# Analysis of Air Quality at Beijing

Nilanjan Debnath
Chennai Mathematical Institute
nilanjan@cmi.ac.in

November 2020

Beijing has constantly ranked as one of the most polluted cities on the planet. In the wake of the 21st century, when global warming is at its worst, and we are constantly being reminded to keep in check the air quality by using community vehicles, renewable and cleaner sources of energy, Beijing, one of the most populated cities in the world has been fighting the war against pollution since 1998.

## Problem Statement

The official report states that the annual concentration of $PM_{2.5}$(atmospheric particulate matter with diameter less than 2.5mm) in 2016 has declined by 9.9% relative to the 2015 level. Here, we analyze the data from 12 different locations in and near Beijing. We have 420768 samples and 17 features in this dataset. 15 of those are numerical variables which include hourly data from 1st March, 2013 to 28th February 2017 of various pollutant levels (PM2.5, PM10, $SO_2$, $NO_2$, CO and $O_3$)[1]. Apart from that information, we also have the temperature, pressure, precipitation, wind direction and wind speed of the data all of which are recorded in accordance to the station where the data was measured.

A substantial part of China is experiencing chronic air pollution with severe fine particulate matter (PM) concentration and PM2.5 in particular. The north China Plain (NCP) that surrounds Beijing endures the most severe air pollution in the country with excessive PM2.5 concentration. In a move to clear up the smog, China's State Council has set a 25% PM2.5 reduction target for the NCP by 2017 relative to the 2012 level, and a specific target of no more than $60\mu gm^{-3}$ for Beijing's annual average. Our target here in this project is to find out how all the efforts at reducing air pollution in the Beijing airspace is effecting the pollutants in the air. We aim to create a time series model which can predict the various pollutant levels for the later months of 2017. We use the instances from 1st March 2013 to 31st December 2016 as training data, and the rest as test data.

[1] all pollutant levels are in $\mu gm^{-3}$

## Exploratory Data Analysis

In the following table, I'll provide a description of the data we are working on.

| Attribute | Description | Data Type |
|---|---|---|
| No. | Row Number | numerical |
| year | year of data in this row | numerical |
| month | month of data in this row | numerical |
| day | day of data in this row | numerical |
| hour | hour of data in this row | numerical |
| PM2.5 | PM2.5 concentration ($\mu gm^{-3}$) | numerical |
| PM10 | PM10 concentration ($\mu gm^{-3}$) | numerical |
| SO2 | SO2 concentration ($\mu gm^{-3}$) | numerical |
| NO2 | NO2 concentration ($\mu gm^{-3}$) | numerical |
| CO | CO concentration ($\mu gm^{-3}$) | numerical |
| O3 | O3 concentration ($\mu gm^{-3}$) | numerical |
| TEMP | Temperature (degree Celsius) | numerical |
| PRES | Pressure (hPa) | numerical |
| DEWP | Dew point temperature | numerical |
| RAIN | Precipitation (mm) | numerical |
| wd | Wind direction | categorical |
| WSPM | Wind Speed | numerical |
| station | Station Name | categorical |

Table 1: Dataset Attribute Description

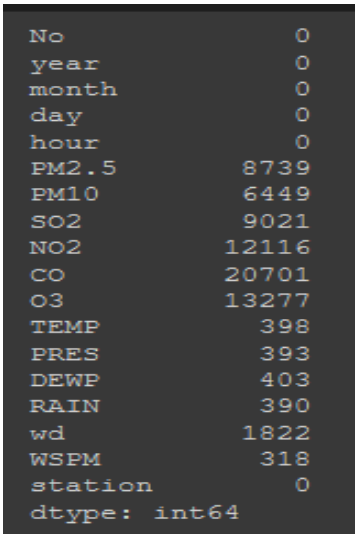


Fig. 1 Number of null values

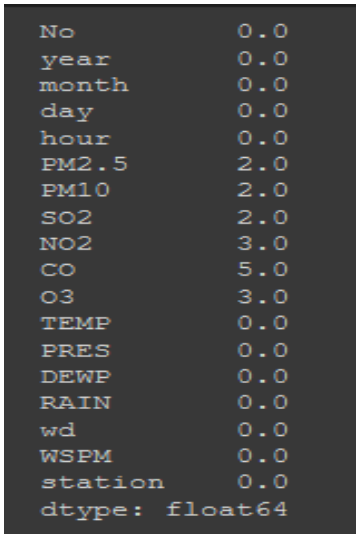


Fig. 2 Percentage of null values

The corresponding figures show the number of null values( Fig. 2) and percentage of null values( Fig. 3) in each attributes. As we can see, that most of the attributes have several null values. There can be a lot of reasons including faulty equipment, battery change etc. To make up for this, we select the median of the day/month of the attribute with missing values and replace the missing value with it. For example if the PM2.5 data for an hour of the 1st of January, 2015 is missing, we will replace it with the median value of the rest of the observed PM2.5 values over the day.

