Clustering of the Air Pollution Standard Index (ISPU) in the Province of DKI Jakarta Using the CLARANS Algorithm

Adelia Ramadhina Azzahra [1*], Nasywa Azzah Nabila [2*], Mohammad Idhom [3*], Trimono [4*]
* Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur
22083010047@student.upnjatim.ac.id [1], 22083010051@student.upnjatim.ac.id [2], idhom@upnjatim.ac.id [3], trimono.stat@upnjatim.ac.id [3]

Article Info

Article history:

Received 2025-06-13
Revised 2025-06-23
Accepted 2025-07-03

Keyword:

Air Pollution Index,
Clustering,
CLARANS,
Jakarta,
Silhouette Score.

ABSTRACT

Air pollution has become a serious global issue. According to IQAir's 2024 report, DKI Jakarta ranked 10th among cities with the worst air quality worldwide, indicating that air pollution in DKI Jakarta has reached a concerning level. This research uses the CLARANS algorithm to cluster daily air quality in DKI Jakarta based on pollution parameters. CLARANS is chosen due to its advantages in terms of big data processing efficiency, outlier resistance, and medoid search capability. The novelty of this research lies in the application of CLARANS to overcome the limitations of clustering algorithms in previous research. This research comprises several stages, including data understanding, data preprocessing, building the CLARANS model, and evaluation using the silhouette score. The CLARANS clustering result using the most optimal parameter combination and k = 3 demonstrates well-separated cluster boundaries, with an overall average silhouette score across all regions and years of 0.6398. The analysis results indicate that air pollution in DKI Jakarta tends to worsen in 2024. Jakarta Barat and Jakarta Pusat are predominantly affected by PM10, CO, and O_3 pollution, whereas Jakarta Selatan and Jakarta Utara are more influenced by SO_2 and NO_2 pollution. On the other hand, air pollution in East Jakarta shows a balanced dominance from both pollutant categories.

This is an open access article under the CC–BY-SA license.

I. Introduction

Air is a fundamental necessity for all living things on Earth. However, the rapid growth of human activity and industrial development has led to a decline in air quality worldwide, making air pollution a serious global issue. The World Health Organization (WHO) reported that air pollution is responsible for approximately 7 million deaths annually [1]. To address this, many countries have developed air quality monitoring systems to control pollutant levels. In Indonesia, air quality is monitored using the Air Pollution Standard Index (ISPU), as stipulated in the Ministerial Decree of the Environment No. KEP 45/MENLH/10/1997 and KEP-107/KABAPEDAL/11/1997 [2]. ISPU is a numerical representation of ambient air quality conditions based on five key parameters, namely particulate matter (PM10), carbon monoxide (CO), sulfur dioxide (SO_2), ozone (O_3), and nitrogen dioxide (NO_2) [3]. The higher the ISPU value, the greater the potential adverse impacts on human health and the environment.

Among Indonesian provinces, DKI Jakarta as the center of economic and governmental activities, experiences relatively high levels of air pollution. According to IQAir's World Air Quality Report in 2024, DKI Jakarta ranked 10th among cities with the worst air quality in the world [4]. This condition indicates that air pollution in DKI Jakarta has reached a concerning level, potentially increasing the number of cases of illness and death due to air pollution. Therefore, a data-driven approach is needed to monitor and understand air pollution patterns in DKI Jakarta. Clustering provides an effective approach to identify daily air quality patterns based on pollution parameters, as well as enabling the analysis of air quality distribution and temporal trends in DKI Jakarta from year to year across various regions.

Previous research has explored various clustering approaches for the Air Pollution Standard Index (ISPU) in DKI Jakarta. For instance, [5] compared the performance of

Fuzzy C-Means (FCM) and Gaussian Mixture Model (GMM) in ISPU clustering from 2019 to 2020. This research used five air pollution parameters, including PM10, CO, SO_2, O_3, and NO_2, across five monitoring stations. The results showed that FCM outperformed GMM in clustering performance, with a higher average Silhouette Index of 0.6443 compared to 0.5748. However, this research primarily focused on algorithm comparison without exploring the temporal or spatial patterns of air pollution in depth, despite the exploration of these patterns being a key objective in clustering-based analysis.

In another research, such as [6] and [7], K-Means and K-Medoids were employed to cluster air quality categories in 2021 based on six pollution parameters, with PM2.5 included as an additional parameter beyond those used in [5]. These research focused on clustering air pollution levels into general categories, such as low, moderate, and high pollution levels. However, both research employed conventional distance-based clustering methods, which are known to be sensitive to outliers and tend to perform poorly on complex or noisy datasets. This limitation is reflected in their clustering evaluation metrics, with K-Means yielding a Davies-Bouldin Index (DBI) of 1.172 and K-Medoids producing a relatively low silhouette score of 0.25. These results indicate suboptimal cluster separation, highlighting the need for a more robust and scalable clustering approach.

