Assessment of Chennai's Ambient Air Quality Data using Multivariate Analysis from 2005 to 2015

Srinivasan Karunanithi[1,*], Thirumalini Perumal[2] and Kabbilawsh[3]

[1] VIT University
Vellore, India

[2] VIT University
Vellore, India

[*] *Corresponding author's email: srinivasan.k [AT] vit.ac.in*

ABSTRACT— ***This paper deals with spatial classification and their respective sources of pollutants within the selected TNPCB and CPCB monitoring station based on a Eleven year Database (2005 –2015). HACA grouped the eight monitoring stations into three different clusters, bases on concentration of air pollutants. HACA results also speculate that a methodological difference lies in the stations operated by CPCB and TNPCB. PCA was performed individually on each cluster to determine the source of pollutant. PCA analysis showed production of wax, fuel, petrochemical feedstock and lube by petrochemical and manufacturing industries, soil dust emission due to construction activities and vehicular movement form the foundation of the pollution base for the first cluster, whereas abundant increase in vehicular fumes which had resulted from drastic increase in personal vehicle trips and decline of public ridership in buses and trains leading to detour was the source for pollutant concentration in second cluster. Illegal commercialization of residential plots, construction of restaurants and complexes, heavy traffic movement explains the result of the third cluster.***

Keywords— TNPCB, CPCB, Hierarchical agglomerative Cluster analysis (HACA), Principal Component analysis (PCA),Oxides of Nitrogen (NO_x) and Sulphur dioxide (SO_2), Total Respirable Particulate matter (TSPM).

1. INTRODUCTION

Urban metropolitan cities of developing countries experience a serious challenge of air pollution due to rapid increase in population, unplanned Industrialization and abnormal increase in personal motorization. Global Burden of disease reported 20 % of death occurs due to outdoor air pollution.

India's contribution to world's population was estimated to be 1033 million and Delhi, Mumbai and Chennai are enlisted in top 10 heavily populated metropolitan cites. Urban and semi-urban parts of Chennai are industrialized without proper planning and traffic congestion have led to declining environmental conditions leading to adverse effects on health of people. Kapoor, S. (1997) has concluded about 18 million children around the globe especially in developing countries having high concentration of lead in their body than the permissible limits, which might have resulted from vehicular emissions [1]. Chennai vehicular population has increased from 5 lakhs in 1991 to 30 lakhs today.

WHO has defined air pollution as those substances, which are circulated in air due to action of mankind in sufficient amounts to cause adverse effects on health, belongings, and productivity of crop or to hinder with the enjoyment of property. Among all pollutants, those which cause considerable amounts of harmfulness are termed as Criteria pollutant. It includes Total suspended particulate matter (TSPM), Carbon monoxide (CO), Ozone (O3), Lead, oxides of Nitrogen and Sulphur (NOx and SOx) [2].

In order to assess the prevailing situation and pollution trend within an area, spatial as well as temporal data gathering by continuous monitoring of constant ambient air quality is the need of the hour. At the same time monitoring of air stations lays an enormous financial burden on the organization and institutions. Monitoring stations must be placed in such a way that they must represent the general pollution exposure. Thereby it is necessary to identify redundant monitoring stations and criteria pollutants that on eliminating render minimum loss of data. There exists no reliable guideline, which specifies any optimal method for selection of monitoring stations. Atsushi Iizuka has provided a methodology to investigate the performance of monitoring networks in Shizuoka, Funabashi and Hiroshima Prefacture in Japan [3].

1.1Multvariate Data Analysis

Multivariate data analysis have been widely used in performance assessment, identification of redundant measurement, modeling, examination of pollutant trend, prediction of concentration, and evaluation of temporal spatial variation and in many studies related to surface and ground water monitoring [4, 5, 6, 7].

In past fifteen years ,multivariate statistical technique such as Principal Component Analysis(PCA),Factor Analysis(FA), Linear Discriminant Analysis(LDA),End Member Analysis(EMA),Canonical Discriminant Analysis(CDA),Cluster analysis(CA),Neural Network(NN),are being applied in order to decrease relatively Large number of real time air quality variables to a smaller number of Orthogonal factors.

1.2 Hierarchical agglomerative Cluster analysis (HACA)

The objective of the cluster analysis is to spill the given data values to groups so that data within group are alike to each other with respect to variables or attributes of interest and the groups themselves differ from one another. Clustering effect summarizes according to degree of proximity among the cluster elements and of the separation among the clusters [1, 8].

1.2 Principal Component analysis (PCA)

Principal component analysis is statistical procedure for identification a smaller number of uncorrelated variables, called "Principal components", which are obtained by transforming the original set of inter-correlated variables. The aim of PCA is to elucidate the maximum amount of variance with the fewest number of principal components. The components are calculated as linear combinations of the original variables. It helps in removing multi-collinearity, or when numbers of predictors relative to the number of observations are more [9, 10].

