Bulletin of Pure and Applied Science
Physics, Vol.42D No.2S (Special Issue)
NCAEST-2023 P.29-39
Original Research Article

Print version ISSN 0970 6569
Online version ISSN 2320 3218
DOI: 10.62427/bpas.2023.42D.2S.4
Available online at www.bpasjournals.com

A Forecasting Approach to Predict the Air Quality Index and Its Related Pollutants

[1]Abhishek Kumar, [2]Rajkishor[*], and[3]Manish Kumar

Author's Affiliations:
[1]B.Tech Graduate, Department of Civil Engineering, Bhagalpur College of Engineering, Bhagalpur, Bihar-813210, India.
E-mail: abhishek13900@gmail.com
[2]Assistant Professor, Department of Civil Engineering, Bhagalpur College of Engineering, Bhagalpur, Bihar-813210, India.
E-mail: rk.bce22@gmail.com
[3]B.Tech Graduate, Department of Civil Engineering, Bhagalpur College of Engineering, Bhagalpur, Bihar-813210, India.
E-mail: manishbce11@gmail.com

***Corresponding author:**
Rajkishor
Assistant Professor, Department of Civil Engineering, Bhagalpur College of Engineering, Bhagalpur, Bihar-813210, India.
E-mail: rk.bce22@gmail.com
Contact No.: +91-8210878137

ABSTRACT

Nowadays, pollution has grown to be a major worry. People ought to be aware of the air they breathe. An approach for determining the current state of air quality is the Air Quality Index (AQI). AQI gives the concept about quality of air or at what degree the air in the particular location is polluted. The present study explores multiple methodologies for estimation and forecasting of the Air Quality Index (AQI) by considering the synergistic effects of key pollutants, specifically PM10, PM2.5, SO_2, and NO_2. The study has been conducted for the regions of New Delhi, India to compare the ambient air quality. In this study various methodologies have been employed such as Center-Line Moving Average (CMA) technique, in addition to conventional AQI estimate techniques together with machine learning techniques to forecast the AQI values for the subsequent month with certain degree of tolerance. The study region's air quality state was classified into good, moderate, satisfactory, and unacceptable classes for different AQI calculations, according to seasonal and daily AQI calculations.

KEYWORDS

Air Quality Index (AQI), PM10, PM2.5, SO2, NO2, Regression Modeling, Machine Learning, Pollution Assessment, CMA etc.

How to cite this article: Kumar A., Rajkishor, and Kumar M. (2023). A Forecasting Approach to Predict the Air Quality Index and Its Related Pollutants. *Bulletin of Pure and Applied Sciences- Physics,* 42D (2 Special Issue), 29-39.

1. INTRODUCTION

The essential element we rely on for life, oxygen, is present in the air that we breathe. It's a crucial component that sustains the functioning of human cells. More than just supporting life, the quality of the air we inhale affects our overall well-being. Poor air quality can have far-reaching health consequences. Contaminated air contains harmful elements that can lead to various respiratory illnesses. Furthermore, air pollution contributes to the concerning issue of global warming. This process involves the trapping of heat in the atmosphere, leading to detrimental effects such as heat-related health problems (Bishoi and Prakash, 2009; (Chakraborty, Debnath and Bose, 2021; Geetha et al., 2020).

The Government of India's Central Pollution Control Board (CPCB) uses the country's Air Quality Index (AQI) for Indian cities which is determined by taking into account 12 air pollutants, which include NO_2 (nitrogen dioxide), SO_2 (sulphur dioxide), CO (carbon monoxide), O_3 (ozone), PM10 (particulate matter with a diameter of less than 10 microns), PM2.5 (particulate matter with a diameter of less than 2.5 microns), NH3 (ammonia), Pb (lead), Ni (nickel), As (arsenic), Benzol, Pyrene, and Benzene "National air quality index report." Central Pollution Control Board (CPCB), 2015.

In the literatures, it is reported that AQI can be assessed based on maximum five sub-index approach using suspended particulate matter (SPM), SO_2, CO, PM10, and NO_2. Assessing air quality is possible through the Air Quality Index (AQI), a numerical representation of air quality. Some of the researchers have explored the air quality prediction technique using machine learning. There is a need to develop a uniform and efficient AQI scheme which provides information about every pollutant (Guttikunda, 2010; Kanchan, Gorai and Goyal, 2015). However, some researchers have developed techniques to predict AQI in literatures (Kumar and Goyal, 2011; Lemes, 2018; Mamta and Bassin, 2010; Taieb and Brahim, 2013).

In the present study, there have been proposed two forecasting approaches using one of the simplest machine learning technique using linear regression model and central moving average (CMA) method to predict the AQI and related pollutant's concentration.

