Time Series Analysis of Coalfield Pollutant Data

Arqam Patel

Introduction

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.

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).

NA data

Out of the 86400 total observations, 13028 are missing.

We can see a breakup of the missing data.

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:

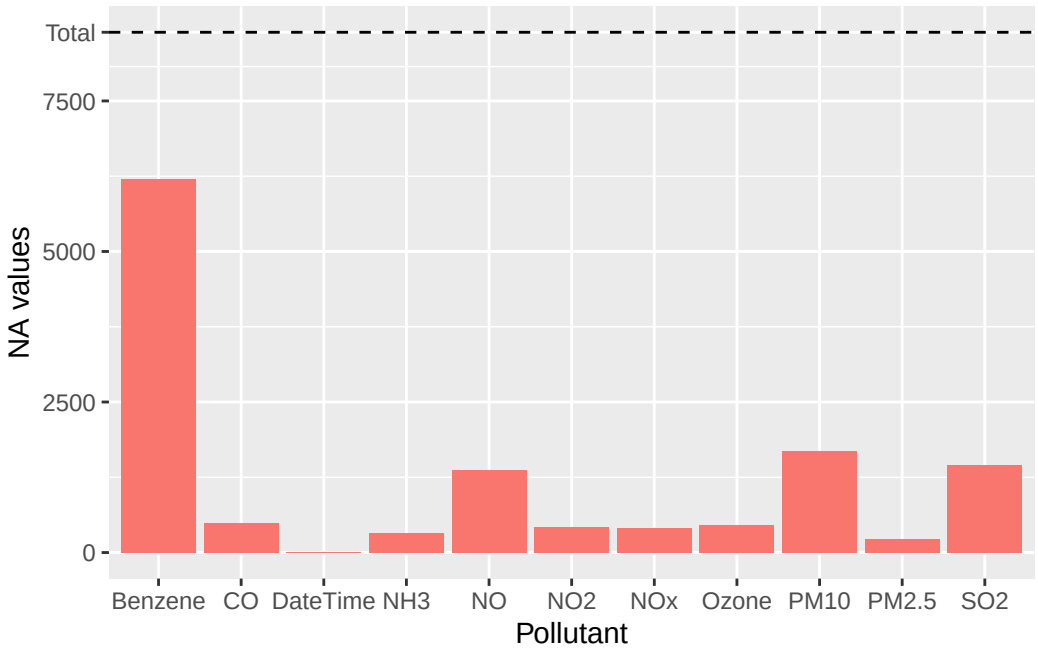
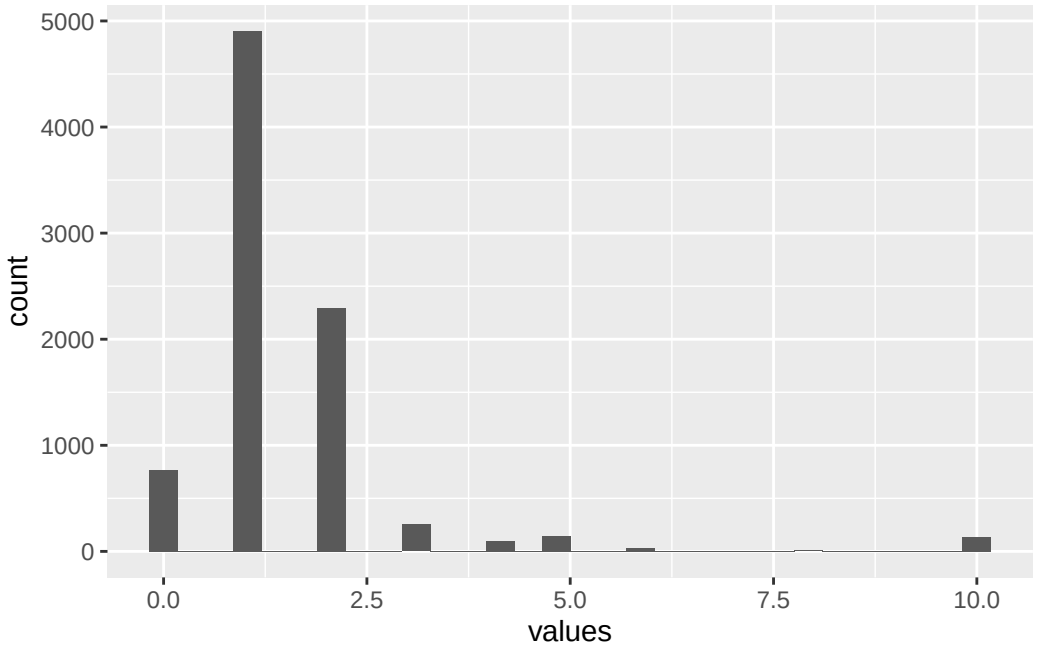


