JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 7, No. 2, April 2023, pp. 399-408

Cluster Analysis of Environmental Pollution in Indonesia Using Complete Linkage Method with Elbow Optimization

Adelia Damayanti[1], Wika Dianita Utami[2], Dian Candra Rini Novitasari[3], Putroue Keumala Intan[4], Mohammad Lail Kurniawan[5]
[1,2,3,4]Department of Mathematic, Islamic State of Sunan Ampel Surabaya University, Indonesia
[5]Statistics of Pasuruan City, Indonesia
adeliadamayanti869@gmail.com[1], wikadianita@uinsby.ac.id[2], diancrini@uinsby.ac.id[3], putroue@uinsby.ac.id[4], mlail@bps.go.id[5]

Article History:
Received : 05-01-2023
Revised : 08-03-2023
Accepted : 10-03-2023
Online : 06-04-2023

Keywords:
Cluster Analysis;
Complete Linkage;
Elbow Method;
Environmental Pollution;
Silhouette Coefficient.

ABSTRACT

The issue of environmental contamination remains unsolved. The problem continues to have a substantial detrimental impact. This research aimed to identify provinces in Indonesia with high or low levels of environmental pollution so that the government may offer treatment to provinces with high levels of pollution and seek a significant reduction in the incidence of environmental pollution in Indonesia. Clustering is required to identify provinces with high and low pollution levels using the complete linkage method because this method can provide tight clusters and is less impacted by outliers. The analysis of the complete linkage method with Elbow optimization revealed two optimal clusters, namely high and low clusters. The high cluster consists of three provinces: Central Java, West Java, and East Java. The low cluster consists of 31 provinces. This research used a Silhouette Coefficient validity test. The value of the Silhouette Coefficient is 0.75. The value indicates that the data object is in the correct cluster and that the cluster structure is relatively strong.

https://doi.org/10.31764/jtam.v7i2.12961

A. INTRODUCTION

Humans have an extremely close interaction with their surroundings. However, humans are unconcerned about the influence on ecosystems. The emergence of pollution and environmental degradation is a significant result of its interaction with environmental concerns (Sipayung et al., 2020). According to Statistics Indonesia, there are 5644 villages in Indonesia that pollute the air, 1499 that pollute the soil, and 10683 that pollute the water (Statistik, n.d.). It has an impact on both short and long-term environmental sustainability. Short-term impacts include visual harm to the environment. Meanwhile, long-term consequences, such as ecosystem disturbance and global warming, are more significant. To control environmental pollution in Indonesia, the community, especially the government should be concerned. To support success in managing environmental pollution in Indonesia, research that can cluster provinces with identical pollution circumstances is required. So that, it can become a suggestion for the government in selecting policies. Cluster analysis is one of the clustering approaches

that can be used (Dinata & Syaputra, 2020). Clustering methods can be used to generate more realistic categorization outputs for datasets in a wide range of fields like physical science, biological sciences, psychological, industry, and text documents, and they are a popular topic of research in data mining (Chen et al., 2015).

Cluster analysis is a method for categorizing some items or data into smaller clusters based on their similarity. In cluster analysis, there are two methods: non-hierarchical and hierarchical ones (Drews et al., 2019). Single linkage, complete linkage, average linkage, and ward's method are examples of hierarchical methods (Widyawati et al., 2020). Meanwhile, the K-Means algorithm is a non-hierarchical method (Ramadhani et al., 2018). This study focuses on the hierarchical technique, specifically the complete linkage, because the hierarchical method processes data more quickly, saving time, and the resulting output is in the form of levels or hierarchies, making it easy to analyze (Hidayat, n.d.).

Previous studies have shown the advantages of complete linkage when combining sub-districts in Sidoarjo Regency based on livestock yield potential. This is based on a comparison of the value of the standard deviation ratio, which shows that the value of the complete linkage is the smallest, that is 0.222, indicating that the complete linkage is the best technique (Mu'afa & Ulinnuha, 2019). Another research looking for ideal clusters utilizing the single linkage, complete linkage, and average linkage approaches based on the Human Development Index indicator in West Kalimantan also indicate the benefits of complete linkage. This was assessed by the accuracy value obtained using valley-tracing, which yielded the highest complete linkage value, that is 0.976 with 5 ideal cluster numbers (Hendra Perdana, Nur Asiska, 2019). Similarly, studies on clustering news articles using four linkage approaches, namely single linkage, complete linkage, average linkage, and average linkage-group, have been conducted. The obtained findings show that complete linkage is the best approach since it has the highest average purity of 0.888 and 0.938 (Wibisono & Khodra, 2018).