2. METHODOLOGY FOR PREDICTION OF AQI AND RELATED POLLUTANT

2.1 First Approach Using Machine Learning (ML) Method:

Machine learning involves mathematical computations where algorithms perform various computations for predicting data (Kumar and Pande, 2022; Soundari, Jeslin and Akshaya, 2019).The accuracy of the same depends upon the accuracy and amount of input data. This approach involves various steps as following-

Step 1: Data Collection-

The AQI and related parameter data have been collected from CPCB official website (https://cpcb.nic.in/) for New Delhi area. The data set contains air quality data and AQI of various parameters in New Delhi. There have been performed machine learning experimentation using Air quality data of New Delhi of duration from April 2018 to August 2023 using algorithms identified through literature review. Figure 1 shows the representation of raw data in graphical format.

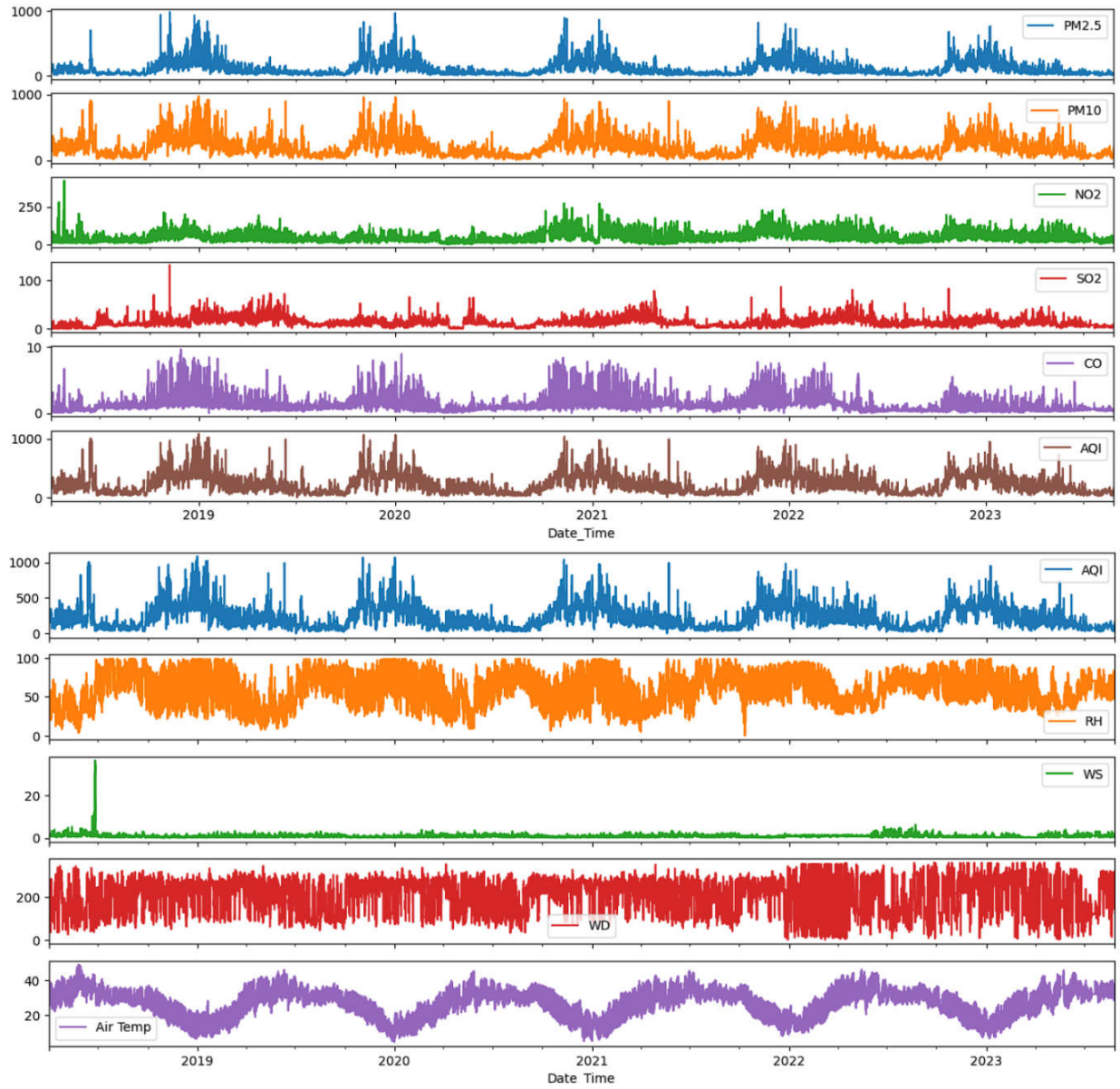

Figure 1: Graphical trend of Dataset used in ML.