To overcome the limitations of previous research, this research proposes the use of the CLARANS algorithm. CLARANS (Clustering Large Application based on RANdomized Search) is a part of the K-Medoids algorithm, developed from the PAM (Partitioning Around Medoids) and CLARA (Clustering Large Application), optimized for large datasets and robust against outliers by using a randomized search [8]. Compared to other K-Medoids algorithms, such as PAM and CLARA, CLARANS is more efficient in finding medoids because it is not limited to only searching local areas, thus producing better cluster quality [9]. This efficiency stems from CLARANS' ability to sample neighbors dynamically during the search process. It allows it to balance exploration and exploitation, contributing to its superior performance on large and complex datasets. Moreover, CLARANS is considered an alternative clustering algorithm because, during the medoid iteration, it selects the medoid point with the minimum sum of distances to other members within its cluster (the minimum cost) [10].

However, the CLARANS algorithm has not been explored in similar case studies, despite its advantages in terms of big data processing efficiency, outlier resistance, and medoid search capability over other clustering algorithms. Previous research has proved that CLARANS can produce effective clustering results across various case studies. According to research by [8], CLARANS successfully handled a dataset of 12287 observations to cluster forest and land fire potential in Indonesia based on fire hotspot distribution. The results obtained 2 clusters with a silhouette score of 0.896, indicating CLARAN's ability to spatially distinguish areas with high and medium fire potential with high precision. Therefore, the novelty of this research lies in the application of CLARANS, which has been limited in prior use for Air Pollution Standard Index clustering. The application of CLARANS aims to improve the silhouette score of previous research by providing more accurate clusters.

This research focuses on implementing the CLARANS algorithm to cluster daily air quality in the Province of DKI Jakarta based on the Air Pollution Standard Index (ISPU). This clustering aims to obtain air quality distribution patterns across various regions of DKI Jakarta. The local government can use the results of this clustering to formulate more effective and targeted air pollution control policies. Furthermore, this research can also serve as a reference for future studies in developing clustering methods, particularly in the application of CLARANS and air quality analysis.

II. METHODS

The overall research flowchart is presented in Figure 1.

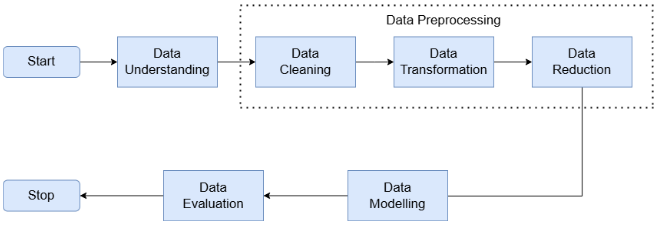

Figure 1. Methods Research Flowchart

A. *Data Understanding*

Data understanding is conducted to identify the data type required for the research. The research data were obtained from the official Open Data Jakarta website, which provides daily records of the Air Pollution Standard Index (ISPU) in the Province of DKI Jakarta from 2023 to 2024 [11]. Given that Open Data Jakarta is a reputable governmental organization, the data utilized in this research are assumed to be accurate and reliable, requiring no additional validation.

B. *Data Preprocessing*

Data preprocessing involves applying various techniques to improve the quality of raw data, ensuring accuracy and optimal results [12]. In this research, multiple preprocessing methods are employed, including:

1) Data Cleaning: This stage involves handling missing values to avoid analysis errors using either the Kalman filter or interpolation method, depending on the time series pattern of each pollutant. The Kalman filter is suitable for datasets with large blocks of missing values and seasonal patterns. This method estimates missing values using observations of previous missing data blocks and underlying seasonal trends [13]. Conversely, interpolation is suitable for time series data exhibiting clear trend patterns, as it estimates missing values based on surrounding data points [14]. By tailoring the imputation method to the underlying temporal structure of each pollutant, the imputed values are more likely to reflect accurate environmental conditions, thereby maintaining the integrity of the subsequent analysis. In addition, this stage involves handling outliers to minimize potential distortions in the analysis by using the *tsoutliers()* function in RStudio software.

2) Data Transformation: Normalization is applied to standardize the scale of air pollution parameters, ensuring fair distance calculations between data points in CLARANS and preventing dominance variables with larger value scales. The normalization process is performed using the Min-Max Scaler. This method converts each value in the air pollution parameter into a range of 0 to 1, ensuring that differences in scale between parameters do not affect the clustering process while maintaining the relative order of values in the original data [15]. The formula of the Min-Max Scaler is shown in Equation (1).

$$x' = \frac{x - min(x)}{range(x)} = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

Where $range(x)$ is the difference between minimum and maximum, $min(x)$ is the minimum, and $max(x)$ is the maximum [16].

3) *Data Reduction*: Uniform Manifold Approximation and Projection (UMAP) is used to reduce the dimensionality of a large dataset. Given that all features demonstrate low linear correlations, UMAP is deemed the most appropriate method. This method aims to preserve both local structures (nearest neighbors) and global structures (overall patterns) in the data without assuming any linear relationship among variables [17].

C. Data Modelling

This stage involves determining and applying clustering methods to the preprocessed dataset. In this research, the CLARANS algorithm is employed for clustering. Unlike traditional k-medoids algorithms that exhaustively search the entire solution space, CLARANS finds k-medoids from a dataset of n data objects by drawing a sample of neighbors in each search step [18]. This approach does not configure the search to a localized area, allowing the algorithm to explore a broader solution space [6]. By balancing local refinement and global exploration, CLARANS effectively navigates the complex solution landscape, making it suitable for high-dimensional data. Its ability to dynamically sample neighbors rather than exhaustively evaluate all possibilities significantly reduces computational overhead while maintaining high clustering quality. Figure 2 shows the exact flow of the CLARANS algorithm.

