Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021]
IEEE Xplore Part Number: CFP21F70-ART; ISBN: 978-1-7281-8501-9

K-Means Clustering of Ambient Air Quality Data of Uttarakhand, India during Lockdown Period of Covid-19 Pandemic

Sandeep Kumar Sunori
Department of ECE
Graphic Era Hill University
Bhimtal Campus
Bhimtal , India
sandeepsunori@gmail.com

Pushpa Bhakuni Negi
Department of Chemistry
Graphic Era Hill University
Bhimtal Campus
Bhimtal , India
pbiochem05@gmail.com
(Corresponding Author)

Sudhanshu Maurya
School of Computing
Graphic Era Hill University
Bhimtal Campus
Bhimtal, India
dr.sm0302@gmail.com

Pradeep Juneja
Department of ECE
Graphic Era University
Dehradun, India
mailjuneja@gmail.com

Anita Rana
Department of Chemistry
Kumaun University
Nainital, India
anitaranachem@gmail.com

Bhawana
Department of Chemistry
Graphic Era University
Dehradun, India
bishtbhawana29@gmail.com

2021 6th International Conference on Inventive Computation Technologies (ICICT) | 978-1-7281-8501-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT50816.2021.9358627

***Abstract*- The analysis of the lockdown effect during covid-19 pandemic on ambient quality of air of Uttarakhand, state of India, has been performed. The combination of SO_2, NO_2 and particulate matter (P.M.10) indicates ambient air quality characteristics. The clustering capability of the K-means clustering technique is investigated with two different approaches of measuring distance using MATLAB. The first approach is termed Euclidean distance and the second one is cosine distance. The data, which is clustered, is the air quality data containing three major components of air pollution such as P.M.10, SO_2, and NO_2 of different major cities of Uttarakhand.**

Keywords*- *Lockdown, covid-19, cluster, air quality, data point

I. INTRODUCTION

Uttarakhand is surrounded by the Himalayas, which is one of the most beautiful states of India that attracts everyone all over the world with its marvelously scenic landscapes. It is one of the rapidly developing states in India wherein 2017, domestic tourist arrival was 34.36 million and the visits of foreign tourist is more than about 0.13 million [1]. Being a young state, developmental activities are growing speedily. Recently, it has been observed that comparatively lesser reachable areas have also become the part of fast urbanization, broadly in terms of connectivity of roads, the advancement of international and domestic tourism, progress in horticulture, globalization of economy and steady move from the primary resource development to secondary sectors because of the nonexistence of land use policy [2]. As a result, there has taken place a huge rise in number of vehicles, industries, constructions, wide usage of fertilizers, consumption of bio and fossil fuels, *etc*. This, in turn, has a great effect on the quality of air, and the weather patterns.

II. LITERATURE SURVEY

The continuous use of fossil and biofuels raised the pollutants in the megacities to alarming levels [3], and the outflow of the pollutant streams from the large cities destroys the environment at both the regional and global levels. The transport vehicles and the emissions released from industries have played a major contribution to air pollution in Dehradun, the capital of Uttarakhand, India, and the nearby locations also according to some previous studies [4]. The combustion taking place in motor vehicles, with the use of diesel fuel or gasoline, is the significant origin of the NO2 component in the atmosphere. NO2 is formed by the quick oxidation of atmospheric NO by the ozone NO2 [5,6].

In the advanced countries, maximum hours of people are spent in the atmosphere which has high specific pollutant concentrations [7], hence inducing a great risk for their health [8]. Indoor pollutants include VOCs (volatile organic compounds). These are mainly released from the building materials, furniture, and various consumer products. The air also has a significant impact on pollutants breathed indoors [9].

Many clustering algorithms have been developed for data mining, such as reviewed by A. Olaode et al. [10]. Different clustering algorithms, or even different ways to use them on the same dataset, can lead to different partition results. None of them have proved to be the best technique in a large number of configurations. Some of these algorithms are useful in the field of electronic nose data clustering [11], each with its respective possibilities and limitations [12,13] depending on the application.

The lockdown period of the covid-19 pandemic has resulted in a drastic reduction in these pollution causing components of air which has shown a remarkable improvement in the air quality index.

