

PENGELOMPOKAN SPASIO-TEMPORAL KUALITAS UDARA PERKOTAAN DI JAKARTA MENGGUNAKAN ALGORITMA K-MEANS BERDASARKAN PARAMETER AQI

Indriani Saputri, Agustia Hananto, Fitria Nurapriani

Sistem Informasi, Universitas Buana Perjuangan Karawang

Jl. H.S. Ronggowaluyo, Teluk Jambe, Desa Sirnabaya, Kecamatan Teluk Jambe Timur,

Kabupaten Karawang, Jawa Barat 41361

si21.indrianisaputri@mhs.ubpkarawang.ac.id

ABSTRAK

Kualitas udara di Jakarta menjadi permasalahan lingkungan yang serius akibat meningkatnya aktivitas transportasi, industri, dan urbanisasi yang memicu tingginya konsentrasi polutan seperti $PM_{2.5}$, PM_{10} , SO_2 , NO_2 , CO, dan O_3 . Penelitian ini bertujuan untuk menganalisis pola kualitas udara menggunakan pendekatan clustering spatio-temporal berbasis algoritma K-Means. Tahapan penelitian meliputi pra-pemrosesan data, normalisasi menggunakan StandardScaler, serta pembentukan fitur temporal dan spasial. Proses clustering dilakukan dengan variasi jumlah cluster ($k=2$ hingga $k=6$), menggunakan metode Elbow untuk menentukan jumlah cluster optimal dan Silhouette Score untuk evaluasi. Hasil penelitian menunjukkan bahwa nilai Silhouette Score tertinggi diperoleh pada $k=2$ sebesar 0,1994, namun $k=3$ dipilih karena lebih representatif dalam menggambarkan variasi data. Clustering menghasilkan tiga kelompok utama, yaitu kualitas udara buruk yang didominasi $PM_{2.5}$, SO_2 , dan NO_2 , kelompok dengan dominasi ozon sebagai polusi sekunder, serta kelompok dengan kualitas udara relatif baik. Distribusi data menunjukkan dominasi cluster dengan karakteristik ozon sebesar 44,37%. Meskipun nilai Silhouette Score relatif rendah, metode K-Means tetap mampu memberikan gambaran pola kualitas udara secara informatif..

Kata kunci : *Kualitas Udara, K-Means, Clustering, AQI, Silhouette Score*

1. PENDAHULUAN

Kualitas udara perkotaan merupakan salah satu isu lingkungan yang krusial karena dampaknya yang signifikan terhadap kesehatan masyarakat dan keberlanjutan pembangunan kota. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa paparan polutan udara seperti particulate matter ($PM_{2.5}$ dan PM_{10}), nitrogen dioksida (NO_2), sulfur dioksida (SO_2), karbon monoksida (CO), dan ozon (O_3) berkontribusi terhadap penyakit pernapasan, gangguan kardiovaskular, serta kematian dini, khususnya di wilayah perkotaan padat penduduk [34],[9]. Hal ini menunjukkan pentingnya pemantauan dan analisis kualitas udara secara efektif dalam konteks perkotaan modern.

Jakarta sebagai ibu kota Indonesia menghadapi permasalahan kualitas udara yang semakin kompleks akibat tingginya laju urbanisasi, pertumbuhan kendaraan bermotor, aktivitas industri, serta kepadatan penduduk. Beberapa penelitian menunjukkan bahwa konsentrasi $PM_{2.5}$ dan PM_{10} di Jakarta memiliki variasi temporal yang signifikan dan dipengaruhi oleh faktor meteorologi seperti suhu, kelembapan, dan curah hujan [29],[4]. Hal ini mengindikasikan bahwa kualitas udara bersifat dinamis dan berubah dari waktu ke waktu.

Selain aspek temporal, faktor spasial juga berperan penting dalam distribusi polusi udara. Penelitian di wilayah Jabodetabek menunjukkan adanya keterkaitan spasial antarwilayah, di mana kondisi kualitas udara di suatu wilayah dapat

memengaruhi wilayah sekitarnya [5],[6]. Oleh karena itu, analisis kualitas udara memerlukan pendekatan yang mampu mengintegrasikan dimensi spasial dan temporal secara simultan.

Air Quality Index (AQI) merupakan indikator yang umum digunakan untuk merepresentasikan kondisi kualitas udara dengan menggabungkan beberapa parameter polutan ke dalam satu nilai yang mudah diinterpretasikan [1]. Namun demikian, meskipun data AQI tersedia secara historis maupun real-time, pemanfaatannya masih sering terbatas pada analisis deskriptif tanpa eksplorasi pola spasial dan temporal yang mendalam [6].

Seiring dengan perkembangan teknologi data dan machine learning, metode unsupervised learning seperti K-Means clustering banyak digunakan untuk mengidentifikasi pola tersembunyi dalam data berskala besar. Algoritma K-Means dikenal karena kesederhanaan, efisiensi komputasi, serta kemampuannya dalam mengelompokkan data numerik multidimensi [6],[7]. Dalam studi kualitas udara, K-Means telah digunakan untuk mengidentifikasi zona polusi dan menganalisis variasi polutan dari waktu ke waktu [12]. Namun, sebagian besar penelitian masih berfokus pada analisis spasial atau temporal secara terpisah, dan hanya sedikit yang mengintegrasikan kedua dimensi tersebut secara bersamaan, khususnya pada kasus Jakarta [13].

