Jurnal Presipitasi
Media Komunikasi dan Pengembangan Teknik Lingkungan
e-ISSN: 2550-0023

Vol 22, No 3, 2025,
Article available at homepage presipitasi

Regional Case Study

Comparing K-Means and K-Medoids for Industrial Air Pollution Analysis in Central Java

Rani Rachma Astining Putri[1], Roifah Fajri[1], Sapta Suhardono[1*], Callista Fabiola Candraningtyas[1], Iva Yenis Septiariva[2]

[1] Department of Environmental Science, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Surakarta, 57126, Indonesia

[2] Graduate Institute of Environmental Engineering, National Central University, Taoyuan City, 32001, Taiwan

* Corresponding Author, email: sapta.suhardono@staff.uns.ac.id

Abstract

Air is a fundamental necessity for all living beings, especially humans. However, human activities whether intentional or unintentional can degrade air quality through pollution. This study compares the performance of the K-Means and K-Medoids clustering algorithms in analyzing the air pollution load from the industrial sector in Central Java in 2021. Using a quantitative approach and R Studio software, the analysis focuses on SO_2 and NO_2 pollution data obtained from the official Central Java BPS website. The results indicate that the K-Medoids algorithm with the silhouette method yields the most optimal clustering performance, with the lowest Davies-Bouldin Index (DBI) value of 0.6201437 and 10 distinct clusters. Notably, Cluster 1 comprises districts with the highest industrial air pollution burden such as Banjarnegara Regency, which recorded 14,472 industries and NO_2 and SO_2 concentrations of 20 µg/m³ and 6 µg/m³, respectively. These findings demonstrate that clustering algorithms not only help reveal spatial pollution patterns but also provide critical insights for prioritizing targeted mitigation efforts and informing environmental policy-making in industrially active regions.

Keywords: Air polution; clustering; data mining; k-means; k-medoids

1. Introduction

Air is a vital necessity for living beings, especially humans, to sustain life. However, various human activities, whether intentional or unintentional, can lead to air pollution (Tiara & Firmawati, 2023). According to Maharani and Aryanta (2023), air pollution occurs when other components enter the air, either directly or indirectly due to human activities. This leads to a decline in air quality to levels that can damage the environment or render it non-functional. Most air pollution is caused by two main sources: mobile sources and stationary sources. These sources include the industrial, domestic, and transportation sectors. High rates of urbanization, population growth, imbalanced spatial development, and low public awareness of air pollution are other factors that indirectly contribute to air pollution (Jannahdita & Cahyonugroho, 2023).

Research conducted in several major cities in Indonesia indicates that the main sources of air pollution are transportation, industry, residential areas, and waste management. The transportation sector, particularly motor vehicles, is the largest contributor to air pollution, accounting for 60% of

pollutant gases and particulates. Meanwhile, the industrial sector contributes 25%, households 10%, and waste 5% (Ridwan et al., 2020). According to Abidin et al. (2019), there are three categories of air pollution sources. The first source originates from industries and urban areas, resulting from technological advancements such as businesses, factories, power plants, and the increasing number of motor vehicles. Research by Romadhon and Mokhtar (2021) indicates that air pollution from industrial activities contributes to emissions of hazardous gases such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), hydrocarbons (HC), carbon monoxide (CO), and dust particles.

In Indonesia, the manufacturing industry is a rapidly growing sector (Harahap et al., 2023). According to data from the Central Bureau of Statistics (BPS) in 2023, the Central Java Province had 204,034 industries. Central Java is known as one of the regions with strong performance in the manufacturing industry, making it the third-largest contributor after West Java and East Java (Soca & Woyanti, 2021). This is supported by the abundant workforce availability in the province (Yulianti et al., 2021). The increasing number of industries in Central Java has led to a rising burden of air pollution (Prayogo et al., 2021). To address the issue of air pollution, careful monitoring, and efficient data management are crucial. Data on the burden of SO_2 and NO_2 pollution from the industrial sector in Central Java in 2021 serves as a critical foundation for developing effective mitigation strategies. One approach to reducing air pollution from industries is to conduct regular monitoring, for instance, every 3-6 months. Additionally, strict law enforcement efforts are also needed (Sugiarto et al., 2024).

In air pollution burden data analysis, data mining techniques are used to uncover new insights or knowledge. Data mining is a relatively new field to explore (Gupta & Chandra, 2020). The results of data mining can be utilized for future decision-making. One widely recognized technique in data mining is clustering (Handoko et al., 2020). Clustering is a method used to find and group data with similar characteristics. Clustering has two main methods: hierarchical clustering and non-hierarchical clustering (Syafrinal & Febrianti, 2023). Algorithms commonly used in clustering analysis include K-Means and K-Medoids (Herman et al., 2022). K-Means clustering is a non-hierarchical method that groups data into one or more clusters (Herviany et al., 2021). K-Medoids is a classical partitioning technique that groups data based on object information indices n into interpretable groups (Adek et al., 2022). K-Medoids use objects as cluster centers, while K-Means use mean values as cluster centers (Meiriza et al., 2023). Both algorithms aim to partition data into several clusters based on shared characteristics (Khan et al., 2023). According to Han et al. (2022), there are differences between these two algorithms. K-Means is advantageous in terms of faster time complexity for handling similar data compared to K-Medoids. Meanwhile, K-Medoids are more robust to outliers. However, each algorithm has its weaknesses, as K-Means cannot handle outliers, and K-Medoids require longer computation time. Comparing these two algorithms can assist in selecting the better algorithm for specific cases.

