Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 5 No. 1 2025, 50 - 60
DOI: *https://doi.org/10.22487/27765660.2025.v5.i1.17489*
eISSN: 2776-5660

CLUSTERING OF REGENCIES/CITIES IN WEST JAVA PROVINCE BASED ON INDICATORS THAT AFFECT AIR QUALITY USING PRINCIPAL COMPONENT ANALYSIS AND K-MEANS CLUSTERING

Muhammad Fairuz Ahnaf[1*], Jasmine Angelia Suriawan[2], Sri Pingit Wulandari[3]
[1,2]Department of Business Statistics, Institut Teknologi Sepuluh Nopember

***e-mail**: [1*]*2043211044@student.its.ac.id*, [2]*2043211048@student.its.ac.id*

ABSTRACT

Air quality in Indonesia has significantly declined over the past two decades, transforming from one of the cleanest countries in 1998 to one of the twenty most polluted by 2016 due to a 171% increase in air particulate pollution concentrations. This study examines factors affecting air quality in West Java Province, one of Indonesia's most populous regions, using Principal Component Analysis (PCA) and K-Means clustering. The analysis includes 14 variables, such as population density, forest area, road length, and vehicle numbers. PCA was used to reduce data dimensions while retaining essential characteristics, identifying two principal components: population mobility and land infrastructure. The cluster analysis revealed two distinct groups: Cluster 1 includes 18 regencies/cities with lower population mobility and land infrastructure, indicating better air quality, while Cluster 2 consists of 9 regencies/cities with higher population mobility and land infrastructure, potentially reflecting worse air quality. Areas in Cluster 2 are concentrated near DKI Jakarta, Bandung, and the eastern border with Central Java, suggesting the influence of urbanization, industrial activities, and cross-border emissions. This study provides a spatial grouping of regencies/cities in West Java based on air quality indicators, offering insights for policymakers to target interventions more effectively. The findings emphasize the need for sustainable urban planning and stricter environmental regulations to address the growing air pollution challenge.

Keywords: *Air Quality Indicators*, *K-Means Clustering, Principal Component Analysis*, *West Java*.

Cite: Ahnaf, M. F., Suriawan, J. A., & Wulandari, S. P. (2025). Clustering of regencies/cities in West Java Province based on indicators that affect air quality using principal component analysis and K-means clustering. *Parameter: Journal of Statistics*, *5*(1), 50–60. https://doi.org/10.22487/27765660.2025.v5.i1.17489

[*]Received: May 2025; Reviewed: May 2025; Published: Jul 2025

INTRODUCTION

Over the past two decades, air quality in Indonesia has undergone significant changes. Between 1998 and 2016, Indonesia transformed from one of the countries with the cleanest air in the world to one of the twenty most polluted countries, due to a 171 percent increase in air particulate pollution concentrations (van Donkelaar, et al., 2016). The largest increase occurred in recent years, with pollution more than doubling between 2013 and 2016. Regardless of the causes, by 2016, 80 percent of Indonesia's 250 million population lived in areas where particulate pollution levels exceeded WHO guidelines (Greenstone & Fan, 2019).

High air pollution now negatively impacts the health of the Indonesian population. In 1998, air pollution almost did not affect life expectancy. Even until 2013, high air pollution only reduced average life expectancy by a few months. However, if pollution concentrations remain at current levels, the average life expectancy will be reduced by up to 1.2 years compared to the life expectancy that could be achieved if the WHO guideline of 10 $\mu g/m^3$ for long-term exposure to fine particles (PM2.5) is met (Greenstone & Fan, 2019).

Various factors potentially affect air quality in a region, one of which is the population density in an area. High population growth rates increase the demand for life-supporting needs, such as space. Therefore, air, as an essential component of life, is vulnerable to changes due to human activities (Karunia, 2019). Other factors that potentially affect air quality levels include population density, migration which also affects population density, forest areas that can absorb carbon dioxide, as well as the length of road sections and the number of motor vehicles that produce emissions from engine combustion. West Java Province, one of the provinces with the largest population in Indonesia, is at high risk of experiencing a decline in air quality.

