Time Series Analysis of Coalfield Pollutant Data

Arqam Patel

Table of contents

Introduction **2**

Dataset description **2**

Descriptive statistics . . . 2

NA data . . . 3

Exploratory analysis **4**

Weekly trajectory . . . 4

Sampled day . . . 5

Cross sectional analysis . . . 5

Interpolation **6**

Characteristics of time series **8**

Stationarity . . . 8

ACFs and PACFs . . . 9

Forecasting **12**

Evaluation . . . 12

Modelling techniques . . . 12

Best models . . . 13

Conclusion **16**

References **16**

Introduction

Despite a push towards green energy, coal is still one of the top sources of energy in India. Indian coal production is dependent on open-pit mining, and this is a cause of significant air pollution due to release of various particulate and gaseous pollutants. I undertake an analysis of the regular readings of various pollutants as time series as part of the assignment of the course EE798 (Foundations of Statistical Inference and Automation), instructed by Dr. Tushar Sandhan.

Due to the number of variables we're dealing with simultaneously, a majority of the analysis was done with the help of this dashboard app that I coded from scratch using an R library called RShiny. It contains detailed plots about the interpolation. forecasting and other analysis techniques used and is a very useful companion to this report.

Dataset description

The given dataset contains 8640 observations each of the levels of 10 pollutants, taken at intervals of 15 minutes from 1 February to 2 May 2023 (i.e. 90 days, 96 observations each). We can view the first few rows of the dataset:

```
             DateTime PM10 PM2.5 NO  NO2  NOx   CO SO2  NH3 Ozone Benzene
1 2023-02-01 00:00:00   95    35 NA 90.1 56.2 0.31  NA 17.7  28.1     0.4
2 2023-02-01 00:15:00   95    35 NA 88.0 55.1 0.33  NA 18.3  27.1     0.4
3 2023-02-01 00:30:00   95    35 NA 87.7 55.2 0.38  NA 19.7  24.9     0.4
4 2023-02-01 00:45:00  122    34 NA 88.9 55.7 0.38  NA 21.3  21.9     0.4
5 2023-02-01 01:00:00  122    34 NA 90.0 55.8 0.38  NA 22.3  16.7     0.4
6 2023-02-01 01:15:00  122    34 NA 90.2 55.9 0.37  NA 22.7  16.1     0.4
```

Descriptive statistics

Using the `summary()` function in R, we can generate a quick overview of the distribution of the various variables.

```
          mean     sd median trimmed    mad  min   max range skew
PM10    181.41 136.02 145.00  160.57 106.75 12.0 847.0 835.0 1.62
PM2.5    75.69  55.25  61.00   67.88  43.00  3.0 474.0 471.0 1.92
NO       14.65  19.22   6.10   10.24   4.15  0.1 157.5 157.4 2.85
NO2      55.76  20.23  53.20   54.78  23.13  0.2 106.9 106.7 0.33
NOx      42.67  22.44  37.70   39.77  20.76  4.2 165.2 161.0 1.26
CO        1.41   0.63   1.42    1.40   0.68  0.1   4.0   3.9 0.19
```

```
SO2       34.23  39.45  25.30   26.26  13.94  0.1 645.6 645.5 4.66
NH3       13.24   6.15  11.00   12.14   2.82  4.6  62.4  57.8 1.76
Ozone     35.63  27.02  32.40   34.01  35.14  0.1 123.8 123.7 0.36
Benzene    0.18   0.10   0.10    0.16   0.00  0.1   0.6   0.5 1.30
```

NA data

Out of the 86400 total observations, 13028 are missing.

We can see a breakup of the missing data.

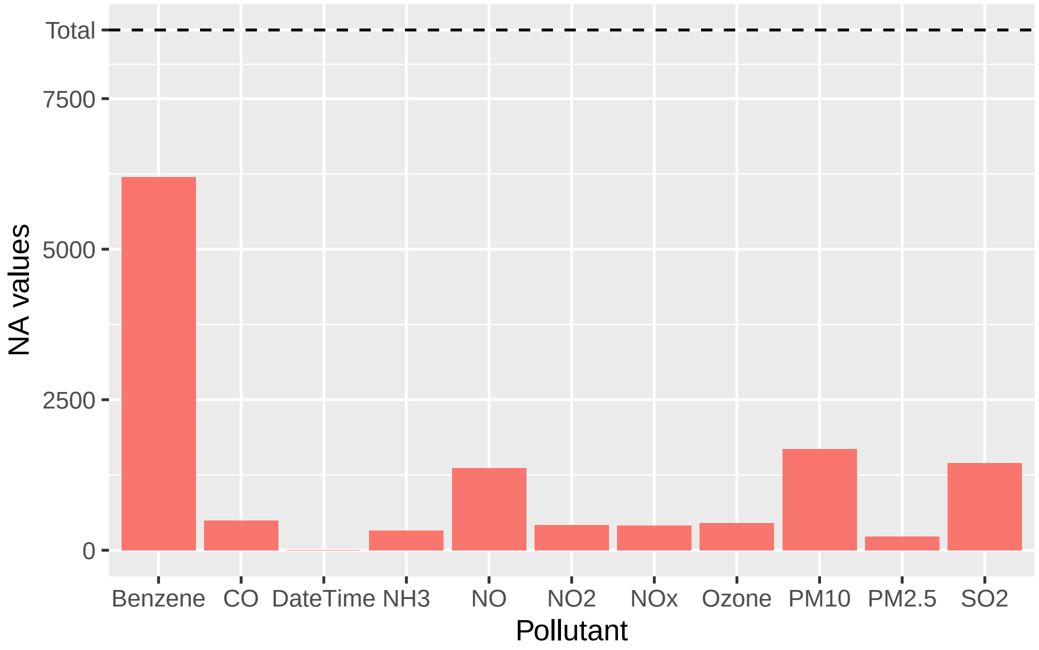
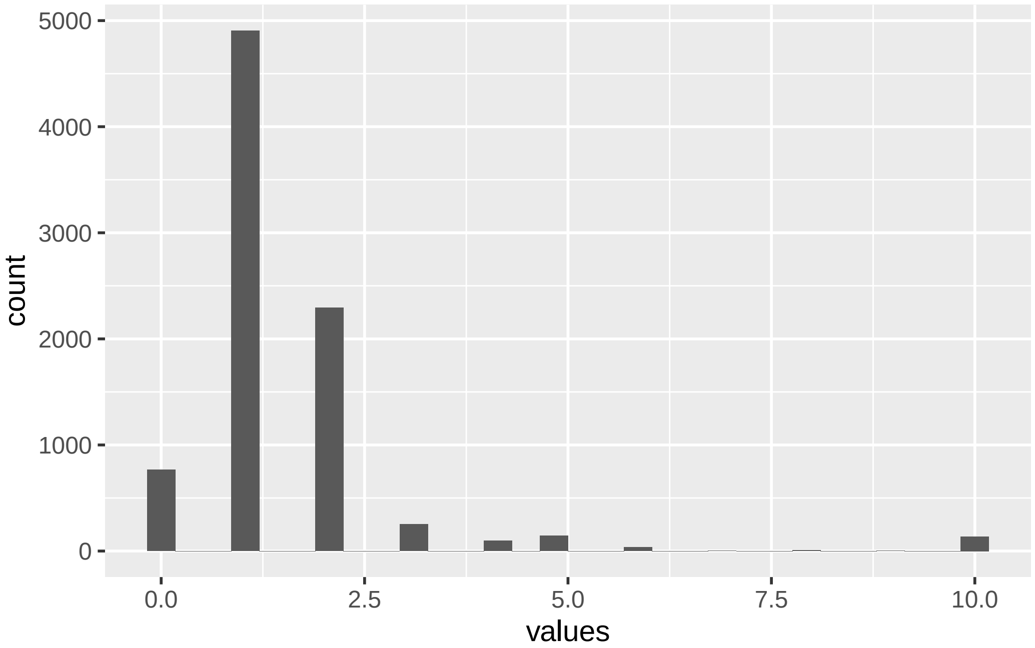