In this study, the use of clustering methods is expected to help identify patterns in SO_2 and NO_2 pollution burden data from the industrial sector in Central Java. The study's objective is to compare the effectiveness of clustering methods between K-Means and K-Medoids in analyzing SO_2 and NO_2 pollution burden data from the industrial sector in Central Java in 2021. The results of this research are expected to provide deeper insights into air pollution patterns and enable the development of more effective and targeted environmental management strategies.

2. Methods

This study employs a quantitative design using a comparative method between K-Means and K-Medoids. Data analysis was conducted using R Studio software. The primary data used were sourced from the official website of the Central Java Bureau of Statistics (BPS), covering the air pollution burden of SO_2 and NO_2 from the industrial sector in Central Java in 2021 (Table 1). The study subjects are pollution burdens, while the variables are SO_2 and NO_2. The data analysis technique utilizes R Studio software to compare the K-Means and K-Medoids algorithms. Based on the data in (Table 1), the data mining process can be carried out using the clustering method. Clustering is a technique used to group data into several

groups based on certain similarities or characteristics. In clustering algorithms, two methods are commonly used, namely K-Means and K-Medoids.

Table 1. Data on sulfur dioxide and nitrogen dioxide air pollution loads in the industrial sector in 2021 in central java

No	Regency / City	Industry Sector	
		NO_2	SO_2
1.	Banjarnegara Regency	20	6
2.	Banyumas Regency	13	11
3.	Batang Regency	13	7
4.	Blora Regency	10	4
5.	Boyolali Regency	19	4
6.	Brebes Regency	13	12
7.	Cilacap Regency	5	3
8.	Demak Regency	32	10
9.	Grobogan Regency	5	10
10.	Jepara Regency	17	13
11.	Karanganyar Regency	13	16
12.	Kebumen Regency	11	3
13.	Kendal Regency	14	11
14.	Klaten Regency	15	5
15.	Kudus Regency	20	8
16.	Magelang Regency	7	8
17.	Pati Regency	8	19
18.	Pekalongan Regency	10	7
19.	Pemalang Regency	12	3
20.	Purbalingga Regency	3	13
21.	Purworejo Regency	11	10
22.	Rembang Regency	10	6
23.	Semarang Regency	36	7
24.	Sragen Regency	14	12
25.	Sukoharjo Regency	7	3
26.	Tegal Regency	12	17
27.	Temanggung Regency	17	13
28.	Wonogiri Regency	13	13
29.	Wonosobo Regency	8	12
30.	Magelang City	7	14
31.	Pekalongan City	18	8
32.	Salatiga City	11	8
33.	Semarang City	26	11
34	Surakarta City	15	6
35.	Tegal City	13	11

Source: Central Java BPS Data on SO2 and NO2 Pollutant Loads in the Industrial Sector, 2019

The first step in completing the data is to collect data regarding the pollutant load of the industrial sector in Central Java Province for 2021 from the official website of the Central Statistics Agency (BPS). This data will be the basis for further analysis and processing. After the data is collected, a preprocessing process is carried out, which includes data cleaning and data selection. Data cleaning aims to eliminate irrelevant data or fill in missing values, while data selection focuses on selecting important

variables that will be used in the analysis. The next step is data processing using the K-Means and K-Medoids algorithms, which begins with a data normalization process. Normalization is carried out to ensure the distribution of values is more even or maintains a consistent scale between variables. This aims to ensure that each feature in the dataset has a comparable range of values so that the analysis results are more accurate.

To determine the optimal number of clusters, two main methods are used: the elbow method and the silhouette method. The elbow method uses the concept of Within Sum of Squares (WSS), which is a comparison of the percentage of the number of clusters with decreasing values that form an elbow pattern on the graph. The exact number of clusters is determined at the point where the graph shows the most significant decrease. On the other hand, silhouette methods evaluate the quality and strength of clusters by measuring how well the objects in the dataset are grouped into appropriate clusters. After the number of clusters is determined, the clustering process is carried out using the K-Medoids method with the Partitioning Around Medoids (PAM) algorithm on normalized data. The clustering results are then visualized in two-dimensional (2D) space to provide a clearer picture of the cluster division.

The next step is to evaluate the clustering results using the Davies-Bouldin Index (DBI). DBI is used to assess cluster quality based on two main aspects: cohesion and separation. Cohesion measures the level of similarity of data in a cluster, while separation measures the distance between cluster centers. The optimal cluster is one that has high separation values and low cohesion. Evaluation using DBI aims to determine the optimal number of clusters and assess how good the separation between clusters is and the homogeneity within clusters. Finally, the results of model evaluation using DBI provide an overview of the quality of the object variables in each cluster. With this approach, the clustering process can be assessed comprehensively, both in terms of grouping accuracy and optimizing the number of clusters.

