IJSA (2024), Vol. 8, No. 2, 191-201
DOI: https://doi.org/10.29244/ijsa.v8i2p191-201

Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

Comparison of K-Means and K-Medoids in Clustering the Regencies/Cities of West Sumatra Province Using Environmental Indicators

Silfi Robiati[1], Dina Fitria[2‡], Dodi Vionanda[3], and Dwi Sulistiowati[4]

[1,2,3,4]Department of Statistics, Universitas Negeri Padang, Indonesia
‡corresponding author: dinafitria@fmipa.unp.ac.id

Abstract

The Environmental Quality Index is an index that describes the condition of environmental management results nationally, and generalizes from all regions and provinces in Indonesia. Although West Sumatra Province's Environmental Quality Index increased, there are still regions in the province that experienced a decline. Therefore, it is necessary to conduct further analysis, one of which is to form a group of regions into a group according to their similarities or characteristics. This study aims to compare the K-Means and K-Medoids methods in grouping regencies/cities in West Sumatra Province based on environmental quality indicators in 2023. The data used in this research is secondary data, which is orginally the publication of Central Bureau of Statistics namely Sumatera Barat Dalam Angka in 2024. The research compares the K-Means cluster method and the K-Medoids cluster method. It concludes K-Means better than K-Medoids methods based on DB index with three clusters. First cluster has 12 regions with a high average air quality index, the second cluster has 6 regions that have small amounts of waste, and the third cluster has 1 city with a high average water quality index and land quality index, but a large amount of waste.

Keywords: Cluster, Environmental, K-Means, K-Medoids.

* Received: Nov 2024; Reviewed: Dec 2024; Published: Dec 2024

1. Introduction

Indonesian The environment is an entity that encompasses everything around us, such as living things and their behaviors that have an impact on life (Ikhsanudin and Wijayanto, 2024). The Environmental Quality Index (IKLH) is an index that describes the condition of environmental management results nationally, and generalizes from all regencies/cities and provinces in Indonesia. This shows that the national IKLH is derived from the regencies/cities, meaning that if the regencies/cities IKLH increases, the national IKLH also increases.

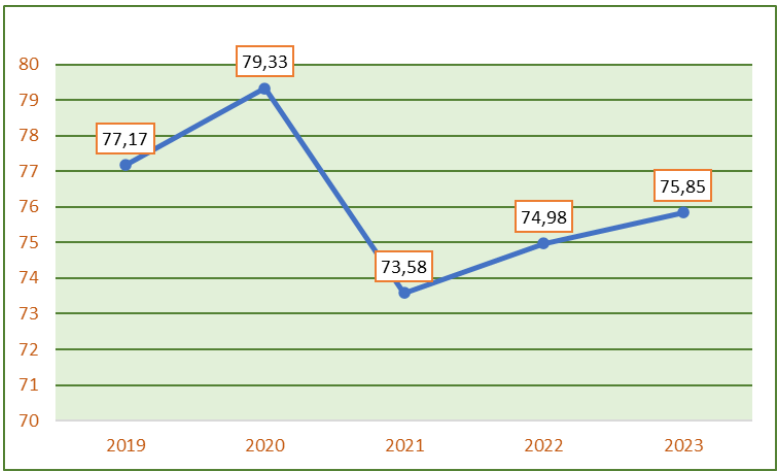


Figure 1: Environmental Quality Index (IKLH) of West Sumatra Province 2019-2023

In Figure 1**.** shows that the IKLH data of West Sumatra Province fluctuates every year, although the IKLH in 2023 has increased compared to last year but there are still regencies/cities in West Sumatra Province that have decreased the environmental quality index. This shows that the quality of the environment in West Sumatra Province has not been evenly distributed and in accordance with the expected target. There are cities in West Sumatra Province that have waste piles of up to 600 ton per day, indicating that waste management has not shown satisfactory results.

The environmental quality index has a relationship with the fulfillment of basic needs, except for the biodiversity index, meaning that if there is an increase in the basic needs of the community, it will result in a decrease in the quality of air, water, and land cover in the area (Suryani, 2018) . In addition, the environmental quality index also shows a relationship with population density, Human Development Index (HDI), and land transport (Hidayati and Zakianis, 2022). This means that the more densely populated an area is, which is in line with the increase in land transport which indirectly increases air pollution, in addition, the environmental quality index is related to the higher the HDI of an area, the environmental quality index also increases. Therefore, it is necessary to group regencies/cities into a group according to their similarities or characteristics, in order to make it easier to determine groups of regencies/cities that have a low environmental quality index and can be used as a reference in equalising

environmental quality in West Sumatra Province. One analysis that can be used for grouping is cluster analysis.

