DOI: https://doi.org/10.34123/jurnalasks.v17i2.618

POLITEKNIK STATISTIKA STIS

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 17 No. 2, December 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

Clustering Regencies/Cities Vulnerable to Air Pollution in the Java Island: Fuzzy Geographically Weighted Clustering

Arya Candra Kusuma[1*], Arie Wahyu Wijayanto[2], Arista[3], Vicka Kharisma Bahar[4], Tifani Husna Siregar[5]

[1]BPS-Statistics Indonesia, Jakarta, Indonesia, [2]Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia, [3]Indiana University Bloomington, Bloomington, Indiana, USA, [4]Asian Development Bank Institute, Japan, [5]King Fahd University of Petroleum and Minerals, Saudi Arabia
**Corresponding Author: E-mail address: arya.candra@bps.go.id*

ARTICLE INFO

Article history:
Received 4 August, 2024
Revised 2 April, 2025
Accepted 12 July, 2025
Published 31 December, 2025

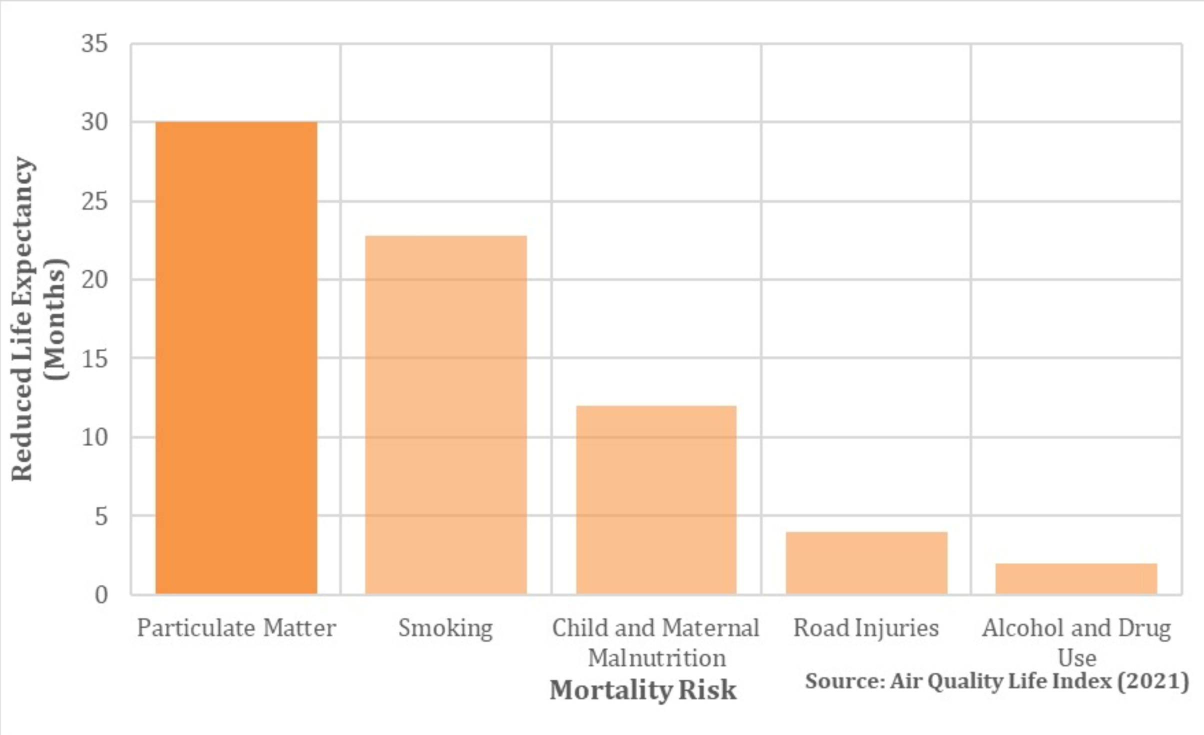
Keywords:
Air Pollution; Artificial Bee Colony (ABC); Environment; Fuzzy C Means (FCM); Fuzzy Geographically Weighted Clustering (FGWC)

Abstract

Introduction/Main Objectives: Air pollution has become a critical global concern with substantial effects on human health and the environment. **Background Problems:** Java Island in Indonesia, recognized for its high population density and industrial activities, necessitates focused effort in resolving this issue. **Novelty:** While air pollution research has been enormous, there has been no effort to cluster regencies or cities on Java Island utilizing spatially-based data. This research seeks to cluster regencies and cities on Java Island according to air pollution levels and to compare geodemographic and non-geodemographic clustering methodologies. **Research Methods:** This study employs secondary data regarding air pollution, obtained from the Openweather API. This study employs a geodemographic clustering technique, namely fuzzy geographically weighted clustering (FGWC), optimized by the artificial bee colony (ABC) algorithm. **Finding/Results:** The study findings indicate that the geodemographic clustering method ABCFGWC surpasses Fuzzy C-Means (FCM) according to the TSS (Tang-Sun-Sun) index. The data reveal that the Greater Jakarta or Jabodetabek area and its adjacent territories are more susceptible to air pollution. The findings of this study are expected to enhance the spatial planning and mapping of air pollution management strategies on Java Island.

1. Introduction

Air pollution has emerged as a global concern due to its harmful impacts on human health and the environment [1], [2], [3]. The deterioration of global air quality is attributable to anthropogenic activities that release pollutants into the atmosphere, including car emissions, industrial emissions, and fossil fuel combustion [2], [4]. The pollutants encompass particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and several other organic chemicals. The World Health Organization (WHO) reports that 91% of the global population lives in regions where air pollution is above the established air quality guidelines [2]. In Indonesia, the average air pollution level exceeds the WHO recommended limit by three times [5]. Consequently, analyzing air pollution is essential for developing effective strategies to alleviate its harmful effects.



Source: Lee & Greenstone [5]

Figure 1. The impact of reduced life expectancy caused by mortality risk in Indonesia

Numerous scientific research papers [6], [7], [8] have provided evidence of the detrimental long-term and short-term health effects of air pollution. From Figure 1, it is evident that air pollution is one of the most significant risk factors contributing to reduced life expectancy, exceeding all other determinants [5]. PM2.5, a category of fine particulate matter, is a highly harmful pollutant closely associated with respiratory and cardiovascular problems [8]. Vulnerable populations, including children, the elderly, and individuals with a history of respiratory or cardiovascular disease, are particularly susceptible to the health impacts of air pollution [9].

In addition to its effects on human health, air pollution also poses a significant threat to the ecosystem and the overall welfare of humanity [2], [4]. Sulfur dioxide (SO2) and nitrogen oxides (NO2) may react chemically to form acid rain, which has harmful consequences for the ozone layer and ultimately contributes to global warming [4]. In terms of human well-being, air pollution has a significant impact on both labor productivity and the cognitive capacities of students [10], [11]. Therefore, resolving air pollution problems will not only tackle health concerns but also address environmental concerns [2].

To alleviate the adverse effects of air pollution, several governments and organizations across different nations have enacted a series of measures, including the establishment of emission regulations, the adoption of renewable energy sources, and the promotion of sustainable transportation [2], [13], [14]. Furthermore, the European Union is implementing the Carbon Border Adjustment Mechanism (CBAM), a scheme intended to charge fees on carbon-intensive goods that will be fully enacted in 2026, following a transitional phase from 2023 to 2026 [12]. In Indonesia, the urgency to address air pollution at the local level is increasing as there are different levels of air pollution in each region [15]. Java Island is a significant area for the assessment of air pollution levels attributable to its concentrated industrial and economic activities.