Figure 1: NA values in dataset

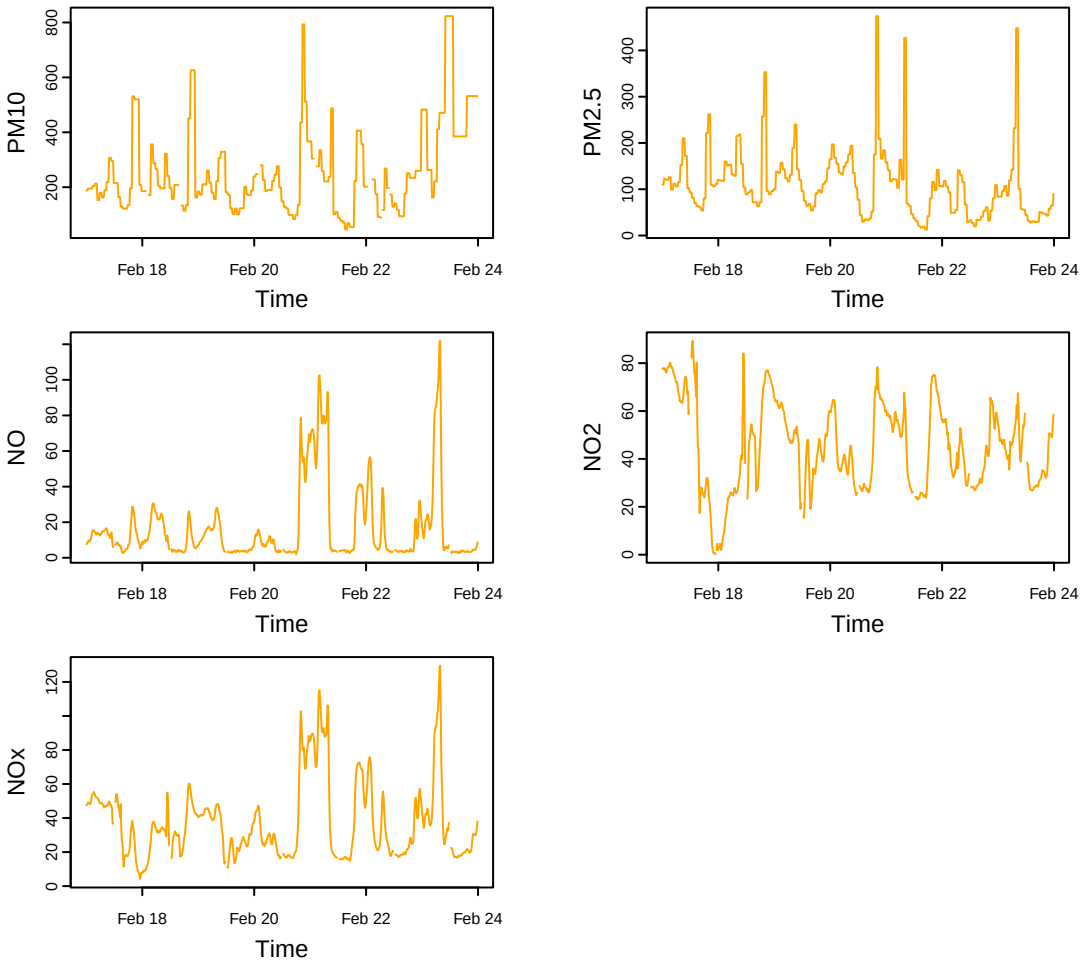


On closer inspection, one day (5th March) had no recorded readings of any of the pollutants.