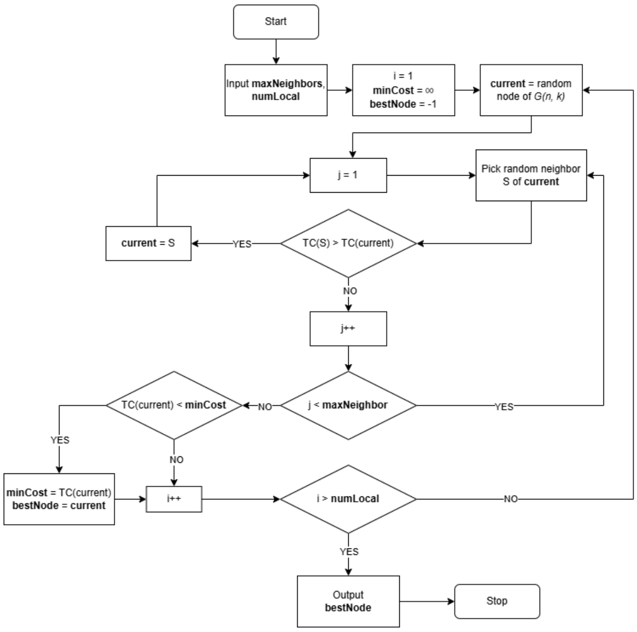

Figure 2. CLARANS Algorithm Flowchart

1) Input parameters *numLocal* (the number of local minima obtained) and *maxNeighbor* (the maximum number of neighbors examined)
2) Initialize *i* to 1, *minCost* to a large number, and *bestNode* to an empty node.
3) Select an arbitrary node *current* in the solution space *G(n, k)*
4) Set *j* to 1
5) Randomly select a neighbor *S* of *current* and calculate the cost difference between them.
6) If *S* has a lower cost, set *current* to *S*, and repeat Step 4.
7) Otherwise, increment *j* by 1. If *j* ≤ *maxNeighbor*, repeat Step 5.
8) When *j* > *maxNeighbor*, compare the cost of *current* with *minCost*. If *current* has a lower cost, set *minCost* to the cost of *current* and *bestNode* to *current*.
9) Increment *i* by 1. If *i* > *numLocal*, output *bestNode* as the optimal medoid solution and stop. Otherwise, repeat Step 3.

The higher the *maxNeighbor* value, the longer the algorithm's search time and the highest quality of the local optimal solution [19]. This research uses *MaxNeighbor* values of 0.025 and 2, while the *NumLocal* parameter uses values of 1 and 2.

D. Model Evaluation

The model's performance is evaluated using the silhouette score, a metric that measures how well each point fits within its assigned cluster compared to others [20]. The silhouette score ranges from -1 to 1, where a higher score indicates better clustering. The silhouette score calculates the average distance between data points within the same cluster and compares it with the average distance to points in the closest neighboring clusters [21]. The formula for calculating the silhouette score is shown in Equation (2).

$$silhouette(x) = \frac{b(x) - a(x)}{max\{a(x), b(x)\}} \qquad (2)$$

Where $a(x)$ is the average distance between the object x and all other objects within the same cluster, and $b(x)$ is the average distance between the object x and all other objects in the nearest neighboring cluster [22].

To evaluate clustering performance using silhouette scores, the impact of preprocessing on the clustering results was first assessed by systematically testing multiple combinations of preprocessing methods, including outlier handling, normalization, and dimensionality reduction. These combinations are evaluated based on their resulting silhouette scores to assess their effect on cluster quality.

In addition, model performance is evaluated across different values of k (ranging from 3 to 6) to identify the optimal number of clusters for the modelling process. The selection of k is made through observation, identifying which value of k yields the highest silhouette score across most datasets, as well as which one achieves the highest average silhouette score overall. Furthermore, a non-parametric Friedman test was conducted to statistically compare clustering results across different values of k [23]. The null hypothesis stated that there is no significant difference in clustering performance among the tested k values. If the null hypothesis is rejected (indicating a significant difference), a post-hoc analysis is performed to identify which specific pairs of k values differ significantly.

III. Result and Discussion

A. Data Understanding

The ISPU dataset of DKI Jakarta from 2023 to 2024 was obtained from the Open Data Jakarta website, which provides daily records of the Air Pollution Standard Index (ISPU). A snippet of this dataset can be seen in Table 1.

TABLE I
ISPU DKI JAKARTA DATASET SNIPPET

Date	Region	CO	PM10	SO_2	NO_2	O_3
1/1/2024	Jakarta Pusat	9	64	52	30	13
1/1/2024	Jakarta Utara	18	67	34	-	26
1/1/2024	Jakarta Selatan	54	65	14	-	13
1/1/2024	Jakarta Timur	20	74	10	10	24
1/1/2024	Jakarta Barat	18	58	40	12	55

The dataset consists of 3655 entries, with seven columns, which are “Date”, “Region,” “PM10,” “SO_2,” “CO,” “NO_2,” and “O_3”. In this stage, all variables were examined for missing values and temporal patterns. It was observed that the variable pairs PM10 and O_3, as well as CO, SO_2, and NO_2, exhibited different temporal patterns. Due to space limitations, PM10 and NO_2 were selected as representative variables for each type of temporal pattern. Figure 3 illustrates the pattern of the PM10 and Figure 4 illustrates the NO_2 pattern.