2. METHODOLOGY

2.1 Monitoring Location

The air quality in Chennai city is being monitored by Tamil Nadu Pollution Control Board (TNPCB) and Central Pollution Control Board (CPCB) as well as by independent academic institutes like IIT Madras. CPCB and TNPCB are nodal agencies for operating the air pollution monitoring stations in Chennai city. CPCB monitoring area consists of Manali, Thiruvottyur and kathivakkam whereas TNPCB are monitoring Adyar, Anna Nagar, Kilpauk, Thiyagaraya Nagar and Vallalar Nagar.(Figure 1) For this study ,four important criteria pollutants – Oxides of Nitrogen and Sulphur dioxide [NO_X and SO_2], and Total Suspended particulate matter [TSPM] and Respirable particulate matter [RSPM] data's were obtained.(Table 1)

2.2 Data Collection

Monthly-recorded data from 2005 to 2015 at the selected air quality monitoring stations were obtained from TNPCB stations which were converted to yearly averages for analysis. In case of CPCB data, while receiving it we got in form of annual average concentration.

In some cases where no data was available for many consecutive months, the data corresponding to those months for all 8 stations were dropped for that year's analysis. Thus the data pertaining to 2008, 2013, 2014 and 2015 was entirely excluded for Cluster analysis due to unavailability of data pertaining to CPCB monitoring stations. Similarly RSPM data pertaining to 2013, 2014 and 2015 for TNPCB was also excluded due to unavailability.

2.3 Standardization of data and software used

Usually data transformation was done before performing multivariate analysis (HACA and PCA) for a variety of reasons to ensure data normality, changing the weights of different variables. Here standardization was not done since all parameters are measured in microgram per millimeter cube ($\mu g/m^3$) making it not necessary for data transformation. The datasets were analyzed by PCA and HACA using commercial software SPSS 16 in this study.

3. RESULTS AND DISCUSSION

3.1 Hierarchical agglomerative Cluster analysis (HACA)

HACA was carried out in order to obtain the spatial variation based on the similarity level .Though sources of each pollutant are distinct in nature, for example oxides of Nitrogen (NO_x) and Sulphur dioxide (SO_2) are by products of combustion reaction while particulate matter have a number of sources which includes wind driven and traffic. Despite this, we have chosen to study the pattern of all pollutants together, rather than to do a cluster analysis of the criteria pollutant individually because in single instant the variation of the all pollutant

belongings to various stations are taken into consideration during the clustering effect. Due to unavalibaility of data pertaining to CPCB stations, cluster analysis was not performed on 2008, 2013, 2014 and 2015.

Figure 2 to Figure 4 shows the main results of HACA application at all stations for pollutants of SO_2, NOx, TSPM and RSPM. Rescaled distance cluster combine (RDCC) was used as tool in deciding the appropriate number of cluster The eight stations was classified into three cluster i.e. Kathivakkam(C1),Manali(C2),Thirvottiyur(C3) always grouped together and formed in a single cluster over the seven years speculating a methodological difference in the stations operated by CPCB and TNPCB .This phenomenon needs to be examined by concerned nodal agencies.

The same problem of two different organization handling the same study area showing differences have been recorded by Saksena et al in 2003 in delhi. In the research done by Saksena *et.al* in 2003 in megacity of Delhi have used Cluster Analysis as a data Classification technique to analyze air quality monitoring network which consists of 9 stations totally out of which 6 are monitored by CPCB and 3 by NEERI They have proved a methodological difference exist between the station operated by NEERI and CPCB [11].

The stations in Cluster 1 constitute an Industrial hub and one of most sensitive air pollution area mainly Manali which is situated 20km north of Chennai, connected by road. The area consists of industries like oil refineries, fertilizer plants, chemicals, fabric yarn and steel, COCL, MFL, TNPP and Manali Petro Chemicals.

The cluster 2 consist of four stations [Anna nagar(T1),Kilpauk(T3),Thiyagaraya Nagar(T4) and Vallalar Nagar(T5) whereas Adyar (T4) emerged as a distinct cluster which individually formed third cluster. TNPCB performs ambient air quality checks twice a week at Adyar, Kilpauk, T Nagar and Anna Nagar.

Concentration of TSPM and RSPM would have been the major reason for splitting of TNPCB monitored station into two clusters since range of value of Adyar monitoring station was between 33 $\mu g/m^3$ to 102 $\mu g/m^3$ from 2005 to 2015, while other stations ranged from a minimum value of 63 $\mu g/m^3$ to 424 $\mu g/m^3$, that only few values were less than permissible limit of 100 $\mu g/m^3$

Respirable suspended Particulate matter has risen from 33$\mu g/m^3$ to 63$\mu g/m^3$ in 2012 and dropped to 58 $\mu g/m^3$ in 2014 in Adyar, according to the Tamil Nadu Pollution Control Board (TNPCB). Another reason which can attributed to the clustering effect is criteria air pollutant concentration is rising in Adyar, whereas there has been a drop in the readings in Anna Nagar, T Nagar and Kilpauk in the same period.

