutually agreed to have 17 goals and 169 targets to be achieved by 2030 [4]–[6]. One of the goals that is of concern to researchers is the 16th goal "Peace, Justice and Strong Institutions". The goal of the 16th goal is to build an inclusive and peaceful society founded on human rights, the rule of law, good governance at all levels, and institutions that are transparent, effective, and responsible [7]. One of the targets in goal 16 is to reduce all forms of violence and related deaths everywhere, to stop abuse and exploitation [8], [9]. The high crime rate will lead to a sense of insecurity and have an impact on economic growth, creating a sense of grudges between communities that can last for generations and so on. A statistical approach can be used to be able to photograph and see the phenomena that occur in relation to the 16th SDGs, especially related to the problem of crime. Cluster analysis is one of statistical method that can be used to view phenomena and obtain mapping results of provinces in Indonesia based on macroeconomic and crime indicators. In this study, researchers used cluster analysis to classify provinces in Indonesia based on macroeconomic indicators and crime.

Multivariate statistical methods like cluster analysis are used to categorize items based on common characteristics [10], [11]. Cluster analysis will group each object that has the

[DOI: 10.14710/JSUNIMUS.11.2.12-21]

closest similarity to other objects in the same group [12], [13]. The algorithm used in the grouping process is K-Means. A non-hierarchical grouping technique called K-Means will divide data into one or more groups [14]–[17]. Several previous studies using the K-Means algorithm include: [15], [18]–[21]. The K-Means grouping algorithm can be applied to various problems, one of which is crime. In this research, K-Means will be used to group provinces in Indonesia based on macroeconomic and crime indicators. There are several studies that serve as references, including:[22]–[24]. Based on the description that has been written, this research will apply the K-Means algorithm to classify provinces in Indonesia based on macroeconomic indicators and crime.

2. LITERATURE REVIEW

2.1. Cluster

Cluster analysis is a set of methods that are automatically used to group objects or data into a cluster based on their similarity[25], [26]. The main goal of cluster analysis is to identify a group of objects that have certain similarities and characteristics that can be separated from other groups[27], [28]. Objects that are in the same cluster are relatively more homogeneous than objects that are in different groups.

2.2. K-Means Algorithm

By dividing the data into various groups, the K-Means algorithm is used to group objects [16], [29], [30]. In general, the K-Means method performs two processes: determining the position of the cluster center and looking for members from each cluster. The stages of the K-Means algorithm are detailed as follows:

a) Set the cluster center to *K* and decide how many clusters you wish to construct.
b) The distance between each data point and the cluster center can then be calculated using Euclidean distance using Equation (1).

$$\boldsymbol{d(x,c)} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{dim}} \boldsymbol{(x_i - c_i)^2}} \tag{1}$$

c) Data should be organized into clusters with the smallest possible distances between the centers.

$$\mathbf{min}\ (\boldsymbol{d(x,c)}) \tag{2}$$

d) The new cluster center update uses the average.
e) If any additional objects are going to another cluster, repeat Steps b through d until they are all moved.

2.3. Silhouette Coefficient

The Silhouette coefficient is one of the evaluation tools for determining the quality and strength of clusters [31], [32]. The Silhoueete calculation stages in general are as follows:

a) Calculate the average distance from item *i* to all objects in the same cluster for each object *i*, $\boldsymbol{a_i}$.

[DOI: 10.14710/JSUNIMUS.11.12.-21]

b) Then for each object *i*, calculate the average distance from an object *i* to all objects in different clusters, then take the minimum value, $\boldsymbol{b_i}$.
c) The Silhouette coefficient $\boldsymbol{S_i}$ is calculated by Equation (3).

$$\boldsymbol{S_i} = \frac{\boldsymbol{b_i - a_i}}{\mathbf{max}\,(\boldsymbol{a_i}, \boldsymbol{b_i})} \tag{3}$$

Silhouette coefficient values range from -1 to 1. Results grouping is said to be good if the silhouette coefficient value is owned equal to 1, meaning that the ith object has joined the correct cluster.

3. METHODOLOGY

3.1. Data and Source Data

The data used in this study is secondary data published by the Badan Pusat Statistik (BPS) through its publication in 2021. The observation units in this study were 34 provinces in Indonesia. The research variables consist of response and predictor variables which are detailed in Table 1.

