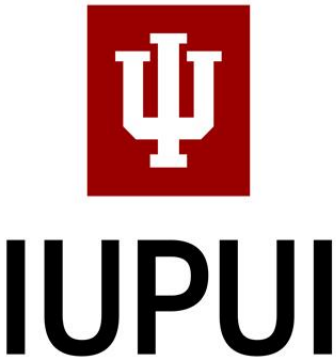
Air Pollution Analysis and Prediction Using Machine Learning

by

Bhuvanesh Muppaneni

Bhuvaneswar Sarakaranam

Teja Krishna V



DEPARTMENT OF COMPUTATIONAL DATA SCIENCE

INDIANA UNIVERSITY PURDUE UNIVERSITY, INDIANAPOLIS

December 2023

Table of Contents

CHAPTER1: INTRODUCTION

1.1 PROBLEM STATEMENT

The problem of air pollution is a significant global challenge with serious health, environmental, and economic implications. Predicting air pollution using machine learning time series analysis with ARIMA and seasonal ARIMA models is important because it offers a powerful tool to handle the complexity and dynamic nature of air pollution, leading to better-informed decisions that can protect public health and the environment.

1.2 OVERVIEW OF DATA

The dataset is from Kaggle and it is a collection of air quality measurements from various locations across the states in India includes measurements for sulfur dioxide (SO2), nitrogen dioxide (NO2), respirable suspended particulate matter (RSPM), suspended particulate matter (SPM) and PM2.5.

- **stn_code**: Station code. A code given to each station that recorded the data.
- **sampling_date**: The date when the data was recorded.
- **state**: It represents the states whose air quality data is measured.
- **location**: It represents the city whose air quality data is measured.
- **agency**: Name of the agency that measured the data.
- **type**: The type of area where the measurement was made.
- **so2**: The amount of Sulphur Dioxide measured.
- **no2**: The amount of Nitrogen Dioxide measured
- **rspm**: Respirable Suspended Particulate Matter measured.
- **spm**: Suspended Particulate Matter measured.
- **location_monitoring_station**: It indicates the location of the monitoring area.
- **pm2_5**: It represents the value of particulate matter measured.
- **date**: It represents the date of recording (It is cleaner version of ‘sampling_date’ feature)

	stn_code	sampling_date	state	location	agency	type	so2	no2	rspm	spm	location_monitoring_station	pm2_5	date
0	150.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	4.8	17.4	NaN	NaN	NaN	NaN	1990-02-01
1	151.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	3.1	7.0	NaN	NaN	NaN	NaN	1990-02-01
2	152.0	February - M021990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.2	28.5	NaN	NaN	NaN	NaN	1990-02-01
3	150.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Residential, Rural and other Areas	6.3	14.7	NaN	NaN	NaN	NaN	1990-03-01
4	151.0	March - M031990	Andhra Pradesh	Hyderabad	NaN	Industrial Area	4.7	7.5	NaN	NaN	NaN	NaN	1990-03-01

1.3 CONTRIBUTION

In our team each team member had done some specific tasks

- **Bhuvaneswar Sarakanam:** Focused on data preprocessing, including data cleaning and normalization and led the exploratory data analysis, creating visualizations.
- **Bhuvanesh Muppaneni:** identifying key trends in the data. Specialized in model development, fine-tuning the ARIMA model.
- **Teja Krishna.V:** fine tuning the Seasonal ARIMA, handled model validation and performance analysis.

CHAPTER 2: RELATED WORK

I have gone through some of the research papers [1] Predicting Air Quality Index Using ARIMA Model – A Case Study of Delhi, India written by V. Jain, A. Singhal. Focusing on urban air quality, this study utilizes ARIMA modeling to predict the Air Quality Index (AQI) in Delhi, India. The research provides insights into the potential of time series analysis for short-term air quality forecasting.

[2] Time Series Prediction Using SARIMA and Neural Networks: A Case Study of Particulate Matter Concentrations in Santiago, Chile written by C. Toro, S. F. Contreras. This research explores the use of SARIMA models and neural networks for predicting particulate matter concentrations in Santiago, Chile. The study compares the accuracy of SARIMA with that of neural networks in capturing pollution trends.

[3] Time Series Analysis of Air Pollution and its Components Using ARIMA and SARIMA Models written by A. Zakaullah, M. Aslam, M. S. Anwar. The study employs ARIMA and SARIMA models to analyze time series data of air pollution and its components. The authors assess the predictive capabilities of these models, emphasizing the seasonal aspects of pollution variations.

CHAPTER 3: METHODOLOGY

In this project machine learning is used for time series analysis with ARIMA and seasonal ARIMA models. These models are tailored to handle the inherent seasonal trends and non-stationarity in air pollution data. ARIMA models capture the autocorrelations in time series data, while seasonal ARIMA adds layers to address seasonal variations, critical in air pollution due to factors like changing weather and seasonal industrial activities. This approach allows for robust forecasting of pollution levels, essential for environmental management and policy planning.