***Step* 2: Data preprocessing-**

Python version 3.9 has been selected as programming environment for data preprocessing. In this steps datasets are transformed into more useful format in a structured way by handling missing data and unwanted vague data in order to enhance the time economy and storage economy.

Step 3: Data analysis-

After data preprocessing, correlation analysis between all the parameters for selection of training parameters has been done. Selected parameters for training of the models are:

PM2.5, PM10, NO2, SO2, CO, Ozone, AQI, RH, WS, WD, Air Temperature. Figure 2 shows correlation metrics between the selected parameters. The description of data is shown in figure 3 based on the co-relation analysis.

	PM2.5	PM10	NO2	SO2	CO	Ozone	AQI
PM2.5	1.000000	0.901302	0.513394	0.306215	0.625606	-0.233993	0.931015
PM10	0.901302	1.000000	0.545926	0.373719	0.607670	-0.141764	0.956644
NO2	0.513394	0.545926	1.000000	0.327444	0.542395	-0.321891	0.519476
SO2	0.306215	0.373719	0.327444	1.000000	0.270489	0.068639	0.350229
CO	0.625606	0.607670	0.542395	0.270489	1.000000	-0.278421	0.589732
Ozone	-0.233993	-0.141764	-0.321891	0.068639	-0.278421	1.000000	-0.128670
AQI	0.931015	0.956644	0.519476	0.350229	0.589732	-0.128670	1.000000

Figure 2: Correlation between parameters of training data.

	PM2.5	PM10	NO2	SO2	CO	Ozone	AQI
count	11838.000000	11838.000000	11838.000000	11838.000000	11838.000000	11838.000000	11838.000000
mean	113.614849	205.418791	53.965487	13.883645	1.369263	43.352311	225.872783
std	114.383447	147.875999	34.341016	8.755720	1.230140	45.463911	157.445865
min	1.000000	1.250000	1.740000	0.100000	0.010000	0.100000	1.000000
25%	36.690000	93.380000	28.112500	7.992500	0.610000	7.810000	101.000000
50%	69.590000	169.280000	45.005000	12.230000	0.990000	25.515000	178.000000
75%	153.750000	278.440000	73.187500	18.050000	1.610000	64.580000	329.000000
max	991.460000	978.800000	424.610000	131.740000	9.670000	199.900000	1086.000000

Figure 3: Data description

Step 4: Training and testing the model-

For training of selected machine learning models the dataset are split into 70% training data and 30% testing data. Matplotlib library has been used to assess the efficiency of predicted values over the actual values using scatter graph and plot graph. Same data and parameters have been used for training of all the selected algorithms. Linear regression machine learning model has been selected for the forecasting of AQI using ML technique as shown in figure 4.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30, random_state = 101)
from sklearn.linear_model import LinearRegression
linreg = LinearRegression()
linreg.fit(X_train,y_train)
```

```
▾ LinearRegression
LinearRegression()
```

```
ypred = linreg.predict(X_test)
print(ypred)
```

```
[165.6605109  159.71288261  31.55312716 ... 218.26687299  57.21647914
 578.85165652]
```

```
linreg.coef_
```

```
array([ 0.18487395, -0.01541584,  0.06818268,  5.71187899,  0.04274358,
        0.44519294,  0.58837237,  0.37190351, -0.01932281, -1.36255538])
```

