ANALISIS PENGELOMPOKAN PROVINSI DI INDONESIA BERDASARKAN JUMLAH DESA/KELURAHAN MENURUT JENIS PENCEMARAN LINGKUNGAN HIDUP TAHUN 2024 MENGGUNAKAN PCA DAN *K-MEANS CLUSTERING*

(*Clustering Analysis of Villages in Each Province of Indonesia Based on Types of Environmental Pollution in 2024 Using PCA and K-Means Clustering*)

Afifah Nisa Rahmadani[1], Revalina Putria Hidayat[2], Fikri Ahmad Riza[3], Ariel Muda Simanungkalit[4], Surya Puspita Sari[5], Magdalena Effendi[6]

Statistika, FSTI, Institut Teknologi Kalimantan, Balikpapan[1,2,3]
Teknik Mesin, FRTI, Institut Teknologi Kalimantan, Balikpapan[4]

E-mail: 16221032@student.itk.ac.id

ABSTRAK

Pencemaran lingkungan merupakan isu penting di Indonesia karena berdampak pada kesehatan masyarakat dan keberlanjutan ekosistem. Variasi tingkat pencemaran antarprovinsi menunjukkan perlunya analisis yang mampu menggambarkan pola spasial secara komprehensif. Penelitian ini bertujuan mengelompokkan 38 provinsi di Indonesia berdasarkan jumlah desa/kelurahan menurut jenis pencemaran lingkungan, meliputi pencemaran air, tanah, udara, serta wilayah yang tidak mengalami pencemaran pada tahun 2024. Data diperoleh dari publikasi resmi Badan Pusat Statistik (BPS) dan dianalisis menggunakan kombinasi *Principal Component Analysis* (PCA) sebagai tahap reduksi dimensi serta *K-Means Clustering* sebagai metode pengelompokan. Tahap awal dilakukan statistika deskriptif, transformasi logaritmik, dan standarisasi data untuk meningkatkan kualitas analisis. PCA menghasilkan dua komponen utama yang mampu menjelaskan 89,41% variasi total data. Penentuan jumlah klaster optimal menggunakan metode *Elbow* dan *Silhouette* menunjukkan bahwa dua klaster memberikan struktur pengelompokan yang paling memadai dibandingkan alternatif jumlah klaster lainnya (*Silhouette* = 0,40). Hasil klasterisasi mengidentifikasi dua kelompok utama, yaitu klaster dengan tingkat pencemaran tinggi yang didominasi provinsi di wilayah barat Indonesia, serta klaster pencemaran rendah yang banyak berada di wilayah timur. Temuan ini memberikan gambaran deskriptif berbasis wilayah mengenai distribusi pencemaran lingkungan antarprovinsi di Indonesia dan dapat menjadi dasar bagi pemerintah dalam merumuskan kebijakan pengendalian lingkungan yang lebih tepat sasaran, khususnya dalam pengelolaan limbah industri, transportasi, dan konservasi sumber daya alam.

Kata kunci: Analisis Spasial, Indonesia, *K-Means Clustering*, PCA, Pencemaran Lingkungan.

ABSTRACT

Environmental pollution is a critical issue in Indonesia due to its impact on public health and ecosystem sustainability. Variations in pollution conditions across provinces indicate the need for analyses that can comprehensively describe spatial patterns. This study aims to classify 38 provinces in Indonesia based on the number of villages and urban villages according to types of environmental pollution, including water, soil, air pollution, and areas without pollution, in 2024. The data were obtained from official publications of Statistics Indonesia (BPS). The analysis employed Principal Component Analysis (PCA) as a dimensionality reduction technique, followed by K-Means Clustering to group provinces with similar pollution characteristics. The initial analysis was supported by descriptive statistical exploration and data standardization. The PCA results show that two principal components explain 89.41% of the total data variance. The optimal number of clusters was determined using the Elbow method and Silhouette coefficient, indicating that a two-cluster solution provides the most appropriate clustering structure (Silhouette score = 0.40). The clustering results reveal differences in environmental pollution characteristics between provinces in western and eastern Indonesia. These findings provide an initial, area-based descriptive overview of environmental pollution distribution in Indonesia and can support regional environmental management and more targeted policy formulation.

***Keywords**: Environmental Pollution, Indonesia, K-Means Clustering, PCA, Spatial Analysis.*

PENDAHULUAN

Pencemaran lingkungan merupakan salah satu persoalan ekologis yang memiliki urgensi tinggi di Indonesia karena berdampak langsung terhadap kesehatan masyarakat, keseimbangan ekosistem, dan kualitas hidup secara keseluruhan. Pencemaran dapat terjadi pada berbagai komponen lingkungan seperti air, udara, maupun tanah, yang umumnya dipicu oleh meningkatnya aktivitas manusia, termasuk industrialisasi, transportasi, pertanian intensif, serta ekspansi kawasan perkotaan. Berdasarkan data Badan Pusat Statistik (BPS) tahun 2024, diketahui bahwa sejumlah desa dan kelurahan di Indonesia telah mengalami pencemaran dengan variasi tingkat keparahan. Pencemaran air umumnya berkaitan dengan pembuangan limbah tanpa pengolahan, pencemaran udara sering dikaitkan dengan emisi kendaraan dan pembakaran bahan bakar fosil, sedangkan pencemaran tanah berkaitan dengan penggunaan pestisida berlebih dan akumulasi limbah padat (Mendes dkk., 2024). Kondisi tersebut menunjukkan perlunya upaya pemetaan pola pencemaran secara terstruktur antarwilayah. Dalam penelitian ini, data pencemaran lingkungan pada tingkat desa dan kelurahan digunakan sebagai indikator agregat untuk menggambarkan karakteristik pencemaran lingkungan pada tingkat provinsi.

Perbedaan intensitas pencemaran antarprovinsi tidak hanya berkaitan dengan sumber pencemar, tetapi juga sering dikaitkan dengan kondisi geografis, tingkat pembangunan ekonomi, kepadatan penduduk, serta pola penggunaan lahan masing-masing wilayah. Hingga saat ini, kajian yang menguraikan pola kemiripan atau perbedaan karakteristik pencemaran antarprovinsi masih relatif terbatas. Pertanyaan mendasar muncul, apakah provinsi yang berdekatan secara geografis menunjukkan karakteristik pencemaran yang serupa atau justru berbeda akibat variasi aktivitas ekonomi dan tata kelola lingkungan (Saputra, 2024) sebagai pertanyaan konseptual yang menjadi dasar eksplorasi pola kemiripan wilayah secara deskriptif. Sebagai ilustrasi, pada aspek pencemaran udara, pada aspek pencemaran udara, Indonesia mencatat rata-rata konsentrasi PM2.5 sebesar 35,5 μg/m³ pada tahun 2024, yang berpotensi meningkatkan risiko gangguan pernapasan dan penyakit kardiovaskular (HEI, 2024). Fakta ini menegaskan bahwa pencemaran lingkungan bersifat multidimensional dan memerlukan pendekatan analitis yang komprehensif.