For the clustering process, two algorithms were employed: K-Means and K-Medoids (PAM algorithm). To determine the optimal number of clusters, two techniques were utilized. The Elbow Method was applied to analyze the Within-Cluster Sum of Squares (WSS) across a range of cluster numbers, aiming to identify the 'elbow' point where additional clusters yield diminishing returns in variance reduction. Meanwhile, the Silhouette Method was used to evaluate the consistency within clusters by measuring how similar an object is to its own cluster compared to others, with higher silhouette scores indicating better-defined clusters. Based on the silhouette analysis, the optimal number of clusters was determined to be 10 for the K-Medoids algorithm. Subsequently, both clustering algorithms were executed, and parameter tuning was conducted by varying the number of clusters (k) from 2 to 15. The clustering results were then assessed using the Davies-Bouldin Index (DBI), which measures clustering quality by balancing intra-cluster cohesion and inter-cluster separation. Lower DBI values signify superior clustering performance.

The research employed a structured sequence of methodological stages. The process commenced with the initiation phase, followed by systematic data collection related to the burden of industrial air pollution in Central Java. The collected data were subsequently processed to ensure completeness, consistency, and readiness for analytical procedures. After preprocessing, the study applied two clustering algorithms K-means and K-medoids to classify the data based on specific characteristics relevant to industrial pollution patterns. Both methods were implemented to evaluate and compare their performance in grouping the datasets. The quality of the clustering outputs was then assessed through a validity testing stage, which examined the robustness and reliability of each algorithm's results. Upon completion of the validation process, the findings were analyzed to interpret the clustering outcomes and to identify the implications of each method in representing industrial air pollution characteristics. Finally, the research concluded with a synthesis of the key results and insights derived from the analytical stages.

3. Result and Discussion

3.1. K-Means Clustering

K-Means Clustering is carried out using two approaches to determine the optimal number of clusters, namely the elbow method and the silhouette method.

3.1.1. K-Means Elbow Method

Figure 2 shows the results of analysis using the Elbow method to determine the optimal number of clusters in the K-Means algorithm. The Elbow method works by calculating the Within-Cluster Sum of Squares (WSS) value for various numbers of clusters. WSS is a measure of variation within each cluster which shows how close the data in a cluster is to the cluster center (centroid). The smaller the WSS value, the better the cluster is at explaining variations in the data within it.

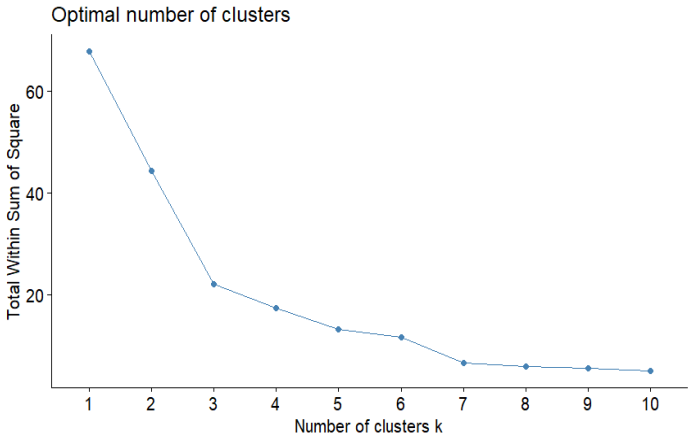

Figure 2. K-Means elbow method results

In the graph of Figure 2, WSS decreases sharply when the number of clusters increases from 1 to 3, indicating that increasing the number of clusters significantly improves the quality of data grouping. After the point k=3, the decrease in WSS becomes more stable, which indicates that the subsequent increase in the number of clusters does not provide a significant increase. This point is called the "elbow point", indicating that k=3 is the optimal number of clusters. With three clusters, data can be grouped with an optimal balance between variation within groups (intra-cluster) and variation between groups (inter-cluster). Research by Prasetyo et al. (2024) show that this method is suitable for data with clear distribution patterns, such as those in this analysis.

Figure 3 depicts the visualization results of the clustering process using the K-Means algorithm with the optimal number of clusters k=3, which are visualized in a two-dimensional plot (2D cluster plot). In this visualization, each cluster is represented with a different color and clear regional boundaries, which helps in understanding the distribution of data within each group (cluster). Each data point in the graph represents a single entity, such as an industrial pollution source, grouped based on similar characteristics.

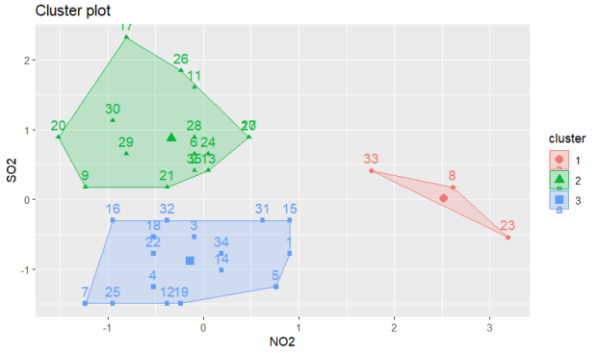

Figure 3. 2D K-means cluster results elbow method

Regional boundaries on the plot indicate areas that approach the respective cluster centers (centroids), where the centroids are represented by certain symbols (often large dots or crosses). Points that are near the centroid show data that is more consistent or homogeneous within the cluster, while points that are further from the centroid may show data that is more varied but still in the same group. This visualization provides insight into the relationships between data in each cluster. By comparing between clusters, different patterns can be seen, indicating that data in different groups have characteristics that vary more from each other (Prasetiyo et al., 2024). This is in accordance with the aim of clustering, namely to maximize similarities within groups (intra-cluster similarity) and minimize similarities between groups (inter-cluster dissimilarity).

