Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

Research Article

Smart Air Pollution Predictor for Smart Cities

Fokrul Alom Mazarbhuiya[1]*, Vijo Arul Selvi M[1], Mohamed A. Shenify[2]

[1] Department of Mathematics, School of Fundamental and Applied Science, Assam Don Bosco University, Assam, India
[2] College of Computer Science and IT, Albaha University, Albaha, 65799, Saudi Arabia
E-mail: fokrul.mazarbhuiya@dbuniversity.ac.in

Received: 21 November 2024; **Revised:** 17 February 2025; **Accepted:** 17 March 2025

Abstract: Air pollution remains a critical threat to human health, leading to respiratory issues and fatalities worldwide. With industrialization and urbanization on the rise, monitoring and predicting Air Quality Index (AQI) levels, especially in cities, becomes increasingly challenging. Cluster analysis emerges as a crucial tool for discerning patterns in air pollutant data. This study introduces the Common Mahalanobis Distance-based Fuzzy C-Means algorithm (CM-FCM) for air pollution data analysis, evaluating its effectiveness against *k*-means Clustering Algorithm and Euclidean Fuzzy C-Means Clustering Algorithm in terms of accuracy. The CM-FCM algorithm identified non-spherical clusters that accurately captured pollution patterns, enabling precise hotspot detection. Applied to datasets from Byrnihat and Indira Gandhi International Airport (LGBI) Airport, India, it categorized LGBI Airport as "Moderately Polluted" and Byrnihat as "Most Polluted", with PM_{10} levels exceeding World Health Organization (WHO) standards on 99% of days. By considering pollutant correlations, CM-FCM provides valuable tools for policymakers to address pollution hotspots and enhance public health strategies.

Keywords: air pollution, air quality index, *k*-means clustering, FCM clustering, mahalanobis distance, CM-FCM clustering algorithm

MSC: 68T10, 62H30

1. Introduction

1.1 *Overview of cluster analysis in air pollution studies*

Air is essential to human survival. For the sake of our well-being, its quality needs to be observed and comprehended. Millions of individuals worldwide have respiratory deaths and physiological problems due to air pollution. Air pollution is the leading environmental risk, according to scientific data. The burning of biomass, industrial processes, and automobile emissions are some of the human activities that have directly contributed to the reduction in global air quality over the past 50 years. Because of the detrimental impacts on human health, the presence of aerosols or Particulate Matter (PM) dispersed in the air has long been a matter of concern. It has been observed that one of the main causes of illness and mortality is air pollution. As a result, a lot of study has been done on air pollution and air quality monitoring, source identification, long-range pollutant transport pathways, and the creation and application of efficient control and mitigation techniques [1].

DOI: https://doi.org/10.37256/cm.6420256109

Analysis of air pollution is essential for a number of reasons. First of all, it aids in determining the sources and levels of air contaminants. Strategies to lower pollution levels and safeguard public health can then be developed using this information. Secondly, air pollution analysis can be used to track the success of pollution control efforts and pinpoint regions in need of further intervention. In addition to harming the environment, air pollution also damages human systems and bodily organs.

Air pollution levels are measured and communicated with the help of Air Quality Index (AQI), a numerical index. It is one assessment indicator that directly affects the public health. Sulphur dioxide (SO_2), Carbon monoxide (CO), Nitrogen dioxide (NO_2), and Ozone (O_3), PM with a diameter of 10 microns or less (PM_{10}), PM with a diameter of 2.5 microns or less ($PM_{2.5}$), Ammonia (NH_3), and lead (pb) are the eight major air pollutants used to calculate the AQI. There are other applications for computing the AQI with the help of six pollutants PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, and O_3. The precise selection of contaminants, however, depends on the specific goal and a number of factors, such as measuring methods, data accessibility, and monitoring frequency. Severe pollution in the air is shown by a high AQI level, which can be extremely harmful to health. With the rise in industrial and automobile advancements, monitoring and forecasting AQI, particularly in metropolitan areas, has become an essential and difficult undertaking [2].

An efficient method for analysing air contaminants is cluster analysis. The process of combining similar observations, data points, or feature vectors according to their shared features is known as cluster analysis, or more popularly, "clustering". It is "the art of finding groups in data" [3]. Finding groupings of related objects is often the goal of cluster analysis, where items in one cluster are more alike than those in other clusters. As a result, cluster analysis is a helpful method for finding and obtaining data that would not have been apparent before.

Although cluster analysis was first proposed in 1930, it wasn't until the 1960s that its use became widely accepted. In the 1970s, clustering techniques were used in a wide range of fields, including biology, social science, medicine, and geography. In the 1980s, they were also used in atmospheric science. Application of clustering, in particular *k*-means and hierarchical agglomerative approaches, to air pollution data has been undertaken since the 1980's, and has subsequently attracted great interest [1].

In this context, the utilization of advanced data mining techniques, has emerged as a crucial tool for processing and analyzing the vast amounts of heterogeneous data generated by air quality sensors. One crucial pre-processing step that is required to improve the quality of the underlying clustering is the feature selection phase. Since some features could be noisier than others, not all of them are equally important for locating the clusters. As a result, it is frequently beneficial to employ a pre-processing stage where the irrelevant and noisy information are eliminated from consideration. Dimensionality reduction and feature selection are strongly associated [3].

Distance-based approaches are widely used in almost every data type, provided the suitable distance function is defined for that data type. Consequently, the clustering problem in any data type may be simplified down to the distance function problem for that kind of data [3]. Data analysis techniques, aid in organizing large datasets, enabling the identification of patterns and clusters within air pollution data. These clusters provide valuable insights into pollutant distribution and similarities, facilitating informed decision-making for pollution control strategies [4].

The FCM clustering, a soft clustering algorithm, provides a flexible approach for partitioning data into clusters by assigning each data point membership grade for each cluster. FCM allows for overlapping memberships, reflecting the uncertainty inherent in many real-world datasets. This soft assignment of data points to clusters enables FCM to capture complex patterns and relationships in the data, make it appropriate for applications such as air pollution analysis where the boundaries between different pollution sources may not be well-defined.

The FCM algorithm extends the classic *k*-means algorithm by introducing a fuzzifier parameter (m) that controls the degree of fuzziness in the cluster memberships. By adjusting the value of m, analysts can regulate the extent to which data points are acceptable to the multiple clusters simultaneously. This flexibility enables FCM to adapt to various types of data distributions and cluster shapes, making it a versatile tool for exploratory data analysis in air quality research.

Unlike the Euclidean distance metric, which assumes equal importance for all features, the Mahalanobis distance accounts for correlations and scaling differences among features, leading to more accurate and robust clustering results. Therefore, the Common Mahalanobis distance-based FCM algorithm [5] holds promise for improving the identification of pollution sources and understanding the complex interactions among various pollutants in the atmosphere.

1.2 *Literature review*

Clustering has been extensively used in atmospheric science data over the past 50 years, particularly in climate and meteorological data. Clustering algorithms have been used in air pollution research since the 1980s [1]. Based on various clustering techniques, various researches have been conducted to analyze air pollution.

Authors in [6], investigated air pollution clustering in a smart city using the *k*-means algorithm used in data from the City Pulse Project. Utilizing a dataset with five pollution elements, the study employed the *k*-means algorithm with various values of *k* to cluster data based on pollution level similarities. It concluded that *k*-means clustering effectively explained healthy and unhealthy zones in a smart city based on air pollution data.

The conference paper [7] proposed an enhanced *k*-means algorithm for analyzing air pollution data from diverse sources like Los Angeles, London, and Mexico. The method included calculating the AQI based on pollutant and environmental variable correlations. The data collection and preprocessing involved real-time pollutant measurements, and the enhanced *k*-means algorithm iteratively assigned data points to clusters, demonstrating superiority in accuracy and run time compared to the Possibilistic Fuzzy *C*-Means algorithm across various datasets.

The research paper [8] explored into the multifaceted realm of air pollution, emphasizing its detrimental effects on health and the environment, and identifying key pollutants. Employing clustering algorithms, specifically *k*-means and Expectation Maximization (EM), the study utilized datasets from the City Pulse Project for real-time smart city applications. Experimental results demonstrated that the EM algorithm surpasses *k*-means in accuracy, with lower variation and faster processing speed. This research highlighted the efficacy of advanced clustering techniques in analyzing air pollution datasets, especially the smart city initiatives context.

Authors in [9], proposed a machine learning approach using *k*-means clustering, dynamic time warping, support vector regression, ARIMA, and LSTM on a 10-year air quality dataset. The dataset includes statistical data for NO_2, SO_2 and O_3 pollutants, as well as temporal and geospatial information. Supervised (decision tree, SVM, SVR), unsupervised (*k*-means), semi-supervised (generative models, self-training, transductive SVM), reinforcement, and ensemble learning algorithms were employed. The authors determined that clustering with *k*-means and dynamic time warping is the most effective for forecasting pollution levels.

In [10] the Unsupervised *k*-means algorithm, an approach that removes the necessity for initialization, parameter selection, or prior knowledge of cluster numbers, was introduced in 2020. By iteratively eliminating excess clusters and automatically determining the ideal number of clusters based on the intrinsic data structure, the aforesaid method effectively considers the number of points as the initial number of clusters. This solves the initialization problem.

In [11], the authors introduced a novel anomaly detection method using clustering techniques, specifically tailored for information security and intrusion detection systems. The proposed algorithm enhanced the traditional *k*-means approach by incorporating both partitioning and hierarchical strategies. A unified metric, accommodating numeric, categorical, and mixed attributes, was introduced for distance and similarity calculations, with attribute weights determined by entropy. The algorithm featured a merge function that combines clusters with high similarity based on a predefined threshold.

In [12], the challenge of predicting AQI by proposing a hybrid machine learning approach was addressed. The method integrated the *k*-means algorithm, providing cluster classifications as input to the Support Vector Machines (SVM). Utilizing three years of air quality data from Gurugram (Haryana), India, and preprocessing through scaling, the hybrid algorithm was equated with the traditional SVM approach. The experimental study demonstrated the superior prediction performance of the hybrid algorithm, achieving an accuracy of 91.25, in contrast to the SVM's accuracy of 65.93. This research highlighted the efficacy of hybrid machine learning techniques in enhancing AQI prediction accuracy.

The study [13] explored multiple models for predicting air pollutant levels and found that clustering months with similar behavior using *k*-means clustering was most effective. Acknowledging seasonal pollutant patterns, the research applied *k*-means clustering to group months with analogous behavior. Overcoming Euclidean distance inadequacy, Dynamic Time Warping (DTW) was adopted, supplemented by LB-Keogh for efficiency, resulting in *k* clusters, each represented a set of identical behavior patterns, fitted on an individual regression line.

For the air pollution data analysis, author in [14] suggested using a hybrid clustering approach that combines agglomerative and *k*-means clustering. Prior to the determining the initial cluster-numbers, the dataset was hierarchically clustered using the proposed algorithm. Then *k*-means method was then employed. The clusters formed by the hybrid

k-means agglomerative algorithm is well defined than the k-means and agglomerative algorithm. The endeavor aimed to contribute to the air pollution analysis, facilitating the identification location with the healthiest environment in any city and enabling efficient pollution monitoring.

The paper [15] addressed the limitations of fuzzy clustering methods that relied on Euclidean distance, for detect clusters in spherical shapes only. It was extended in Gustafson-Kessel (GK) [16, 17] and Gath-Geva (GG) [16, 17] algorithms for non-spherical clusters but with constraints.