Untuk mengatasi keterbatasan tersebut, penelitian ini mengusulkan pendekatan clustering spatio-temporal menggunakan algoritma K-Means

untuk menganalisis kualitas udara di Jakarta berdasarkan parameter AQI dari berbagai stasiun peman-tauan dalam periode waktu tertentu. Penentuan jumlah kluster optimal dilakukan menggunakan metode Elbow, sedangkan evaluasi kualitas kluster dilakukan menggunakan Silhouette Score. Hasil penelitian ini diharapkan dapat memberikan pemahaman yang lebih komprehensif terhadap pola kualitas udara, mengidentifikasi zona polusi yang homogen, serta mendukung pengambilan keputusan berbasis data dalam upaya mitigasi pencemaran udara di lingkungan perkotaan.

2. TINJAUAN PUSTAKA

2.1. Kualitas Udara dan Air Quality Index (AQI)

Kualitas udara perkotaan merupakan isu krusial dalam pembangunan berkelanjutan dan kesehatan masyarakat, terutama di kota besar seperti Jakarta yang mengalami peningkatan pencemaran akibat urbanisasi, pertumbuhan populasi, dan aktivitas industri [28]. Polutan utama seperti PM_{2.5}, PM₁₀, SO₂, NO₂, CO, dan O₃ sering melampaui ambang batas yang direkomendasikan oleh WHO dan berdampak pada meningkatnya risiko penyakit pernapasan serta kardiovaskular [34]. Selain itu, kualitas udara dipengaruhi oleh faktor temporal dan meteorologis seperti suhu, kelembapan, dan curah hujan, yang menunjukkan adanya dinamika perubahan polusi dari waktu ke waktu [33].

Di sisi lain, aspek spasial menunjukkan bahwa polusi udara bersifat lintas wilayah, terutama di kawasan aglomerasi seperti Jabodetabek, sehingga memerlukan pendekatan analisis spatio-temporal [30]. Meskipun metode K-Means telah banyak digunakan untuk mengelompokkan kualitas udara, sebagian penelitian masih terbatas pada wilayah atau periode tertentu dan belum mengkaji keterkaitan antar lokasi serta dinamika temporal secara komprehensif [8]. Oleh karena itu, penelitian ini mengusulkan analisis kluster spatio-temporal berbasis multi-stasiun dan multi-periode untuk memperoleh pemahaman yang lebih menyeluruh terkait pola dan distribusi kualitas udara di Jakarta.

2.2. Analisis Spatio-Temporal pada Data Lingkungan

Analisis spatio-temporal merupakan pendekatan yang mempertimbangkan dimensi ruang dan waktu secara simultan untuk memahami variasi fenomena lingkungan [19]. Dalam konteks kualitas udara, analisis ini memungkinkan identifikasi perubahan konsentrasi polutan dari waktu ke waktu dan perbedaan antar wilayah [39]. Pendekatan ini sering digunakan untuk:

- Mengidentifikasi pola distribusi polutan;
- Menentukan sumber utama polusi; dan
- Membuat prediksi spasial berbasis waktu (spatial forecasting).

Penelitian-penelitian terkini juga mengintegrasikan machine learning dengan data spasial untuk mengenali pola yang kompleks [4].

2.3. Konsep dan Prinsip K-Means Clustering

Algoritma K-Means merupakan salah satu metode unsupervised learning yang banyak digunakan untuk analisis kluster pada data numerik [12]. Prinsip dasar algoritma ini adalah meminimalkan jumlah kuadrat jarak antara setiap titik data dengan pusat kluster (centroid).

Algoritma K-Means clustering digunakan untuk mengelompokkan pelanggan berdasarkan pola penjualan, dengan prinsip meminimalisasi jarak antar data terhadap centroid cluster [22].

Secara matematis, fungsi objektif K-Means dirumuskan sebagai berikut:

$$J = \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \quad \dots\dots\dots (1)$$

Keterangan :

J : fungsi objektif (objective function) K-Means yang diminimalkan,

K : jumlah kluster yang ditentukan,

C_i : himpunan data yang termasuk ke dalam kluster ke-*i*,

X_j : data ke-*j* yang berada pada kluster C_i,

μ_i : centroid (rata-rata) dari kluster ke-*i*,

$\|x_j - \mu_i\|^2$: jarak Euclidean kuadrat antara data x_j dan centroid kluster μ_i.

Di mana *k* adalah jumlah kluster, x_j merupakan titik data ke-*j*, C_i adalah kluster ke-*i*, dan μ_i merupakan centroid dari kluster C_i. Proses K-Means terdiri dari empat tahap utama: inisialisasi centroid, penetapan titik ke kluster terdekat, pembaruan centroid, dan iterasi hingga konvergen [24]. Dalam penelitian kualitas udara, K-Means sering digunakan untuk:

- Mengelompokkan wilayah berdasarkan tingkat pencemaran [28].
- Mengidentifikasi zona kualitas udara homogen [40] dan
- Menganalisis pola temporal fluktuasi polutan [6].