The following figures shows the change in the level of pollutant level over the years. For this, we select the median value of the entire year.

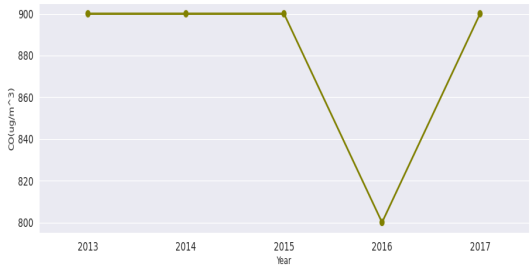


Fig. 3 Carbon Monoxide emission over the years

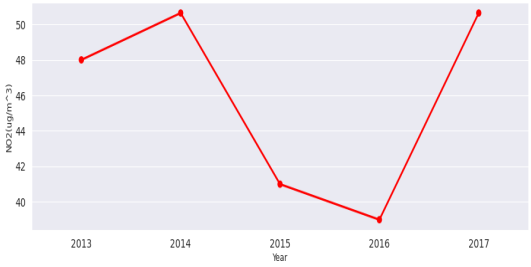


Fig. 4 Nitrogen Dioxide emission over the years

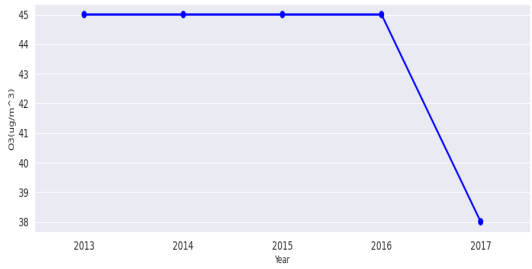


Fig. 5 Ozone emission over the years

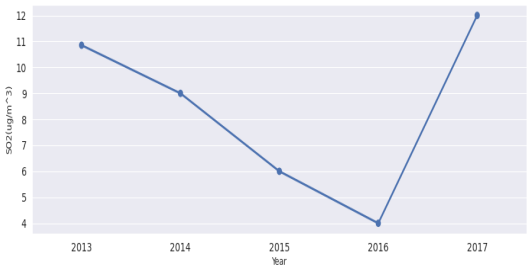


Fig. 6 Sulphur Dioxide emission over the years

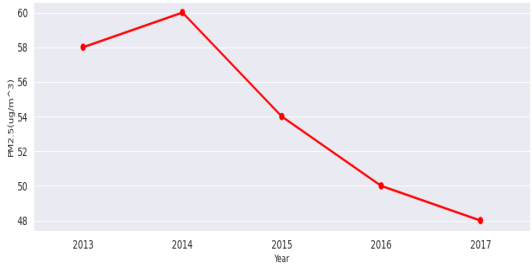


Fig. 7 PM2.5 emission over the years

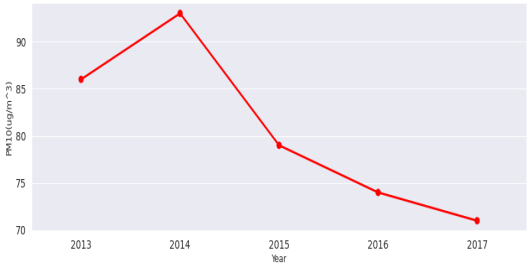


Fig. 8 PM10 emission over the years

Upon examination, while all the pollutants except Ozone show a decrease from 2015 to 2016 which the Beijing Municipal Corporation correctly announced, there is a massive increase in the release of pollutants in 2017. While we have the data of only 2 months of 2017, let's check if the change of pollutants is a seasonal effect.

