International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

A Study on Air Pollution Trends in Sanathnagar, Hyderabad using Principle Component and Cluster Analysis

Bhavana Hemavani[1], Dr. G. V. R. Srinivasa Rao[2]

[1]Research Scholar, Department of Civil Engineering, Andhra University, Visakhapatnam, India

[2]Professor, Department of Civil Engineering, Andhra University, Visakhapatnam, India

Abstract: ***Principle Component Analysis (PCA) and Cluster Analysis (CA) are used to understand the correlations among various air pollutants and the meteorological parameters in an industrial and commercial area at Hyderabad, India. The analysis is done using a decade long data and the correlations are obtained seasonally. The results from the both analysis have shown similar trends. The KMO test sampling adequacy and the Bartlett test significance values have proved that the results of analysis are satisfactory.***

Keywords: Principle Component Analysis, Cluster Analysis, Multivariate Analysis, SPSS, Hyderabad

1. Introduction

Recent years, research related to forecasting air pollution trends using statistical analysis have seen an increase. Main factors influencing the changes in air pollution dispersion are Meteorological [Precipitation, wind speed, wind direction, temperature and boundary layer height] and Pollutant sources [Transportation, Industrial, agricultural, natural]. Both the factors are to be thoroughly studied to predict the air quality trends [1].

PCA (Principle component analysis) and CA (cluster analysis) are one of the well-used statistical methods to analyse and study trends in ambient air pollution. These statistical methods are used by researchers worldwide to explain the patterns and trends between pollutants and meteorological factors. PCA is an analytical technique which reduces dimensions to create new variables know as Principle components (PCs) [2] [3], which are linear combinations of original variables. Then again, CA is a classification method used to divide the data into a set of clusters. The main objective of a cluster is to identify objects with similar characters and differentiate them from others in a different cluster. A dendogram or tree diagram represents a cluster between n numbers of variables. Main objectives of this paper are to identify the relationship between meteorological factors and pollution parameters, affecting pollution dispersion and obtaining the influences in air pollution trends by applying a Statistical approach such as PCA and CA.

2. Literature Survey

Air quality in Hyderabad, India often exceeds the national ambient air quality standards and the prominent pollutants being PM (particulate matter) which has a high tendency to enter the human respiratory tracks. Prolonged exposure to vehicular air pollutants is harmful to health, precisely in "lungs". Relationship between exposure time to vehicular exhaust was given by studying blood samples of Traffic police, who are having high risk and an increase in oxidant stress, decrease in levels of antioxidants, nitric oxide which may create an imbalance in the body, leading to lung damage [4][5].

The association between pollution parameters and its influencing factors (meteorological factors) are required to understand the air quality trends, recent advancements in the statistical analysis made it possible to study their relationship effortlessly with the support of software's available

Cluster analysis predicted maximum ozone value in Houston by using air pollutant and meteorological parameters [6]. Many other researchers [6] [7] [8] used cluster analysis to determine the AQI pattern of different monitoring stations along with pollution dispersion & spatial variations [9].

3. Methodology

Area description: Hyderabad is one of the largest metropolitan city in India. Hyderabad experiences a minimum temperature of 11.60^{o} c and a maximum of 40.50^{o} c. The present study based in sanathnagar, the centre of the city.

Data: Hourly data of PM, gases and meteorological parameters of sanathnagar, CPCB station (2007-2017) was collected. The data comprised of missing values by more than 5 %, which implies they can neither be removed nor neglected, else it will reduce the sample size. The missing values were replaced by IBM SPSS 26. Since P-value from the MCAR test was higher than 0.05, therefore multiple imputation method is the best fit one. The data divided seasonally, i.e., summer, pre-monsoon, monsoon, winter and PCA, CA, are being used in the analysis.

The parameters consist of the following: 1) PM (particulate matter) µg/m3 2) Gases SO_2, CO, NO_x, O_3(µg/m^3). 3) Meteorological features- Wind Speed (WS, m/s), Temperature (AT^{o} C), Relative Humidity (RH, %), Wind direction (deg), Radiation (SR, W/m^2), Barometric Pressure (BP, mmHg).