III. METHODOLOGY

At present, the clustering capability of the K-means clustering technique is examined with two different methods of estimating the distance using MATLAB. The first method is termed Euclidean distance and the second method is cosine distance. The data is clustered in the air quality and taken from the available online source [14], containing three major components of air pollution such as P.M.10, SO_2, and NO_2, of different major cities of Uttarakhand, for the year 2020 (during lockdown period of covid-19) and the preceding year 2019.

The Euclidian distance between these two vectors R and S is given by equation (1) and the cosine distance is given by equation (2) [15, 16].

$$E = \sqrt{\Sigma(r_j - s_j)^2} \quad (1)$$

$$C = 1 - \cos(R, S) \quad (2)$$

$$\text{Where, } cos(R, S) = \frac{\Sigma r_j s_j}{\sqrt{\Sigma r_j^2 \Sigma s_j^2}} \quad (3)$$

This is an iterative method in which data points are assigned to the clusters in such a way that the summation of distances of each data point from the respective cluster center is the minimum.

At every iteration, this sum is decreased and cluster centers are updated. This algorithm continues to execute until this sum becomes minimum. The flow chart of this algorithm is presented in Fig. 2. Now the clusters found by this algorithm are compared with the actual known clusters, named 'polluted air' and 'clean air', to determine the clustering capability of this algorithm.

IV. SIMULATION AND RESULTS

The K-means clustering algorithm is applied on 140 data points (70 points for the year 2019 and 70 points for the year 2020). Initially, the algorithm is executed with the Euclidean distance approach. The corresponding Silhouette plot is depicted in Fig.1. The extent of separation of the resulting clusters can be assessed from the Silhouette plot. This plot gives an estimate of the closeness of all data points in one cluster to the data points in nearby clusters. A Silhouette value (SV) close to 1 indicates a good separation between clusters. A Small S.V. implies that the respective point is very close to the points of other clusters. Therefore this Silhouette plot indicates a good separation between the two clusters as so many points of both clusters has SV not very far from 1.

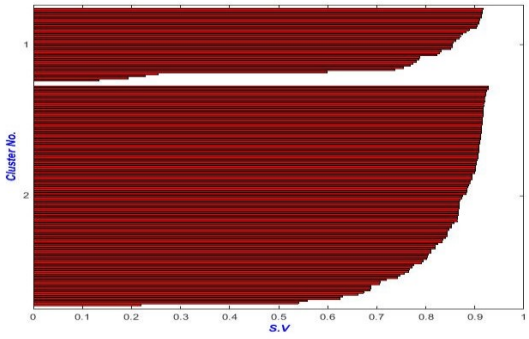

Fig.1 Silhouette plot with Euclidian distance

Table I indicates that only in 4 iterations, the algorithm reaches the best (minimum) value of sum which is equal to 69799.6.