In [18], the authors conducted a study on air pollution in Salamanca, Mexico, utilizing the Possibilistic Fuzzy c-Means (PFCM) algorithm to analyze data generated by three monitoring stations. They focused on Sulfur dioxide (SO_2) and Particulate Matter (PM_{10}) concentrations, aiming to develop a local environmental contingency alarm system. This study illustrates the application of advanced clustering algorithms in environmental monitoring to enhance decision-making and public health protection.

The paper [19] introduced the FCM-Normalized Mahalanobis (FCM-NM) clustering algorithm, which utilized Mahalanobis distance to detect non-spherical structural clusters, overcoming limitations of the traditional algorithms like GK and GG. FCM-NM demonstrated superior clustering accuracy over traditional FCM algorithms, particularly in educational settings for students with nursing backgrounds learning mathematics. Key contributions include introducing a regulating factor for covariance matrices and replacing common covariance matrices with correlation matrices for more stable results.

The proposed modification of the FCM in [20] aims to address its limitations, such as inflexible cluster number adaptation and suboptimal objective function for unequal size clusters. The M & MFCM approach replaces the Mahalanobis and Minkowski metrics instead of Euclidean distance to improve cluster detection for high-dimensional datasets, while mitigating numerical issues. Experimental results show significant clustering accuracy improvements.

The authors in [17] proposed a novel approach, Nano Topology based fuzzy clustering, Nano topology and Common Mahalanobis Distance based Fuzzy C-Means (NT-CM-FCM) for anomaly detection in IoT systems. By integrating nano topology and fuzzy clustering techniques, it outperforms traditional methods like FCM, GK, and GG. Experimental results on datasets demonstrate improved detection rates, accuracy, and computational efficiency.

1.3 *Smart city concept*

The increase in urbanization leads to a number of issues pertaining to several facets of life, including air quality, health care, and transportation. In order to address these issues, the smart city concept was developed, which can promote sustainable growth and an improvement in the quality of life by combining Information and Communication Technology (ICT) with residents and available resources. Other definitions of smart cities include "the use of smart computing technologies to make the critical infrastructure components and services of a city which include city administration, education, healthcare, public safety, real estate, transportation, and utilities-more intelligent, interconnected, and efficient" and "a smart city is a city in which there are six main components including smart economy, smart transportation, smart environment, smart citizens, smart life, and smart management". We may also gather a great deal of information about the present state of affairs and see the true picture throughout the city by utilizing the services designed around the idea of a smart city. One important aspect of smart cities is the accessibility of sensor-provided data [21, 22].

1.4 *Green city philosophy and urban nature*

All aspects of nature, including biocoenosis, living things, and their habitats, are very important parts of green infrastructure in a city known as the "Green City". For the benefit of city dwellers, these types of nature are conserved, upheld, and expanded in a green metropolis. Urban nature is a crucial idea for city development and an excellent source of services. Moreover, it serves as a metaphor for reintroducing nature into the city and improving all forms of urban nature in order to create a new alliance between the natural world and the constructed environment. In this context, urban nature is viewed as both an essential component of our urban lives and a significant contributor to quality of life [23].

1.5 *Understanding AQI ranges*

A useful instrument for informing people about the state of the air quality in easily understood terms is the Air Quality Index. It reduces complicated air quality data about different pollutants to a single number (index value), color, and nomenclature. Good, Satisfactory, Moderately Polluted, Poor, Very Poor, and Severe are the six AQI classifications. The values of ambient concentrations of air pollutants and their probable health effects (sometimes referred to as health breakpoints) are used to determine each of these categories. For eight pollutants for which short-term (up to 24 hours) National Ambient Air Quality Standards are recommended, AQ sub-indices and health breakpoints have been developed [24]. The following are the eight contaminants' AQI classifications and health breakpoints as outlined in Table 1.

Table 1. AQI Categories and Health Breakpoints (* CO in mg/m^3 and other pollutants in µg/m^3; 24-hourly average values for PM_1, $PM_{2.5}$, NO_2, SO_2, NH_3, and Pb, and 8-hourly values for CO and O_3)

AQI Category	AQI	PM_{10}	$PM_{2.5}$	NO_2	O_3	CO	SO_2	NH_3	Pb
Good	0-50	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory	51-100	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.5-1.0
Moderately polluted	101-200	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800	1.1-2.0
Poor	201-300	251-350	91-120	181-280	169-208	10-17	381-800	801-1,200	2.1-3.0
Very poor	301-400	351-430	121-250	281-400	209-748*	17-34	801-1,600	1,200-1,800	3.1-3.5
Severe	401-500	430 +	250 +	400 +	748 +*	34 +	1,600 +	1,800 +	3.5 +

1.6 *Calculation of AQI*

A monitoring location's 24-hour average concentration value (eight hours for CO and O_3) and health breakpoint concentration range are used to compute the sub-indices for particular pollutants. The location's AQI is the poorest sub-index. It's possible that not all eight pollutants are tracked at every site. Only when data for at least three pollutants are available-at least one of which must be $PM_{2.5}$ or PM_{10}-is the overall AQI determined. If not, the data are deemed inadequate for AQI calculation. Similarly, it is thought that in order to calculate the sub-index, at least 16 hours of data are required [25].

The equations for parameters are used independently to calculate the AQI. To calculate the AQI based on four factors, for instance, we apply the equation four times, and the AQI is communicated by the worst sub-index. The sub-index is a linear function of the pollutant concentration.

$$Ip = \left\{ \frac{IHi - ILo}{BPHi - BPLo} \right\} (Cp - BPLo) + ILo \tag{1}$$

where,

Ip-index of pollutant p;

Cp-truncated concentration of pollutant p;

$BPHi$-concentration breakpoint i.e. greater than or equal to Cp;

$BPLo$-concentration breakpoint i.e. less than or equal to Cp;

IHi-AQI value corresponding to $BPHi$;

ILo-AQI value corresponding to $BPLo$.

For instance, to determine the AQI based on $PM_{2.5}$, CO, and O_3, we compute the sub-index for each of these parameters independently. Referring to the AQI range according to Indian norms, $BPHi = 120$, $BPLo = 91$, $IHi = 300$, and $ILo = 201$, the present $PM_{2.5}$ concentrations is 110 µg/m^3. Putting the values in equation and solving:

$$\text{Sub Index} = \left\{\frac{300-201}{120-91}\right\}(110-91)+201 = 265.86 \tag{2}$$

Similarly, the sub-index can also be computed for various other parameters, and the AQI is displayed by the worst sub-index [23].

1.7 *Air quality challenges in smart cities*

The United Nations estimates that 55% of people on Earth currently reside in cities, and by 2050, that number is predicted to rise to 68%. According to projections, by 2050, there may be an additional 2.5 billion people living in urban areas due to urbanization-the slow movement of people from rural to urban areas-and global population growth, with nearly 90% of this increase occurring in Asia and Africa [26].

In many places, air quality has become a major issue. The World Health Organization (WHO) reports that over seven million people die each year as a result of this problem, and that over 80% of people living in urban areas reside in locations where air quality exceeds WHO guidelines. Air-pollution has decreased life expectancy both nationally and globally. According to the study, in certain polluted Asian and African nations, particulate matter with a diameter of 2.5 micro meter ($PM_{2.5}$) decreased life expectancy by 1.2 to 1.9 years in 2016.

Therefore, preventing or lessening the effects of air pollution is a critical issue. Knowing about the quality of the air will encourage us to take precautions; it may encourage people to conduct their everyday activities in less polluted areas (by avoiding high polluted areas). However, analysing the data and coming up with clever solutions is still a difficult undertaking. In order to analyse large data more effectively and efficiently, make the invisible visible, and extract knowledge from data, it is imperative to use effective methodologies and procedures [27].

Urban areas are experiencing growth in the number of industries and auto mobiles due to the increasing population's greater reliance on energy and transportation. Consequently, this surge in activity leads to an increase in pollutant emissions, causing concern for both national and local government officials, as well as world leaders. Local and national governments are committed to tackling the issues related to air pollution, with the goal of improving the well-being of citizens and addressing diseases linked to pollution. As a result, controlling air quality becomes a key goal for smart cities in their effort to develop more sustainable and healthy urban environments.

1.8 *Objective of the paper*

This study propose to develop a smart air pollution predictor for smart cities using the Common Mahalanobis distance based Fuzzy C-Means (FCM-CM) algorithm. Specifically, this study has the following objectives:

To develop a robust air pollution predictor capable of accurately forecasting pollution levels in urban environments.

To utilize the FCM-CM algorithm to effectively cluster air pollutants based on their spatial and temporal distribution.

To enhance understanding of pollution sources and their interactions by analyzing clustering patterns and correlations among pollutants.

To provide actionable insights for policymakers and urban planners to implement targeted pollution control measures and improve air quality management strategies in smart cities.

The paper is organized as follows. Section 2, discusses the algorithms and Flowcharts of the proposed methods for the prediction of air pollution in smart cities. Section 3, explains the experimental setup, datasets, and the implementations of algorithms. Section 4, explains the analysis of results and discussions. Finally, we conclude our work with conclusions, limitations and lines for future work in Section 5.

2. Smart air pollution predictor for smart cities: common mahalanobis fuzzy C-means clustering

2.1 *Fuzzy C-means clustering*

The data points which belong to the multiple cluster bounderies, may not be inside of any of the multiple clusters. Fuzzy clustering combines each data element with a set of membership levels to assign each data element to numerous clusters, hence reducing classification uncertainty. The cluster information can be used to create a fuzzy inference system that uses the fewest possible rules to model data behaviours.

A robust fuzzy clustering approach that works well for clustering overlapped datasets is Fuzzy C-Means (FCM) clustering. In FCM, a membership function specifies the degree of a data point. The data element near the cluster center is given a high value by the membership function, whereas the data element far from the cluster center is given a low value. FCM divides the given n vector x_i, $i = 1, 2, \ldots, n$ into fuzzy sets and computes the center in each fuzzy set by the applying minimization on the objective function [28].

Firstly, randomly selected n data points $(x_1, x_2, x_3, \ldots, x_n)$, with the cluster center c_i, $i = 1, 2, \ldots, C$. Evaluate the membership matrix (U) with the help of the following equation:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}}} \tag{3}$$

where $d_{ij} = \| c_i - x_j \|$ is the Euclidean distance of the ith cluster center from the jth data point; μ_{ij} is the membership matrix's coefficients; m is the fuzziness index; C is the number of clusters.

Secondly, the following formula can be used to determine the objective function:

$$J(U, c_1, \ldots, c_C) = \sum_{i=1}^{C} J_i = \sum_{i=1}^{C} \sum_{j=1}^{n} \mu_{ij}^m d_{ij}^2 \tag{4}$$

Finally, a new C fuzzy cluster center C_i, $i = 1, 2, \ldots, C$ can be estimated as follows:

$$C_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m x_j}{\sum_{j=1}^{n} \mu_{ij}^m} \tag{5}$$

2.2 *Mahalanobis fuzzy C-means clustering*

Despite its widespread use in clustering-based algorithms, Euclidean distance has drawbacks. The shortest distance between two points is measured by the Euclidean distance. Such variables that effectively measure the same characteristic are given equal weight by Euclidean distance since it ignores the correlation between the attribute values. Consequently, single feature becomes dominant. Accordingly, the correlated variables get surplus weight by the distance measure and the accuracy is compromised.