Several previous studies have examined air pollution using different methodologies. A comprehensive analysis incorporating spatiotemporal assessment, clustering, and correlation techniques has been applied to air pollution data in northern China [16]. An enhanced variant of the K-means algorithm has been proposed to improve accuracy and computational efficiency in determining the air quality index (AQI) [17]. Cluster analysis combined with spatial correlation methods has also been employed to examine air pollution patterns in China [18]. In addition, the K-means clustering technique has been used to classify air pollution regions in Makassar [19]. Despite extensive study on air pollution, no studies have focused on clustering areas by air pollution at the regency or city level specifically for Java Island, including geographical characteristics. Moreover, previous studies mostly focused on clustering, neglecting the relationship among distinct regions. Clustering analysis, which incorporates spatial characteristics, might serve as a methodological approach to formulate policies that acknowledge regional variability [20]. Geodemographic clustering enhances the identification of inter-regional linkages and supports spatially explicit analysis, hence assisting in policy formulation [20], [21]. This approach has been successfully implemented in several regions [22] as well as in other domains [23], [24], [25].

Thus, this study aims to cluster regencies/cities on Java Island based on air pollution conditions. In addition, this research also aims to compare clustering results between clustering methods that consider spatial aspects and clustering methods that do not consider spatial aspects. By clustering regencies/cities, it is possible to identify pollution levels and air pollution characteristics between regions. Utilizing this knowledge, regional development strategies can be designed to mitigate air pollution by taking into account the unique characteristics of each location. This research contributes to the field of air pollution analysis and management in Java.

2. Material and Methods

2.1. Air Pollution Data

The data utilized in this research was acquired using the Application Programming Interface (API) from the website https://openweathermap.org/. The study utilized regencies/cities on Java Island as observation units, specifically 119 regencies/cities. The study represents air pollution in each regency/city by using the centroid point of the regency/city area. Data collection occurred daily for one week, from June 15, 2023, to June 22, 2023, at 12:00 WIB. The data collection is specifically targeted at periods of high air pollution, which are typically associated with increased population activity. The acquired data is then averaged to determine the pollution level in the regency/city. The grouping of vulnerable regencies/cities is achieved by using air pollution measurement variables, as shown in Table 1.

Table 1. Research variables

Variable	Definition	Units
CO	CO (Carbon monoxide) concentration	$\frac{\mu g}{m^3}$
NO	NO (Nitrogen monoxide) concentration	$\frac{\mu g}{m^3}$
NO_2	NO_2 (Nitrogen dioxide) concentration	$\frac{\mu g}{m^3}$
O_3	O_3 (Ozone) Concentration	$\frac{\mu g}{m^3}$
SO_2	SO_2 (Sulfur dioxide) concentration	$\frac{\mu g}{m^3}$
$PM_{2.5}$	$PM_{2.5}$ (Fine particulate matter) concentration	$\frac{\mu g}{m^3}$
PM_{10}	PM_{10} (Coarse particulate matter) concentration	$\frac{\mu g}{m^3}$
NH_3	NH_3 (Ammonia) Concentration	$\frac{\mu g}{m^3}$

2.2. Fuzzy C Means (FCM)

FCM (Fuzzy C-Means) is one of the fuzzy clustering algorithms that assigns a membership level to each observation in each cluster, ranging from zero to one [23]. The membership level for each observation in the cluster is determined by optimizing an objective function. The most commonly used and studied objective function for FCM clustering is minimizing the least squares error, as in equation (1) [26].

$$J_m(\mathbf{U}, \boldsymbol{v}) = \sum_{i=1}^{N} \sum_{j=1}^{c} \mu_{ij}^{m} d_{ij}^{2} \tag{1}$$

With: $\mathbf{Y}$ represents the set of all N observations, with each observation a vector in D-dimensional space, i.e. $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\} \subset R^D$. c is the number of clusters in the set of all observations $\mathbf{Y}$; $2 \leq c < N - 1$, m is the fuzziness weighting exponent with a value of $1 \leq m < \infty$, μ_{ij} is the element in the i-th row and j-th column of the fuzzy membership matrix, indicating the degree of membership of observation i in cluster j, $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_c)$ is the cluster centroid vector, $\boldsymbol{v}_j = (v_{j1}, v_{j2}, \cdots, v_{jD})$ is the centroid for cluster j, d_{ij}^2 is the squared distance between observation i and cluster j. The fuzzy membership matrix $\mathbf{U} = [\mu_{ij}]$ is initialized using equation (2).

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\| \boldsymbol{y}_i - \boldsymbol{v}_j \|}{\| \boldsymbol{y}_i - \boldsymbol{v}_k \|} \right)^{\frac{2}{m-1}}} \tag{2}$$

The cluster centroids, $\boldsymbol{v}_j$, are obtained using equation (3).

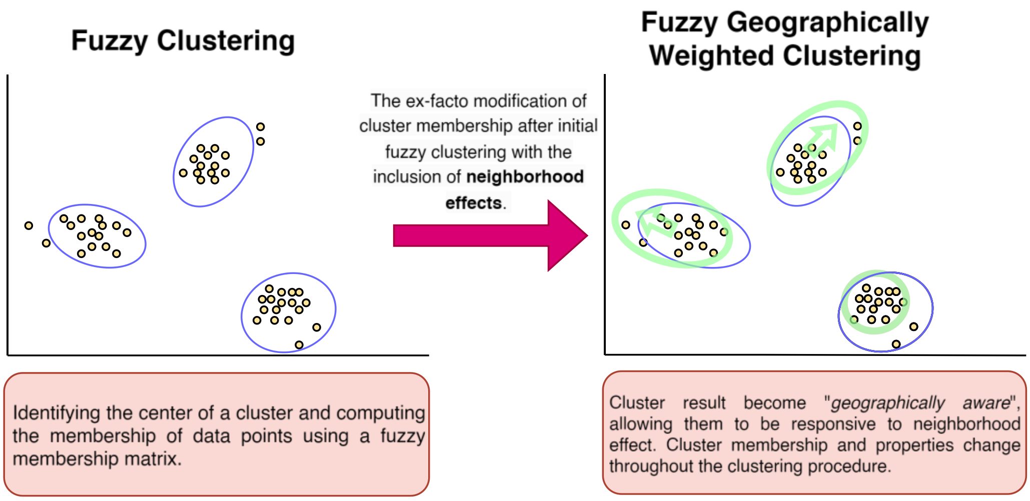
$$\boldsymbol{v}_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m \boldsymbol{y}_i}{\sum_{i=1}^{N} \mu_{ij}^m} \tag{3}$$

The optimal form of fuzzy clustering for dataset $\mathbf{Y}$ is defined as the pair of values $(\mathbf{U}, \boldsymbol{v})$ that minimizes the objective function of equation (1). The optimal pair is obtained through an iteration process. The iteration process will stop when the condition in equation (4), which indicates the achievement of the local optimum [27]. With: ε is the stopping criterion for the iteration, ranging from 0 to 1, t is the iteration index.

$$\| \mu_{ij}^{t+1} - \mu_{ij}^{t} \| < \varepsilon \tag{4}$$

2.3. Fuzzy Geographically Weighted Clustering (FGWC)

The fuzzy c-means (FCM) algorithm was further developed through the integration of spatial features [28]. The resulting algorithm is called Fuzzy Geographically Weighted Clustering (FGWC), which is able to combine the effects of population and distance in cluster analysis in a geographical context so as to enrich the existing fuzzy C-means algorithm [29]. The addition of spatial elements is based on the understanding that cluster centers depend not only on variable values, but also on fuzzy values associated with cluster parameters [28]. Therefore, the spatial component is added by giving weights to the cluster membership formed so as to produce clusters that also consider spatial aspects [28]. FGWC has been applied to various studies that consider spatial aspects, particularly social-environmental studies [23], [24], [25], [29].