Figure 1: NA values in dataset

We can see that too many Benzene values (>70%) are missing to interpolate, or predict with any degree of usefulness. Thus, we leave Benzene out of all subsequent analysis.

Next, we review the occurrence of NA values by time (i.e. how many days had insufficient data for x pollutants):

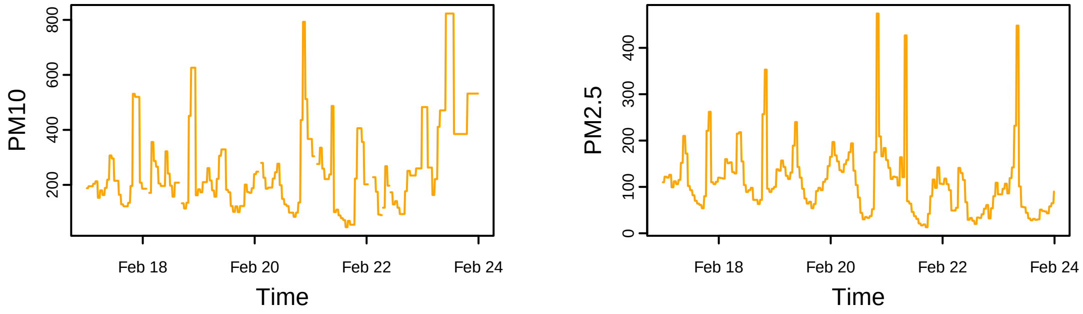


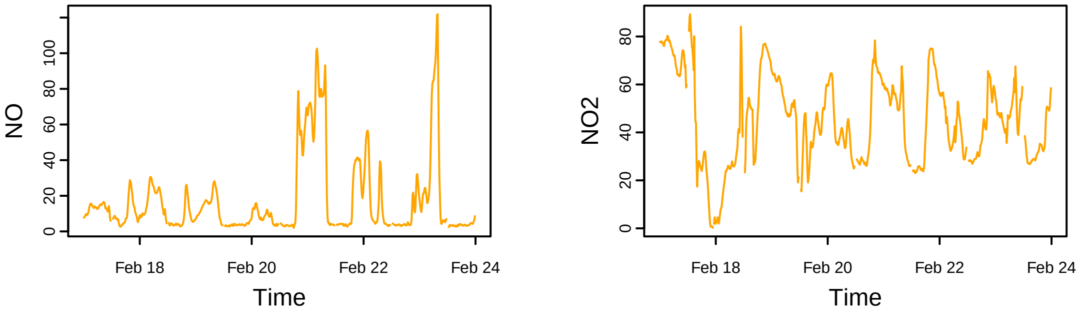
On closer inspection, one day (5th March) had no recorded readings of any of the pollutants.