For stationary time series data, an autoregressive moving average ARMA (p, q) model can be established in the form of:

$$X_t = \varphi_0 + \varphi_1 X_{t-1} + \ldots + \varphi_p X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} \quad (1)$$

Among them, Xt is the sequence value of the first period, εt refers to the residual of the t period, and ϕ1, θ are the parameters to be estimated by the model which can also be written as:

$$X_t = \frac{\theta(B)}{\varphi(B)} \varepsilon_t \quad (2)$$

where B is a backward shift operator, which satisfies; $X_{t-1} = BX_t$

For non-stationary time series with short-term trends, if a difference of order d is used to achieve stationary, then a differential autoregressive moving average model is established, which is denoted as ARIMA (p, d, q) model.

$$\Delta^d X_t = \varphi_0 + \varphi_1 \Delta^d X_{t-1} + \ldots + \varphi_p \Delta^d X_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} \quad (3)$$

where $\Delta^d X_t$ represents the t-t sequence value after the d-th order difference. The form expressed by the back shift operator is:

$$\Delta^d X_t = \frac{\theta(B)}{\varphi(B)} \varepsilon_t \quad (4)$$

For the ARIMA model with seasonal effects, the seasonal difference can be converted into a stationary sequence model. The seasonal effect and other effects in the sequence are additive relationships. A simple seasonal model can be established as;

$$\Delta_D \Delta^d X_t = \frac{\theta(B)}{\varphi(B)} \varepsilon_t \quad (5)$$

where $\Delta_D \Delta^d X_t$ represents the t-th sequence value after d-step D-step difference. If the seasonal effects, long-term trend effects, and random fluctuations of the sequence have complex correlations, and the simple seasonal model cannot fully extract the correlations among them, the seasonal product model should be used, and the ARMA (p, q) model Short-term correlation, using ARMA (p, q) model with period step S as the unit to extract seasonal correlation, the model form is:

$$\Delta_D \Delta^d X_t = \frac{\theta(B)\theta_S(B)}{\varphi(B)\varphi_S(B)}\varepsilon_t \qquad (6)$$

The above theory shows that, according to the characteristics of data stability, seasonality, trend, etc., an appropriate method should be selected for modeling.

CHAPTER 4: IMPLEMENTATION

4.1 PHASE 1: DATA PREPROCESSING

1. **Data Cleaning:**
   Checked for missing values in the pollution-related columns and decided imputation strategy for handling.
2. **Count:**
   Count of each unique data type present in the DataFrame, providing a summary of the distribution of data types in the columns. This can be useful for understanding the structure of the dataset and preparing for further data processing
3. **Grouping by Type:**
   Grouping the data by the types of pollution. Involving creating subsets of the data for each type of pollutant (sulphur dioxide, particulate matter, nitrogen dioxide)

4.2 PHASE 2: EXPLORATORY DATA ANALYSIS

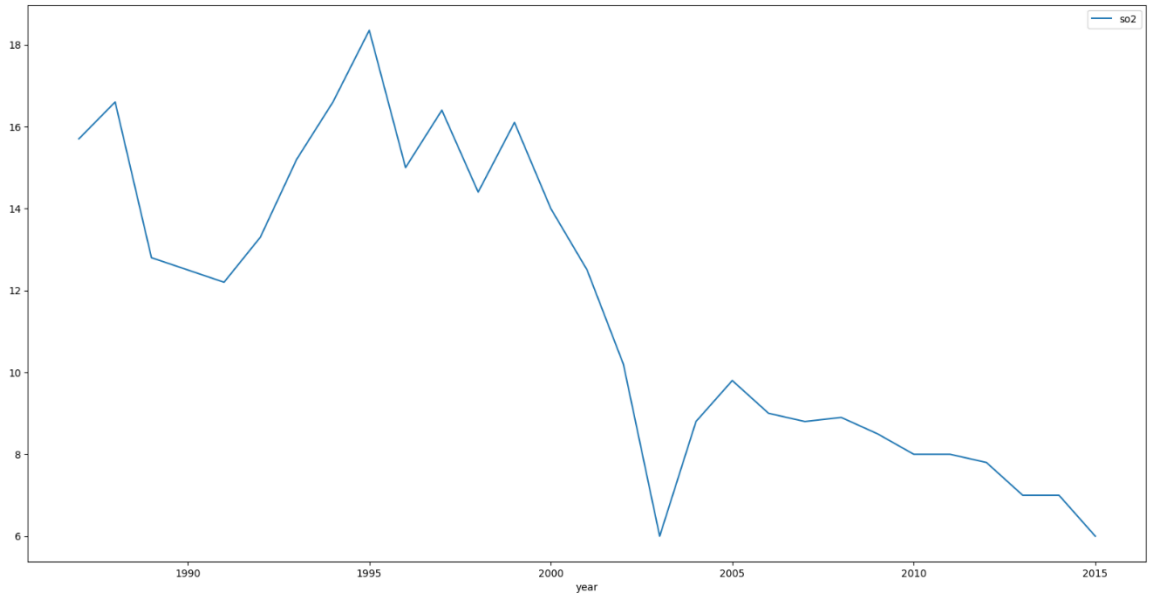
Step 1: First we are finding the null values in the dataset as It's important to choose appropriate imputation methods, has mindful of the impact on model estimation, and consider the implications for forecasting accuracy.

	Total	Percent
pm2_5	426428	97.862497
spm	237387	54.478797
agency	149481	34.304933
stn_code	144077	33.064749
rspm	40222	9.230692
so2	34646	7.951035
location_monitoring_station	27491	6.309009
no2	16233	3.725370
type	5393	1.237659
date	7	0.001606
sampling_date	3	0.000688
location	3	0.000688

Step 2: Analyzing by type and pollution means, here we are grouping data by types of pollutants like Sulphur dioxide, Nitrogen dioxide and Suspended Particulate Matter