```
plt.scatter(y_test, ypred)
```

Figure 4: Linear regression model for AQI prediction

2.2 Second Approach Using Centre Line Moving Average (CMA) Method:

Step-1: The available AQI and its related pollutants concentration data for New Delhi (data of 6 years) have been collected from the web-source of CPCB, Govt. of India as shown in figure 1.

Step-2: It has been observed that as season changes throughout the year the variation pattern of various pollutant concentration changes. Hence, the forecasting approach has been developed based on their seasonal variation.

Step-3: A forecasting approach called Centre-line Moving Average (CMA) method has been adopted in which under-mentioned factors are determined to forecast the pollutant concentration at a given date/time:-

- All the available data (y_t) has been arranged in sequence manner and have been assigned rank (R) on chronological basis.
- These datas have been then grouped on month basis and then their moving average (MA) has been calculated.
- Adopting CMA technique, moving average of previously calculated MA datas have been obtained.
- Irregularity trend (I_t) of the known conc. data is then determined as-

$$I_t = y_t / CMA \quad (1)$$

- Seasonality trend (S_t) of the known conc. data is then determined by monthly average of Y_t.
- Then De-seasonalized factor (DS_t) is obtained by y_t/S_t.
- A regression analysis has been performed on rank data and De-seasonlized factor in order to obtain their variation trend factors in terms of slope (M) and intercept (C) using Data analysis option available in MS Excel for estimating the trend component of the dataset.