3.1.2. K-Means Silhouette Method

Figure 4 shows the results of determining the optimal number of clusters using the Silhouette method in the K-Means algorithm. The Silhouette method is used to evaluate the quality of grouping by calculating the average value of the silhouette index, which shows the extent to which objects in one cluster are more similar to objects in their own cluster compared to objects in other clusters (Nugraha et al, 2024). On the graph, the silhouette value is in the range -1 to 1, where a value close to 1 indicates that the data is well grouped, while a value close to -1 indicates a grouping error.

Based on the graph in Figure 4, the optimal number of clusters was found to be seven (k=7), because at this point the average silhouette value reaches its peak. This shows that dividing the data into seven groups provides the best clustering results, with a clear cluster structure and significant differences between groups. These findings are consistent with the results identified in the study of Sowan et al. (2023), who stated that the silhouette method is an effective tool for evaluating the optimal number of clusters, especially on datasets with complex distributions.

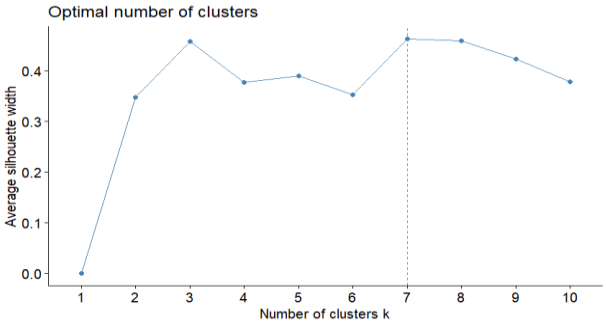

Figure 4. K-Means silhouette method results

Figure 5 presents the results of clustering visualization using the K-Means algorithm with k=7, visualized in a 2D cluster plot. In this graph, each group (cluster) is represented with a different color and has clear regional boundaries. The cluster center (centroid) is indicated by a certain symbol, which functions as a representation of the average position of data within each cluster.

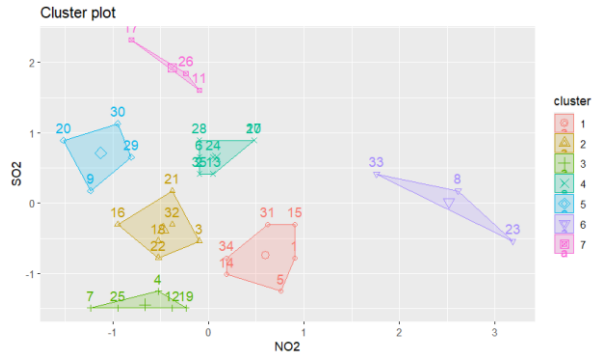

Figure 5. 2D K-Means cluster results silhouette method

This visualization shows specific data distribution patterns in each cluster, where data in one group has more similar characteristics compared to data in other groups. For example, in industrial pollution analysis, each cluster may represent a group of industries with certain levels of pollution or characteristics, such as fuel type, emission output, or geographic location. Jentner & Keim's research (2021), supports this approach, where clustering visualization helps provide an intuitive understanding of data patterns that are not visible in higher dimensions.

3.1.3. K-Medoids Elbow Method

Figure 6 shows the results of analysis using the elbow method, which is used to determine the optimal number of clusters based on the Within-Cluster Sum of Squares (WSS) value. The elbow curve describes the relationship between the number of clusters and WSS, where WSS measures the total variation in each cluster (Utari, 2021). In this graph, you can see a sharp decrease in WSS as the number of clusters increases, but this decrease begins to slow down after the number of clusters is three (k=3). The point at which the decline slows is what is known as the "elbow point," indicating the optimal number of clusters. Thus, in this study, the optimal number of clusters was three, because adding more clusters after this point did not provide a significant reduction in WSS.

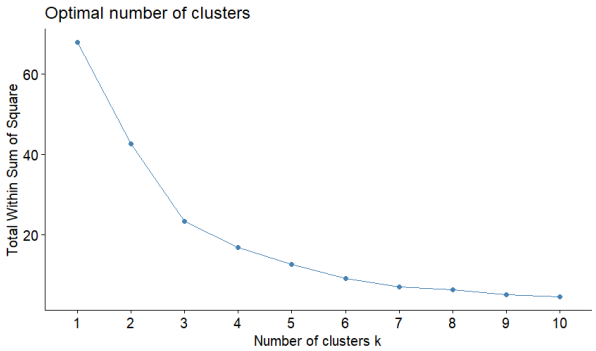

Figure 6. Results of k-medoids elbow method

Determining the optimal number of clusters using the elbow method is an important step to avoid under-clustering or over-clustering (Dang et al., 2024). If the number of clusters is too small, then the data variation within each cluster becomes too large, which causes the loss of important information. Conversely, if the number of clusters is too large, the risk of overfitting increases, where the data becomes too segmented and therefore difficult to interpret meaningfully. The study by Seikholeslami et al. (2019), supports the importance of the elbow method, which is said to be a simple but powerful approach for analyzing complex environmental data, such as industrial air pollution data.