Source: Mason & Jacobson [28]

Figure 2. Illustrative comparison of FGWC and fuzzy clustering

Figure 2 shows an illustration of FGWC. In the process of cluster formation, FGWC adopts an approach that considers the influence of a region on other regions through population factors and the distance separating them [28]. The mathematical equation for FGWC can be described as follows. For a new cluster membership μ_i' of area i, it can be expressed as a linear combination of the previous cluster membership μ_i and the membership weight w_{ij} with respect to the weighting factors α and β. Mathematically, the FGWC formulation can be written as in equation (5) [28].

$$\mu_i' = \alpha\mu_i + \beta\frac{1}{A}\sum_{j=1}^{n} w_{ij}\mu_j \tag{5}$$

Where μ_i' is the new cluster membership for area i, μ_i is the previous cluster membership for area i, α and β are the scale parameters of the proportion between the previous cluster membership and the membership weight. The values of α and β must fulfill the requirement $\alpha + \beta = 1$, A is a scale factor calculated for all clusters such that the total membership of an area in each cluster remains 1. In addition, the regional weight w_{ij} in equation (5) can be defined as in equation (6). Where: m_i and m_j represent the population of areas i and j respectively. d_{ij} is the distance between areas i and j, a are parameters defined by the researcher.

$$w_{ij} = \frac{\left(m_i m_j\right)^b}{d_{ij}^a} \tag{6}$$

2.4. *Artificial Bee Colony (ABC)*

FGWC has the disadvantage of being prone to getting trapped in local optima [23], [25], [29]. Artificial Bee Colony (ABC) is one of the optimization approaches that can be used to minimize this limitation [30], [31]. ABC is an optimization algorithm inspired by the behavior of a group of bees in search of optimal food sources [32], [33]. The search for the optimal value in ABC is analogous to the search for the optimal food source by bees in the search space [32]. The ABC algorithm consists of three stages involving three types of bees, namely employed bees, onlooker bees, and scout bees. The stages of the ABC algorithm can be explained as follows [32].

First, employed bees travel back and forth between the nest and the food source location. Each employed bee is associated with a specific food source and remembers its location. They not only return to the hive with the food source, but also conduct local exploration to find better food sources.

Second, onlooker bees observe the behavior of employed bees returning from the food source search process. Onlooker bees are not associated with a particular food source location, and they probabilistically observe the behavior of employed bees to determine the best food source location. This selection probability can be calculated using equation (7) [34]. Where fit_i is the fitness value of the i-th solution that is directly proportional to the quality of the food source at the i-th location, and n_{emp} is the number of employed bees.

$$p_i = \frac{fit_i}{\sum_{q=1}^{n_{emp}} fit_q} \tag{7}$$

Third, scout bees observe the behavior of employed bees when returning to the nest. Scout bees are not associated with a specific food source location, and their job is to find new food sources. If scout bees observe that employed bees do not find progressively better food sources, they randomly search for new food source locations in the search space. In the ABC algorithm, the search for a new food source is expressed in equation (8) [34].

$$x_{ij}^* = x_{ij} + \phi_{ij}(x_{ij} - x_{lj}) \tag{8}$$

Where x_{ij}^* is a new candidate food source location calculated by comparing the previous food source x_{ij} with a random food source location, $l \in \{1,2,\cdots,n_{emp}\}$, and $j \in \{1,2,\cdots,D\}$. The notation ϕ_{ij} represents a number in the range $[-1,1]$ that is randomly generated.

2.5. Analysis Flow

This study employs FGWC parameters $m = 2$, $\alpha = 0.7$, $\beta = 0.3$, $a = 1$, $b = 1$, and $\varepsilon = 10^{-5}$, which are the default settings in the naspaclust package [35]. Meanwhile, for the ABC algorithm, the parameters $n_{food} = 10, n_{onlk} = 5$, and limit $= 4$ were used. The analysis process was carried out using Rstudio 4.2.1 software by utilizing several statistical analysis packages and data visualization. The flow of analysis in this study can be described as follows.

1. First, the data was retrieved through the openweather API by determining the regency/city center point as the area to be retrieved from the API. The data is then preprocessed using Rstudio.
2. Next, cluster analysis was conducted by forming 2 to 6 clusters using both clustering methods. The range of the number of clusters formed up to 6 refers to the categorization of AQI, which also amounts to 6 categories [36]. The determination of the optimal number of clusters in both methods is done using the cluster validity index, the TSS (Tang-Sun-Sun) Index [37]. The TSS index is a development of the Xie-Beni Index and Kwon Index to overcome the problem of monotonically increasing or monotonically decreasing index values by imposing a penalty on the calculation of index validity, formulated in equation (9) [37].

$$I_{TSS} = \frac{1}{\min\limits_{j\neq k}\left\|\boldsymbol{v}_j - \boldsymbol{v}_k\right\|^2 + \frac{1}{c}}\left(\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^2\left\|\boldsymbol{x}_i - \boldsymbol{v}_j\right\|^2 + \frac{1}{c(c-1)}\sum_{j=1}^{c}\sum_{k=1,k\neq i}^{c}\left\|\boldsymbol{v}_j - \boldsymbol{v}_k\right\|^2\right) \tag{9}$$

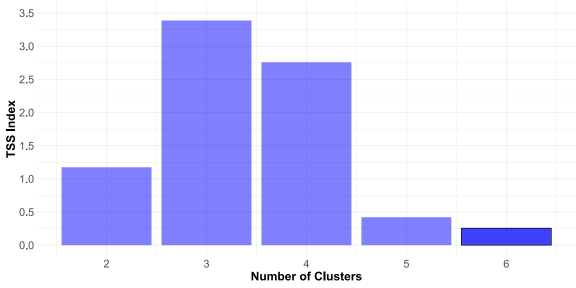
   Where μ_{ij} i is the degree of membership of the i-th observation to the j-th cluster, $\mathbf{x}_i$ is a vector indicating the value for the i-th observation, $\mathbf{v}_j$ and $\mathbf{v}_k$ are vectors indicating the centroid of the j-th and k-th clusters, n indicates the number of observations, and c is the number of clusters. The lower the I_{TSS} value, the better the cluster formed [37].
3. After the clusters are formed, profiling is done on the results of the clusters formed. Profiling is done by interpreting the cluster centroid results and showing biplots to support clustering.
4. The comparison between clustering methods is then carried out using the TSS index [37] and the Rand index [38]. The Rand index is a measure of the similarity of cluster results from two different clustering methods on the same data [38]. The Rand index is formulated in equation (10) [38].

$$c(Y, Y') = \frac{\binom{N}{2} - \frac{1}{2}\left\{\sum_i\left(\sum_j n_{ij}\right)^2 + \sum_j\left(\sum_i n_{ij}\right)^2\right\} - \sum_i\sum_j n_{ij}^2}{\binom{N}{2}} \tag{10}$$

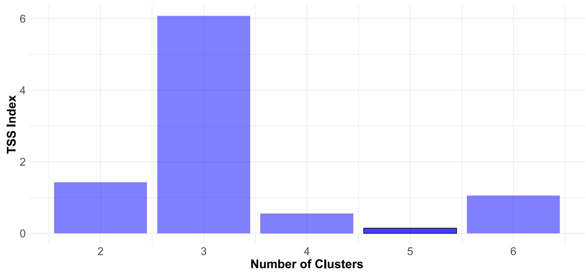
   Where: n_{ij} is the number of data grouped into the i-th cluster in the clustering method Y and grouped into the j-th cluster in the clustering method Y'. In addition, the notation N represents the number of clustered data. The Rand index value is between 0 and 1 with 1 denoting that the cluster results between the two methods are identical [38]. The better clustering method is then interpreted and discussed further.