Facing the many factors that can affect air quality, this study uses Principal Component Analysis (PCA) to reduce the dimensions of these factors without losing the characteristics of the original data. Principal Component Analysis (PCA) is an analytical method used to group several variables into a smaller group of variables based on the similarity of properties or characteristics possessed by the data (Baroroh, 2013). A variable that will be grouped into a factor, so that the variable correlates with other variables that fall into a certain factor group, is called Principal Component Analysis (Santoso, 2012). The principal components produced will later be used to group the regencies/cities of West Java Province into several groups with different characteristics. This analysis is carried out using K-Means Clustering. Cluster analysis is an analytical technique that aims to group objects into several groups that have different characteristics from one group to another. The K-Means method refers to a method based on distance calculations, which divides data into several clusters (Silvi, 2018; Metisen & Sari, 2015). The formed groups will be examined for their characteristics to identify which regencies/cities have air quality risks.

Several studies in Indonesia have explored the use of Principal Component Analysis (PCA) and clustering techniques to address environmental issues, though their focus and scope vary. Magriarty et al. (2023) applied PCA and K-Means clustering to identify waste management zones in Tapin Regency, showing how PCA can reduce dimensions and help classify regions based on vulnerability. However, the study did not focus on air quality. Annas et al. (2022) used K-Means and Self-Organizing Maps (SOM) to cluster areas in Makassar City based on pollutant concentrations such as CO, NO_2, and SO_2, supported by GIS visualization. While this study addressed air pollution, it did not apply PCA or examine broader environmental factors like population density or forest cover. Meanwhile, Rahmah et al. (2022) used Affinity Propagation to cluster air quality data in Pekanbaru, but the analysis lacked dimensionality reduction, and the clustering results were categorized as weak based on the Silhouette Coefficient (0.264).

Despite these contributions, research that integrates both PCA and clustering techniques such as K-Means to analyze multivariate environmental determinants of air quality in West Java Province is still very limited. Most existing studies either focus solely on pollutant concentration or apply clustering to non-air-quality issues. This study addresses that gap by using PCA to reduce and interpret complex variables affecting air quality—such as population density, vehicle numbers, forest area, and infrastructure—and applying K-Means clustering to group regencies and cities based on these characteristics. Through this approach, a more comprehensive understanding of regional air quality risk can be achieved.

The results of this study are presented as a classification map of regencies and cities in West Java Province based on environmental indicators related to air quality. These findings are expected to assist local governments and stakeholders in understanding the underlying environmental patterns

contributing to pollution, enabling more informed decision-making. By identifying priority areas for intervention, provincial and city-level authorities can design targeted, evidence-based policies to reduce air pollution and improve environmental sustainability. Additionally, this study serves as a reference for future research on regional environmental analysis and air quality management.

MATERIALS AND METHODS

The data used in this study is secondary data published by Badan Pusat Statistika (BPS) of West Java Province. This data includes factors affecting air quality in the 27 regencies/cities of West Java Province. The research variables used in this study are shown in the table below.

Table 1. Research Variables

Variable	Description	Unit
X_1	Population	People
X_2	Population Density	People/Km^2
X_3	Number of In-migrations	People
X_4	Number of Out-migrations	People
X_5	Limited Production Forest	Ha
X_6	Permanent Production Forest	Ha
X_7	Convertible Production Forest	Ha
X_8	Length of National Roads	Km
X_9	Length of Provincial Roads	Km
X_{10}	Length of Regency Roads	Km
X_{11}	Number of Passenger Cars	Units
X_{12}	Number of Buses	Units
X_{13}	Number of Trucks	Units
X_{14}	Number of Motorcycles	Units

This study uses variables with different year periods, namely population, population density, road length including national, provincial, and regency, as well as the number of passenger cars, buses, trucks, and motorcycles in 2023. The variables of the number of in-migration, out-migration, and area of limited production forest, area of permanent production forest, and area of convertible production forest in 2022. The year limitation on each variable is due to data limitations and was chosen with consideration to obtain results that are representative of current conditions. This restriction is expected to provide a more relevant picture of the factors analyzed in this study.

Before conducting Principal Component Analysis (PCA), assumption testing was performed to ensure the data met the required criteria for multivariate analysis. The following assumptions were tested.

1. Multivariate Normality

The multivariate normal distribution test is conducted to assess the distribution of data in a group of data or variables, whether the data distribution is multivariate normally distributed or not. This needs to be done for residual analysis after matching various modes (Hajarisman, 2008). Multivariate normal distribution testing is conducted to determine whether the data follows a multivariate normal distribution pattern or not. Multivariate normal distribution testing uses the proportion T-test. The proportion T hypothesis is as follows.