TABLE I. PROGRESS OF ALGORITHM

Iteration No.	Sum
1	75481.9
2	70177.9
3	69799.6
4	69799.6

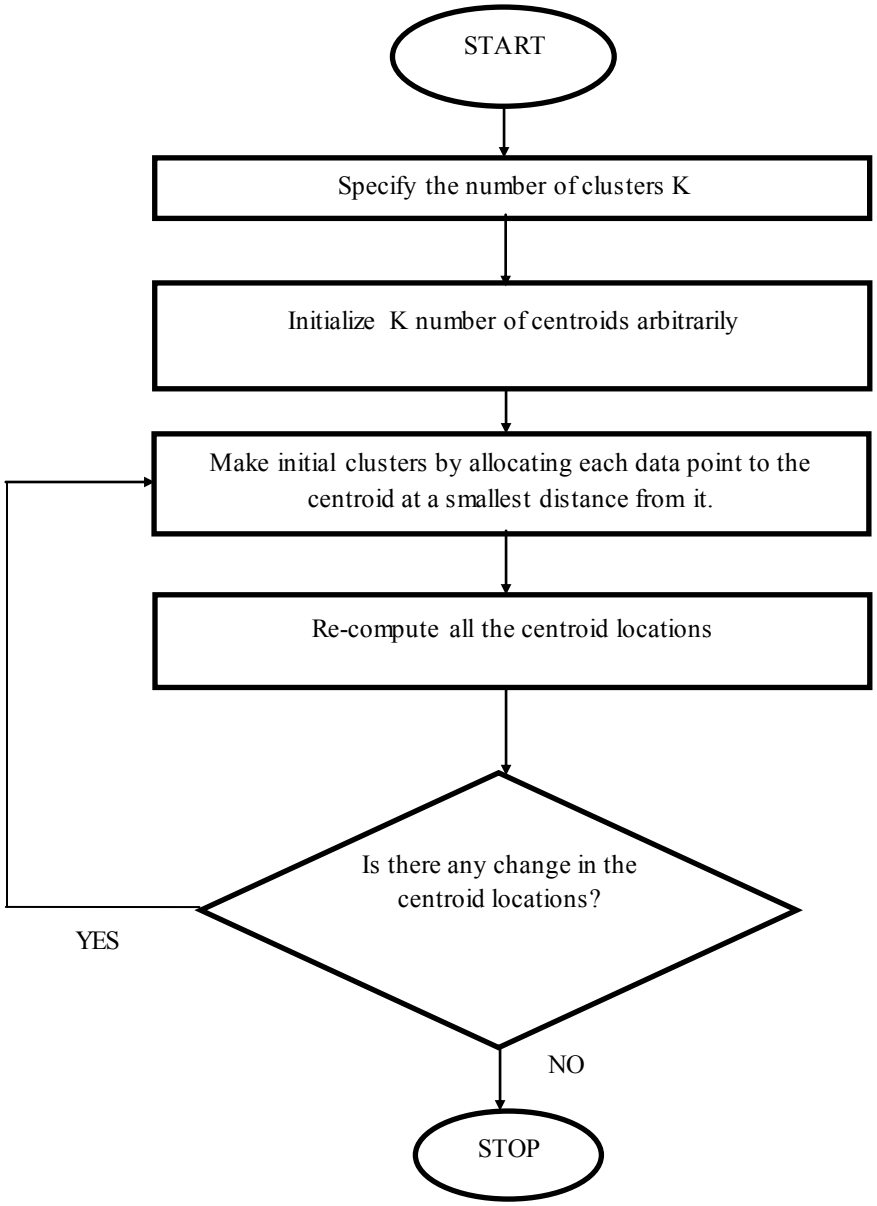

Fig.2. Flow chart of K-means clustering technique [17]

The final two clusters formed using the Euclidian distance approach is shown in Fig. 3.

The algorithm arrives at the following final centroids of cluster 1 and cluster 2,

Cluster 1 → [185.006, 24.812, 28.706]

Cluster 2 → [109.215, 15.372, 22.827]

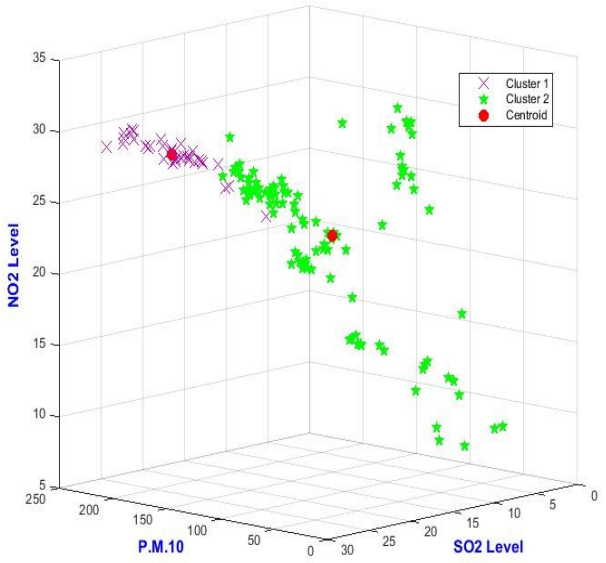

Fig.3. Clusters formed using the Euclidian distance approach

Now the result of this clustering approach (Euclidean distance) is compared with the actual known clusters ('polluted air' and 'clean air') as depicted in Fig.4. It is found that out of 140 points only 47 points don't fall in the correct cluster and the remaining points are allocated the correct cluster by this approach.