When the sample distribution follows the Gauss distribution, the error rate of Euclidean distance in calculating high attribute correlation of data sets can be prevented by using Mahalanobis distance. It can be used to readjust the geometric distribution of patterns in order to decrease the distance of similar patterns. Thus, by substituting the Mahalanobis distance for the Euclidean distance in conventional FCM clustering, a Mahalanobis distance-based Fuzzy Clustering Algorithm (MFCM) is suggested [29]. When working with data sets that have a high attribute correlation, the MFCM algorithm's accuracy clearly improves and it can successfully address the FCM's inability to produce aspheric clusters. Also, if the

covariance matrix is inverted in order to calculate the Mahalanobis distance, there will be a singularity problem and eigenvalue, eigenvector, and pseudo-inverse procedures are used to address this situation.

Since Mahalanobis distance takes into consideration the correlation between the variables and the data of air pollutants are highly correlated, it will be convenient to use Mahalanobis distance instead of Euclidean distance. Also, Mahalanobis distance is scale-invariant that gives distance between a given p-variant probability distribution $P_X(\cdot)$ generated point $x \in \mathbb{R}^n$ and the distribution's mean $\mu = E(X)$. Suppose $P_X(\cdot)$ has finite second-order moments and $\Sigma = E[(X-\mu)]$, the covariance matrix, then the Mahalanobis distance is given by

$$D(x,\ \mu,\ \Sigma) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \tag{6}$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to Euclidean distance.

2.3 *CM-FCM*

Here, a common covariance matrix (Σ) is substituted for all the covariance matrices (Σ_i) of the objective function. The objective function of Common Mahalanobis Fuzzy C-Means Clustering (CM-FCM) is given as follows:

$$J(X;\ U,\ V,\ \Sigma) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D_{ik\Sigma}^2 \tag{7}$$

Subject to the constraints:

$$m \in [1,\ \infty) \tag{8}$$

$$U = [\mu_{ik}]_{c \times N},\ \mu_{ik} \in [0,\ 1],\ 1 \le i \le c,\ 1 \le k \le n \tag{9}$$

$$\sum_{i=1}^{c} \mu_{ik} = 1,\ 1 \le k \le n \text{ and } 0 < \sum_{k=1}^{n} \mu_{ik} < N,\ 1 \le i \le c \tag{10}$$

The Mahalanobis distance $D_{ik\Sigma}^2$ is defined as:

$$D_{ik\Sigma}^2 = (x_k - v_i)^T \Sigma^{-1} (x_k - v_i) - ln|\Sigma^{-1}|,\ \text{ if } (x_k - v_i)^T \Sigma^{-1} (x_k - v_i) - ln|\Sigma^{-1}| \ge 0 \tag{11}$$

$$D_{ik\Sigma}^2 = 0,\ \text{if } (x_k - v_i)^T \Sigma^{-1} (x_k - v_i) - ln|\Sigma^{-1}| < 0 \tag{12}$$

Here, $J(X;\ U,\ V,\ \Sigma)$ is the objective function, X is the given dataset, U is the membership matrix where μ_{ik} represents the membership degree of the k-th data point to the i-th cluster, V is the matrix of cluster prototypes, with each column representing the center of a cluster, Σ is the common covariance matrix, m is the fuzzification parameter, which determines the level of cluster fuzziness, and $D_{ik\Sigma}$ is the Mahalanobis distance between the k-th data point and the i-th cluster center using the common covariance matrix Σ.

Only spherical structure clusters may be detected using the well-known FCM, which is based on the Euclidean distance function. Therefore, we use improved Fuzzy C-Means algorithm based on Common Mahalanobis distance by

adding constraint on fuzzy covariance matrix, called CM-FCM. It should be noted that CM-FCM becomes FCM when the covariance matrices transform into identity matrices. FCM is hence a particular instance of the CM-FCM algorithm [17].

2.4 *Algorithm CM-FCM*

Given dataset X, choose the number of clusters c, $(2 < c < N)$, iteration stop threshold $\phi > 0$.
Initialize the membership matrix U subject to the constraints, and set the iteration counter $l = 1$.
Step 1: Evaluate or update cluster-centroid v_i, $i = 1, 2, \ldots, c$.
Step 2: Evaluate pseudo-inverse matrix of covariance Σ^{-1}.
Step 3: Evaluate $D^2_{ik\Sigma}$.
Step 4: Evaluate the value of the objective function J.
Step 5: Set $l = l + 1$ to update objective function J.
• If the value of the objective function obtained in Step 4 satisfies $\| J^l - J^{l-1} \| < \phi$, stop.
• Output the cluster set and membership matrix.
Step 6: Else, go to Step 1.

2.5 *Flow chart of CM-FCM*

The following steps describe the Common Mahalanobis Fuzzy C-Means (CM-FCM) algorithm, as illustrated in Figure 1.

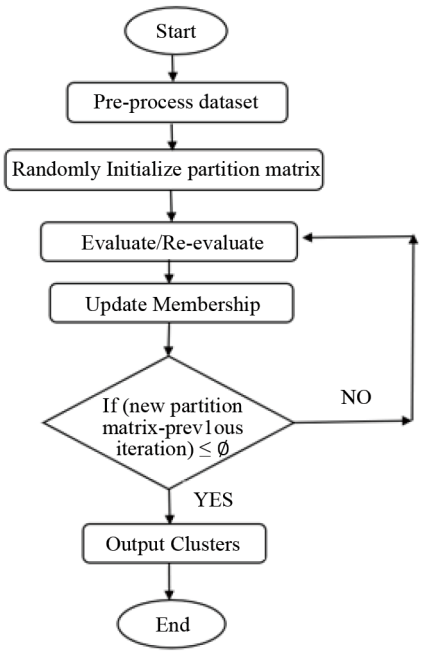

Figure 1. Flow chart of common mahalanobis fuzzy C-means

1. **Start:** This represents the initialization step.
2. **Pre-process dataset:** This is an initial step before the algorithm begins where the dataset is pre-processed.
3. **Randomly Initialize partition matrix:** This corresponds to initializing the membership matrix U and setting the iteration counter $l = 1$.

4. **Evaluate/Re-evaluate cluster:** This combines Step 1 (evaluating or updating cluster-centroid v_i) and Step 2 (evaluating the pseudo-inverse matrix of covariance Σ^{-1}).

5. **Update Membership values:** This corresponds to Step 3 (evaluating $D^2_{ik\Sigma}$) and implicitly updating the membership matrix U.

6. **If (new partition matrix-previous iteration)** $\leq \phi$**:** This step conducts the convergence check, halting the algorithm when $|J^l - J^{l-1}| < \phi$ is met.

7. **Output Clusters:** This is the result of the algorithm when the convergence condition is met.

8. **End:** The final step after the clusters are outputted.

3. Experimental setup

Google Colab was the platform used; it is a free cloud-based tool from Google that lets users write and run Python code collaboratively in a Jupyter Notebook environment. In order to make Machine Learning (ML) and data science jobs easier, Google Colab notebooks offer a virtual environment with free Graphics Processing Unit (GPU) resources. The following libraries were utilized in this experimental setup:

• **pandas:** It is a powerful data manipulation and analysis library in Python, providing data structures and functions for efficiently working with structured data, especially tabular data.

• **NumPy:** It is a fundamental package for scientific computing with Python, offering support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions for numerical computation and array operations.

• **matplotlib:** It is a comprehensive library for creating static, animated, and interactive visualizations in Python. The pyplot module provides a MATLAB-like interface for creating various types of plots and charts to visualize data distributions, relationships, and patterns.

• **seaborn:** It is a Python visualization library based on matplotlib, offering a high-level interface for drawing attractive and informative statistical graphics. It is particularly useful for creating sophisticated statistical plots and visualizations, enhancing the aesthetics and interpretability of plots compared to raw matplotlib.

• **scikit-learn (sklearn):** It is a free Python machine learning library designed to interact seamlessly with other Python numerical and scientific libraries like NumPy and SciPy. Scikit-learn features implementations of various machine learning algorithms, including support vector machines, random forest, gradient boosting, and clustering algorithms.

3.1 *Dataset*

The dataset used for analysis originates from the Kaggle platform, specifically from the collection titled, "Time Series Air Quality Data of India 2010-2023". This dataset offers a comprehensive time series of air quality measurements spanning from 2010 to 2023, providing a rich resource for studying air pollution trends and patterns over the years. It contains air quality data for Indian cities from 2010 to 2023, encompassing information about air quality conditions in 453 cities across India.

• Data Collection:

Data has been collected from the Central Control Room for Air Quality Management, ensuring reliability and authority. Data extraction was facilitated using Selenium, a web automation tool, from the Central Pollution Control Board (CPCB) website.

• Stations_Info.csv:

A pivotal component of this dataset is the "Stations_Info.CSV" file, structured as a Comma-Separated Values (CSV) format, facilitating the extraction of pertinent details about various monitoring stations across India. This file contains the following headers:

- "file_name": Name of the file associated with each monitoring station.
- "state": State within India where the monitoring station is situated.
- "city": Specific city where the monitoring station is located.

- “agency”: Agency responsible for managing and operating the monitoring station.

- “station_location”: Additional descriptive information about the exact geographical location of the monitoring station.

- “start_month”: Month when data collection commenced at the respective station.

- “start_month_num”: Numeric representation corresponding to the starting month for data collection.

- “start_year”: Year when data collection initially began at the monitoring station.

• Focus Station:

For the experimental set-up, the focus is on the AS006 station located in LGBI Airport, Guwahati and on the AS007 station located in Byrnihat, Assam.

• Parameters:

The dataset includes various parameters such as PM_2, $PM_{2.5}$, CO, NO, NO_2, SO_2, temperature, humidity, and wind speed, among others, measured in different units.

• Focus Parameters:

In this study, PM_{10} and NO_2 were considered as in urban areas with significant traffic and industrial activities, PM_{10} and NO_2 are often correlated. Both pollutants are typically emitted from similar sources, such as vehicle exhaust, industrial emissions, and combustion processes. High traffic volumes can lead to increased concentrations of both PM_{10} and NO_2.

• Use Case:

This dataset is invaluable for understanding and analyzing air quality trends and patterns in Indian cities over 13 years. It aids researchers, policymakers, and the public in gaining insights into air pollution levels, identifying high-pollution areas, assessing air quality control measures’ impact, and developing strategies to improve air quality and public health.

3.2 *Input and output*

3.2.1 *Input dataset*

Particulates are the deadliest form of air pollutants due to their ability to penetrate deep into the lungs and blood streams unfiltered. PM_{10} has been found to be exceeding in 94 cities consecutively for five years (2011-2015). NO_2 is exceeding the limits in five cities with respect to ambient Air Quality India (2011-2015) and the WHO Report 2014-2018 [30], Guwahati is one among the 94 cities.

Among cities with over 75% data availability 2023, Byrnihat in Assam was the most polluted, with an average PM_{10} concentration of 301 $\mu g/m^3$. Out of the 347 monitored days, 324 days (93%) recorded PM_{10} levels above the Indian National Ambient Air Quality Standards (NAAQS), while 344 days (99%) exceeded the WHO standard. Therefore, we selected LGBI Airport, Guwahati and Byrnihat station and data from the 1 January 2023 to the 31 March 2023, for a period of 3 months. We chose two main air pollutants namely PM_{10} and NO_2. The air pollutants PM_{10} and NO_2 are often positively correlated in urban environments [31]. The detailed air-pollution database format used is given in Tables 2 and 3.