H_0: Data is multivariate normal distributed

H_1: Data is not multivariate normal distributed

With a significant level of α, H_0 is rejected if the proportion T value is outside the range of values $0.45 < T < 0.55$. The proportion T test statistic is shown in Equation 1.

$$T = n(\bar{x} - \mu_0)^T \sum^{-1} (\bar{x} - \mu_0) \quad (1)$$

Where:

n: Number of data or sample size

$\bar{\boldsymbol{x}}$: sample mean vector

$\boldsymbol{\mu_0}$: Hypothesized population mean vector

$\boldsymbol{\Sigma}$: Variance-covariance matrix

2. Data Dependency

The Bartlett Test of Sphericity is used to determine if the correlation matrix is not the identity matrix. The population correlation matrix is the identity matrix. This test is used to test the hypothesis that the variables are not correlated in the population. The value for each variable that is perfectly correlated with itself is R = 1 and for variables that are not correlated with other variables, R ≠ 1 (Supranto, 2004). The hypothesis in the Bartlett test is shown as follows.

H_0: **ρ = I** (The correlation matrix is identical to the identity matrix or the correlation between variables is independent).

H_1: **ρ ≠ I** (The correlation matrix is not identical to the identity matrix or the correlation between variables is dependent).

With a significant level of α, H_0 is rejected if the value of Bartlett's test statistic is greater than $\chi^2_{(1-\alpha;\, g-1)}$. Bartlett's test statistic is shown in Equation 2.

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right)\ln |\mathbf{P}| \tag{2}$$

Where:
n: Number of data
p: Number of variables
P: Correlation Matrix

3. Sampling Adequacy

Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy is an index that compares the magnitude of the observed correlation coefficient with the magnitude of the partial coefficient. The number generated by the KMO Measure of Sampling Adequacy must be greater than 0.5 so that factor analysis can be processed further (Santoso, 2012). KMO values can be categorized based on the criteria in Table 2. The KMO value is obtained from the calculation in Equation 3.

Table 2. KMO Categories

Criteria	Categories
KMO ≥ 0,9	Very Good
0,8 ≤ KMO < 0,9	Good
0,7 ≤ KMO < 0,8	Satisfactory
0,6 ≤ KMO < 0,7	Poor
0,5 ≤ KMO < 0,6	Bad
KMO < 0,5	Not Accepted

$$KMO = \frac{\sum_{j=1}^{p}\sum_{i=1}^{p} r_{ji}^2}{\sum_{j=1}^{p}\sum_{i=1}^{p} r_{ji}^2 + \sum_{j=1}^{p}\sum_{i=1}^{p} a_{ji}^2} \tag{3}$$

Where:
r_{ij}: Correlation coefficient between variables i and j
a_{ij}: Partial correlation coefficient between variables i and j
p: Number of Variables

4. Data Feasibility

Data eligibility can be checked using the Measure of Sampling Adequacy (MSA), which is an examination to determine the eligibility of each variable to be used in factor analysis (Santoso, 2012). Three assessment criteria that can be used to determine the feasibility of the variables used are described in Table 3.

Table 3. MSA Criteria

Criteria	Description
MSA = 1	The variable is predicted without any error from other variables.
0,5 < MSA < 1	The variable can still be used for further analysis.
MSA ≤ 0,5	The variable cannot be used for further analysis.

Principal Component Analysis (PCA) was first introduced by Karl Pearson in 1901 in the field of biology. Later, it was independently rediscovered by Karhunen and further developed by Loeve in 1963, becoming known as the Karhunen-Loeve Transform in telecommunications. PCA is a statistical technique used to reduce the dimensionality of a dataset by transforming a large set of variables into a smaller set of new, uncorrelated variables (principal components) while retaining as much information as possible. This method simplifies data interpretation by identifying patterns and grouping variables with similar characteristics (Faisal, Dinata, & Sari, 2023; Baroroh, 2013). According to Santoso (2012), PCA involves grouping variables into factors, where each factor represents a subset of variables that are correlated. The outcome of PCA is a set of principal components that collectively explain the variability in the original data.