3.2 Principle component analysis

PCA was applied as a data reduction technicque on four parameters of the given data sets in order to determine which pollutant contribute much to the variations in each cluster produced by HACA.Generally the first Principle component explains the most of varience in the data.The PCA results consist of scree plot and component matrix which is obtained by Varimax roation with kaiser normalisation (Table2).

PC's are ratoted by varimax roatation because it maximizes the squared factor loadings in each component. In other words it simplifies the columns of the component matrix. In each component the large loadings are increased and the small ones are decreased so that each component only has a few variables with large loadings.

A scree plot is graph drawn between the Eigen values associated with prinicpal component in descending order versus the number of prinicapl components. It helps in visually assessing the components which explains the most of the variability in the data. The The ideal pattern in a scree plot is a steep curve, followed by a bend and then a flat or horizontal line. Those components or factors in the steep curve before the first point that starts the flat line trend need to be retained for explainig the varaiability.The remaining factors explain a very small proportion of the variability and are likely unimportant (Figure 5).

Correlation matrix was used to calculate the prinicpal components though the variables are measured in same scale because there is appreciable difference in variances between the four criteria pollulants.

3.2 PCA result of Cluster 1

Kathivakkam(C1),Manali(C2),Thirvottiyur(C3) formed the first cluster. Among the three, Manali is an Industrial zone in which Chennai Petroleum Corporation Limited (CPCL) is located and it is one of the most elaborate and modernistic refineries in India with facilities for production of wax, fuel, feedstock pertaining to petrochemicals and lube. The manufactured product include LPG, Motor Spirit, Superior Kerosene, A Turbine Fuel, High Speed Diesel, Naphtha, Bitumen, Lube Base Stocks, Paraffin Wax, Fuel Oil, Hexane and Petrochemical feed stock.

Manali is a home for various industries which have existed for more than twenty years. There are twenty-eight types of industries which are classified as twenty belonging to major industries and rest eight constitute medium scale industries.

Kathivakkam lies near to Ennore Thermal power plant. Thiruvottiyur is a densely populated area and it is surrounded by mighty industrial establishments namely Royal Enfield Motors, Greaves Cotton LTD, CPCL, SRF, Ashok Leyland and Ennore foundries.It is also supported by numerous small scale industries.

This scree plot (Figure 5 (A)) shows that two factors are enough to explain the variable nature of this cluster since the line begins to unbend after factor-2 .Principal components with Eigen value greater than one were used for analysis. The first PC explained 50.461% of original data variance which had an Eigen value of 2.279 and showed strong positive loadings on $S0_2$ (0.986) and NOx (0.917) concentration (Table2).

In the research made by Pereira et al (2007) [12] in Oporto-MA of Portugal he had showed source of SO_2 concentration in atmosphere is because of presence of oil refinery, one petrochemical plant, one thermoelectric plant working with natural gas, one incineration unit, and one international shipping port. Nevertheless, motor vehicle traffic is estimated to be responsible for a significant amount of pollutants emitted to the atmosphere. Similarly high concentration of SO_2 can be attributed to presence of number of industries in Manali, Thiruvottiyur and Kathivakkam.

Strong positive loadings for NO_x, which are evidently indicative of emissions from automobile engines of cars, lorries, buses and boats which moves in and out of this Cluster zone. Janssen et al [14] has assessed the exposure to traffic related air pollution of children attending schools and concluded that both indoor and outdoor air, concentrations of $PM_{2.5}$significantly increases with increasing truck traffic density and significantly decreased with increasing distance and similarly NO_x concentrations significantly increased with increasing car traffic.

In past ten years, Thiruvottiyur which belongs to this cluster is undergoing enormous bloom in real estate business mainly because of its locational proximity to Central railway station of Chennai and relatively lower residential plots with decent amenities. Increase in the number of plots coupled with digging and dust from construction activities are responsible for the increase in particulate matter.

The second principal component explains 31.948% of the total variance which shows moderate loading on TSPM (0.732) and strong loading on RSPM (0.883) which are evident from open burning and soil dust emissions around the sampling station.

3.3 PCA result of Cluster 2

Anna nagar (T1), Kilpauk (T3), Thiyagaraya Nagar (T4) and Vallalar Nagar (T5) formed the second cluster. The scree plot (Figure 5 (B)) shows that two factors are enough to explain the variable nature of this cluster since the line begins to unbend after factor-2. It has two PC's explaining nearly 84.67% of the total variance. The first PC explained 56.972 % of the variance and has higher loadings on TSPM (0.917) and RSPM (0.940) whereas second PC explained 27.696% of the total variance and has positive loading on Sulphur dioxide (0.749) negative loading on NO_x (-0.712) (Table 2).

Anna Nagar consists of West and East bus terminuses, from where buses ply to different parts of the city especially from West line. Design of roads in Anna Nagar is based on a matrix system which bears similarity to those in developed countries. Roads are aligned parallel and perpendicular to each other. Traffic congestion is more in the Roundtana intersection which is located in 2^{nd} and 3^{rd} avenue due to presence of flourishing high scale commercial buildings. Vehicular Congestion level has increased several folds in the Roundtana in past twenty years and also consumed percentage of road length for parking. In Study made by IIT madras on source apportionment they narrowed down that vehicles contributes one sixth of the particulate matter

The first PC showed strong positive loadings on TSPM (0.917) and RSPM (0.940) concentration (Table2).