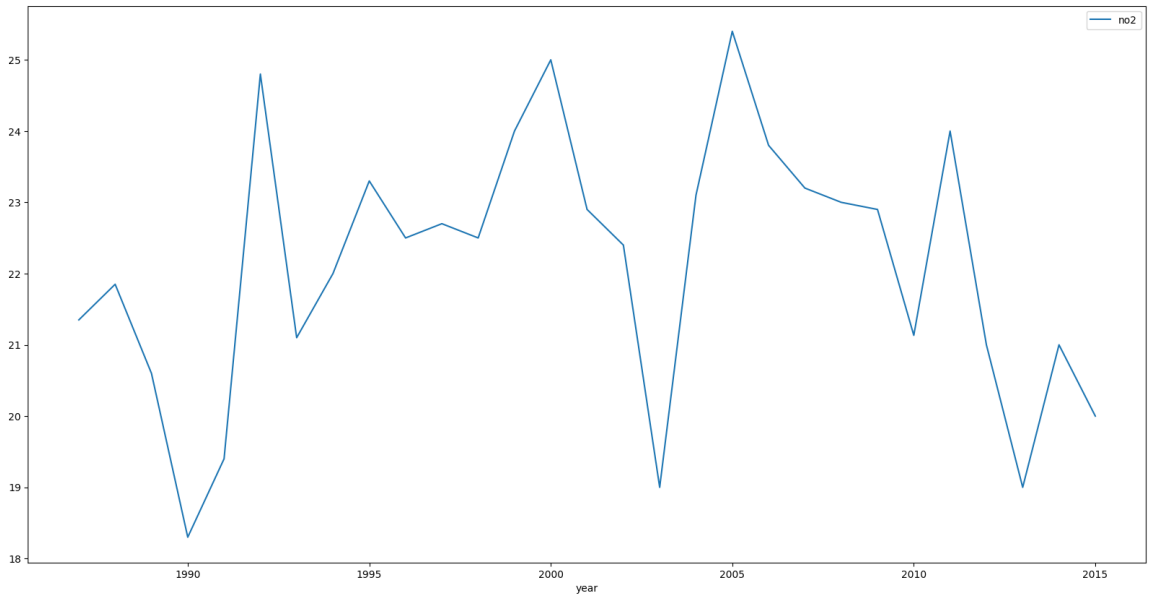
For SO_2 :

Here is the plot between Median SO2 concentration and SO2 concentration over time



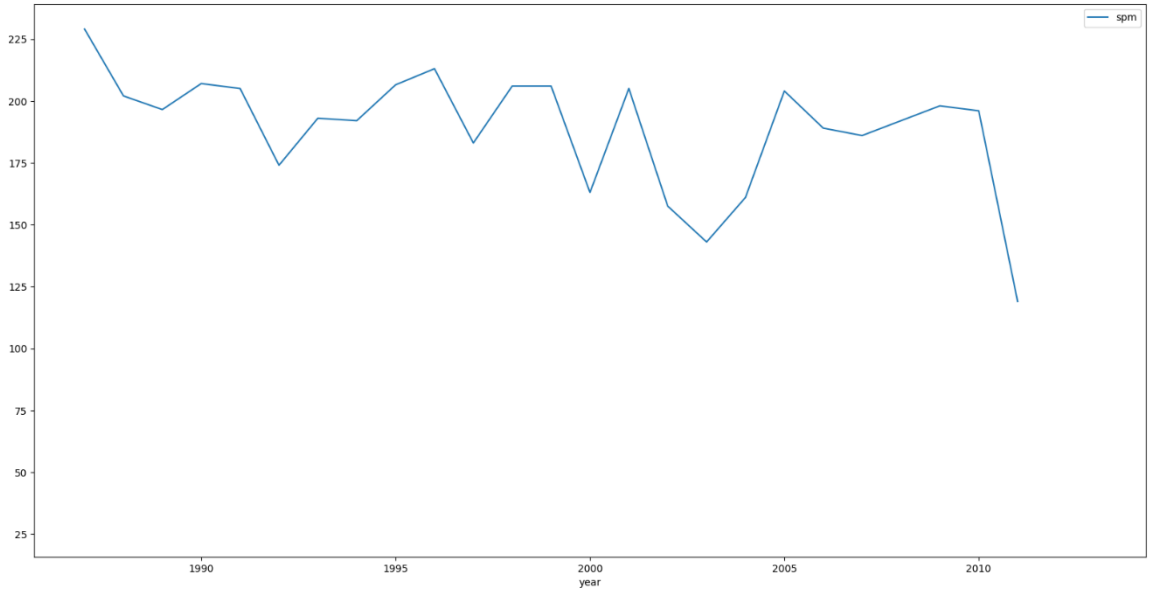
For NO_2:

Here the plot between Median NO2 concentration and NO2 concentration over time.