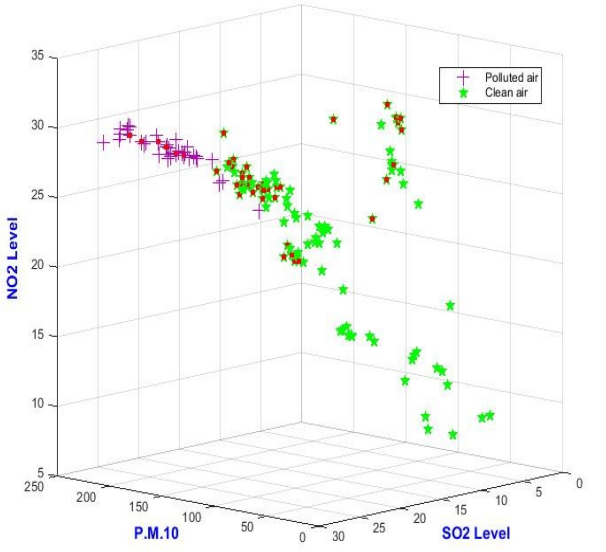

Fig.4. Comparison of formed clusters using Euclidian distance with actual clusters

The cosine distance approach of K-means clustering is used to check its clustering capability on the same data set of 140 points. The corresponding Silhouette plot, Clusters formed and the comparison with the actual clusters are presented in Figures 5, 6, and 7 respectively.

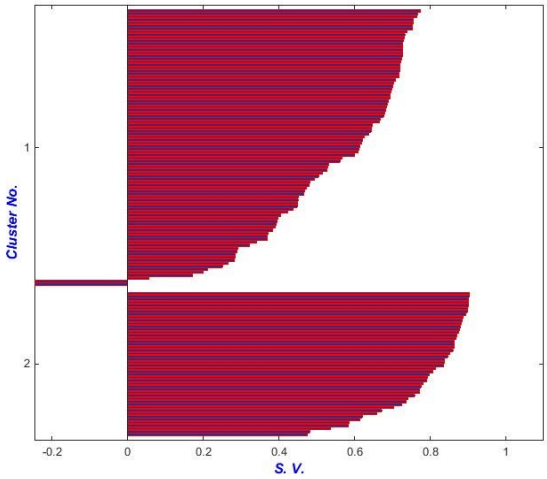

Fig. 5. Silhouette plot with cosine distance

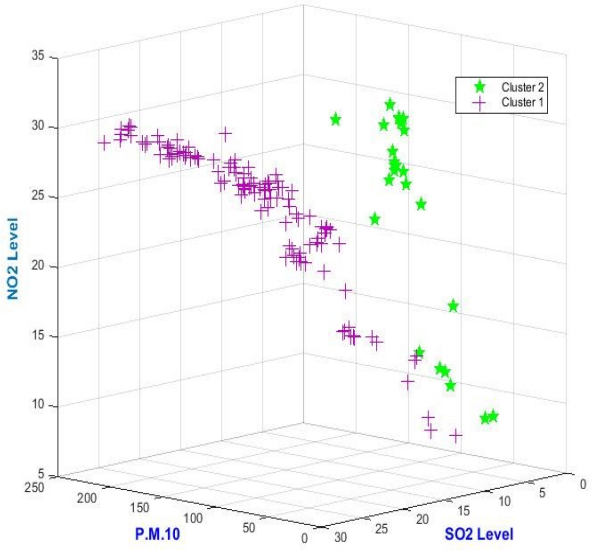

Fig. 6. Clusters formed using cosine distance approach

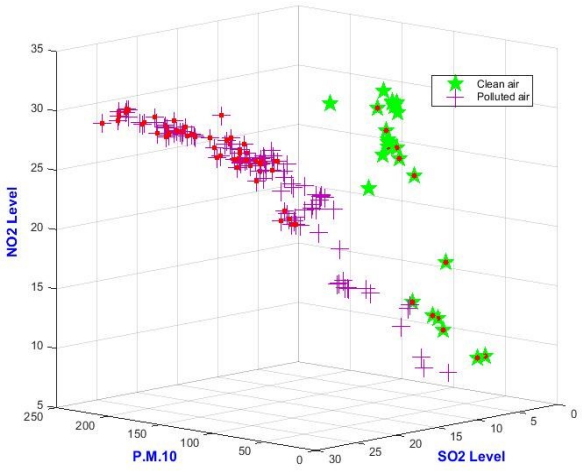

Fig.7. Comparison of formed clusters using cosine distance with actual clusters

The comparison shows out of 140 points 76 points do not appear in the correct cluster and the remaining points are allocated the correct cluster by this approach. So, the cosine distance approach of K-means clustering does not come up with a satisfactory clustering result as more than 50 percent of data points missed the actual cluster.