3. Results and Discussion

This section explains the results and discusses cluster formation in the case of air pollution in Java. To compare the two cluster formation methods, the optimal number of clusters for each method is first determined. The determination of the optimal number of clusters is done using the TSS index. Clusters will be created using both methods, with the TSS index determining the optimal number of clusters. Next, the cluster results will be compared. After obtaining the best method, the better cluster method is interpreted, and the cluster results are discussed to analyze the case of air pollution on Java Island.



(a)



(b)

Figure 3. Determining the number of clusters with TSS index in clustering methods (a) Fuzzy C Means; (b) ABCFGWC

3.1. *Optimal Number of Clusters*

Figure 3 shows the TSS index graph for clusters formed using the FCM and ABCFGWC methods, with cluster numbers ranging from 2 to 6. In the FCM method, it can be seen that the lowest TSS index value in the cluster range of 2 to 6 is the TSS index for 6 clusters. As a result, the optimal number of clusters using FCM is six. Thus, six clusters were formed using the FCM method. On the other hand, in the ABCFGWC method, it can be seen that the lowest TSS index value for the range of clusters 2 to 6 is the TSS index for 5 clusters. As a result, using the ABCFGWC method, the optimal number of clusters is five. Thus, five clusters were formed using the ABCFGWC method.

3.2. *Fuzzy C Means (FCM) Method*

Based on the TSS index, a total of six clusters were formed using FCM. Interpretation of the cluster results can be done using the center of the cluster results. Table 2 displays the cluster centers of the 6 clusters formed using the FCM method.

Table 2. Centroid fuzzy c means (FCM) clustering result

Cluster	CO	NO	NO_2	O_3	SO_2	$PM_{2.5}$	PM_{10}	NH_3
1	432.94	0.32	4.07	76.64	5.29	18.77	26.78	5.33
2	1416.28	0.92	21.35	179.71	20.95	104.10	124.82	16.96
3	3282.25	2.85	63.72	281.21	49.14	243.95	275.12	5.33
4	7809.41	3.46	116.40	639.47	69.80	721.95	832.20	1.53
5	833.50	0.43	9.34	100.76	8.50	52.69	69.99	13.18
6	2077.22	0.99	2.36	193.18	32.79	174.00	200.78	12.95

Source: Authors' Calculation

According to Table 2, Cluster 4 tends to have the highest air pollution compared to other clusters, except for the NH_3 pollutant. The next cluster that has the highest air pollution after Cluster 4 is Cluster 3. In contrast to Cluster 4, Cluster 3 tends to have lower air pollution values for all air pollutants except for the NH_3 pollutant. When it comes to the NH_3 pollutant, Cluster 3 actually has a higher average value than Cluster 4. Meanwhile, the other clusters exhibit their own unique patterns. For example, Cluster 6 has high pollution levels, ranking just below clusters 3 and 4 in most pollutants but has the lowest NO_2 pollutant among all clusters. Another example is Cluster 2, which has the highest NH_3 pollutant but is not very prominent in other pollutants.

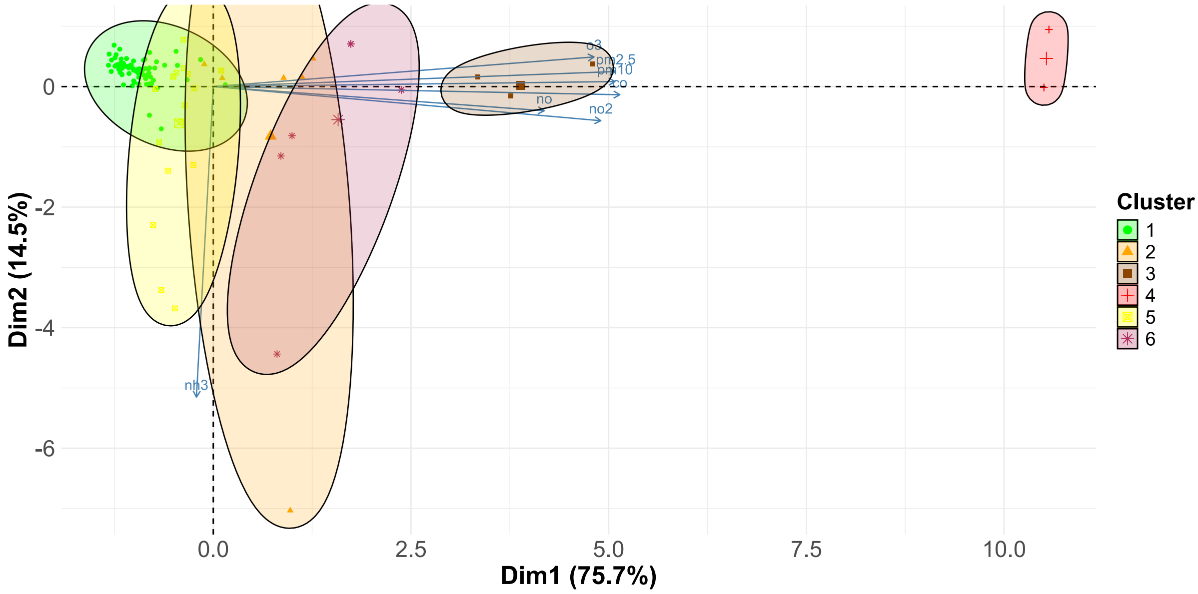


Figure 4. Biplot of principal component analysis (PCA) fuzzy c means (FCM) clustering results

In addition to using cluster centroids, biplots can be used to support profiling and show clustering in the clusters formed. Figure 4 shows a biplot using the PCA technique based on the cluster results obtained with the FCM method. Using PCA, dimension 1 is able to explain 75.7% of the data variation, while dimension 2 is able to explain 14.5% of the data variation. Thus, using only these two dimensions, the data has been well explained because it has explained more than 90% of the data variation. The majority of air pollution is explained by dimension 1 in the PCA results, except for NH_3 pollutants.

Examining Dimension 1 yields similar results, indicating that Cluster 4 typically has the highest air pollution compared to the other clusters. Figure 4 also supports Table 3 by showing that Cluster 3 is the most vulnerable cluster to air pollution after Cluster 4. On the other hand, Cluster 1 has the lowest air pollution compared to other clusters. Cluster 1, with its cluster center closest to the origin, demonstrates this result. When considering dimension 2, it can also be seen that Cluster 2, Cluster 5, and Cluster 6 are clusters characterized by high NH_3 pollutants. In addition, Cluster 2 also appears to have relatively higher NH_3 pollutants than the other clusters, as indicated by Cluster 2's elongated ellipse in dimension 2. Thus, the biplot shows similar results to the cluster centroid in Table 3.

3.3. Artificial Bee Colony (ABC) Fuzzy Geographically Weighted Clustering (FGWC) Method

The next clustering method implemented in this research is ABCFGWC. Based on the TSS index, a total of 5 clusters were formed using ABCFGWC. Similar to FCM, the interpretation of the cluster results is done using the center of the cluster results. Table 3 displays the cluster centers of the 5 clusters formed using the ABCFGWC method.

Table 3. Centroid artificial bee colony (ABC) fuzzy geographically weighted clustering (FGWC) result

Cluster	CO	NO	NO_2	O_3	SO_2	$PM_{2.5}$	PM_{10}	NH_3
1	1962.00	1.02	27.74	191.95	29.46	159.75	186.28	16.00
2	900.00	0.44	10.15	109.25	9.61	59.53	77.31	13.12
3	3264.53	2.78	62.34	282.00	48.61	243.32	275.23	5.42
4	7692.77	3.46	115.38	628.55	69.42	709.67	817.87	1.61
5	460.99	0.35	4.58	78.13	5.65	20.82	29.44	5.88

Source: Authors' Calculation

Table 3 displays the cluster centers of the 5 clusters formed using the ABCFGWC method. According to Table 3, Cluster 4 tends to have the highest air pollution compared to other clusters, except for NH_3 pollution. The next cluster that has the highest air pollution after Cluster 4 is Cluster 3. In contrast to Cluster 4, Cluster 3 tends to have high air pollution values for all air pollutants, except for the NH_3 pollutant. When it comes to the NH_3 pollutant, Cluster 3 actually has a higher average value than Cluster 4. The ABCFGWC method yields similar results to the FCM method. Other clusters exhibit characteristics unique to each cluster.