Table 1. Research Variable

Variables	Notation	Explanation
Crime Rate	x_1	Crime rate is a number that can indicate the level of vulnerability of a crime in a certain area at a certain time
Percentage of Poor Population	x_2	The percentage of poor people represents residents who have an average monthly per capita expenditure below the poverty line.
Open Unemployment Rate	x_3	The Open Unemployment Rate is the ratio of the number of jobless to the entire labor force.
Gini Ratio	x_4	The Gini Ratio describes overall equity and inequality, from income to distribution

3.2. Steps of Analysis

The steps in the K-Means algorithm data analysis method for identifying Indonesian provinces are described below:

1. Exploring data with descriptive statistics and making spatial mapping for each variable used.
2. As a first step before utilizing the K-Means method to group the data, standardize the data using the Z-Score. Standardization is used to make the data scale for each variable within the same range.
3. Apply the K-Means algorithm to the grouping process. K=2, 3, 4,..., 10 groups were examined, and the Euclidean Distance was employed to determine similarity.
4. Determine the optimal number of groups with Silhouette coefficients.
5. Profiling the results of grouping provinces in Indonesia by visualizing them with spatial mapping.

[DOI: 10.14710/JSUNIMUS.11.12.-21]

4. RESULTS AND DISCUSSION

We'll go over the analysis's findings in this section, along with a discussion of how the K-Means algorithm was used to group Indonesia's provinces.

4.1. Descriptive Statistics

The data exploration will be presented by displaying descriptive statistics for each variable in Table 2.

Table 2. Research Variable

Variables	Notation	Minimum Value	Maximum Value	Average
Crime Rate	x_1	29	328	138.97
Percentage of Poor Population	x_2	4.53	26.8	10.29
Open Unemployment Rate	x_3	2.34	8.31	4.96
Gini Ratio	x_4	0.26	0.46	0.34

Based on Table 2, it can be seen that each province in Indonesia tends to have different characteristics for all variables. As an example variable, namely the Crime Rate. West Papua Province is the Province with the highest Crime Rate, with a crime rate of 328 (per 100,000 population), followed by Maluku and North Sulawesi Provinces.

