International Journal of Mathematics and Statistics Invention (IJMSI)
E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759
www.ijmsi.org Volume 11 Issue 4 || July - August, 2023 || PP 17-27

Statistical Innovations and Strategies for Air Pollution Challenges

[1]Preethi Jayarama Shetty, [2]Satyanarayana

1. *Corresponding author: Department of statistics, Mangalore University, Karnataka*
Email: preethishetty997@gmail.com
2. *Department of PG studies and research in statistics, Mangalore University, Karnataka*
Email: sathya1301@gmail.com

Abstract: *Air pollution is continuously causing major damage to the Earth's environment. It will also severely affect ecosystems and disturb ecological balance. Air Pollution is a serious problem and most people are still unaware of Air pollution causes and its effects. These challenges are not only causing trouble to the human beings but also to animals and plants. Hence appropriate measures must be taken to tackle this hazard. Air quality Index (AQI) is the tool for identifying the severity of the air pollution. In this paper innovative applications of statistical procedure are suggested for the analysis and modelling of Air pollution of Delhi. Main focus of the study is to identify the major risk factors and best predictive model for estimating the air pollution. This study shows that due to the natural variation in the phenomena it is better to estimate the model parameters in the presence of both multicollinearity and autocorrelation in order to get the precise estimators. Best predictive model is selected on the basis RMSE value and it is used to predict the AQI level. Traditional Time series models and Artificial Neural Network are used to forecast the major risk factors for the air pollution. Based on the study policy measures were given which can be used by the local administrators to curve the AQI level.*

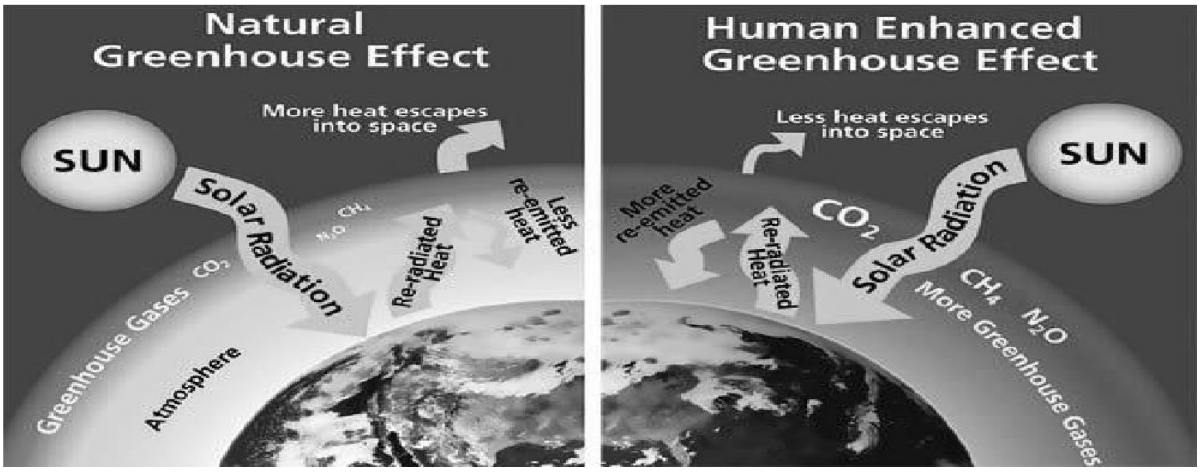
Keywords: *Air Quality, Regression model, autocorrelation, Multicollinearity, Generalized Ridge Type Estimator.*

Date of Submission: 03-08-2023 Date of acceptance: 15-08-2023

I. Introduction

Global warming is being emerged as one of the most important environmental issues that threatens both developed and developing countries. Global warming can be defined as rise in the average temperature on the earth's surface. It is observed that warming of climate system is caused by the increasing concentration of greenhouse gases such as carbon monoxide, Ammonia, nitrogen dioxide, Nitric Oxide, PM_{10}, $PM_{2.5,}$ etc. This is the result of human activities such as emission of gases from industries and transport vehicles, Urbanisation and constructions of buildings in large number, deforestation and burning of fossil fuels for electricity. This leads to the production of gases like Methane, Carbon dioxide and Nitrous oxides. As a result, the temperature increases and disasters like flood and drought seem to occur more frequently.

The loss of trees and other vegetation can cause severe damage to the environment. This causes depletion of oxygen in atmosphere which in turn increases the unwanted components and damages the quality of air. The presence of the substances which has harmful or poisonous effects will cause air pollution. This can be suspected to have harmful effects on human health and environment. When nitrogen oxide and sulphur dioxide particles in the air, mix with water and oxygen in the atmosphere results in acid rain. Hence there can be observable imbalance in the climate system too.



There are several cities facing the problem of pollution and global warming, among which India's capital, Delhi is one. Delhi is listed under the top most polluted cities of India, having poor Air Quality Index. One of the causes to this can be the emission of Carbon monoxide from the automobiles that is increasing day by day, as a result of which, not only human health but also the monuments are damaged. The best example can be The Taj Mahal, which is slowly decaying and discolouring the marble and turning it to yellowish-brown colour. There may be several influencing factors like: heavy traffic, growing population, increasing industries etc. The Ambient Air Pollution (AAP) report of 2014 revealed that Delhi had $PM_{2.5}$ pollution levels, which is second highest in the world. This is supposed to be a very serious matter that can cause respiratory diseases and other health problems like lung cancer. The level of Nitrogen dioxide has also been increasing. The current serious issue of global warming and air pollution trending on the media must be taken as a motivation by every citizen in order to analyse the root cause and overcome the situation.