Cluster analysis is an analytical technique used to group objects into distinct clusters, where objects within a cluster are as similar as possible, while objects in different clusters exhibit significant differences. Each cluster is internally homogeneous, aiming to minimize variation within the group (Silvi, 2018). Cluster analysis methods are broadly categorized into hierarchical and non-hierarchical approaches. The non-hierarchical method involves grouping objects without relying on a hierarchical structure. This method requires the number of clusters to be predetermined and typically involves random. Selection of initial cluster centers. Despite its drawbacks, such as dependency on initial centroids and the need to specify the number of clusters, the non-hierarchical approach is more efficient for handling large datasets compared to hierarchical methods. A key challenge in this method is selecting an appropriate centroid or starting point for forming clusters (Andiani, Septiani, & Riana, 2022).

One common non-hierarchical technique is the K-Means method, which partitions data into clusters based on distance calculations. This algorithm is restricted to datasets with numeric attributes (Metisen & Sari, 2015). The algorithm in cluster analysis with K-Means is as follows.

1. Determine k as the number of clusters to be formed.
2. Generate k centroids as initial cluster centers randomly.
3. Calculate the distance of each object to each centroid of each cluster with Euclidean Distance as follows.

$$d_{(x,y)} = \| x - y \| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}; i = 1, 2, 3, ..., n \quad (4)$$

Where:
x_i: The i-th variable of object x
y_i: The i-th variable of object y
n: Total number of samples

4. Allocate each object to the nearest centroid.
5. Perform iteration and determine the new centroid with the following formula.

$$v = \frac{\sum_{i=1}^{n} x_i}{n}; i = 1, 2, 3, ..., n \quad (5)$$

Where:
v: New centroid of the cluster
x_i: The i-th object in the cluster
n: The total number of objects in the cluster

6. **Repeat the third step if the centroid position is not the same.**

To determine the best clustering method for the observed data, it is essential to select the optimum number of clusters. One common approach is the Silhouette Method, which evaluates cluster quality by combining cohesion and separation. Cohesion is assessed by counting all objects in a cluster, while separation is measured by calculating the average distance between each object in the cluster and its closest cluster. The silhouette value for all data with k number of clusters, expressed as sil(c), is calculated using the formula for the average silhouette value of all clusters (Rokach, 2015). The silhouette score would always lie between -1 to 1, with 1 representing better clustering. The formula for the silhouette score of a single object i is shown in Equation 6.

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{6}$$

Where:

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \tag{7}$$

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \tag{8}$$

RESULTS AND DISCUSSION

This chapter will explain the clustering of air quality in West Java Province based on the results of principal component analysis using descriptive statistics, testing the assumptions of principal component analysis which includes normal distribution testing, independence test using Barlett, and checking sampling adequacy with KMO, followed by principal component analysis. The results of the principal component analysis will be clustering using k-means, where the optimum cluster is selected using silhouette scores and then the results of cluster mapping based on the optimum cluster are described.

Descriptive Statistical Analysis of Factors Affecting Air Quality in West Java Province

The characteristics of factors affecting air quality in West Java Province using descriptive statistics are presented in Table 4.

Table 4. Descriptive Statistics

Variable	*Mean*	Standard Deviation	Minimum	Maximum
X_1	1865	1228	210	5682
X_2	3911	4668	385	15176
X_3	13015	8260	2243	34389
X_4	13395	8274	2427	35282
X_5	6378	11885	0	47545
X_6	6559	9852	0	39831
X_7	7930	9228	0	31318
X_8	65,7	54,7	4,0	211,0
X_9	87,5	82,1	7,0	315,0
X_{10}	904	526	105	1958
X_{11}	81395	93762	5953	372806
X_{12}	1005	1130	47	5512
X_{13}	20877	17160	2501	65987
X_{14}	510584	383031	59135	1433350

Table 4 shows that the data characteristics on variables X_1, X_3, X_4, X_8, X_9, X_{10}, X_{13}, and X_{14} have a standard deviation value that is smaller than the average value, which indicates that in these variables the data variance is relatively small (close to the average value). Meanwhile, the other variables, namely X_2, X_5, X_6, X_7, X_{11}, and X_{12} have a standard deviation value greater than the average value, so it can be said that the data variance is classified as large.

Principal Component Analysis Assumption Testing

Testing the assumptions of principal component analysis on data on factors affecting air quality in West Java Province consists of a multivariate normal distribution test, an independence test between variables (Bartlett test), a sampling adequacy test (KMO test), and an Anti-Image correlation test. The multivariate normal distribution test on the data of factors affecting air quality in West Java Province uses the following hypothesis.