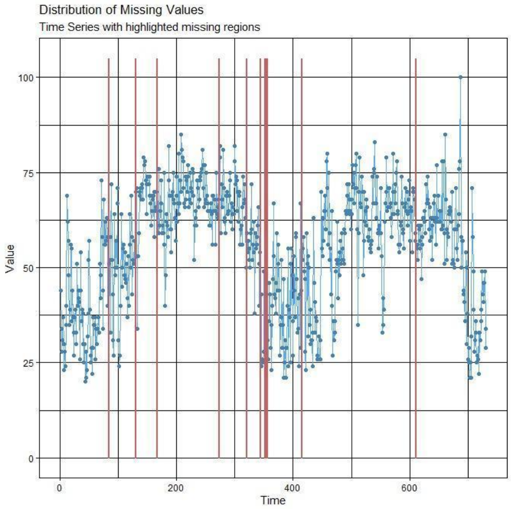

Figure 3. PM10 Patterns

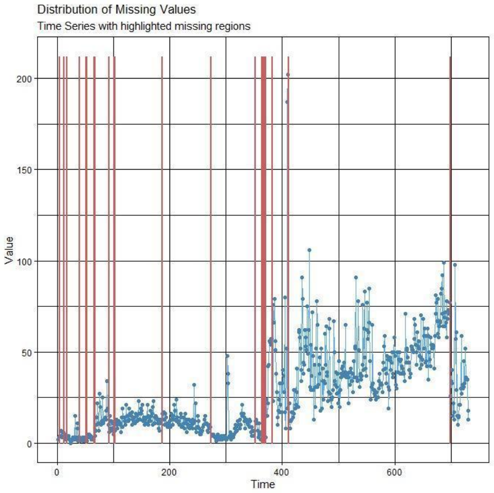

Figure 4. NO_2 Patterns

The results showed that parameters PM10 and O_3 exhibited seasonal patterns, while parameters CO, SO_2, and NO_2 exhibited slight trend patterns. Seasonal patterns were identified based on recurring cycles in the data points, whereas trend patterns were observed through gradual increases or consistent directional changes over time. These differences in temporal behavior may be attributed to the distinct sources and environmental behaviors of the pollutants. For instance, PM10 and O_3 are often influenced by meteorological factors and seasonal activities, whereas CO, SO_2, and NO_2 are more closely linked to continuous anthropogenic emissions, such as traffic and industrial activities. Additionally, the red blocks in these figures indicate missing values, highlighting periods where data was not recorded or lost. This emphasized the need for proper handling before the datasets can be reliably used for clustering analysis.

B. Data Preprocessing

The dataset contained missing values and outliers in several air pollution parameters. For PM10 and O_3, which exhibited substantial blocks of missing values and seasonal patterns, imputation was performed using the Kalman filter. Conversely, interpolation was utilized for CO, SO_2, and NO_2, as these parameters exhibited slight trend patterns. After imputation, outliers were addressed using the *tsoutliers()* function, followed by Min-Max normalization and dimensionality reduction using UMAP. The normalized dataset can be seen in Table 2.

TABLE 2
DATA NORMALIZATION RESULT

Date	Region	CO	PM10	SO_2	NO_2	O_3
1/1/2024	Jakarta Pusat	0.17	0.81	1.00	0.39	0.16
1/1/2024	Jakarta Utara	0.55	0.45	0.77	0.33	0.51
1/1/2024	Jakarta Selatan	0.85	0.65	0.17	0.02	0.26
1/1/2024	Jakarta Timur	0.42	0.22	0.28	0.24	0.39
1/1/2024	Jakarta Barat	0.50	0.73	0.41	0.38	0.59

C. Data Modelling

Clustering was performed using the CLARANS algorithm. Various combinations of *MaxNeighbor* and *NumLocal* were tested, using varying k values from 3 to 6 for each region and year. The configuration with the highest silhouette score was selected for final clustering, resulting in k = 3 being chosen as the optimal number of clusters.

Among the other regions and years, Jakarta Selatan 2023 is presented below as a representative case, as its clustering structure is consistent with the overall trend observed across the other subsets. The CLARANS clustering result using the most optimal parameter combination and k = 3 demonstrates well-separated cluster boundaries, indicating high intra-cluster similarity and inter-cluster dissimilarity based on pollution patterns. The representative visualization of this result is shown in Figure 4.