Figure 7 shows a visualization of clustering results using the K-Medoids algorithm on data that has been grouped into three clusters. Each cluster has a clear distribution, showing that the K-Medoids algorithm is successful in separating data based on the characteristics of industrial air pollution in Central Java. According to the study by Wibawa et al. (2021), K-Medoids has advantages in minimizing the influence of outliers compared to K-Means, making it more suitable for complex datasets such as air pollution.

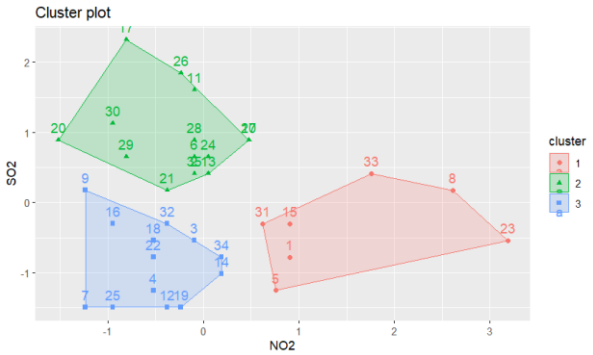

Figure 7. 2D K-Medoids elbow method cluster results

Determining optimal clusters and visualizing clustering results using the K-Medoids algorithm provides deep insight in analyzing industrial air pollution loads. This is in line with the study of Madbouly et al. (2022), who found that the K-Medoids-based clustering approach was superior in environmental data analysis due to its ability to handle uneven data distribution.

3.1.4. K-Medoids Silhouette Method

Figure 8 displays the results of analysis using the silhouette method, which was used to determine the optimal number of clusters in this research. The silhouette method evaluates the quality of grouping by comparing the average distance between one data and other data in its own cluster (cohesion) and the average distance to data in the nearest cluster (separation) (Nugraha et al., 2024). The greater the average silhouette width value, the better the clustering quality because it shows that the data within the cluster is more homogeneous while the distance between clusters is greater. Based on Figure 8, the highest average silhouette width value is found in the number of clusters of 10 (k=10), so this number is considered the optimal configuration for grouping data in this study.

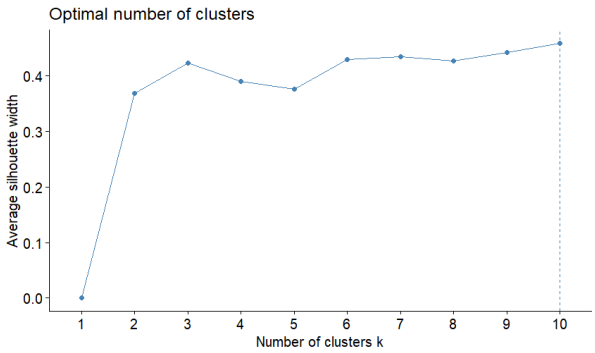

Figure 8. Results of K-Medoids silhouette method

The optimal value k=10 indicates that industrial air pollution data in Central Java has high complexity, with significant differences between clusters. Grouping into 10 clusters allows this research to capture more detailed data variations compared to a smaller number of clusters, thereby providing deeper insight into the distribution of air pollution. The study of Huang et al. (2021) also show that more detailed clustering configurations often provide greater benefits in location-based analysis, especially in grouping diverse environmental data.

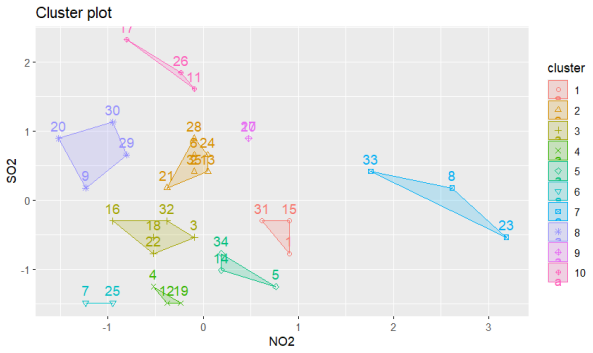

Figure 9. Results of K-Medoids silhouette method

In Figure 9, visualization of clustering results with the K-Medoids algorithm for 10 clusters is shown. The diagram shows the distribution of data in two-dimensional space, with each cluster represented in a different color. The K-Medoids algorithm was chosen because of its ability to minimize the influence of outliers and provide more stable results than the K-Means algorithm, as supported by research by Wibawa et al. (2021). The clear distribution in Figure 9 shows that the K-Medoids algorithm is successful in separating data based on the characteristics of industrial air pollution in Central Java. Each cluster can represent a different level of emissions, which is very relevant for understanding air pollution patterns in different industrial regions.

The interpretability of the clusters suggests that regions with higher pollution levels are typically characterized by dense population, rapid industrialization, and limited environmental regulation enforcement. Socio-economic factors such as low income levels, poor infrastructure, and unplanned urban expansion may further exacerbate pollution levels in these areas. Understanding these underlying factors provides deeper insight into the spatial distribution of pollution and supports the formulation of targeted interventions.