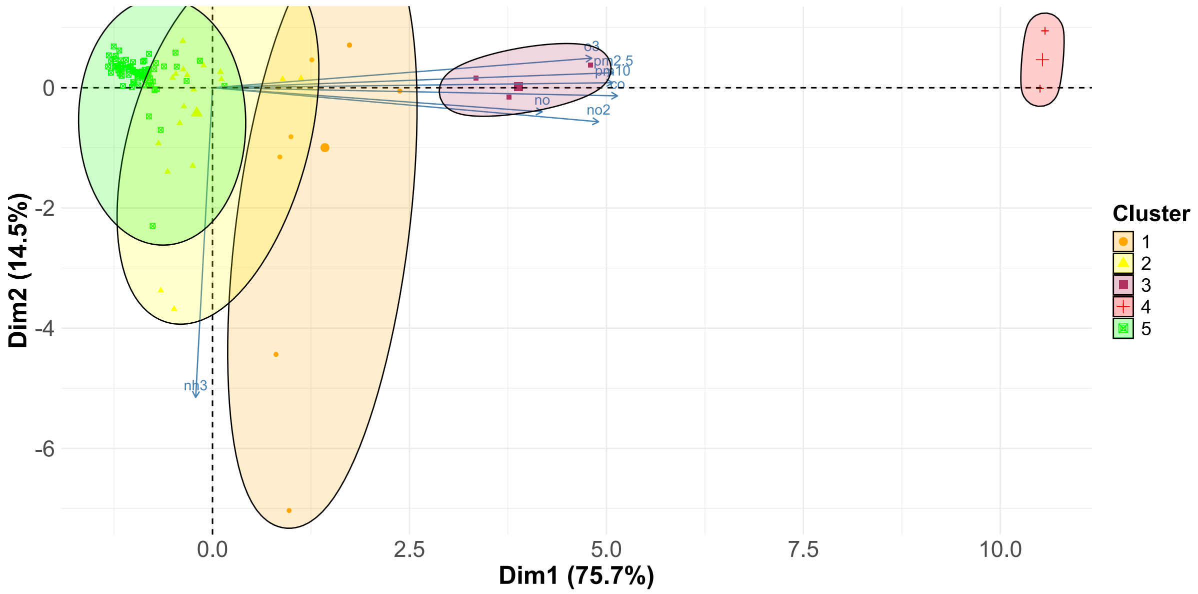


Figure 5. Biplot of principal component analysis (PCA) ABCFGWC (Artificial Bee Colony Fuzzy Geographically Weighted Clustering) results

Figure 5 illustrates the biplot using the PCA technique, based on the cluster results obtained through the ABCFGWC method. Using PCA, dimension 1 can explain 75.7% of the data variation, while dimension 2 can explain 14.5% of the data variation. Similar to the previous findings, the majority of air pollution is explained by dimension 1 in the PCA results, except for NH3 pollutants. Examining dimension 1 yields similar results once more, suggesting that Cluster 4 typically experiences the highest levels of air pollution. Figure 5 also supports Table 3 by showing that Cluster 3 is the most vulnerable cluster to air pollution after Cluster 4. The results with the ABCFGWC method are similar to those obtained with the FGWC method, but there is a difference: in the ABCFGWC method, the separation between clusters is more visible. In addition, it should be noted that the numbering between the two clustering methods is the result of random initialization.

3.4. Comparison of the Two Clustering Methods

A comparison between the two clustering methods is done using the TSS index and Rand index values. The TSS index value is used to determine the suitability between the clusters formed with a better clustering method that will produce a smaller TSS index value. According to Table 4, the lowest TSS index values for the FCM and ABCFGWC methods are 0.257 and 0.145, respectively, in the range of clusters 2 to 6. As a result, the clusters formed by the ABCFGWC method outperform FCM based on the TSS index.

Table 4. Comparison of clustering methods by TSS index and rand index

Clustering Method	TSS Index for Number of Clusters					Rand Index
	2	3	4	5	6	
FCM	1.175	3.389	2.759	0.422	0.257	0.945
ABCFGWC	1.429	6.078	0.547	0.145	1.056	

Source: Authors' Calculation

However, the cluster membership formed using this method does not show a significant difference. The Rand index can be used to determine the suitability of members between the two clusters that formed. The Rand Index has a range of values between 0 and 1, with the higher values indicating more identical clusters [38]. Using the Rand Index, as many as six clusters generated using FCM are compared with as many as five clusters generated using ABCFGWC. The resulting Rand Index value is close to

1, indicating that these two clusters are quite similar despite having a different number of optimal clusters.

However, the Rand Index method measures the suitability of cluster members without considering the fuzzy membership degree but directly uses the cluster membership results [38]. Meanwhile, the FCM and ABCFGWC methods have different TSS index values, even showing a difference in the optimal number of clusters between the two methods. With the TSS index, the degree of cluster membership is considered for each observation. Therefore, the difference in TSS index value shows that there is a difference in the degree of membership between the two clustering methods. The lower TSS index value in the ABCFGWC method indicates that the results of clustering air pollution improve when the neighborhood effect is taken into account and the fuzzy membership degree is adjusted accordingly. Previous researchers found similar results, indicating that FGWC clustering outperforms FCM in the presence of regional aspects [39].

3.5. Discussion of Cluster Results

The clustering results of these two methods are depicted in a map visualization. Figure 6 depicts the visualization map of the clustering outcomes obtained from the FCM method. Meanwhile, Figure 7 depicts the visualization map of the clustering results obtained from the ABCFGWC method. It is evident that the two maps in Figure 6 and Figure 7 show a similar pattern of results. It should be noted, however, that the numbering between the two clustering methods is the result of random initialization. The difference in results between the two methods can be seen in a small number of regencies/cities in the eastern and western parts of Java Island. The absence of striking differences between the two clustering methods also supports the results of the rand index calculation.

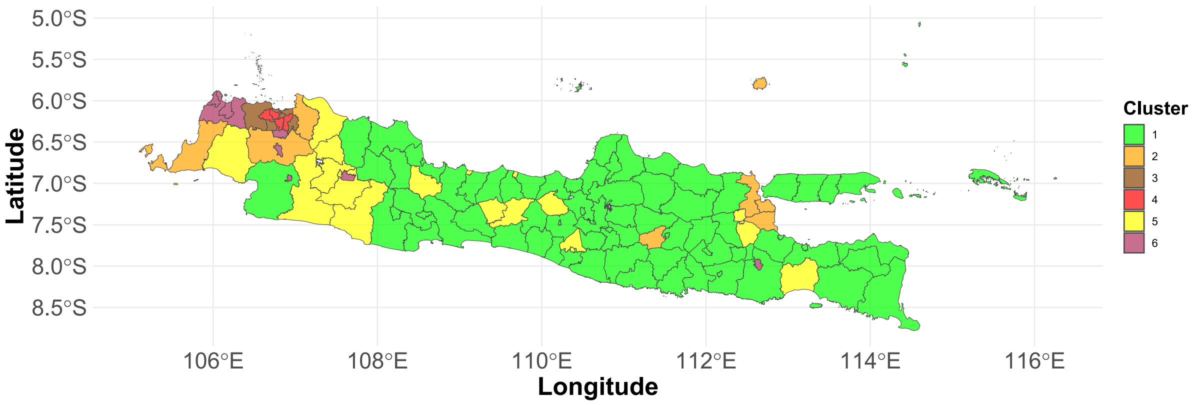


Figure 6. Map of cluster results of regency/city in Indonesia based on air pollution by fuzzy c means (FCM) approach