4. Results and Discussions

4.1*Principle Component Analysis* for eleven parameters of the data set is conducted to determine the factors influencing the air pollution trends (seasonally) in each cluster produced by cluster analysis. The Output of SPSS 26 for PCA consists of a) Descriptive statistics b) Correlation matrix c) Rotated Component matrix. For all the four seasons KMO test sampling adequacy is found to be higher than 0.5, and Significance from Bartlett's test is lesser then 0.00 which implies the analysis is satisfactory

Summer season – The major influencer is BP mm/Hg with maximum mean=708.93 from descriptive statics. Out of all the parameters in the correlation matrix, SR has the highest correlation with Temperature. The maximum and minimum extracted values were temperature=0.750, respectively SO_2 = 0.230.Total Variance % ranged from 21.905, 20.885, 16.726 for Component 1,2, and 3.The curve flattens between components 3 and 4 (shown in fig : 1). Since component has Eigenvalues less than one, only three components retained. Rotated Component matrix observed that the following parameters (table no-1) overloaded as Principle Components (PC).

Table 1: PCs for summer seasons

Components	Parameters
PC1	WD
PC2	Temp, SR
PC3	Temp, RH, SR, NO_x, CO

Pre-monsoon –Descriptive statics states that BP and PM are most influencing factors with the highest mean. Correlation matrix observed that out of all parameters NO_x is Found to be most influencing and having the highest correlation with CO. Communalities analysis gave maximum value as O_3= 0.898, minimum value WD= 0.368. % Total variance for components 1, 2, 3 respectively is 30.616, 20.133, and 14.275. Eigenvalues- Screen plot (Fig- 2) states that only three PCs retained as the curve flattens at component 3&4. The following table shows the parameters overloaded as PCs.

Table 2: PCs for Pre-monsoon

Components	Parameters
PC1	NO_x, CO, SO_2, PM
PC2	SR
PC3	Temp, WD

Monsoon- BP is the most influencing factor with the highest mean of 709.30. Observations from the co-relation matrix, which states that out of all parameters NO_x had the highest correlation with CO. The maximum and minimum extracted values are WD=0.914, respectively BP=0.519.Total variance for components 1,2,3,4 & 5 are 28.465, 16.338, 11.781, 11.281, 9.766 respectively. Eigenvalues plotted by scree plot (fig-3) states that curve flattens at component 5 & 6. The rotated component matrixes overloaded as Principle components are in given in a table form.

Table 3: PCs for Monsoon

Components	Parameters
PC1	SR,BP
PC2	SO_2, NO_x
PC3	CO, NO_x
PC4	Temp
PC5	WD

Winter season- Parameters having the highest mean are BP (714.966), and PM (115.266) the influencing factors (fig 4). NO_x has the highest correlation with CO. The maximum and minimum values of extractions are SR (0.716) and WS (0.494). Total variance varies from 29.757, 20.023, and 11.468 respectively for 1, 2, 3&4 components. Three PCA components were extracted and loaded as PCs.

Table 4: PCs for Winter Season

Components	Parameters
PC1	CO,PM
PC2	O_3, AT
PC3	SR

4.2 Cluster Analysis

It is a statistical approach for sorting cases, observations, or variables for a set of data. In the present study, Hierarchical Agglomerative Cluster Analysis (HACA) required for analysing the data, output included proximity matrix, agglomerative schedule and a dendogram. The proximity matrix gives the shortest and the longest distance between any two cases. Shortest distance implies the two cases joined in the first cluster, and the longest distance indicating a sudden change (Jump) in coefficient between any two cases. Agglomeration schedule followed by proximity matrix in the output display how HACA progressively clusters the cases or observations. At the end of SPSS output, a dendogram appears and examined from Left to Right. The vertical lines in a Dendogram represent the grouping of clusters and stages. After identifying a set of meaningful subgroup, a final step that can be taken place for further validity is to compare CA to PCA analysis.CA analysis for all the four seasons gave a total of 11 stages, and 10 cases. Summer season (Dendogram fig-5) has two clusters identified as Cluster 1 (CO, WS, SO_2, O_3,NO_x, Temp, PM, RH) highly correlated and Cluster 2 (WD, SR, BP) as moderately correlated. The minimum distance among the clusters was for CO –WS (6729.8) forming as stage one in agglomeration table, and maximum length was between CO-BP (612512234.63) seen as the last stage. In Pre-Monsoon (Fig-6)& Monsoon(Fig-7) highly correlated parameters are identified as CO, WS, SO2, AT0 C, RH, O3, NO_x, and moderately correlated as PM, WD, SR. The minimum distance in stage one was CO- WS, and maximum CO- BP. Winter season(fig-8) has two clusters, i.e. highly correlated parameters- CO, WS, SO_2, AT^{o}C, RH, O_3 and moderately parameters- PM, WD, SR, NO_x.