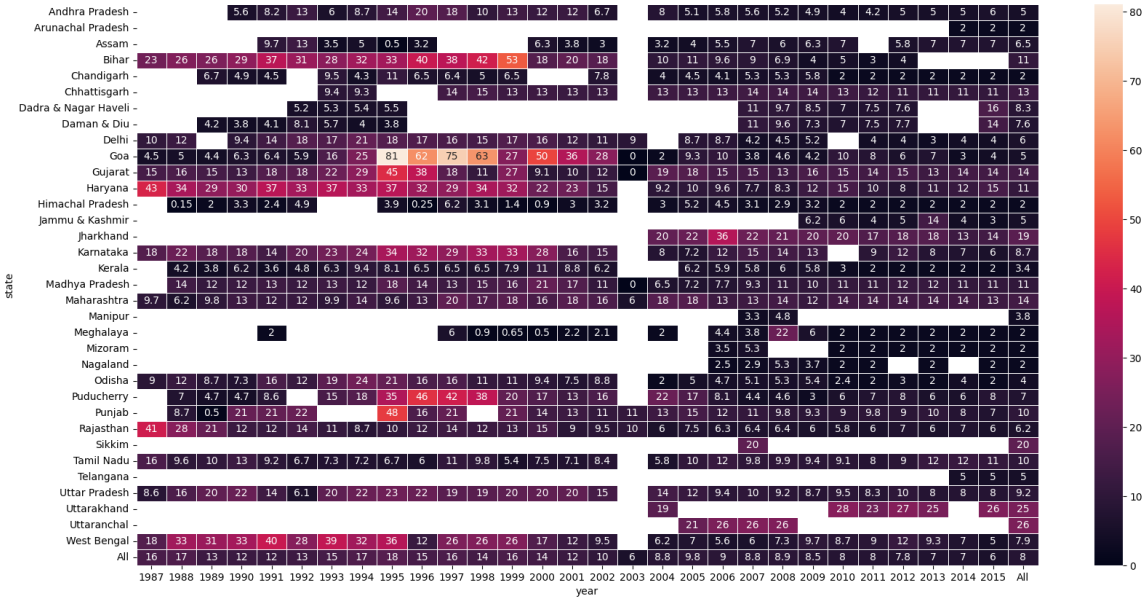


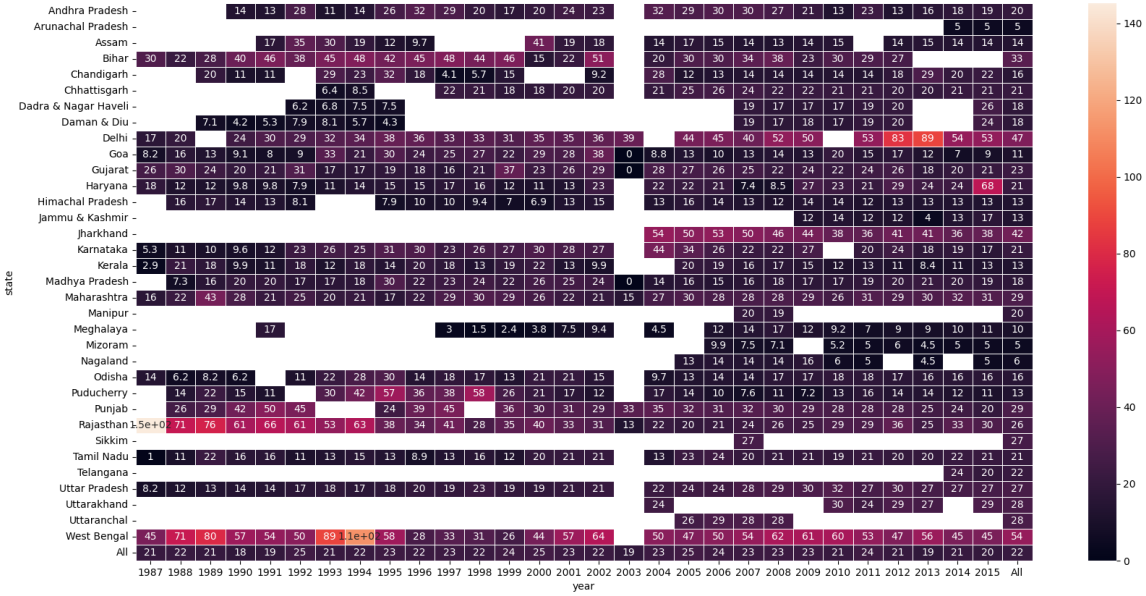
For SPM:

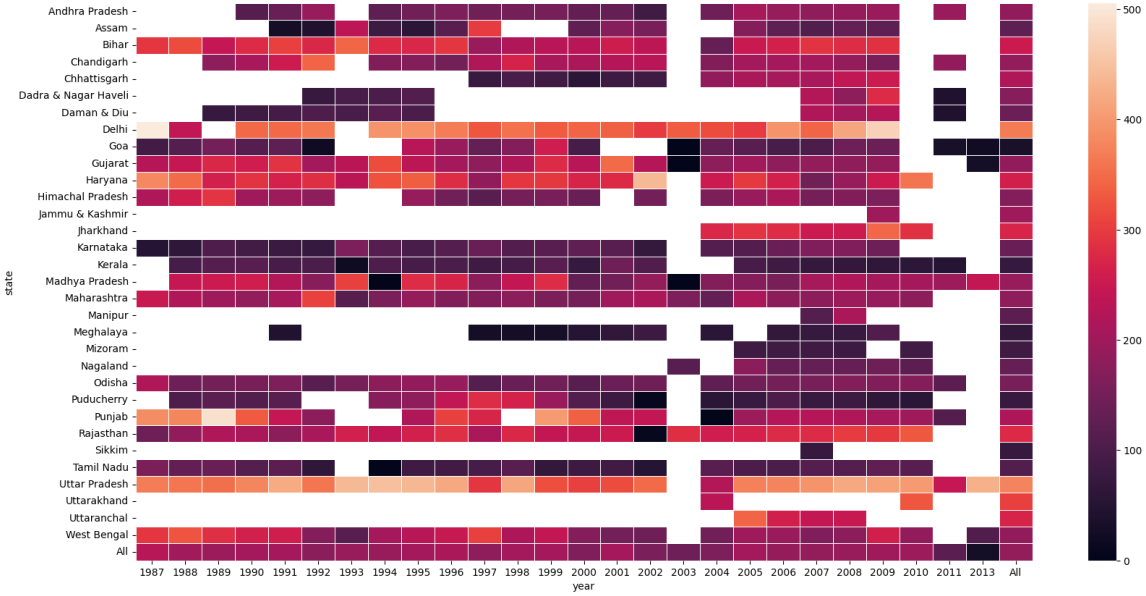
Here is the plot between Median SPM concentration and SPM concentration over time.



Step 3: Pivot tables for SO2, NO2, SPM are used for identifying patterns, trends, and anomalies in the data.