The vehicular emissions were the main reason for high levels of RSPM in T Nagar .Bad roads and increase in the number of vehicles coupled with digging and dust from construction activities are responsible for the increase in particulate matter.

The four areas which come under this cluster have become more sophisticated by construction of number of flyovers, foot over bridges and construction of roads which are free of stop signs or regulatory signals enabling the vehicles to move in higher traveling speeds has decreased the accessibility to walking. It has motivated more motorized travels. In the research done by Centre for science and Environment, the effect of detour has increased the vehicle distance from few meters of walk to kilometers in car which has made one petrol to eject 24 milligram of extra PM and a diesel car to emit 240 milligrams of PM. In past twenty years ridership on bus and train has decreased drastically paving way for personal motorization. This accumulative impact of the detour and

conversion of emission free walking to unwanted polluting motorized trips on traffic volume over years can be attributed for strong loadings on TSPM, RSPM and SO_2.

3.3 PCA result of Cluster 3

It includes only Adyar monitoring station. Unlike the location of other monitoring station it is neither an Industrial hub nor has any type of power plant. It is home to number of education and research institutions. Educational institutions include both schools and colleges. Colleges in this neighborhood consists of IIT Madras, CLRI, Anna University, TTTI, NIFT and Asian college of Journalism. It includes nearly 12 prominent well-established schools and newer ones are being constructed. There always exists enormous vehicular movement in and around these places.

This scree plot (Figure 5 (C)) shows that 2 of those factors explain most of the variability because the line starts to straighten after factor 2.The first PC explained 40.640 % of original data variance which had an Eigen value of 1.626 and shows strong positive loadings on Particulate matter.

Particulate matter which includes particles of aerodynamic diameter of less than 10 microns mainly consists of substances like debris formed during construction activities, soot discharged into air after combustion from vehicles, fly ash, cement and pollen grains. Commercialization and construction are the main reasons for the increase.

Inhabitants of Adyar in the past ten years have been leaving the house for marketable and commercial purpose leading to spatial congestion and pollution. The Chennai Metropolitan Development authority which is nodal agency for all types of activities in field of construction have completely prohibited those kinds of malpractices and people do not follow them. In spite of that letting out houses for commercial purposes in places like Kalakshetra colony, is leading to congestion and pollution. Construction of restaurants and complexes is increasing in Adyar. This has led to an increase in traffic mainly pointing to a restaurant being built on Gandhi Nagar Second Avenue. The smoke from the Vehicle fumes, construction dust cause drop in Adyar air quality

The second PC explained 32.74% of original data variance which had an eigen value of 1.310 and strong negative loading on SO_2 (-0.809) which can attributed to that fact that SO_2 is emitted by the burning of fossil fuels by power plants, industrial units, Since Adyar is primarily a residential zone and absence of power plants explains the lower concentration of SO_2.

In the research undertaken by Ghazali et al. (2009), Janssen et al (2001) and Morawska et al. (2002) [13, 14, 15] they have shown NO2 is outcome of heavy traffic and motor vehicle emissions. Strong positive loading on NO2 may be attributed to rapid commercialization of Adyar in turn leading to dense development of Traffic.

4. CONCLUSION

The productiveness of combining the Principal component analysis with Cluster analysis is demonstrated in this study. HACA application of the data produced three clusters based on four parameters observed from different local surroundings. PCA analysis showed production of wax, fuel, petrochemical feedstock and lube by petrochemical and manufacturing industries, soil dust emission due to construction activities and vehicular movement form the foundation of the pollution base for the first cluster, whereas abundant increase in vehicular fumes which had resulted from drastic increase in personal vehicle trips and decline of public ridership in buses and trains leading to detour was the source for pollutant concentration in second cluster. Illegal commercialization of residential plots, construction of restaurants and complexes, heavy traffic movement explains the result of the third cluster. Chennai people need to adapt themselves with use of public transport rather than private vehicles, integrate multi-modal transport options and government needs to enforce four wheeler restraint policies.

5. REFERENCES

[1] Kapoor, S., (1997.)Vehicular Exhaust: Killing us Softly but Surely, Down to Earth, p. 63, March 15, 1997

[2] World Health Organisation (WHO),(1996). Regional Health Report, Regional Office for South East Asia, NewDelhi, 1996.

[3] Atsushi Iizuka, Shintaro Shirato, Atsushi Mizukoshi, Miyuki Noguchi, Akihiro Yamasaki, Yukio Yanagisawa., 2014. A Cluster Analysis of Constant Ambient Air Monitoring Data from the Kanto Region of Japan. International Journal of Environmental Research and Public Health,11(7),6844-6855

[4] Shrestha, S., Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji River basin, Japan. Environmental Modeling and Software 22 (4), 464-475.