Exploratory analysis

Weekly trajectory

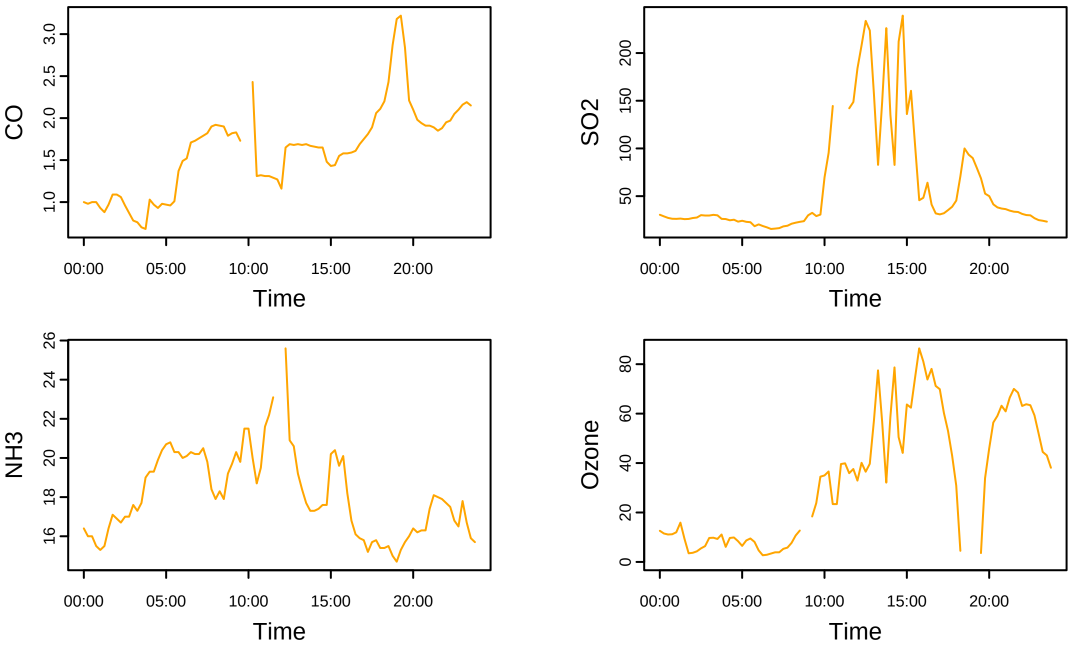
We can inspect the weekly trajectory of a few of to affirm our hypothesis of seasonal daily behaviour. We can see that the behaviour is somewhat cyclic but also displays significant variation.





Sampled day

Let us first explore what a (not randomly) sampled daily trajectory of the levels of various pollutants looks like. We have selected 17 February 2023, after visual inspection as it seemed to have nearly complete observations.



We need to bear in mind that since these plots are of a single day, they are not necessarily representative of all days in our dataset.

Cross sectional analysis

Correlations

Now, let us compute and visualize the correlations between levels of various pollutants.

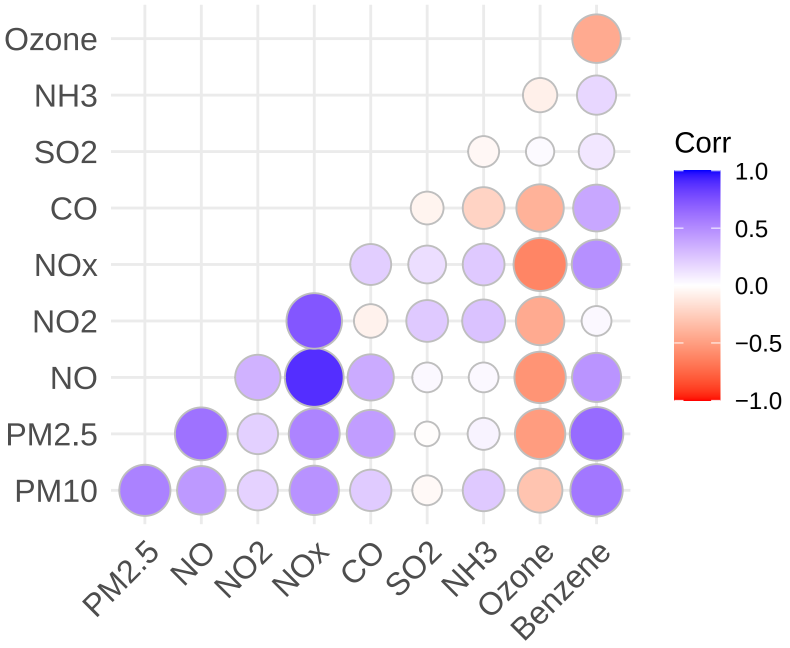


Figure 2: Correlations between various pollutant amounts

We can observe that there is a moderate level of positive correlation between some pollutants.

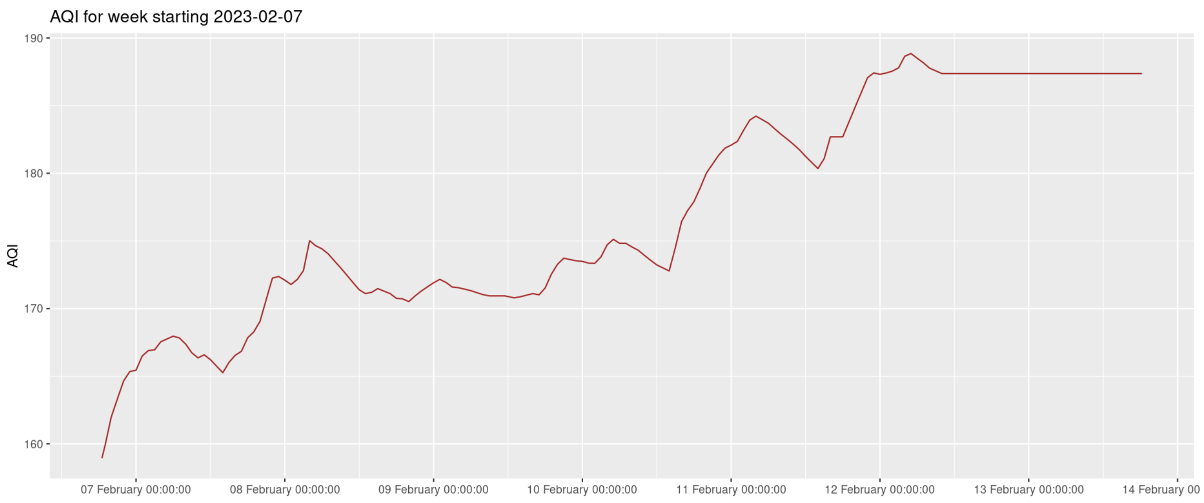
Exceptions to this observation are CO, Ozone, SO2 and NH3, which are negatively correlated (or uncorrelated) with most other pollutants.