Beberapa penelitian sebelumnya telah menggunakan metode analisis cluster untuk mengelompokkan wilayah berdasarkan tingkat pencemaran lingkungan, dengan tujuan memahami pola distribusi pencemaran dan prioritas penanganannya. Azmi dkk. (2025) menunjukkan bahwa metode *K-Means*, *K-Medoids*, dan *Fuzzy C-Means* dapat membentuk kelompok wilayah dengan tingkat pencemaran rendah, sedang, dan tinggi. Namun, studi tersebut umumnya hanya menganalisis satu atau dua jenis pencemaran, seperti udara atau air, sehingga belum mencerminkan hubungan simultan antara pencemaran air, udara, dan tanah. Kesenjangan ini menjadi dasar penting dilakukannya penelitian ini, yaitu pengembangan model pengelompokan provinsi yang mempertimbangkan ketiga jenis pencemaran secara bersamaan untuk menghasilkan pengelompokan wilayah yang komprehensif dan informatif bagi pengelolaan lingkungan.

Selain itu, sebagian besar penelitian terdahulu menggunakan seluruh variabel pencemaran secara langsung dalam proses pengelompokan tanpa melakukan reduksi dimensi terlebih dahulu. Pendekatan ini berpotensi menimbulkan bias akibat korelasi tinggi antar variabel, misalnya antara kadar logam berat dan total padatan tersuspensi pada air, atau antara partikulat PM2.5 dan CO_2 pada udara. Untuk mengatasi hal tersebut, penelitian ini menerapkan *Principal Component Analysis* (PCA) sebagai tahap pra-pemrosesan. PCA mengekstraksi komponen utama yang mewakili sebagian besar keragaman data, sehingga menghilangkan redundansi antar variabel dan menyederhanakan struktur data tanpa kehilangan informasi penting.

Sebelum melakukan pengelompokan, penelitian ini juga menerapkan statistika deskriptif untuk memberikan gambaran umum data yang digunakan. Analisis ini mencakup penghitungan nilai pemusatan, penyebaran data, bentuk distribusi, serta visualisasi melalui grafik dan diagram untuk menunjukkan pola dasar data tanpa inferensi statistik (Yusuf Nalim & Salafudin Tarmudi, 2012). Setelah tahap PCA, analisis *cluster* diterapkan sebagai teknik untuk mengelompokkan provinsi berdasarkan tingkat kemiripan karakteristik (Fa'rifah & Pramesti, 2022). Provinsi yang memiliki karakteristik mirip akan ditempatkan dalam *cluster* yang sama, sedangkan provinsi dengan karakteristik berbeda akan

masuk *cluster* yang berbeda (Sitepu & Gultom, 2011). Dengan kombinasi PCA dan *K-Means*, klasterisasi provinsi diharapkan menjadi lebih stabil, representatif, dan bebas dari multikolinearitas.

Penelitian ini menggunakan metode Non-Hierarki *Clustering* yang dilanjutkan dengan algoritma *K-Means*. Meskipun metode hirarki seperti dendrogram dapat menunjukkan struktur bertingkat antar provinsi (Johnson dkk., 2011), K-Means dipilih karena algoritma ini lebih efisien untuk memperoleh kelompok yang stabil melalui proses iterasi. Penentuan jumlah cluster optimal dilakukan dengan metode *Elbow* yang mempertimbangkan penurunan nilai *Sum of Squared Errors* (Muningsih, 2018). Algoritma kemudian mengelompokkan provinsi berdasarkan kedekatan setiap provinsi dengan *centroid* hingga nilai pusat *cluster* tidak berubah lagi (Irwansyah, 2015; Witten, 2015). Metode ini bekerja efektif pada *dataset* berukuran besar, meskipun berpotensi terjebak pada *local optimum* (Suyanto, 2017).

Penelitian ini bertujuan untuk mengidentifikasi dan mengelompokkan provinsi di Indonesia berdasarkan karakteristik pencemaran air, udara, dan tanah pada tahun 2024, sehingga menghasilkan pemetaan pencemaran yang komprehensif dan informatif secara spasial. Hasil penelitian diharapkan menjadi bahan pertimbangan bagi pemerintah dan pemangku kepentingan dalam menyusun strategi pengendalian pencemaran yang lebih terarah sesuai kondisi lokal setiap provinsi, serta berkontribusi pada pengembangan kajian ilmiah di bidang lingkungan dan analisis spasial.

METODE

Sumber Data

Penelitian ini menggunakan data sekunder dari publikasi resmi Badan Pusat Statistik (BPS) berjudul "Banyaknya Desa/Kelurahan Menurut Jenis Pencemaran Lingkungan Hidup (Desa), tahun 2024." Data ini bersifat nasional dan mencakup 38 provinsi di Indonesia, sehingga mampu memberikan gambaran menyeluruh mengenai kondisi pencemaran lingkungan di tingkat desa dan kelurahan. Variabel yang dianalisis terdiri dari empat jenis, yaitu pencemaran air, tanah, udara, serta kategori tidak mengalami pencemaran. Seluruh data merupakan *cross-section*, karena hanya merepresentasikan kondisi pada periode tahun 2024.

Statistika Deskriptif

Analisis ini digunakan untuk memberikan gambaran awal mengenai kondisi pencemaran di setiap provinsi. Tahap ini mencakup perhitungan nilai rata-rata, penyebaran data, serta penyajian dalam bentuk tabel dan grafik, sehingga pola dasar pencemaran dapat lebih mudah dipahami (Yusuf Nalim & Salafudin Tarmudi, 2012).

Standarisasi Data

Standarisasi dilakukan menggunakan metode *z-score* agar semua variabel berada pada skala yang sebanding. Langkah ini penting untuk menghindari bias akibat perbedaan satuan atau skala nilai (Suyanto, 2017; Johnson, Wichern, & Sharma, 2011).

Principal Component Analysis

Principal Component Analysis (PCA) adalah teknik reduksi dimensi yang mengubah sekumpulan variabel yang saling berkorelasi menjadi sejumlah komponen baru yang tidak berkorelasi (orthogonal). Proses PCA didasarkan pada dekomposisi matriks kovarian dan analisis *eigenvalue–eigenvector* (Jolliffe, 2002). Komponen ini disusun berdasarkan jumlah variansi yang dijelaskan dari data asli (Jolliffe & Cadima, 2016).

PCA sangat efektif digunakan ketika variabel memiliki korelasi tinggi, karena dapat menekan multikolinearitas dan menjaga sebagian besar informasi penting. Dalam konteks data penyebab perceraian yang memiliki banyak variabel dengan korelasi kuat, PCA membantu memusatkan informasi menjadi beberapa komponen utama untuk mempermudah proses klasterisasi (Putra et al., 2021).