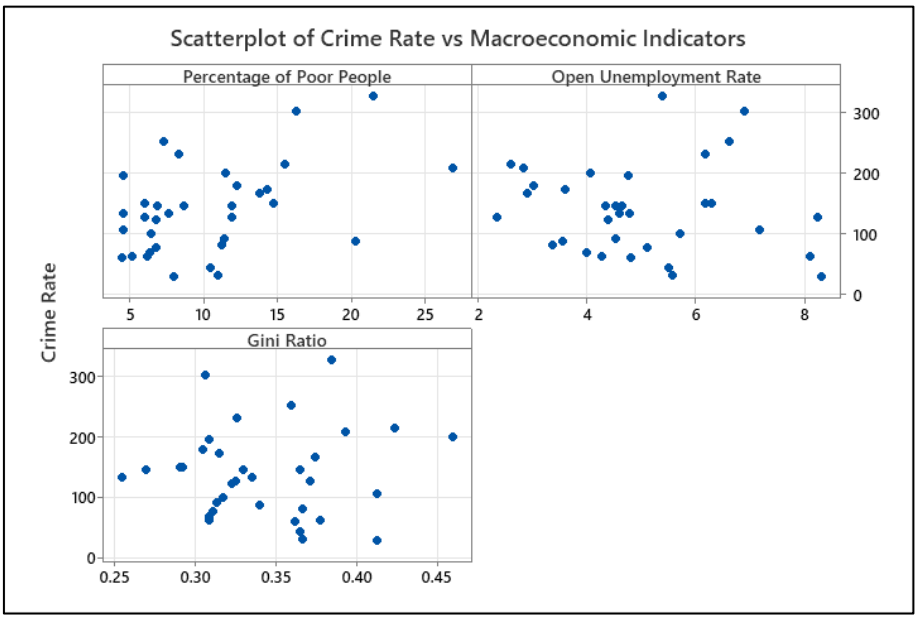


Fig 1. Scatter Plot for Macroeconomic Variables vs Crime Rate

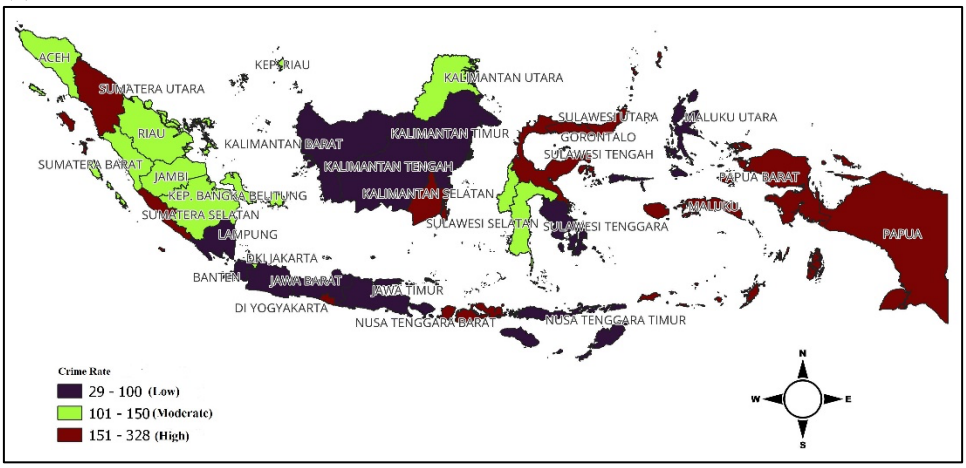
Based on Fig 1, it can be seen that there is a tendency for there to be a positive relationship between the percentage of poor population variable and the crime rate variable, a negative relationship between the open unemployment rate variable and the crime rate variable, and a positive relationship between the Gini ratio variable.

[DOI: 10.14710/JSUNIMUS.11.12.-21]

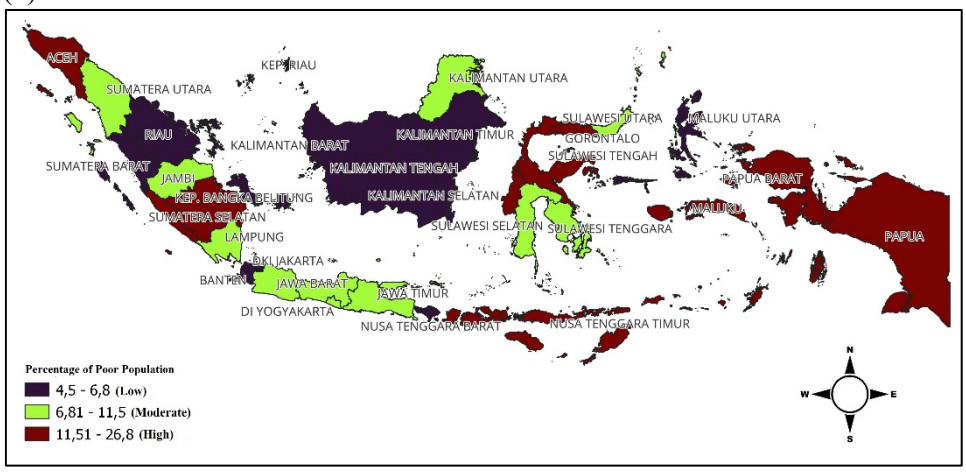
4.2. Spatial Mapping

Considering that in this study, the unit of observation is the provinces in Indonesia, which are location-based, so that spatial mapping is carried out as a data exploration technique. Spatial mapping for each variable is presented in Fig. 2.

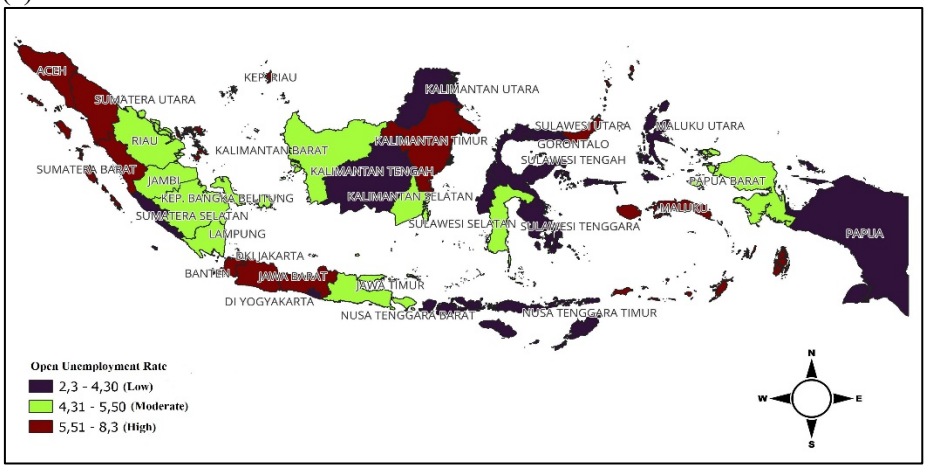
(a)