The K-Means method is one of the clustering techniques often used in unsupervised learning. According to (Ikhsanudin and Wijayanto, 2024), research comparing the clustering of provinces in Indonesia based on environmental quality with the Hierarchy and Partition methods shows that K-Means is the best method compared to K-Medoids, Fuzzy C-Means, and Ward's Hierarchy. Another research by (Az-Zahra and Wijayanto, 2024) also found that K-Means provided optimal results in analysing welfare in Indonesia's border areas in 2021. However, (Hidayat, Ghoni and Tyas, 2023) conducted a study showing that K-Medoids is superior in clustering districts/cities based on the Human Development Index in Central Java Province, with a smaller Davies Bouldin Index (DB) value than K-Means. The DB index is used to assess the quality of clustering, where the smaller the value, the better the cluster results. This finding suggests that K-Means is not always the best clustering method. This study aims to first determine what method is better between K-Means and K-Medoids through the DBI value obtained, second to find out how many clusters are formed and which regencies/cities are included in the cluster, and third to find out the characteristics of each cluster.

2. Methodology

2.1 Data Sources and Data Variables

The data used in this research is secondary data, which is orginally the publication of Central Bureau of Statistics namely Sumatera Barat Dalam Angka in 2024 (BPS Provinsi Sumatera Barat, 2024). The method used in this research is cluster analysis by comparing the K-Means method and K-Medoids cluster. Based on the publication book, the Environmental Quality Index (IKLH) data variables can be presented as follows.

Table 1: Research Variables

Variable	Description	Types
IKA	Water Quality Index	Numeric
IKU	Air Quality Index	Numeric
IKL	Land Quality Index	Numeric
Waste	Midden (Ton/day)	Numeric

In Table 1. shows the variables that will be used in the study, the calculation of the Environmental Quality Index is calculated based on 3 indicators with a weight of 30% IKA, 30% IKU, and 40% IKL. According to (Kemlhk, 2010), the Environmental Quality Indicators used to calculate IKLH consist of 3 indicators, namely the Water Quality Index (IKA) which is measured based on the parameters TSS, DO, BOD, COD, Total Phosphate, Fecal Coli, and Total Coliform. The Air Quality Index (AQI) is measured based on SO2 and NO2 parameters, and the Land Cover Quality Index (LQI) is measured based on forest cover area. According to Ikhsanudin & Wijayanto (2024:156),

there are 3 indicators to assess the level of public environmental awareness, namely water consumption, energy consumption, and waste management.

2.2 Data Analysis Technique

The stages of cluster analysis using K-Means and K-Medoids include the following.

1. Detection of outliers using Boxplot

Before continuing the analysis, perform an outlier check to determine if the data contains outliers. This is important because the medoids method is used to place each object into the cluster that has the closest medoids, so it is able to handle data containing outliers, (Fadlurohman and Nur, 2023).

2. Data standardisation

Euclidean distance is one of the most common distance metrics, but it is sensitive to different scales of variables, (Gere, 2023). Therefore, data standardisation is necessary if the variables in the data have significant differences in unit size. So before doing cluster analysis, the first step that must be done is to standardize the data.

$$z = \frac{x_i - \bar{x}}{s} \quad (1)$$

Description:
x_i = observed value
$\bar{x}$ = average value
s = standard deviation

3. Determine the number of k groups

To determine the number of clusters, one of the methods used is the Silhouette Coefficient. The Silhouette coefficient formula used is (Subbaswamy, 1977).

$$S\ (i) = \frac{b_i - a_i}{\max\ (b_i, a_i)} \quad (2)$$

$$a_i = \frac{1}{n(C(i))} \sum_{j \in C(i)} dist\ (i, j) \quad (3)$$

$$b_i = \min_{Ck \in C\ C(i)} \quad \sum_{j \in C_k} \frac{dist(i,j)}{n(C_k)} \quad (4)$$

