International Journal of Psychosocial Rehabilitation, Vol. 23, Issue 03, 2019
ISSN: 1475-7192

Air Pollution in Karnataka: A Causal Analysis

Dr. Chandrasekhar Subramanyam, Sidharth Khatua, Enna, Dr. SiddharthMisra*

Abstract---

Purpose: This research aims to highlight Major contributing factors towards air pollution in Karnataka and provide necessary recommendation to government and other stakeholders.

Design/Methodology/ Approach: The data was collected from data.gov.in of sample size 2400 observation and 12 variables. The framework that was used to do the analysis were problem identification, Data exploration, Model building, Solution implementation and monitoring. Training the sample data and validating the results on the test data were the approach taken furthermore programming tool R, Excel and SPSS was used to explore, recode data and build model.

Findings:From the analysis,it is observed that chemical components likeNO2and So2, cities like Devengree, Bangalore,Tumkur winter months like from October to February were the top contributors to air pollution in Karnataka.

Practical Implications:Using the analysis derived, Government stakeholders can identify the cities that are impacting air pollution, understand the reason for the cause and apply stringent norms to reduce the pollutants in the air. Also educating the younger generation by providing awareness regarding the ill effects of air pollutants on heath was also recommended.

Originality/ Values:Stepwise linear regression was used to estimate causal effect between PM10 and other variables, Decision Tree (CHAID) regressionused for machine learning to reduce the overall MAPE furthermore random forest regression tree was used to identify the top 10 contributing factors cause air pollution in Karnataka adding worthiness and novelty to this study.

Keywords--- *Karnataka;Air pollution; PM10; Nitrogen dioxide; Sulphur dioxide; Stepwise linear regression; Decision Tree (CHAID).*

I. INTRODUCTION

Air quality in Karnataka has been slowly deteriorating due to the rampant exposure of vehicle emission, unscientific and poor waste management. According to Karnataka state pollution control board out of nineteen habitations in the state ten cities and town, have (particulate matter less than 10 microns) PM10 way beyond permissible limit (Times of India, 2018).Air quality has become one of the major issues across various countries andcities around the globe. India, being a developing country is in the radar of world health organisation (WHO). WHO ranks India ranks 177 out of 180 in environmental quality index and is considered as one of the most polluted countries in the world and approx. out of 20 polluted cities, 13 exists in India itself. Major components that affect air

Dr. Chandrasekhar Subramanyam, SeniorProfessor, IFIM Business School, Bangalore. Email:chandrasekhar@ifimbschool.com
Sidharth Khatua, PGDM Student, IFIM Business School, Bangalore. Email:sidharth.khatua@ifimbschool.com
Enna, PGDM Student, IFIM Business School, Bangalore. Email:enna@ifimbschool.com
Dr. SiddharthMisra, Adjunct Professor, IFIM Business School, Bangalore. Email:siddharth.misra@ifim.edu.in*

Received: 12 Mar 2019 | Revised: 29 Mar 2019 | Accepted: 25 Apr 2019

pollution are power plants, fuelwood, biomass burning, natural disasters, vehicle emission and traffic congestion (Analytics Vidya, 2016). Air pollutants that affect the human health which possesses a great threat to the environment are ozone(O_3) carbon monoxide (CO) RSPM (both pm10 and pm2.5) SO_2 NO_2. Higher concentration of such pollutants pose great harm to life causing breathing issues, dizziness, headaches and in worst cause can also lead to cancer and heart attacks. (Pope et al. 1995).

Fuel wood is the primary reason for haze creation and smoke in India which affects health and causes various chronological diseases. When fuel woods are burned in higher volume it releases high levels of (particulate matter less than 10 microns) PM10, fog, smoke, NO_2, $SO_2$and other air pollutants. (Smith and Mehta 2003) Traffic congestion is a serious issue in Indian cities and states especially in Bangalore. Increase in the no of vehicles in the roads, traffic accidents and peak hours have been some major reason for traffic congestion. Due to traffic congestion traffic speeds reduces on an average and due to low speeds vehicle burn their fuel inefficiently which pollutes the earth and the environment.(Watson et al. 1996) Inhalation of PM 2.5 and PM 10 for a long time can lead to suffocation, difficulty in breathing, chronological disease like asthma, lung infection, Bronchitis and cardiovascular diseases like lung cancer and heart attack in the premature age.

It is estimated that there were about approx. 620,000 premature deaths occurred in India in 2010 due to air toxins and air related diseases. It is important to control air particulate matters like nitrogen dioxide and Sulphur dioxide because they have damaging effects on human health.(Analytics Vidya 2016).

Due to the high impact on health by PM10, Awareness of air quality in both emerging and emerged countries has increased over time.(Balakrishan et al. 2014).The central pollution control board (CPCB) highlighted in a study which connects air pollutants like sulphur dioxide, Nitrogen dioxide, PM10 is the cause for these diseases. (Gupta et al. 2006) Recently the central regulatory adopted stringent rules and regulation formany air toxins and pollutants to deal with such issues (Nagendra et al. 2007).Agencies like the European Environmental Agency, Ministry of Environment and Forests has already established their work on providing guidelines and policies to regulate air population all over the world. (K.R and Mehta 2003).

In the past, we have observed that various reliable models like 3D Eulerian chemistry transport models like Hemispheric model of Danish, Community multi-scale air quality have been used to analyse air quality studies for regulatory purposes (Kumar and Goyal, 2011). Though they are reliable they were prone to physical bias and detailed information about factors causing air pollution cannot be known or explained properly by these models. In order to overcome all these shortcomings leveraging Analytics and statistical models is a better option which will help in understanding who are the top drivers impacting the air pollutants. With the help of analytics, one can convert the raw data into meaningful insights and patterns which aid in decision making.

The main motive of this paper is to leverage the power of predictive analytics, open source tools to identify the contributing factor that affect air pollution in Karnataka state and to predict the air quality, understand the coefficients and provide a suitable recommendation to the government and other stakeholders.

II. LITERATURE REVIEW

To address the impact of air pollution in many countries, various discriminant and Gaussian models exists which can help evaluate the PM10 and PM 2.5(particulate matter less than 2 micron) dispersion in both rural and urban areas which help in predicting higher concentration of particulate matter dispersion also various studies were carried out based on the statistical models on different regions to identify the drivers of air pollution and what are the factors that strongly associate with PM10.