II. Literature Review:

As the countries start developing, the impact of pollution seems to be more severe. As a result of potential impact of climate change and health risks, efforts are taken in reducing pollution. M. Venkatramanan and Smitha (2011) talks about 'Causes and effects of global warming'. It is said that, the reduction in demand for fossil fuels can reduce global warming by using energy more wisely. In a study, "Global warming: Its cause and effect in context to India", by S.K Bhartiya and B.K Choudhary (2012) reveals some deadliest effect of global warming. Umair Shahzad (2015) explained the Global Warming: Causes, Effects and Solutions. He mentioned that BURNING of Fossil fuels and Deforestation are major cause to global warming and it's causes and hazards. Also, presents some solutions to solve this hot issue. Agriculture being the backbone of Indian economy, Ruchita Shah &Rohit Srivastava (2017) study can be used as guide to understood the effect of climate change on Indian Agriculture.

III. Materials and Methods:

3.1 Parametric Regression:

Consider the multiple linear regression model $Y = X\beta + \varepsilon$ (1)

where Y is an response variable, X is a regressor, β is regression coefficient and Ɛ is error term representing random variables which are assumed to normal with finite mean and constant variance. Under the assumption that Rank(X)=k, X and ε are independently distributed and the errors are non autocorrelated, the Ordinary Least Square (OLS) estimator of β is $\hat{\beta} = (X'X)^{-1}X'Y$ with the covariance matrix of $\hat{\beta}$ is obtained as $cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$. In case of multicollinearity OLS estimator exist and unbiased, but has very large variance.

Ridge Regression: The violation of the basic ideal condition independent regressors leads to multicollinearity. If $Rank(X) < k$ then columns of X cab written as exact linear combination of the remaining regressor and hence OLSE estimator does not exist. This situation is called as perfect multicollinearity and it is very rare in practice. In other hand some regressors are nearly related to the remaining regressors. This is called near multicollinearity. In case of near multicollinearity, Rank(X)=k and hence OLS estimator exist. With strongly interrelated pairs of regressors, variance of OLS estimator becomes large which leads to forcefully accept the hypotheses. The ridge estimator (ordinary ridge estimator) of β is

$$\hat{\beta}_{RR} = (X'X + kI)^{-1}X'Y \quad (2)$$

where the constant k >0 is known as "ridge" parameter. Lawles and wang(1976) , Horel Kennard and Baldwin ,Khalaf and Shukur(2005) are the leading method for estimating value of k.

Autocorrelation:

Correlation between successive observations in linear regression model is leads to Autocorrelation. When the error satisfies the first order autocorrelation it is represented by

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \,; t = 1,2,3, \dots, n \quad \text{.....(3)}$$

u_t are iid random variable with mean zero and variance σ_u^2. The variance covariance matrix of Y is D(Y) = D(ε) =$\sigma^2\Omega \neq \sigma^2 I$, where $\sigma^2 = \sigma_u^2 \frac{1}{1-\rho^2}$

The generalized least squares estimator of β is

$$\widehat{\beta_{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \qquad (4)$$

If the parameter ρ in (3) is known then we can write ρ

$$\Omega^{-1} = \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix} \qquad (5)$$

Generalised Ridge Type Estimator: Consider a general linear regressive model with errors satisfying relation in equation (1) and (3) respectively) and the regressors exhibiting near multicollinearity. As seen earlier, in case of autocorrelation D(Y) = D(ε) =$\sigma^2\Omega \neq \sigma^2 I$. Autocorrelation is special case of heteroscedasticity and hence GLS can be applied to estimate the model parameters. Further, once there's multicollinearity, usually used technique is that the ridge regression as mentioned in (2). Combining these two methods, propose a new method for the autocorrelated model with multicollinearity a generalized ridge type estimator represented as $\widehat{\beta_{GR}} = (X'^{\Omega^{-1}}X + kI)^{-1}X'\Omega^{-1}Y$ where Ω^{-1} where Ω^{-1} is as defined in (5). Hence the model under consideration contains the unknown parameters k, μ, σ^2 and β. Methods for estimating ridge parameter k defined in equation (2a), (2b) and (2c).

3.2 Non-Parametric Regression:

Non-parametric regression used for estimating a regression curve while not creating sturdy assumptions regarding the form of actuality regression perform. These techniques are helpful for building and checking parametric form and data description. The non-parametric model is of the form Y = m (X) +ε ,where Y is the response variable, m(X) is the mean response of the regression function. E(Y/X=x) = m(X) is assumed to be smooth and ε is the independently and identically distributed random error with mean zero and constant variance.

Decision tree-based regression

A **decision tree** could be a flowchart-like tree structure, wherever every **internal node** (non-leaf node) denotes a test on an attribute, every **branch** represents an outcome of the test, and each **leaf node** (or *terminal node*) holds a category label. The topmost node in tree is the root node. Given a set observation in X for which class label is unknown, the attribute values of the observation is tested against the constructed tree. A path is traced from the root to a leaf node, that holds the class prediction/predicted values for that tuple.

Since decision tree is a non- parametric algorithm it doesn't requires any domain knowledge and therefore it is an effective method for exploratory knowledge discovery. The average of dependent variable values in a tuple is taken as the predicted value for all those tuples. CART algorithm applies the Gini index as the attribute selection measure.