5. Figures

Fig 1 to 4 Scree Plots for Four seasons (PCA)

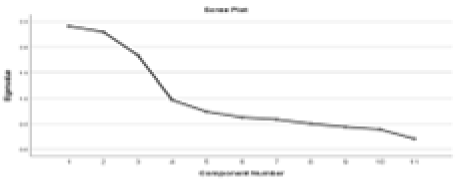

Fig no 1- Scree plot for Summer season

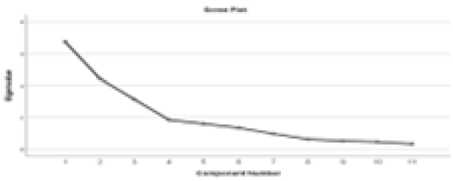

Fig no 2- Scree plot for Pre-monsoon season

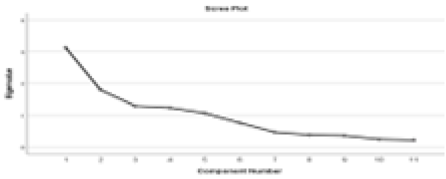

Fig no3- Scree plot for Monsson season

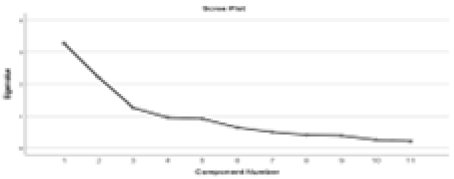

Fig no3- Scree plot for Monsson season

Fig no 5 to 8 Dendogram (cluster analysis)

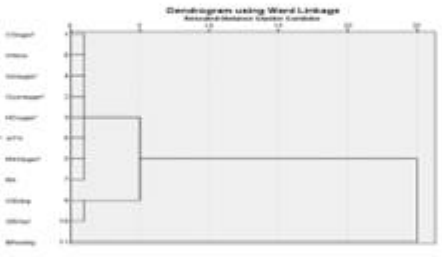

Fig no 5- Dendogram for summer season

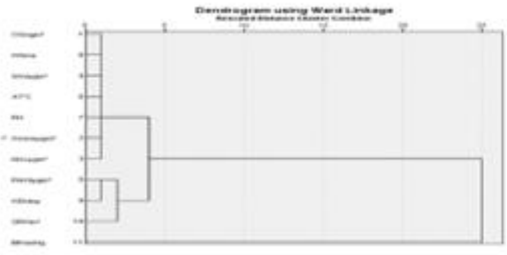

Fig no 6- Dendogram for Pre-monsoon season

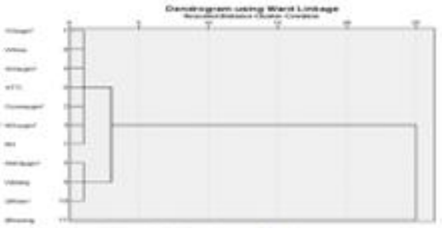

Fig no 7- Dendogram for Monsoon season

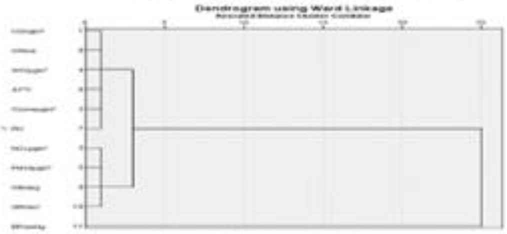

Fig no 8- Dendogram for Winter season

6. Conclusion

Assessment of air quality monitoring data of sanathnagar using statistical analysis (PCA & CA) showed a significant influence of meteorological conditions on air pollution. The results from PCA compared with CA showed similar trends. CA showed parameters such as BP, Temp, WD, WS are significant influencers on the pollutants. Gases like NO_x highly correlated with CO, SO_2, and O_3, and Particulate matter also acts as influencers in all the seasons. Through an analysis of four seasons for a period of 2007-2017 proved the existence of a correlation between PM, Gases, and meteorological factors. Limitations exist in this area of research. Analysis of different terrains for a more extended period would have explained the variations accurately. More factors like Precipitation, Sunlight duration, different types of terrain and geographical locations could have given a more in-depth analysis. An analysis considering the influence of different altitudes on the diffusion of air pollutants from vertical and horizontal directions could have given much better results.