According to previous research, complete linkage is better than other methods. Complete linkage produces compact clusters with great precision (Nabiilah Ardini Fauziyyah & Sholikhah, 2021). Based on the preceding discussion, this research will be conducted to cluster provinces in Indonesia that suffer environmental pollution using the complete linkage method, so that it may be utilized as reference material and recommendations for the government in reducing incidents of environmental pollution.

B. METHODS

The data in this research is the number of villages in Indonesia that harm the environment per province. The data was gathered from Statistics Indonesia official website. The variables in this research are water pollution (X1), soil pollution (X2), and air pollution (X3). A hierarchical clustering method using complete linkage. The steps in this study are described in a flowchart, as shown in Figure 1.

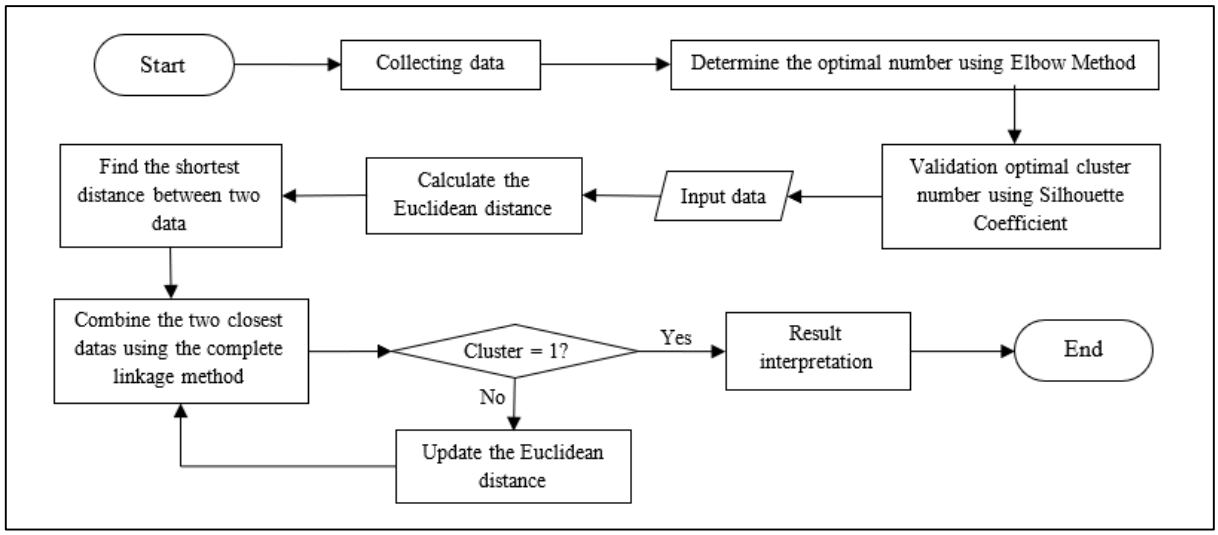

Figure 1. Cluster Analysis Flowchart

The research steps can be better understood with the flowchart above. The explanation for these actions is provided below.

1. Collecting Data

This research used secondary data gathered from the Statistics Indonesia. The data is presented in the Table 1.

Table 1. The Amount of Pollution in Indonesia in 2021

No.	Province	Water Pollution	Soil Pollution	Air Pollution
1	Aceh	350	23	481
2	North Sumatra	673	72	339
3	West Sumatra	193	21	60
4	Riau	250	9	148
5	Jambi	390	16	37
6	South Sumatera	440	73	229
7	Bengkulu	163	8	71
8	Lampung	308	23	210
9	Bangka Belitung Island	100	26	35
10	Riau Island	16	4	22
11	DKI Jakarta	78	10	42
12	West Java	1217	129	556
13	Central Java	1310	224	781
14	DI Yogyakarta	76	8	41
15	East Java	1152	154	777
16	Banten	257	40	197
17	Bali	82	5	15
18	West Nusa Tenggara	152	18	79
19	East Nusa Tenggara	79	35	199
20	West Kalimantan	715	121	155
21	Central Kalimantan	610	125	91
22	South Kalimantan	396	39	140
23	East Kalimantan	227	26	89
24	North Kalimantan	99	31	47
25	North Sulawesi	161	27	85