Thun et al.(1995), has done a study on the Air quality as a main cause of mortality in United States Adults they used models like cohort studies, Time series model, cross-sectional season studies to observe the relationship between mortality and air pollution. Schwartz (1991), conducted a study on particulate air pollution and daily mortality in Detroit where he observed that there is an increasing amount of daily mortality in Donora, Pennsylvania and United Kingdom London in1948 and 1952 respectively. This is due to an increase in the higher concentration of smog and sulphur dioxide concentration, especially in London. He also observed that there was a strong relationship between average smoke released and inhaled and number of death rates in London which lead to bronchitis. Li and Shue (2004), used data mining techniques to predict air quality in Taiwan using two-layer neural networks. Their process includes scrapping the data from the website which captured data from 71 monitoring stations in Taiwan later data pre-processing stage where the numerical data were normalised which in turn is feed in a neural network to obtain the prediction. (Samet et al. 2000). The study was not only on-air pollution but there were studies that were conducted on the factors that lead to water pollution. Beck (1987), analysed the assessment of water quality in London by creating a mathematical model to understand the root cause and later used it for prediction furthermore he analysed four problem areas in detail and there risk involved in model structure uncertainty: in parameters and coefficient values, validation of residual errors the framework of the experiment which will help in reducing the critical issues and uncertainties associated with the model. The type of data that they have collected was time-bound and hence, Time series analysis was used to do predictions in order to reduce the uncertainties and understand the risk and error involved. To tackle this issue Monte Carlo simulation was used to map the uncertainties.

Furthermore, the study was extended to Asian countries like Nepal and the state of India like Delhi and Lucknow to analyse the factors contributing to the air pollution. Simkhada et al. (2005), has assessed air quality in Nepal Kathmandu to analyse the particulate matters which contributed to the PM10, sulphur dioxide and oxides of nitrogen. They have also analysed and identified that microbial flora and fungi are present in very high concentration states like Teku Donovan and Shovabhagavti areas. These areas showed a very high level of air toxics.PM10 level of all the sites that they surveyed showed that air quality was very harmful and hazardous. Traffic congestion, Heavy traffic and few roads also contributed to the health hazards. Sites observed were Kathmandu: Here they observed that due to bowl topography the valley prevents the wind movements most of the particulate matters remain in the atmosphere. This is often absorbed during winter and is one of the major issues faced by the citizens of Kathmandu In the past ten years the PM10 emission has gone up by five times which is alarming and dangerous. In the recent study, it is estimated that vehicular emission in Kathmandu contributes to sixty-seven per cent of air particulate matters. Due to Bowl like topography the state is vulnerable to a higher concentration of air toxics. Teku Donovan: This site was observed with dumps of wastes which created a lot of nuisance and smell in the

environment. The site was filled with garbage and solid wastes. Kalimati Bridge: In this site, they have observed many vehicles on roads. Emission from these vehicles contributed to the large number of nitrogen oxides, sulphur dioxide in the atmosphere. Apart from these Municipal sewers contributed to stagnant water, Bad smell and uneasiness. Kankeswori: This site consisted of Slaughterhouse and was full of dead animal wastes and dumb. Pungent smell was observed from this site which contributed to air pollution. Shiva Bhagwati temple: This site was observed with very traffic and many vehicles just like kaneshwori site. New Bus Park: This sites road is busy for almost twenty-four hours and data was collected on the roof of the guardhouse. Balaju: This site is one the busiest places in Kathmandu. Vehicular emission was the top contributor in this area and in addition to that Balaiju industrial District was also a source.

They used Low volume air sampler (LVAS) Model of Pawan Tara an indoor air sampling equipment which helps in getting the sample ambient air which has at least 8 hours of battery to capture air quality sample. The equipment was installed at a height of 10 to 25 feet above from the ground level and samples of air were taken which contained particulate matters like PM10, sulphur dioxide and nitrogen dioxide. (Chirstopher et al. 2006).

In India, Cropper et al. (1991), explained that important reason that is impacting health of people are due to the inhalation of dangerous air toxins likeCarbon mono oxide (CO), Sulphur dioxide ($S0_2$) Nitrogen dioxide ($N0_2$) in higher concentration furthermore with this study he has observed that total no trauma deaths in Delhi, India is less compared to the effects observed in the United States. No such recommendation on policies and regulations were observed in their research paper. Kumar and Goyal (2009), used statistical models like autoregressive integrated moving average, (ARIMA), principal component regression and a combination of both the models to analyse the daily air quality index in Delhi. Based on the analysis the following conclusion was drawn. All seven days of the weeks were considered. They observed that weekdays and weekends were not statistically significant variables. Out of the three models, a combination of ARIMA and PCR performed better compared to the other models and is used to forecast the air quality of urban cities in India. Though the model performed well there was uncertainness associated with the model. Using Ensemble methods (Kumar& Goyal 2011), also used unsupervised learning model like Principle component analysis (PCA) and black box model like neural networks to analyse the air quality index of New Delhi since Delhi is one among most polluted cities in India. Using the correlation matrix, the association between the target and predictor variables are observed which a part of model validation. (Gupta and Rai 2013), have conducted a study in which Principal component analysis (PCA) was used to estimate the source of air pollution furthermoredecision Tree Ensemble methods were used to predict the air quality of Indian state Lucknow. They used data from the meteorological database of period 5 years to conduct this study. In their study using PCA algorithm, they have identified that fuel combustion and emission from vehicles contributed to air pollution. They have used various models to predict the air quality in rural and urban areas. Model of both classification and regression tree was used and even to improve the accuracy they have used ensemble methods like Bagging and Boosting tree to predict the air quality. Bagging is an ensemble method which is used to reduce the variance in the model here random sample is taken from the population and decision tree is fitted in each random bootstrapped sample (GuttikundaandJawahar 2012). Prediction of bagging is the average of all the decision tree outputs for the regression model and majority vote of outputs for the classification model. But decision tree of Bagging is highly

correlated to each other which will increase the variance of the output in order to tackle this Random forest algorithm is used where random bootstrapped samples and random variables are taken to minimize the correlation between the decision trees. Boosting algorithm is sequentially process where a decision tree is fitted to the entire data and next decision tree is built giving more weight is given to misclassified error. The process repeats sequentially until the error or loss function is totally minimised. It uses gradient descent to minimise the loss function or error.

Support vector machine was also used along with the ensemble methods. In evaluating the model's performance, they have used misclassification error, sensitivity, specificity, and confusion matrix as the metric for evaluation and for regression trees they have used actual vs. predicted scatter plot graph to evaluate the model (Smith and Huang 1995). Among the model's used Tree-based ensemble methods were highly accurate in predicting air quality which is thus used as a primary model to improve the management and decision making. The study was further extended to indoor air pollution.At least 3 billion people rely on Biomass that is dung, plant residues and fuel woods for the primary source of energy. Ezzati and Kammen (2002), describedmostly about the impacts of indoor air pollution and their source like combustion of Biomass, leads to morbidity, lower lifespan and chronic lung infection. In this study they try to find the relationship between combustion that are emitted from indoor air pollution and disease and how do they affect the health and identify the possible intervention and strategies to reduce the exposure.