K-Nearest Neighbours based Regression (KNN)

K-nearest neighbour is based on learning by analogy, that is by comparing a given test tuple with a training tuple that similar to it. When given an unknown tuple, a K- nearest neighbour classifier searches the pattern space for the k-training tuple that is closest to the unknown tuple. Closeness is defined in terms of distance matric, such as Euclidean distance. The unknown tuple is assigned to the foremost class among its K-nearest neighbours in KNN. K-nearest neighbour classifiers are also used for prediction, that is, to return a real-valued prediction for a given unknown tuple and, in this instance, the classifier returns the average value of the real-valued labels associated with the K-nearest neighbours of the unknown tuple. A good value for K, number of nearest neighbours, can be found experimentally or k may be taken as k = $\sqrt{number\ of\ training\ tuples}$

3.3 Time Series Analysis

A Time Series (TS) is a sequence of observations ordered in equally spaced, discrete time intervals. A basic assumption in any time series analysis/modelling is that some aspects of past pattern will continue to remain the same in the future. Suitable forecasting time series model can be developed which gives minimum forecasting error. At least 50 observations are necessary for performing Time Series analysis, as propounded by Box Jenkins who were pioneers in Time Series modelling. The four main objectives of Time Series Analysis are: *Description, Explanation, Prediction and Control.* We start by plotting the time series data and look for non-stationary components. We then eliminate these components using various methods, in order to get a stationary data. After identifying a suitable probability model for the time series, this model can be used for prediction. Analysis of time series consist of time profile, Making series stationary, Model building, Diagnostic checking, Forecasting.

TESTING FOR THE PRESENCE OF TREND COMPONENT

Mann-Kendall trend test:

The Mann-Kendall trend test is a nonparametric test used to identify a trend in a series, even if a seasonal component exists. The hypothesis for this test is as follows:

H_0: There is no trend in the data.

H_1: There is monotonic trend in the data.

The Mann-Kendall test is based on the calculation of Kendall's tau measure of association between two samples, which itself is based on the ranks of the samples. Here the main assumption is that observations are independent.

TESTING FOR THE PRESENCE OF SEASONALITY COMPONENT

Rank-sum test

The Rank-sum test is a test used to identify seasonality in a series. The hypothesis for this test is as follows:

H_0: There is no seasonal variation in the data

H_1: There is seasonal variation in the data

Test statistic is $\chi_0{}^2=\frac{12\sum_{j=1}^{D}(M_j-\frac{C(D-1)}{2})^2}{CD(D+1)} \sim \chi^2_{(D-1)}$

D-seasonality periods, C-total number of years, M_j-sum of the ranks for the j^{th} period. If chi-square calculated is more than chi-square critical value, then we reject H_0 and conclude that there is seasonal variation is there in the data.

Autoregressive Integrated Moving Average (ARIMA) Model:

This model is one of the most popular and frequently used stochastic time series model the fundamental assumption created during this model is that the considered time series is linear and follows a specific known statistical distribution, such as the normal distribution. In this model a non-stationary time series is formed stationary by applying finite differencing of the data points.

Let $\{X_t, t \in I\}$ denote a non-stationary time series, non-stationary due to trend component. Let $\{\in_t, t=\pm1,\pm2, ...\}$ be a sequence of white noise. Then $\{X_t, t\in I\}$ is said to follow autoregressive integrated moving average process if it is has the following representation

$\emptyset(B)(1-B)^d X_t = \theta(B)\in_t$

Where $\emptyset(B)=1-\beta_1 B-\ldots\ldots\ldots-\beta_p B^p$

$\theta(B)=1-\alpha_1 B-\ldots\ldots\ldots-\alpha_q B^q$

$\alpha_1,\alpha_2,\ldots\ldots,\alpha_q$ are MA parameters and $\beta_1,\beta_2,\ldots\ldots\ldots,\beta_p$ are AR parameters. Where d is the number of time difference taken to make the series stationary. This model is also known as Box-Jenkins model.

Seasonal Autoregressive Integrated Moving Average (SARIMA)Model

The ARIMA model is for non-seasonal non-stationary data whereas SARIMA model is designed to deal with seasonality. Here, non-stationarity from the series is removed by seasonal differencing of appropriate model. The difference between an observation and the corresponding observation from the previous year constitutes first order seasonal difference and is calculated as $z_t = y_t - y_{t-1}$ for monthly and quarterly time series s = 12 and s = 4 respectively.

This model is generally termed as $SARIMA\ (p,d,q)(P,D,Q)_{[s]}$ model. Where p and q denote non seasonal ARMA coefficient, d denotes number of non-seasonal difference. P-Number of seasonal AR term, Q-Number of multiplicative seasonal MA term, D-Number of seasonal differencing required to remove trend in seasonality, s-Seasonal period. $SARIMA\ (p,d,q)(P,D,Q)_{[s]}$ has the representation

$$\emptyset(B)\ \emptyset(L)\ (1-B)^d\ (1-L)^D\ X_t = \theta(B)\ \theta(L)\ \varepsilon_t$$

Where , $L=B^S$

$$\emptyset(B) = 1-\beta_1 B-\ldots\ldots-\beta_p B^p$$
$$\theta(B) = 1-\alpha_1 B-\ldots\ldots-\alpha_q B^q$$