Exploratory analysis

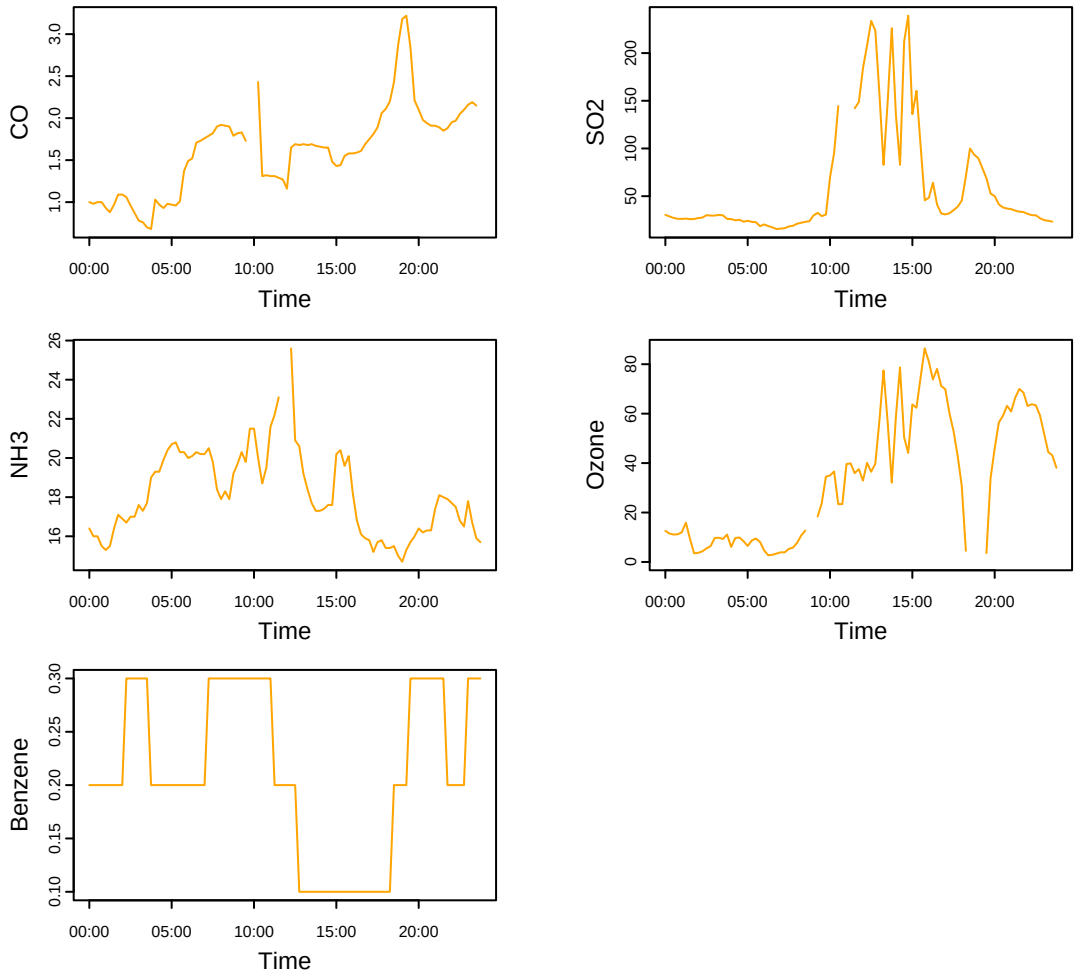
Weekly trajectory

We can inspect the weekly trajectory of a few of to affirm our hypothesis of seasonal daily behaviour.



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. However

Cross sectional analysis

Correlations

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

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.