Langkah PCA secara matematis:

1. Standarisasi Variabel
   Untuk setiap variabel X:
$$Z = \frac{X-\mu}{\sigma} \quad (1)$$
2. Membentuk Matriks Kovarian
$$S = \frac{1}{n-1} Z^{\top} Z \quad (2)$$
3. Menghitung *Eigenvalue* dan *Eigenvector*
$$Se_k = \lambda_k e_k \quad (3)$$
   dimana:
   λ_k= *eigenvalue* (proporsi variansi)
   e_k= *eigenvector* (arah komponen)
4. Membentuk Komponen Utama
   Komponen utama ke-k dihitung menggunakan:
$$PC_k = Ze_k \quad (4)$$
5. Menentukan Banyak Komponen Berdasarkan Variansi
   Variansi yang dijelaskan oleh PC_k:
$$PVE_k = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} \quad (5)$$

Metode Elbow

Digunakan untuk menentukan jumlah klaster optimal pada *K-Means Clustering*. Titik optimal ditentukan dari perubahan nilai *Sum of Squared Errors (SSE)*, di mana penurunan mulai melandai (Maori & Evanita, 2023).

$$SSE = \Sigma_{k=1}^{k} \ \Sigma_{x_i \in ck} \ (x_i - \varphi_k)^2 \quad (6)$$

dimana:
Ck = K *cluster* yang terbentuk
k = banyak *cluster*
xi = data x pada fitur ke-i
φk = rata-rata K *cluster* pada nilai k (k=1,2,3,...,K)

Silhouette Score

Penilaian kualitas *cluster* menggunakan rumus *Silhouette* (Rousseeuw, 1987):
$$s(i) = \frac{b(i)-a(i)}{max\,(a(i),b(i))} \quad (7)$$

dimana:
$a(i)$ = rata-rata jarak objek i dengan anggota *cluster*-nya
$b(i)$ = jarak rata-rata objek i ke *cluster* terdekat lainnya

Nilai:
$s(i) \approx 1$: sangat cocok dengan klasternya
$s(i) \approx 0$: berada di batas antar klaster
$s(i) < 0$: salah pengelompokkan

K-Means Clustering

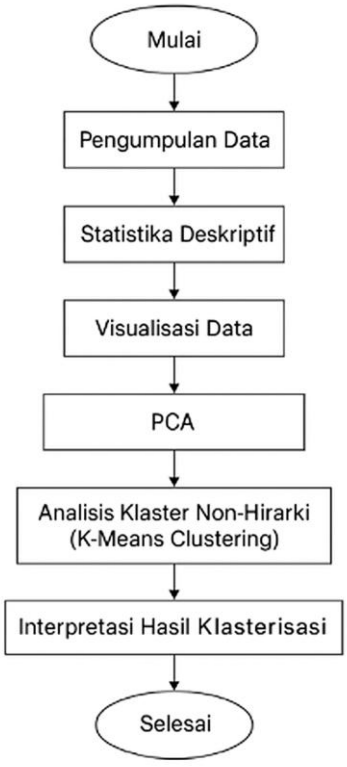
Metode *K-Means* digunakan untuk mengelompokkan provinsi berdasarkan kemiripan karakteristik pencemaran. Proses dilakukan secara iteratif hingga pusat klaster (*centroid*) stabil dan menghasilkan kelompok yang representatif (Irwansyah, 2015; Fa'rifah & Pramesti, 2022).

Prosedur Analisis Data

Langkah-langkah analisis dilakukan secara berurutan agar hasil penelitian lebih sistematis dan mudah dipahami. Berikut tahapan yang digunakan:

1. Pengumpulan Data
   Mengambil data sekunder dari publikasi BPS tahun 2024 tentang banyaknya desa/kelurahan menurut jenis pencemaran lingkungan hidup.
2. Statistika Deskriptif
   Memberikan gambaran awal tentang kondisi pencemaran di tiap provinsi melalui ukuran seperti rata-rata, maksimum, minimum, dan standar deviasi (Yusuf Nalim & Salafudin Tarmudi, 2012).
3. Visualisasi Data
   Data divisualisasikan dalam bentuk diagram dan peta tematik agar pola pencemaran antarprovinsi mudah dilihat (Saputra, 2024).
4. Analisis Komponen Utama (PCA)
   PCA digunakan untuk menyederhanakan variabel dengan cara mengubahnya menjadi beberapa komponen utama tanpa mengurangi informasi penting. Tahapan ini membantu mengurangi dimensi data sebelum dilakukan proses klasterisasi (Jolliffe & Cadima, 2016).
5. Analisis Klaster Non-Hirarki (*K-Means Clustering*)
   Metode *K-Means* digunakan untuk mengelompokkan provinsi berdasarkan kesamaan karakteristik pencemaran. Proses dilakukan secara iteratif hingga posisi pusat klaster (*centroid*) stabil (Irwansyah, 2015; Fa'rifah & Pramesti, 2022).
6. Interpretasi Hasil Klasterisasi
   Hasil klasterisasi diinterpretasikan untuk mengetahui kelompok provinsi dengan tingkat pencemaran tinggi, sedang, dan rendah.

Diagram Alir



Gambar 1. Diagram Alir.

HASIL DAN PEMBAHASAN

Statistik Deskriptif

Analisis statistik deskriptif dilakukan untuk memberikan gambaran awal mengenai distribusi data pencemaran di 38 provinsi di Indonesia. Hasil analisis menunjukkan bahwa variabel Tidak Ada Pencemaran memiliki nilai rata-rata tertinggi, yaitu sebesar 1.855,11 desa/kelurahan sedangkan variabel dengan rata-rata terendah adalah Pencemaran Tanah sebesar 24,92 desa/kelurahan. Kondisi ini menunjukkan bahwa jumlah desa/kelurahan yang tidak mengalami pencemaran relatif lebih besar dibandingkan dengan desa/kelurahan yang mengalami pencemaran, khususnya pencemaran tanah. Dari sisi variasi data, seluruh variabel menunjukkan nilai simpangan baku yang cukup besar, terutama pada variabel Tidak Ada Pencemaran dan Pencemaran Air, yang mengindikasikan adanya perbedaan karakteristik pencemaran antarprovinsi. Rentang nilai minimum dan maksimum yang cukup lebar pada masing-masing variabel juga menunjukkan ketimpangan distribusi pencemaran lingkungan di tingkat provinsi. Secara keseluruhan, hasil statistik deskriptif ini memberikan indikasi awal mengenai heterogenitas kondisi pencemaran lingkungan antarprovinsi di Indonesia, yang selanjutnya menjadi dasar dilakukannya analisis lanjutan menggunakan metode klasterisasi.

Tabel 1. Statistik Deskriptif Variabel Pencemaran Lingkungan.

Statistik	Pencemaran Air	Pencemaran Tanah	Pencemaran Udara	Tidak Ada Pencemaran
count	38	38	38	38
mean	289,97	24,92	125,11	1.855,11
std	345,99	28,41	139,19	1.806,32
min	7	0	0	157
25%	74	6	39,5	824,25
50%	171	15,5	72,5	1.126
75%	316,25	27	159,5	2.209,25
max	1.366	122	583	7.103

Transformasi Data

Untuk menstabilkan variansi data dan mengurangi pengaruh nilai ekstrem, seluruh variabel pencemaran ditransformasi menggunakan transformasi logaritmik natural (ln). Transformasi ini bertujuan untuk mengurangi tingkat kecondongan (*skewness*) data serta memperkecil perbedaan rentang nilai antarprovinsi. Hasil transformasi menunjukkan bahwa penyebaran data menjadi lebih homogen dibandingkan dengan data awal, sehingga lebih memenuhi asumsi untuk analisis multivariat selanjutnya.