Description:
a_i = average distance between object i and all objects in the same cluster
b_i= average distance between object i and objects in the nearest cluster
C(i) = Cluster with observation i
dist (i, j) = Euclidean distance between objects i and j
n(C) = Cardinality cluster C

4. K-Means Algorithm

K-Means is one of the algorithms in clustering that is included in Non-Hierarchical, which is a clustering method whose groups have been determined

before cluster analysis. K-Means is one of the clustering methods that uses a simple partitioning based clustering approach, which groups objects into k clusters, (Ikhsanudin and Wijayanto, 2024). The stages of clustering the Regencies/Cities using K-Means analysis are as follows:

a. Randomly determine the initial centroid
b. Calculating the Eucludean distance
c. Calculate the distance matrix D using the Euclidean formula. Euclidean distance can be calculated by the formula:

$$D = d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (5)$$

Description:

d_{ij} = Euclidean distance between object i and object j
x_{ik} = Observed value between object i of variable k
x_{jk} = Observed value between object j of variable k
n = Number of variables

d. Find the average of each cluster (centroid)
e. Calculate the distance between each object and each centroid and put the object into the cluster corresponding to the closest distance.
f. Determine the centroid of the new cluster
g. Repeat steps b through e until no objects are reassigned between clusters

5. K-Medoids Algorithm

K-Medoids is one of the methods known as Partitioning Around Medoids (PAM), a development of K-Means that is sensitive to outliers. This method uses individual objects (medoids) as cluster centres, (Hoerunnisa *et al.*, 2024). The stages of grouping districts using K-Medoids analysis are as follows (Martha, Permana and Fitri, 2024).

a. Determine the cluster center (medoid) as random as the k groups that have been obtained.
b. Calculate the distance matrix D using the Euclidean formula (5).
c. Group objects into medoids based on the smallest distance using distance D.
d. Calculate the distance of each object in each cluster with the new medoids members
e. Calculate the total deviation by calculating the sum value of the new distance - the old distance. If S < 0, then swap objects with cluster data to form new medoids.
f. Repeat steps c-e until there is no change in the members of the medoids.

6. Cluster Method Evaluation

Davies Bouldin Index (DBI) is used to see the number of clusters and the best clustering, the smaller the DBI value, the better the cluster results. The formula used to calculate DBI (Suraya and Wijayanto, 2022) is:

$$\text{DB} = \frac{1}{k}\sum_{p=1}^{k} R_p \quad (6)$$

Description:

DB = davies bouldin index

R_p = cluster similarity measure (maximum)

3. RESULTS AND DISCUSSION

3.1 Checking Outliers

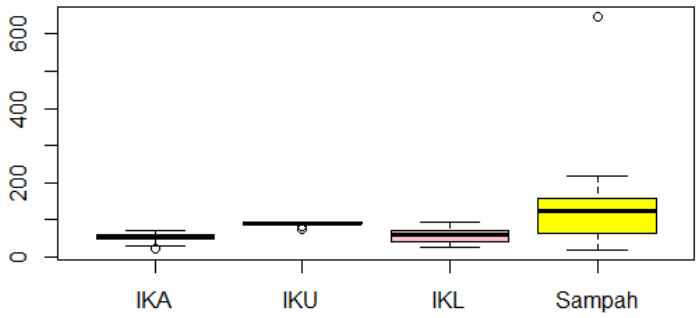


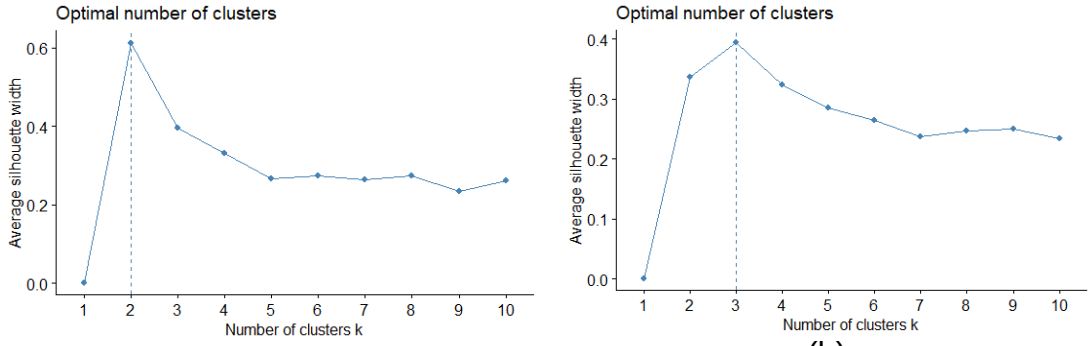
Figure 2: Outlier checking on each variable