(b)



(c)



[DOI: 10.14710/JSUNIMUS.11.12.-21]

(d)

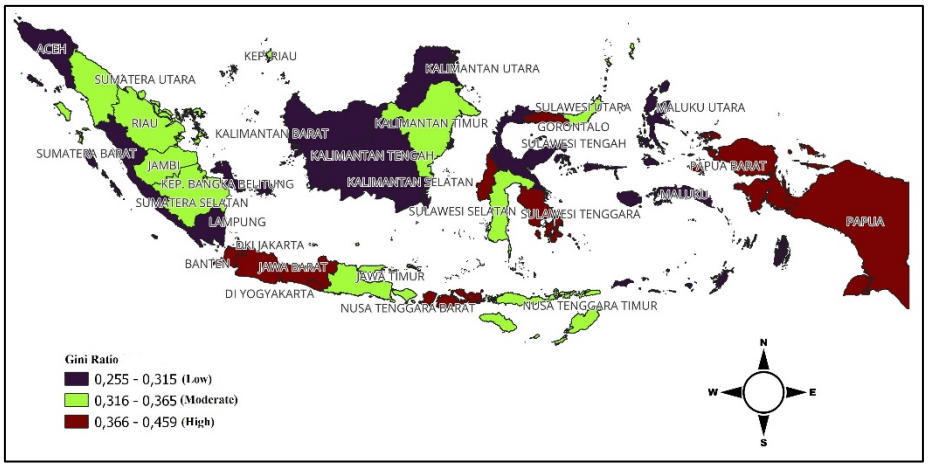


Fig 2. Spatial Mapping: (a) Crime Rate, (b) Percentage of Poor Population, (c) Open Unemployment Rate, (d) Gini Ratio

Based on Figure 2, it can be seen that for the crime rate variable, provinces on the island of Papua tend to have higher crime rates than others. This is similar to what happened on Papua Island for the variables of percentage of poor population and Gini ratio

4.3. Grouping with the K-Means Algorithm

The process of grouping provinces in Indonesia using the K-Means Algorithm is carried out using the help of R software with the packages "tidyverse", "cluster", and "factoextra". Silhouette value is used as a criterion for the goodness of grouping results. In this study, the distance for measuring similarity used was the Euclidean distance and the number of groups tested was K=2,3,4,…,10.

Table 3. Silhouette Coefficient Value for each K tested

Number of Groups (K)	Silhouette Coefficient
2	0.55
3	0.54
4	**0.57**
5	0.52
6	0.51
7	0.53
8	0.54
9	0.56
10	0.55

Furthermore, can be visualized in Fig 3.

[DOI: 10.14710/JSUNIMUS.11.12.-21]

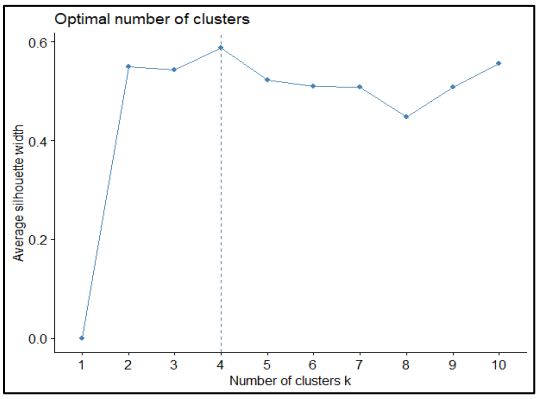


Fig 3. Visualization of Silhouette Coefficient

Based on the Silhouette, the optimal number of groupings is 4 clusters / group, because it has the highest Silhouette value. Table 4 displays the findings of profiling the Indonesian provinces based on Fig. 4.