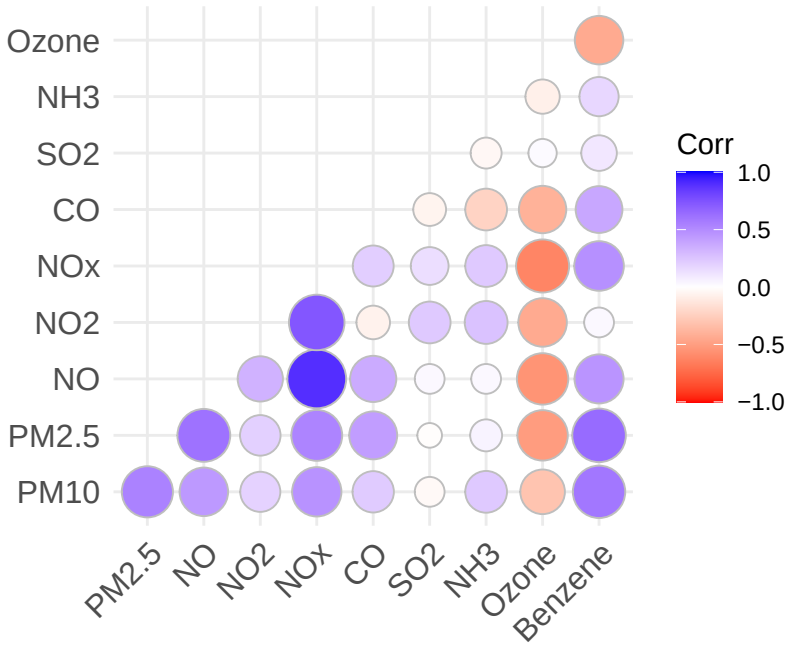


Figure 2: Correlations between various pollutant amounts

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 in .

Interpolation

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

Modelling

Evaluation

Data split

```
train <- cubic_int_df[1:6720,]
test <- cubic_int_df[6721:8640,]
```

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 11 April (70 days) while the test set comprises data from 12 April to 2 May (20 days).