Figure 2 shows that checking for outliers using boxplots identified outliers in the water and air quality variables that slightly exceeded the maximum limit, while outliers in the waste variable were very far apart. This indicates that there are Regencies/Cities that have different waste characteristics compared to other Regencies/Cities. Padang City as the capital of the Province has the lowest air quality index and the highest waste generation of 647.39 tonnes per day. This condition is caused by population density, heavy traffic that increases air pollution, and waste management that is not yet optimal.

3.2 Data standardisation

The first step before conducting cluster analysis is to standardise the data using z-score, so that all variables have the same units. Standardised data has a consistent scale and distribution, making it easier to analyse and compare. The results of standardisation show differences in the z-score values of each Regencies/Cities based on the variables IKA, IKU, IKL, and waste generation. Regencies/Cities with positive z-scores on the IKA, IKU, and IKL variables indicate better than average environmental performance, while negative values indicate environmental problems that require more attention. For the waste variable, a positive z-score reflects a high amount of waste that needs to be addressed, while a negative value indicates a more manageable amount of waste.

3.3 Determine k clusters

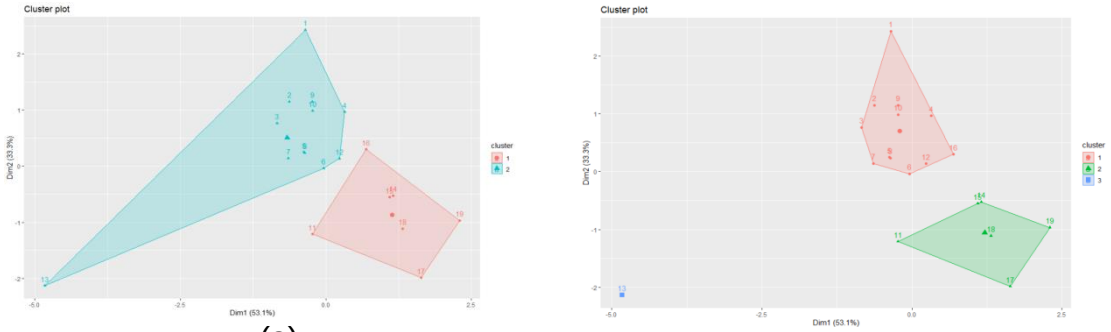


(a) (b)

Figure 3: (a) Silhouette plot of K-Means and (b) Silhouette plot of K-Medoids.

In Figure 3. (a), the silhouette method shows the largest average silhouette width value at k = 2, so the optimum number of clusters for K-Means is two clusters. Figure 3.3.(b) shows the largest average value of silhouette width obtained at k = 3, making the optimum number of clusters for K-Medoids three clusters. Based on the optimum number of clusters from both methods, the next analysis will use two clusters and three clusters for both methods to compare which one is the best.

3.4 K-Means analysis using R studio software.

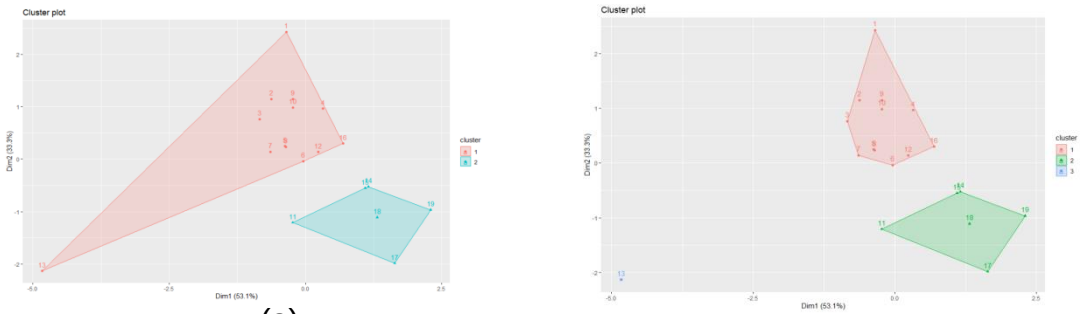


(a) (b)

Figure 4: (a) K-Means analysis results with 2 clusters and (b) K-Means analysis results with 3 clusters

Based on K-Means analysis, clustering of objects is done based on the similarity distance between regencies/cities using the Euclidean distance to the nearest centroid. Clustering with two clusters shows a wider distribution, while clustering with three clusters is better able to capture variations in the distribution of data characteristics.