There are studies where various machine learning techniques and ensemble methods were used to accurately predict the values but interpreting the coefficients were not observed in such studies. Athanasiadis et al. (2003), have applied Machine learning techniques to provide decision making support for air quality of the data their work describes model know as novel classifier FLNMAP for predicting ozone concentration level in the atmosphere Their novel classifier does well from other machine learning algorithms like decision tree algorithm or C4.5 algorithm and back propagation neural network. They have incorporated a framework which helps them analyse and build model and they are as follows they gather and monitor the air index quality data using data sensors which are provided and monitored by Data agents these are then given to data management agent who clean and make the data model ready further it goes to the prediction stage when data is predicted and coefficients are interpreted. Alarming agents are identified, and suitable recommendation is put forth. Similarly, (Niska et al 2004), said neural networks have been used to tackle nonlinear relationship and high dimensional sample space. But the selection of neural network architecture is very difficult and consumes a lot of time for model selection. In this paper, they have used techniques like parallel genetic algorithm for selecting inputs and designing multilayer perceptron architecture for predicting air quality of the data. Sample data was collected from the APPETISE1 (Modelling tools for improved smog management database). Using this paper, they used a Genetic algorithm to design multilayer perceptron which helped them come up with better prediction as absorbed in their Actual vs. predicted scatterplot graph as it showed very high correlation. They used pre-processed data from stations like Helsinki Vantaa airport and Helsinki _Isobaric.They observed the relationship between predictors and feature variables shows a nonlinear pattern which is why they have selected black box model like neural networks with a multilayer architecture to increase their accuracy and decrease the root mean square error (RMSE). A genetic algorithm was used as a grid search to obtain the right architecture which can help in further tasks.

After reviewing most of the papers that have been published in this field it was observed that none of the studies was on the Karnataka state where the vehicle growth rate and unscientific waste management have increased over the years especially in the year 2016- 2017.with the help of the study it can be estimated that the top 10 factors that affect PM10 in the Karnataka. Another gap that has been observed that there were fewer model diagnostics used to evaluate the models. It's a great initiative by of honourable Prime Minister Mr Narendra Singh Modi to introduce the digital India initiative – Open Government Data (OGD). It is the platform used by government departments to publish documents, services, and datasets to increase the transparency and open platforms for an innovative solution for the problems identified (data.gov.in 2018).

III. RESEARCH OBJECTIVES

- To identify the factors contributing towards Air pollution in Karnataka.
- To analyse the dependency of PM10 on other variables.
- To recommend the insights generated to government and other stakeholders.

IV. METHODOLOGY

Air Quality Data

Sample air quality data of nitrogen dioxide, sulphur dioxide, respirable suspended particulate matters, the location where it was monitored; city and type of location were collected for the period 2015 from Karnataka pollution control board which was listed on the website data.gov.in. The reason we have selected Karnataka because growth of two wheelers has increased by ten per cent from 2014 to 2016 and an increase of twelve percent in the year 2016 to 2017. Growth of four wheelers has increased by ten per cent in the year 2014 to 2015, a nine per cent increase in the year 2015 to 2016 and an eleven per cent increase in the year 2016 to 2017 (Times of India 2018).

Framework

To conduct this study, following framework is used to arrive at the solution and the tool used to do the analysis in R programming language (Data manipulation, Data Visualization and Model building), Excel for column recoding.

Problem identification includedAir pollution and the factors contributing to PM10 in Bangalore.

Solution Identification

In this stage, the business problem is converted into an analytical problem by identifying the target or response variable and feature variables. Target variable in this study is PM10 and the feature variables are a city, village (one column), type of location (either rural, urban, Industrial or other areas), the location of monitoring station (27 monitoring station in Bangalore), sampling date, Nitrogen dioxide, sulphur dioxide. The goal is to identify the relationship between PM10 and other feature variables. Since the dependent variable is numerical supervised learning model like linear and stepwise linear regression and Decision tree regression is used.

Received: 12 Mar 2019 | Revised: 29 Mar 2019 | Accepted: 25 Apr 2019

Data Exploration

In this Stage, nature and distribution of each column are investigated and identifying the quality of the data i.e. no of missing values in the data whether to impute them or delete them depending upon the portion of missing values. It is observed that No2, So2 variables were highly skewed towards right and hence log transformation was used to reduce the skewness in the data type. Location for monitoring station was recoded to 1, 2,3…27 as seen in the figure for easy interpretability in graphs since the texts were large.

Table I Recoded Value for monitoring stations

Location	Recode
K.R.Circle, Visvesvaraya Bldg., Mysore	1
Graphite India, White Field Road, Bangalore	2
AMCO Batteries, Mysore Road, Bangalore	3
KSPCB Bldg. Hebbal Ind. Area, Mysore	4
KHB Industrial Area, Near R.R. Founders, Yelahanka, Bangalore	5
Peenya Industrial Area, Bangalore	6
Victoria Hospital, Bangalore	7
Rani Chennamma Circle, Hubli	8
Lakkamanahalli Industrial Area, Dharwar	9
Yeshwanthpura, Bangalore	10
Narashima Raja Circle, Hassan	11
Government Hospital, Gulbarga	12
Karnataka SPCB Office Building, Belgaum	13
Stides Premises, Baikampady Industrial Area, Mangalore	14
Department of Environmental Science, Jnanabharathi Campus, Bangalore University	15
Vittal Medi health Care Pvt Ltd, Domlur	16
KSPCB Office Premises, Devengere	17
Mothi Talkies P.B. Road, Devengere	18
HPF Ltd., Water intake well, Ranebennur	19
KSPCB Office Premises, Mandya	20

Received: 12 Mar 2019 | Revised: 29 Mar 2019 | Accepted: 25 Apr 2019

KSPCB Office Premises, Raichur	21
KSPCB Office Premises, Bijapur	22
KSPCB Office Premises, Chitradurga	23
The VISL, Oxygen Plant, Shimoga	24
KSPCB Office Premises, Kolar	25
KSPCB Office Premises, Bidar	26
Tumkur University, Tumkur	27

Notes: Data has been recorded from 27 stations in Karnataka from data.gov.in.

Model Creation and Model validation

After cleaning and transforming the data it is then split into Training and Test set to build and validate the model. The splitting ratio is 80:20 i.e. 80% of the entire data is used to train a model and 20% of the remaining data is used to validate the model. Model Used were supervised parametric multiple linear regression and Decision Tree Regression (CHAID).

Stepwise Linear Regression

Stepwise Linear regression is a technique of building regression models by selecting the variables that better explain the variation In the response variable using some pre-specified criterion like Adjusted R square, Bayesian information criterion(BIC), Akaike Information Criterion(AIC),. The model is trained either by adding new variables from the model or subtracting the variables from the existing model. Lower the AIC better is the model.

Decision Tree (CHAID):Chi square of Automatic interaction detector (CHAID) is a type of decision tree which uses chi square test of association method to split the entire dataset into sub groups. The dataset splits into two or three child nodes and the process continues till there is no statistical difference between the groups and the dependent variable (Ture et al. 2009).

Model Validation

In order to test the model and coefficients, we use the following metrics.

R square: It ranges from 0 to 1. Closer to 1 indicates better fit.

Adjusted R square: Increase of Adjusted R square determines if the new variables added is statistically significant or not. It ranges from 0 to 1 closer to 1 indicates better fit.

For the above metric to be valid, the basic assumption of the residuals or the error is to be met. The assumptions for the residuals are: There should be a linear relationship between x and y variables.Residuals should be independent of each other and there should not be any autocorrelation. Residuals should be normally distributed with mean Zero. The error should have constant variance or Homoskedacious.Mean absolute percentage error (MAPE) were calculated on the test set to analyse the performance of the model.