Tabel 2 menyajikan hasil transformasi logaritmik pada variabel pencemaran air, tanah, udara, dan kategori tidak ada pencemaran untuk masing-masing provinsi. Transformasi ini dilakukan sebelum tahap deteksi *outlier* dan standarisasi data, agar evaluasi sebaran data serta perhitungan jarak pada tahap PCA dan *K-Means* dilakukan pada data yang telah memiliki variansi yang lebih stabil.

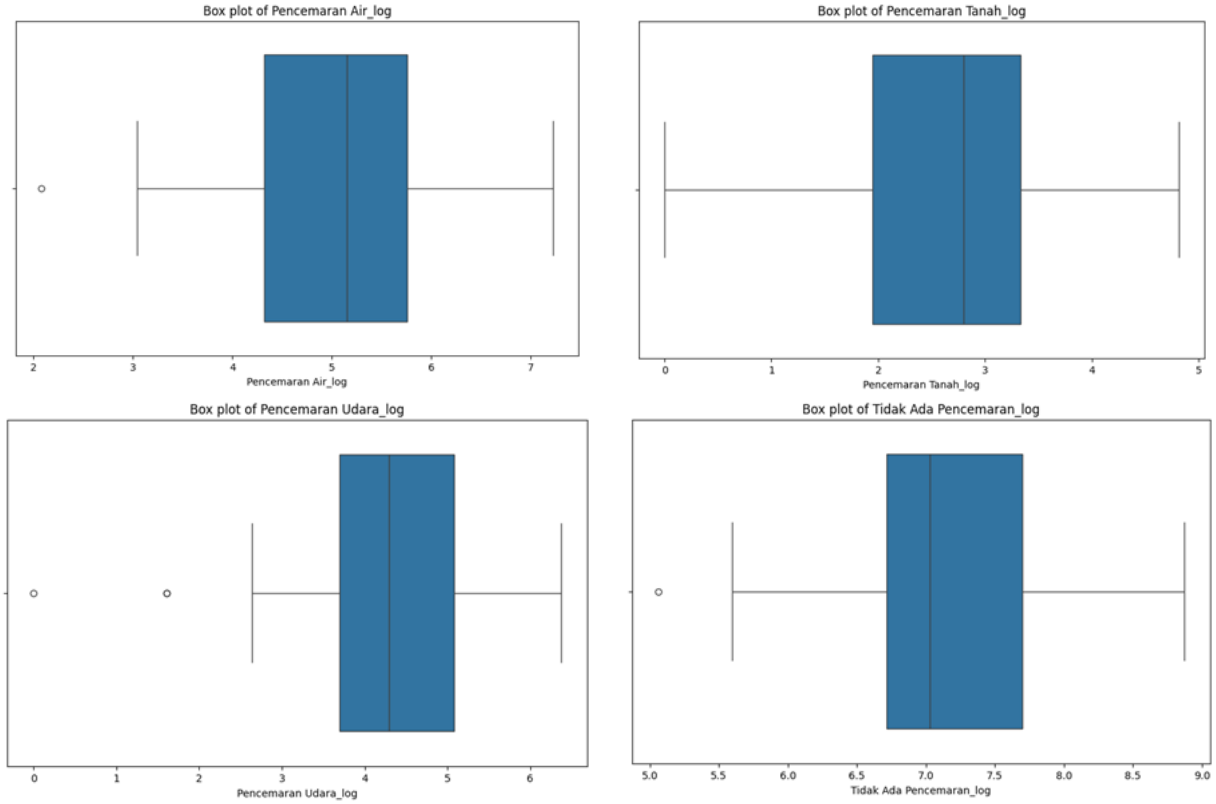
Tabel 2. Hasil Transformasi Logaritmik Variabel Pencemaran.

Provinsi	Air	Tanah	Udara	Tidak Ada Pencemaran	Air log	Tanah log	Udara log	Tidak Ada log
ACEH	308	14	168	6.088	5,733	2,708	5,13	8,714
SUMATERA UTARA	800	20	229	5.184	6,686	3,045	5,438	8,554
SUMATERA BARAT	238	27	158	950	5,476	3,332	5,069	6,858
RIAU	248	6	123	1.535	5,517	1,946	4,82	7,337
JAMBI	466	27	55	1.082	6,146	3,332	4,025	6,987

Deteksi Outlier

Gambar 2 menampilkan boxplot dari empat variabel pencemaran lingkungan yang telah melalui transformasi logaritmik, yaitu Pencemaran Air, Pencemaran Tanah, Pencemaran Udara, dan Tidak Ada Pencemaran. Visualisasi ini digunakan untuk mengidentifikasi keberadaan *outlier* serta mengevaluasi karakteristik sebaran data antarprovinsi setelah transformasi. Secara umum, variabel Pencemaran Air menunjukkan sebaran yang relatif lebar, yang mengindikasikan adanya variasi tingkat pencemaran air antarprovinsi, disertai dengan beberapa nilai ekstrem yang merepresentasikan provinsi dengan tingkat pencemaran air yang tinggi. Sebaliknya, variabel Pencemaran Tanah memiliki sebaran paling sempit dengan jumlah outlier yang terbatas, menunjukkan bahwa pencemaran tanah relatif rendah dan lebih merata di sebagian besar provinsi.

Variabel Pencemaran Udara memperlihatkan distribusi yang mendekati simetris, dengan keberadaan beberapa nilai ekstrem pada sisi atas distribusi, yang mencerminkan adanya provinsi dengan tingkat pencemaran udara lebih tinggi dibandingkan provinsi lainnya. Sementara itu, variabel Tidak Ada Pencemaran memiliki median yang relatif tinggi dan rentang data yang cukup luas, menggambarkan perbedaan yang cukup signifikan antarprovinsi dalam jumlah desa/kelurahan yang tidak mengalami pencemaran. Secara keseluruhan, hasil boxplot menunjukkan bahwa transformasi logaritmik mampu menyeimbangkan distribusi data, memperkecil pengaruh nilai ekstrem, serta meningkatkan kelayakan data untuk analisis lanjutan menggunakan metode PCA dan *K-Means*.



Gambar 2. Boxplot Variabel Pencemaran Lingkungan (Setelah Transformasi Logaritmik).

Standarisasi Data

Standarisasi data dilakukan untuk menyamakan skala antarvariabel agar tidak ada satu variabel yang mendominasi hasil analisis. Proses ini mengubah setiap variabel menjadi memiliki rata-rata nol dan simpangan baku satu, sehingga perbedaan satuan maupun rentang nilai antarvariabel tidak memengaruhi hasil perhitungan jarak pada tahap PCA dan *K-Means*. Berdasarkan Tabel 3, nilai hasil standarisasi pada seluruh variabel sudah berada di sekitar nol, yang menandakan bahwa data telah berada pada skala yang sebanding. Dengan demikian, standarisasi penting agar setiap variabel pencemaran memberikan kontribusi yang proporsional dalam proses analisis multivariat selanjutnya.