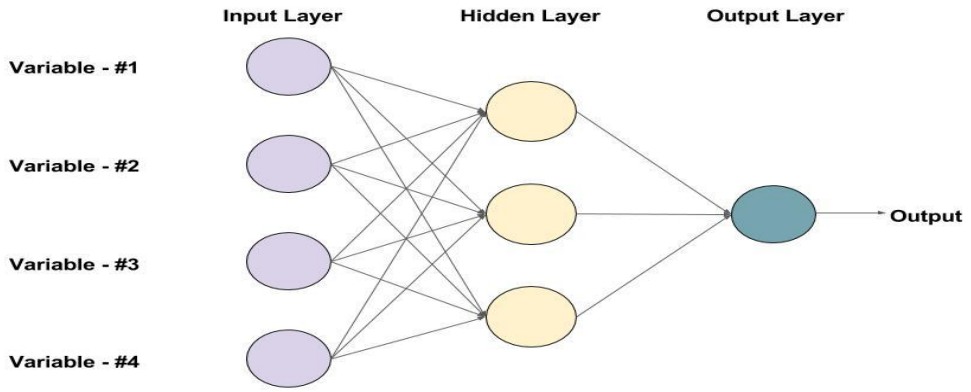
$$\emptyset(L) = 1 - \emptyset_1\text{-}\ldots\text{-}\emptyset_P L^P$$
$$\theta(L) = 1 - \theta_1\text{-}\ldots\text{-}\theta_Q L^Q$$

ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN) is a computational model which is inspired from the structure and functions of biological neural networks. Information that is passed through the network affects the structure of the ANN, because a neural network changes based on that input and output. One of the most recognised advantages of ANN is that it can actually learn from observing the data. ANN is employed as a random perform approximation tool and these forms of tools facilitate to estimate the foremost cost-efficient and ideal strategies for inward at solutions while defining computing functions or distributions. The three layers of ANN are interconnected. The first layer is named as input neurons which sends data onto the second layer, which in turn sends the output neurons to the third layer.

Multilayer perceptron (MLP)

A multilayer feed-forward neural network consists an associate in nursing input layer, one or more hidden layers and an associate in nursing an output layer. A multilayer feed-forward neural network is associated interconnections of perceptron's during which knowledge and calculations flow in exceedingly single directions, from the computer file to the outputs. The number of layers in a neural network is the number of layers of perceptron's. The best neural network is one with one input layer which is associated to an output layer of perceptron's, successive most intricate neural network is one with 2 layers. This additional layer is mentioned as a hidden layer. In general, there's no restriction on the quantity of hidden layers. The back propagation algorithmic program performs learning on a multilayer feed-forward neural network. It iteratively learns a collection of weights for prediction of the category label of tuples.



An example of a Feed-forward Neural Network with one hidden layer (with 3 neurons)

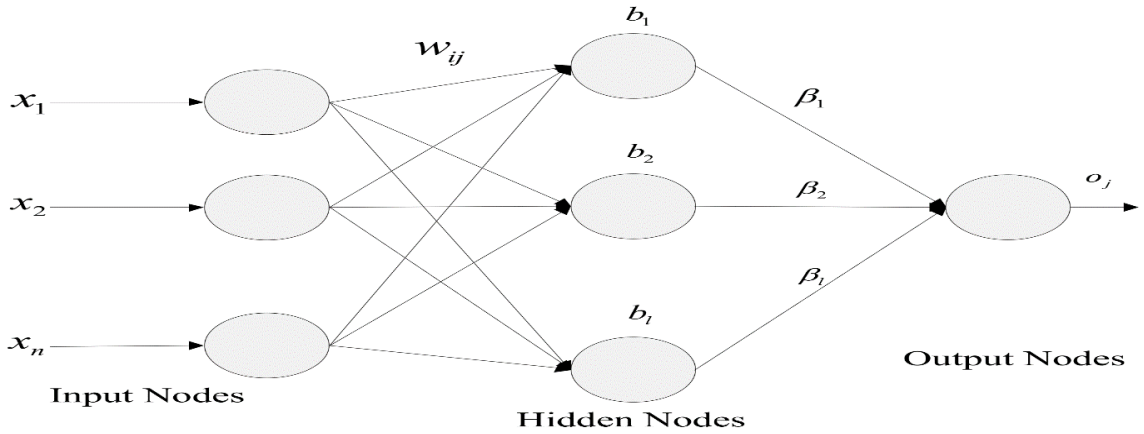
Each layer consists of units. The inputs to the network correspond to the attributes measured for every training tuple. The inputs are fed at the same time into the units creating the input layer. These inputs undergo the input layer and are then weighted and fed at the same time to a second layer of "neuronlike" units, called a hidden layer. The outputs of the hidden layer units are inputs to a different hidden layer, and so on. The weighted outputs of the hidden layer are inputs to units creating the output layer, which emits the network's prediction for the given tuples. Each output unit takes, as input, a weighted total of the outputs from units in the previous layer.

Extreme Learning Machine (ELM)

Extreme learning machine are feed-forward neural networks for classification, regression, clustering and prediction, compression and feature learning with a single layer or multiple layers of hidden nodes, wherever the parameters of hidden nodes needn't be tuned. These hidden nodes often at randomly assigned and never updated or can be inherited from their ancestors without being changed. In most of the cases, the output weights of hidden nodes are often computed in a single step, which basically amounts to learning a linear model.

In most instance, ELM is employed as one hidden layer feed-forward network. These models are able to produce smart generalization performance and learn thousands of times faster than networks trained using back propagation.

Accuracy Measures



In forecasting, our objective is to produce and optimum forecast that has no error or as little error as possible, which leads us to the minimum mean square error forecast. This forecast can produce an optimum future value with the minimum error in terms of the MSE criterion. We compared different models based on RMSE, MAE and MAPE.