3.5 K-Medoids analysis using R studio software.



(a) (b)

Figure 5: (a) K-Medoids analysis results with 2 clusters and (b) K-Medoids analysis results with 3 clusters

Based on K-Medoids analysis, clustering of objects is done based on the similarity distance between regencies/cities using the Euclidean distance to the nearest centroid. Clustering with two clusters shows a wider distribution, while clustering with three clusters is better able to capture variations in the distribution of data characteristics.

3.6 Evaluation of the Best Cluster Method

After comparing the K-Means and K-Medoids methods, the next step is to determine the best method between them using the DB Index, as shown in Table 2 below.

Table 2: Validation test of the best method

Klaster	K-Means	K-Medoids
2	1,312364	1,815269
3	0,716584	1,043985

Based on Table 2, the K-Means method shows a lower DBI value in each cluster compared to K-Medoids. In cluster 2, the DBI value of K-Means is 1.312364, while K-Medoids reaches 1.815269. For cluster 3, K-Means is also lower with a value of 0.716584 compared to K-Medoids which is 1.043985. This DBI value indicates that the K-Means method with 3 clusters is more effective in capturing variations in data characteristics, as shown in Figure 4. (b) which shows a smaller spread of data in each cluster. Therefore, the next analysis will use the K-Means method with 3 clusters. Here are the clustering results using K-Means with 3 clusters.

Table 3: Results of K-Means Cluster

Cluster	Regencies/Cities	Total
1	Kepulauan Mentawai, Pesisir Selatan, Solok, Sijunjung, Tanah Datar, Padang Pariaman, Agam, Lima Puluh Kota, Pasaman, Solok Selatan, Pasaman Barat, dan Padang Panjang.	12
2	Padang Dharmasraya, Kota Solok, Sawahlunto, Bukittinggi, Payakumbuh, dan Pariaman.	6
3	Padang	1

Based on Table 3. for the first cluster there are 12 regencies/cities, this shows that the first cluster is dominated by regencies located in West Sumatra Province. Then the second cluster has 6 regencies/cities which are dominated by cities in West Sumatra Province, and the third cluster has one city, Padang City, which is the capital of West Sumatra Province.

Table 4: Average of Each IKLH Indicator by Cluster

IKA	**IKU**	**IKL**	**Waste**	***Cluster***
59,21	92,41	64,46	133,19	1
41,51	89,51	36,1	73,35	2
69,77	74,11	73,2	647,39	3

In Table 4, after cluster analysis of regencies/cities in West Sumatra Province using K-Means with an optimum number of clusters of two clusters, it is found that the first cluster consists of 12 regencies/cities with a high average air quality index. This can be seen from the number of regencies in West Sumatra Province that are far from urban areas that are included in the first cluster. The second cluster has 6 regencies/cities with a low amount of waste, this can be seen from the cities in West Sumatra Province that are included in the second cluster and have shown effective waste management. The third cluster consists of 1 city, Padang City, which has a high average water quality index and land quality index, but also has a large amount of waste, indicating that Padang City is good at managing water and land quality, but not good enough at handling waste management issues.

4. Conclusions

After the analysis, the cluster evaluation results show that the K-Means method with three clusters is the best cluster method in grouping regenciec/cities in West Sumatra Province based on the environmental quality index because it has the smallest DB index when compared to the K-Medoids method, while the K-Means cluster results with three clusters. The first cluster are regencies/cities that have a high Air Quality Index, but have low Water Quality and Land Quality as well as large waste piles, namely Kepulauan Mentawai Regency, Pesisir Selatan Regency, Solok Regency, Sijunjung Regency, Tanah Datar Regency, Padang Pariaman Regency, Agam Regency, Lima Puluh Kota Regency, Pasaman Regency, Solok Selatan Regency, Pasaman Barat Regency, dan Padang Panjang City. The second cluster are regencies/cities that have a high Air Quality Index, but have low Water Quality and Land Quality as well as large waste piles, namely Kepulauan Mentawai Regency, Pesisir Selatan Regency, Solok Regency, Sijunjung Regency, Tanah Datar Regency, Padang Pariaman Regency, Agam Regency, Lima Puluh Kota Regency, Pasaman Regency, Solok Selatan Regency, Pasaman Barat Regency, dan Padang Panjang City. The third cluster is a city that has high water quality and land quality, but has a large amount of waste, namely Padang City.