Kelebihan K-Means adalah kesederhanaan dan efisiensi komputasi, namun metode ini memiliki kelemahan yaitu sensitif terhadap nilai awal dan asumsi bentuk kluster yang spherical [21].

Meskipun demikian, K-Means tetap relevan digunakan dalam penelitian ini karena tujuan analisis adalah mengidentifikasi pola dan zonasi kualitas udara spatio-temporal secara eksploratif, bukan untuk mendeteksi anomali atau kluster berbasis kepadatan.

2.4. Integrasi K-Means dengan Analisis Spatio-Temporal

Penggabungan clustering dengan analisis spasial dan temporal banyak digunakan untuk mendeteksi fenomena lingkungan kompleks [41]. Dalam konteks kualitas udara, pendekatan spatio-temporal K-Means diterapkan untuk menilai variasi antar lokasi

pengukuran sekaligus pergeseran waktu [32]. Metode ini bekerja dengan dua skema:

- Klasterisasi berdasarkan lokasi (spatial clustering), untuk memetakan wilayah dengan karakteristik polusi serupa; dan
- Klasterisasi berdasarkan waktu (temporal clustering), untuk melihat pola perubahan kualitas udara. Integrasi keduanya memberikan wawasan lebih kaya dibanding pendekatan tunggal karena dapat menyoroti hubungan antar waktu dan wilayah yang berdekatan secara geografis [20].

2.5. Penelitian Terkait

Berbagai penelitian terdahulu telah dilakukan untuk mengkaji penerapan algoritma clustering, khususnya K-Means, dalam menganalisis pola kualitas udara perkotaan. Beberapa studi juga mengombinasikan K-Means dengan metode lain, seperti Principal Component Analysis (PCA), DBSCAN, maupun algoritma pembelajaran mesin lainnya, guna meningkatkan akurasi dan interpretabilitas hasil klasterisasi. Berikut adalah table penelitian terdahulu:

Table 1. Penelitian Terdahulu

No.	Judul Penelitian	Metode	Hasil Pdednelitian
1.	Zhang et al. (2022). Spatio-Temporal Clustering of PM _{2.5} Concentrations in Chinese Cities – Science of the Total Environment	K-Means, PCA	Menemukan 4 klaster utama polusi dengan pola musiman signifikan
2.	Rahman et al. (2021). Air Quality Grouping Using Unsupervised Learning in Dhaka – Environmental Science and Pollution Research	K-Means	Menghasilkan tiga zona polusi dominan berdasarkan PM _{2.5} dan CO
3.	Tang et al. (2022). Temporal Variation and Clustering of Air Pollutants in Asian Cities – Environmental Research Letters	K-Means, DBSCAN	Kombinasi algoritma meningkatkan akurasi identifikasi periode polusi.
4.	Gupta & Singh (2023). Machine Learning for Urban Air Pollution Pattern Analysis – Environmental Modelling & Software	K-Means, Random Forest	K-Means efektif sebagai pra-proses segmentasi sebelum klasifikasi.
5.	Li et al. (2024). Spatio-Temporal Dynamics of AQI in Southeast Asia – Sustainable Cities and Society	K-Means, Spatial Autocorrelation	Menunjukkan efek spasial lintas-wilayah di kota-kota besar Asia Tenggara.
6.	Liu et al. (2023). Urban Air Quality Spatio-Temporal Clustering with Deep Learning – IEEE Access	K-Means, Autoencoder	Integrasi K-Means dan deep feature learning meningkatkan stabilitas klaster

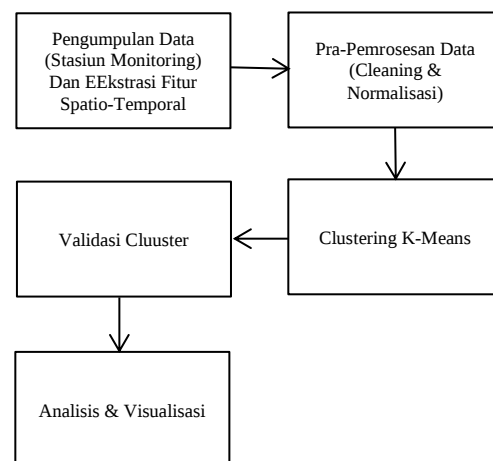
Tabel tersebut menampilkan perbandingan judul penelitian, metode yang digunakan, serta hasil utama yang diperoleh, sehingga dapat menjadi dasar dalam menentukan posisi dan kontribusi penelitian ini terhadap studi-studi sebelumnya.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis data-driven analysis dengan metode machine learning, khususnya unsupervised learning menggunakan algoritma K-Means clustering. Pendekatan ini dipilih karena data kualitas udara berupa data numerik kontinu yang dianalisis secara objektif menggunakan pemodelan matematis dan statistik [6]. Metode clustering digunakan untuk mengidentifikasi pola laten dalam data yang tidak memiliki label, serta mampu mengelompokkan data mul-tidimensi secara efektif. Selain itu, penelitian ini mengadopsi pendekatan spa-tio-temporal dengan mengintegrasikan dimensi ruang dan waktu secara simultan [5].