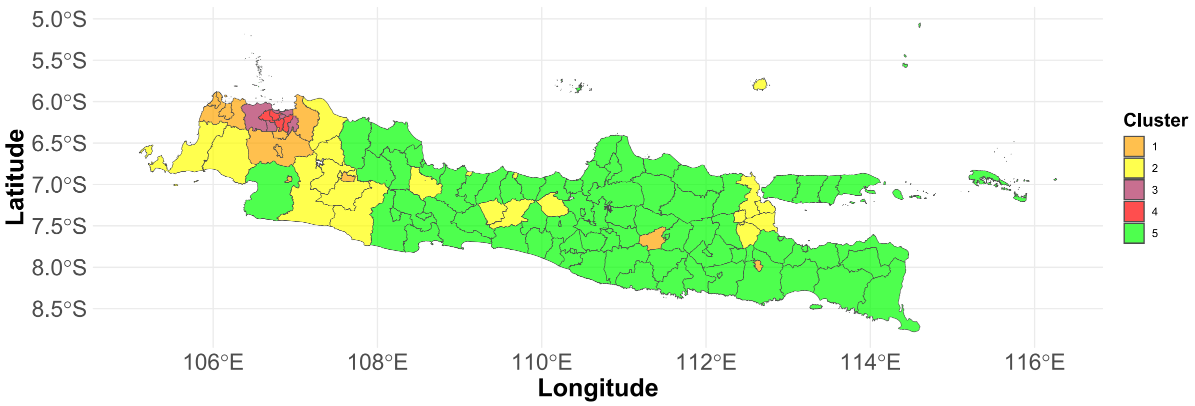


Figure 7. Map of cluster results of regency/city in Indonesia based on air pollution by artificial bee colony (ABC) fuzzy geographically weighted clustering (FGWC) approach

The FCM method, as shown in Figure 6, creates clusters solely based on data values. This approach leads to some differences in results for certain areas in the eastern and western parts of Java Island compared to the ABCFGWC method depicted in Figure 7. Meanwhile, the ABCFGWC method forms clusters by considering the distance between each region. This method makes it appear as though the regencies and cities in the eastern and western parts of Java Island are grouped together. In addition, the TSS Index result in the ABCFGWC method is lower, so the interpretation and discussion of the cluster results will be carried out in the ABCFGWC clustering method. This result is supported by the nature of air pollution, which has inter-regional linkages that are considered in the ABCFGWC method but not considered in the FCM method.

Figure 7 shows the distribution of regencies/cities on Java Island based on air pollution using ABCFGWC. The Jabodetabek area is classified as a member of Cluster 4, which, according to the profiling results, is a regency/city with highly vulnerable air pollution characteristics. Greater Jakarta or Jabodetabek, comprises Jakarta Special Capital Region and the surrounding regencies/cities of Bogor, Depok, Tangerang, and Bekasi. With dense economic and government activities, the Jabodetabek area is classified as a cluster that has high air pollution, which was also found in a related study [40]. This result aligns with Tian et al. [16], who found that air pollution is more closely linked to industrial structure than to geographical and socioeconomic conditions. However, measures to address high air pollution in industrial areas cannot be done by directly halting the industrial process. Reducing industrial and transportation activities will help reduce air pollution, but such action is not sufficient if the area is experiencing critical levels of pollution [41].

The results reveal that, in addition to the Jabodetabek area, several regencies/cities in the vicinity also fall into air pollution-vulnerable clusters, albeit not as vulnerable as Jabodetabek. In general, the eastern part of Java Island does not have any regions or cities that are susceptible to air pollution. Regencies/cities in the eastern part of Java Island mostly belong to Cluster 1, and only a small number belong to clusters apart from Cluster 1. Regencies/cities in the eastern part of Java Island tend to cluster close to the city of Surabaya, which is influenced by the dense economic and business activities. However, Figure 7 shows that the western part of Java Island is highly vulnerable to air pollution, especially the Jabodetabek area. The relationship between pollution and income typically follows an inverted U-shaped curve [42]. It starts in areas with low income and minimal pollution. When investments flow into such underdeveloped regions, it often happens because these areas have regulations allowing industry to be free to pollute. As a result, both income and pollution increase until they reach a peak. At this stage, the region becomes wealthy enough to prioritize environmental concerns. Over time, continued economic growth leads to reduced pollution as sustainable practices and cleaner technologies are adopted.

In Indonesia, particularly in Java, the region is still in the developing phase. This means the relationship between economic development and pollution has not yet reached its peak. Jakarta and Surabaya, which have higher economic output and GDP, also experience significantly higher pollution levels compared to other regions.

Table 5. Details of air quality index (AQI) and pollutant concentration range

AQI Category	Range of Pollutant Concentrations $\left(\frac{\mu g}{m^3}\right)$							
	CO	NO	NO_2	O_3	SO_2	$PM_{2.5}$	PM_{10}	NH_3
Good	0-1,000	0-40	0-40	0-50	0-40	0-30	0-50	0-200
Satisfactory	1,000-2,000	41-80	41-80	51-100	41-80	31-60	51-100	201-400
Moderately Polluted	2,000-10,000	81-180	81-180	101-168	81-380	61-90	101-250	401-800
Poor	10,000-17,000	181-280	181-280	169-208	381-800	91-120	251-350	801-1200
Very Poor	17,000-34,000	281-400	281-400	209-748	801-1,600	121-250	351-430	1,201-1,800
Severe	34,000+	400+	400+	748+	1,600+	250+	430+	1,800+

Source: Kumar & Yadav [36]

Pollution in the Jabodetabek area can be said to be quite serious when compared to the AQI (Air Quality Index) settings for each pollutant. Table 5 shows the settings for the air quality index for each pollutant [36]. By examining the ABCFGWC cluster centroid results in Table 3 and comparing them with each pollutant, it can be seen that some pollutants in Cluster 4 are rated as moderately polluted or worse. On the other hand, NO, SO_2, and NH_3 pollutants in Cluster 4 are classified as satisfactory or good. Of all pollutants, the pollutants that need to be the main concern in Cluster 4 are $PM_{2.5}$ and PM_{10}, which are classified as severe. These two pollutants (particulate matter) are frequently mentioned by several sources as one of the most dangerous because of their carcinogenic nature and links to diseases of several organ systems [43]. As a result, policies to reduce air pollution must be applied to clusters of areas with high pollutant values, namely Cluster 4 and Cluster 3, with a particular emphasis on $PM_{2.5}$ and PM_{10} pollutants.

Some regions have implemented policies to reduce air pollution. For example, countries that have implemented policy measures to address air pollution are China and India. In 2014, China began implementing the "war against pollution" program and has successfully curtailed particulate pollution by 30% [13]. China established a policy of reducing air pollution through green technology reform and the use of green energy [14]. On the other hand, India implemented the National Clean Air Program (NCAP) in the form of city-specific air pollution policies along with a national strategy [44]. In addition to policies, the WHO proposes whole-community collaborative measures to reduce air pollution. WHO [2] conducts training for health workers to have the skills to implement policies and strategies for reducing air pollution. These health workers are then expected to become trainers and community mobilizers, ready to conduct innovative programs to address air pollution; this approach is called train-the-trainer [2].