No.	Province	Water Pollution	Soil Pollution	Air Pollution
26	Central Sulawesi	126	25	60
27	South Sulawesi	308	39	229
28	Southeast Sulawesi	140	23	162
29	Gorontalo	62	7	27
30	West Sulawesi	98	10	46
31	Maluku	53	5	25
32	North Maluku	71	15	63
33	West Papua	39	13	16
34	Papua	292	95	90
	Total	10683	1499	5644

2. Elbow Method

The Elbow approach is a methodology for determining the amount of clusters that should be used according to the proportion of comparison values between the cluster numbers. The Elbow method calculate the SSE (Sum of Square Error) for every cluster result (Muningsih & Kiswati, 2018). The formula of SSE value is as follows.

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left\|x_i^{(j)} - c_i\right\|^2 \tag{1}$$

Where k is number of clusters, i is data index, j is a cluster, n is number of data, $x_i^{(j)}$ is data i in cluster j, and c_i is centreoid or average of the data in a cluster. SSE decreases as expected when k is less than the optimal quantity of clusters. If k approaches the optimal number of clusters, then SSE will decrease dramatically and continue to be stable as k increases. As a result, the drop of SSE will be significant before becoming flat. In other words, the correlation curve between SSE and k has the structure of an elbow, as well as the value of the associated k at this elbow indicates the real cluster number of the data. (Liu & Deng, 2021).

3. Silhouette Coefficient

The Silhouette Coefficient is a method that determines the proximity of relationships between objects in a cluster. This method is used to calculate the proximity of objects in one cluster to those in another. The Silhouette Coefficient value is between -1 and 1. A value of 1 means the objects have been appropriately clustered, and a value of -1 shows that objects have not been efficiently clustered (Ogbuabor & F. N, 2018). The Silhouette Coefficient of an i object is calculated using two variables, a_i and b_i. The formula for calculating a_i and b_i are as follows (Hidayati et al., 2021):

$$a_i = \frac{1}{|P|-1}\sum_{v\in A, v\neq 1} d(u,v) \tag{2}$$

$$b_i = \min_{Q\neq P}\frac{1}{|Q|}\sum_{v\in C} d(u,v) \tag{3}$$

Where *P* is the number of data in cluster *P*, *Q* is the number of data in cluster *Q*, *u* and *v* are the data index, and *d (u, v)* is the distance between the *u* data and the *v* data in one cluster. The Silhouette Coefficient is calculated by the equation below (Widyawati et al., 2020):

$$S_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \tag{4}$$

Where S_i is object *i*'s Silhouette Coefficient value in one cluster, a_i is object *i*'s average length from all other objects within the same cluster, b_i is the minimal value of the mean distance between *i* object and all objects in other cluster that vary from *i* object. The indicators of Silhouette Coefficient value presented in Table 2 (Swindiarto et al., 2018), is required to determine the quality of the final clustering, as shown in Table 2.

Table 2. Kauffman Table

Silhouette Coefficient Value	Structure
0.7 – 1	Strong
0.5 – 0.7	Good
0.25 – 0.5	Weak
≤ 0.25	Bad

4. Complete Linkage

Complete linkage computes the greatest dissimilarity between two objects, as opposed to single linkage. The maximum distance between any two items belonging to separate clusters defines the proximity of two clusters. This method of linkage produces tight clusters and is less affected by outliers (Govender & Sivakumar, 2020). Especially, measuring distance is a crucial part of the clustering process (Cao et al., 2020). The first step of complete linkage is determining the distance matrix between objects. The Euclidean distance, whose formula is provided below, is one of the way for estimating the closeness distance between objects (Cui, 2020):

$$d\ (x, y) = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - y_{jk}\right)^2} \tag{5}$$