3.2. Comparison of DBI Values in K-Means and K-Medoids Clustering

The Davies-Bouldin Index (DBI) is an index used to determine the most optimal number of clusters. To find out the best clustering in DBI, you need to know the smallest DBI value from the available options or a DBI value that is closer to 0 indicates a better cluster quality. The lower the DBI value, the more optimal the clustering results obtained (Riani et al., 2023).

Table 2. Comparison results of the k-means and k-medoids methods

Methods	K-Means		K-Medoids	
	Cluster	DBI Value	Cluster	DBI Value
Metode Elbow	k = 3	0.7847946	k = 3	0.838663
Metode Silhouette	k = 7	0.6708402	**k = 10**	**0.6201437**

Based on the results of Table 2, it shows that the smallest value of the two algorithms with two different methods, namely the K-Medoids algorithm with the silhouette method with a DBI value of 0.6201437. This shows that the best method for clustering air pollutant loads with SO_2 and NO_2 variables is the K-Medoids silhouette clustering method. Selection of the optimal number of clusters based on the silhouette method provides advantages over other approaches, such as the elbow method, because it not only considers variations within clusters but also inter-cluster relationships. This is in accordance with findings by Herman et al. (2022), who emphasize that the silhouette method is very reliable in validating clustering results on large and heterogeneous datasets. In the context of this study, datasets covering various air pollution parameters, such as gas emissions, geographic location, and industrial production capacity, show high heterogeneity. Therefore, the choice of this evaluation method is very relevant to ensure the quality of the grouping results.

3.3. Clustering of Air Pollution Loads in the Industrial Sector of Central Java Province

In the analysis listed in Table 3, it can be seen that cluster 1 accommodates the districts that experience the highest levels of air pollution from the industrial sector. For example, Banjarnegara recorded 14,472 industries, with pollutant levels of NO2 of 20 µg/m3) and SO2 of 6 (µg/m3). Likewise, Kudus has 3,339 industries, with NO2 pollutant levels of 20 (µg/m3) and SO2 of 8 (µg/m3), and Pekalongan with 15,237 industries, and NO2 pollutant levels of 10 (µg/m3) and SO2 of 7 (µg/m3) (BPS, 2023). This fact confirms that increasing the number of industries can directly increase the level of air pollution (Antarasari, 2019). Therefore, the importance of effective industrial waste management is a necessity in order to minimize its negative impact on the ecosystem.

Table 3. K-medoids clustering results using the silhouette method

Cluster	Regency / City
1	Banjarnegara, Kudus, and Pekalongan
2	Banyumas, Brebes, Kendal, Purworejo, Sragen, Wonogiri, and Tegal
3	Batang, Magelang, Pekalongan, Rembang, and Salatiga
4	Blora, Kebumen, and Pemalang
5	Boyolali, Klaten, and Surakarta
6	Cilacap and Sukoharjo
7	Demak, Semarang, and Semarang
8	Grobogan, Purbalingga, Wonosobo, and Magelang
9	Jepara and Temanggung
10	Pati and Tegal

An important step that can be taken is to reduce industrial waste emissions. These programs are often initiated by the government or environmental agencies with the main aim of reducing the impact of industrial pollution on ecosystems and human health (Rahmansyah et al., 2024). Findings by Sulasminingsih et al. (2024) show that the application of industrial exhaust gas emission control technology can significantly reduce air pollution levels and improve air quality around industrial sites. On the other hand, cluster 10 shows the lowest level of air pollution from the industrial sector, especially seen in Pati and Tegal with each number of industries amounting to 2,976 with pollutant levels of NO2 reaching 8 (µg/m3) and SO2 of 19 (µg/m3) and 1,790 with NO2 pollutant levels reaching 12 (µg/m3) and SO2 of 17 (µg/m3). This indicates that the two districts are likely to have implemented pollutant load control technology.

These findings are consistent with prior studies that identified urban and industrial areas as major contributors to environmental pollution (Ridwan et al., 2020; Romadhon & Mokhtar, 2021). However, our results reveal distinct regional variations that were not captured in earlier works, emphasizing the importance of spatial clustering. This divergence may be attributed to differences in data resolution, methodology, and recent socio-economic developments, underscoring the added value of our clustering approach in capturing nuanced patterns (Herman et al., 2022; Huang et al., 2021).

Based on the clustering results, region-specific mitigation strategies can be formulated to address the varying degrees of industrial air pollution across Central Java. For high-risk clusters—such as regions with dense industrial activities and high concentrations of SO_2 and NO_2—policy recommendations include the enforcement of stricter emission standards through regular inspections, mandatory pollution control technologies (e.g., scrubbers, electrostatic precipitators), and the integration of green industrial zones. These areas should also prioritize the development of green infrastructure, such as urban forests and vegetative buffers, to help absorb airborne pollutants. Furthermore, public-private partnerships could be initiated to promote cleaner production techniques, supported by targeted subsidies or tax incentives.

For moderate-risk clusters, installing continuous air quality monitoring systems and establishing early warning mechanisms are essential to ensure timely response to pollution spikes. These areas would also benefit from implementing incentive-based policies that encourage industries to adopt energy-efficient and low-emission technologies, including renewable energy sources. Education campaigns and capacity-building programs should also be tailored to local communities and industrial stakeholders to raise awareness and promote behavioral change.