Table 4. Province Grouping

Cluster	Province
1	Maluku and Papua Barat
2	Kep. Bangka Belitung, Aceh, Jambi, Riau, , Kep. Riau, Nusa Tenggara Barat, Kalimantan Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Barat, Sumatera Barat, and Sulawesi Selatan
3	Gorontalo, Papua, Kalimantan Selatan, Sulawesi Utara, Sumatera Utara, and DI Yogyakarta
4	Sulawesi Tenggara, Nusa Tenggara Timur, Banten, DKI Jakarta, Bali, Kalimantan Barat, Kalimantan Tengah, Kalimantan Timur, Lampung, Jawa Barat, Jawa Tengah, Jawa Timur, and Maluku Utara

Based on Table 4, it can be visualized the results of grouping Provinces in Indonesia in the form of spatial mapping in Fig. 4.

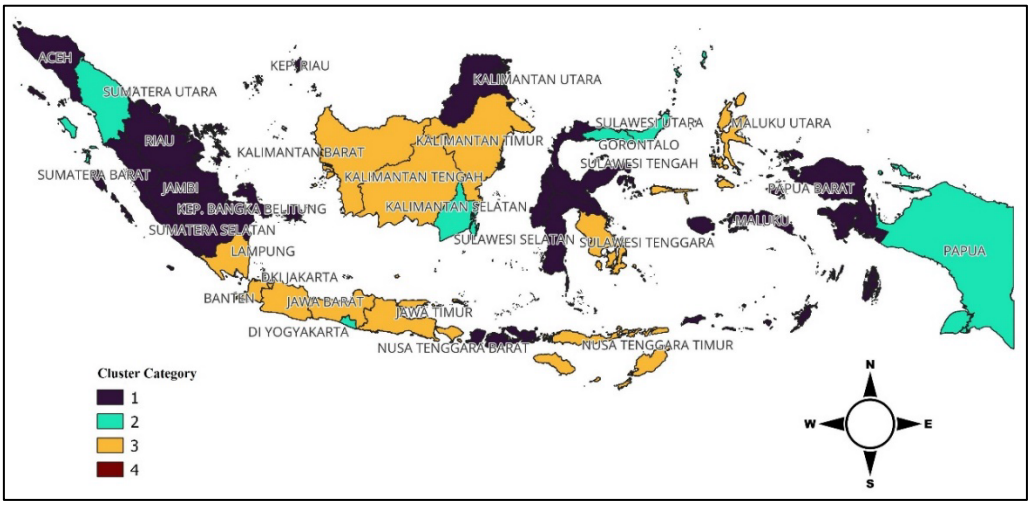


Fig 4. Visualization of Grouping Provinces

[DOI: 10.14710/JSUNIMUS.11.12.-21]

Based on the grouping results, we can know if:

1. Cluster 1 is the provinces with the highest Percentage of Poor Population, Open Unemployment Rate, Gini Ratio, and Crime Rate. Cluster 1 is a group of provinces with high crime rates and other economic problems that follow.
2. Cluster 2 is the provinces with the Percentage of Poor Population, Open Unemployment Rate, Gini Ratio, and Crime Rate which also need attention, although the numbers are not as high as in Cluster 1 and Cluster 3.
3. Cluster 3 is the provinces with the second highest Percentage of Poor Population, Gini Ratio, and Crime Rate after Cluster 1.
4. Cluster 4 is the provinces with the lowest percentage of poor people and crime rates compared to other provinces. However, the Open Unemployment Rate and the Gini Ratio are still problematic in this cluster.

5. CONCLUSION

As a result of the data analysis and discussion, the best number of clusters, according to the Silhouette value, is four, with each cluster possessing traits that distinguish it from the others. Indonesian provinces are divided into (4) four clusters, with details: cluster 1 has (2) two provinces, cluster 2 has (13) thirteen provinces, cluster 3 has (6) six provinces, and cluster 4 has (13) thirteen provinces.

The recommendation for the government based on the research results is to pay attention to provinces that are included in the crime-prone group. Based on the cluster results, the government can prioritize handling for early mitigation of problems related to macroeconomics and crime. Cluster 1 is the cluster that requires the most attention, including the provinces of Maluku and West Papua.

ACKNOWLEDGMENT

This research was funded in part by DIPA BLU-PNBP FMIPA, Mulawarman University, Samarinda, Indonesia [No: 1700 / UN17.7 / LT/ 2023].

REFERENCES

[1] M. F. Cordova and A. Celone, “SDGs and innovation in the business context literature review,” *Sustainability (Switzerland)*, vol. 11, no. 24, Dec. 2019, doi: 10.3390/su11247043.