Where *d (x, y)* is the separation between objects *x* and *y*, x_{ik} is the value of object *i* in the *k* data, y_{jk} is the value of object *j* in the *k* data, and *n* is the number of objects. Second, from the distance matrix computation, identify the object that produces the shortest or least distance. The third uses the following formula to get the combined cluster distance with the smallest distance (Ramadhani et al., 2018).

$$d_{(XY)A} = max\{d_{XA}, d_{YA}\} \tag{6}$$

Where d_{XA} and d_{YA} are the longest distances between clusters X and A and clusters Y and A, respectively. Fourth, based on prior computations, update the distance matrix. The second

through fourth steps are repeated (n-1) times. All items will collect into a single cluster at the end of the operation.

5. Centroid Value

The centroid or center point for each cluster is calculated by taking the average of all data values in the cluster. Here is the formula for determining the centroid (Rizal, 2013):

$$C = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{7}$$

Where n represents the number of data in a cluster, j represents the j data index in the cluster, and x_j represents the value of data j in a cluster.

C. RESULT AND DISCUSSION

Based on Table 1, Central Java is the province with the most cases of environmental pollution in Indonesia throughout 2021, followed by West Java and East Java. Meanwhile, Riau Islands have fewer pollution cases of water and soil pollution, and Bali province has fewer cases of air pollution. This objective result will be followed by cluster analysis to see the grouping results with the processes below.

1. Determination of the Cluster Optimal Number Using Elbow

The Elbow approach was utilized in this study to identify the appropriate number of cluster assumptions (k). The Elbow curve for environmental pollution in Indonesia showed in Figure 2.

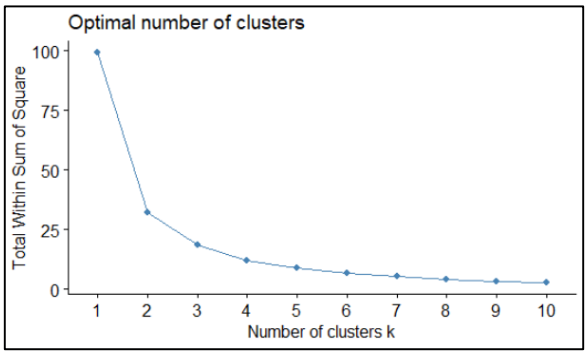

Figure 2. Elbow Curve

We can easily conclude that cluster 2 (k = 2) is the ideal number of cluster because the elbow curve is formed at point 2. The cluster number is assumed to be 2 (k = 2) and 3 (k = 3) to see all choices and assure the optimal number of clusters. Furthermore, to find the ideal number of clusters, these assumptions will be tested using the Silhouette Coefficient technique.

2. Optimal Cluster Number Validation

After assuming the ideal number of clusters using the Elbow method, the assumption result will be confirmed using the Silhouette Coefficient method. The following is a description of the Silhouette Coefficient, as shown in Figure 3 and Figure 4.