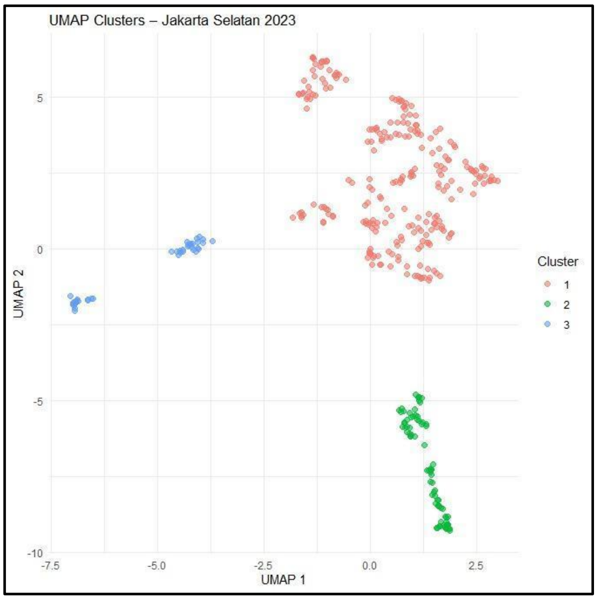

Figure 4. UMAP Visualization Cluster

D. Model Evaluation

The silhouette score was used as the evaluation metric to assess clustering performance. Initially, to evaluate the influence of preprocessing methods on clustering performance, several combinations of outlier handling, normalization, and dimensionality reduction were tested on selected datasets. Table 3 shows a comparison of silhouette scores across different preprocessing pipelines applied to the Jakarta Timur 2023 dataset.

TABLE 3
PREPROCESSING METHODS EVALUATION

Outlier Handling	Min-Max Normalization	UMAP	Silhouette Score
No	No	No	0.359
Yes	No	No	0.375
Yes	Yes	No	0.455
Yes	Yes	Yes	0.678

As shown above, the application of UMAP after both outlier handling and Min-Max normalization improved clustering performance. This trend was consistent in most other datasets. Therefore, the final pipeline involved handling outliers, Min-Max normalization, and UMAP for dimensionality reduction. The consistent improvement in silhouette scores justifies the use of these preprocessing steps prior to clustering.

Next, the model performance was evaluated across different values of k (ranging from 3 to 6) to determine the optimal number of clusters for the modelling process. Table 4 shows the silhouette scores obtained for each k value across five regions in Jakarta for the years 2023 and 2024.

TABLE 4
SILHOUETTE SCORE RESULTS

Region & Year	k=3	k=4	k=5	k=6
Jakarta Pusat 2023	**0.572**	0.540	0.524	0.533
Jakarta Pusat 2024	0.639	0.679	0.696	**0.708**
Jakarta Utara 2023	**0.677**	0.626	0.623	0.623
Jakarta Utara 2024	0.595	0.602	**0.620**	0.609
Jakarta Selatan 2023	**0.717**	0.602	0.622	0.617
Jakarta Selatan 2024	**0.634**	0.586	0.618	0.608
Jakarta Timur 2023	0.678	**0.817**	0.693	0.726
Jakarta Timur 2024	0.671	**0.689**	0.673	0.627
Jakarta Barat 2023	**0.636**	0.601	0.579	0.579
Jakarta Barat 2024	0.579	0.556	**0.598**	0.587
Average	**0.6398**	0.6298	0.6246	0.6217

As shown in Table 4, it can be observed that most regions achieved their highest silhouette scores when the number of clusters (k) was set to 3. Furthermore, the overall average silhouette score across all regions and years was also highest at k = 3, reaching a score of 0.6398. Although the Friedman test was conducted to assess the statistical significance of clustering performance across various values of k, the results did not reveal statistically significant differences ($p > 0.05$). Therefore, in the absence of strong statistical evidence favoring a specific k value, k = 3 was selected as the optimal number of clusters. This decision was based on the consistent clustering performance, facilitating interpretation and comparative analysis across regions and time periods.

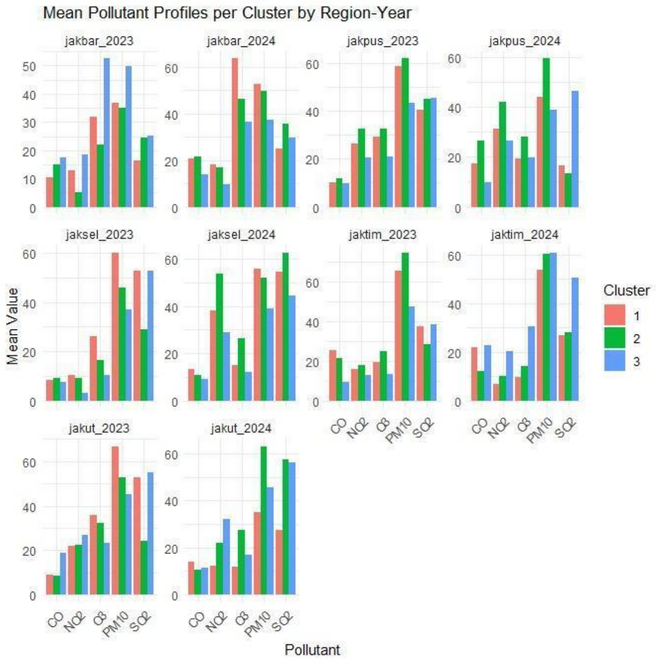

Figure 5. Mean Pollutant Profiles for Each Cluster

Figure 5 presents the average concentrations of CO, NO_2, O_3, PM10, and SO_2 for each region-year cluster, allowing for easy identification of similar pollutant profiles. Clusters with the same numerical ID do not necessarily have similar profiles in different locations or years; for example, Cluster 1 in Jakarta Selatan 2024 differs from Cluster 1 in Jakarta Utara 2024. Therefore, to facilitate comparison across regions and years, the clusters were regrouped into six meta-profiles based on the similarity of their pollutant profiles. The number six was chosen because it provides a balance between detail and simplicity, more than six results in excessive fragmentation, while fewer than six can disguise important profiles.