In contrast, low-risk clusters can focus on preventive strategies by maintaining current emission levels through zoning regulations that limit new industrial developments in environmentally sensitive areas. Additionally, fostering inter-regional collaboration can facilitate the exchange of best practices and ensure that mitigation strategies are both scalable and contextually relevant. Overall, these differentiated policy recommendations not only align with the cluster-specific pollution profiles but also enable efficient resource allocation, ensuring that environmental management efforts are both equitable and impactful.

4. Conclusions

In this research, a comparison was carried out between two clustering methods, namely K-Means and K-Medoids. The research results show that the K-Medoids method with a k value of 10 is effective in clustering air pollutant loads in the industrial sector. The clustering results were evaluated and analyzed using the Davies-Bouldin Index (DBI), and a value of 0.6201437 was obtained, which is close to 0, indicating good clustering results. This analysis reveals that the district/in Central Java that experiences the highest level of pollution from the industrial sector is Banjarnegara with a NO_2 pollution level of 20 µg/m3 and SO_2 of 6 µg/m3 Kudus with NO_2 of 20 µg/m3 and SO_2 of 8 µg/m3 , and Pekalongan with NO_2 of 10 µg/m3 and SO_2 of 7 µg/m3. Therefore, efforts are needed to implement industrial exhaust gas emission control technology to significantly reduce the level of air pollution and improve air quality around industrial sites. Thus, the clustering method is not only useful for analyzing pollution data but also for formulating effective mitigation strategies.

References

Abidin, J, Hasibuan, F, A, and Kunci, K. 2019. Pengaruh dampak pencemaran udara terhadap kesehatan untuk menambah pemahaman masyarakat awam tentang bahaya dari polusi udara. Prosiding SNFUR-4 2(2), 978-979.

Adek, R, T, Aidilof, H, A, K, Mukhlis, M, and Nur, K. 2022. Penerapan algoritma K-medoid dalam perbandingan daya serap akademik siswa sekolah peran dan sekolah pedesaan selama masa pandemi. Jurnal Tekno Kompak 16(2), 85-97.

Antasari, D, W. 2020. Implementasi green economy terhadap pembangunan berkelanjutan di Kediri. Jurnal Ekonomi Pembangunan STIE Muhammadiyah Palopo 5(2), 80-88.

Azzahrah, F, Annas, S, and Rais, Z. 2022. Hybrid hierarchical clustering dalam pengelompokan daerah rawan bencana tanah longsor di Sulawesi Selatan. VARIANSI: Journal of Statistics and Its application on Teaching and Research 4(3), 153-161.

Badan Pusat Statistik. 2023. Jumlah industri dan tenaga kerja menurut kabupaten/ di provinsi jawa tengah 2019 - 2022.

Cahyonugroho, O, H, and Jannahdita, D, U. 2023. Analisis pengaruh beban pencemar terhadap kualitas udara ambien dari kegiatan pengembangan universitas X di Surabaya. Prosiding Esec 4(1), 340-345.

Dang, J, Xiao, D, Zhang, X, Jia, R, and Jiao, Y. 2024. Optimal configuration of retired battery reconfigurable network considering switching losses. Journal of Energy Storage 101, 113735.

Dewi, D, A, I, C, and Pramita, D, A, K. 2019. Analisis perbandingan metode elbow dan silhouette pada algoritma clustering K-medoids dalam pengelompokan produksi kerajinan bali. Matrix: Jurnal Manajemen Teknologi dan Informatika 9(3), 102-109.

Gupta, M, K, and Chandra, P. 2020. A comprehensive survey of data mining. International Journal of Information Technology 12(4), 1243-1257.

Han, J, Pei, J, and Tong, H. 2022. Datamining: Concepts and techniques. Morgan Kaufmann.

Handoko, S, Fauziah, F, and Handayani, E, T, E. 2020. Implementasi data mining untuk menentukan tingkat penjualan paket data telkomsel menggunakan metode K-means clustering. Jurnal Ilmiah Teknologi dan Rekayasa 25(1), 76-88.

Harahap, N, A, P, Al Qadri, F, Harahap, D, I, Y, Situmorang, M, and Wulandari, S. 2023. Analisis perkembangan industri manufaktur indonesia. El-Mal: Jurnal Kajian Ekonomi & Bisnis Islam 4(5), 1444-1450.

Herman, E, Zsido, K, E, and Fenyves, V. 2022. Cluster analysis with k-mean versus k-medoid in financial performance evaluation. Applied Sciences 12(16), 1-19.

Herviany, M, Delima, S, P, Nurhidayah, T, and Kasini, K. 2021. Perbandingan algoritma K-means dan K-medoids untuk pengelompokkan daerah rawan tanah longsor pada provinsi Jawa Barat: Comparison of K-means and K-medoids algorithms for grouping landslide prone areas in west java province. MALCOM: Indonesian Journal of Machine Learning and Computer Science 1(1), 34-40.

Huang, H, Yao, X, A, Krisp, J, M, and Jiang, B. 2021. Analytics of location-based big data for smart cities: Opportunities, challenges, and future directions. Computers, Environment and Urban Systems 90, 101712.