IV. Data Analysis and Results:

It is known that almost entire city of Delhi is enveloped by pollution layer all around with the contribution from multiple sources within the city, nearby region and even from long distances. In a broader sense the air is more toxic as it contains larger contribution of combustion products. The levels of PM particles and CO are statistically higher. Therefore, in this research paper we focus on the identification and estimation of influential factors of AQI of Delhi and forecasting the future AQI and its risk factors. The data set was collected from the website of 'Central Pollution Control Board of India'. Variables included in the study are CO: Carbon Monoxide in mg/m3, NH3: Ammonia in ug/m3, $PM_{2.5}$: Particular Matter 2.5-micrometer in ug/m3, PM_{10}: Particular Matter 10-micrometer in ug/m3, NO: Nitric oxide in ug/m3, NO2: Nitrogen dioxide in ug/m3, NOx: Nitric oxide-x in ppb, Benzene: Benzene in ug/m3, SO2: Sulphur Dioxide in ug/m3, O3: Ozone in ug/m3, Toluene: Toluene in ug/m3, AQI: Air quality index

Table 1: Model parameters, Significant regressors

	Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	67.27355	15.00783	4.483	3.61E-05	***
Benzene	5.69768	4.02657	1.415	0.1625	
CO	8.31893	1.60312	5.189	2.92E-06	***
O3	-0.10039	0.24149	-0.416	0.679276	
SO2	-0.4171	0.75871	-0.55	0.584757	
NO	-0.39651	0.39909	-0.994	0.324883	
NH3	-1.08409	0.38472	-2.818	0.00663	**
NO2	0.46104	0.34394	1.34	0.1854	
NOx	0.43024	0.13497	3.188	0.00233	**
Toluene	-0.90165	0.32052	-2.813	0.00672	**
PM10	0.28054	0.09025	3.108	0.00293	**

PM2.5	0.8746	0.13368	6.543	1.83E-08	***

Residual standard error: 23.68 on 57 degrees of freedom. Multiple R-squared: 0.9506, Adjusted R-squared: 0.9437. i.e, 94% variation in AQI can be explained by these regressors. F-statistic: 137.1 on 8 and 57 DF, p-value: < 2.2e-16.

From Table 1, we observed that CO, NH3, NOx, Toluene, PM10, PM2.5 are the significant variables for the Air quality. Among which NH3 and Toluene has negative effect and remaining variable has positive effect, in that CO and PM 2.5 has large positive impact on air pollution compared to all other variables. From the F-statistic we conclude that the overall model is good fit. The coefficient of determination R^2=0.9506 implies that 95% of the variation in AQI is explained by significance regressors.

The Durbin Watson test shows that there is autocorrelation in the dataset. Correlation matrix and VIF values shows the existence of multicollinearity between the regressors. Therefore, we applied ridge regression, generalised least square estimator to deal with above two problems independently. Generalised ridge regression was also applied to handle multicollinearity and autocorrelation simultaneously.

Table 2: Comparing the Performance of Parametric and Non-Parametric regressors

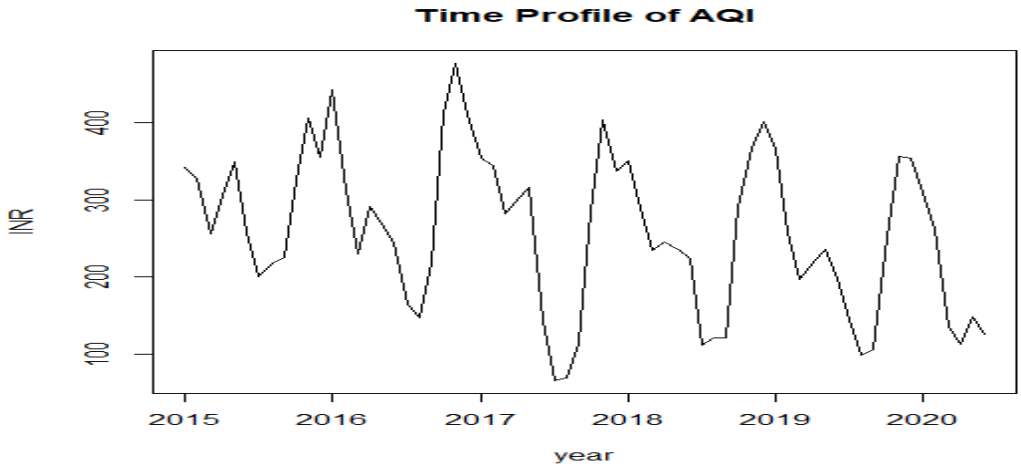
SI. No		Methods		RMSE
1.	PARAMETRIC REGRESSION	MULTIPLE LINEAR REGRESSION		27.67023
2.		MULTICOLLINEARITY	Ridge trace	31.41226
3.			Lawless and Warg 1976	284.1541
4.			Hoerl, Kennard and Baldwin	31.22202
5.			Khalaf and Shukur 2005	28.3508
6.		AUTOCORRELATION	GLS	23.38767
7.		**GENERALIZED RIDGE TYPE ESTIMATOR**	Ridge trace	23.75822
8.			**Lawless and Warg 1976**	**23.07448**
9.			Hoerl, Kennard and Baldwin	23.6935
10.			Khalaf and Shukur 2005	23.11178
11.	NON-PARAMETRIC REGRESSION	REGRESSION TREE		34.64482
12.		K-NEAREST NEIGHBOURS REGRESSION		28.46958

From above Table 2, it is clear that Generalised Ridge Type estimator under Lawless and Warg method performance better than all other Parametric and Non-Parametric methods based on minimum Root mean square error.

TIME SERIES:

In the analysis, data from Jan-2015 to June-2019 is taken as train data and July-2019 - June-2020 data is taken as test data.