We

Trend

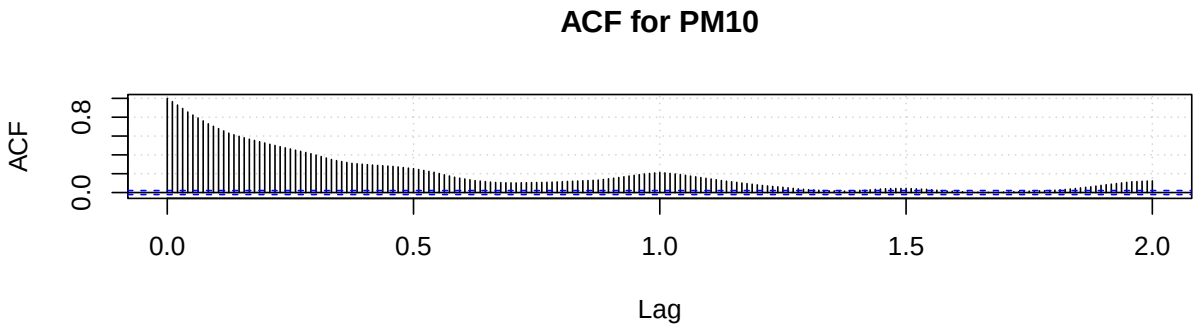
Seasonality

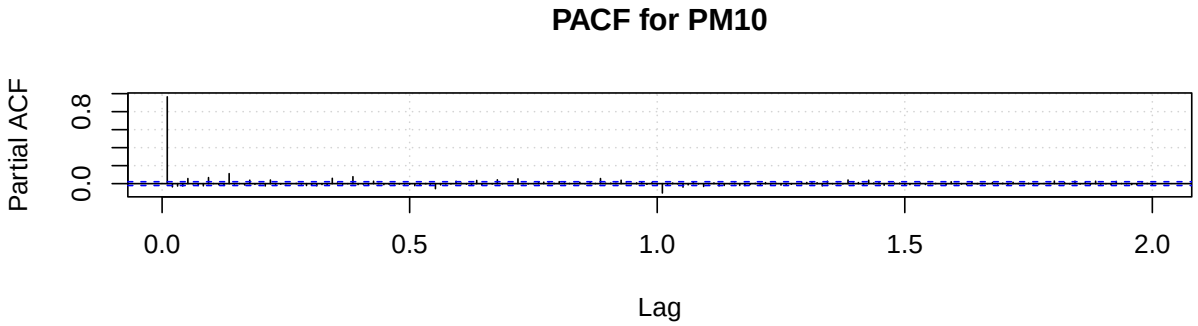
Stationarity

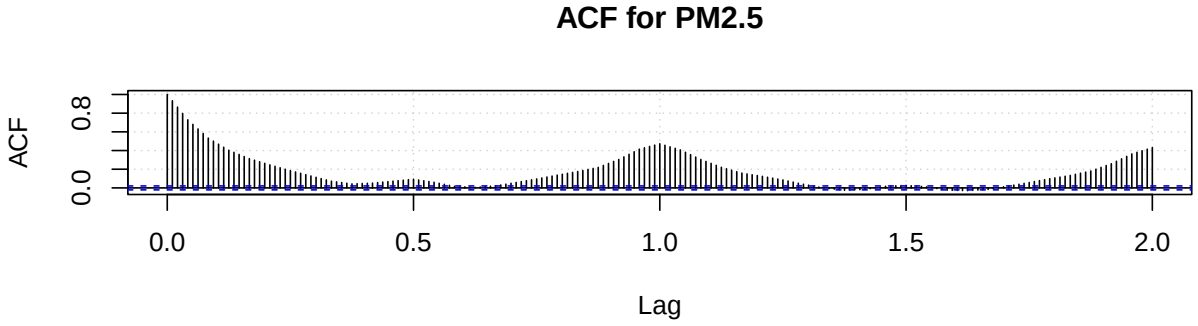
We apply the Augmented Dickey-Fuller test to check whether each of the time series are stationary. All of them have a p-value of less than 0.01 hence we go with the alternative hypothesis; i.e. the series are all stationary.

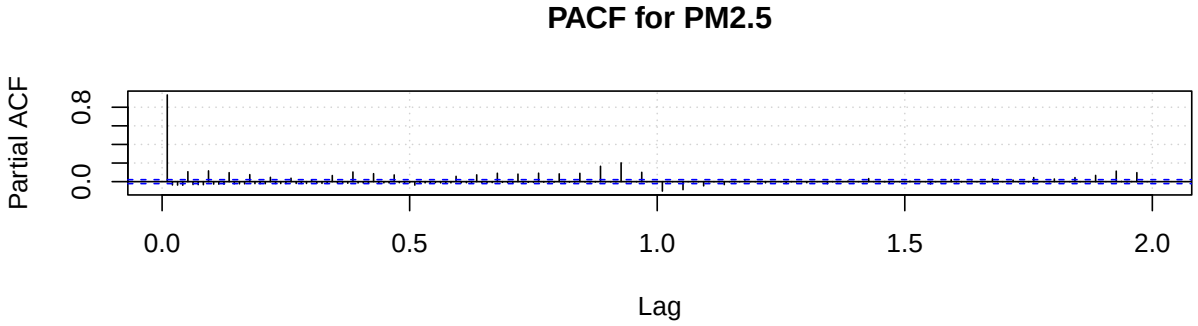
ACFs and PACFs

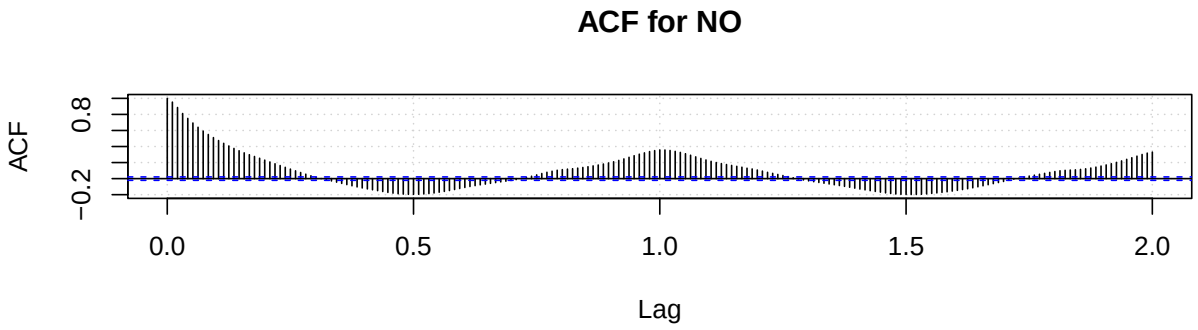
For most pollutants, we see a cyclical pattern in the ACFs, suggesting that an ARIMA model may be a good fit. In many of these, recorded values of nearly 1 day previously .