Solution monitoring and controlling

The model builds on the training set is now used on the validation or the test set to check if it generalises well on the unseen dataset.

V. DATA ANALYSIS AND INTERPRETATION

Analysis was done on the sample data collected and the following results were observed.

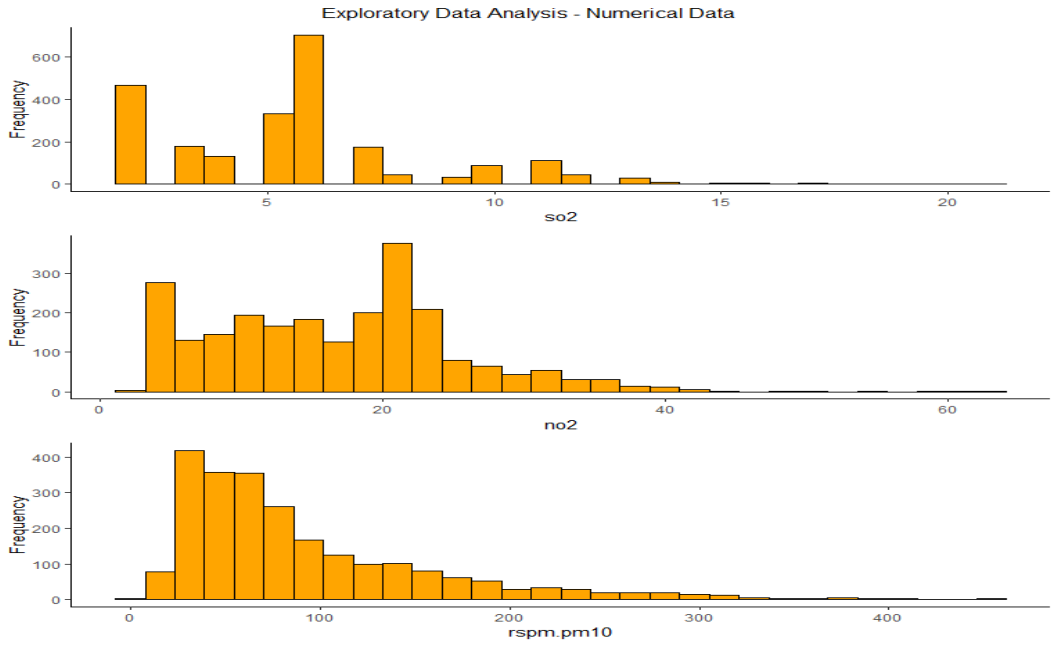

Fig 1Exploratory data analysis using numerical data

From the above graph we have observed that SO2, NO2, PM10 are skewed towards right indicating the positively skewed distribution and hence log transformation is taken for the following variables. Generally, linear regression work well with data which are normally distributed.

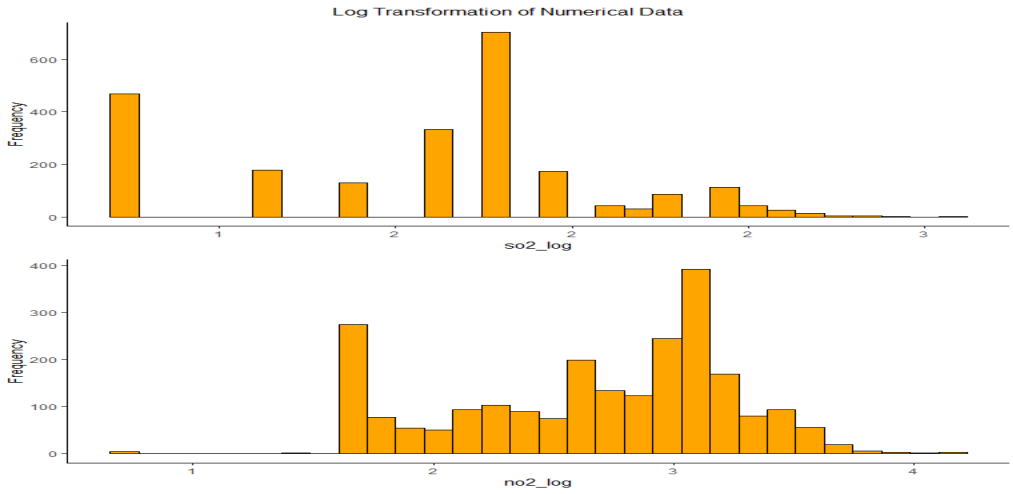

Fig 2Log transformations of numerical data

From the above graph it is observed that after taking the log transformation the observation is approx. normally distributed. It is observed that PM10 is positively correlated with SO2 and No2.The correlation between PM10 and S02 observed were .3248 and between PM10 and NO2 is .064.

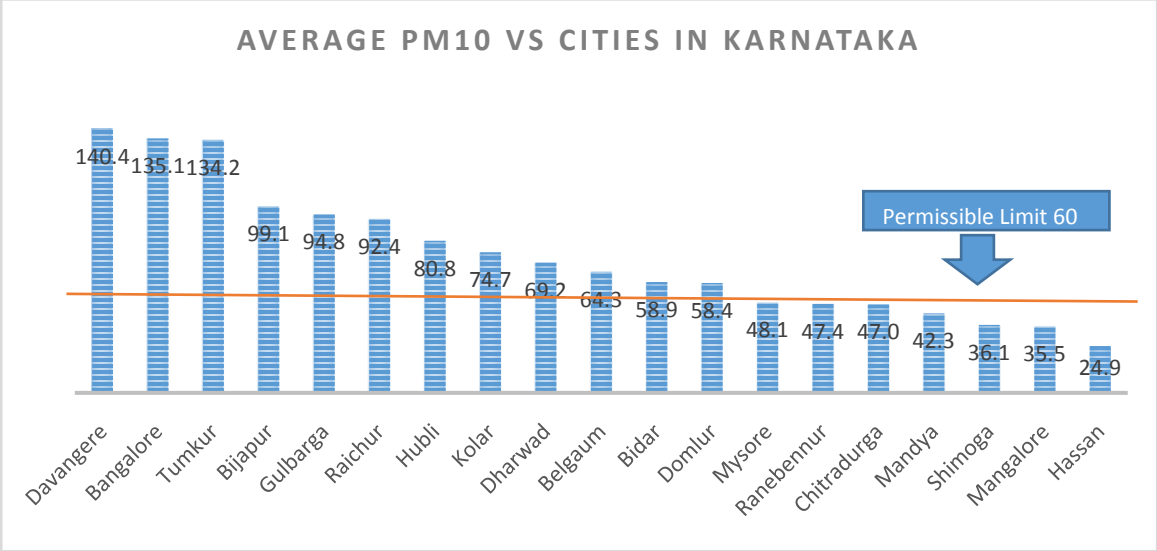

Fig 3 Average PM10 versus cities in Karnataka

From the above graph 3, it is observed that On an Average PM10 is highest in Davangere followed by Bangalore,Tumku and Gulbarga. Permissible limit in Karnataka is 60 microgram per metre cube and it is observed that these 3 cities way beyond the permissible limit. Out of nineteen cities in Karnataka it is observed that 10 cities are beyond permissible limit above 60.On an average PM10 is highest in a sensitive area. Out of 19 cities in Karnataka 11 cities are above the permissible limit.