Penelitian dilakukan pada wilayah DKI Jakarta yang mencakup Jakarta Pusat, Jakarta Utara, Jakarta Selatan, Jakarta Timur, dan Jakarta Barat dengan rentang wak-tu 2020–2024. Data yang digunakan merupakan data sekunder yang diperoleh dari Dinas Lingkungan Hidup DKI Jakarta, BMKG, dan platform OpenAQ. Parameter kualitas udara mengacu pada Air

Quality Index (AQI), yaitu PM_{2.5}, PM₁₀, SO₂, NO₂, CO, dan O₃ [1]. Data berbentuk time-series multivariat dengan informasi koordinat geografis. Tahapan penelitian secara umum ditunjukkan sebagai berikut :



Adapun tahapan penelitian adalah sebagai berikut:

- Pengumpulan data kualitas udara multi-parameter dari berbagai stasiun monitoring.
- Pra-pemrosesan data (data cleaning, normalisasi, dan penanganan missing values).
- Pembentukan fitur spatio-temporal.

- d. Pemodelan menggunakan algoritma K-Means.
- e. Evaluasi model menggunakan metode Elbow dan Silhouette Score.
- f. Visualisasi dan interpretasi hasil.

3.1. Pra-pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk memastikan kualitas data sebelum dilakukan analisis. Tahapan ini meliputi data cleaning, normalisasi, dan feature engineering.

a. Data Cleaning

Table 2. Proses Data Cleaning

Langkah	Deskripsi
Menghapus duplikasi data	Menghilangkan data yang muncul lebih dari satu kali
Penyelarasan interval waktu	Mengubah data hourly rata-rata daily
Penanganan missing value	Menggunakan metode linear interpolation

Tahap ini bertujuan untuk memastikan data bersih, konsisten, dan memiliki resolusi temporal yang seragam. Metode linear interpolation digunakan karena efektif dalam menangani data hilang pada time-series kontinu [3].

b. Normalisasi Data

Table 3. Normalisasi Data

Teknik	Rumus	Alasan Penggunaan
Min-Max Normalization	$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	Menyetarakan skala fitur agar tidak terjadi bias

Normalisasi dilakukan karena setiap parameter memiliki rentang nilai yang berbeda. Metode Min-Max digunakan untuk mengubah skala data ke rentang 0–1 sehingga meningkatkan performa algoritma berbasis jarak seperti K-Means [9]

c. Feature Engineering

Table 4. Feature Engineering

Kategori Fitur	Komponen
Temporal	Bulan, tahun
Spasial	Latitude, longitude
Kualitas udara	PM _{2.5} , PM ₁₀ , SO ₂ , NO ₂ , CO, O ₃

Feature engineering dilakukan untuk memperkaya struktur data agar mampu merepresentasikan pola spasial dan temporal secara simultan.

3.2. Pembentukan spasio temporal

Pembentukan fitur dilakukan dengan menggabungkan:

- a. Fitur temporal (bulan dan tahun),
- b. Fitur spasial (latitude dan longitude),
- c. Parameter kualitas udara (AQI).

Tujuan tahap ini adalah agar model mampu menangkap hubungan antarwilayah serta dinamika perubahan kualitas udara dari waktu ke waktu [5]

3.3. Pemodelan k-means clustering

Algoritma K-Means digunakan untuk mengelompokkan data berdasarkan fitur spatio-temporal yang telah dibentuk. Data input meliputi parameter kualitas udara, fitur temporal, dan koordinat geografis.

Proses clustering dilakukan menggunakan jarak Euclidean dan berjalan secara iteratif hingga konvergen. Fungsi objektif K-Means adalah meminimalkan nilai Sum of Squared Error (SSE) antara data dan centroid kluster [8].

Penentuan jumlah kluster optimal dilakukan menggunakan:

- a. Elbow Method
- b. Silhouette Score

Output dari proses ini berupa label kluster yang merepresentasikan pola kualitas udara.

3.4. Evaluasi model

Evaluasi dilakukan untuk mengukur kualitas hasil clustering:

- a. Elbow Method digunakan untuk menentukan jumlah kluster optimal berdasarkan nilai SSE.
- b. Silhouette Score digunakan untuk mengukur kohesi dan separasi antar kluster [9].