Tabel 3. Hasil Standarisasi Variabel Pencemaran.

Provinsi	Air_log	Tanah_log	Udara_log	Tidak Ada_log
ACEH	0,574	0,037	0,727	1,793
SUMATERA UTARA	1,393	0,318	0,993	1,610
SUMATERA BARAT	0,353	0,559	0,675	–0,326
RIAU	0,388	–0,600	0,461	0,221
JAMBI	0,929	0,559	–0,223	–0,177

Analisis Principal Component Analysis (PCA)

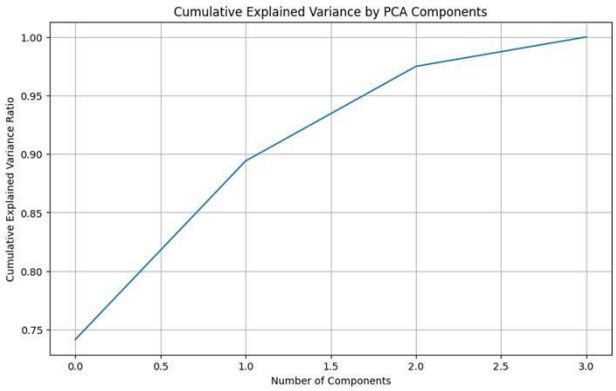
Analisis *Principal Component Analysis* (PCA) diterapkan untuk mereduksi dimensi dari empat variabel pencemaran lingkungan yang saling berkorelasi menjadi beberapa komponen utama yang saling bebas orthogonal. Tujuan dari proses ini adalah menyederhanakan struktur data tanpa menghilangkan sebagian besar informasi penting yang terkandung di dalam variabel asal, sehingga data lebih efisien untuk dianalisis pada tahap klasterisasi. Hasil analisis menunjukkan bahwa dari empat variabel awal terbentuk empat komponen utama dengan nilai *explained variance ratio* yang disajikan pada Tabel 4.

Tabel 4. Nilai *Explained Variance Ratio* Tiap Komponen PCA.

Komponen	*Explained Variance Ratio*
Komponen 1	0,7416
Komponen 2	0,1525
Komponen 3	0,0808
Komponen 4	0,0251

Dari tabel tersebut terlihat bahwa komponen pertama (PC1) memiliki kontribusi terbesar dalam menjelaskan keragaman data, yaitu sebesar 74,16%, diikuti oleh komponen kedua (PC2) sebesar 15,25%, sedangkan dua komponen lainnya hanya menjelaskan sebagian kecil variasi. Secara kumulatif, dua komponen pertama telah mampu menjelaskan 89,41% total variasi, sehingga keduanya dianggap cukup representatif untuk menggambarkan struktur utama data pencemaran antarprovinsi.

Visualisasi Cumulative Explained Variance pada Gambar 3 memperlihatkan bahwa kurva mulai melandai setelah komponen kedua, yang menunjukkan bahwa tambahan komponen berikutnya hanya memberikan peningkatan informasi yang terbatas. Oleh karena itu, dua komponen utama pertama dipilih untuk digunakan dalam analisis klaster berikutnya.



Gambar 3. Visualisasi *Cumulative Explained Variance* oleh Komponen PCA.

Berdasarkan tabel tersebut, komponen utama pertama (PC1) memiliki *loading* relatif tinggi pada variabel Pencemaran Air, Pencemaran Tanah, dan Pencemaran Udara. Hal ini menunjukkan bahwa PC1 merepresentasikan dimensi tingkat pencemaran lingkungan secara umum, di mana provinsi dengan skor PC1 tinggi cenderung memiliki tingkat pencemaran lingkungan yang lebih besar. Sementara itu, komponen kedua (PC2) didominasi oleh variabel Tidak Ada Pencemaran dengan nilai *loading* sebesar 0,889, yang mencerminkan dimensi kondisi lingkungan yang relatif bersih. Komponen ini berperan dalam membedakan provinsi dengan kondisi lingkungan yang lebih baik dari provinsi yang memiliki tingkat pencemaran tinggi. Berdasarkan hasil *explained variance* dan

loading factors, PC1 dan PC2 dipilih sebagai komponen utama yang paling representatif dalam menggambarkan variasi data pencemaran antarprovinsi. Kombinasi kedua komponen tersebut selanjutnya digunakan sebagai input pada tahap analisis *K-Means Clustering*, karena telah mampu menangkap pola utama data dengan kehilangan informasi yang minimal.

Tabel 5. *Loading Factors* (Kontribusi Variabel terhadap Komponen Utama PCA).

Variabel	PC1	PC2	PC3	PC4
Pencemaran Air_log	0,552	–0,215	0,076	0,802
Pencemaran Tanah_log	0,502	–0,391	–0,666	–0,388
Pencemaran Udara_log	0,522	–0,099	0,716	–0,454
Tidak Ada Pencemaran_log	0,413	0,889	–0,195	–0,027

Dari hasil tersebut, PC1 dan PC2 dipilih sebagai komponen terbaik untuk mewakili pola variasi utama data pencemaran antarprovinsi. Kombinasi keduanya digunakan sebagai input pada tahap analisis *K-Means Clustering*, karena sudah mampu menangkap karakteristik data secara menyeluruh dengan kerugian informasi yang minimal.

Analisis K-Means Clustering

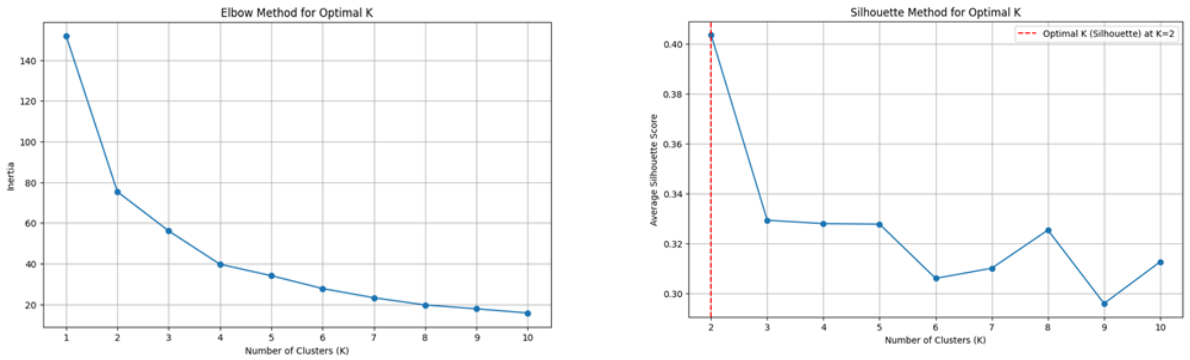
Penentuan Jumlah Klaster dengan Metode Elbow dan Silhouette

Penentuan jumlah klaster optimal dalam algoritma *K-Means* dilakukan menggunakan dua pendekatan, yaitu metode *Elbow* dan metode *Silhouette*. Metode *Elbow* bertujuan untuk mengamati perubahan nilai *inertia* (jumlah kuadrat jarak antar data terhadap pusat klasternya). Nilai *inertia* akan terus menurun seiring bertambahnya jumlah klaster, namun setelah titik tertentu penurunan tersebut tidak lagi signifikan. Titik di mana terjadi perlambatan penurunan disebut titik *elbow*, dan menandakan jumlah klaster yang optimal. Sementara, metode *Silhouette* mengukur seberapa baik setiap data dikelompokkan. Nilai *Silhouette* berkisar antara -1 hingga 1, dengan nilai mendekati 1 menunjukkan bahwa objek berada dalam klaster yang tepat dan memiliki jarak yang jauh dari klaster lain.