Indonesia has implemented several measures to address air pollution, particularly particulate matter. Most of the policies in Indonesia focus on the transportation sector, as this sector is the biggest source of air pollution [45]; for example, the Indonesian government requires Euro-4 standard fuel from September 2018 [5]. Additionally, Indonesia has addressed air pollution by resolving peatland and forest fires and implementing coal regulations [5]. Nonetheless, an innovative step needs to be implemented to further accelerate the process of addressing air pollution problems in Indonesia. This innovative step can be achieved by involving all community groups in an air pollution reduction measure [2]. For instance, previous studies [3], [46] have proposed the integration of air pollution reduction and environmental sustainability into education. In addition, other steps can also be taken by transforming the transportation sector in Indonesia, for example, using renewable energy transportation [47].

4. Conclusion

We conducted a clustering analysis of regencies and cities on Java Island using the fuzzy clustering approach. This study further compares the performance of the fuzzy c-means (FCM) method with the ABCFGWC method, which incorporates regional factors. The findings indicate that the ABCFGWC method, due to its consideration of regional aspects, performs better in addressing air pollution cases than the FCM method. The ABCFGWC approach identified five clusters, with the Jabodetabek region classified as part of a cluster highly vulnerable to air pollution, exceeding the Air Quality Index (AQI) threshold. This work is intended to provide stakeholders with a basis for developing regional strategies to mitigate air pollution on Java Island.

There remains significant potential for further development of this study in terms of data, methodologies, and substantive focus. Future work can be advanced by employing panel data to examine the temporal dynamics of air pollution. Additionally, subsequent studies could broaden the geographic scope to include all regencies and cities across Indonesia or explore alternative clustering methods. Future studies may also investigate the interplay between air pollution issues and various socioeconomic challenges.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Arya Candra Kusuma: Conceptualization, Methodology, Data Collection, Data Analysis, Writing–Original Draft. **Arie Wahyu Wijayanto**: Conceptualization, Methodology, Manuscript Review, Research Advisor. **Arista**: Editing, Writing–Review. **Vicka Kharisma Bahar**: Editing, Writing–Review. **Tifani Husna Siregar**: Editing, Writing–Review.

References

[1] G. Shaddick, M. L. Thomas, P. Mudu, G. Ruggeri, and S. Gumy, "Half the world's population are exposed to increasing air pollution," *NPJ Clim Atmos Sci*, vol. 3, no. 1, Dec. 2020, doi: 10.1038/s41612-020-0124-2.

[2] WHO, "Mapping opportunities for training in air pollution and health for the health workforce," 2021.

[3] Y.-J. Huo, K.-T. Shih, and C.-J. Lin, "The study on integrating air pollution environmental education into the teaching personal and social responsibility model in physical education," *IOP Conf Ser Earth Environ Sci*, vol. 576, no. 1, p. 012006, Nov. 2020, doi: 10.1088/1755-1315/576/1/012006.

[4] A. A. Almetwally, M. Bin-Jumah, and A. A. Allam, "Ambient air pollution and its influence on human health and welfare: an overview," *Environmental Science and Pollution Research*, vol. 27, no. 20, pp. 24815–24830, Jul. 2020, doi: 10.1007/s11356-020-09042-2.

[5] K. Lee and M. Greenstone, "Polusi udara indonesia dan dampaknya terhadap usia harapan hidup [Indonesia's air pollution and its impact on life expectancy]," 2021. Accessed: Jun. 04, 2023. [Online]. Available: https://aqli.epic.uchicago.edu/wp content/uploads/2021/09/AQLI_IndonesiaReport-2021_IND-version9.7.pdf

[6] D. A. Glencross, T. R. Ho, N. Camiña, C. M. Hawrylowicz, and P. E. Pfeffer, "Air pollution and its effects on the immune system," *Free Radic Biol Med*, vol. 151, pp. 56–68, May 2020, doi: 10.1016/j.freeradbiomed.2020.01.179.

[7] A. J. White, J. P. Keller, S. Zhao, R. Carroll, J. D. Kaufman, and D. P. Sandler, "Air pollution, clustering of particulate matter components, and breast cancer in the sister study: a U.S.-wide cohort," *Environ Health Perspect*, vol. 127, no. 10, Oct. 2019, doi: 10.1289/EHP5131.

[8] H. Kim *et al.*, "Cardiovascular effects of long-term exposure to air pollution: a population-based study with 900 845 person-years of follow-up," *J Am Heart Assoc*, vol. 6, no. 11, Nov. 2017, doi: 10.1161/JAHA.117.007170.

[9] L. A. Rodriguez-Villamizar, A. Magico Bsc, A. Osornio-Vargas, and B. H. Rowe, "The effects of outdoor air pollution on the respiratory health of Canadian children: a systematic review of epidemiological studies," *Can Respir J*, vol. 22, no. 5, pp. 282–292, 2015, doi: 10.1155/2015/263427.

[10] A. Ebenstein, V. Lavy, and S. Roth, "The long-run economic consequences of high- stakes examinations: evidence from transitory variation in pollution," *Am Econ J Appl Econ*, vol. 8, no. 4, pp. 36–65, 2016, doi: 10.1257/app.20150213.

[11] R. Hanna and P. Oliva, "The effect of pollution on labor supply: evidence from a natural experiment in Mexico City," *J Public Econ*, vol. 122, pp. 68–79, Feb. 2015, doi: 10.1016/j.jpubeco.2014.10.004.

[12] M. Dobranschi, D. Nerudová, V. Solilová, and K. Stadler, “Carbon border adjustment mechanism challenges and implications: the case of Visegrád countries,” *Heliyon*, vol. 10, no. 10, p. e30976, May 2024, doi: 10.1016/j.heliyon.2024.e30976.

[13] M. Greenstone and C. Fan, “Is China winning its war on pollution?,” 2020. Accessed: Jun. 04, 2023. [Online]. Available: https://aqli.epic.uchicago.edu/wp-content/uploads/2021/02/China-Report_2020updateGlobal.pdf

[14] P. Wang, “China’s air pollution policies: progress and challenges,” *Curr Opin Environ Sci Health*, vol. 19, Feb. 2021, doi: 10.1016/j.coesh.2020.100227.

[15] M. Greenstone and Q. Fan, “Indonesia’s worsening air quality and its impact on life expectancy,” 2019. Accessed: Jun. 08, 2023. [Online]. Available: https://aqli.epic.uchicago.edu/wp-content/uploads/2019/03/Indonesia-Report.pdf

[16] D. Tian *et al.*, “Characteristic and spatiotemporal variation of air pollution in Northern China based on correlation analysis and clustering analysis of five air pollutants,” *Journal of Geophysical Research: Atmospheres*, vol. 125, no. 8, Apr. 2020, doi: 10.1029/2019JD031931.

[17] G. R. Kingsy, R. Manimegalai, D. M. S. Geetha, S. Rajathi, K. Usha, and B. N. Raabiathul, “Air pollution analysis using enhanced k-means clustering algorithm for real time sensor data,” in *2016 IEEE Region 10 Conference (TENCON)*, Singapore: Institute of Electrical and Electronics Engineers (IEEE)., Feb. 2017, pp. 1945–1949. doi: 10.1109/TENCON.2016.7848362.

[18] Z. Xu *et al.*, “Classification of urban pollution levels based on clustering and spatial statistics,” *Atmosphere (Basel)*, vol. 13, no. 3, Mar. 2022, doi: 10.3390/atmos13030494.

[19] S. Annas, U. Uca, I. Irwan, R. H. Safei, and Z. Rais, “Using k-means and self organizing maps in clustering air pollution distribution in Makassar City, Indonesia,” *Jambura Journal of Mathematics*, vol. 4, no. 1, pp. 167–176, Jan. 2022, doi: 10.34312/jjom.v4i1.11883.

[20] A. B. Santoso, A. C. Candra, R. Nooraeni, and A. W. Wijayanto, “Development of a hybrid fuzzy geographically weighted k-prototype clustering and genetic algorithm for enhanced spatial analysis: application to rural development mapping,” *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol. 16, no. 2, pp. 122–139, Dec. 2024, doi: 10.34123/jurnalasks.v16i2.789.