Seasonal trends

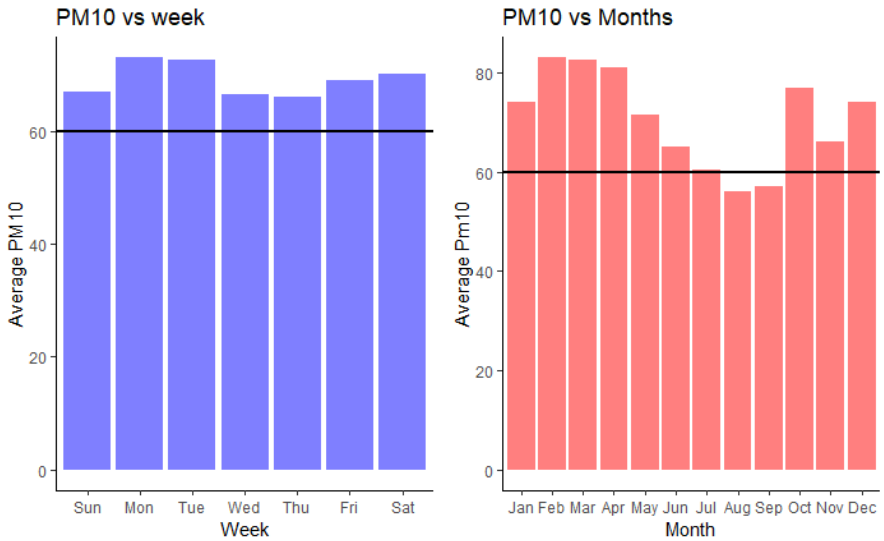

Fig 4 Month and days versus PM10

Received: 12 Mar 2019 | Revised: 29 Mar 2019 | Accepted: 25 Apr 2019

Frome the above graph it is observed that Months have a seasonal trend on an average the Pm10 were high in winter especially October, November, December, January, Februaryand lowest in the month August. The days graph shows sinusoidal pattern with higher PM10 on Monday.

Model

Linear regression was used initially using the 80% of data and then bi-directional stepwise linear regression was used to create a parsimonious model.

Table II Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.811[a]	.657	.651	40.32343229944

Table III ANOVA

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	7190904.280	45	159797.873	98.278	.000[b]
Residual	3749508.018	2306	1625.979		
Total	10940412.298	2351			

Table IV Regression Analysis

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-1.581	9.519		-.166	.868
location_recode.2	145.279	6.925	.467	20.979	.000

Received: 12 Mar 2019 | Revised: 29 Mar 2019 | Accepted: 25 Apr 2019

location_recode.3	104.092	6.889	.322	15.111	.000
location_recode.4	-13.437	5.614	-.041	-2.394	.017
location_recode.5	99.043	7.104	.299	13.941	.000
location_recode.6	100.805	6.845	.320	14.727	.000
location_recode.7	90.255	6.908	.296	13.065	.000
location_recode.8	44.623	6.718	.135	6.643	.000
location_recode.9	32.630	6.722	.099	4.854	.000
location_recode.10	94.972	6.892	.312	13.781	.000
location_recode.11	-8.698	6.546	-.027	-1.329	.184
location_recode.12	99.519	9.541	.186	10.431	.000
location_recode.13	54.883	8.621	.136	6.366	.000
location_recode.14	12.268	6.578	.036	1.865	.062
location_recode.15	-4.355	9.308	-.006	-.468	.640
location_recode.16	17.020	9.983	.026	1.705	.088
location_recode.17	49.761	8.024	.149	6.202	.000
location_recode.18	201.927	6.679	.609	30.232	.000
location_recode.19	38.170	7.659	.115	4.983	.000

Received: 12 Mar 2019 | Revised: 29 Mar 2019 | Accepted: 25 Apr 2019

location_recode.20	-9.123	5.695	-.027	-1.602	.109
location_recode.21	67.445	7.772	.156	8.678	.000
location_recode.22	85.932	8.545	.211	10.057	.000
location_recode.23	42.660	8.064	.129	5.290	.000
location_recode.24	29.969	7.933	.090	3.778	.000
location_recode.25	22.936	11.432	.028	2.006	.045
location_recode.26	14.193	10.045	.021	1.413	.158
location_recode.27	98.571	7.331	.270	13.445	.000
week.Mon	3.096	3.109	.016	.996	.319
week.Sat	-.551	3.145	-.003	-.175	.861
week.Sun	-2.358	3.161	-.012	-.746	.456
week.Thu	4.729	3.197	.024	1.479	.139
week.Tue	-1.039	3.245	-.005	-.320	.749
week.Wed	-1.418	3.219	-.007	-.441	.660
months.Aug	-29.212	4.478	-.106	-6.523	.000
months.Dec	-1.227	4.491	-.005	-.273	.785
months.Feb	12.106	4.252	.051	2.847	.004
months.Jan	10.581	4.178	.046	2.533	.011

Received: 12 Mar 2019 | Revised: 29 Mar 2019 | Accepted: 25 Apr 2019

months.Jul	-26.174	4.156	-.108	-6.298	.000
months.Jun	-25.313	4.122	-.104	-6.141	.000
months.Mar	-4.131	4.159	-.018	-.993	.321
months.May	-17.138	4.157	-.070	-4.123	.000
months.Nov	-17.542	4.373	-.071	-4.011	.000
months.Oct	-9.227	4.368	-.036	-2.112	.035
months.Sep	-28.270	4.199	-.113	-6.732	.000
so2	3.596	.677	.151	5.310	.000
no2	1.274	.168	.160	7.600	.000

a. Dependent Variable: pm10

Notes: Causal Analysis of PM10 with independent variables.

Model Diagnostic

In order to rely on the above metrics, error diagnostic is performed because linear regression follows the following assumption and they are:

- Errors should be normally distributed
- Residuals should have constant variance
- Residuals should be independent and random scattered.