Air Quality Index:



From the above figure we observe that there is clear seasonal variation in the data. applied Mann Kendall test to test for the presence of trend and rank sum test to test for presence of seasonal variations and obtain there is seasonality and no trend in the data respectively. Hence removed seasonality by carrying out seasonal difference to make the series stationary. Therefore, D=1and d=0.

From ACF, we observed that the value of seasonal MA order Q=1 and non-seasonal MA order q=1 are obtained and from PACF plot AR order P=2 and non-seasonal AR order p=1 are obtained. We fitted SARIMA model based on different combinations of (P, D, Q) and (p, d, q)

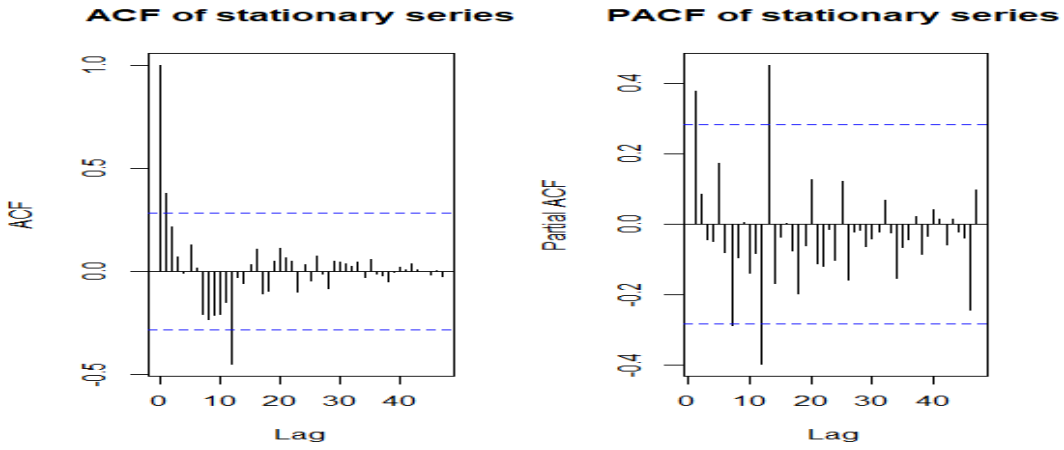
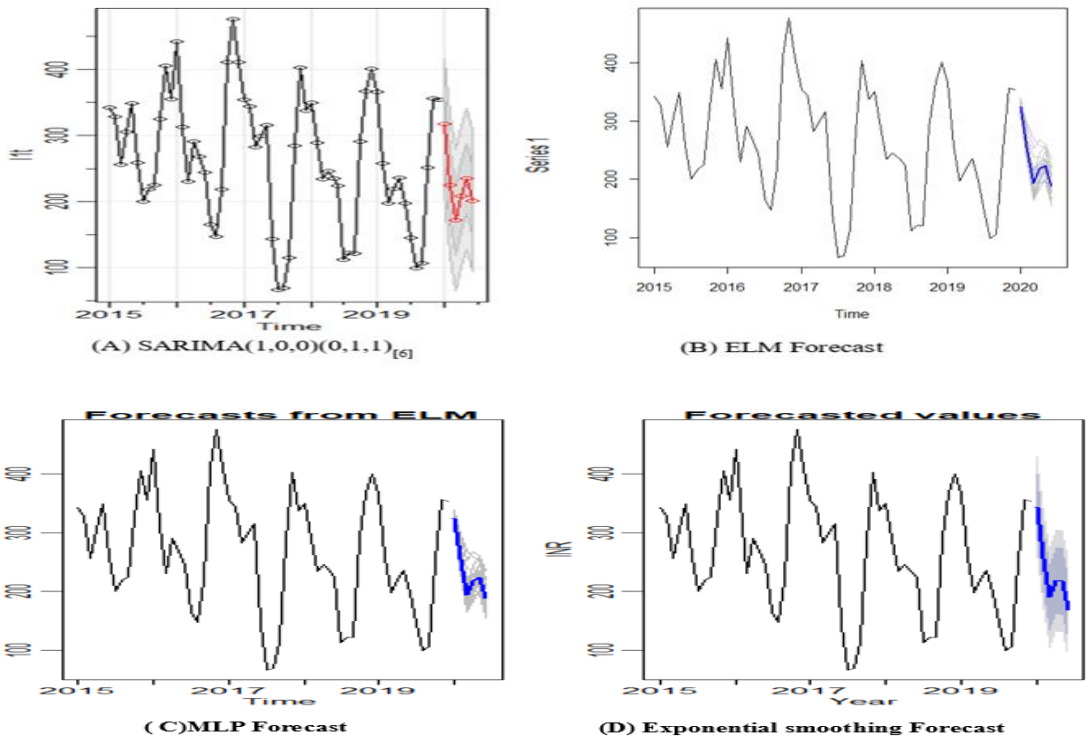


Table 3: SARIMA model based on different combinations of (P, D, Q) and (p, d, q)