[5] Singh,K.P., Malik, A., Mohan, D., Sinha, S., 2004. Multivariate statistical techniquesfor the evaluation of spatial and temporal variations in water quality of Gomti River (India)- a case study. Water Research 38 (18), 3980-3992.

[6] Singh, K.P., Malik, A., Sinha, S., 2005.Water quality assessment and apportionment of pollution sources of Gomti River (India) using multivariate statistical technique-a case study.AnalyticaChimicaActa 538 (1-2), 355-374

[7] Yindana, S.M., Ophori, D., Banoeng-Yakubo, B., 2008.A multivariate statistical analysis of surface water chemistrydata—The Ankobra Basin, Ghana. Journal of Environmental Management 86, 80–87.

[8] Kaufman L., and RousseeuwP. J., ,Finding groups in data, Wiley Interscience, New York, 1990.

[9] Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG. Management of air quality monitoring using principal component and cluster analysis Part I: $SO_2$and PM10. Atmos Environ 2008a;42:1249-60.

[10] Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG. Management of air quality monitoring using principal component and cluster analysis Part II: CO, NO2and O3. Atmos Environ 2008b; 42:1261-74

[11] Saksena, S., Joshi V., Patil R S., 2003.Cluster analysis of Delhi's ambient air quality data. Journal of Environmental Montoring 5, 491-499.

[12] Pereira, M.C., Santos, R.C., Alvim-Ferraz, M.C.M., 2007.Air quality improvements using European environment policies: a case study of SO2in a coastal region in Portugal. Journal of Toxicology and Environment Heal the Part A: Current Issue 70, 347-351

[13] Janssen, N.A.H., Van Vliet, P.H.N., Aarts, F., Harssema, H., Brunekreef, B., 2001 Assessment of exposure to traffic-related air pollution of children attending school near motorways. Atmospheric Environment 35, 3875-3884

[14] Ghazali, N.A., Ramli, N.A., Yahaya, A.S., 2009. A study to investigate and model the transformation of nitrogen dioxide into ozone using time series plot. European Journal of Scientific Research 37 (2), 192-205.

[15] Morawska, L., Jayaratne, E.R., Mengersen, K., Jamriska, M., Thomas, S., 2002. Differences in airborne particle and gaseous concentration in urban air between weekdays and weekends. AtmosphericEnvironment 36, 4375-4383.

Table 1: Site Characteristics of the Air quality Monitoring network of Chennai Area

STATION NAME	STATION ID	CATEGORY
Kathivakkam	C1	Industrial Area
Manali	C2	Industrial Area
Thirvottiyur	C3	Mix of Industrial and Residential Area
Anna Nagar	T1	Residential Area, Traffic
Adyar	T2	Residential Area, Traffic
Kilpauk	T3	Residential Area, Traffic
Thiyagaraya Nagar	T4	Traffic, Commercial Area
Vallalar Nagar	T5	Mix of Industrial ,Residential Area and Traffic

Table 2: Loadings of various pollutants after PCA- Varimax rotation at the study area

VARIABLES	CLUSTER 1		CLUSTER 2		CLUSTER 3	
	PC1	**PC2**	**PC1**	**PC2**	**PC1**	**PC2**
SO_2	**0.986**	0.026	0.519	**0.749**	0.133	**-0.809**
NO_X	**0.917**	-0.126	0.533	**-0.720**	0.165	**0.813**
TSPM	-0348	**0.732**	**0.917**	-0.125	**0.908**	-0.053
RSPM	0.170	**0.883**	**0.940**	0.117	**0.864**	-0.023
EIGEN VALUE	2.279	1.108	2.279	1.108	1.626	1.310
VARIABILITY (%)	50.461	31.948	56.972	27.696	40.640	32.741
CUMULATIVE VARIANCE (%)	50.461	82.409	56.972	84.667	40.640	73.381