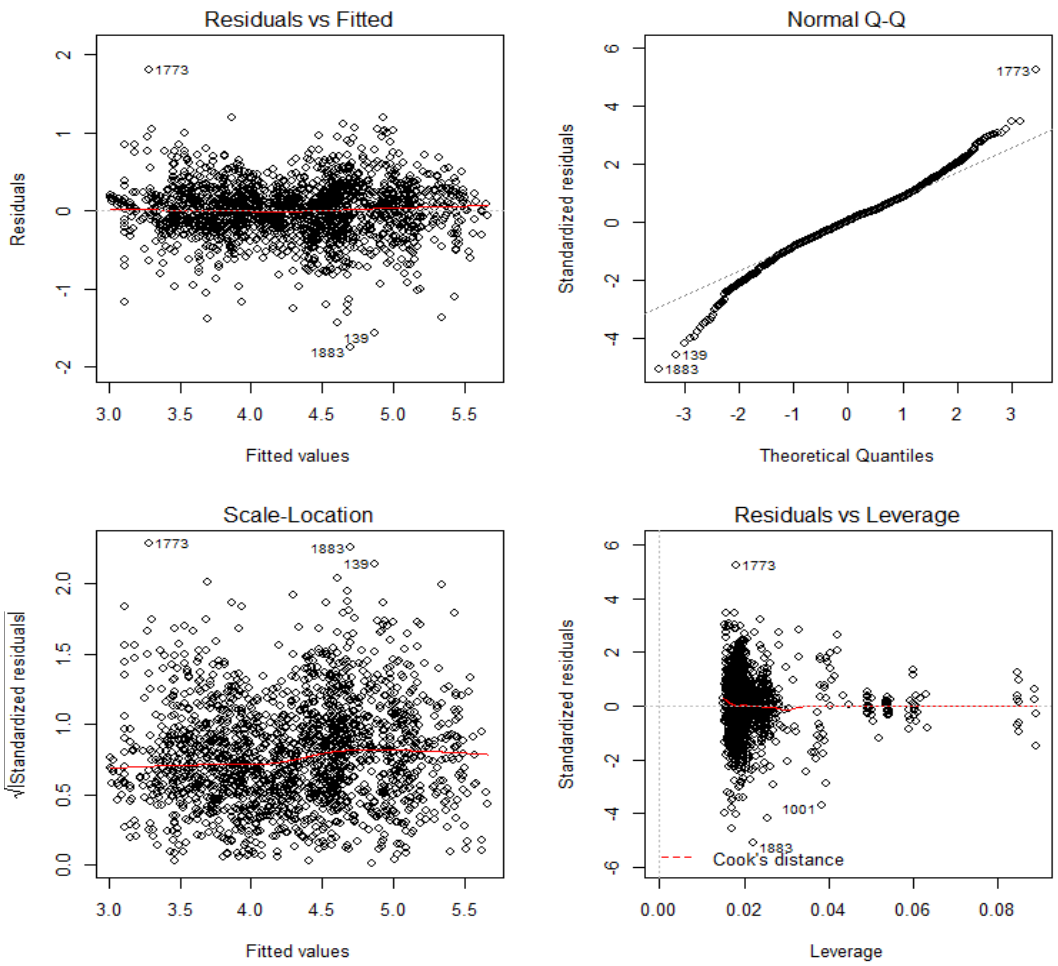

Fig 5 Model diagnostics

The first graph shows that residuals are randomly scattered with a constant variance which validates the 2nd and 3rd assumption.The second graph is approx. normal indicating that residuals are normally distributed.

Similarly, Decision Tree CHAID was run on the training data and tree was produced.