Model	(p,d,q)(P,D,Q)	AIC	Ljung-Box p-vale	Box –Pierce p-vale
1	(0,0,1)(0,1,1)	510.6074	0.4074209	0.42912668
2	(0,0,1)(1,1,0)	508.3981	0.7578631	0.81021876
3	(0,0,1)(1,1,1)	510.3595	0.5920386	0.65715031
4	(0,0,1)(2,1,0)	510.3623	0.8397678	0.8694508
5	(0,0,1)(2,1,1)	510.6074	0.4074209	0.42912668
6	(1,0,0)(0,1,1)	510.6074	0.2727363	0.29382440
7	(1,0,0)(1,1,0)	510.6074	0.4074209	0.52912668
8	(1,0,0)(1,1,1)	510.6074	0.2727363	0.29382440
9	(1,0,0)(2,1,0)	510.6074	0.1752759	0.26554567
10	**(1,0,0)(2,1,1)**	**510.3623**	**0.8397678**	**0.90694508**
11	(1,0,1(0,1,1))	510.6074	0.4074209	0.42912668
12	(1,0,1)(1,1,0)	510.6074	0.2727363	0.29382440
13	(1,0,1)(1,1,1)	510.6074	0.4074209	0.42912668
14	(1,0,1)(2,1,0)	510.6074	0.2727363	0.39382440
15	(1,0,1)(2,1,1)	510.6074	0.1752759	0.26554567



Forecast diagram of AQI variable

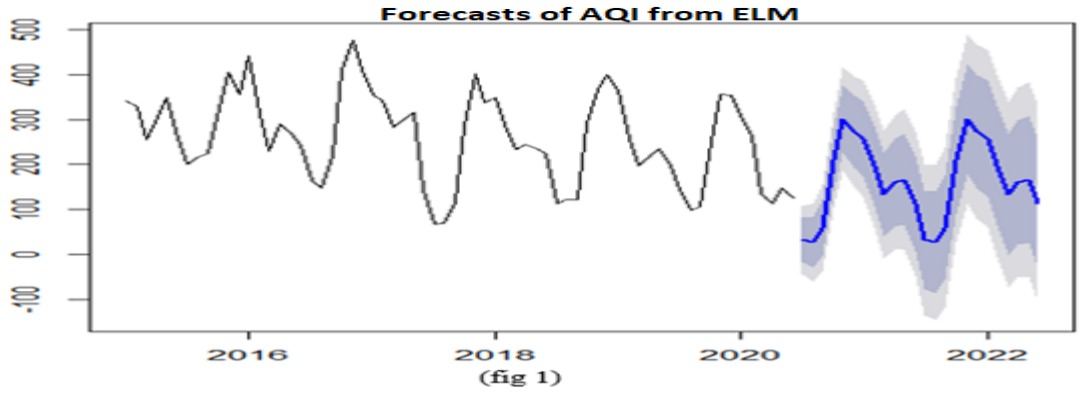
From Table 3 we obtain, **SARIMA**$(1,0,0)(2,1,1)_{[12]}$**is performance better than all other model** based on Minimum AIC value and Maximum Box–Pierce P-value.

Table 4: Accuracy measures RMSE, MAE and MAPE:

Model	RMSE	MAE	MAPE
SARIMA	44.2295	28.7656	18.9223
MLP	38.76033	27.7338	17.90521
ELM	**35.08217**	**27.10994**	**16.91449**

From the table 4, it is clear that on the value of RMSE, MAE, MAPE is minimum for **ELM** model compared to SARIMA and MLP model. Therefore, ELM model is the best model for forecasting the Air Quality Index. The best model is used to forecast the AQI for the next 24 months.

Figure 1: Represents the plot of actual and forecasted values of AQI using ELM method



(fig 1)

Time series analysis of PM2.5:

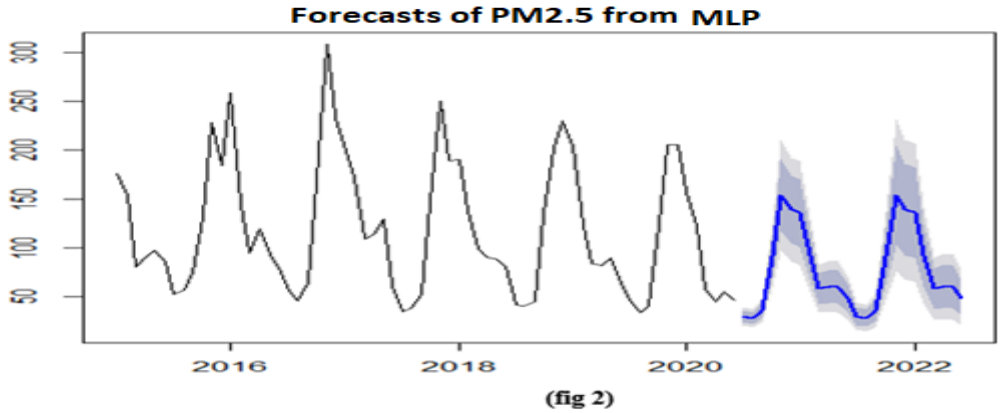
Table 5: Accuracy measures RMSE, MAE and MAPE

Model	RMSE	MAE	MAPE
SARIMA	31.91908	28.07359	36.82567
MLP	**21.18835**	**19.04739**	**20.87819**
ELM	22.46495	18.56267	21.90037

From the table 5, it is clear that on the value of RMSE, MAE, MAPE is minimum for **MLP** model compared to SARIMA and ELM model. Therefore, MLP model is the best model for forecasting the $PM_{2.5}$. The best model is used to forecast the $PM_{2.5}$ for the next 24 months.