References

Az-Zahra, A. and Wijayanto, A.W. (2024) ‘Tinjauan Kesejahteraan di Daerah Perbatasan Republik Indonesia Tahun 2021: Penerapan Analisis Klaster K-Means

dan Hierarki', *Jurnal Sistem dan Teknologi Informasi*, 12(1), pp. 55–64. Available at: https://doi.org/10.26418/justin.v12i1.69040.

BPS Provinsi Sumatera Barat (2024) 'Provinsi Sumatera Barat Dalam Angka', 54, p. 929.

Fadlurohman, A. and Nur, I.M. (2023) 'Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Perumahan dan Kesehatan Lingkungan Menggunakan Metode K-Medoids', *Prosiding Seminar Nasional UNIMUS*, 6, pp. 1168–1180.

Gere, A. (2023) 'Recommendations for validating hierarchical clustering in consumer sensory projects', *Current Research in Food Science*, 6(February), p. 100522. Available at: https://doi.org/10.1016/j.crfs.2023.100522.

Hidayat, N.W., Ghoni, U. and Tyas, F.A. (2023) 'Perbandingan Algoritma K-Means dan K-Medoids dalam Pengelompokkan Kabupaten/Kota Berdasarkan Indeks Pembangunan Manusia di Provinsi Jawa Tengah', *Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media*, 3, pp. 663–672.

Hidayati, Z.A. and Zakianis (2022) 'Analisis Faktor-Faktor Yang Mempengaruhi Indeks Kualitas Lingkungan Hidup (IKLH) Di Indonesia Tahun 2017-2019', *Jurnal Medika Utama*, 3(2), p. 2329. Available at: http://jurnalmedikahutama.com/index.php/JMH/article/view/456.

Hoerunnisa, A. *et al.* (2024) 'Komparasi Algoritma K-Means Dan K-Medoids Dalam Analisis Pengelompokan Daerah Rawan Kriminalitas Di Indonesia', *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(1), pp. 103–110. Available at: https://doi.org/10.36040/jati.v8i1.8249.

Ikhsanudin, M.R. and Wijayanto, A.W. (2024) 'Perbandingan Pengelompokkan Provinsi di Indonesia Menurut Kualitas Lingkungan Hidup Menggunakan Metode Hierarki dan Partisi Comparing Province Clustering in Indonesia Based on Environmental Quality Using Hierarchical and Partition Methods', *Jurnal Sistem dan Teknologi Informasi*, 12(1), pp. 155–163. Available at: https://doi.org/10.26418/justin.v12i1.71495.

Kemlhk (2010) *Indeks Kualitas Lingkungan Hidup 2009*, *Jakarta: Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia*.

Martha, Z., Permana, D. and Fitri, F. (2024) 'K-Medoids Cluster Analysis for Grouping Provinces in Indonesia Based on Agricultural Households ST2023', 2, pp. 324–329.

Subbaswamy, K.R. (1977) 'Brillouin Scattering From Thermal Fluctuations in Superionic Conductors.', *Solid State Communications*, 21(4), pp. 371–372. Available at: https://doi.org/10.1016/0038-1098(77)91248-0.

Suraya, G.R. and Wijayanto, A.W. (2022) 'Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting', *Indonesian Journal of Statistics and Its Applications*, 6(2), pp. 180–201. Available at: https://doi.org/10.29244/ijsa.v6i2p180-201.

Suryani, A.S. (2018) ‘Pengaruh Kualitas Lingkungan Terhadap Pemenuhan Kebutuhan Dasar di Provinsi Banten’, *Jurnal Aspirasi*, 9(1), pp. 35–63. Available at: https://doi.org/10.22212/aspirasi.v9i1.991.