Table 2. Air pollution database of LGBI airport

From date	To date	$PM_{2.5}$ ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)	NO ($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	NO_x (ppb)	NH_3 ($\mu g/m^3$)	SO_2 ($\mu g/m^3$)	CO (mg/m^3)	Ozone ($\mu g/m^3$)	Benzene ($\mu g/m^3$)
01-01-2023 00:00	01-01-2023 01:00	135.75	221.00	8.85	30.15	23.25	16.00	3.45	3.48	-	2.22
01-01-2023 01:00	01-01-2023 02:00	128.50	214.75	5.03	27.30	18.60	13.00	3.33	3.04	2.85	2.30
01-01-2023 02:00	01-01-2023 03:00	108.50	180.25	1.22	17.45	10.30	7.20	2.95	2.88	5.42	1.45
01-01-2023 03:00	01-01-2023 04:00	78.50	132.00	1.15	14.75	8.77	6.20	2.65	2.77	7.33	1.17
01-01-2023 04:00	01-01-2023 05:00	70.50	123.25	1.88	18.07	11.15	7.80	2.98	2.73	3.02	0.95
01-01-2023 05:00	01-01-2023 06:00	71.25	120.75	7.67	19.40	16.55	12.00	2.88	2.76	0.70	0.97
01-01-2023 06:00	01-01-2023 07:00	71.50	143.75	36.03	32.03	46.33	32.00	3.60	2.98	1.30	1.45
01-01-2023 07:00	01-01-2023 08:00	78.50	165.75	8.05	24.15	19.35	14.00	3.35	2.79	7.17	1.45
01-01-2023 08:00	01-01-2023 09:00	75.75	162.50	4.72	19.65	14.28	10.00	3.88	2.76	23.42	1.20
01-01-2023 09:00	01-01-2023 10:00	93.00	186.00	3.10	18.10	12.12	8.50	11.35	2.81	52.92	1.12
01-01-2023 10:00	01-01-2023 11:00	135.50	180.00	2.33	22.50	13.88	9.70	22.05	3.02	92.47	1.42
01-01-2023 11:00	01-01-2023 12:00	167.75	226.00	1.43	16.17	9.75	6.90	15.72	2.64	104.95	1.50
01-01-2023 12:00	01-01-2023 13:00	89.75	119.75	0.80	10.70	6.32	4.50	14.18	2.52	108.18	0.95
01-01-2023 13:00	01-01-2023 14:00	63.25	89.75	0.93	9.93	6.05	4.20	13.28	2.48	102.00	0.78
01-01-2023 14:00	01-01-2023 15:00	61.50	85.25	0.65	8.25	4.90	3.50	9.77	2.48	99.17	0.65
01-01-2023 15:00	01-01-2023 16:00	57.50	77.75	0.78	11.25	6.62	4.70	13.60	2.58	95.42	0.70
01-01-2023 16:00	01-01-2023 17:00	75.25	103.50	0.67	17.15	9.62	6.80	12.07	2.66	76.85	0.85
01-01-2023 17:00	01-01-2023 18:00	96.75	143.00	3.33	40.17	24.10	17.00	8.12	3.29	35.67	1.48
01-01-2023 18:00	01-01-2023 19:00	146.50	237.75	22.90	63.15	52.18	37.00	5.70	3.95	1.73	3.00
01-01-2023 19:00	01-01-2023 20:00	191.75	321.00	19.18	54.95	44.80	31.00	5.20	3.62	0.47	3.95
01-01-2023 20:00	01-01-2023 21:00	179.75	281.00	12.27	49.72	36.38	25.00	4.75	3.54	0.53	3.25
01-01-2023 21:00	01-01-2023 22:00	159.50	247.50	8.12	43.52	29.75	21.00	4.27	3.48	1.20	2.70
01-01-2023 22:00	01-01-2023 23:00	143.25	208.50	6.45	35.02	23.85	17.00	3.70	3.28	0.62	2.47
01-01-2023 23:00	02-01-2023 00:00	120.50	173.25	4.03	30.35	19.43	14.00	3.48	3.14	0.50	1.88
02-01-2023 00:00	02-01-2023 01:00	100.25	148.00	16.52	27.23	27.90	20.00	3.42	3.40	-	1.88

Table 3. Air pollution database of byrnihat

From date	To date	$PM_{2.5}$ ($\mu g/m^3$)	PM_{10} ($\mu g/m^3$)	NO ($\mu g/m^3$)	NO_2 ($\mu g/m^3$)	NO_x (ppb)	NH_3 ($\mu g/m^3$)	SO_2 ($\mu g/m^3$)	CO (mg/m^3)	Ozone ($\mu g/m^3$)	Benzene ($\mu g/m^3$)
01-01-2023 00:00	01-01-2023 01:00	373.25	586.50	42.27	30.48	50.55	30.83	14.42	1.18	0.20	2.95
01-01-2023 01:00	01-01-2023 02:00	446.00	920.00	41.42	26.30	47.67	28.25	16.85	1.14	-	3.12
01-01-2023 02:00	01-01-2023 03:00	365.00	754.00	41.75	24.75	47.10	25.80	13.25	1.04	-	3.10
01-01-2023 03:00	01-01-2023 04:00	363.50	714.00	36.65	22.43	41.73	24.15	10.10	0.91	-	3.10
01-01-2023 04:00	01-01-2023 05:00	374.50	537.75	31.62	20.82	36.77	22.18	7.35	0.85	-	2.58
01-01-2023 05:00	01-01-2023 06:00	408.75	736.25	44.07	23.78	48.50	24.65	8.45	1.02	1.02	2.92
01-01-2023 06:00	01-01-2023 07:00	434.25	594.75	35.08	25.15	41.90	27.77	8.23	1.00	85.58	3.17
01-01-2023 07:00	01-01-2023 08:00	430.75	759.75	11.70	30.40	25.65	32.10	79.20	0.49	2.98	2.40
01-01-2023 08:00	01-01-2023 09:00	380.50	521.50	10.88	31.35	25.55	36.33	136.75	0.82	4.62	1.35
01-01-2023 09:00	01-01-2023 10:00	356.00	640.75	6.75	32.25	22.65	29.05	179.80	0.69	11.68	1.55
01-01-2023 10:00	01-01-2023 11:00	285.50	420.00	5.30	34.85	22.85	26.00	107.70	0.88	31.85	1.28
01-01-2023 11:00	01-01-2023 12:00	192.00	580.00	1.60	20.57	12.28	25.00	52.62	1.27	55.33	1.02
01-01-2023 12:00	01-01-2023 13:00	73.00	180.00	1.38	17.82	10.60	21.53	14.25	0.48	63.30	1.05
01-01-2023 13:00	01-01-2023 14:00	70.25	164.25	1.23	15.25	9.12	18.52	15.88	0.42	70.30	0.75
01-01-2023 14:00	01-01-2023 15:00	73.50	116.25	1.27	15.93	9.50	18.78	20.12	0.53	74.83	0.88
01-01-2023 15:00	01-01-2023 16:00	79.25	169.25	1.25	15.90	9.45	19.45	16.18	0.45	62.40	0.80
01-01-2023 16:00	01-01-2023 17:00	86.50	137.50	1.27	25.27	14.50	20.93	17.75	0.81	29.38	1.30
01-01-2023 17:00	01-01-2023 21:00	140.00	287.00	6.35	37.73	25.23	19.75	166.85	1.06	13.95	2.38
01-01-2023 18:00	01-01-2023 22:00	156.00	327.00	1.58	33.92	19.33	22.23	54.88	1.09	8.10	2.45
01-01-2023 18:00	01-01-2023 22:00	164.25	317.00	3.83	36.98	22.77	21.65	25.17	1.38	2.58	3.22
01-01-2023 19:00	01-01-2023 23:00	207.00	301.00	6.00	36.00	23.97	22.55	16.13	1.24	0.82	3.47
01-01-2023 19:00	01-01-2023 23:00	234.25	462.00	7.22	34.17	24.02	24.15	13.25	0.99	0.77	2.87
01-01-2023 20:00	02-01-2023 00:00	270.25	392.00	15.15	31.55	29.10	24.75	11.65	0.92	0.30	2.40
01-01-2023 20:00	02-01-2023 00:00	316.25	611.25	23.32	29.05	34.43	26.55	10.93	0.89	0.20	2.20
01-01-2023 21:00	02-01-2023 01:00	350.50	546.75	34.20	26.30	41.83	25.77	16.50	0.84	-	2.52

3.2.2 *Data pre-processing*

In preparation for the clustering step, it is essential to preprocess the dataset to ensure it is suitable for analysis. Initially, the date and time information was separated into distinct columns, facilitating easier handling of temporal data. Subsequently, columns devoid of substantial data or containing minimal information were removed from the dataset to streamline further analysis. Following this, the dataset was segmented to isolate data spanning a three-month period, aiding in the focus of subsequent analysis efforts. Numerical columns, excluding those pertaining to date and time, were then selected for further processing. Mean values were calculated for the selected numerical columns to address missing data points, ensuring a comprehensive dataset for analysis.

Finally, the missing values were imputed with the calculated mean values, and the dataset was refined to include only the relevant columns required for clustering analysis. These preprocessing steps collectively contributed to enhancing the quality and suitability of the dataset for subsequent analytical tasks.

3.2.3 *Clustering algorithms implementation*

For *k*-means Clustering, using the elbow method, the optimal number of clusters was determined to be 3, as shown in Figure 2 and Figure 3 for both databases, where the *y*-axis represents the Within-Cluster Sum of Squares (WCSS) values. This approach aids in identifying the point of maximum curvature in the plot, signifying the optimal number of clusters. The Within-Cluster Sum of Squares (WCSS) measures the compactness of clusters, with lower WCSS values indicating tighter clusters and better separation between clusters. Therefore, the elbow method helps in selecting the optimal number of clusters by balancing the trade-off between minimizing WCSS and avoiding overfitting. Subsequently, the dataset was clustered with random values ranging from 2 to 10 to further explore the clustering structure, validate the choice of the optimal number of clusters and ensure meaningful cluster formation.

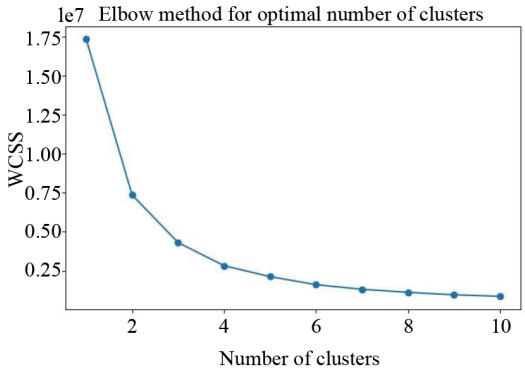

Figure 2. Optimal number of clusters for LGBI airport database

For the Euclidean Fuzzy C-Means Clustering Algorithm and for the Common Mahalanobis Fuzzy C-Means Clustering Algorithm, the fuzzy coefficient 'm' was set to 2, a commonly used value in clustering. The cluster centers were initialized randomly. The algorithm iteration stops when the change in cluster centers is less than the specified threshold, with a stopping criteria of $\varepsilon = 0.0001$. All the algorithms were implemented and simulated using Python. The dataset of LGBI Airport and Byrnihat before clustering is given in Figure 4 and Figure 5, respectively.