- The trend component (T_t) is then formulated as-

$$T_t = R \times M + C \quad (2)$$

- Finally, the forecasted value (F_{PP}) of a given pollutant concentration is obtained by-

$$F_{PP} = S_t \times T_{t(1.3)} \quad (3)$$

- Using the equation (1) and (2), the sub-index of all the pollutant parameters are determined and then AQI is taken as maximum of all the sub-indices calculated above for the desired date.

3. RESULTS AND DISCUSSION

3.1. CMA approach of forecasting:

The Figure 5 to figure 8shows the graphical comparison among observed and calculated using CMA forecasting approach for AQI, PM2.5, PM10, and $SO_2$for New Delhi city area. The figure 5 to 8 shows the very good agreement between observed and forecasting AQI and related pollutant's concentration values with allowable tolerance limit.

Also the forecasting for the AQI is made for the next one month period i.e 01st August 2023 to 31st August 2023 (Table 1). The average deviation is found to be 12.8%. Some deviations are also observed for some values which may be due to vagueness of the data arised due to many reasons such as festival etc. It has been observedthat data varies in a particular pattern in all summer, winter and rainy season. In winter season air pollution is maximum and in rainy season the air pollution is minimum. It is found that pollutant PM2.5, PM10 and NO_2 are the prominent pollutant.

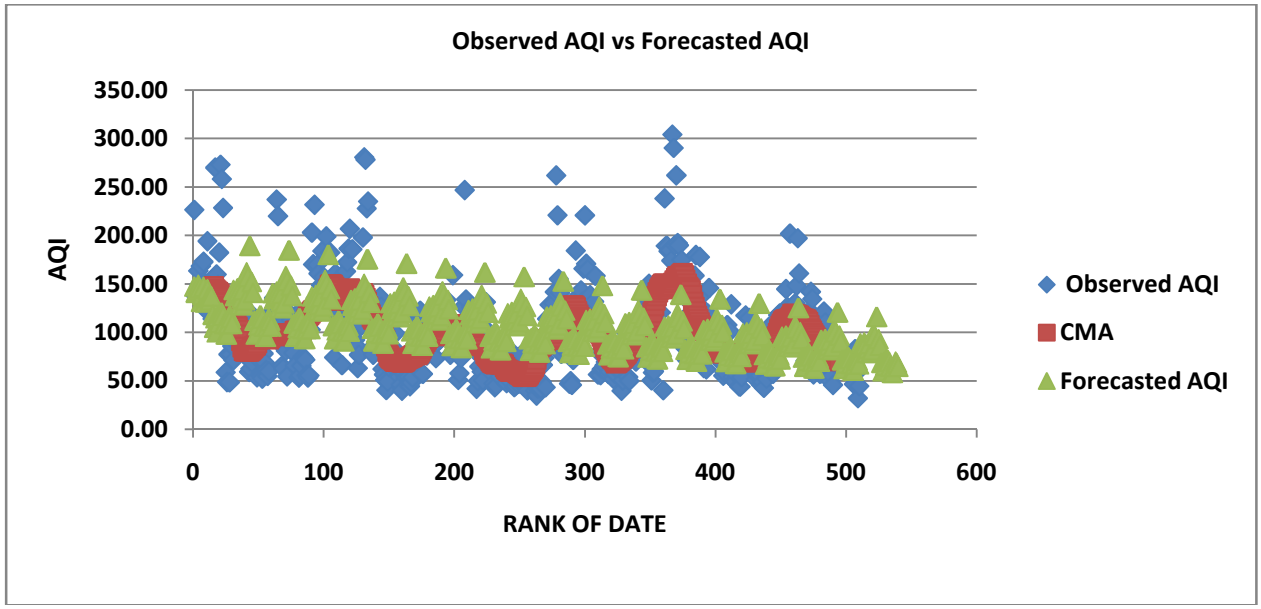

Figure 5: Comparison between observed AQI and Forecasted AQI for Delhi City.