Tabel 6. Nilai *Inertia* dan *Silhouette* untuk Setiap Jumlah Klaster.

K	*Inertia*	*Silhouette*
1	152,00	–
2	75,33	0,40
3	56,06	0,33
4	39,68	0,33
5	34,06	0,33
6	27,73	0,31
7	23,19	0,31
8	19,67	0,33
9	17,76	0,30
10	15,72	0,31

Berdasarkan Tabel 6, nilai *inertia* mengalami penurunan tajam hingga k = 2 dan mulai melandai setelahnya. Hasil ini sejalan dengan metode *Silhouette*, di mana nilai *Silhouette Score* tertinggi juga diperoleh pada k = 2 sebesar 0,40. Temuan tersebut diperkuat melalui visualisasi metode *Elbow* dan *Silhouette* pada Gambar 4. Dengan demikian, meskipun kedua metode sama-sama menunjukkan jumlah klaster optimal yang sama, metode *Silhouette* dinilai lebih unggul dalam konteks penelitian ini karena tidak hanya mempertimbangkan efisiensi pembagian data, tetapi juga kualitas pemisahan antar klaster. Nilai *Silhouette Score* sebesar 0,40 menunjukkan struktur klaster yang cukup jelas dan stabil antarprovinsi, sehingga jumlah klaster optimal ditetapkan sebanyak dua kelompok.



Gambar 4. Visualisasi Metode *Elbow* dan *Silhouette* untuk Penentuan Jumlah Klaster Optimal.

Hasil Pengelompokan K-Means Clustering

Setelah jumlah klaster ditetapkan (k = 2), algoritma *K-Means* diterapkan pada data hasil reduksi PCA untuk mengelompokkan 38 provinsi di Indonesia berdasarkan karakteristik pencemaran lingkungan. Proses iterasi menghasilkan dua klaster utama dengan distribusi 22 provinsi pada klaster 1 dan 16 provinsi pada klaster 0. Tabel 7 menyajikan contoh lima provinsi awal hasil pengelompokan *K-Means* beserta nilai variabel pencemaran yang telah distandarisasi dan logaritmikan. Kelima provinsi tersebut seluruhnya tergolong dalam klaster 1, yang secara umum ditandai oleh nilai relatif tinggi pada variabel pencemaran air, tanah, dan udara.

Tabel 7. 5 Provinsi awal Hasil Pengelompokan *K-Means*.

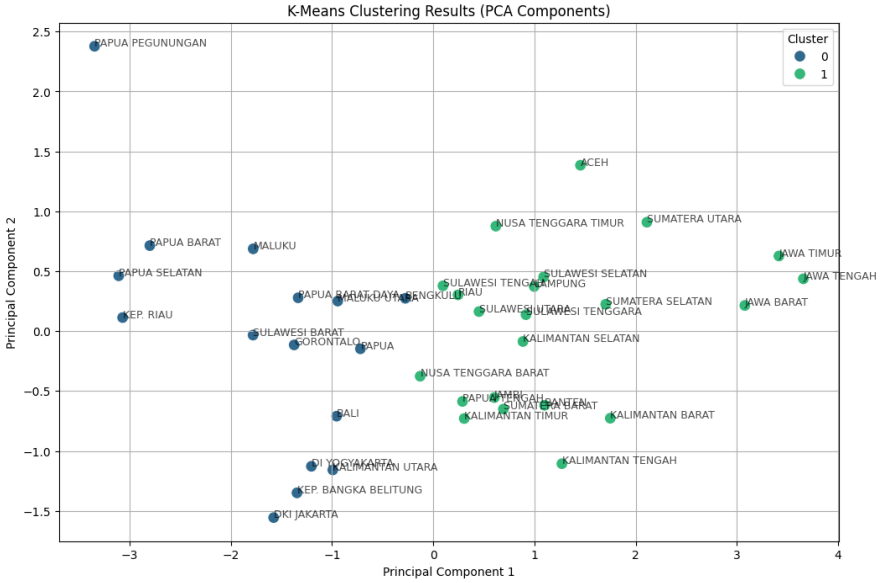
Provinsi	Air_log	Tanah_log	Udara_log	Tidak Ada_log	Cluster
ACEH	0,574	0,037	0,727	1,793	1
SUMATERA UTARA	1,393	0,318	0,993	1,610	1
SUMATERA BARAT	0,353	0,559	0,675	–0,326	1
RIAU	0,388	–0,600	0,461	0,221	1
JAMBI	0,929	0,559	–0,223	–0,177	1

Gambar 5 memperlihatkan hasil pengelompokan provinsi di Indonesia menggunakan algoritma *K-Means Clustering* berdasarkan dua komponen utama hasil *Principal Component Analysis* (PCA). Visualisasi tersebut menunjukkan pembagian provinsi ke dalam dua kelompok yang relatif jelas, yang merepresentasikan perbedaan karakteristik tingkat pencemaran lingkungan antarprovinsi. Provinsi yang termasuk dalam klaster 1 (ditandai dengan warna hijau) cenderung memiliki nilai pencemaran yang lebih tinggi pada media air, tanah, dan udara, sehingga mencerminkan tingkat pencemaran lingkungan yang relatif tinggi. Beberapa provinsi yang tergolong dalam klaster ini antara lain Jawa Timur, Jawa Tengah, Jawa Barat, Sumatera Utara, dan Aceh, yang umumnya memiliki aktivitas ekonomi dan kepadatan penduduk yang relatif tinggi serta tekanan pemanfaatan lahan yang besar.