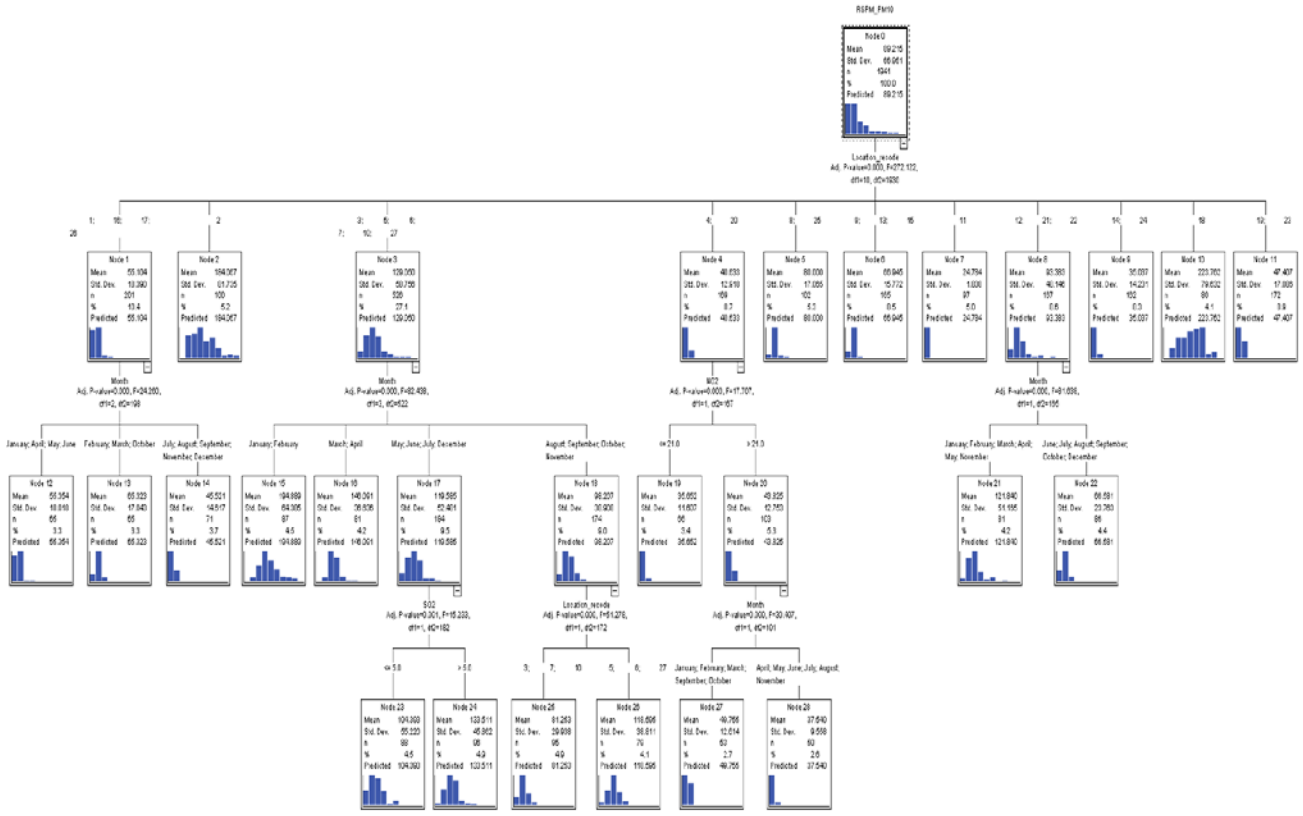

Fig 6 CHAID Diagram of Train Data Set

Model Validation

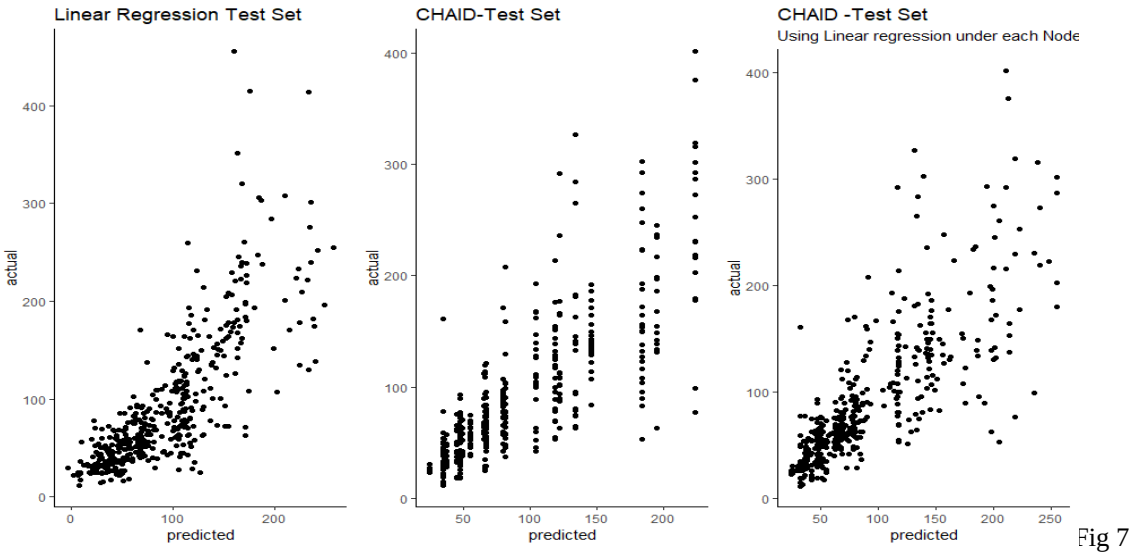

Fig 7

Actual versus Predicted

Model were then used to predict on the validation set and the following graph and MAPE were calculated. From the above graph, it is observed that actual values and Predicted value are positively correlated indicating that the

model has fairly done a good amount of prediction both by CHAID and linear regression. The Mean absolute percentage error (MAPE) obtained for linear regression is 40.6% and for Decision Tree (CHAID) is 32% furthermore in order to decrease the overall error linear regression were performed on each node of CHAID Model. After performing the analysis it's observed that the MAPE of CHAID model is 29%.

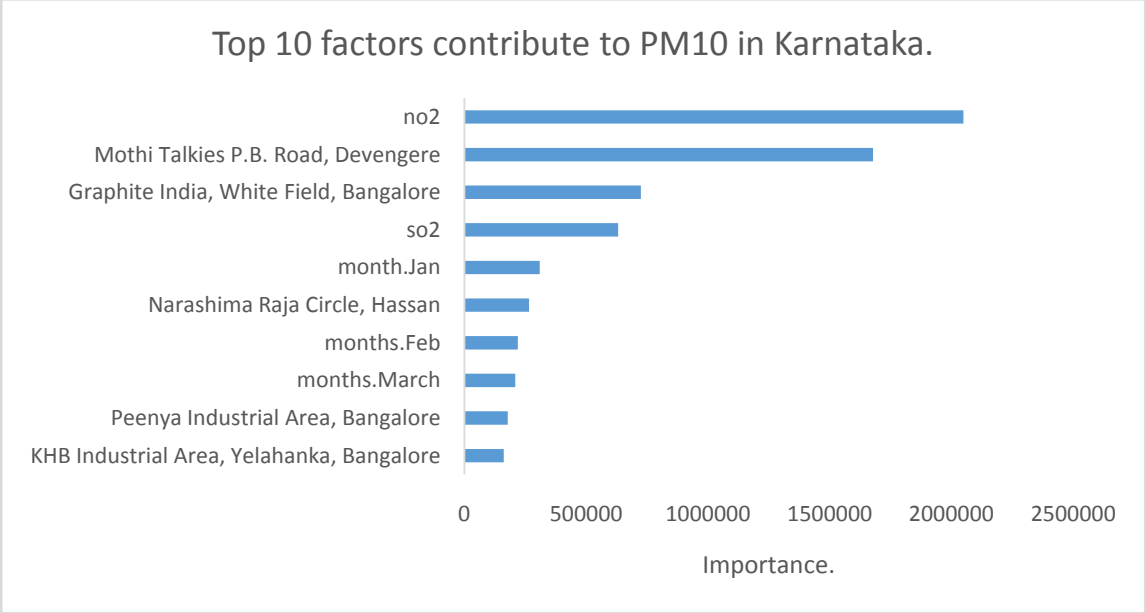

Fig 8 Top 10 Factors contributing to PM10

From the above graph 8, Factors contributing to PM10 in Karnataka is arranged in the descending order and top 10 variables or factors are highlighted.

VI. DISCUSSION

This study highlights the factors contributing to PM10 in Karnataka. In the sample data it observed that there is positive correlation between PM10 to Sulphur dioxide. As the sulphur di oxide increase in the atmosphere pm10 also increases. About 95% of SO_2in the atmosphere is due to human sources. Industrial process and material that contain sulphur for example production of electricity from coal, oil and burning of fossil fuel which contain sulphur are some of the major sources. (Nagendra et al. 2007) Sulphur di oxide is also released due to fuel combustion and vehicular emission.SO_2affects human health when it is inhaled. It blocks the airways irritates the nose and throat which causes sneezing cough breathe shortness and suffocation around the chest. Person can quickly feel the affect within in 10 to 15 minutes after inhaling. This can lead to worse conditions like asthma, similar correlation is observed with PM10 and NO_2just like $SO_2NO_2$is also an most important air pollutants that leads to photochemical smog some sources of $NO_2$is burning of fossilfuels, manufacturingindustrialareas, petrol. Inhalation of NO_2 for a long period can cause inflammation in the lungs,bronchitis, reduces immunity to lung infections. In above observed seasonal trend graph it's observed that in winter i.e. (October to February) there are high level of PM10. The reason for increased PM10 in the month of October and November is because of festive season in India especially Durga puja, Christmas day, New Year there is a large number of vehicles purchase and vehicular emission, bursting of

crackers which increases the NO_2 and SO_2 in the atmosphere furthermore household heating and vehicular emission are really high during the winters (Airlief 2017). In week graphit is observed that on an average PM10 is high on Mondays this is due the reason most of IT organisation in Karnataka operate in the weekdays especially on Monday large number vehicles are observed on the road which leads to vehicular emission and traffic congestion .In fig 3 graph it is observed that city that have high level of PM10 are Tumkur, Banglore, Davangree and Gulbarga. This is due to increase usage of plastics which remains uncontrolled and unchecked. Violation of industrial norms set by the government and fast pace in industrial activity cause high PM10 in the respected areas. Expansion and urbanizationareanother reason for the cause of air pollution. The rate of expansion in these cities are increasing in an alarming rate, Addition of new vehicles especially in Bangalore, vehicular density, increased emission, increase in small scale industries, increase usage of Diesel generators. Increase In the disposable income of a person is also an indirect reason for air pollution. Deforestation increase usage of air conditioning and emission of ozone from such equipment can also be another factor. In Bangalore in the year 2014 – 2015 the levels of pm10 has increased by 23% and levels has Nox gradually increases over time. The important areas in Bangalore that impact air pollution is Graphite India Whitefield Victoria hospital Bangalore. From the last graph the top 10 factors that contribute to PM10 are Nitrogen Dioxide, Nashima raja circle Hassan, SO_2, Mothi Talkies P.B. Road, Devengere, Graphite India, White Field Road, Bangalore,Peenaya Industrial Area Bangalore,KHB Industrial Area Yelahanka,Bangalore,February, January and march months. Most of the areas that were on the top factors are industrial areas and festive months.