V. NOVELTY

After performing a persuasive literature survey, it has been observed that many clustering algorithms have been adopted by many researchers. It has been observed that for a very large data set the clustering result of the K-means clustering technique is much better than other techniques. So, in this work, for the large air quality data, this technique has been adopted which gives much better results than other clustering techniques used in previous research. This technique does not require vast statistical analysis, therefore, its implementation is comparatively modest. In addition to this, it executes very fast as compared to many other clustering techniques.

VI. CONCLUSION

In this research work, the mixed air quality data of two years i.e. 2019 (normal period) and 2020 (lockdown period of covid-19) were considered for Uttarakhand state, which has been clustered into two clusters namely polluted air and clean air using K-means clustering algorithm in MATLAB. Two different approaches to measuring distance have been used and their results are compared. It is concluded that for the given data set the Euclidian distance approach of K-means clustering gives a much better clustering result than the cosine distance approach.

REFERENCES

[1]. https://www.ibef.org/download/uttarakhand-dec-2018.pdf.

[2]. Ghosh, P., "Urbanization-A Potential threat to the Fragile Himalayan Environment," Current Science, 93 (2), 2007, pp. 126-127.

[3]. Butler T M, Lawrence M G, Gurger B R, van Aardenne J, Schultz M and Lelieveld J., "The representation of emissions from mega cities in global emission inventories," Atmos. Environ. 42(4), 2008, pp. 703-719.

[4]. Chauhan A, Pawar M, Kumar R and Joshi P C, "Ambient air quality status in Uttarakhand (India): A case study of Haridwar and Dehradun using air quality index," J. Am. Sci., 6(9), 2010, pp. 565-574.

[5]. Chaney AM, Cryer DJ, Nicholl EJ, Seakins PW, "NO and NO2 interconversion downwind of two different line source in suburban environments," Atmospheric Environment, Vol. 45, Issue 32, 2011, pp.5863-5871.

[6]. Palmgren F, Berkowicz R, Hertel O, Vignati E, "Effect of reduction of NOx on the NO2 levels in urban streets," The Science of the Total Environment, Vol. 189-190, 1996, pp.409-415

[7]. G.A. Ayoko, H. Wang, "Volatile Organic Compounds in Indoor Environments," Indoor Air Pollution, Springer, Berlin, Heidelberg, 2014, pp. 69-107.

[8]. K.W. Tham, "Indoor air quality and its effects on humans—a review of challenges and developments in the last 30 years," Energy and Buildings, Vol.130 , 2016, pp. 637-650.

[9]. M. Verriele, C. Schoemaecker, B. Hanoune, N. Leclerc, S. Germain, V.Gaudion, N. Locoge, "The mermaid study: indoor and outdoor average pollutant concentrations in 10 low-energy school buildings in France," Indoor Air, Vol. 26 , 2016, pp. 702-713.

[10]. A. Olaode, G. Naghdy, C. Todd, "Unsupervised classification of images: a review," International Journal of Image Processing (IJIP), Vol. 8, Issue 5, 2014, pp.325-342.

[11]. S. Marco and A. Gutierrez-Galvez, "Signal and Data Processing for Machine Olfaction and Chemical Sensing: A Review," IEEE Sensors Journal, Vol. 12, No. 11, Nov. 2012, pp. 3189-3214.

[12]. A. Hierlemann, R. Gutierrez-Osuna, "Higher-order chemical sensing," Chem. Rev., 108 , 2008, pp. 563-613.

[13]. M. Bicego, G.Tessari, G.Tecchiolli, M.Bettinelli, "A comparative analysis of basic pattern recognition techniques for the development of small size electronic nose," Sensors and Actuators B: Chemical, Vol. 85, Issues 1–2, 2002, pp. 137-144.

[14]. Official Website of Uttarakhand Pollution Control Board, Government of Uttarakhand (https://ueppcb.uk.gov.in/pages/display/95-air-quality-data).

[15]. Dibya Jyoti Bora, Dr. Anil Kumar Gupta, "Effect of different measures on the performance of K-means algorithm: an experimental study in Matlab," IJCSIT, Vol. 5(2), 2014, pp. 2501-2506.

[16]. M. Thangarasu, Dr. H.Hannah Inbarani, "Analysis of K-means with multi viewpoint similarity and cosine similarity measures for clustering the document," IJAER, Vol.10, No.9, 2015, pp. 6672-6675.

[17]. P. Sasikumar and S. Khara, "K-Means Clustering in Wireless Sensor Networks," Fourth International Conference on Computational Intelligence and Communication Networks, Mathura, 2012, pp. 140-144.