Figure 2: Represents the plot of actual and forecasted values of $PM_{2.5}$ using MLP method

Time series analysis of Carbon Monoxide:



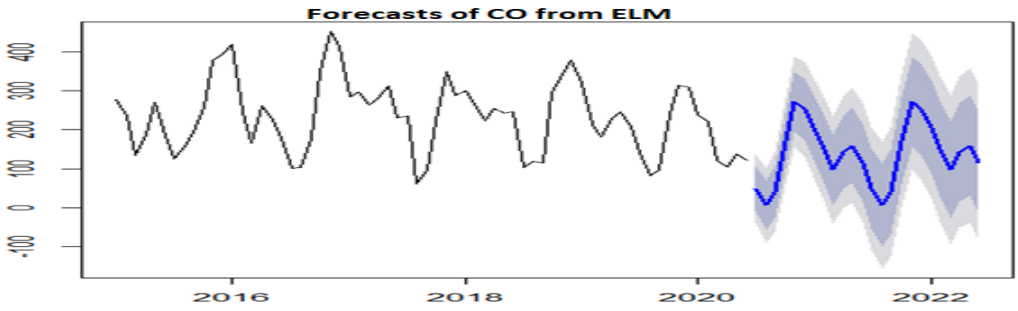
(fig 2)

Table 6: Accuracy measures RMSE, MAE and MAPE of different models

Model	RMSE	MAE	MAPE
SARIMA	38.14983	31.14321	21.42326
MLP	30.0751	28.87803	17.05054
ELM	**29.10294**	**26.11007**	**16.94926**

From the table 6, it is clear that RMSE, MAE, MAPE is minimum for **ELM** model compared to SARIMA and MLP model. Therefore, **ELM** model performance better than all other models for forecasting the CO. This model is used to forecast the CO for the next 24 months.

Figure 3: Represents the plot of actual and forecasted values of CO using ELM method



The above forecasted graph can be used as a guide to reduce the emission of carbon monoxide. This graph also helps us to identify the months during which emission of CO is more and proper measures to curb CO can be implemented.

V. Conclusion:

Global warming challenges is a serious problem and most people are still unaware of global warming causes and its effects. These challenges are not only causing trouble to the human beings but also to animals and plants. Hence appropriate measures must be taken to tackle this hazard. Air quality Index (AQI) is the tool for identifying the severity of the air pollution. In this paper innovative applications of statistical procedure are suggested for the analysis and modelling of Air pollution of Delhi. The regression analysis results show that CO, NH_3, NOx, Toluene, PM_{10} and $PM_{2.5}$ are significant variable for the AQI. Among these, $PM_{2.5}$ and CO shows large positive impact on AQI. Therefore, more attention is needed to curb the CO level so that air pollution can be controlled in significant level. This paper successfully handles the multicollinearity in presence of autocorrelation rather than handling them separately. The results show that in this situation Generalized Ridge Type estimator performs better than all other types of estimator. In ordered to control Air Pollution we need to control the emission of its significant pollutants. Therefore, we used time series analysis and forecast the AQI and its Significant risk factors for the next two years.

On the basis of results obtained from the statistical techniques, to embark upon Air Pollution Challenges, we advocate the following **policy measures** for the administrators.

i. The administrator needs to establish climate policies aimed at curbing global emissions in ordered to improve the air quality.

ii. Importance must be given to the use of alternative energy sources such as solar, wind, hydro, geothermal, bio mass to avoid continues burning of coal which in turn reduces the produce of carbon dioxide, methane and nitrous oxides.

iii. Awareness must be given to use the recycled items to reduce Deforestation and Administrators must take initiatives such as "One Student One Tree" as taken by UGC.

iv. Government should divert their attention to reduce emission of gases from vehicles by giving more privilege to electronic vehicles and also need to control over the emission of gases from industry.

v. Government need to take project for improving efficiency in the conversion of fuel to electricity which leads to reduction in pollution.

REFERENCES:

[1]. Box GEP and Jenkins G.M(1976): Time Series Analysis: Forecasting and Control, Holden-day, San Franscisco.
[2]. Central Pollution Control Board of India: www.cpcb.nic.in
[3]. Chatfield C.(1996): The Analysis of Time Series An Introduction, Chapman & Hall.
[4]. Dilip M. Nachane (2006) "Econometrics- Theoretical Foundations and Empirical Perspectives", OUP India
[5]. Ismail B., Manjula Suvarna: Estimation of Linear Regression Model with Correlated Regressors in the Presence of Autocorrelation
[6]. Peter Schmidt, 1976,. Econometrics, New York: Marcel dekker Inc.
[7]. Hoerl, A. E. and Kennard, R.W, 1970, Ridge regression: biased estimation for nonorthogonal problems, Technometrics, 12:9-82.
[8]. Jalika V.N. and Patil B.L(2015): Onion price forecasting in Hubli market of Northern Karnataka using ARIMA technique, Karnataka, Journal Agricultural Science, 28(2): 228-231.
[9]. Jiawei Han, Micheline Kamber(2002): Data Mining-Concepts and Techniques, Morgan Kaufman Publishers, U.S.A
[10]. Johnston J. (1984): Econometric Methods, 3rd Edition., McGraw Hill.
[11]. M. Venkataramanan and Smitha(2011): "Causes and Effects of Global Warming", Indian Journal of Science and Technology, Vol.4, issue 3
[12]. Satyanarayana, Swathi, Ismail B. (2018): Statistical Analysis of Food Grain Prices Karnataka, research Reviews International Journal of Multidisciplinary, Vol.3 (8).
[13]. Simon Haykin (2012): Neuarl Networks and Learning Machines, 3rd edition, PHI Learning Private Limited, New Delhi
[14]. Sunil, Satyanarayana, Sachin Acharya, Arun Kumar (2019): Application of Hybrid Model for Forecasting Prices of Jasmine Flower in Bangalore, India. International Journal of Scientific & Technology Research, 8, No. 11.