Jentner, W, and Keim, D, A. 2019. Visualization and visual analytic techniques for patterns. Springer International Publishing, 303-337.

Khan, A, S, S, Fatekurohman, M, and Dewi, Y, S. 2023. Perbandingan algoritma K-medoids dan K-means dalam pengelompokan kecamatan berdasarkan produksi padi dan palawija di jember. Jurnal Statistika dan Komputasi 2(2), 67-75.

Madbouly, M, M, Darwish, S, M, Bagi, N, A, and Osman, M, A. 2022. Clustering big data based on distributed fuzzy K-medoids: An application to geospatial informatics. IEEE Access 10, 20926-20936.

Maharani, S, and Aryanta, W, R. 2023. Dampak buruk polusi udara bagi kesehatan dan cara meminimalkan resikonya. Jurnal Ecocentrism 3(2), 47-58.

Muningsih, E, Maryani, I, and Handayani, V, R. 2021. Penerapan metode K-means dan optimasi jumlah cluster dengan index davies bouldin untuk clustering propinsi berdasarkan potensi desa. Evolusi: Jurnal Sains dan Manajemen 9(1), 95-100.

Nugraha, M, F, Martano, M, and Hayati, U. 2024. Clustering data indonesian food delivery menggunakan metode K-means pada gofood product list. JATI (Jurnal Mahasiswa Teknik Informatika) 8(3), 3484-3492.

Prasetiyo, D, Lestari, W, and Atima, V. 2024. Penerapan clustering dengan K-means untuk pemilihan menu favorit di tetra coffeeshop. JATISI (Jurnal Teknik Informatika dan Sistem Informasi) 11(3).

Prayogo, W, Marhamah, F, Fauzan, H, A, Azizah, R, N, and Va, V. 2021. Strategi pengendalian pencemaran industri untuk pengelolaan mutu air sungai dan tanah di DAS diwak, Jawa Tengah. Jurnal Sumberdaya Alam dan Lingkungan 8(3), 123-132.

Riani, A, P, Voutama, A, and Ridwan, T. 2023. Penerapan K-means clustering dalam pengelompokan hasil belajar siswa dengan metode elbow. Jurnal Teknologi Informasi dan Sistem Komputer TGD 6(1), 164-172.

Ridwan, M, Situmorang, C, and Darpito, H. 2020. Pengaruh car free day terhadap penggolongan kualitas udara parameter so2 dan no2 di depan mesjid raya sumatera barat padang. Jurnal TechLINK 4(2), 1-10.

Romadhon dan Mokhtar, A. 2021. Model penanggulangan pencemaran udara pada mesin asphalt mixing plant. Program Studi Persatuan Insinyur Indonesia 1(1).

Sheikholeslami, R, Razavi, S, Gupta, H, V, Becker, W, and Haghnegahdar, A. 2019. Global sensitivity analysis for high-dimensional problems: How to objectively group factors and measure robustness and convergence while reducing computational cost. Environmental Modelling & Software 111, 282-299.

Soca, N, and Woyanti, N. 2021. Pengaruh unit usaha, nilai output, biaya input, dan upah minimum terhadap penyerapan tenaga kerja industri besar dan sedang di provinsi jawa tengah. BISECER (Business Economic Entrepreneurship) 4(2), 27-37.

Sowan, B, Tzung-Pei H, Ahmad A, Mohammad A, and Nasim, M. 2023. Ensembling validation indices to estimate the optimal number of clusters. Applied Intelligence 53(9), 9933-9957.

Sugiarto, I, R, Hartono, E, D, Widjaja, J, M, Ariffa, R, and Wijaya, J. 2024. Penyelesaian masalah (studi komparasi pengaturan polusi udara menurut hukum negara indonesia dan swiss). Innovative: Journal Of Social Science Research 4(2), 4284-4296.

Sulasminingsih, S, Juwariyah, T, Siahaan, Y, Putri, B, H, and Putra, N, A. 2024. Penerapan tema sdgs kehidupan sehat dan sejahtera untuk menangani polusi udara di jakarta. IKRA-ITH Teknologi Jurnal Sains dan Teknologi 8(1), 18-26.

Syafrinal, I, and Febrianti, E, L. 2023. Penerapan algoritma K-means pada aplikasi data mining untuk menentukan pola penjualan (studi kasus: Zahra mart). Jurnal Digit: Digital of Information Technology 13(1), 31-40.

Tiara, M, I, and Firmawati, N. 2023. Rancang bangun prototipe sistem kontrol penutup ventilasi dan pembersih udara otomatis berbasis mikrokontroler. Jurnal Fisika Unand 12(3), 356-362.

Utari, D, T. 2021. Analisis karakteristik wilayah transmisi covid-19 dengan menggunakan metode K-means clustering. Jurnal Media Teknik Dan Sistem Industri 5(1), 25-32.

Wibawa, A, P, Miftahuddin, F, and Suyono, S. 2021. K-medoids clustering untuk pembentukan database stopword bahasa jawa. Ranah: Jurnal Kajian Bahasa 10(2), 261-269.

Yulianti, A, and Sasana, H. 2021. Analisis peningkatan upah minimum terhadap penyerapan tenaga kerja di provinsi jawa tengah. Jurnal Ekonomi Pembangunan 10(3), 134-143