Based on the visual observation, clusters that have low PM10 and O_3, but high on SO_2 were labeled "Low PM10 & O_3, High SO_2", clusters that have high O_3 and CO, but low on SO_2 were labeled "High O_3 & CO, Low SO_2", clusters that have high CO and PM10 were labeled "High CO & PM10", clusters that have high PM10 and SO_2, but low on CO were labeled "High PM10 & SO_2, Low CO", clusters that have low PM10, SO_2, and NO_2 were labeled "Low PM10, SO_2 & NO_2", and clusters that have high SO_2 and NO_2 are labeled "High SO_2 & NO_2".

Figure 6 illustrates the proportional distribution of pollutant profiles for the region and year subset. This illustration provides a comparison of air pollution profiles across the five regions in DKI Jakarta for 2023 and 2024.

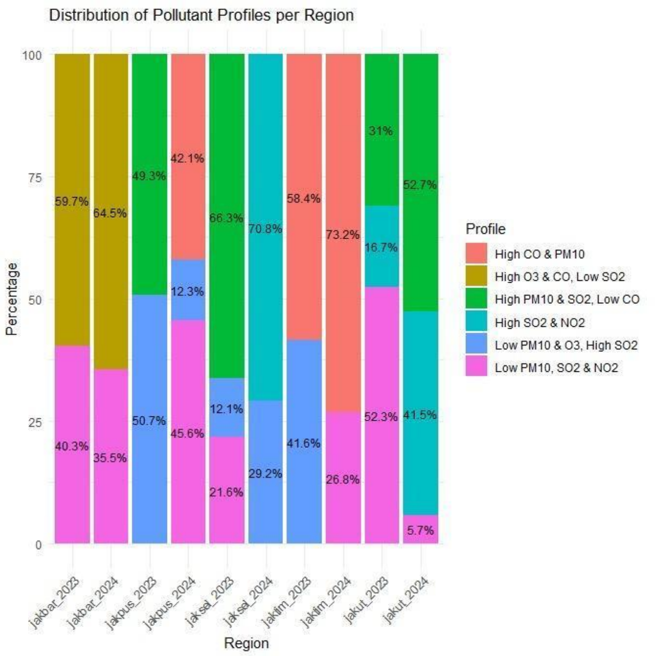

Figure 6. Distribution of Pollutant Profiles for Each Region

As shown in Figure 6, the air quality cluster distribution in Jakarta Barat is primarily dominated by high levels of O_3 and CO, with consistently low levels of SO_2. In 2023, 59.7% of this location belonged to the "High O_3 & CO, Low SO_2" profile, which increased to 64.5% in 2024. Meanwhile, the "Low PM10, SO_2 & NO_2" profile declined from 40.3% to 35.5%. This trend indicates a worsening of air quality, likely driven by increased emissions from gasoline-powered vehicles and traffic congestion, both of which significantly contribute to elevated levels of CO and O_3. Conversely, the consistently low SO_2 levels suggest a limited presence of primary stationary sources, such as heavy industries in Jakarta Barat. This aligns with the area's land use profile, which is predominantly residential and commercial rather than industrial.

In Jakarta Pusat, the distribution in 2023 was relatively balanced between the "Low PM10 & O_3, High SO_2" profile at 50.7% and the "High PM10 & SO_2, Low CO" profile at 49.3%, indicating elevated SO_2 pollution primarily driven by emissions from heavy industrial activities. In 2024, a "Low PM10, SO_2 & NO_2" profile emerged at 45.6%, indicating that some areas within this region still maintained better air quality. However, there was a dominant shift toward the "High CO & PM10" profile at 42.1%, accompanied by a sharp decline in the "Low PM10 & O_3, High SO_2" profile to just 12.3%. This shift indicates a change in the primary sources of pollution from industrial emissions to vehicular emissions, which became the main contributors to CO and PM10 levels in Jakarta Pusat in 2024. This trend aligns with the region's land use, which serves as a governmental and economic hub, resulting in high volumes of both private and public transportation. Furthermore, traffic congestion significantly contributes to elevated CO and PM10 emissions. Therefore, pollution control efforts should shift focus from industrial emissions to transportation emission management and traffic congestion mitigation.

Jakarta Selatan shows a growing dominance of SO_2 and NO_2 pollution. In 2023, the most significant pollution contribution came from the "High PM10 & SO_2, Low CO" profile, at 66.3%, followed by the "Low PM10, SO_2 & NO_2" profile at 21.6%, and the "Low PM10 & O2, High SO_2" profile at 12.1%. In 2024, the landscape shifted significantly, with the "High SO_2 & NO_2" profile decreasing to 70.8% and the "Low PM10 & O_3, high SO_2" profile increasing to 29.2%. This trend indicates the growing impact of industrial activities and diesel-fueled vehicles, which contribute to high levels of SO_2 and NO_2 pollution in Jakarta Selatan. Elevated SO_2 pollution suggests the presence of primary stationary sources, such as factories or industrial areas. Meanwhile, the significant NO_2 pollution is primarily attributed to the high volume of diesel-powered vehicles. In addition, the relatively low levels of PM10 and CO indicate that emissions from gasoline-fueled vehicles are less dominant in this region compared to the city center area.