References

[1] Liu, Chong, Xuguang Wang, and Qingchuan Wang. "Air Pollution Quality Prediction Model Based on Clustering and Multivariate Regression." *EkolojiDergisi* 107 (2019).

[2] Pandey, Bhanu, Madhoolika Agrawal, and Siddharth Singh. "Assessment of air pollution around coal mining area: emphasizing on spatial distributions, seasonal variations and heavy metals, using cluster and Principle component analysis." *Atmospheric pollution research* 5, no. 1 (2014): 79-86.

[3] Wang, Shengwei, and Fu Xiao. "AHU sensor fault diagnosis using Principle component analysis method." *Energy and Buildings* 36, no. 2 (2004): 147-160.

[4] Suresh, Y., MM Sailaja Devi, V. Manjari, and U. N. Das. "Oxidant stress, antioxidants and nitric oxide in traffic police of Hyderabad, India." *Environmental Pollution* 109, no. 2 (2000): 321-325.

[5] Gupta, Sharat, Shallu Mittal, Avnish Kumar, and Kamal D. Singh. "Respiratory effects of air pollutants among nonsmoking traffic policemen of Patiala, India." *Lung*

India: Official Organ of Indian Chest Society 28, no. 4 (2011): 253.

[6] L. S. Darby, Cluster Analysis of Surface Winds in Houston, Texas, and the Impact of Wind Patterns on Ozone, Journal of Applied Meteorology, 44 (2005), 1788-1806.

[7] Saithanu, Kidakan, and Jatupat Mekparyup. "Air quality assessment in the urban areas with multivariate statitical analysis at the east of Thailand." *International Electronic Journal of Pure and Applied Mathematics* 7, no. 4 (2014).

[8] K. Saithanu, J. Mekparyup, Clustering of Air Quality and Meteorological Variables Associated with High Ground Ozone Concentration in the Industrial Areas, at the East of Thailand, International Journal of Pure and Applied Mathematics, 3, NO. 81 (2012), 505-515.

[9] Banerjee, T., S. B. Singh, and R. K. Srivastava. "Development and performance evaluation of statistical models correlating air pollutants and meteorological variables at Pantnagar, India." *Atmospheric Research* 99, no. 3-4 (2011): 505-517.

[10] Nagendra, SM Shiva, and Mukesh Khare. "Principle component analysis of urban traffic characteristics and meteorological data." *Transportation Research Part D: Transport and Environment* 8, no. 4 (2003): 285-297.

[11] Guttikunda, S.K., Kopakka, R.V., Dasari, P. and Gertler, A.W., 2013. Receptor model-based source apportionment of particulate pollution in Hyderabad, India. *Environmental monitoring and assessment*, *185*(7), pp.5585-5593.

[12] Yim, Odilia, and Kylee T. Ramdeen. "Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data." *The quantitative methods for psychology* 11, no. 1 (2015): 8-21.

[13] Pires, J. C. M., S. I. V. Sousa, M. C. Pereira, M. C. M. Alvim-Ferraz, and F. G. Martins. "Management of air quality monitoring using Principle component and cluster analysis—Part I: SO2 and PM10." *Atmospheric Environment* 42, no. 6 (2008): 1249-1260.

[14] He, Jianjun, Sunling Gong, Ye Yu, Lijuan Yu, Lin Wu, Hongjun Mao, Congbo Song et al. "Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major Chinese cities." *Environmental pollution* 223 (2017): 484-496.

[15] Choi, Jaesung, Yong Shin Park, and Ju Dong Park. "Development of an aggregate air quality index using a PCA-based method: a case study of the US transportation sector." *American Journal of Industrial and Business Management* 5, no. 02 (2015): 53.

[16] Trivedi, Dinesh Kumar, Kaushar Ali, and GufranBeig. "Impact of meteorological parameters on the development of fine and coarse particles over Delhi." *Science of the Total Environment* 478 (2014): 175-183.

Author Profile

Mrs. Bhavana Hemavani (Ph.D.), Research Scholar, Department of Civil Engineering, Andhra University.

Dr. G. V. R. Srinivasa Rao, Professor, Department of Civil Engineering, Andhra University College of Engineering (A), Andhra University, Visakhapatnam – 530003.