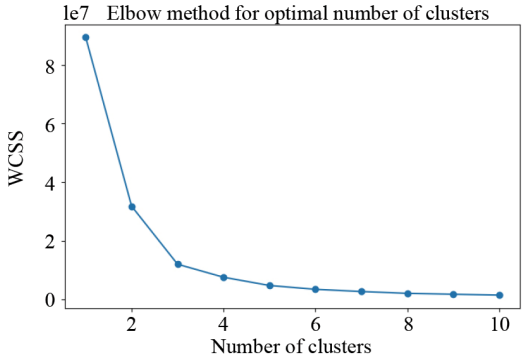

Figure 3. Optimal number of clusters for byrnihat database

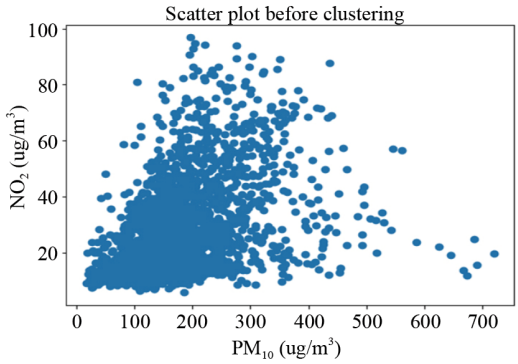

Figure 4. Dataset of LGBI airport before clustering

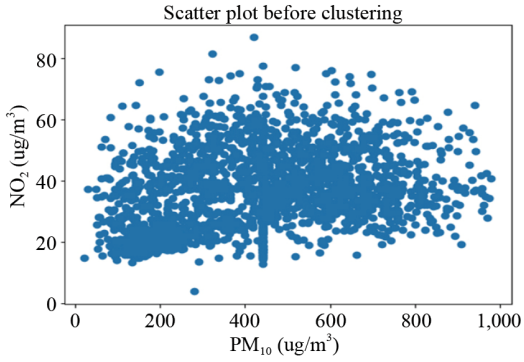

Figure 5. Dataset of byrnihat airport before clustering

3.2.4 *Generated outputs*

From Figure 6 to Figure 10, the comparison of clustering results for k-means, Euclidean FCM, and CM-FCM with cluster sizes ranging from $k = 2$ to $k = 6$ is presented for the LGBI Airport dataset. Similarly, from Figure 11 to Figure 15, the clustering comparison for k-means, Euclidean FCM, and CM-FCM with $k = 2$ to $k = 6$ is provided for the Byrnihat dataset.

The results that we obtained for the LGBI Airport dataset are given below:

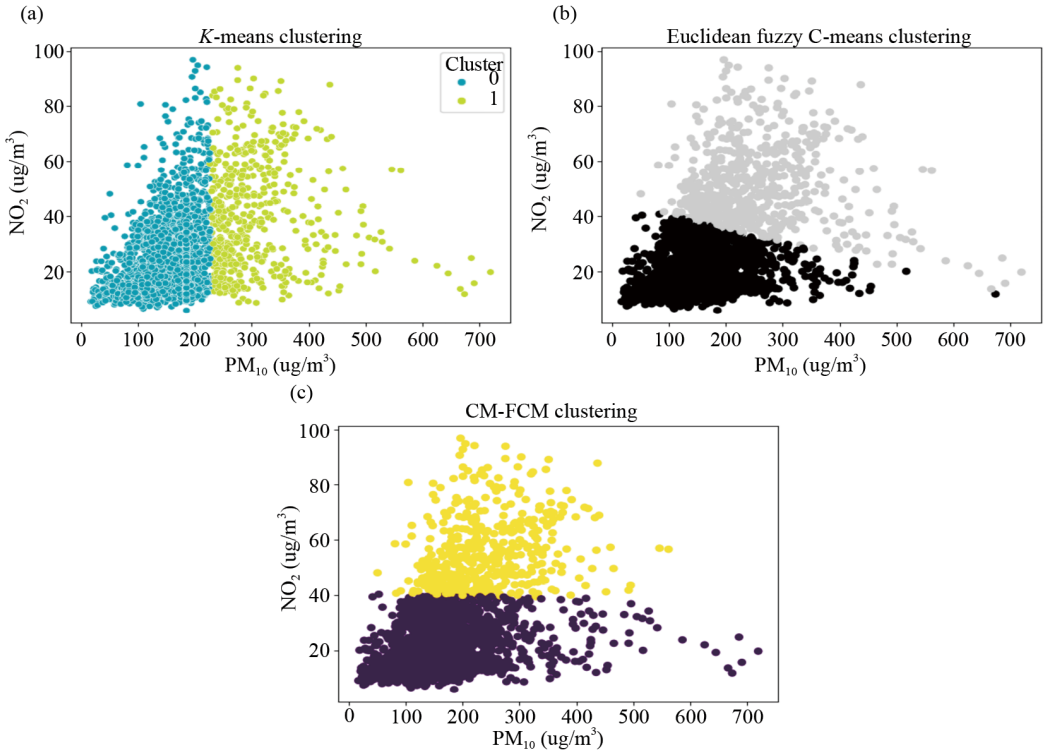

Figure 6. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 2$ for LGBI Airport dataset

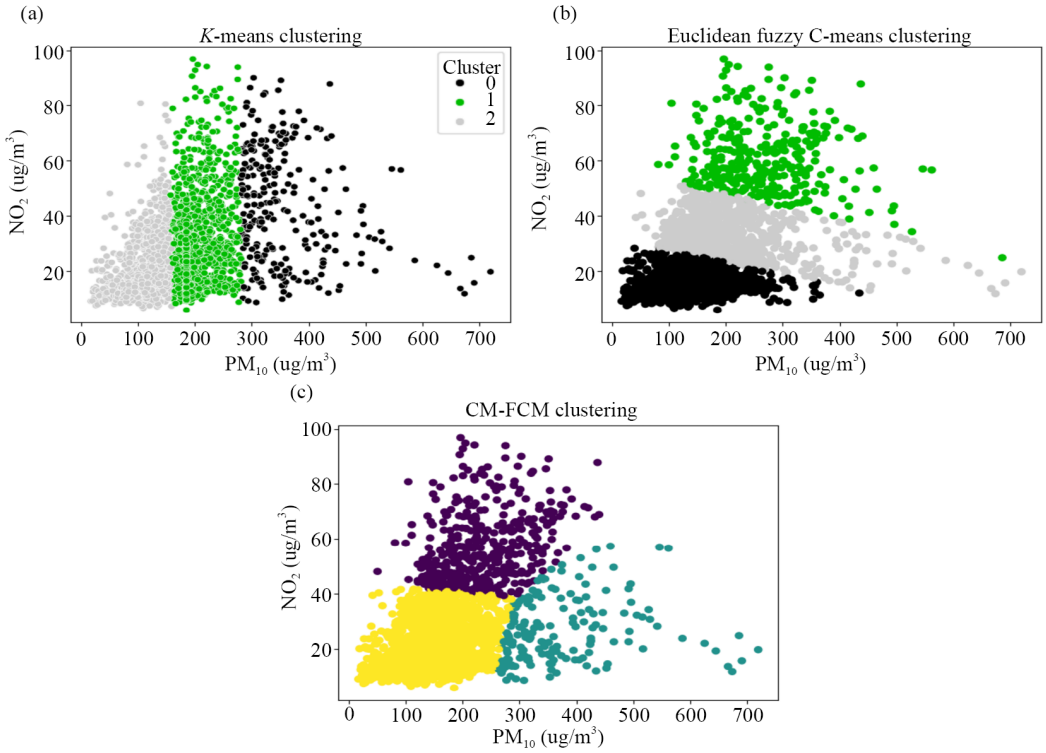

Figure 7. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 3$ for LGBI Airport dataset

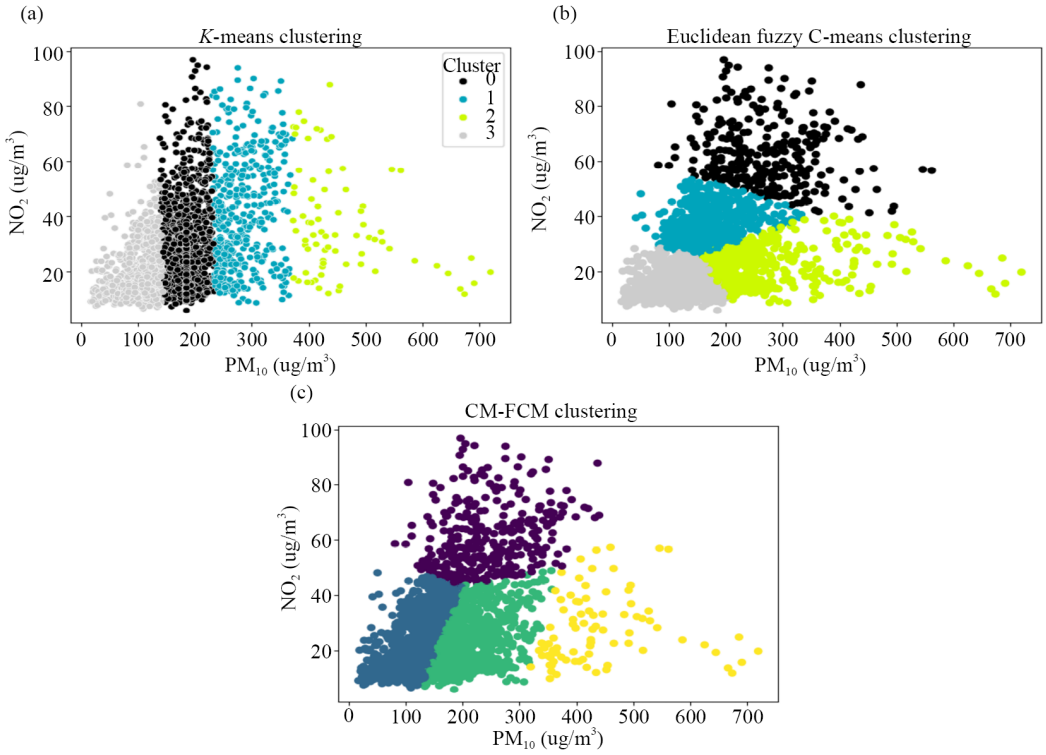

Figure 8. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 4$ for LGBI Airport dataset