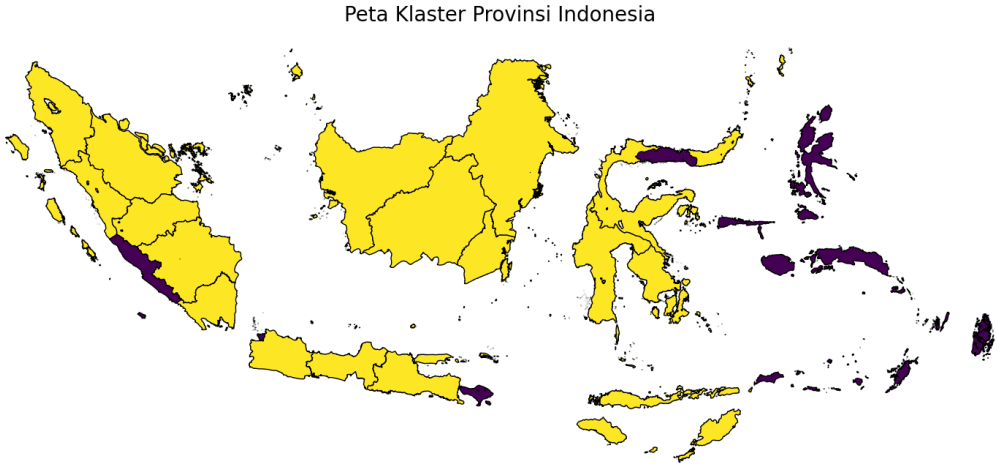
Sebaliknya, klaster 0 (ditandai dengan warna biru) merepresentasikan provinsi-provinsi dengan tingkat pencemaran lingkungan yang lebih rendah, yang menunjukkan kondisi lingkungan yang relatif lebih bersih. Provinsi dalam klaster ini antara lain Papua, Maluku, Nusa Tenggara, dan Kepulauan Riau, yang cenderung memiliki intensitas aktivitas industri dan kepadatan penduduk yang lebih rendah. Secara keseluruhan, pola klasterisasi ini menunjukkan bahwa wilayah barat Indonesia cenderung berada pada kelompok dengan tingkat pencemaran yang lebih tinggi dibandingkan wilayah timur. Hasil ini menegaskan bahwa pendekatan PCA dan *K-Means* mampu membedakan karakteristik pencemaran lingkungan antarprovinsi secara cukup jelas berdasarkan komponen utama pencemarannya.

Tabel 8. Rata-Rata Variabel pada Setiap Klaster.

Cluster	Pencemaran Air_log	Pencemaran Tanah_log	Pencemaran Udara_log	Tidak Ada Pencemaran_log
0	–0,883872	–0,851890	–0,835301	–0,754754
1	0,642816	0,619556	0,607491	0,548912



Gambar 5. Hasil Visualisasi Pengelompokan *K-Means Clustering* Berdasarkan PC = 2 (PCA).



Gambar 6. Visualisasi Peta Indonesia Pengelompokan *K-Means Clustering* Berdasarkan PC = 2 (PCA).

Peta klaster pada Gambar 6 menunjukkan hasil pengelompokan provinsi di Indonesia menggunakan algoritma *K-Means Clustering* dengan dua komponen utama PCA (PC = 2). Visualisasi tersebut memperlihatkan adanya dua kelompok besar provinsi yang terbentuk berdasarkan kemiripan karakteristik data. Provinsi yang ditandai dengan warna kuning membentuk klaster pertama, yaitu daerah yang umumnya memiliki nilai variabel lebih tinggi dalam dimensi tertentu misalnya aktivitas ekonomi, kepadatan penduduk, atau tingkat pengembangan wilayah. Klaster ini banyak didominasi oleh provinsi di wilayah barat Indonesia serta beberapa kawasan dengan aktivitas pembangunan yang intensif. Sementara itu, provinsi yang ditampilkan dengan warna ungu membentuk klaster kedua, yaitu wilayah yang menunjukkan karakteristik berbeda, seperti intensitas aktivitas ekonomi yang lebih rendah, kondisi lingkungan yang relatif lebih baik, atau tingkat tekanan pembangunan yang tidak sebesar klaster pertama.

Pola spasial yang terbentuk menunjukkan adanya perbedaan yang cukup jelas antara Indonesia bagian barat dan timur dalam hal karakteristik pencemaran lingkungan. Secara keseluruhan, peta ini mempertegas bahwa pendekatan PCA yang dikombinasikan dengan *K-Means Clustering* mampu memetakan perbedaan kondisi pencemaran antarprovinsi secara spasial dan informatif, sehingga memberikan gambaran yang lebih komprehensif mengenai pola pencemaran lingkungan di Indonesia.

IMPLIKASI KEBIJAKAN

Hasil analisis klaster menunjukkan bahwa wilayah Indonesia bagian barat, khususnya provinsi-provinsi di Pulau Jawa dan Sumatera, termasuk dalam klaster dengan tingkat pencemaran lingkungan yang relatif tinggi. Temuan ini sejalan dengan penelitian Nugraha dan Fitriani (2022) yang menegaskan bahwa kepadatan penduduk

dan tingginya aktivitas industri berperan signifikan dalam meningkatkan tingkat polusi udara dan air. Kondisi tersebut mencerminkan tekanan pembangunan yang lebih besar di wilayah barat Indonesia dibandingkan wilayah lainnya.

Implikasinya, diperlukan kebijakan pengendalian pencemaran yang bersifat spesifik lokasi (*location-specific policy*), terutama melalui penguatan pengelolaan limbah industri, pengembangan sistem transportasi ramah lingkungan, serta peningkatan pengawasan kualitas udara dan air secara berkelanjutan (Suryani & Hartanto, 2021). Pendekatan kebijakan yang terarah menjadi penting agar upaya pengendalian pencemaran dapat lebih efektif dan sesuai dengan karakteristik wilayah.

Sebaliknya, provinsi-provinsi di wilayah timur Indonesia, seperti Papua, Maluku, dan Nusa Tenggara, tergolong dalam klaster dengan tingkat pencemaran yang relatif rendah. Untuk menjaga kondisi lingkungan tersebut, pemerintah daerah perlu diarahkan pada penerapan kebijakan pembangunan hijau dan berkelanjutan, antara lain melalui peningkatan investasi pada energi terbarukan, perlindungan ekosistem, serta penguatan konservasi sumber daya alam (OECD, 2021).

Selain itu, hasil penelitian ini menunjukkan bahwa pemanfaatan data spasial dan metode analitik multivariat, seperti PCA dan *K-Means Clustering*, berpotensi diintegrasikan ke dalam sistem *early warning* pencemaran lingkungan nasional. World Bank (2023) menekankan bahwa pendekatan kebijakan berbasis data mampu meningkatkan efektivitas pengendalian lingkungan hingga 30 persen. Oleh karena itu, penguatan sistem pemantauan berbasis bukti menjadi langkah strategis ke depan.

Lebih lanjut, penguatan kelembagaan lingkungan juga perlu diarahkan untuk mendukung implementasi kebijakan berbasis data ilmiah. Kementerian Lingkungan Hidup dan Kehutanan (KLHK, 2024) dapat mempertimbangkan perluasan sistem *Environmental Quality Index* (EQI) dengan memasukkan indikator hasil analisis klaster spasial, sehingga strategi pengendalian pencemaran dapat dilakukan secara lebih terarah dan adaptif. Dengan demikian, hasil penelitian ini dapat menjadi dasar pengambilan keputusan bagi pemerintah pusat dan daerah dalam merumuskan kebijakan pengendalian pencemaran lingkungan yang berbasis wilayah dan evidensi ilmiah.