H_0: Data is multivariate normal distributed
H_1: Data is not multivariate normal distributed

The significant level (α) is 0.05 with a rejection area of rejecting H_0 if T-Proportion ≤ 45% or T-Proportion ≥ 55% which will be proven by the test statistics in Table 5.

Table 5. Multivariate Normal Test Results

T-Proportion	T-Table
0,516	45% ≤ α ≤ 55%

The multivariate normal distribution test conducted with a significant level of 0.05 has a rejection area if the T-proportion is outside the range of 45% to 55%. The results of the calculation of test statistics from the multivariate normal distribution test found that the data on factors affecting air quality in West Java Province has a t-proportion value of 51.6% which is in the range of 45% to 55% so it was decided to fail to reject H_0. Thus, the data on factors affecting air quality in West Java Province are multivariate normally distributed.

Testing the independence assumption is done using the Barlett test to determine the homogeneity of variance in the influence of factors that affect air quality in West Java Province. Independence testing is done using the following hypothesis.

H_0: $\boldsymbol{\rho = I}$ (The correlation matrix is identical to the identity matrix or the correlation between variables of factors affecting air quality in West Java Province is independent).

H_1: $\boldsymbol{\rho \neq I}$ (The correlation matrix is identical to the identity matrix or the correlation between variables of factors affecting air quality in West Java Province is dependent).

The significant level (α) is set at 0.05 with the H_0 rejection area if $\chi^2 > \chi^2_{(0.05;91)}$ or P-Value < 0.05 which will be proven by the test statistics in Table 6.

Table 6. Bartlett Test Results

$\boldsymbol{\chi^2}$	$\boldsymbol{\chi^2_{(0.05;91)}}$	P-Value
453.593	114.268	0.000

Table 6 shows that 453.592 is greater than $\chi^2_{(0.05;91)}$ of 114.268 and reinforced by a p-value of 0.000 where the value is smaller than 0.05 so that it can be decided to reject H_0, which means that the correlation matrix is identical to the identity matrix or the correlation between variables of factors affecting air quality in West Java Province is dependent. So, in this case, the variable data of the factors affecting air quality in West Java Province fulfills the dependent assumption.

The data sufficiency check is used to test whether it is sufficient to factorize, as it is needed in principal component analysis using KMO. Based on the analysis results, the KMO test value is 0.719, which is greater than 0.5, which means that the data on factors affecting air quality in West Java Province are sufficient to be factored in the range of 0.7 - 0.79 in the normal category.

The check to measure the correlation adequacy of each variable in the data on factors affecting air quality in West Java Province uses the Anti-Image correlation test. This check will be met if the Measures of Sampling Adequacy (MSA) value is greater than 0.5. However, if the MSA value is less than 0.5, then the assumption is not met and the variable cannot be analyzed further so it needs to be excluded. The results of the Anti-Image correlation test on data on factors affecting air quality in West Java Province are shown in Table 7.

Table 7. MSA Value

Variable	MSA Value	Variable	MSA Value
X_1	0.738	X_8	0.774
X_2	0.780	X_9	0.844
X_3	0.716	X_{10}	0.634
X_4	0.667	X_{11}	0.698
X_5	0.702	X_{12}	0.668
X_6	0.774	X_{13}	0.737
X_7	0.860	X_{14}	0.670

Table 7 shows that the MSA value for each variable in the data on factors affecting air quality in West Java Province is greater than 0.5, so it can be decided to conduct further analysis with all variables, without eliminating them.

Principal Component Analysis Results

The principal component analysis of the data of factors affecting air quality in West Java Province will consist of total variance explained, scree plot, and component naming. The results of total variance are explained on data on factors affecting air quality in West Java Province to determine the results of how many factors are formed based on eigenvalues, variance, and cumulative values in Table 8.