In Jakarta Timur, air quality is heavily influenced by CO and PM10. Air pollution patterns in 2023 were characterized by the "High CO & PM10" profile at 58.4% and the "Low PM10 & O_3, High SO_2" profile at 41.6%, indicating a mixed pollution from both transportation emissions and industrial activities. This aligns with the land use in Jakarta Timur, which consists of densely populated residential areas and heavy industry. In 2024, the dominance of the "High CO & PM10" profile further intensified, increasing significantly to 73.2%, indicating growing emissions from transportation sources. In addition, a new profile "Low PM10, SO_2 & NO_2" emerged at 26.8%, suggesting that certain areas of the region still managed to maintain relatively better air quality despite the overall deterioration.

Jakarta Utara exhibits a notable shift in air quality patterns, transitioning from relatively better conditions in 2023 to deteriorated in 2024. In 2023, the region was dominated by the "Low PM10, SO_2 & NO_2" profile, but the proportion of this profile dropped sharply to just 5.7% in 2024. This indicates an increase in PM10, SO_2, and NO_2 pollutants during this period. The rise in pollution is also reflected in the increase of the "High SO_2 and NO_2" profile from 16.7% to 41.5%, and the "High PM10 & SO_2, Low CO" profile from 31% to 52.7%. These conditions align with the characteristics of North Jakarta as a heavy industrial and port area, which are the primary sources of PM10, NO_2, and SO_2 emissions. Therefore, improving air quality in this region requires targeted mitigation strategies focusing on industrial emission control.

These findings highlight how pollutant composition varies not only by region but also over time. Based on the analysis, air pollution in DKI Jakarta tends to worsen in 2024. PM10, CO, and O_3 pollution are most prominent in Jakarta Barat and Jakarta Pusat, primarily driven by high motor vehicle emissions resulting from traffic congestion. This condition necessitates urban traffic management policies that aim to reduce congestion, along with the development of green infrastructure to help mitigate these pollutants. In contrast, SO_2 and NO_2 pollution are more prevalent in Jakarta Selatan and Jakarta Utara, primarily attributed to heavy industrial activities. Therefore, stricter industrial emission controls, the enforcement of cleaner fuel standards, and related environmental regulations are necessary. Meanwhile, air pollution in Jakarta Timur reflects a balanced contribution from both vehicular emissions and industrial activities, indicating the need for an integrated strategy that addresses both emission sources simultaneously.

IV. CONCLUSION

The final stage of preprocessing pipeline consisting of imputation using the Kalman Filter and interpolation, outlier handling, Min-Max normalization, and dimensionality reduction using UMAP. The clustering process performed using the CLARANS algorithm with the most optimal parameter combination and k = 3 produced well-separated cluster boundaries, as reflected by an overall average silhouette score of 0.6398 across all regions and years. Notably, most regions also achieved their highest silhouette scores when the number of clusters was set to 3. Based on the analysis results, air pollution in DKI Jakarta tends to worsen in 2024. PM10, CO, and O_3 pollution dominated Jakarta Barat and Jakarta Pusat. Conversely, SO_2 and NO_2 pollution dominated Jakarta Timur and Jakarta Selatan. Meanwhile, air pollution in Jakarta Timur demonstrated a balanced contribution from both categories of pollutants.

Despite the clear results, several limitations in this research must be acknowledged. One key limitation is the use of daily average data, which may obscure short-term pollution spikes, such as hourly peaks, potentially missing extreme patterns or localized incidents. Additionally, the dataset contains a substantial number of missing values, requiring imputation to complete the data. However, the imputed values may not fully reflect actual conditions. Based on these limitations, future research is recommended to use data with higher time resolutions, such as hourly data, to more accurately capture the dynamics of air pollution. The adoption of more advanced imputation techniques should also be considered to improve the accuracy of estimated values. Furthermore, exploring other machine learning algorithms as alternatives to clustering may provide valid performance comparisons. Lastly, integrating optimization techniques could support automatic selection of k values and tuning of other parameters, thereby improving clustering quality as measured by silhouette scores.

REFERENCES

[1] S. Annas, U. Uca, I. Irwan, R. H. Safei, and Z. Rais, "Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia," *Jambura Journal of Mathematics*, vol. 4, no. 1, pp. 167–176, Jan. 2022, doi: 10.34312/jjom.v4i1.11883.

[2] P. Alusvigayana, A. S. Yuwono, M. Yani, and S. Syarwan, "Evaluation of the Air Pollutant Standard Index (ISPU) parameter concentration limits in industrial estates on Java Island," *Jurnal Pengelolaan Sumberdaya Alam dan Lingkungan (Journal of Natural Resources and Environmental Management)*, vol. 13, no. 4, pp. 537–548, Dec. 2023, doi: 10.29244/jpsl.13.4.537-548.

[3] V. Deandra, F. Hamami, and I. Darmawan, "Analisis Klasifikasi Kualitas Udara Menggunakan Metode Algoritma K-Nearest Neighbor Pada Provinsi Dki Jakarta," *e-Proceeding of Engineering*, vol. 11, no. 4, pp. 3692–3698, 2024.