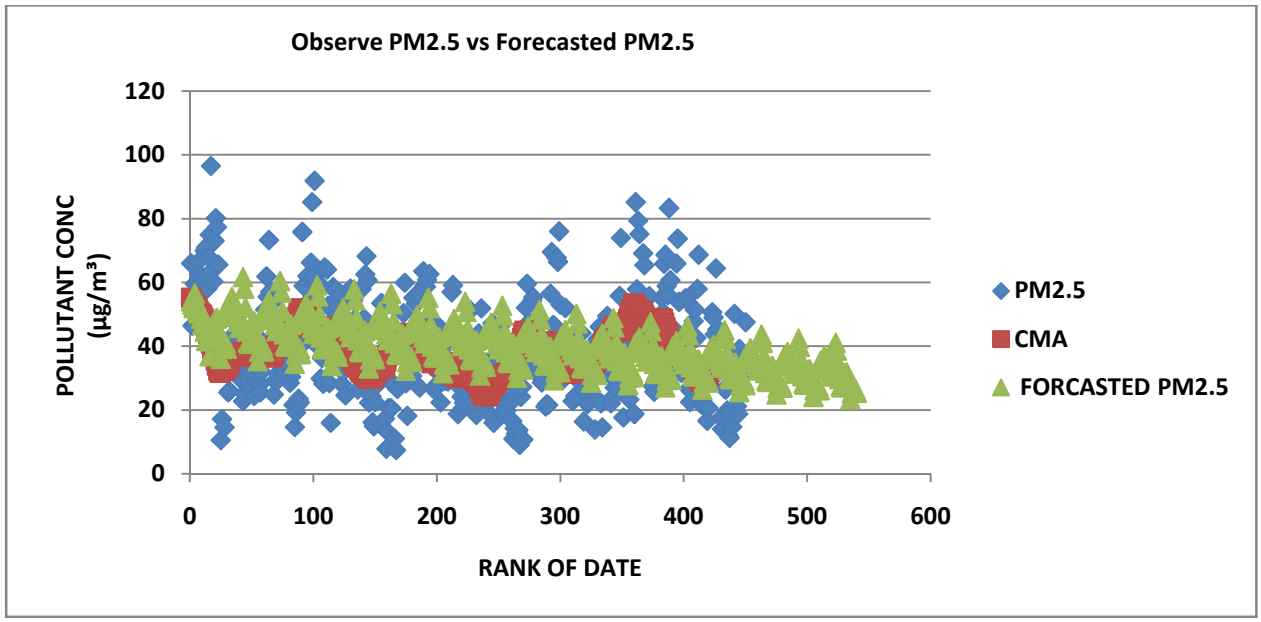

Figure 6: Comparison between observed PM2.5 and Forecasted PM2.5 for Delhi City.

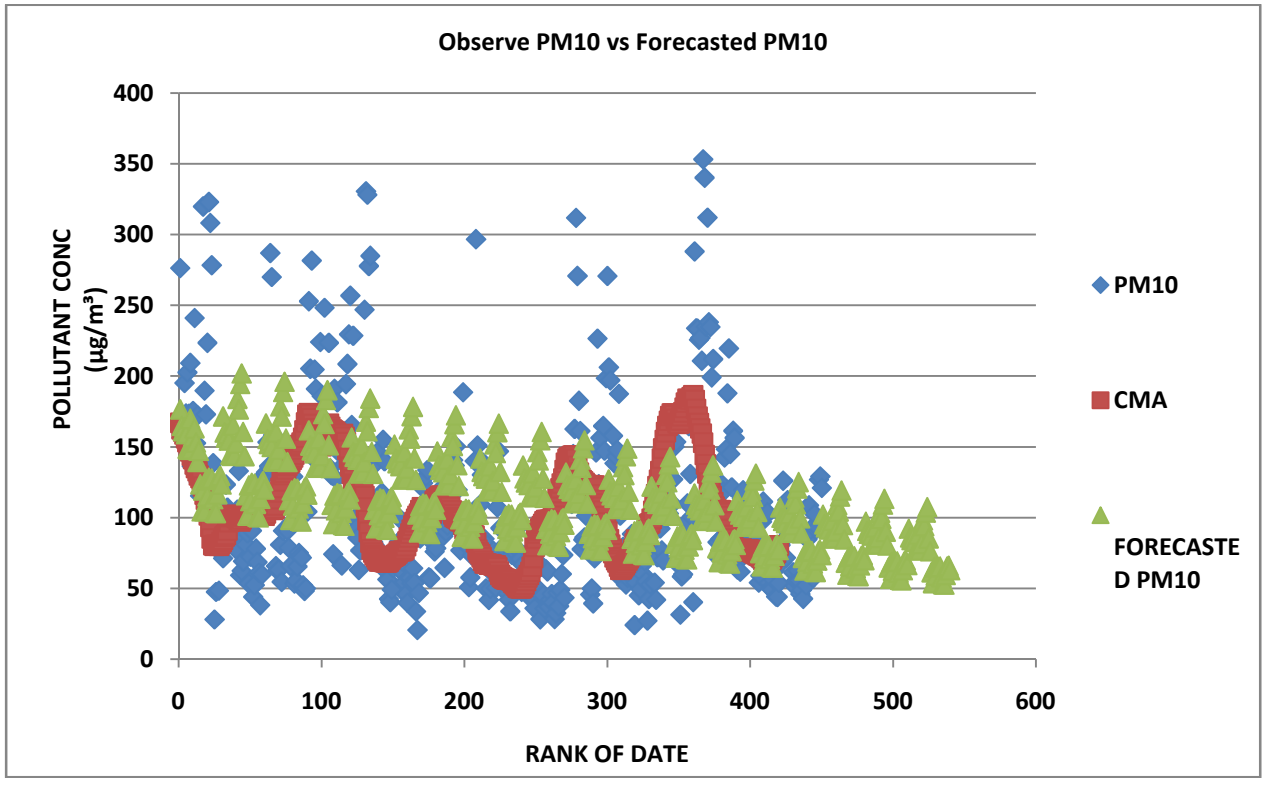

Figure 7: Comparison between observed PM10 and Forecasted PM10 for New Delhi City.

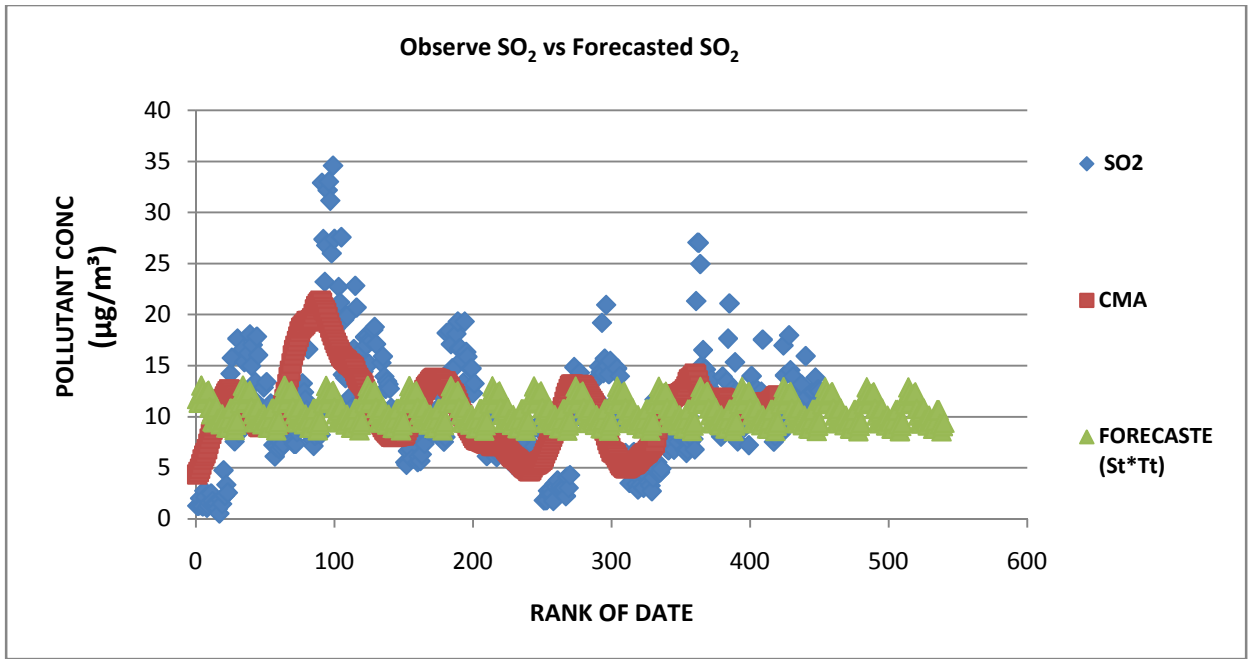

Figure 8: Comparison between observed SO_2 and Forecasted SO_2 for New Delhi City.

Table 1: Comparison of AQI obtained by various approaches For New Delhi

DATE	PM2.5	PM 10	$S0_2$	CO	NO_2	O_3	NH_3	AQI Actually Observed	AQI As Per Proposed Method	% Difference
01-08-2023	35.48	92.32	11.60	0.77	50.39	45.65	43.90	**97**	89	8.2
02-08-2023	34.79	86.15	11.24	0.65	44.38	48.78	44.11	**94**	84	10.6
03-08-2023	34.06	84.16	12.11	0.63	49.03	40.37	45.38	**90**	86	4.4
04-08-2023	37.04	85.48	12.89	0.68	51.76	44.50	42.43	**121**	89	26.4
05-08-2023	35.57	82.84	12.34	0.70	48.55	42.27	43.52	**94**	87	7.4
06-08-2023	33.39	77.49	11.43	0.67	47.05	41.18	46.45	**82**	79	3.7
07-08-2023	32.91	76.95	12.12	0.73	47.86	40.39	43.97	**56**	83	-48.2
08-08-2023	33.80	88.37	12.27	0.73	48.52	41.34	39.49	**97**	88	9.3
09-08-2023	31.83	77.10	12.40	0.68	46.60	37.18	38.88	**114**	78	31.6
10-08-2023	30.50	81.48	10.78	0.68	44.85	35.24	45.70	**123**	81	34.1
11-08-2023	31.88	98.18	10.13	0.68	46.30	30.96	44.25	**116**	99	14.7
12-08-2023	32.01	94.13	9.46	0.58	40.44	31.75	41.27	**121**	96	20.7
13-08-2023	41.09	103.63	9.36	0.61	42.88	30.34	36.53	**101**	115	-13.9
14-08-2023	38.67	107.47	10.52	0.63	42.29	33.54	37.15	**95**	115	-21.1

15-08-2023	34.67	85.44	10.82	0.58	38.14	35.09	37.64	**106**	93	12.3
16-08-2023	29.74	76.27	9.68	0.60	40.48	31.54	37.11	**111**	86	22.5