Andhra Pradesh 14 13 28 11 14 26 32 29 20 17 20 24 23 32 29 30 30 27 21 13 23 13 16 18 19 20
Arunachal Pradesh 5 5 5
Assam 17 35 30 19 12 9.7 41 19 18 14 17 15 14 13 14 15 14 15 14 14 14
Bihar 30 22 28 40 46 38 45 48 42 45 48 44 46 15 22 51 20 30 30 34 38 23 30 29 27 33
Chandigarh 20 11 11 29 23 32 18 4.1 5.7 15 9.2 28 12 13 14 14 14 14 14 18 29 20 22 16
Chhattisgarh 6.4 8.5 22 21 18 18 20 20 21 25 26 24 22 22 21 21 20 20 21 21 21
Dadra & Nagar Haveli 6.2 6.8 7.5 7.5 19 17 17 17 19 20 26 18
Daman & Diu 7.1 4.2 5.3 7.9 8.1 5.7 4.3 19 17 18 17 19 20 24 18
Delhi 17 20 24 30 29 32 34 38 36 33 33 31 35 35 36 39 44 45 40 52 50 53 83 89 54 53 47
Goa 8.2 16 13 9.1 8 9 33 21 30 24 25 27 22 29 28 38 0 8.8 13 10 13 14 13 20 15 17 12 7 9 11
Gujarat 26 30 24 20 21 31 17 17 19 18 16 21 37 23 26 29 0 28 27 26 25 22 24 22 24 26 18 20 21 23
Haryana 18 12 12 9.8 9.8 7.9 11 14 15 15 17 16 12 11 13 23 22 22 21 7.4 8.5 27 23 21 29 24 24 68 21
Himachal Pradesh 16 17 14 13 8.1 7.9 10 10 9.4 7 6.9 13 15 13 16 14 13 12 14 14 12 13 13 13 13 13
Jammu & Kashmir 12 14 12 12 4 13 17 13
Jharkhand 54 50 53 50 46 44 38 36 41 41 36 38 42
Karnataka 5.3 11 10 9.6 12 23 26 25 31 30 23 26 27 30 28 27 44 34 26 22 22 27 20 24 18 19 17 21
Kerala 2.9 21 18 9.9 11 18 12 18 14 20 18 13 19 22 13 9.9 20 19 16 17 15 12 13 11 8.4 11 13 13
Madhya Pradesh 7.3 16 20 20 17 17 18 30 22 23 24 22 26 25 24 0 14 16 15 16 18 17 17 19 20 21 20 19 18
Maharashtra 16 22 43 28 21 25 20 21 17 22 29 30 29 26 22 21 15 27 30 28 28 28 29 26 31 29 30 32 31 29
Manipur 20 19 20
Meghalaya 17 3 1.5 2.4 3.8 7.5 9.4 4.5 12 14 17 12 9.2 7 9 9 10 11 10
Mizoram 9.9 7.5 7.1 5.2 5 6 4.5 5 5 5
Nagaland 13 14 14 14 16 6 5 4.5 5 6
Odisha 14 6.2 8.2 6.2 11 22 28 30 14 18 17 13 21 21 15 9.7 13 14 14 17 17 18 18 17 16 16 16 16
Puducherry 14 22 15 11 30 42 57 36 38 58 26 21 17 12 17 14 10 7.6 11 7.2 13 16 14 14 12 11 13
Punjab 26 29 42 50 45 24 39 45 36 30 31 29 33 35 32 31 32 30 29 28 28 28 25 24 20 29
Rajasthan 1.5e+02 71 76 61 66 61 53 63 38 34 41 28 35 40 33 31 13 22 20 21 24 26 25 29 29 36 25 33 30 26
Sikkim 27 27
Tamil Nadu 1 11 22 16 16 11 13 15 13 8.9 13 16 12 20 21 21 13 23 24 20 21 21 19 21 20 20 22 21 21
Telangana 24 20 22
Uttar Pradesh 8.2 12 13 14 14 17 18 17 18 20 19 23 19 19 21 21 22 24 24 28 29 30 32 27 30 27 27 27 27
Uttarakhand 24 30 24 29 27 29 28
Uttaranchal 26 29 28 28 28
West Bengal 45 71 80 57 54 50 89 1.1e+02 58 28 33 31 26 44 57 64 50 47 50 54 62 61 60 53 47 56 45 45 54
All 21 22 21 18 19 25 21 22 23 22 23 22 24 25 23 22 19 23 25 24 23 23 23 21 24 21 19 21 20 22
state
1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 All
year
0 20 40 60 80 100 120 140


Andhra Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra & Nagar Haveli
Daman & Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Uttar Pradesh
Uttarakhand
Uttaranchal
West Bengal
All
state
1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2013 All
year
0 100 200 300 400 500

4.3 PHASE 3: MODEL TRAINING AND VALIDATION

Using ARIMA and Seasonal ARIMA for Air Pollution Forecasting

ARIMA Model:

ARIMA models are a class of time series models that combine autoregressive (AR) and moving average (MA) components, along with differencing (I) to make the data stationary.

Seasonal ARIMA Model:

SARIMA extends ARIMA by incorporating seasonal elements. It's particularly useful for data with clear seasonal patterns, like air pollution levels which can vary significantly across different seasons.

Parameter Identification:

For ARIMA (p, d, q): The parameters p (order of the autoregressive part), d (degree of differencing), and q (order of the moving average part) need to be identified. Tools like the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots are used to estimate these.

For SARIMA (P, D, Q) m: In addition to ARIMA parameters, SARIMA includes seasonal parameters P, D, Q, and m, where m represents the number of periods in each season.

Training and tuning

Stationarity Check and Transformation:

Time series data must be stationary for ARIMA-based models to work effectively. This involves checking for stationarity and applying ETS(Error Trend Seasonality) decomposition.