[4] "2024 World Air Quality Report," 2024. Accessed: May 15, 2025. [Online]. Available: https://www.iqair.com/us/newsroom/waqr-2024-pr.

[5] M. H. S. Situmorang, B. I. Nasution, M. E. Aminanto, Y. Nugraha, and J. I. Kanggrawan, "Air Pollution Index (API) Analysis at Jakarta in 2019-2020 using Fuzzy C-Means and Gaussian Mixture Model," in Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications, New York, NY, USA: ACM, Nov. 2022, pp. 174–178. doi: 10.1145/3575882.3575916.

[6] I. Mahendrasyah, A. Diana, Rusdah, and D. Mahdiana, "PENERAPAN ALGORITMA K-MEANS UNTUK KLASTERISASI INDEKS STANDAR PENCEMARAN UDARA," Teknologi, vol. 14, no. 2, pp. 146–156, Dec. 2024, doi: 10.26594/teknologi.v14i2.4088.

[7] H. al AZIES, "Air Pollution in Jakarta, Indonesia Under Spotlight: An AI-Assisted Semi-Supervised Learning Approach," Proceedings of The International Conference on Data Science and Official Statistics, vol. 2023, no. 1, pp. 150–161, Dec. 2023, doi: 10.34123/icdsos.v2023i1.348.

[8] S. Wisa Fitri, Z. Martha, Y. Kurniawati, and Zilrahmi, "Pengelompokan Potensi Kebakaran Hutan/Lahan di Indonesia Berdasarkan Sebaran Titik Panas Menggunakan Metode CLARANS," UNP Journal of Statistics and Data Science, vol. 2, no. 3, pp. 273–278, Aug. 2024, doi: 10.24036/ujsds/vol2-iss3/182.

[9] R. T. Ng and Jiawei Han, "CLARANS: a method for clustering objects for spatial data mining," IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 5, pp. 1003–1016, Sep. 2002, doi: 10.1109/TKDE.2002.1033770.

[10] A. Vatresia, F. P. Utama, I. P. Hati, and L. Z. Mase, "Discovering Bengkulu Province Earthquake Clusters with CLARANS Methods," Journal of Soft Computing in Civil Engineering, vol. 8, no. 3, pp. 71–86, 2024.

[11] "Data Indeks Standar Pencemar Udara (ISPU) di Provinsi DKI Jakarta," Satu Data Jakarta. Accessed: May 15, 2025. [Online]. Available: https://satudata.jakarta.go.id

[12] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Frontiers in Energy Research*, vol. 9, Mar. 2021, doi: 10.3389/fenrg.2021.652801.

[13] P. Bansal, P. Deshpande, and S. Sarawagi, "Missing Value Imputation on Multidimensional Time Series," *Proceedings of the VLDB Endowment*, 2021.

[14] C. Oh, S. Han, and J. Jeong, "Time-Series Data Augmentation based on Interpolation," *Procedia Computer Science*, vol. 175, pp. 64–71, 2020, doi: 10.1016/j.procs.2020.07.012.

[15] V. Sharma, "A Study on Data Scaling Methods for Machine Learning," *International Journal for Global Academic & Scientific Research*, vol. 1, no. 1, Feb. 2022, doi: 10.55938/ijgasr.v1i1.4.

[16] S. Sinsomboonthong, "Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification," *International Journal of Mathematics and Mathematical Sciences*, vol. 2022, pp. 1–9, Apr. 2022, doi: 10.1155/2022/3584406.

[17] I. Stolarek, A. Samelak-Czajka, M. Figlerowicz, and P. Jackowiak, "Dimensionality reduction by UMAP for visualizing and aiding in classification of imaging flow cytometry data," *iScience*, vol. 25, no. 10, p. 105142, Oct. 2022, doi: 10.1016/j.isci.2022.105142.

[18] Y. FAKIR, R. ELAYACHI, and B. MAHI, "Clustering objects for spatial data mining: a comparative study," *Journal of Big Data Research*, vol. 1, no. 3, pp. 1–11, Mar. 2023, doi: 10.14302/issn.2768-0207.jbr-23-4478.

[19] J. Zhang and H. Wang, "Analysis of CLARANS Algorithm for Weather Data Based on Spark," *Computers, Materials & Continua*, vol. 76, no. 2, pp. 2427–2441, 2023, doi: 10.32604/cmc.2023.038462.

[20] I. K. Khan *et al.*, "Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm," *Egyptian Informatics Journal*, vol. 27, p. 100504, Sep. 2024, doi: 10.1016/j.eij.2024.100504.

[21] S. Renaldi. S, D. A. Prasetya, and A. Muhaimin, "Analisis Klaster Partitioning Around Medoids dengan Gower Distance untuk Rekomendasi Indekos (Studi Kasus: Indekos di Sekitar Kampus UPNVJT)," *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 3, pp. 2060–2069, Jul. 2024, doi: 10.33379/gtech.v8i3.4898.

[22] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy*, vol. 23, no. 6, p. 759, Jun. 2021, doi: 10.3390/e23060759.

[23] A. M. Ikotun and A. E. Ezugwu, "Boosting k-means clustering with symbiotic organisms search for automatic clustering problems," *PLoS ONE*, vol. 17, no. 8, Aug. 2022, doi: 10.1371/journal.pone.0272861.