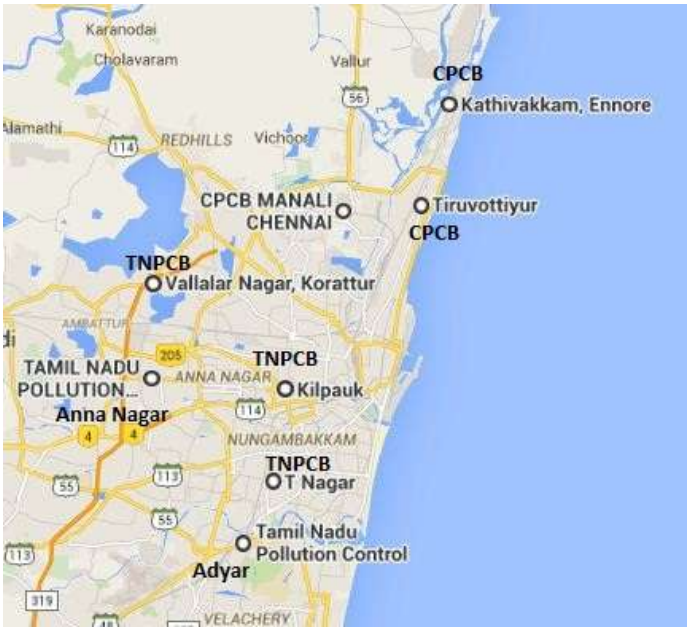

Figure 1: Location of various monitoring stations under CPCB and TNPCB

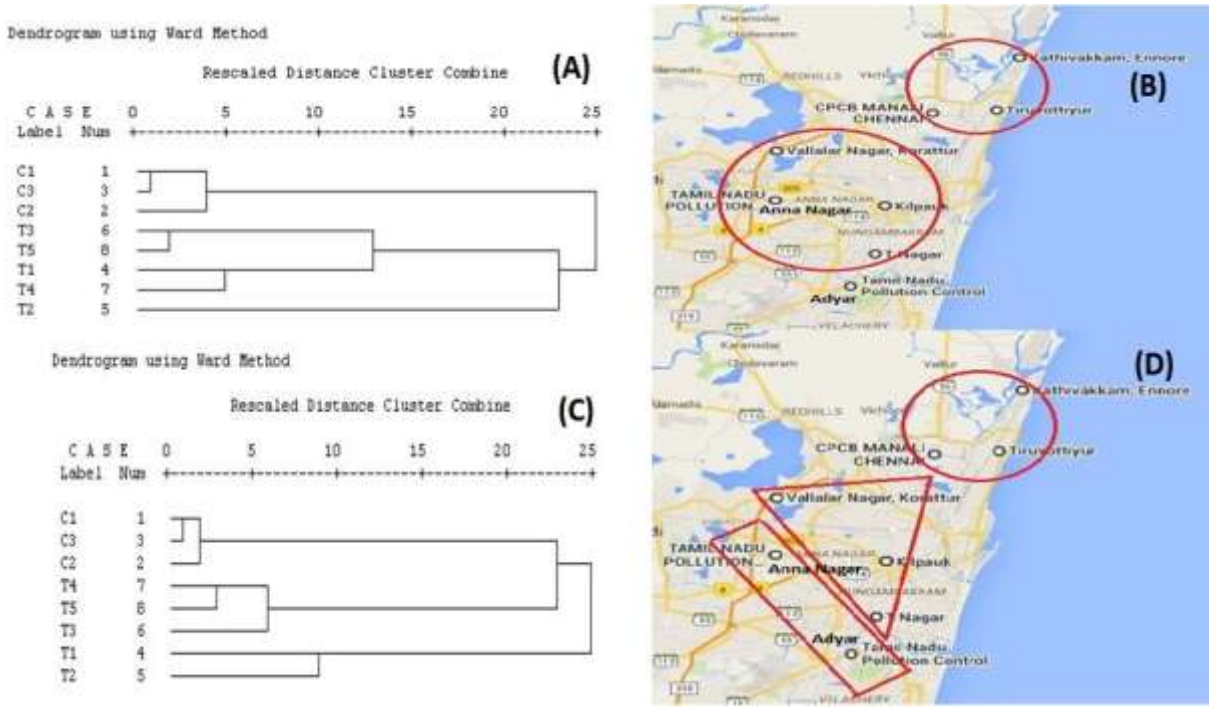

Figure 2: **(A)** Dendrogram of Different Cluster of Air monitoring station of the year 2005
(B) Spatial Location of Cluster formation of year 2005
(C) Dendrogram of Different Cluster of Air monitoring station of the year 2006
(D) Spatial Location of Cluster formation of year 2005