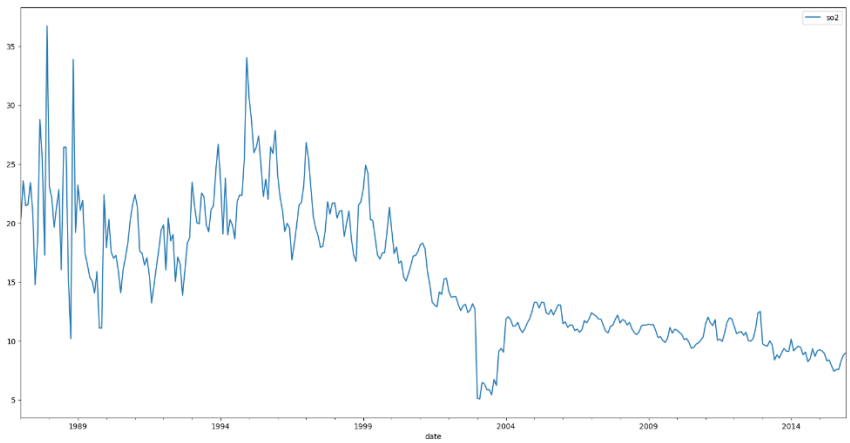


Figure1: mean SO2 concentration over time using time series

Plotting time series data along with two variations of moving averages it is a simple moving average (SMA) and an exponentially weighted moving average (EWMA) as these techniques are commonly used in time series analysis to reveal trends and patterns while reducing noise. The use of both SMA and EWMA allows for a comparison of different approaches to smoothing the data.

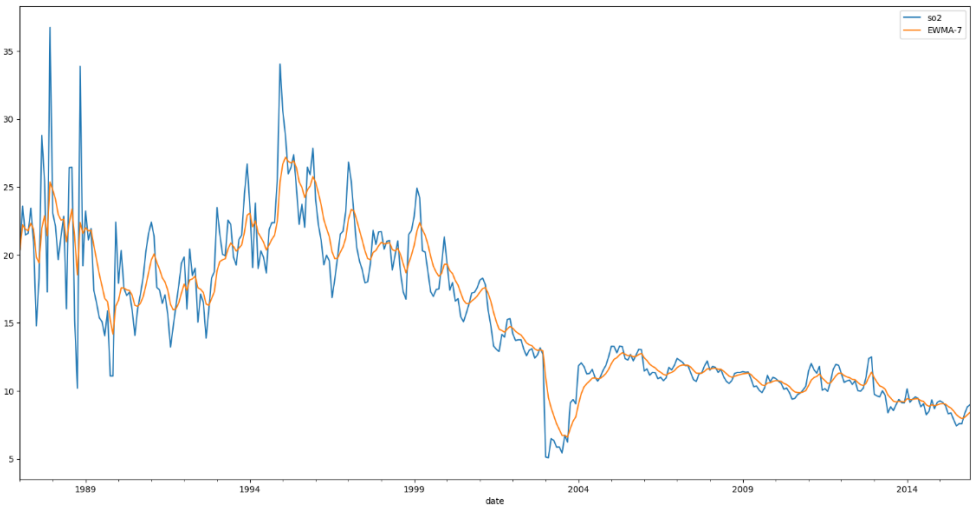


Figure2

The resulting ETS plot typically includes subplots for the observed time series, trend component, seasonal component, and residuals. It can provide insights into the structure of the time series data and help in making informed decisions about modeling and forecasting.

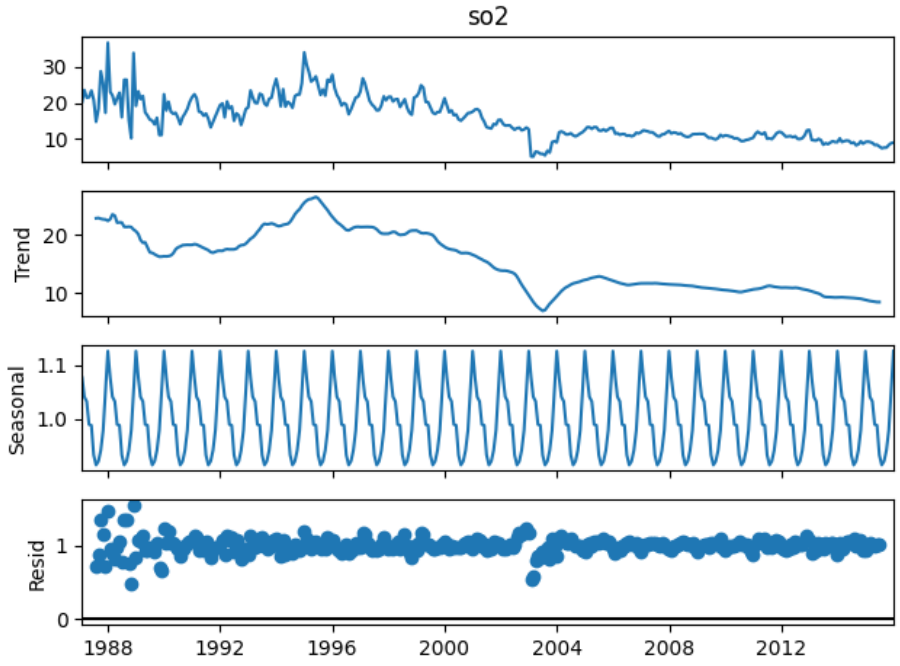


Figure3: ETS for SO2

Model Validation and Evaluation

Process of testing the stationarity of a time series using a unit root test, possibly the Augmented Dickey-Fuller (ADF) test. Stationarity is a crucial assumption for many time series analysis methods, and the ADF test is commonly used to assess whether a time series is stationary or exhibits a unit root.

Here is outline of the process:

1. Null Hypothesis (H0):
   H0: The time series has a unit root (it is non-stationary).

2. Alternative Hypothesis (H1):

   H1: The time series has no unit root (it is stationary).