AQI

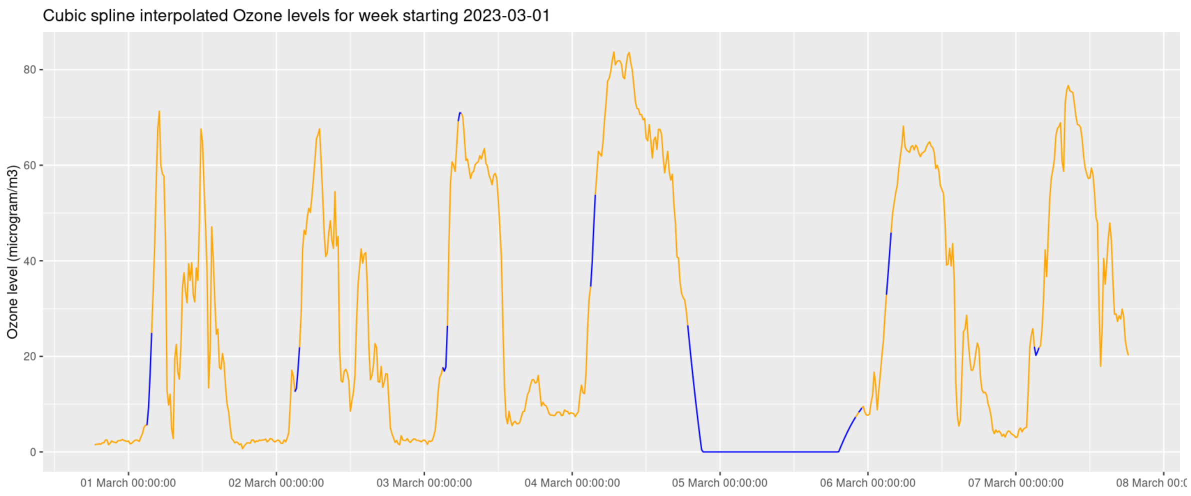
We also a devise a function to compute the AQI. However, since it has a relatively long context window (24 hours or 8 hours), AQI is relatively less sensitive to instantaneous changes and thus cannot be used as a single variable proxy for pollution at a time instant.

Interpolation

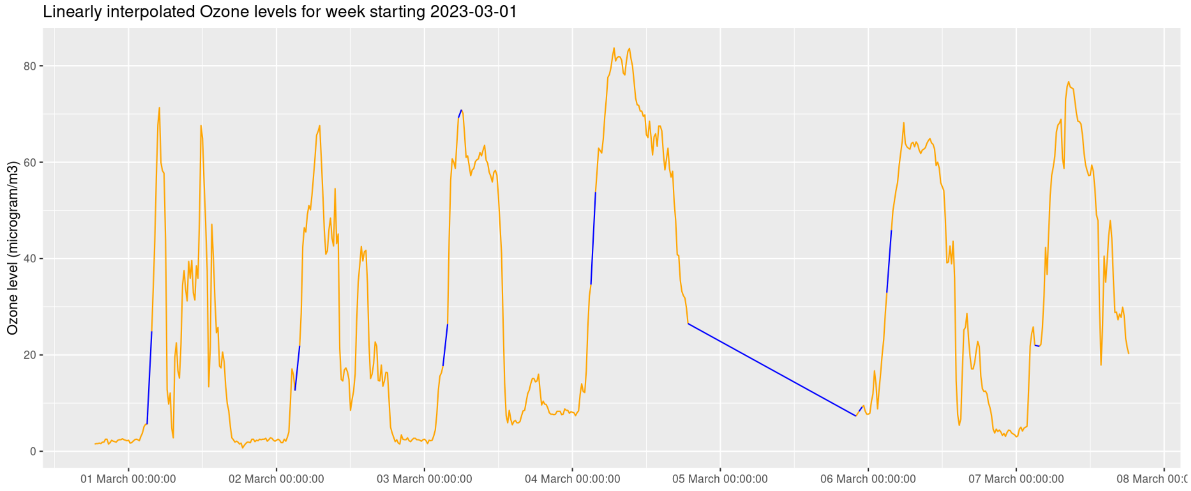
We tested two kinds of techniques for interpolation: linear, and cubic spline. Upon inspection, cubic spline seemed to better fit NA values in a few cases so we used it. We used the `na_interpolation()` function from the `imputeTS` library in R.



AQI for week starting 2023-02-07



Cubic spline interpolated Ozone levels for week starting 2023-03-01



In cubic spline interpolation, in many cases the interpolated value fell below zero, so this had to be normalized.

Characteristics of time series

From visual inspection, we can conclude that there is no clear trend, but the time series display seasonality with a period of 1 day.

Stationarity

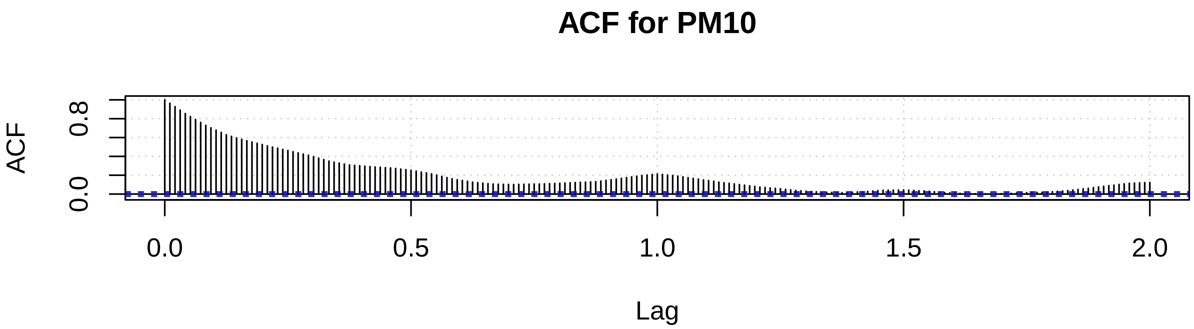
We apply the Augmented Dickey-Fuller test to check whether each of the time series are stationary. All of them have a p-value of less than 0.01 hence we go with the alternative hypothesis; i.e. the series are all stationary.