VII. CONCLUSION

Air pollution has always been a lame stream issue which draws high attention from the government of every nation. It is observed that it's amongst the leading cause of premature deaths in the world. Similarly, Karnataka is encountering with various levels of air polluted particles and it's observed that nitrogen and sulphur di oxide are major contributors towards it. Time constraint is one of the limitations that werefaced during the research. Enough data was not available on daily basis which restricted this study to do cellular level analysis on PM10. Other limitation is that the Data fails to capture PM2.5 which is one of the major factors towards air pollution. Due to its small size the particles tend to stay longer in the atmosphere because of which it is difficult to capture the data regarding this particle. By the advancement in technology, finest particle like PM2.5 can be captured for further studies and understand the factors that contribute to PM2.5 furthermore with advancement in field of analytics sophisticated models like random forest, deep learning, extreme gradient boosting algorithms can be performed to increase the accuracy and decease overall error. Similar analysis can be performed on all the states of India to understand the top cities that contribute to air pollution. TheKarnataka state pollution control board (KSPCB) has taken numerous initiatives towards the quality of air by regulating the permissible limit to 60 ug/m^3. The recommendation given are as follows enhancing fuel quality in vehicles by automobile industries, Banning of commercial vehicles which have a life span of 10 years, Avoid usage of diesel generators by no frequent power cuts, Opting Compressed natural gas (CNG) over petrol and diesel, frequent increase in vehicle maintenance and improvement, Awareness activities should be promoted in every educational institutes by introducing full credit course in the curriculum.

ACKNOWLEDGEMENT

The satiation and euphoria that accompany the successful completion of this research would be incomplete without the mention of the people who made it possible. We thank the research team of Accendere Knowledge Management Services, CL Educate Ltd. for their unflinching guidance, continuous encouragement and support to successfully complete this research work.

REFERENCES

[1] Badami,M.GTransport and urban air pollution in India. *Environmental Management* (2005).
[2] Begum, A., HariKrishna, S., & Khan, I. Chemical composition of rainwater in South Bangalore, Karnataka. *Rasayan J. Chem* (2008).
[3] Kumar, A., & Goyal, P, Forecasting of air quality in Delhi using principal component regression technique. *Atmospheric Pollution Research* (2011)
[4] Kumar, A., & Goyal, PForecasting of air quality index in Delhi using neural network based on principal component analysis. *Pure and Applied Geophysics* (2013).
[5] Li, S. T., & Shue, L. Y. Data mining to aid policy making in air pollution management. *Expert Systems with Applications* (2004).
[6] Pope, C. A., Thun, M. J., Namboodiri, M. M., Dockery, D. W., Evans, J. S., Speizer, F. E., & Heath, C. W. Particulate air pollution as a predictor of mortality in a prospective study of US adults. *American journal of respiratory and critical care medicine*(1995).
[7] Schwartz, JParticulate air pollution and daily mortality in Detroit. *Environmental Research*, (1991).
[8] Schwartz, J., & Marcus, A.Mortality and air pollution j london: a time series analysis. *American journal of epidemiology*, (1990).
[9] Simkhada, K., Murthy, K. V., & Khanal, S. NAssessment of ambient air quality in Bishnumati corridor, Kathmandu metropolis. *International Journal of Environmental Science &Technology*, (2005).
[10] Singh, K. P., Gupta, S., & Rai, P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, (2013).
[11] Balakrishnan, K., Cohen, A., & Smith, K. R. Addressing the burden of disease attributable to air pollution in India: the need to integrate across household and ambient air pollution exposures. *Environmental health perspectives*, (2014).
[12] Smith, K. R., & Mehta, S. The burden of disease from indoor air pollution in developing countries: comparison of estimates. *International journal of hygiene and environmental health*, (2003).
[13] Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., ... & Speizer, F. E. An association between air pollution and mortality in six US cities. *New England journal of medicine*, . (1993).
[14] Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I., & Zeger, S. L. Fine particulate air pollution and mortality in 20 US cities, 1987–1994. *New England journal of medicine*, (2000).
[15] Guttikunda, S. K., & Jawahar, P. Application of SIM-air modelling tools to assess air quality in Indian cities. *Atmospheric Environment*, (2012).
[16] Gupta, P., Christopher, S. A., Wang, J., Gehrig, R., Lee, Y. C., & Kumar, NSatellite remote sensing of particulate matter and air quality assessment over global cities. *Atmospheric Environment*. (2006).
[17] Mohan, D. Traffic safety and health in Indian cities. *Journal of Transport and Infrastructure*, (2002).
[18] Nagendra, S. S., Venugopal, K., & Jones, S. L. Assessment of air quality near traffic intersections in Bangalore city using air quality indices. *Transportation Research Part D: Transport and Environment*, (2007).
[19] Chow, J. C., Watson, J. G., Lu, Z., Lowenthal, D. H., Frazier, C. A., Solomon, P. A., ... & Magliano, K.Descriptive analysis of PM2. 5 and PM10 at regionally representative locations during SJVAQS/AUSPEX. *Atmospheric Environment*, (1996).
[20] Smith, V. K., & Huang, J. C. Can markets value air quality? A meta-analysis of hedonic property value models. *Journal of political economy*, (1995).
[21] Ture, M., Tokatli, F., & Kurt, I. Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications*, *36*(2), 2017-2026 (2009).
[22] Complete study of factors contributing to Air pollution

[23] https://www.analyticsvidhya.com/blog/2016/10/complete-study-of-factors-contributing-to-air-pollution/ Oct 29, 2016, accessed Nov 9, 2018.
[24] Air pollution levels in Karnataka's 11 cities and towns cross limits
[25] https://timesofindia.indiatimes.com/city/bengaluru/air-pollution-levels-in-karnatakas-11-cities-and-towns-cross-limits/articleshow/63036922.cms Feb 23 2018, accessed Oct 28 2018.

Abbreviations

AIC	Akaike Information Criterion
ARIMA	Autoregressive integrated moving average
BIC	Bayesian information criterion
CO	Carbon monoxide
CHAID	Chi square of Automatic interaction detector
CNG	Compressed natural gas
CPCB	Central pollution control board
FLNMAP	Classifier
IT	Information Technology
KSPCB	Karnataka state pollution control board
LVAS	Low volume air sampler
MAPE	Mean absolute percentage error
NO2	Nitrogen dioxide
OGD	Open Government Data
O3	Ozone
PM	Particulate Matter
PCA	Principle component analysis
PCR	Principle component regression
RSPM	Respirable suspended particulate matter
RMSE	Root mean square Error
SPSS	Statistical Package for the Social Sciences
So2	Sulphur dioxide
WHO	World Health Organization

Received: 12 Mar 2019 | Revised: 29 Mar 2019 | Accepted: 25 Apr 2019