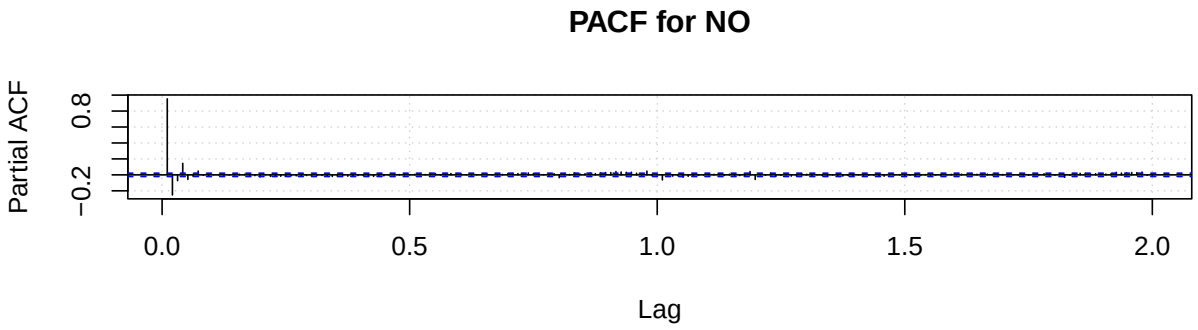

ACF for PM10
ACF
0.0
0.8
0.0
0.5
1.0
1.5
2.0
Lag

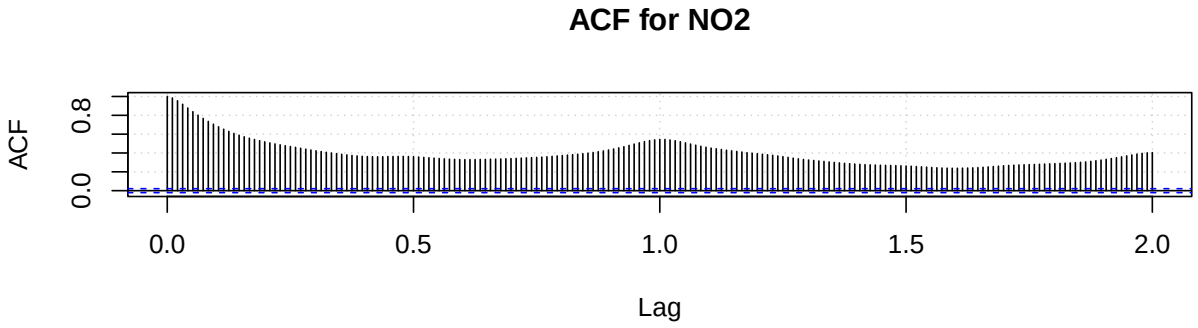

PACF for PM10
Partial ACF
0.0
0.8
0.0
0.5
1.0
1.5
2.0
Lag

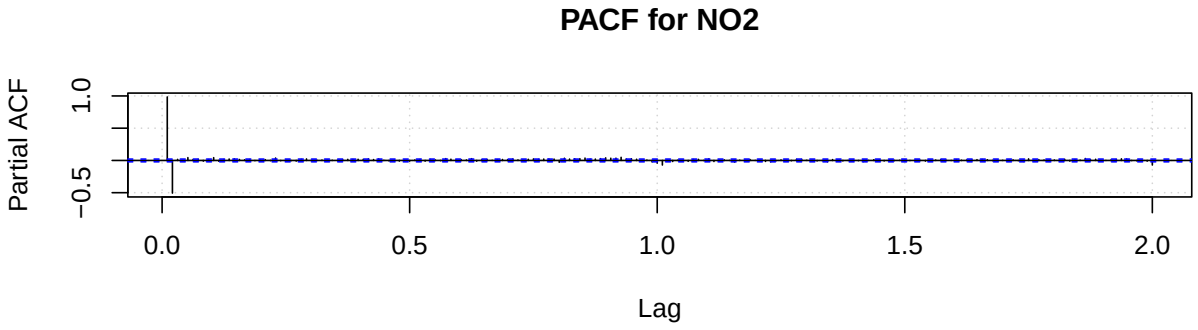

ACF for PM2.5
ACF
0.0
0.8
0.0
0.5
1.0
1.5
2.0
Lag


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

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.

Residual analysis

Conclusion

Using these relatively simple searching tools,

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.