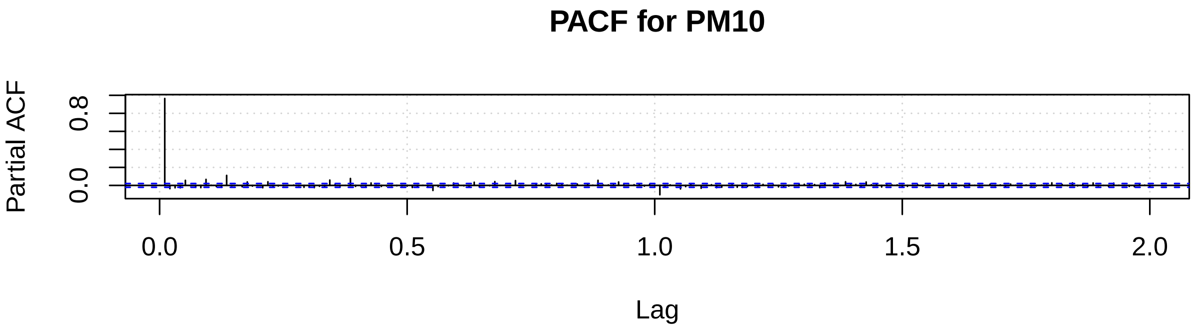
```
[1] "Stationarity test for PM10 series"

    Augmented Dickey-Fuller Test

data:  train[[i]]
Dickey-Fuller = -9.878, Lag order = 20, p-value = 0.01
alternative hypothesis: stationary

[1] "p value less than 0.01 hence stationary series."
[1] ""
[1] "Stationarity test for PM2.5 series"
```

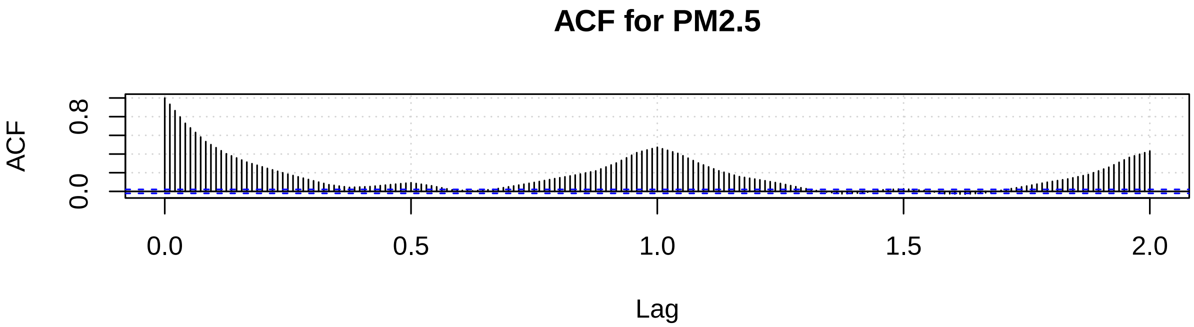
```
    Augmented Dickey-Fuller Test

data:  train[[i]]
Dickey-Fuller = -13.086, Lag order = 20, p-value = 0.01
alternative hypothesis: stationary

[1] "p value less than 0.01 hence stationary series."
[1] ""
[1] "Stationarity test for NO series"

    Augmented Dickey-Fuller Test

data:  train[[i]]
Dickey-Fuller = -14.294, Lag order = 20, p-value = 0.01
alternative hypothesis: stationary

[1] "p value less than 0.01 hence stationary series."
[1] ""
```

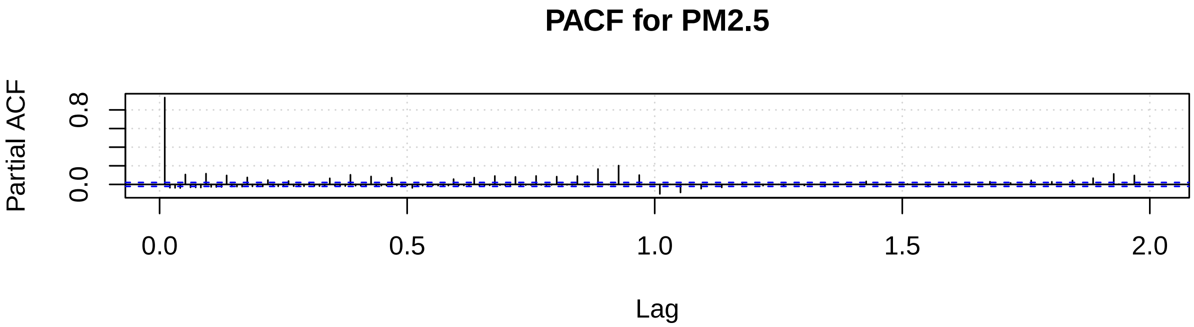
ACFs and PACFs

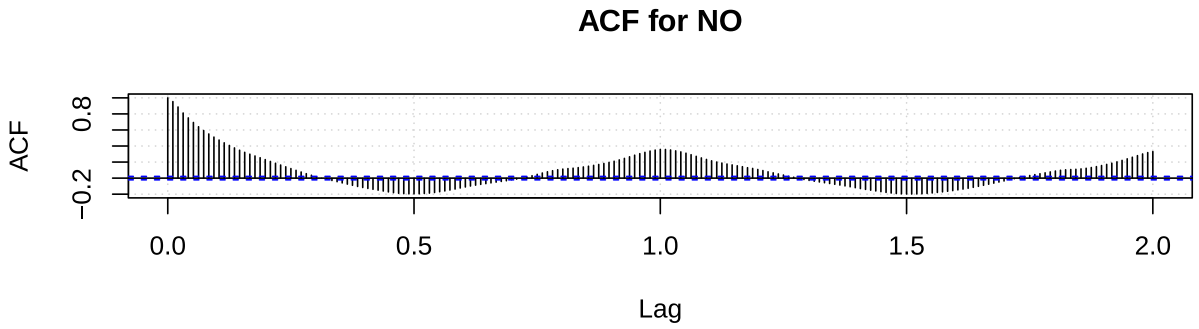
For most pollutants, we see a cyclical pattern in the ACFs, suggesting that an ARIMA model may be worth a try. In many of these, recorded values of nearly 1 day previously show significant PACF, showing that there is some recurrence-type relationship.