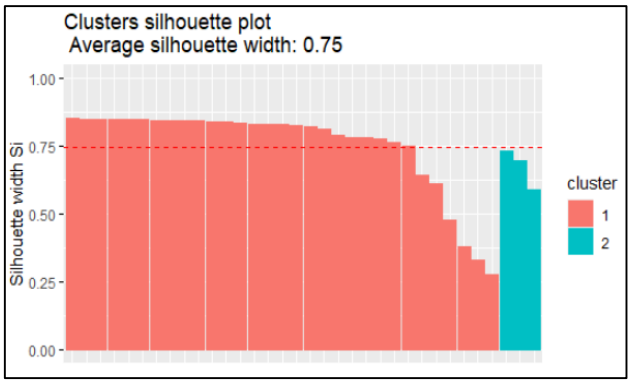

Figure 3. Silhouette Coefficient of 2 Cluster

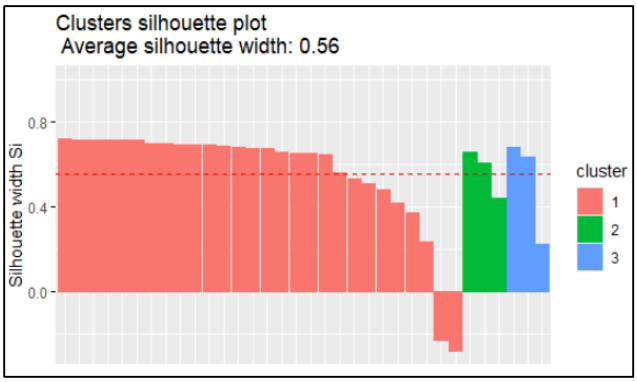

Figure 4. Silhouette Coefficient of 3 Cluster

Each data object in a cluster is represented by a bar chart above. Objects of the same color belong to the same cluster. The average silhouette width is the value of the Silhouette Coefficient of clustering, which is indicated by a dotted line in the graph's center that showed in Figure 3 and Figure 4. A graph that is close to 1 suggests that the data object is in the correct cluster. If it is around -1, the item is not in the proper cluster. (Wijaya et al., 2021). Both of the above figures show that Figure 4 has a more excellent Silhouette Coefficient value than Figure 3, which is 0.75. Based on the Kauffman Table, Table 2 shows a strong cluster structure. There are no data objects near -1, implying that two clusters (k = 2) are the ideal number of clusters for clustering in this research.

3. Complete Linkage Cluster Analysis

The proximity matrix is generated by calculating the Euclidean distance using the formula Equation (5) as the initial step in cluster creation. The results are presented in Table 3.

Table 3. Euclidean Distance Matrix

	P1	**P2**	**P3**	**....**	**P34**
P1	0.0	1.5	2,2	....	2,4
P2	1.5	0.0	2,2	....	1,7
P3	2,2	2,2	0.0	....	1.5
⋮	⋮	⋮	⋮	⋱	⋮
P34	2,4	1,7	1.5	....	0.0

P_i = Provinces on the Table 1, i = 1, 2, ..., 34

The Euclidean distance is used to find the shortest distance between two items. Then, using the formula Equation (6), merge the two closest objects using the complete linkage approach. The Euclidean distance between objects should then be updated to interpret the proximity between the new clusters and the remaining clusters. The method is repeated indefinitely until only one cluster remains. Previously, the Elbow and Silhouette Coefficient approaches yielded the optimal number of clusters as two. As a result, the estimated cluster analysis findings will be divided into two clusters and displayed in Figure 5.