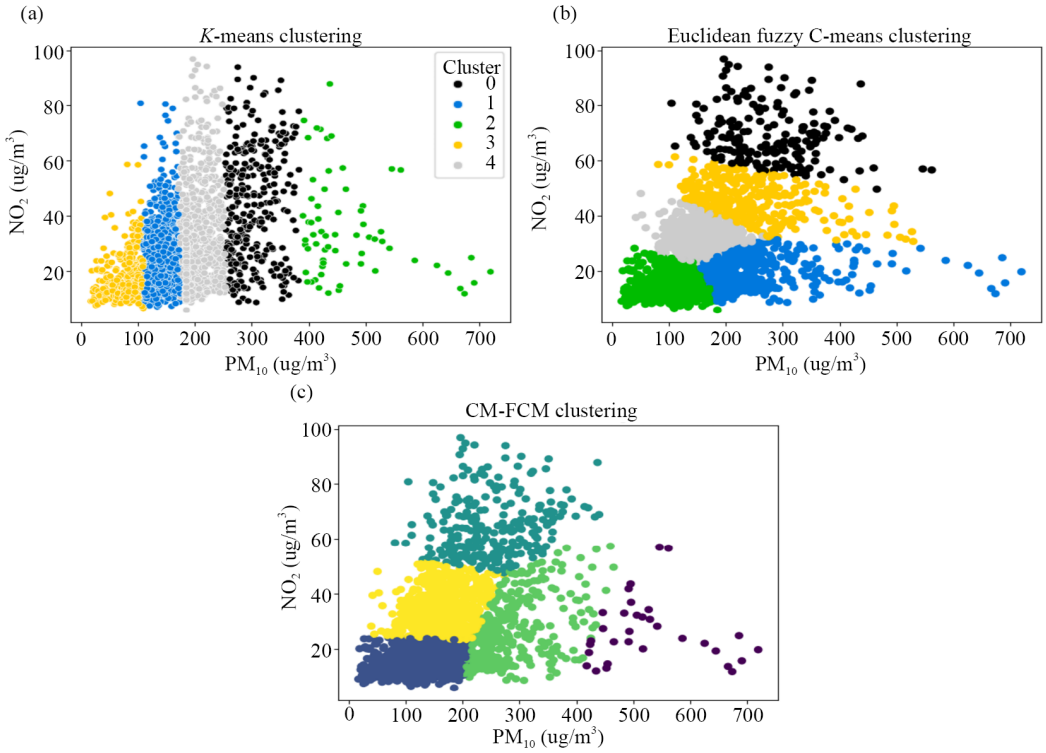

Figure 9. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 5$ for LGBI Airport dataset

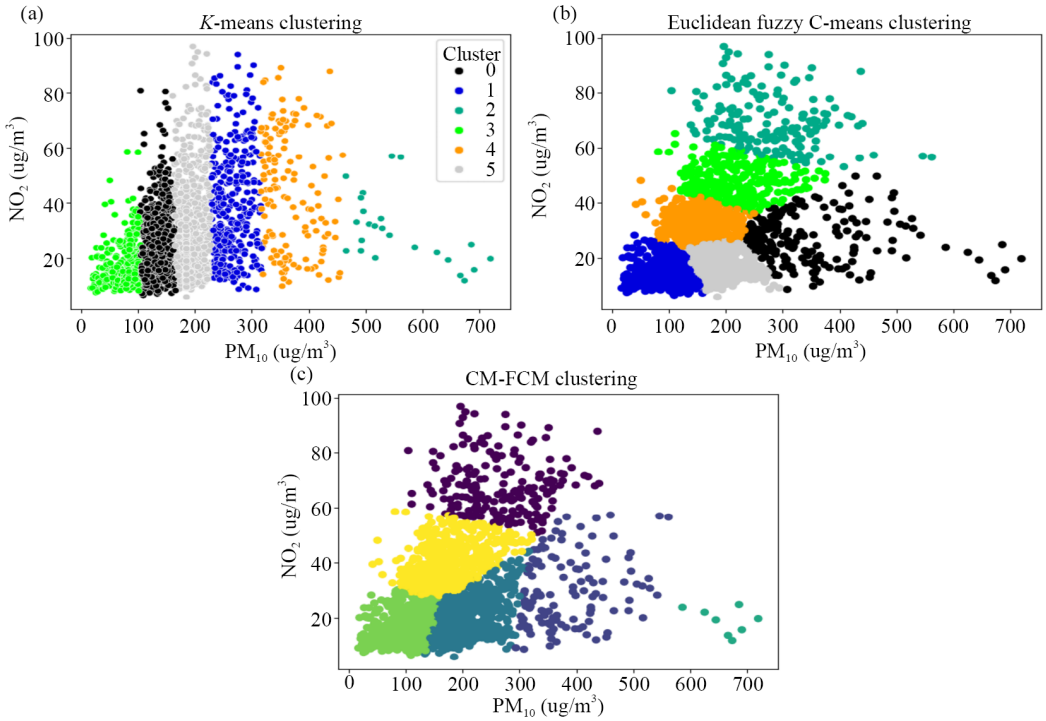

Figure 10. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 6$ for LGBI Airport dataset

The results that we obtained for Byrnihat dataset are given in Figures 11-15 below:

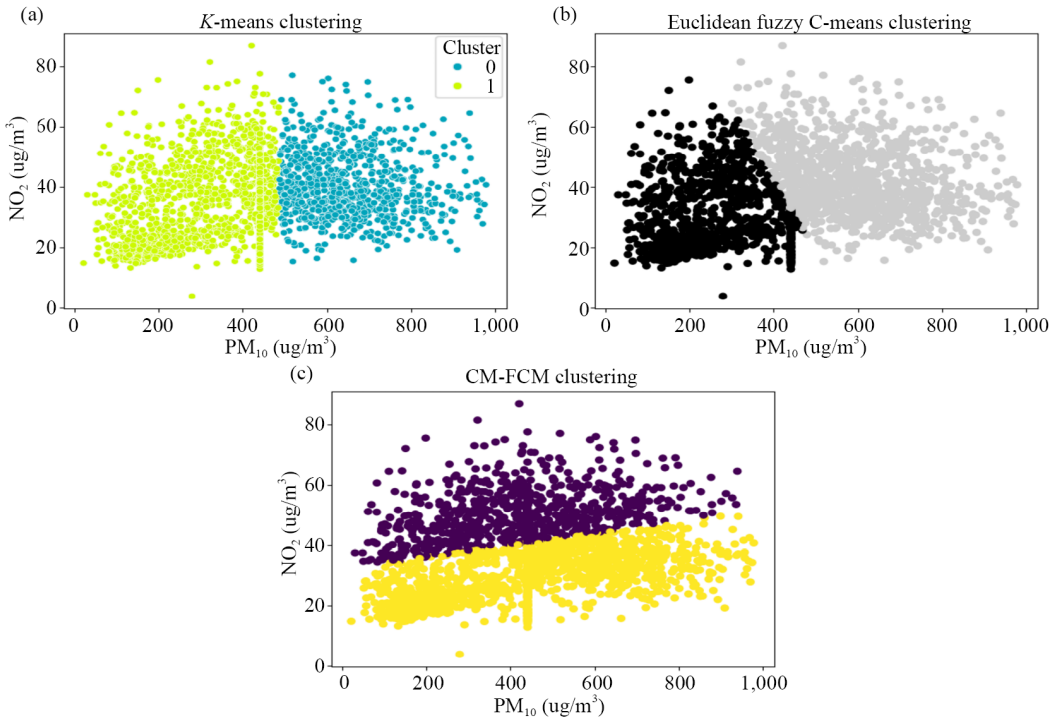

Figure 11. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 2$ for Byrnihat dataset

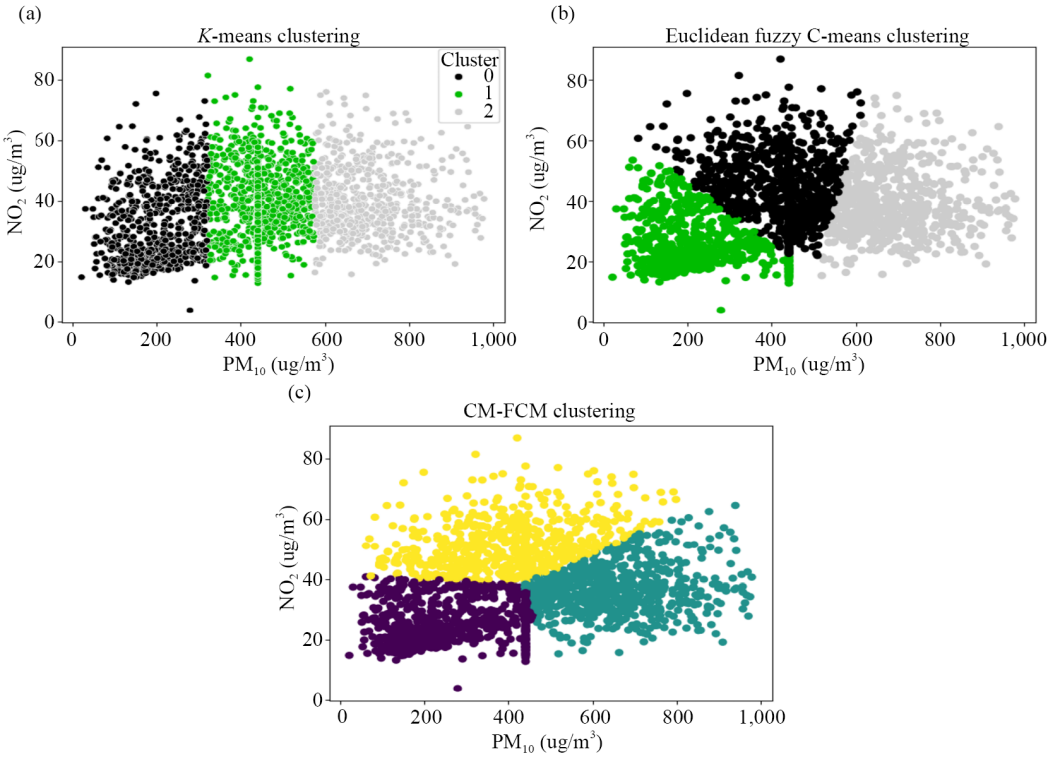

Figure 12. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 3$ for Byrnihat dataset

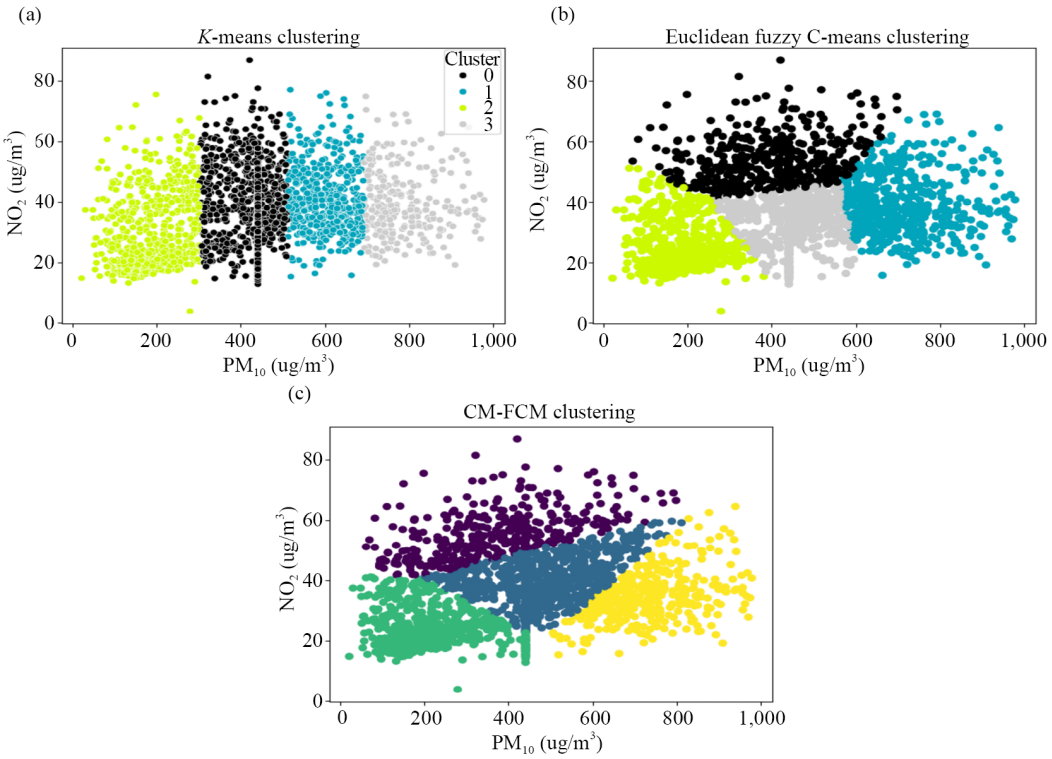

Figure 13. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 4$ for Byrnihat dataset