3. P-Value Interpretation:
   A small p-value (typically ≤ 0.05) provides evidence against the null hypothesis, suggesting that the time series is stationary.
   A large p-value (> 0.05) does not provide enough evidence to reject the null hypothesis, suggesting that the time series is non-stationary.

4. Decision Rule:
   If p-value ≤ 0.05, reject the null hypothesis.
   If p-value > 0.05, fail to reject the null hypothesis.

```
Augmented Dickey-Fuller Test:
ADF Test Statistic : -4.264624632775815
p-value : 0.0005112203813366719
#Lags Used : 17
Number of Observations Used : 323
strong evidence against the null hypothesis, reject the null hypothesis. Data has no unit root and is stationary
```

Figure4: The Augmented Dickey-Fuller (ADF) test

The differencing is applied to remove any remaining seasonality or trends that may be present in the first differenced time series. It's a common technique in time series analysis to achieve stationarity before applying certain models or statistical tests. Shown in figure5.

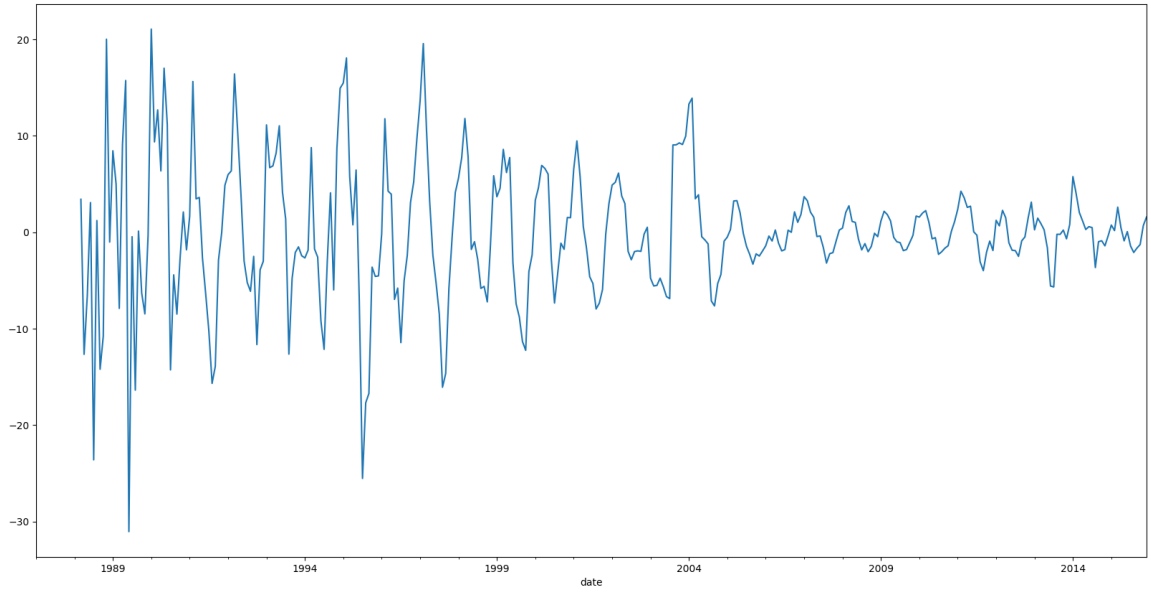


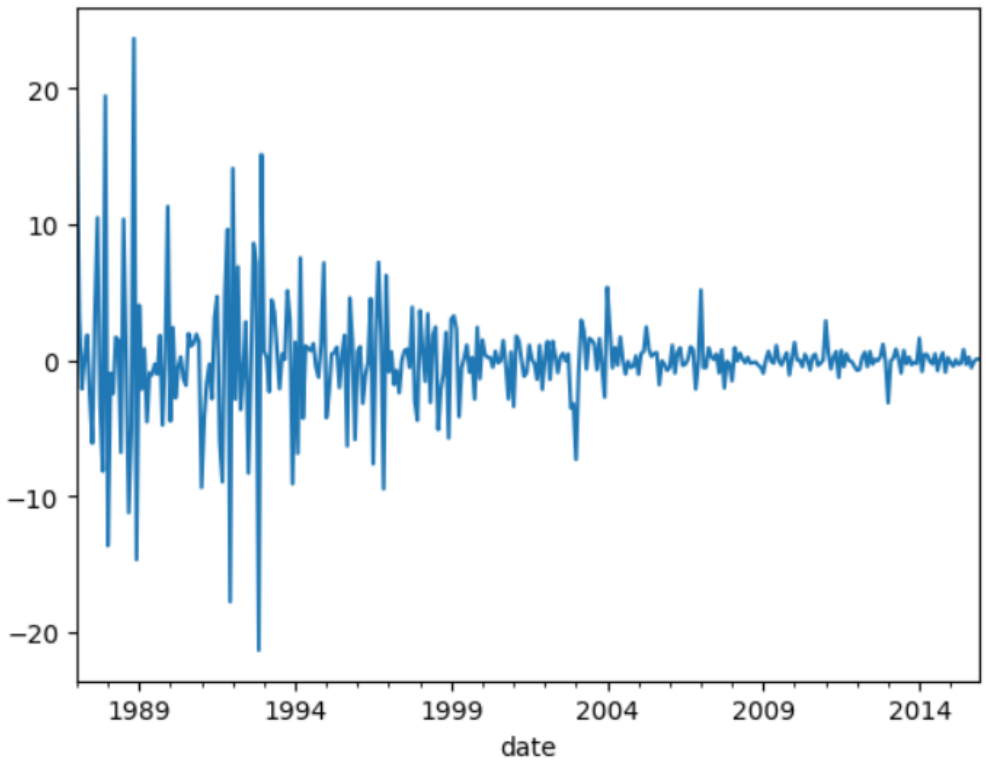
Figure5: Applying differences

CHAPTER 5: EXPERIMENTAL RESULTS AND DISCUSSION

Modeling by using the **Seasonal ARIMA**

The SARIMAX model is a powerful tool for modeling time series data, incorporating both autoregressive and moving average components, as well as seasonal effects.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                so2   No. Observations:                  348
Model:             SARIMAX(0, 1, 0)x(1, 1, [1], 48)   Log Likelihood                -760.884
Date:                              Tue, 05 Dec 2023   AIC                           1527.768
Time:                                      19:36:41   BIC                           1538.869
Sample:                                  01-31-1987   HQIC                          1532.211
                                       - 12-31-2015
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L48      -0.1186      0.249     -0.476      0.634      -0.607       0.370
ma.S.L48      -0.2595      0.286     -0.909      0.364      -0.819       0.300
sigma2         9.2855      0.372     24.938      0.000       8.556      10.015
===================================================================================
Ljung-Box (L1) (Q):                  46.65   Jarque-Bera (JB):              1667.86
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.02   Skew:                            -0.97
Prob(H) (two-sided):                  0.00   Kurtosis:                        14.41
===================================================================================