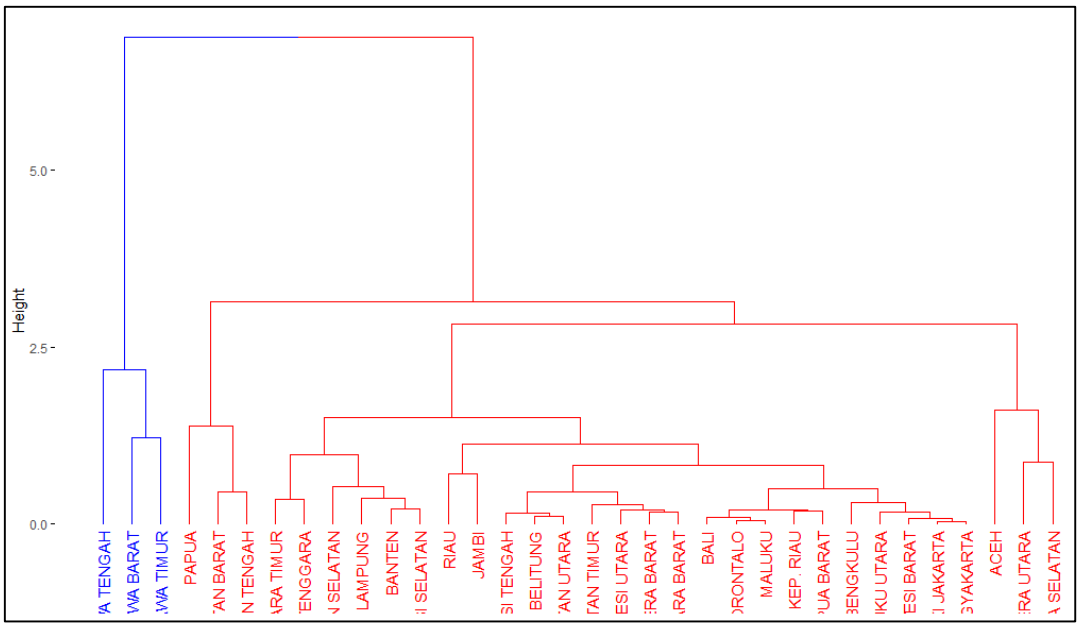

Figure 5. Dendogram of Cluster Analysis Results with 2 Clusters

The dendogram displays the objects that constitute a hierarchy and two clusters. Objects in the same cluster join to form a larger hierarchy, until all objects are gathered in one hierarchy. The blue color represents the first cluster, while the red color represents the second. Furthermore, the cluster state can be determined using the cluster center (centroid). The province with the highest pollution status will be represented by the centroid with the highest score. Meanwhile, the centroid with the lowest score is going to be cluster of low-pollution provinces. The centroid value is derived by averaging the objects in a cluster (Ais et al., 2022). The centroid score of every cluster from each variable is displayed in Table 4.

Table 4. Centroid Value

Cluster	Water pollution	Soil Pollution	Air pollution
1	1226.33	169	704.67
2	225.94	32	113.87

The centroid value of water, soil, and air pollution in cluster 1 is greater than that of cluster 2, indicating that the province in cluster 1 has a higher pollution status than the province in cluster 2. As a result, cluster 1 is a province with a high pollution status in 2021. Cluster 2 is a collection of provinces with low pollution levels. High cluster provinces include Central Java, West Java, and East Java. The following shows the members of each cluster, as shown in Table 5.

Table 5. Result of Cluster Analysis

Cluster	Cluster Members (Province)	Number of Cluster Members	Status
1	Central Java, West Java, East Java	3	High
2	Aceh, North Sumatra, South Sumatra Papua, West Kalimantan, Central Kalimantan NTT, Southeast Sulawesi, South Kalimantan, Lampung, Banten, South Sulawesi, Riau, Jambi, Central Sulawesi, Kep. Bangka Belitung, North Kalimantan, East Kalimantan, North Sulawesi, West Sumatra, NTB, Bali, Gorontalo, Maluku, Kep. Riau, West Papua, Bengkulu, North Maluku, West Sulawesi, DKI Jakarta, DI Yogyakarta	31	Low

D. CONCLUSION AND SUGGESTIONS

According to the clustering algorithm's estimate, the optimal number of clusters created is two. The two clusters are divided into high and low pollution clusters. The high cluster comprises three provinces: Central Java, West Java, and East Java. The low cluster is made up of 31 provinces. A Silhouette Coefficient value is 0.75, indicating that data objects were in the correct cluster and cluster structure was classed as strong.

The results of this research can be used by governments and citizens to reduce environmental pollution. The government can implement citizen activities to encourage a healthy environment. It is suggested that future research implement non-hierarchical methods like K-Means. Besides the silhouette coefficient, it can also include internal validation procedures in determining the accuracy of the research outcomes.

REFERENCES

Ais, C., Hamid, A., Candra, D., & Novitasari, R. (2022). Analysis of Livestock Meat Production in Indonesia Using Fuzzy C-Means Clustering. *Jurnal Ilmu Komputer Dan Informasi (Journal of Computer Science and Information)*, *15*(1), 1–8.

Cao, R., Li, B., Wang, Z., Peng, Z. R., Tao, S., & Lou, S. (2020). Using a Distributed Air Sensor Network to Investigate the Spatiotemporal Patterns of PM2.5 Concentrations. *Environmental Pollution*, *264*, 114549. https://doi.org/10.1016/j.envpol.2020.114549

Chen, M., Wang, P., Chen, Q., Wu, J., & Chen, X. (2015). A Clustering Algorithm for Sample Data Based on Environmental Pollution Characteristics. *Atmospheric Environment*, *107*, 194–203. https://doi.org/10.1016/j.atmosenv.2015.02.042

Cui, M. (2020). *Introduction to the K-Means Clustering Algorithm Based on the Elbow Method*. 5–8. https://doi.org/10.23977/accaf.2020.010102

Dinata, E., & Syaputra, H. (2020). Penerapan Metode Agglomerativ Hirarchical Clusturing Untuk Klasifikasi Dokumen Skripsi. *Bina Darma Conference on Computer Science (BDCCS)*, *2*(2), 412–422.