[2] S. Panuluh Meila Riskia Fitri, “Perkembangan Pelaksanaan Sustainable Development Goals (SDGs) di Indonesia,” 2015. [Online]. Available: www.infid.org

[3] F. Irhamsyah, “Sustainable Development Goals (SDGs) dan Dampaknya Bagi Ketahanan Nasional,” *Jurnal Kajian Lemhannas RI*, no. 38, pp. 45–54, 2019, [Online]. Available: www.unsplash.com

[4] D. A. Sari *et al.*, “Performance Auditing to Assess the Implementation of the Sustainable Development Goals (SDGs) in Indonesia,” *Sustainability (Switzerland)*, vol. 14, no. 19, Oct. 2022, doi: 10.3390/su141912772.

[5] N. Rulandari, “Study of Sustainable Development Goals (SDGS) Quality Education in Indonesia in the First Three Years,” *Budapest International Research and Critics Institute (BIRCI-Journal): Humanities and Social Sciences*, vol. 4, no. 2, pp. 2702–2708, May 2021, doi: 10.33258/birci.v4i2.1978.

[6] S. Syamsuri, Y. Sa'adah, and I. A. Roslan, "Reducing Public Poverty Through Optimization of Zakat Funding as an Effort to Achieve Sustainable Development Goals (SDGs) in Indonesia," *Jurnal Ilmiah Ekonomi Islam*, vol. 8, no. 1, p. 792, Mar. 2022, doi: 10.29040/jiei.v8i1.3872.

[7] S. Leite, "Using the SDGs for global citizenship education: definitions, challenges, and opportunities," *Globalisation, Societies and Education*, vol. 20, no. 3, pp. 401–413, 2022, doi: 10.1080/14767724.2021.1882957.

[8] K. Daerah Istimewa Yogyakarta Pendekatan Ekonomi, "Analisis Faktor-Faktor yang Mempengaruhi Kriminalitas di."

[9] E. Yulia Purwanti, J. Ilmu Ekonomi Studi Pembangunan FEB Undip Eka Widyaningsih, and J. Ilmu Ekonomi Studi Pembangunan FEB Undip, "Analisis Faktor Ekonomi yang Mempengaruhi Kriminalitas Di Jawa Timur," vol. 9, no. 2, 2019, [Online]. Available: http://jurnal.untirta.ac.id/index.php/Ekonomi-Qu

[10] R. Novidianto and A. T. R. Dani, "Analisis Klaster Kasus Aktif COVID-19 Menurut Provinsi di Indonesia Berdasarkan Data Deret Waktu," *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol. 5, pp. 15–24, 2020.

[11] D. Widyadhana, R. B. Hastuti, I. Kharisudin, and F. Fauzi, "Perbandingan Analisis Klaster K-Means dan Average Linkage untuk Pengklasteran Kemiskinan di Provinsi Jawa Tengah," *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 4, pp. 584–594, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[12] G. D. Rembulan, T. Wijaya, D. Palullungan, K. N. Alfina, and M. Qurthuby, "Kebijakan Pemerintah Mengenai Coronavirus Disease (COVID-19) di Setiap Provinsi di Indonesia Berdasarkan Analisis Klaster," *JIEMS (Journal of Industrial Engineering and Management Systems)*, vol. 13, no. 2, Sep. 2020, doi: 10.30813/jiems.v13i2.2280.

[13] Z. He, X. Xu, and S. Deng, "Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach," no. October 2005, 2005, [Online]. Available: http://arxiv.org/abs/cs/0509011

[14] L. F. Marini and C. D. Suhendra, "Penggunaan Algoritma K-Means Pada Aplikasi Pemetaaan Klaster Daerah Pariwisata," *Jurnal Media Informatika Budidarma* , vol. 7, no. 2, pp. 707–713, 2023, doi: 10.30865/mib.v7i2.5558.

[15] A. Wahyu and Rushendra, "Klasterisasi Dampak Bencana Gempa Bumi Menggunakan Algoritma K-Means di Pulau Jawa," *Jurnal Edukasi dan Penelitian Informatika*, vol. 8, no. 1, pp. 175–179, 2022.

[16] R. Madhuri, M. R. Murty, J. V. R. Murthy, P. V. G. D. P. Reddy, and S. C. Satapathy, "Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms," *Advances in Intelligent Systems and Computing*, vol. 249 VOLUME, pp. 137–144, 2014, doi: 10.1007/978-3-319-03095-1_15.