Penggunaan kedua metode ini bertujuan meningkatkan validitas hasil clustering

3.5. Visualisasi Dan interpretasi

Hasil clustering divisualisasikan dalam beberapa bentuk:

- a. Peta spasial distribusi kluster
- b. Grafik temporal perubahan polutan
- c. Ringkasan karakteristik tiap kluster

4. HASIL DAN PEMBAHASAN

4.1. Pemahaman Data

Pada tahap ini dilakukan pemahaman terhadap dataset yang digunakan dalam penelitian. Dataset yang digunakan merupakan data kualitas udara di Jakarta yang terdiri dari beberapa parameter utama Air Quality Index (AQI), yaitu PM_{2.5}, PM₁₀, SO₂, NO₂, CO, dan O₃. Selain itu, dataset juga memuat informasi waktu (tanggal) dan lokasi (stasiun) yang digunakan dalam pembentukan fitur spatio-temporal.

	tanggal;stasiun;pm25;pm10;so2;co;o3;no2;max;critical;kategori
0	01/01/2010;DKI1 (Bunderan HI);;60.0;4.0;73.0;2...
1	02/01/2010;DKI1 (Bunderan HI);;32.0;2.0;16.0;3...
2	03/01/2010;DKI1 (Bunderan HI);;27.0;2.0;19.0;2...
3	04/01/2010;DKI1 (Bunderan HI);;22.0;2.0;16.0;1...
4	05/01/2010;DKI1 (Bunderan HI);;25.0;2.0;17.0;1...

Gambar 1. Struktur Dataset

Proses awal dilakukan dengan membaca dataset menggunakan Google Colab serta mengidentifikasi struktur data seperti jumlah baris, jumlah kolom, dan tipe data.

4.2. Pembersihan Data

Tahap ini dilakukan untuk memastikan data bebas dari kesalahan dan siap digunakan dalam proses analisis. Penanganan missing value dilakukan dengan metode pengisian nilai rata-rata pada setiap parameter numerik. Selain itu, dilakukan perbaikan format data terutama pada kolom tanggal.

	0
tanggal	0
stasiun	0
pm25	0
pm10	0
so2	0
co	0
o3	0
no2	0
max	0
critical	2
categori	0
bulan	0
tahun	0
stasiun_kode	0
cluster	0

dtype: int64

Gambar 1. Hasil Pembersihan Data

4.3. Transformasi Data

Pada tahap ini dilakukan pembentukan fitur tambahan berupa Bulan dan tahun (temporal) dan Kode stasiun (spasial). Transformasi ini bertujuan agar model mampu menangkap pola perubahan kualitas udara berdasarkan waktu dan lokasi

	tanggal	bulan	tahun	stasiun_kode
0	2010-01-01	1	2010	0
1	2010-02-01	2	2010	0
2	2010-03-01	3	2010	0
3	2010-04-01	4	2010	0
4	2010-05-01	5	2010	0

Gambar 3. Transformasi Data

4.4. Normalisasi Data

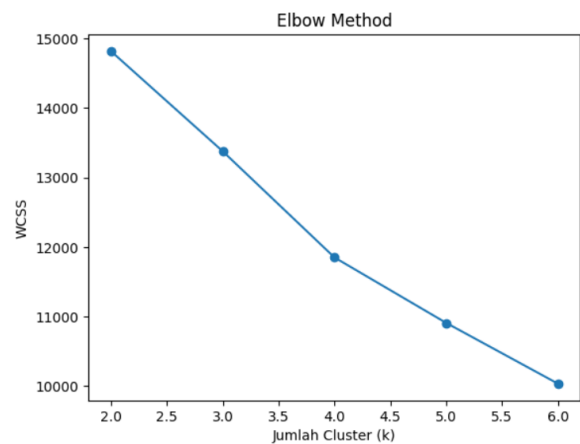
Normalisasi dilakukan menggunakan metode StandardScaler untuk menyamakan skala antar variabel, sehingga tidak terjadi dominasi fitur tertentu dalam proses clustering.

```
array([[ 1.10155294e-15,  2.13060835e-01, -1.18655753e+00,
        -3.46182124e-02,  4.47366171e+00, -9.79305590e-01,
        -1.59325501e+00, -1.80602464e+00],
       [ 1.10155294e-15, -1.23041257e+00, -1.31682777e+00,
        -5.66077407e-01, -3.72279492e-01, -8.61682088e-01,
        -1.30357228e+00, -1.80602464e+00],
       [ 1.10155294e-15, -1.48817568e+00, -1.31682777e+00,
        -5.66077407e-01, -1.17229955e-01, -1.11653301e+00,
        -1.01388955e+00, -1.80602464e+00],
       [ 1.10155294e-15, -1.74593879e+00, -1.31682777e+00,
        -8.84952924e-01, -3.72279492e-01, -1.21455259e+00,
        -7.24206824e-01, -1.80602464e+00],
       [ 1.10155294e-15, -1.59128092e+00, -1.31682777e+00,
        -6.72369246e-01, -2.87262980e-01, -1.21455259e+00,
        -4.34524095e-01, -1.80602464e+00]])
```

Gambar 4. Hasil Normalisasi Data

4.5. Penentuan Jumlah Cluster

Penentuan jumlah cluster optimal dilakukan menggunakan metode Elbow dengan melihat hubungan antara jumlah cluster (k) dan nilai WCSS.