Table 8. Total Variance Explained

Component	Eigen Value	% of Variance	% Cumulative
X_1	6.124	43.742	43.742
X_2	4.589	32.781	76.523
X_3	0.848	6.058	82.581
X_4	0.596	4.254	86.836
X_5	0.524	3.743	90.579
X_6	0.467	3.332	93.911
X_7	0.256	1.830	95.741
X_8	0.237	1.690	97.431
X_9	0.142	1.015	98.446
X_{10}	0.129	0.921	99.367
X_{11}	0.044	0.312	99.679
X_{12}	0.030	0.213	99.892
X_{13}	0.010	0.075	99.967
X_{14}	0.005	0.033	100

Table 8 shows that the results of the first 2 components have an eigenvalue of more than 1, so based on this value, 3 factors are formed from 14 components. Components 1 and 2 can explain the variables by 43.742% and 32.781% respectively, so that the cumulative percentage of the three components is 76.523%. The results of the scree plot on the data of factors affecting air quality in West Java Province are shown in Figure 1.

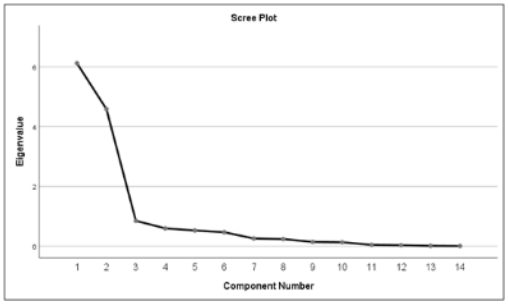

Figure 1. Scree Plot

Figure 1 shows the movement from component 1 to component 14, where there is a significant decrease in component 1 to component 3 which illustrates the difference in eigenvalues. Meanwhile, after component 3 there is no significant decrease. Furthermore, the grouping of components in the data of factors affecting air quality in West Java Province is shown in Table 9.

Table 9. Component Grouping

Variable	Component		Variable	Component	
	1	2		1	2
X_1	0.127	0.096	X_8	0.003	0.189
X_2	0.068	-0.153	X_9	-0.003	0.182

X_3	0.147	0.022	X_{10}	0.075	0.111
X_4	0.151	0.016	X_{11}	0.157	-0.077
X_5	0.022	0.175	X_{12}	0.138	-0.076
X_6	-0.024	0.176	X_{13}	0.163	-0.024
X_7	-0.022	0.158	X_{14}	0.158	-0.002

Table 9 shows that the variables X_1, X_2, X_3, X_4, X_{11}, X_{12}, X_{13}, X_{14} have a value of component 1 that is greater than the value of component 2, which means that these variables are included in component 1. Meanwhile, other variables, namely X_5, X_6, X_7, X_8, X_9, and X_{10} are included in component 2. The model formed from the principal component analysis is shown as follows.

$$PC_1 = 0{,}127X_1 + 0{,}068X_2 + 0{,}147X_3 + 0{,}151X_4 + 0{,}022X_5 - 0{,}024X_6 - 0{,}022X_7 + 0{,}003X_8 - 0{,}003X_9 + 0{,}075X_{10} + 0{,}157X_{11} + 0{,}138X_{12} + 0{,}163X_{13} + 0{,}158X_{14} \quad (9)$$

$$PC_2 = 0{,}096X_1 - 0{,}153X_2 + 0{,}022X_3 + 0{,}016X_4 + 0{,}175X_5 + 0{,}176X_6 + 0{,}158X_7 + 0{,}189X_8 + 0{,}182X_9 + 0{,}111X_{10} - 0{,}077X_{11} - 0{,}076X_{12} - 0{,}024X_{13} - 0{,}002X_{14} \quad (10)$$

In the formation of Principal Components, the variables included in the Principal Components and their new component names are obtained as follows.

Table 10. Principal Components Produced

Principal Component (PC)	Variables	New Component Name
PC_1	X_1, X_2, X_3, X_4, X_{11}, X_{12}, X_{13}, X_{14}	Population mobility
PC_2	X_5, X_6, X_7, X_8, X_9, X_{10}	Land infrastructure

Clustering Analysis Results

Based on the model from the principal component analysis, each principal component value will be calculated for regencies/cities in West Java and then clustering analysis will be conducted. Cluster analysis will use the k-means method, where the process starts from calculating the optimum number of clusters, division of regencies/cities in West Java based on the results of the most optimum silhouette score, characteristics of the cluster division results for each main component, and cluster mapping. The result of calculating the optimum silhouette score is shown in Figure 2.