[17] H. Sofyan, M. Iqbal, M. Marzuki, and M. Muhammad, "The comparison of k-modes clustering and ROCK clustering to the poverty indicator in Samadua Subdistrict, South Aceh," *IOP Conf Ser Mater Sci Eng*, vol. 1087, no. 1, p. 012085, 2021, doi: 10.1088/1757-899x/1087/1/012085.

[18] C. Purnama, W. Witanti, and P. N. Sabrina, "Klasterisasi Penjualan Pakaian untuk Meningkatkan Strategi Penjualan Barang Menggunakan K-Means," *Journal of Information Technology*, vol. 04, no. 1, pp. 35–58, 2022.

[19] A. Munawar *et al.*, "Cluster Application with K-Means Algorithm on the Population of Trade and Accommodation Facilities in Indonesia," *J Phys Conf Ser*, vol. 1933, no. 1, 2021, doi: 10.1088/1742-6596/1933/1/012027.

[20] E. Banjarnahor, A. Bustamam, T. Siswantining, and P. Tampubolon, "Analyzing Kinship in Severe Acute Respiratory Syndrome Coronavirus 2 DNA Sequences Based on Hierarchical and K-Means Clustering Methods Using Multiple Encoding Vector," *Int J Adv Sci Eng Inf Technol*, vol. 12, no. 6, pp. 2237–2247, 2022, doi: 10.18517/ijaseit.12.6.15582.

[21] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl Eng*, vol. 63, no. 2, pp. 503–527, 2007, doi: 10.1016/j.datak.2007.03.016.

[22] H. S. Firdaus, A. L. Nugraha, B. Sasmito, M. Awaluddin, and C. A. Nanda, "Perbandingan Metode Fuzzy C-Means dan K-Means Untuk Pemetaan Daerah Rawan Kriminalitas Di Kota Semarang," *Elipsoida*, vol. 04, no. 01, pp. 58–64, 2021.

[23] A. D. Lestari, T. W. Utami, and M. Al Haris, "Pengelompokan Provinsi Di Indonesia Berdasarkan Kriminalitas Menggunakan Metode Ward Dan K-Medoids." [Online]. Available: http://repository.unimus.ac.id

[24] R. N. Fahmi, M. Jajuli, N. Sulistiyowati, and U. S. Karawang, "Analisis Pemetaan Tingkat Kriminalitas Di Kabupaten Karawang Menggunakan Algoritma K-Means Mapping Analysis Of Criminality Level In Karawang Using K-Means Algorithm," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 4, no. 1, 2021, [Online]. Available: www.pasundanekspres.co

[25] Z. He, X. Xu, and S. Deng, "A cluster ensemble method for clustering categorical data," *Information Fusion*, vol. 6, no. 2, pp. 143–151, 2005, doi: 10.1016/j.inffus.2004.03.001.

[26] J. C. Gower, "A Comparison of Some Methods of Cluster Analysis," *Biometrics*, vol. 23, no. 4, p. 623, 1967, doi: 10.2307/2528417.

[27] S. Sarumathi, P. Ranjetha, C. Saraswathy, M. Vaishnavi, and S. Geetha, "A Review and Comparative Analysis on Cluster Ensemble Methods," *International Journal of Computer and Information Engineering*, vol. 15, no. 6, pp. 385–394, 2021.

[28] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J Intell Inf Syst*, vol. 17, no. 2–3, pp. 107–145, 2001, doi: 10.1023/A:1012801612483.

[29] E. Herman, K. E. Zsido, and V. Fenyves, "Cluster Analysis with K-Mean versus K-Medoid in Financial Performance Evaluation," *Applied Sciences (Switzerland)*, vol. 12, no. 16, 2022, doi: 10.3390/app12167985.

[30] J. I. M. Araujo *et al.*, "Non-hierarchical cluster analysis for determination of resistance to worm infection in meat sheep," *Trop Anim Health Prod*, vol. 53, no. 1, 2021, doi: 10.1007/s11250-020-02484-3.

[31] R. Hidayati, A. Zubair, A. Hidayat Pratama, L. Indana, P. Studi Sistem Informasi, and F. Teknologi Informasi, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering," 2021.

[32] S. Zhou, F. Liu, and W. Song, "Estimating the Optimal Number of Clusters Via Internal Validity Index," *Neural Process Lett*, vol. 53, no. 2, pp. 1013–1034, 2021, doi: 10.1007/s11063-021-10427-8.