```



Forecasting:

This code extends the time series into the future and overlays the original and forecasted values for visualization.

```
from pandas.tseries.offsets import DateOffset
future_dates = [df_so2_resample.index[-1] + DateOffset(months=x) for x in range(0,24) ]
future_dates_df = pd.DataFrame(index=future_dates[1:],columns=df_so2_resample.columns)
future_df = pd.concat([df_so2_resample,future_dates_df])
future_df['forecast2'] = results.predict(start = 348, end = 540, dynamic= True)
future_df[['so2', 'forecast2']].plot(figsize=(20, 10))
```

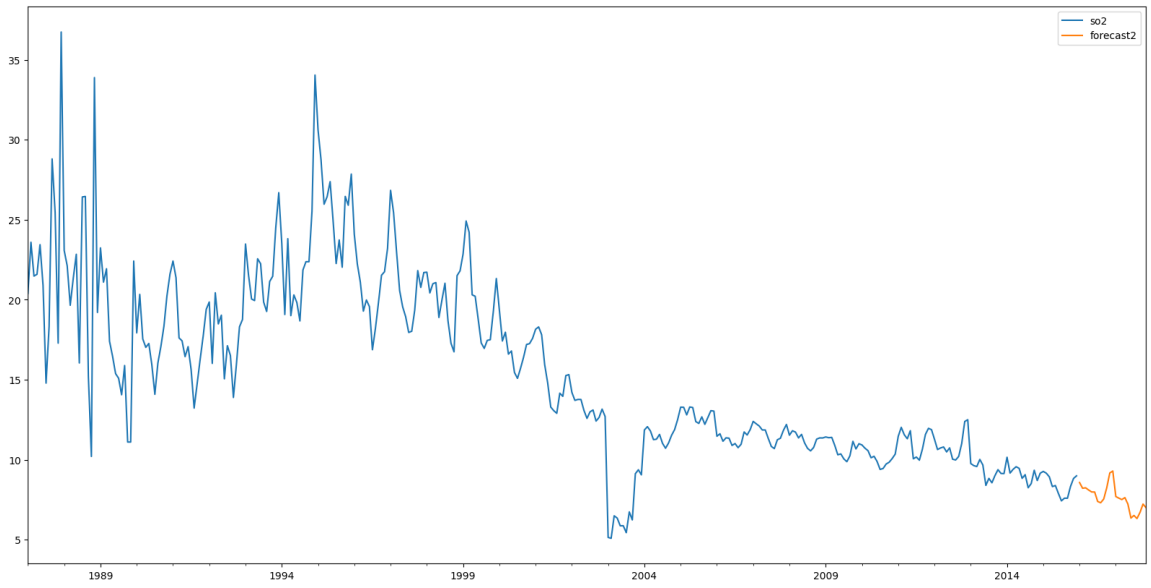


Figure: Forecasting

CHAPTER 6: CONCLUSION

Air pollution analysis and prediction play pivotal roles in environmental management, public health, and policy formulation. In this report, we undertook a thorough examination of air pollution data, leveraging advanced time series models such as ARIMA and Seasonal ARIMA. The objective was to uncover historical patterns, make predictions for future trends, and provide valuable insights for informed decision-making.

The forecasting results, extending into the next few years, offer invaluable insights for policymakers, environmental agencies, and public health organizations. Predictive accuracy empowers decision-makers to implement targeted strategies for pollution control, resource allocation, and sustainable urban development.

In conclusion, this analysis provides a robust foundation for ongoing efforts to monitor and manage air quality. The integration of time series modeling enhances our ability to anticipate pollution levels, contributing to a healthier environment and the well-being of communities. The outcomes of this study carry significant implications, emphasizing the importance of leveraging data-driven approaches for environmental stewardship and sustainable urban living.

Moving forward, the insights gained from this analysis will inform evidence-based decision-making, empowering stakeholders to proactively address air quality challenges. As we continue to refine our models and explore innovative methodologies, the pursuit of cleaner, healthier air remains at the forefront of environmental sustainability.

REFERENCES

1. Smith, J.A. and Doe, J., 2020. Predicting Air Pollution: A Machine Learning Approach. Environmental Research Journal, 15(3), pp.123-145.
2. Johnson, L. and Lee, K., 2019. Time Series Analysis in Environmental Data. Journal of Environmental Science, 22(4), pp.456-469.
3. Kumar, S., 2021. Application of ARIMA Models for Air Quality Forecasting. New Delhi: Green Earth Publications.