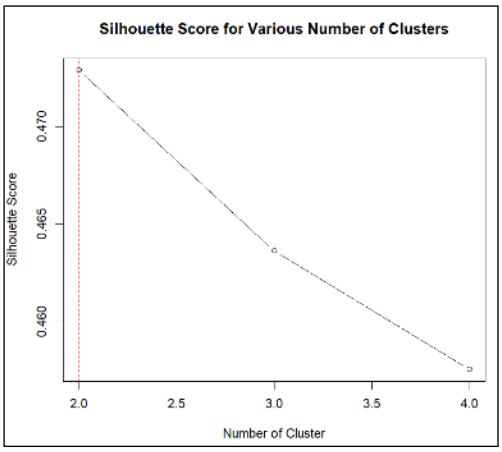

Figure 2. Silhouette Score

Based on Figure 2, it is found that the optimum number of clusters is 2 clusters with a silhouette score of 0.472. All regencies/cities in West Java will be divided into clusters with fellow regencies/cities in West Java that have similar characteristics. The results of the division of regencies/cities in West Java in the two clusters are shown in Table 11.

Table 11. Cluster Division

Cluster	Number of Members	Regency/City
1	18	Cianjur, Tasikmalaya, Cirebon, Sumedang, Indramayu, Subang, Purwakarta, Karawang, Pangandaran, Bogor City, Sukabumi City, Bandung City, Cirebon City, Bekasi City, Depok City, Cimahi City, Tasikmalaya City, Banjar City
2	9	Bogor, Sukabumi, Bandung, Garut, Ciamis, Kuningan, Majalengka, Bekasi, Bandung Barat

Based on cluster classification using the K-Means method by combining the results of principal component analysis, it is known that cluster 1 is 18 regencies/cities and cluster 2 is 9 regencies/cities of West Java. Moreover, based on the clustering results, we will identify the characteristics of each cluster on the factors affecting air quality in West Java as shown in the mean, minimum, and maximum values. Cluster characteristics are shown in Table 12.

Table 12. Cluster Characteristics

Variable	Clusters	Mean	Minimum	Maximum
PC_1	1	-0.524	-1.270	0.533
	2	1.050	-0.081	2.690
PC_2	1	-0.248	-1.250	0.594
	2	0.497	-1.300	2.820

Table 12 shows that cluster 1 has an average value of PC_1 and PC_2 factors that are lower than the average value of PC_1 and PC_2 factors in cluster 2. This shows that regencies/ cities in West Java that are included in Cluster 1 have average population mobility and land infrastructure that is less than the mean compared to regencies/cities in West Java that are included in Cluster 2, potentially indicating better air quality in regencies/ cities on cluster 1 and worse in cluster 2. Based on the clustering results, a grouping mapping of regencies/cities in West Java Province based on air quality indicators will be conducted, as shown in Figure 3.

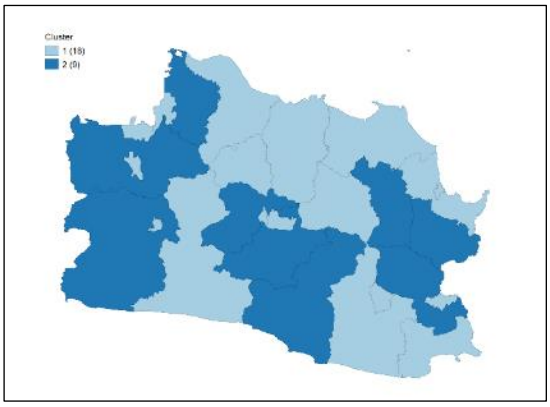

Figure 3. Clustering of Regencies/cities Based on Air Quality Indicators

Figure 3 illustrates the administrative boundaries of regencies and cities in West Java Province, grouped into two clusters based on air quality indicators. The light blue areas represent cluster 1, while the darker blue areas indicate cluster 2, which has a worse air quality indicator. Cluster 2 regions are primarily located near the border with DKI Jakarta Province, around Bandung City, and in the eastern part of West Java bordering Central Java. This distribution suggests that air quality in cluster 2 may be influenced by factors such as the proximity to the national capital (DKI Jakarta), industrial and urban activities around Bandung (the provincial capital), and inter-provincial emissions from Central Java.

CONCLUSION

Analyzing the conditions across West Java's regencies and cities revealed clear patterns in the factors affecting air quality. Using Principal Component Analysis (PCA), two main components were identified that represent the underlying patterns in the data: population mobility and land infrastructure. These components highlight how movement of people and the extent of physical development can contribute to differences in air quality between regions.