The following figures (Fig.9-14) shows the change in the level of pollutant level over the months. For this, we select the median value of the entire month.
As we can see, that pollutant emission is significantly higher in winter than in the middle of the year. Hence, it makes sense that just taking data from the first 2 months of the year 2017 gives us a higher median than the other years.
Thus, while fitting the time series model, we have to take in context the seasonal variation of the data since as is evident it affects the data in a big manner.

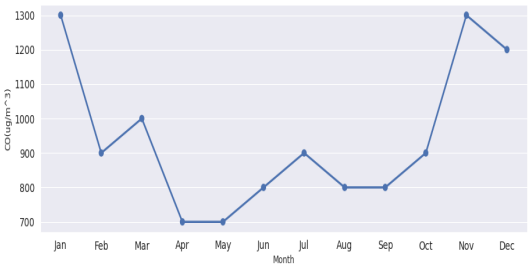


Fig. 9 Carbon Monoxide emission over the months

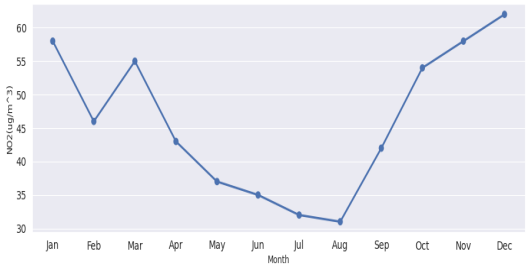


Fig. 10 Nitrogen Dioxide emission over the months

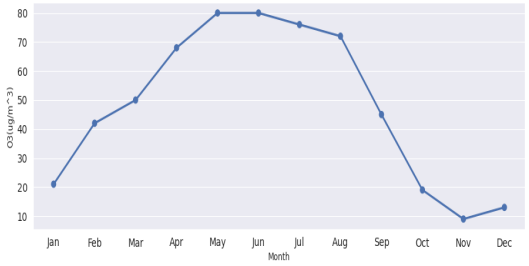


Fig. 11 Ozone emission over the years

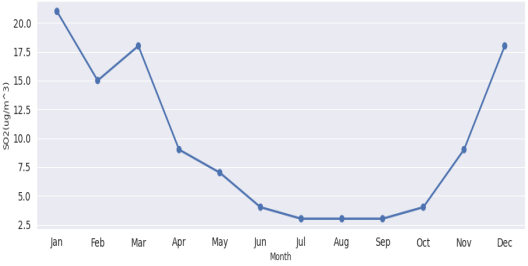


Fig. 12 Sulphur Dioxide emission over the years

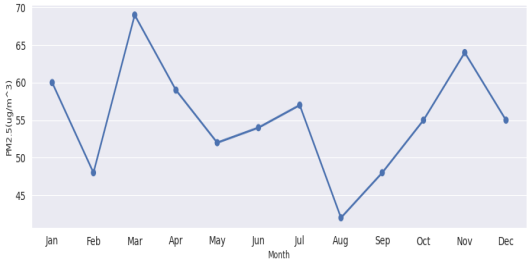


Fig. 13 PM2.5 emission over the years

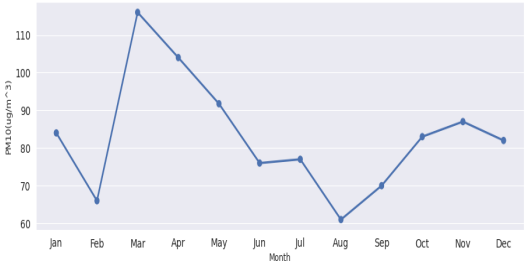


Fig. 14 PM10 emission over the years

Finally we'll look at the hourly median emission of pollutant over the years (Fig. 15-20).
We can see in most of the graphs, increase in pollutant level between mid day and evening as is expected because of daytime working hours. However, surprisingly, PM2.5, PM10 show a increase in the night rather than the daytime.
This is indicative of the fact some activities in powerplants, industries and automobiles are even going on at night. Also PM2.5 and PM10 are the main components of smog in the morning which goes in alignment to news articles as such. Life in China, Smothered by Smog, NY Times

## Time Series Forecasting

In this section we will fit a ARIMA and/or SARIMAX model for each pollutant from March 2013 to October 2016. We then will test it against the data from November 2016 till February 2017. The model which performs better, will be used to forecast data from March 2017 till June 2017.
As we can see that we have hourly data, we took the mean of each day and the median for the month. We then assigned the median of every month to the last day of the same