[21] A. Páez, M. Trépanier, and C. Morency, “Geodemographic analysis and the identification of potential business partnerships enabled by transit smart cards,” *Transp Res Part A Policy Pract*, vol. 45, no. 7, pp. 640–652, 2011, doi: 10.1016/j.tra.2011.04.002.

[22] E. Austin, B. A. Coull, A. Zanobetti, and P. Koutrakis, “A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition,” *Environ Int*, vol. 59, pp. 244–254, Sep. 2013, doi: 10.1016/j.envint.2013.06.003.

[23] B. I. Nasution, R. Kurniawan, T. H. Siagian, and A. Fudholi, “Revisiting social vulnerability analysis in Indonesia: an optimized spatial fuzzy clustering approach,” *International Journal of Disaster Risk Reduction*, vol. 51, Dec. 2020, doi: 10.1016/j.ijdrr.2020.101801.

[24] B. I. Nasution, F. M. Saputra, R. Kurniawan, A. N. Ridwan, A. Fudholi, and B. Sumargo, “Urban vulnerability to floods investigation in Jakarta, Indonesia: a hybrid optimized fuzzy spatial clustering and news media analysis approach,” *International Journal of Disaster Risk Reduction*, vol. 83, p. 103407, Dec. 2022, doi: 10.1016/j.ijdrr.2022.103407.

[25] R. E. Caraka *et al.*, “Micro, small, and medium enterprises’ business vulnerability cluster in Indonesia: an analysis using optimized fuzzy geodemographic clustering,” *Sustainability*, vol. 13, no. 14, Jul. 2021, doi: 10.3390/su13147807.

[26] J. C. Bezdek, R. Ehrlich, and W. Full, “FCM: The fuzzy c-means clustering algorithm,” *Comput Geosci*, vol. 10, no. 2–3, pp. 191–203, 1984, doi: 10.1016/0098-3004(84)90020-7.

[27] G. Grekousis, “Local fuzzy geographically weighted clustering: a new method for geodemographic segmentation,” *International Journal of Geographical Information Science*, vol. 35, no. 1, pp. 152–174, 2021, doi: 10.1080/13658816.2020.1808221.

[28] G. A. Mason and R. D. Jacobson, “Fuzzy Geographically Weighted Clustering,” 2006, Accessed: Aug. 04, 2023. [Online]. Available: https://www.researchgate.net/publication/255650957

[29] A. W. Wijayanto, A. Purwarianti, and L. H. Son, “Fuzzy geographically weighted clustering using artificial bee colony: an efficient geo-demographic analysis algorithm and applications to the analysis of crime behavior in population,” *Applied Intelligence*, vol. 44, no. 2, pp. 377–398, Mar. 2016, doi: 10.1007/s10489-015-0705-7.

[30] A. W. Wijayanto, S. Mariyah, and A. Purwarianti, “Enhancing clustering quality of fuzzy geographically weighted clustering using ant colony optimization,” in *4th International Conference on Data and Software Engineering (ICoDSE 2017)*, Palembang, Indonesia: institute of electrical and electronics engineers (IEEE), 2018. doi: 10.1109/ICODSE.2017.8285858.

[31] A. W. Wijayanto and A. Purwarianti, “Improvement design of fuzzy geo-demographic clustering using artificial bee colony optimization,” in *2014 International Conference on Cyber and IT Service Management (CITSM)*, 2014, pp. 69–74. doi: 10.1109/CITSM.2014.7042178.
[32] D. Simon, *Evolutionary optimization algorithms*, 1st ed. Wiley, 2013.
[33] D. Karaboga and B. Basturk, “On the performance of artificial bee colony (ABC) algorithm,” *Applied Soft Computing Journal*, vol. 8, no. 1, pp. 687–697, Jan. 2008, doi: 10.1016/j.asoc.2007.05.007.
[34] F. De Rango, N. Palmieri, and M. Tropea, “Multirobot coordination through bio-inspired strategies,” in *Nature-Inspired Computation and Swarm Intelligence*, 1st ed., X.-S. Yang, Ed., Academic Press, 2020, ch. 19, pp. 361–390. doi: 10.1016/b978-0-12-819714-1.00030-0.
[35] B. I. Nasution, R. Kurniawan, and R. E. Caraka, “naspaclust: Nature-inspired spatial clustering,” 2021. [Online]. Available: https://cran.r-project.org/package=naspaclust
[36] S. Kumar and S. Yadav, “Air quality index and criteria pollutants in ambient atmosphere over selected sites: impact and lessons to learn from covid-19,” in *Environmental Resilience and Transformation in Times of Covid-19*, 1st ed., Elsevier, 2021, ch. 15, pp. 153–162. doi: 10.1016/c2020-0-02703-9.
[37] Y. Tang, F. Sun, and Z. Sun, “Improved validation index for fuzzy clustering,” in *Proceedings of the American Control Conference*, 2005, pp. 1120–1125. doi: 10.1109/acc.2005.1470111.
[38] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *J Am Stat Assoc*, vol. 66, no. 336, pp. 846–850, 1971, doi: 10.2307/2284239.
[39] A. W. Wijayanto, “Improvement of fuzzy geo-demographic clustering using metaheuristic optimization on Indonesia population census,” 2015. Accessed: Jun. 12, 2023. [Online]. Available: https://www.researchgate.net/publication/279977406
[40] S. Kautsar, R. R. Ridwansyah, and N. Priscilla, “Formulation of greenhouse gas emission index and strategy towards net zero emission for cities on the java island in support of golden Indonesia 2045,” *Seminar Nasional Official Statistics*, vol. 2024, no. 1, pp. 591–602, Nov. 2024, doi: 10.34123/semnasoffstat.v2024i1.2119.
[41] P. Wang, K. Chen, S. Zhu, P. Wang, and H. Zhang, “Severe air pollution events not avoided by reduced anthropogenic activities during COVID-19 outbreak,” *Resour Conserv Recycl*, vol. 158, Jul. 2020, doi: 10.1016/j.resconrec.2020.104814.
[42] B. Özcan and I. Öztürk, *Environmental Kuznets Curve (EKC)*. Elsevier, 2019. doi: 10.1016/C2018-0-00657-X.
[43] P. J. Landrigan *et al.*, “The lancet commission on pollution and health,” *The Lancet Comission*, vol. 391, no. 10119, pp. 462–512, Feb. 2018, doi: 10.1016/S0140-6736(17)32345-0.
[44] M. Greenstone and Q. Fan, “India’s ‘war against pollution’: an opportunity for longer lives,” 2019. Accessed: Jun. 08, 2023. [Online]. Available: https://aqli.epic.uchicago.edu/wp-content/uploads/2019/01/India.NCAP_.Global.English.pdf
[45] P. Lestari, M. K. Arrohman, S. Damayanti, and Z. Klimont, “Emissions and spatial distribution of air pollutants from anthropogenic sources in Jakarta,” *Atmos Pollut Res*, vol. 13, no. 9, p. 101521, Sep. 2022, doi: 10.1016/j.apr.2022.101521.
[46] E. Sáez de Cámara, I. Fernández, and N. Castillo-Eguskitza, “A holistic approach to integrate and evaluate sustainable development in higher education. the case study of the university of the basque country,” *Sustainability*, vol. 13, no. 1, p. 392, Jan. 2021, doi: 10.3390/su13010392.
[47] R. Kurniawan, A. C. Kusuma, B. Sumargo, P. U. Gio, S. K. Wongsonadi, and K. Sasmita, “The role of renewable energy and foreign direct investment toward environmental degradation convergence to achieve sustainability: evidence from ASEAN countries,” *International Journal of Energy Sector Management*, May 2024, doi: 10.1108/IJESM-02-2024-0012.