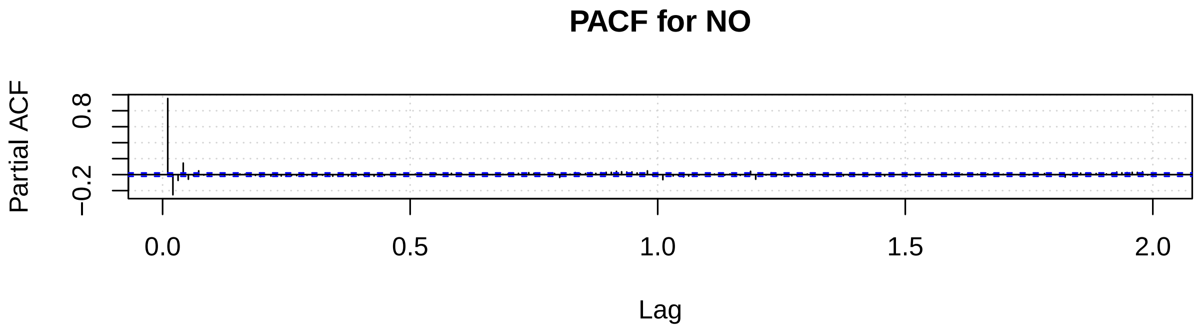

ACF for PM10
ACF
0.0
0.8
0.0
0.5
1.0
1.5
2.0
Lag

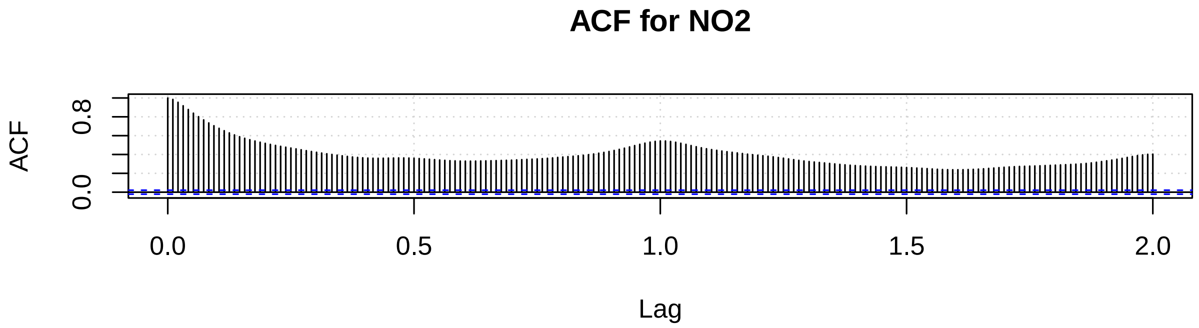

PACF for PM10
Partial ACF
0.0
0.8
0.0
0.5
1.0
1.5
2.0
Lag

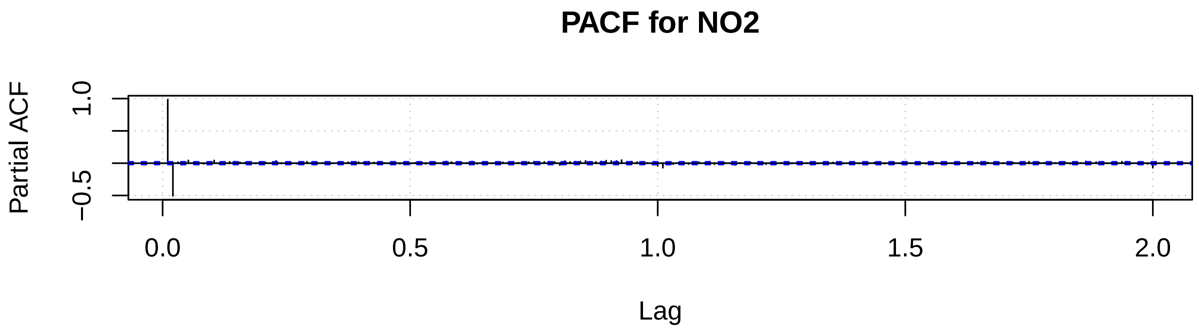

ACF for PM2.5
ACF
0.0
0.8
0.0
0.5
1.0
1.5
2.0
Lag


PACF for PM2.5
Partial ACF
0.0
0.8
0.0
0.5
1.0
1.5
2.0
Lag


ACF for NO
ACF
0.8
−0.2
0.0
0.5
1.0
1.5
2.0
Lag


PACF for NO
Partial ACF
0.8
−0.2
0.0
0.5
1.0
1.5
2.0
Lag


ACF for NO2
ACF
0.8
0.0
0.0
0.5
1.0
1.5
2.0
Lag


PACF for NO2
Partial ACF
1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Lag

Forecasting

Evaluation

Data split

To evaluate our models without bias, we will bifurcate the data into two sets, one each for training and testing. The training set comprises the data from 1 February to 29 April (88 days) while the test set comprises data from 30 April and 1 May (2 days). Thus, we are essentially targeting a two day forecast window.

Modelling techniques

With each of the modelling tools, we use the very useful `forecasts` package in R.

Model 1: Auto ARIMA

We use Auto ARIMA, which is a tool that searches the hyperparameter space of ARIMA models and chooses the best valued model according to AIC.