Drews, S., Savin, I., & van den Bergh, J. C. J. M. (2019). Opinion Clusters in Academic and Public Debates on Growth-vs-Environment. *Ecological Economics*, *157*(October 2018), 141–155. https://doi.org/10.1016/j.ecolecon.2018.11.012

Govender, P., & Sivakumar, V. (2020). Application of K-Means and Hierarchical Clustering Techniques for Analysis of Air Pollution: A Review (1980–2019). In *Atmospheric Pollution Research* (Vol. 11, Issue 1). Turkish National Committee for Air Pollution Research and Control. https://doi.org/10.1016/j.apr.2019.09.009

Hendra Perdana, Nur Asiska, N. S. (2019). Pencarian Cluster Optimum Pada Single Linkage, Complete Linkage dan Average Linkage. *Bimaster : Buletin Ilmiah Matematika, Statistika Dan Terapannya*,

8(3), 393–398.

Hidayat, A. (n.d.). *Penjelasan Lengkap tentang Analisis Klaster*. Statistikian.

Hidayati, R., Zubair, A., Hidayat Pratama, A., & Indana, L. (2021). Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering. *Techno.Com*, *20*(2), 186–197.

Liu, F., & Deng, Y. (2021). Determine the Number of Unknown Targets in Open World Based on Elbow Method. *IEEE Transactions on Fuzzy Systems*, *29*(5), 986–995. https://doi.org/10.1109/TFUZZ.2020.2966182

Mu'afa, S. F., & Ulinnuha, N. (2019). Perbandingan Metode Single Linkage, Complete Linkage Dan Average Linkage dalam Pengelompokan Kecamatan Berdasarkan Variabel Jenis Ternak Kabupaten Sidoarjo. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *4*(2).

Muningsih, E., & Kiswati, S. (2018). Sistem Aplikasi Berbasis Optimasi Metode Elbow Untuk Penentuan Clustering Pelanggan. *Joutica*, *3*(1), 117.

Nabiilah Ardini Fauziyyah, & Sholikhah, I. (2021). *Introduction to Hierarchical Clustering*.

Ogbuabor, G., & F. N, U. (2018). Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value. *International Journal of Computer Science and Information Technology*, *10*(2), 27–37. https://doi.org/10.5121/ijcsit.2018.10203

Ramadhani, L., Purnamasari, I., & Amijaya, F. D. T. (2018). Penerapan Metode Complete Linkage dan Metode Hierarchical Clustering Multiscale Bootstrap (Studi Kasus: Kemiskinan Di Kalimantan Timur Tahun 2016). *Eksponensial*, *9*(2016), 1–10.

Rizal. (2013). *Metode Klasterisasi K-Means*.

Sipayung, A. T., Saifullah, & Winanjaya, R. (2020). Penerapan Metode K-Means Dalam Mengelompokkan Banyaknya Desa/ Kelurahan Menurut Jenis Pencemaran Lingkungan Hidup Berdasarkan Provinsi. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, *1*(1), 104–111.

Statistik, B. P. (n.d.). *Banyaknya Desa/Kelurahan Menurut Jenis Pencemaran Lingkungan Hidup (Desa), 2014-2021*. Badan Pusat Statistik.

Swindiarto, V. T. P., Sarno, R., & Novitasari, D. C. R. (2018). Integration of Fuzzy C-Means Clustering and TOPSIS (FCM-TOPSIS) with Silhouette Analysis for Multi Criteria Parameter Data. *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, ISemantic 2018*, 463–468.

Wibisono, Y., & Khodra, M. L. (2018). *Pengelompokan Artikel Berita Berbahasa Indonesia dengan Agglomerative Clustering*. *2014*, 11–13.

Widyawati, W., Saptomo, W. L. Y., & Utami, Y. R. W. (2020). Penerapan Agglomerative Hierarchical Clustering Untuk Segmentasi Pelanggan. *Jurnal Ilmiah SINUS*, *18*(1), 75.

Wijaya, A., Ar, F., & Rusyana, A. (2021). Perbandingan Metode Gerombol Pautan Lengkap dan Pautan Rataan untuk Pengelompokan Kemiskinan Kabupaten/Kota di Indonesia. *Journal of Data Analysis*, *3*(1), 13–25.