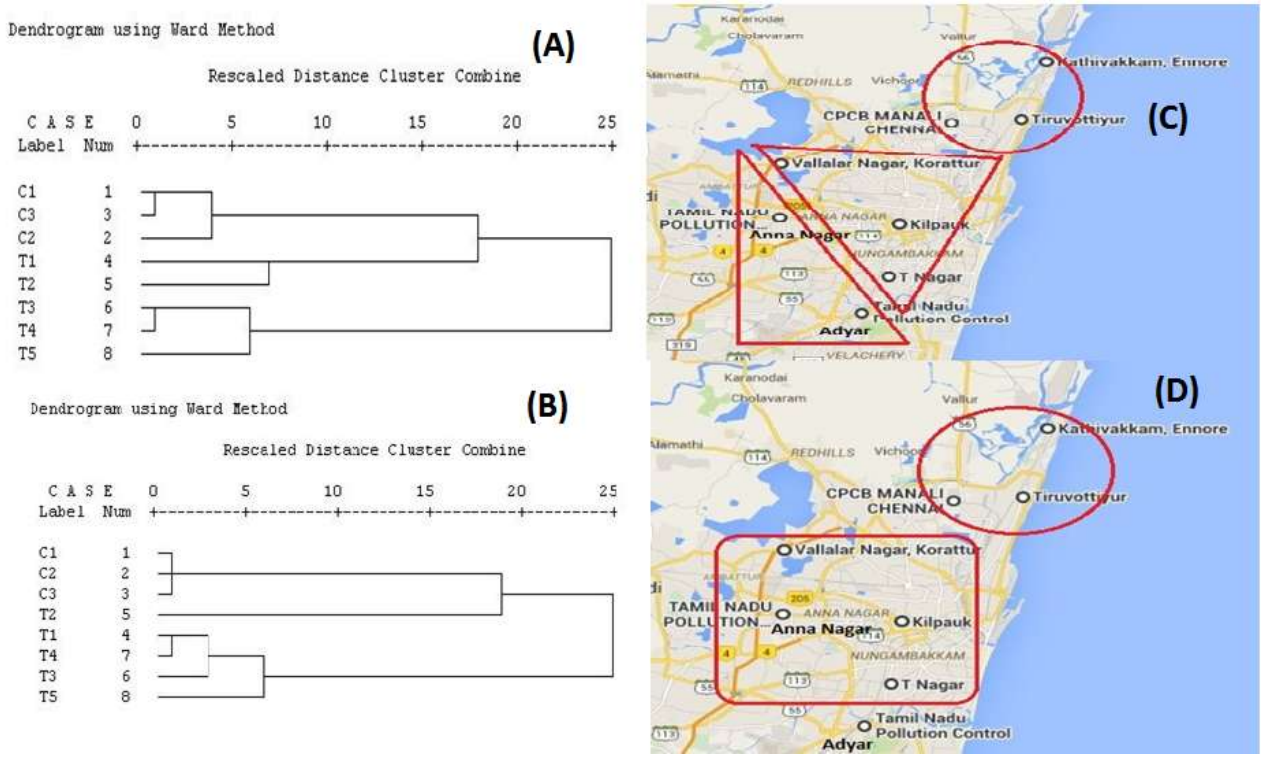

Figure 3: **(A)** Dendrogram of Different Cluster of Air monitoring station of the year 2007
(B) Spatial Location of Cluster formation of year 2007
(C) Dendrogram of Different Cluster of Air monitoring station of the year 2009
(D) Spatial Location of Cluster formation of year 2009