Many of the pollutants were giving a flat line in the ARIMA model. Some techniques I tried to improve fit were to use only the past 1 or 2 months data, but these yielded worse results than the full dataset. So, while useful as a benchmark, it only yielded good results with Ozone. We use the `auto.arima` tool from the `tseries` package in R.

Model 2: Auto Complex Exponential Smoothing

Complex Exponential Smoothing is a variant of the Exponential Smoothing method that is designed to handle time series data with complex patterns, such as seasonality and trend. It uses a combination of smoothing parameters and complex exponential smoothing equations to forecast future values. We use the `auto.ces` tool from the `smooth` package in R. This was the best performing technique, predicting the trajectory of a majority of the pollutants.

Model 3: Auto Multiple Seasonal ARIMA

Multiple Seasonal State Space ARIMA (MS ARIMA or MSSS-ARIMA) is an advanced time series forecasting model that extends the capabilities of traditional ARIMA models to handle multiple seasonal patterns. It is particularly useful for time series data that exhibit multiple seasonal cycles, such as daily, weekly, and yearly patterns We use the `auto.msarima` tool from the `smooth` package in R. MSARIMA proved much better than plain ARIMA models at predicting the time series.

Best models

Here, we present the plots of the best fitting models out of the 3 options explored, for each pollutant. Out of all the pollutants, NOx was the most difficult to model, with all 3 architectures failing to get a good forecast.

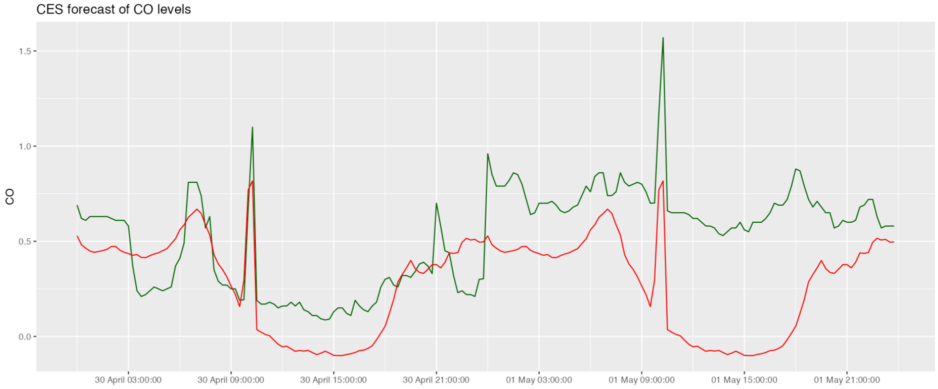


Figure 3: CO(CES)

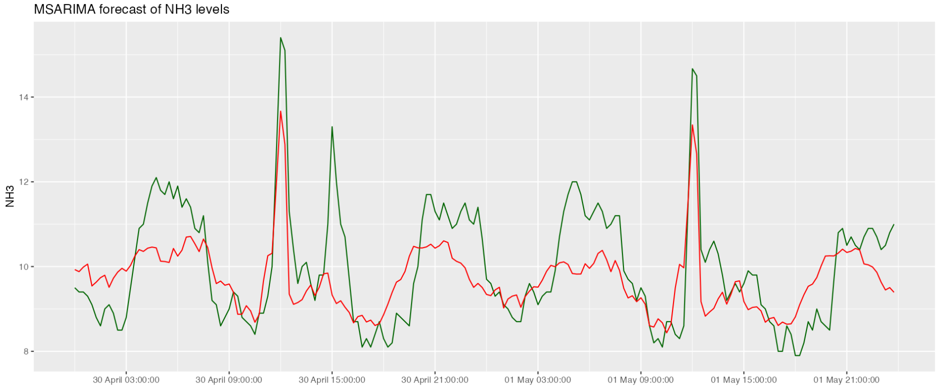


Figure 4: NH3 (MSARIMA)

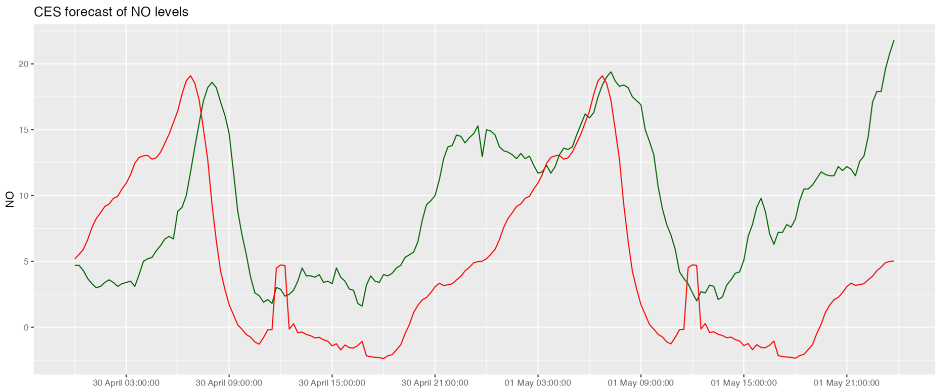


Figure 5: NO (CES)

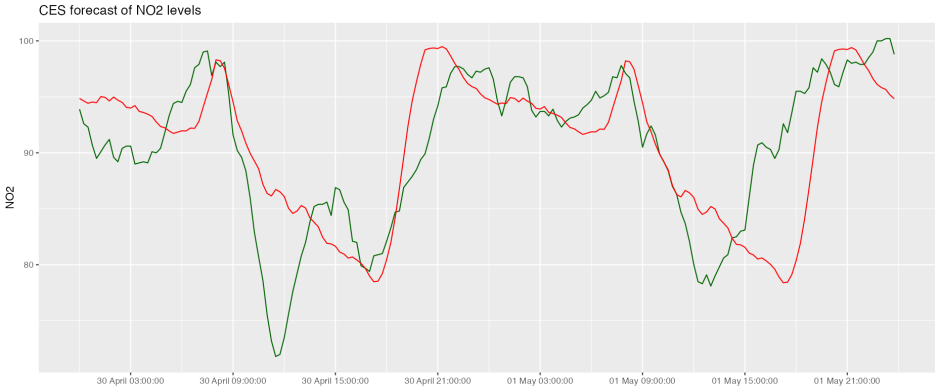


Figure 6: NO2 (CES)