Through K-Means clustering, the regions were grouped into two clusters with distinct characteristics. Cluster 1 consists of areas with lower population movement and infrastructure development, which may be associated with better air quality. Cluster 2, on the other hand, includes areas with higher levels of human activity and land use—many of which are located near major urban and industrial centers such as Jakarta and Bandung. This suggests that higher mobility and infrastructure density may contribute to lower air quality.

The findings of this research provide valuable insight for local governments in West Java. By understanding the environmental characteristics that influence air quality, policymakers can better identify priority areas for intervention and formulate more effective strategies to address pollution and promote sustainable development.

REFERENCES

Andiani, D., Rahayu, S. D., & Riana, A. (2022). Analisis Teknik non-Hierarki untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Barat Berdasarkan Indikator Kesejahteraan Rakyat 2020. JRMST | Jurnal Riset Matematika Dan Sains Terapan, 2(1), 21–28.

Annas, S., Uca, U., Irwan, I., Safei, R. H., & Rais, Z. (2022). Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia. Jambura Journal of Mathematics, 4(1), Article 1. https://doi.org/10.34312/jjom.v4i1.11883

Baroroh, A. (2013). Analisis Multivariat dan Time Series dengan SPSS 21. PT Elex Media Kumputindo.

Faisal, M., Dinata, S. A. W., & Sari, D. R. (2023). ANALISIS KOMPONEN UTAMA PADA DINAS KETENAGAKERJAAN BAGIAN PENEMPATAN DAN PERLUASAN KERJA MENCARI PEKERJAAN MENURUT GOLONGAN PEKERJAAN. Journal of Innovation Research and Knowledge, 2(12), Article 12. https://doi.org/10.53625/jirk.v2i12.5627

Greenstone, M., & Qing, F. (2019). Kualitas Udara Indonesia yang Memburuk dan Dampaknya terhadap Harapan Hidup (p. 10). AQLI. https://aqli.epic.uchicago.edu/wp-content/uploads/2019/03/AQLI_Indonesia_Report_v04_Digital_id_27032019_LowRes.pdf

Hajarisman, N. (2008). Statistika Multivariat. Program Studi Statistika, Universitas Islam Bandung.

Karunia, D. (2019). PENGARUH AKTIVITAS MANUSIA TERHADAP PERUBAHAN KUALITAS UDARA. OSF. https://doi.org/10.31227/osf.io/rxejg

Magriaty, R., Murtilaksono, K., & Anwar, S. (2023). Analisis K-Means Cluster untuk Identifikasi Kawasan Pengelolaan Sampah di Kabupaten Tapin Provinsi Kalimantan Selatan. Journal of Regional and Rural Development Planning (Jurnal Perencanaan Pembangunan Wilayah Dan Perdesaan), 7(1), Article 1. https://doi.org/10.29244/jp2wd.2023.7.1.79-90

Metisen, B. M., & Sari, H. L. (2015). ANALISIS CLUSTERING MENGGUNAKAN METODE K-MEANS DALAM PENGELOMPOKKAN PENJUALAN PRODUK PADA SWALAYAN FADHILA. JURNAL MEDIA INFOTAMA, 11(2). https://doi.org/10.37676/jmi.v11i2.258

Rahmah, M., Candra, A., & Sembiring, R. W. (2022). Identifikasi Predikat Hasil Pengelompokan Data Kualitas Udara dengan Menggunakan Affinity Propagation dan Silhouette Coefficient. InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan, 6(2), Article 2. https://doi.org/10.30743/infotekjar.v6i2.4670

Rokach, L., & Maimon, O. (2014). Data Mining with Decision Trees: Theory and Applications (2nd ed., Vol. 81).

Santoso, S. (2012). Analisis SPSS pada Statistik Parametrik. PT. Elex Media Komputindo.

Silvi, R. (2018). Analisis Cluster dengan Data Outlier Menggunakan Centroid Linkage dan K-Means Clustering untuk Pengelompokkan Indikator HIV/AIDS di Indonesia. Jurnal Matematika MANTIK, 4(1), Article 1. https://doi.org/10.15642/mantik.2018.4.1.22-31

Supranto, J. (2004). Analsis Multivariat. Alfabeta.

van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M., & Winker, D. M. (2016). Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. Environmental Science & Technology, 50(7), 3762–3772. https://doi.org/10.1021/acs.est.5b05833