17-08-2023	27.45	61.49	9.55	0.61	43.27	30.96	37.16	**116**	68	41.4
18-08-2023	27.11	53.84	9.17	0.65	44.37	29.99	41.50	**133**	90	32.3
19-08-2023	30.72	65.12	10.11	0.69	46.59	32.99	40.39	**151**	99	34.4
20-08-2023	28.63	59.22	9.68	0.68	49.39	35.14	42.62	**93**	95	-2.2
21-08-2023	31.69	65.95	9.56	0.71	49.69	33.98	40.89	**66**	72	-9.1
22-08-2023	31.76	65.14	8.85	0.69	45.63	34.07	43.95	**83**	72	13.3
23-08-2023	29.31	65.54	8.83	0.68	41.20	32.20	43.74	**68**	70	-2.9
24-08-2023	27.50	57.97	9.67	0.67	45.33	30.38	44.00	**68**	65	4.4
25-08-2023	23.43	53.13	10.67	0.64	50.14	28.48	44.81	**93**	78	16.1
26-08-2023	25.42	52.83	10.67	0.68	43.82	27.62	43.49	**124**	88	29.0
27-08-2023	25.99	59.64	9.31	0.67	42.90	29.68	45.33	**158**	105	33.5
28-08-2023	26.25	65.54	8.60	0.64	44.15	27.32	43.88	**148**	110	25.7
29-08-2023	26.53	66.16	9.11	0.65	45.74	30.89	47.17	**161**	120	25.5
30-08-2023	25.44	62.97	9.43	0.65	44.41	25.05	45.60	**126**	100	20.6

Note:- The minus sign indicates the under estimation of the AQI value.

3.2 Results analysis for AQI obtained from machine learning techniques:-

Table 2 reflects the performance of linear regression algorithm used in modeling which have been used to evaluate the performance of the models. The figure 9 shows a scatter plot graph.

Table 2: Performance of linear algorithm used in modeling

	MAE	**RMSE**	**R2**
Linear Regression	22.3241	36.2575	0.8986

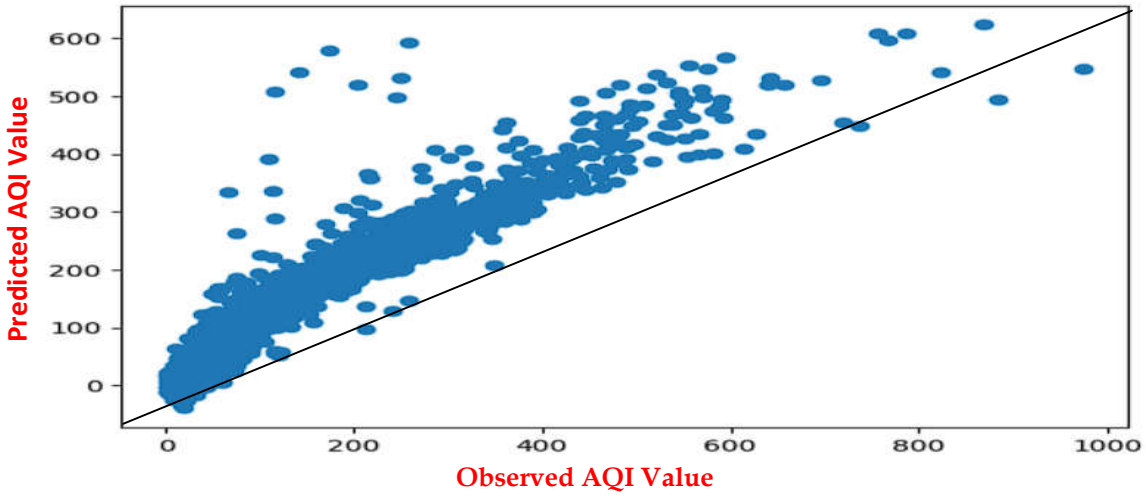

Figure 9: **Scatter plot for linear regression model**

The figure 10 shows the calculation of performance metrices.