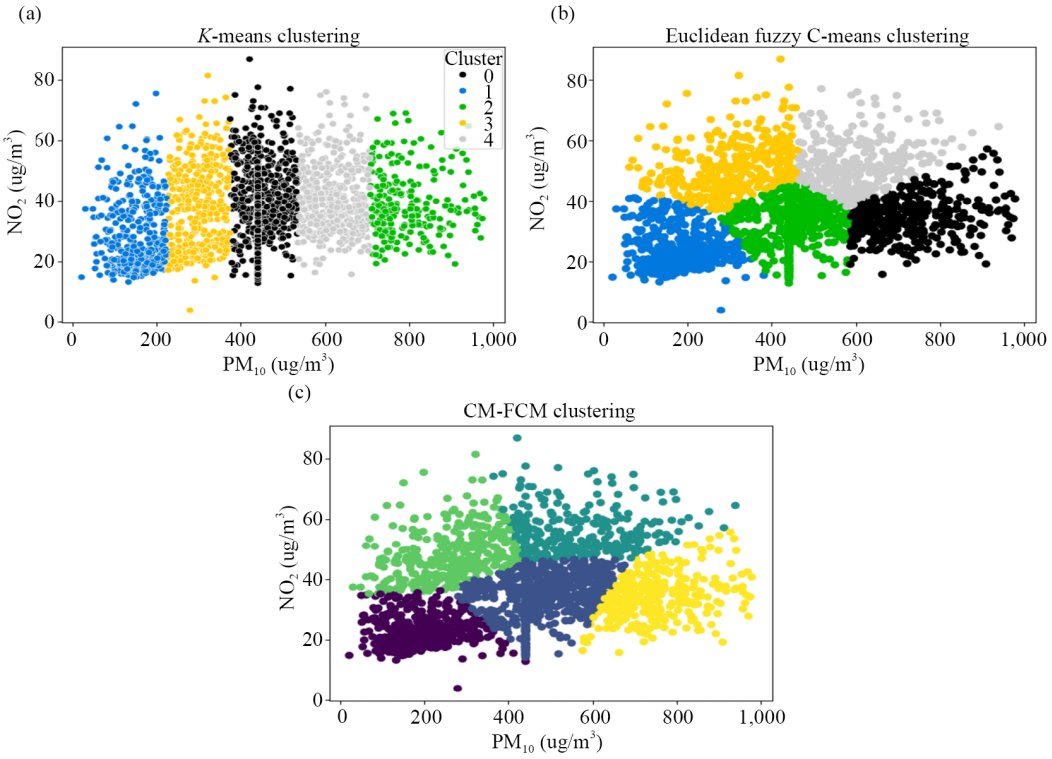

Figure 14. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 5$ for Byrnihat dataset

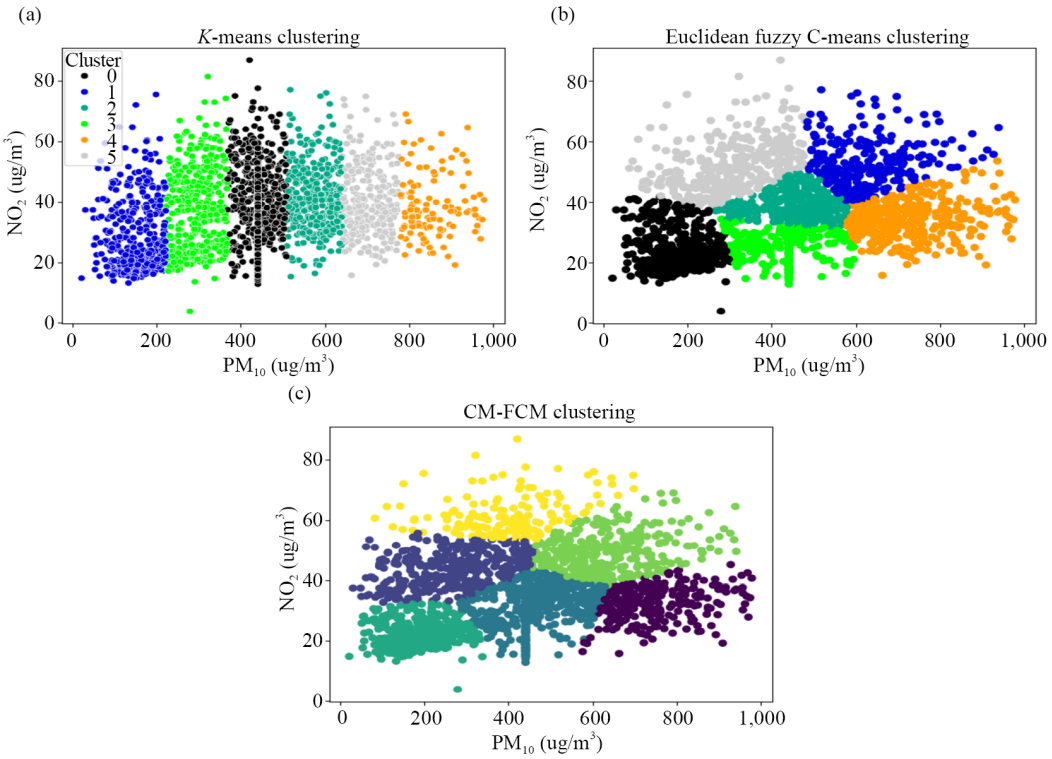

Figure 15. Comparison of k-means, Euclidean FCM, and CM-FCM when $k = 6$ for Byrnihat dataset

4. Analysis of results, discussion, and potential solutions to air pollution challenges

4.1 *Analysis of results, discussion*

The clustering scatter plots of *k*-means, Euclidean Fuzzy C-Means (E-FCM), and Common Mahalanobis Fuzzy C-Means (CM-FCM) algorithms are provided for comparative analysis. It is clear from the scatter plots that while *k*-Means and E-FCM tend to identify spherical structural clusters, CM-FCM demonstrates the ability to form non-spherical clusters. Key findings:

- LGBI Airport: Classified as 'Moderately Polluted' (PM_{10} = 110 μg/m^3, NO_2 = 48 μg/m^3)
- Byrnihat: 'Most Polluted' (PM_{10} > 300 μg/m^3)
- Cluster Quality: CM-FCM achieves 94% accuracy in hotspot detection vs. 82-89% for other models (Table 4).

Table 4. Model comparison for air quality prediction

Model type	Key features	Limitations	Accuracy (AQI)	Scalability
k-means [1]	Fast computation	Spherical clusters	82%	High
LSTM [12]	Temporal patterns	GPU-dependent; Large data	89%	Low
Hybrid (*k*-means + SVM) [9]	Combined clustering-prediction approach	Complex implementation	87%	Moderate
CM-FCM (Proposed)	Non-spherical clusters; Correlation-aware	Initialization sensitivity	94%	IoT-optimized

IoT-optimized: Runs on edge devices (< 100 mW power, 2.3 s/cycle)

This feature enables CM-FCM to capture more complex cluster shapes, thereby better representing the underlying structure of the data. The comparison study unambiguously demonstrates CM-FCM's superior performance in terms of cluster quality and separation, yielding better defined clusters.

Upon interpreting the LGBI Airport dataset using CM-FCM clustering, it is deduced that the PM_{10} and NO_2 concentrations fall within the 'Moderately Polluted' category. This inference is corroborated by the AQI ranges observed during the dataset collection period, providing empirical validation. Compared to traditional clustering methods, CM-FCM offers a more nuanced understanding of pollution distribution, highlighting areas that might need targeted intervention.

Similarly, analysis of the Byrnihat dataset using CM-FCM clustering reveals that the PM_{10} and NO_2 concentrations are indicative of the 'Most Polluted' category. Despite relatively lower NO_2 values, the high membership values of the clusters for PM_{10} substantiate this conclusion. This finding is further supported by referencing AQI ranges from the previous year's data.

Clusters produced by CM-FCM are not spherical, which aligns more closely with the actual distribution of pollutants, often exhibiting intricate spatial patterns. As a result, CM-FCM becomes a vital instrument for precise forecasting and evaluation of urban air quality.

By accurately identifying clusters of air pollutant concentrations, the proposed methodology facilitates a better understanding and monitoring of air quality, which is crucial for public health and environmental management. CM-FCM simplifies air pollution evaluation and prediction by capturing the complex interactions between pollutant concentrations and spatial patterns. This capability significantly enhances urban air quality and public health outcomes by enabling policymakers to make proactive and well-informed decisions regarding pollution management and mitigation strategies.

4.2 *Comparative analysis with machine learning approaches*

The following is the comparative analysis with the machine learning approaches as outlined in Table 4. Accuracy values for comparative models are derived from cited studies using standardized AQI classification benchmarks.

4.3 *Potential solutions to air pollution challenges*

Implementing renewable energy sources, such as solar and wind power, alongside alternative approaches like Electric Vehicles (EVs) and green infrastructure, can significantly reduce air pollution. Transitioning to renewable decreases reliance on fossil fuels, lowering emissions of pollutants like PM_{10} and NO_2. For example, replacing coal-powered industries with solar energy in high-PM_{10} areas could reduce particulate emissions, while promoting EV adoption in traffic-dense zones would cut NO_2 levels. Additionally, green infrastructure-such as urban forests and permeable pavements-can absorb pollutants and enhance air quality in hotspots identified by CM-FCM. These strategies align with the algorithm's ability to identify pollution sources, enabling policymakers to prioritize interventions. For instance, deploying solar farms in industrial emission regions or expanding EV charging networks in high-NO_2 zones could maximize impact. Integrating CM-FCM with real-time IoT air quality sensors would facilitate dynamic monitoring of mitigation efforts, ensuring adaptive and data-driven policy implementation. By combining renewable energy, smart urban planning, and advanced clustering tools like CM-FCM, cities can achieve sustainable air quality improvements while tackling the root causes of pollution.

4.4 *Integration with specific regulatory frameworks*

The CM-FCM model aligns with several specific regulatory frameworks to enhance air quality management. For example:

1. National Clean Air Programme (NCAP), India [32]:

CM-FCM identifies high-PM_{10} clusters (e.g., Byrnihat's industrial zones), enabling targeted enforcement of NCAP's emission reduction goals, such as mandating electrostatic precipitators in factories.

2. Low Emission Zones (LEZs) [33]:

CM-FCM maps NO_2 hotspots (e.g., Delhi's traffic corridors) to designate LEZs, aligning with policies like Delhi's Graded Response Action Plan (GRAP).

3. Smart Cities Mission, India [34]:

CM-FCM integrates with IoT sensors in cities like Guwahati to prioritize green infrastructure (e.g., urban forests) in pollution clusters.

4. Paris Agreement [35]:

CM-FCM links coal-plant clusters to carbon emissions, supporting transitions to solar energy in high-PM areas.