Gambar 5. Grafik Elbow Method

Berdasarkan grafik Elbow, terlihat bahwa penurunan nilai WCSS terjadi secara signifikan pada beberapa nilai awal, kemudian mulai melandai. Hal ini menunjukkan bahwa jumlah cluster optimal berada pada rentang tertentu.

4.6. Evaluasi Clustering (Silhouette Score)

Evaluasi dilakukan menggunakan Silhouette Score untuk mengukur kualitas cluster.

Table 5. Nilai Silhouette Score

Jumlah Cluster (K)	Silhouette Score
2	0.1994
3	0.1405
4	0.1685
5	0.1529
6	0.1576

Berdasarkan hasil evaluasi, nilai tertinggi diperoleh pada k=2. Namun, penelitian ini menggunakan k=3 karena memberikan interpretasi yang lebih representatif terhadap kategori kualitas udara.

4.7. Hasil Clustering K-Means

Berdasarkan hasil evaluasi, model K-Means dibangun dengan jumlah cluster sebanyak 3. Setiap data kemudian diberikan label cluster yang menunjukkan keanggotaan data terhadap kelompok tertentu.

	tanggal	stasiun	pm25	pm10	so2	co	o3	no2	max	critical	categori	bulan	tahun	stasiun_kode	cluster
0	2010-01-01	DK01 (Bunderan HI)	87.861204	60.0	4.0	73.0	27.0	14.0	73.0	CO	SEDANG	1	2010	0	2
1	2010-02-01	DK01 (Bunderan HI)	87.861204	32.0	2.0	16.0	33.0	9.0	33.0	O3	BAIK	2	2010	0	1
2	2010-03-01	DK01 (Bunderan HI)	87.861204	27.0	2.0	19.0	20.0	9.0	27.0	PM10	BAIK	3	2010	0	1
3	2010-04-01	DK01 (Bunderan HI)	87.861204	22.0	2.0	16.0	15.0	6.0	22.0	PM10	BAIK	4	2010	0	1
4	2010-05-01	DK01 (Bunderan HI)	87.861204	25.0	2.0	17.0	15.0	8.0	25.0	PM10	BAIK	5	2010	0	1

Gambar 6 Hasil Clustering Data

4.8. Analisis Karakteristik Cluster

Table 5. Rata-rata Parameter Tiap Cluster

Cluster	PM2.5	PM10	SO2	NO2	CO	O3
0	98.09	62.05	42.14	24.52	17.24	39.69
1	87.90	63.67	14.46	10.99	22.07	108.90
2	82.06	42.66	20.67	12.74	20.03	58.08

- Cluster 0 (Kualitas Udara Buruk)
Memiliki nilai PM_{2.5}, SO₂, dan NO₂ yang tinggi, menunjukkan tingkat pencemaran yang signifikan.
- Cluster 1 (Dominasi Polusi Ozon)
Ditandai dengan nilai O₃ yang sangat tinggi, menunjukkan adanya polusi sekunder akibat reaksi fotokimia.
- Cluster 2 (Kualitas Udara Lebih Baik)
Memiliki nilai parameter polutan yang relatif lebih rendah dibandingkan cluster lainnya.

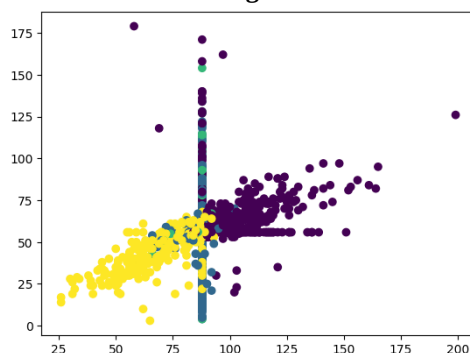
4.9. Distribusi Data Tiap Cluster

Table 6. Distribusi Cluster

Cluster	Jumlah	Persentase
0	437	20.01%
1	969	44.37%
2	778	35.62%

Distribusi data menunjukkan bahwa cluster 1 memiliki persentase terbesar, yang mengindikasikan dominasi polusi ozon dalam kualitas udara Jakarta.

4.10. Visualisasi Clustering



Gambar 7. Visualisasi Clustering (PM2.5 vs PM10)

Visualisasi menunjukkan adanya pola pengelompokan data berdasarkan parameter PM_{2.5} dan PM₁₀, meskipun masih terdapat beberapa tumpang tindih antar cluster.

Hasil penelitian menunjukkan bahwa algoritma K-Means mampu mengelompokkan data kualitas udara berdasarkan kemiripan karakteristik parameter polutan. Pola yang terbentuk menunjukkan adanya perbedaan yang cukup jelas antara cluster dengan tingkat polusi tinggi, sedang, dan rendah.

Dominasi cluster dengan konsentrasi ozon tinggi menunjukkan bahwa kualitas udara di Jakarta tidak hanya dipengaruhi oleh emisi langsung, tetapi juga oleh reaksi fotokimia di atmosfer. Selain itu,

keberadaan cluster dengan tingkat polusi tinggi menunjukkan adanya kontribusi signifikan dari aktivitas kendaraan dan industri. Meskipun nilai Silhouette Score relatif rendah, metode K-Means tetap mampu memberikan gambaran pola distribusi kualitas udara yang cukup representatif..

5. KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa metode K-Means mampu mengelompokkan data kualitas udara di Jakarta secara efektif berdasarkan parameter Air Quality Index (AQI) yang meliputi PM_{2.5}, PM₁₀, SO₂, NO₂, CO, dan O₃, serta mempertimbangkan aspek spatio-temporal. Berdasarkan evaluasi menggunakan Silhouette Score, nilai tertinggi diperoleh pada k=2 sebesar 0,1994, namun penelitian ini menggunakan k=3 karena memberikan interpretasi yang lebih representatif terhadap kategori kualitas udara, yaitu baik, sedang, dan buruk. Hasil clustering menghasilkan tiga kelompok utama, di mana cluster 0 menunjukkan kondisi kualitas udara buruk dengan dominasi PM_{2.5}, SO₂, dan NO₂, cluster 1 didominasi oleh tingginya konsentrasi ozon (O₃) yang mengindikasikan polusi sekunder, dan cluster 2 merepresentasikan kondisi kualitas udara yang relatif lebih baik.

Distribusi data menunjukkan bahwa cluster 1 memiliki persentase terbesar sebesar 44,37%, diikuti oleh cluster 2 sebesar 35,62%, dan cluster 0 sebesar 20,01%. Hal ini mengindikasikan bahwa sebagian besar kondisi kualitas udara di Jakarta dipengaruhi oleh polusi sekunder, khususnya ozon, yang terbentuk melalui reaksi fotokimia di atmosfer. Selain itu, nilai Silhouette Score yang relatif rendah menunjukkan adanya tumpang tindih antar cluster, yang mencerminkan kompleksitas dan dinamika kualitas udara di wilayah perkotaan, sehingga batas antar kategori tidak selalu terpisah secara tegas.

Berdasarkan hasil tersebut, penelitian selanjutnya disarankan untuk menggunakan metode clustering lain seperti DBSCAN atau Hierarchical Clustering untuk meningkatkan kualitas pengelompokan. Selain itu, penambahan variabel eksternal seperti faktor meteorologi (suhu, kelembaban, dan kecepatan angin) serta penggunaan metode prediksi berbasis time series seperti LSTM dapat memberikan analisis yang lebih komprehensif. Hasil penelitian ini juga diharapkan dapat menjadi bahan pertimbangan bagi pemerintah dalam merumuskan kebijakan pengendalian pencemaran udara, khususnya dalam mengurangi emisi prekursor ozon seperti NO_x dan senyawa organik volatil (VOC).

DAFTAR PUSTAKA

- [1] A. Hidayat, R. Pratama, and B. Susilo, "Air quality clustering in Jakarta using K-means algorithm," *Journal of Engineering and Applied Science*, vol. 41, no. 3, pp. 55–64, 2023.
- [2] A. Jain and R. Dubey, "An updated review on the applications of K-means algorithm,"