KESIMPULAN

Penelitian ini berhasil mengelompokkan 38 provinsi di Indonesia berdasarkan karakteristik pencemaran air, tanah, dan udara dengan menggunakan kombinasi metode *Principal Component Analysis* (PCA) dan *K-Means Clustering*. Hasil analisis menunjukkan terbentuknya dua klaster utama, yaitu klaster dengan tingkat pencemaran relatif tinggi yang didominasi oleh wilayah Indonesia bagian barat, serta klaster dengan tingkat pencemaran lebih rendah yang umumnya berada di wilayah Indonesia bagian timur. Temuan ini menegaskan adanya perbedaan spasial yang jelas dalam pola pencemaran lingkungan antarprovinsi.

Dua komponen utama hasil PCA mampu menjelaskan sekitar 89,41 persen total variasi data, yang menunjukkan bahwa variabel pencemaran air, tanah, dan udara memiliki keterkaitan yang kuat dalam membentuk pola pencemaran antarwilayah. Secara teoretis, hasil ini memperkuat penggunaan pendekatan statistik multivariat sebagai alat yang efektif dalam mengidentifikasi struktur data kompleks serta pola spasial pencemaran lingkungan.

Dari sisi praktis, penelitian ini memberikan kontribusi dalam mendukung upaya pemerintah melakukan pemetaan risiko pencemaran lingkungan secara nasional berbasis data. Kombinasi PCA dan *K-Means* terbukti mampu menghasilkan model klasifikasi yang efisien, terukur, dan adaptif, sehingga berpotensi diaplikasikan pada berbagai kajian lingkungan lainnya. Secara keseluruhan, penelitian ini menegaskan pentingnya pendekatan analitik kuantitatif dalam pengelolaan lingkungan hidup di Indonesia serta dapat menjadi landasan bagi perumusan kebijakan pengendalian pencemaran yang lebih tepat sasaran guna menjaga keberlanjutan ekosistem dan kualitas hidup masyarakat.

DAFTAR PUSTAKA

Azmi, R., Fathurrahman, M., & Nurhidayati, S. (2025). *Clustering tingkat pencemaran lingkungan menggunakan K-Means, K-Medoids, dan Fuzzy C-Means*. Jurnal Sains dan Data, 8(1), 45–56.

Badan Pusat Statistik. (2024). *Banyaknya desa/kelurahan menurut jenis pencemaran lingkungan hidup (desa), tahun 2024*. BPS RI.

Fa'rifah, L., & Pramesti, Y. (2022). *Penerapan metode K-Means clustering untuk pengelompokan data polusi udara di Indonesia*. Jurnal Statistika dan Komputasi, 9(2), 115–124.

Fa'rifah, M., & Pramesti, D. (2022). *Penerapan metode K-Means untuk pengelompokan data. Jurnal Matematika Integratif*, 18(2), 123–132.

Health Effects Institute. (2024). *State of global air 2024: A special report on global exposure to air pollution and its health impacts*. Health Effects Institute

Irwansyah, E. (2015). *Data mining: Algoritma dan implementasi*. Penerbit Informatika.

Irwansyah, M. (2015). *Analisis algoritma K-Means untuk pengelompokan data mahasiswa berdasarkan nilai akademik.* Jurnal Ilmiah Informatika, 4(2), 73–80.

Johnson, R. A., Wichern, D. W., & Sharma, M. (2011). *Applied multivariate statistical analysis (6th ed.).* Pearson Education.

Jolliffe, I. T. (2002). *Principal component analysis (2nd ed.).* Springer.

Jolliffe, I. T., & Cadima, J. (2016). *Principal component analysis: A review and recent developments.* Philosophical Transactions of the Royal Society A, 374(2065), 20150202.

Kementerian Lingkungan Hidup dan Kehutanan. (2024). *Laporan status lingkungan hidup Indonesia tahun 2024*. KLHK Press.

Maori, N. A., & Evanita, S. (2023). *Metode Elbow dalam optimasi jumlah cluster pada K-Means clustering*. Jurnal SIMETRIS, 14(2), 705–716.

Maori, C., & Evanita, S. (2023). *Penentuan jumlah klaster optimal menggunakan metode Elbow dan Silhouette pada algoritma K-Means*. Jurnal Teknologi Informasi, 5(1), 45–56.

Mendes, R., Alwan, N., & Putri, S. (2024). *Kajian pencemaran lingkungan dan dampaknya terhadap kesehatan masyarakat di Indonesia.* Jurnal Ekologi dan Pembangunan, 13(1), 22–34.

Muningsih, D. (2018). *Penentuan jumlah klaster optimal menggunakan metode Elbow pada analisis K-Means clustering.* Jurnal Matematika dan Sains, 23(2), 89–97.

Nugraha, R. A., & Fitriani, D. (2022). *Analisis spasial pencemaran lingkungan di Indonesia menggunakan pendekatan statistik multivariat*. Jurnal Ilmu Lingkungan, 20(3), 215–228.

OECD. (2021). *Environmental policy tools and evaluation: Promoting green growth in developing economies.* OECD Publishing.

Putra, A. D., Pratiwi, H., & Nugroho, W. (2021). *Penerapan principal component analysis pada data sosial yang memiliki multikolinearitas tinggi.* Jurnal Statistika dan Sains Data, 9(1), 12–25.

Rousseeuw, P. J. (1987). *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 20, 53–65.

Saputra, A. (2024). *Perbandingan pola pencemaran antarprovinsi di Indonesia berdasarkan aktivitas ekonomi dan kepadatan penduduk*. Jurnal Lingkungan Hidup dan Pembangunan, 6(1), 14–25.

Saputra, R. (2024). *Visualisasi data dengan peta tematik: Konsep dan implementasi*. Deepublish.

Sitepu, R., & Gultom, D. (2011). *Analisis pengelompokan menggunakan metode K-Means pada data polusi udara Kota Medan*. Jurnal Teknologi Informasi, 7(2), 88–95.

Suryani, N., & Hartanto, B. (2021). *Evaluasi penerapan kebijakan pengendalian pencemaran udara di wilayah metropolitan Indonesia.* Environmental Policy and Management Review, 8(2), 89–102.

Suyanto. (2017a). *Data mining untuk klasifikasi dan klasterisasi data.* Andi Publisher.

Suyanto. (2017b). *Pengantar machine learning menggunakan R.* Informatika.

Witten, I. H., Frank, E., & Hall, M. A. (2015). *Data mining: Practical machine learning tools and techniques (3rd ed.).* Morgan Kaufmann.

World Bank. (2023). *Data-driven environmental governance: Leveraging analytics for sustainable policy implementation.* World Bank Group.

Yusuf, N., & Tarmudi, S. (2012a). *Statistika deskriptif: Teori dan aplikasi dalam penelitian sosial.* Deepublish.

Yusuf, N., & Tarmudi, S. (2012b). *Statistika deskriptif untuk penelitian.* Graha Ilmu.