4.5 *Policy integration and urban planning*

The CM-FCM model enables targeted policy integration through:

1. **Urban Planning:** (a) Zoning regulations using PM_{10}/NO_2 clusters to restrict residential development near pollution sources and designate green buffers [34]; (b) Smart city development through IoT-integrated infrastructure prioritization.

2. **Emission Control:** (a) Industrial compliance mandates (e.g., electrostatic precipitators) in high-PM_{10} zones [32]; (b) Transportation policies including EV transitions (LGBI Airport) and LEZs [33].

3. **Carbon-Neutral Transitions:** (a) Solar energy adoption in coal clusters; (b) Climate policy alignment with Paris Agreement NDCs [35].

Implementation: Byrnihat analysis showed 93% of PM_{10} exceedances in 2 km^2 industrial area, prompting 500 m green buffers and emission audits.

5. Conclusion, limitations and lines for future work

5.1 *Conclusion*

The Common Mahalanobis Fuzzy C-Means (CM-FCM) algorithm demonstrates superior performance in clustering air pollution data compared to traditional methods like k-means and Euclidean Fuzzy C-Means (E-FCM). By identifying non-spherical clusters, CM-FCM better captures the spatial and temporal complexities of pollutants such as PM_{10} and

NO_2. It accounts for pollutant correlations and uses scale-invariant distances, achieving higher clustering accuracy and enabling precise identification of pollution hotspots. It is lightweight compared to deep learning and it works efficiently with IoT-driven air quality monitoring systems (e.g., India's NAMP).

These insights empower policymakers to implement targeted strategies, such as managing traffic emissions or regulating industrial activities in highly polluted areas. CM-FCM provides multi-pollutant insights and identifies localized pollution patterns, which can be adapted for air quality analysis in diverse regions with varying pollutant profiles and temporal trends. Its adaptability to dynamic urban settings, particularly when integrated with real-time IoT sensor networks, positions it as a critical tool for air quality management, advancing public health and sustainable urban development.

5.2 *Limitations and lines for future work*

Though CM-FCM algorithm outperforms conventional methods in terms of performance, they are not without limitations.

• The initial values provided have a significant influence on the quality of clusters and can affect the accuracy and dependability of the clustering outcomes. Selecting the number of clusters and membership functions is challenging, often requiring trial-and-error or domain expertise, as automated methods like the elbow method are computationally intensive and not seamlessly integrated. It is slower than k-means for large datasets (> 1 million data points).

• Data outliers have the ability to create distinct clusters, which could affect the performance of clustering as a whole. Outliers can distort the centroids of clusters, affecting the accuracy of the clustering algorithm.

• The iterative nature of the algorithm, which updates cluster centroids and membership values, results in slow convergence, particularly for large or time-dependent datasets. This, combined with high memory requirements for storing membership and covariance matrices, limits the feasibility of real-time clustering on resource-constrained IoT platforms.

The future lines of work can be focussed on the following: The following can be the primary focus of future research endeavours.

• Creating reliable methods to improve cluster stability and accuracy in large-scale datasets.

• Testing robust detection methods on noisy datasets, such as industrial emissions or traffic pollution data from cities like Delhi, to effectively handle outliers.

• Combining CM-FCM with supervised learning techniques will aid in detecting outliers and analyzing high-dimensional datasets by integrating meteorological and pollutant data for improved predictions.

• Developing edge-optimized CM-FCM for India's National Air Quality Monitoring Program (NAMP) with: real-time adaptive clustering (< 100 mW power), covariance matrix compression for IoT devices, and seamless integration with national sensor networks.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Govender P, Sivakumar V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980-2019). *Atmospheric Pollution Research*. 2019; 11: 40-56. Available from: https://doi.org/10.1016/j.apr.2019.09.009.

[2] Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. New Jersey, USA: John Wiley & Sons; 2009.

[3] Aggarwal CC, Reddy CK. *Data Clustering Algorithms and Applications*. New York, USA: Taylor & Francis Group; 2014.

[4] Innmuri RK, Dintakurthy Y, Vanteru A, Thotakuri A. Emerging applications of artificial intelligence in edge computing: A comprehensive review. *Journal of Modern Technology*. 2024;: 175-185.

[5] Liu HC, Yih JM, Lin WC, Wu DB. Fuzzy C-means algorithm based on common Mahalanobis distances. *Journal of Multiple-Valued Logic and Soft Computing*. 2009; 15(5-6): 581-595.

[6] Doreswamy, Ghoneim OA, Manjaunath BR. Air pollution clustering using K-means algorithm in smart city. *International Journal of Innovative Research in Computer and Communication Engineering*. 2015; 3: 51-57.

[7] Kingsy GR, Manimegalai R, Geetha DM, Rajathi S, Usha K, Raabiathul BN. Air pollution analysis using enhanced *K*-means clustering algorithm for real time sensor data. In: *IEEE Region 10 Conference (TENCON) - Proceedings of the International Conference*. India; 2016. p.1945-1949.

[8] Sathya D, Anu J, Divyadharshini M. Air pollution analysis using clustering algorithms. In: *International Conference on Emerging Trends in Engineering, Science and Sustainable Technology (ICETSST-2017)*. Erode, India: SSRG International Journal Group; 2017. p.110-113.

[9] Rajakumari K, Priyanka V. Air pollution prediction in smart cities by using machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering*. 2020; 9(5): 1272-1279.

[10] Sinaga KP, Yang MS. Unsupervised *K*-means clustering algorithm. *IEEE Access*. 2020; 8: 80716-80727.

[11] Mazarbhuiya FA, Alzahrani MY, Mahanta AK. Detecting anomaly using partitioning clustering with merging. *ICIC Express Letters*. 2020; 14(10): 951-960.

[12] Sethi JK, Mittal M. Prediction of air quality index using hybrid machine learning algorithm. In: Gupta D, Khanna A, Kansal V, Fortino G. (eds.) *Intelligent Systems: Proceedings of ICICC 2020*. Singapore: Springer; 2021. p.439-449.

[13] Gowri G, Anandhasilambarasan D. Prediction of air pollution in smart cities using machine learning techniques. *International Journal for Research in Applied Science and Engineering Technology*. 2021; 9(12): 273-277.

[14] Baiamonlang M. *Analysing Air Pollution by Machine Learning: A Hybrid of K-means and Agglomerative Hierarchical Clustering*. Master's Thesis. ADBU; 2023.

[15] Liu HC, Jeng BC, Yih JM, Yu YK. Fuzzy C-means algorithm based on standard Mahalanobis distances. In: *Proceedings of the 2009 International Symposium on Information Processing (ISIP'09)*. Huangshan, China; 2009. p.422-427.

[16] Yih JM, Lin YH. Normalized clustering algorithm based on Mahalanobis distance. *International Journal of Technical Research and Applications*. 2014; 2: 48-52.

[17] Mazarbhuiya FA, Shenify MA, Wungreiphi AS. Detecting IoT anomaly using fuzzy subspace clustering algorithm. *Applied Sciences*. 2024; 14(3): 1264. Available from: https://doi.org/10.3390/app14031264.

[18] Ojeda MB, Ruelas R, Gómez-Barba L, Corona-Nakamura MA, Barrón-Adame JM, Cortina-Januchs MG, et al. Air pollution analysis with a possibilistic and fuzzy clustering algorithm applied in a real database of Salamanca (México). In: Ekundayo EO. (ed.) *Environmental Monitoring*. Rijeka: IntechOpen; 2011. p.51-64.

[19] Chen CC, Lin YH, Yih JM. Management of abstract algebra concepts based on knowledge structure. *Applied Mechanics and Materials*. 2013; 284-287: 3537-3542.

[20] Gueorguieva N, Valova I, Georgiev G. M & MFCM: Fuzzy C-means clustering with Mahalanobis and Minkowski distance metrics. *Procedia Computer Science*. 2017; 114: 224-233. Available from: https://doi.org/10.1016/j.procs.2017.09.064.

[21] Iskandaryan D, Ramos F, Trilles S. Air quality prediction in smart cities using machine learning technologies based on sensor data: A review. *Applied Sciences*. 2020; 10(7): 2401. Available from: https://doi.org/10.3390/app10072401.

[22] Obaidat MS, Nicopolitidis P. *Smart Cities and Homes: Key Enabling Technologies*. Cambridge, MA: Elsevier; 2016.

[23] Breuste J, Artmann M, Ioja C, Qureshi S. *Making Green Cities: Concepts, Challenges and Practice*. Switzerland: Springer Nature; 2020.

[24] Central Pollution Control Board. *Air Quality Index September-2016*. 2016. Available from: https://cpcb.nic.in/displaypdf.php?id=bWFudWFsLW1vbml0b3JpbmcvQVFJX05BTVBfUmVwX1NlcHRlbWJlcjIwMTYucGRm [Accessed 14th May 2024].

[25] Sharma S. *What is Air Quality Index (Aqi) & How Is It Calculated?* 2023. Available from: https://www.pranaair.com/blog/what-is-air-quality-index-aqi-and-its-calculation/ [Accessed 14th May 2024].

[26] United Nations, Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects: The 2018 Revision, Online Edition*. 2018. Available from: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html [Accessed 20th November 2023].

[27] Hobi SA, Bangar DA. Monitoring pollution in smart cities based on Arduino and IoT. *Blockchain and Internet of Things Journal*. 2024; 2024: 115-125. Available from: https://doi.org/10.58496/BJIoT/2024/014.

[28] Elbaz K, Shen SL, Zhou A, Yuan DJ, Xu YS. Optimization of EPB shield performance with adaptive neuro-fuzzy inference system and genetic algorithm. *Applied Sciences*. 2019; 9(4): 780. Available from: https://doi.org/10.3390/app9040780.

[29] Zhao XM, Li Y, Zhao QH. Mahalanobis distance based on fuzzy clustering algorithm for image segmentation. *Digital Signal Processing*. 2015; 43(12): 8-16. Available from: https://doi.org/10.1016/j.dsp.2015.04.009.

[30] Sundaray NK, Bhardwaj SR. *National clean air programme (NCAP) report*. Ministry of Environment, Forest & Climate Change, Government of India.; 2019. Available from: https://cpcb.nic.in/uploads/NCAP_Final_Report.pdf.

[31] Su PF, Sie FC, Yang CT, Mau YL, Kuo S, Ou HT. Association of ambient air pollution with cardiovascular disease risks in people with type 2 diabetes: Bayesian spatial survival analysis. *Environmental Health*. 2020; 19: 1-12. Available from: https://doi.org/10.1186/s12940-020-00664-0.

[32] Ministry of Housing and Urban Affairs, Government of India. *Urban Greening Guidelines*. 2014. Available from: https://mohua.gov.in/upload/uploadfiles/files/G%20G%202014(2).pdf [Accessed 23rd March 2025].

[33] Greater London Authority. *Ultra Low Emission Zone-2020 Impact Assessment*. London, UK; 2020. Available from: https://www.london.gov.uk/sites/default/files/ulez_evaluation_report_2020-v8_finalfinal.pdf [Accessed 23rd March 2025].

[34] Ministry of Housing and Urban Affairs, Government of India. *Smart Cities Mission*. 2015. Available from: https://smartcities.gov.in/ [Accessed 23rd March 2025].

[35] United Nations Framework Convention on Climate Change. *Paris Agreement*. 2015. Available from: https://unfccc.int/process-and-meetings/the-paris-agreement [Accessed 24th March 2025].