- International Journal of Scientific Research in Computer Science*, vol. 9, no. 4, pp. 98–105, 2020.
- [3] A. Rahman, et al., “Robust air quality preprocessing methods in time series modeling,” 2022.
 - [4] Chen, et al., “Hybrid spatial-temporal framework for air quality prediction,” 2023.
 - [5] F. Pedregosa, et al., “Scikit-learn: Machine learning in Python,” 2022.
 - [6] Gupta and S. Singh, “Machine learning for urban air pollution pattern analysis,” *Environmental Modelling & Software*, vol. 165, 105678, 2023.
 - [7] H. Wang and J. Xu, “Impact of feature scaling on clustering accuracy in environmental data,” 2024.
 - [8] Hidayat, A. Pratama, R., and B. Susilo, “Air quality clustering in Jakarta using K-means algorithm,” *Journal of Engineering and Applied Science*, vol. 41, no. 3, pp. 55–64, 2023.
 - [9] J. Lelieveld, J. Evans, and A. Schneider, “Air pollution and public health impacts in Southeast Asia,” *The Lancet Planetary Health*, vol. 7, no. 4, pp. e322–e334, 2023.
 - [10] J. Park and S. Lee, “Comparative study of clustering methods for air quality monitoring networks,” *Environmental Informatics*, vol. 29, no. 1, pp. 14–26, 2023.
 - [11] J. Xiao, Y. Huang, and F. Li, “Spatio-temporal variations of urban air pollution in Beijing,” *Atmospheric Environment*, vol. 254, 118401, 2021.
 - [12] Jain and R. Dubey, “An updated review on the applications of K-means algorithm,” *International Journal of Scientific Research in Computer Science*, vol. 9, no. 4, pp. 98–105, 2020.
 - [13] K. Nguyen and M. Do, “Evaluating K-means clustering performance for air pollution monitoring in Ho Chi Minh City,” *Environmental Data Science*, vol. 5, no. 3, pp. 299–310, 2023.
 - [14] L. Chen, et al., “Hybrid spatial-temporal framework for air quality prediction,” 2023.
 - [15] L. Han, C. Zhou, and T. Wong, “Dynamic cluster analysis of urban air quality in smart cities,” *Journal of Cleaner Production*, vol. 422, 138624, 2024.
 - [16] L. Tang, X. Wang, and D. Chen, “Temporal variation and clustering of air pollutants in Asian cities,” *Environmental Research Letters*, vol. 17, no. 6, 064012, 2022.
 - [17] L. Zhang and Y. Liu, “GIS-based clustering of AQI using K-means for smart urban monitoring,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 17, 10822, 2022.
 - [18] L. Zhao, H. Pan, and Z. Ren, “Spatial-temporal clustering for environmental big data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 12, pp. 4461–4473, 2021.
 - [19] Li, et al., “Handling missing environmental data using interpolation methods,” 2023.
 - [20] Liu, Z. Yang, and S. Shaw, “Urban air quality spatio-temporal clustering with deep learning,” *IEEE Access*, vol. 11, pp. 9012–9026, 2023.
 - [21] MacQueen, J., et al., “Rethinking centroid sensitivity in modern K-means optimization,” *Journal of Computational Algorithms*, vol. 18, no. 2, pp. 44–59, 2020.
 - [22] Manek, S. Faisal, and B. Priyatna, “Penerapan K-means clustering untuk mengelompokkan pelanggan berdasarkan data penjualan ayam,” *Techno Xplore: Jurnal Ilmu Komputer dan Teknologi Informasi*, vol. 3, no. 2, pp. 85–92, 2018.
 - [23] OpenAQ, “OpenAQ air quality public dataset API documentation,” 2024.
 - [24] Pedregosa, F, et al., “Scikit-learn documentation version 1.2: Machine learning in Python,” 2022.
 - [25] Q. Liu, Z. Yang, and S. Shaw, “Urban air quality spatio-temporal clustering with deep learning,” *IEEE Access*, vol. 11, pp. 9012–9026, 2023.
 - [26] R. Gupta and S. Singh, “Machine learning for urban air pollution pattern analysis,” *Environmental Modelling & Software*, vol. 165, 105678, 2023.
 - [27] R. Siregar, S. Abdullah, and E. Riyadi, “Spatial interdependence of air pollution in the Jabodetabek metropolitan area,” *Sustinere: Journal of Environment and Sustainability*, vol. 7, no. 1, pp. 55–69, 2023.
 - [28] Rahman, A., et al., “Robust air quality preprocessing methods in time series modeling,” 2022..
 - [29] S. Utami and L. Sari, “Temporal analysis of PM₁₀ concentration in Jakarta and its relationship with meteorological parameters,” *Jurnal Lingkungan Hidup Indonesia*, vol. 44, no. 1, pp. 77–89, 2022.
 - [30] Siregar, S. Abdullah, and E. Riyadi, “Spatial interdependence of air pollution in the Jabodetabek metropolitan area,” *Sustinere: Journal of Environment and Sustainability*, vol. 7, no. 1, pp. 55–69, 2023.
 - [31] T. Zhang, R. Wang, and H. Xu, “Spatio-temporal clustering of PM_{2.5} concentrations in Chinese cities,” *Science of the Total Environment*, vol. 828, 154529, 2022.
 - [32] Tang, X. Wang, and D. Chen, “Temporal variation and clustering of air pollutants in Asian cities,” *Environmental Research Letters*, vol. 17, no. 6, 064012, 2022.
 - [33] Utami, S., and L. Sari, “Temporal analysis of PM₁₀ concentration in Jakarta and its relationship with meteorological parameters,” *Jurnal Lingkungan Hidup Indonesia*, vol. 44, no. 1, pp. 77–89, 2022..
 - [34] World Health Organization, *Global air quality guidelines: Particulate matter (PM_{2.5} and PM₁₀)*,

- ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: WHO Press, 2021.
- [35] X. Li, Y. Tan, and W. Chong, "Spatio-temporal dynamics of AQI in Southeast Asia," *Sustainable Cities and Society*, vol. 102, 104556, 2024.
- [36] X. Zhang, et al., "Spatial clustering of AQI data using integrated feature systems," 2024.
- [37] Y. Chen, H. Zhang, and L. Wu, "Hybrid spatial-temporal machine learning for air quality analysis," *Atmospheric Pollution Research*, vol. 14, no. 2, pp. 101–112, 2023.
- [38] Y. Chen, H. Zhang, and L. Wu, "Hybrid spatial-temporal machine learning for air quality analysis," *Atmospheric Pollution Research*, vol. 14, no. 2, pp. 101–112, 2023.
- [39] Zhang and Y. Liu, "GIS-based clustering of AQI using K-means for smart urban monitoring," *International Journal of Environmental Research and Public Health*, vol. 19, no. 17, 10822, 2022.
- [40] Zhang, R. Wang, and H. Xu, "Spatio-temporal clustering of PM_{2.5} concentrations in Chinese cities," *Science of the Total Environment*, vol. 828, 154529, 2022.
- [41] Zhao, H. Pan, and Z. Ren, "Spatial-temporal clustering for environmental big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 12, pp. 4461–4473, 2021