```
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np
pred = linreg.predict(X_train)

print('Mean absolute error: {}'.format(mean_absolute_error((y_train), (pred))))
print('Root mean square error: {}'.format(np.sqrt(mean_squared_error((y_train), (pred)))))
print('R-squared: {}'.format(r2_score((y_train), (pred))))
```

```
Mean absolute error: 22.324152584932776
Root mean square error: 36.2575893810397
R-squared: 0.8986627507844618
```

Figure 10: **Calculation of performance metrices**

4. CONCLUSION

Based on the results of this analytical study, the following conclusions may be drawn:

- The accuracy of the proposed forecasting approach can further be developed by using more dataset and deep level regression technique.
- The data have been preprocessed and successfully trained the linear regression and Centre Line Moving Average.
- Based on the performance of the models, it can be concluded that the models Support Vector Regression have shown better accuracy among all whereas Ridge regression and Linear regression have shown better performance with lower MAE, RMSE and higher r-squared techniques.
- It can be concluded that data varies in a particular pattern in all summer, winter and Rainy season. In winter season air pollution is maximum and in rainy season the air pollution is minimum.
- In present study, the linear regression model has been adopted in CMA method, if Multiple linear regression model in CMA could enhance the accuracy level in forecasting the AQI.
- The proposed methods provide the safest path for the public during their journey depending on AQI value. This air quality information will help the people who want to travel specified routes and also identify the impacts on health if they travel in the polluted route where air quality is poor.

5. ACKNOWLEDGEMENT

The authors gratefully express their thanks to department of civil engineering, Bhagalpur College of Engineering, Bhagalpur, Bihar, India for providing necessary platform and support & a big thanks to Mr. Kaiser Azad, B.Tech Graduate, Bhagalpur College of Engineering, Bhagalpur and Mr. Diwakar Kumar, B.Tech Graduate, Bhagalpur College of Engineering, Bhagalpur for their sincere help which has been thoroughly required for carrying out this present research work.

REFERENCES

1. Bishoi, B. and Prakash, A. (2009). A comparative study of Air Quality Index based on factor analysis and US-EPA methods for an Urban Environment, Bishoi et al., Aerosol and Air Quality Research, 9(1), 1-17.
2. Chakraborty, M., Debnath, S., and Ghosh, S. (2021). A Study during Lockdown Period Based on AQI over Indian Mega cities during COVID-19. J. Phys.: Conf. Ser. 1797 012056. DOI: 10.1088/1742-6596/1797/1/012056
3. CPCB (2015). National Air Quality Index Report, Central Pollution Control Board, 2015.
4. Geetha, A., Ramya P. S., Sravani, C. and Ramesh, M. (2020). Real Time Air Quality Index from Various Locations International Journal of Recent Technology and Engineering (IJRTE), 9(2). DOI:10.35940/ijrte.B3493.079220

5. Guttikunda, S. (2010). Air Quality Index (AQI) for Delhi, India: Trend Analysis & Implications for the CWG 2010 and Beyond. Sarath Guttikunda SIM-air working paper series # 35-2010.

6. Kanchan, Gorai, A. K. and Goyal, P. (2015). A Review on Air Quality Indexing System. Asian Journal of Atmospheric Environment, 9(2), 101-113. DOI: http://dx.doi.org/10.5572/ajae.2015.9.2.101

7. Kumar, A. and Goyal, P. (2011). Forecasting of air quality in Delhi using principal component regression technique, Atmospheric Pollution Research 2 (2011) 436-444. DOI: https://doi.org/ 10.5094/APR.2011.050

8. Kumar, K. and Pande, B. P. (2022). Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science and Technology. DOI: https://doi.org/10.1007/s13762-022-04241-5

9. Lemeš, S. (2018). Air quality index (AQI) – comparative study and assessment of an appropriate model for B&H'. 12th Scientific/Research Symposium with International Participation, metallic and nonmetallic materials. B&H, 19th-20th April 2018.

10. Mamta, P. and Bassin, J. K. (2010). Analysis of ambient air quality using air quality index – a case study.IJAET/Vol.I/ Issue II/July-Sept.,2010/106-114.

11. Soundari, A. G., Jeslin, J. G., and Akshaya. (2019). Indian Air Quality Prediction And Analysis Using Machine Learning. International Journal of Applied Engineering Research, 14(11), (Special Issue).

12. Taieb, D. and Brahim, A. B. (2013).Methodology for developing an air quality index (AQI) for Tunisia. Int. J. Renewable Energy Technology, 4(1), 2013. DOI: 10.1504/IJRET.2013.051067

****