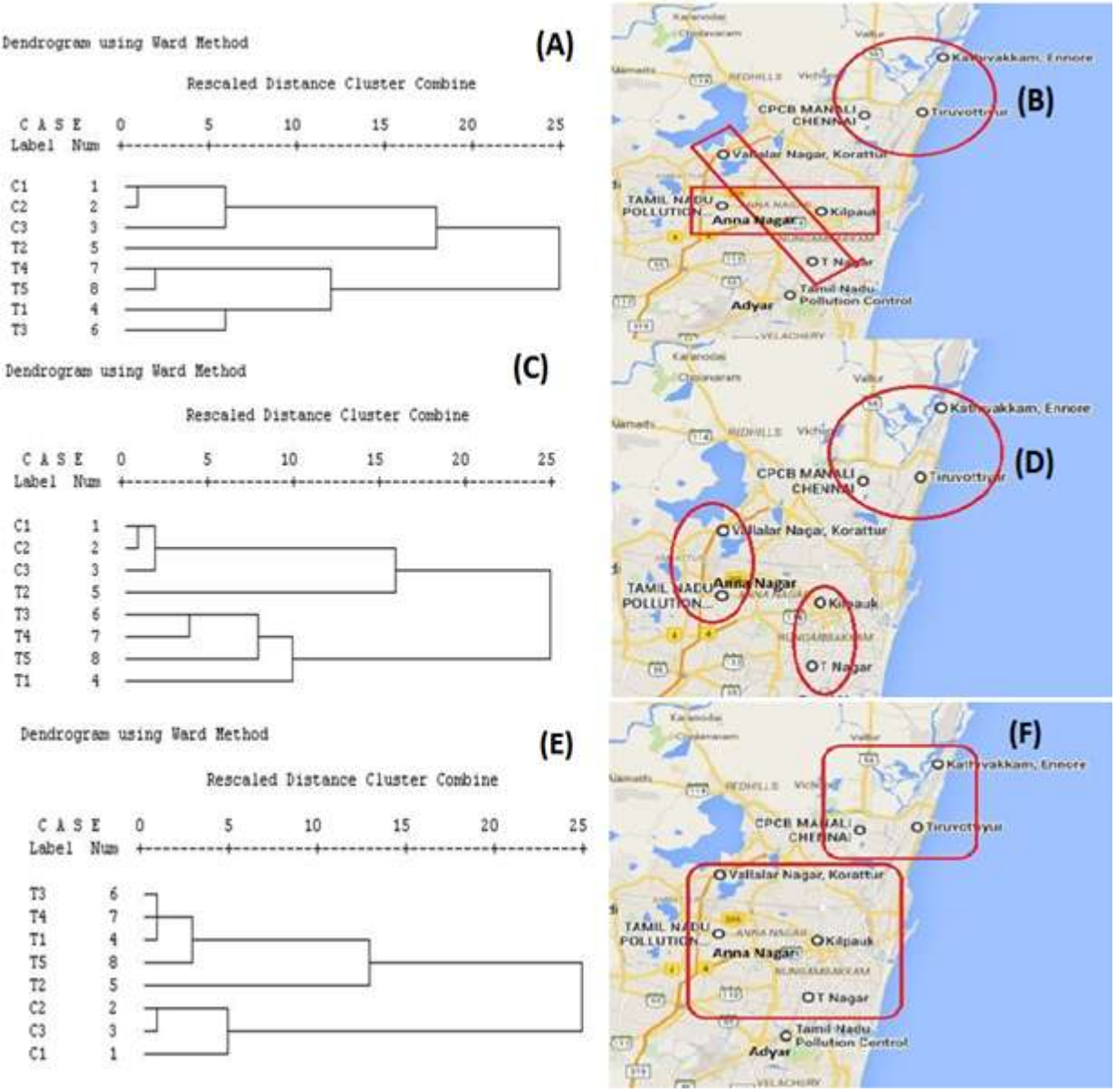

Figure 4: **(A)** Dendrogram of Different Cluster of Air monitoring station of the year 2010
(B) Spatial Location of Cluster formation of year 2010
(C) Dendrogram of Different Cluster of Air monitoring station of the year 2011
(D) Spatial Location of Cluster formation of year 2011
(E) Dendrogram of Different Cluster of Air monitoring station of the year 2012
(F) Spatial Location of Cluster formation of year 2012

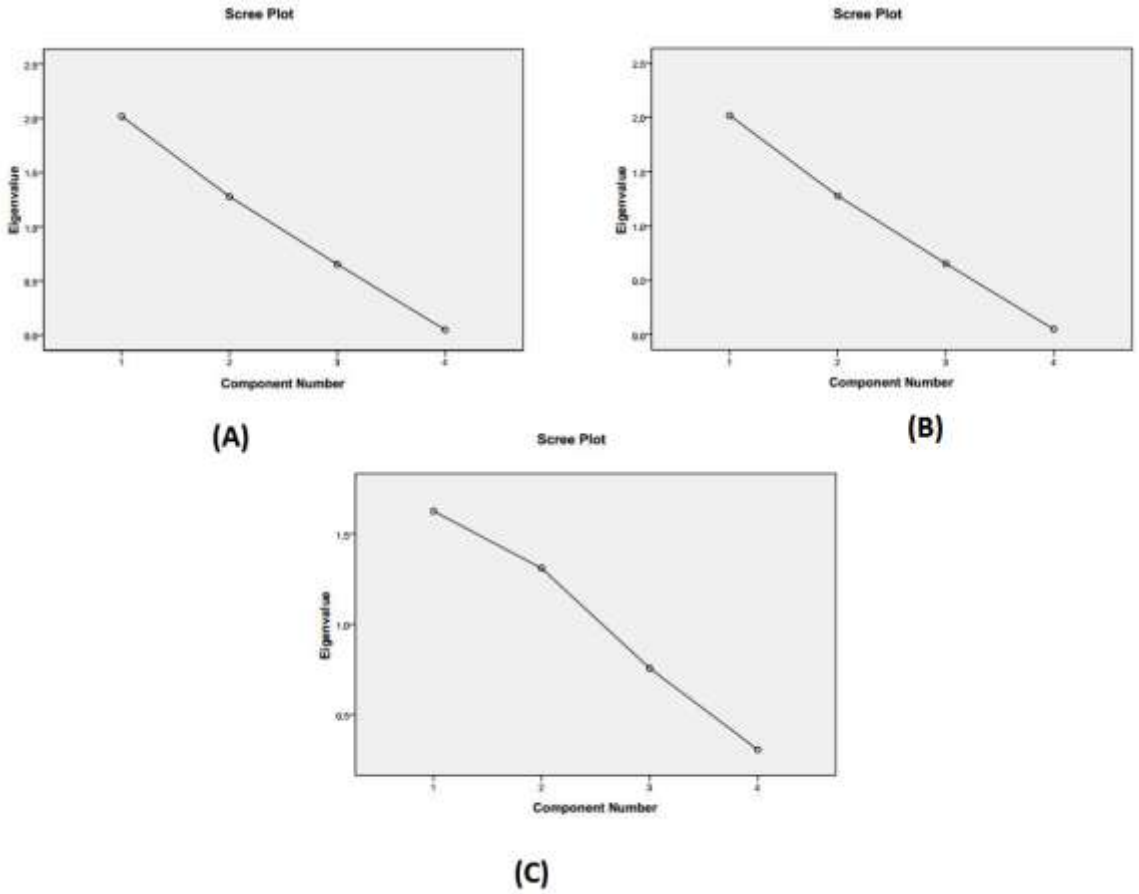

Figure 5 : **(A)** Scree plot for extraction of Eigen values of PCA loadings of Cluster -1
(B) Scree plot for extraction of Eigen values of PCA loadings of Cluster -2
(C) Scree plot for extraction of Eigen values of PCA loadings of Cluster -3