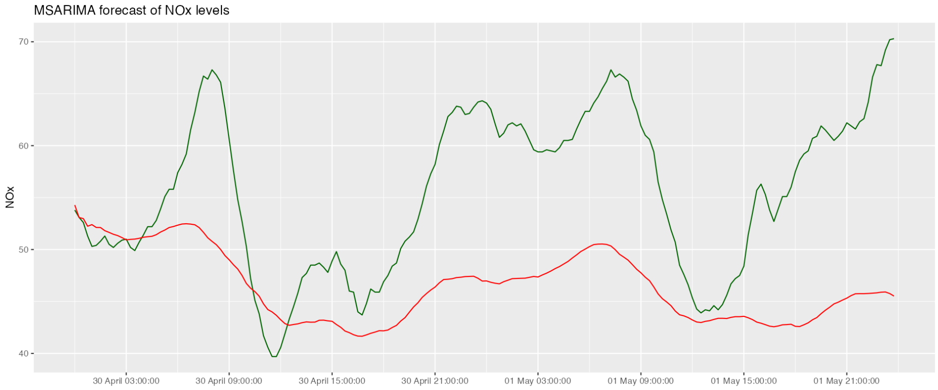


Figure 7: NOx (MSARIMA)

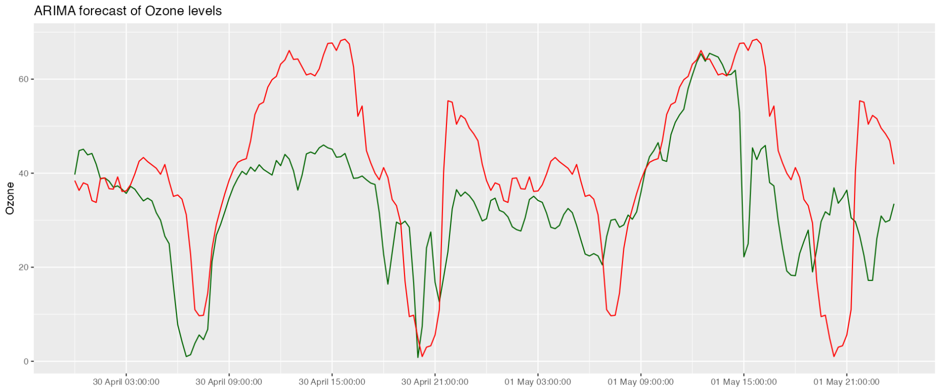

ARIMA forecast of Ozone levels

Figure 8: Ozone (ARIMA)

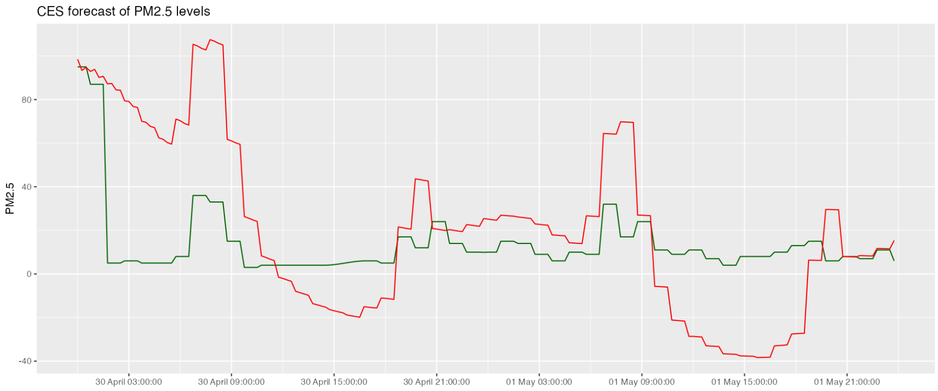

CES forecast of PM2.5 levels

Figure 9: PM2.5 (CES)

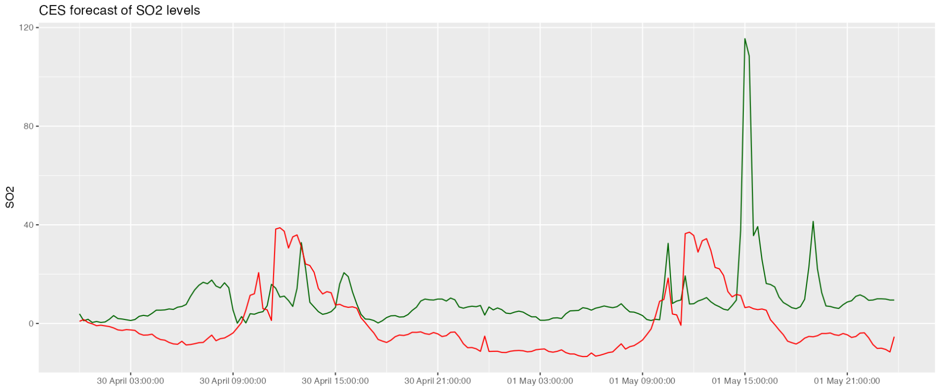

CES forecast of SO2 levels

Figure 10: SO2 (CES)

Conclusion

Using these relatively simple searching tools, we were able to generate relatively good predictive models for the pollutants, for a couple of days. Such modelling techniques will enable us to forecast as well as accurately interpolate NA values. We can ensemble these tools to predict the pollutants more accurately in the future.

References

Choudhary, Arti & Kumar, Pradeep. (2022). Estimation of Air Pollutants using Time Series Model at Coalfield Site of India. Proceedings of The International Conference on Data Science and Official Statistics. 2021. 10.34123/icdsos.v2021i1.59.