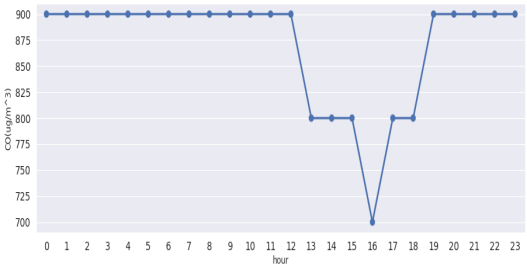


Fig. 15 Carbon Monoxide emission over the day

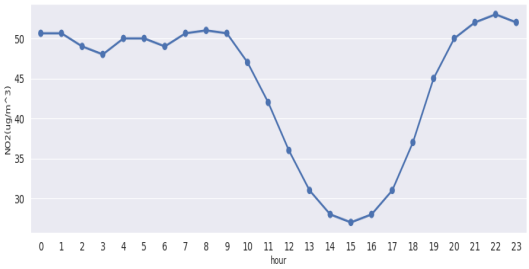


Fig. 16 Nitrogen Dioxide emission over the day

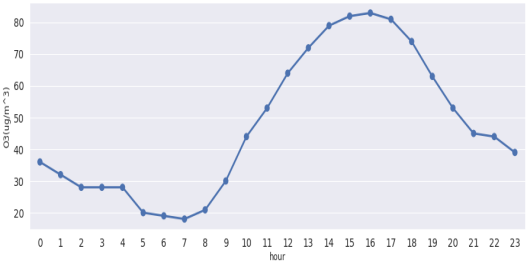


Fig. 17 Ozone emission over the day

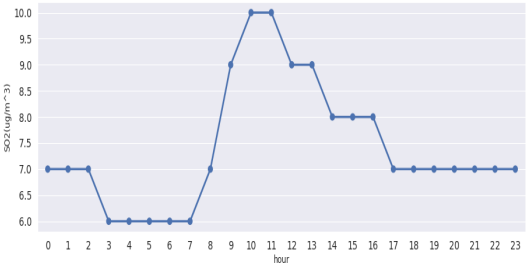


Fig. 18 Sulphur Dioxide emission over the day

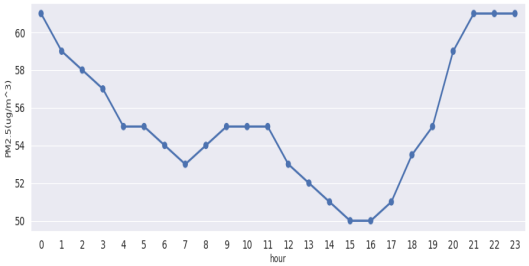


Fig. 19 PM2.5 emission over the day

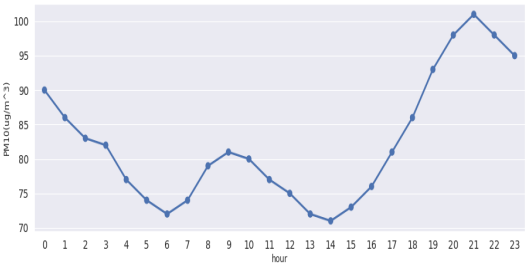


Fig. 20 PM10 emission over the day

month just for uniformity.

## Particulate Matter 2.5 (PM2.5)

Fig. 21 showcases the pollutant graph over the years while fig. 22 is the seasonal decomposition of the graph over the years.

As we can see that there from the trend graph in Fig. 22 that there is a decreasing trend of PM2.5 pollutants in the air over Beijing over the years which is indicative of the fact that Beijing's fight against PM2.5 is working. We can see a distinct seasonality in the graphs where every March to July every year there is a drop in PM2.5 pollutant level and an increase in the levels in July-August before decreasing again and finally making the jump up during winter months.

First we test a SARIMA model taking the first 44 months as training data and the last 4 months as test data(Fig. 23).(We take the model with the lowest AIC). Next we test the best ARIMA model on the same dataset(Fig. 24). We move on ahead with the SARIMA model since it was giving a better result(RMSE(SARIMA) = $11.36\mu gm^{-3}$ against RMSE(ARIMA) = $16.31\mu gm^{-3}$). In Fig. 25 we can see a few outliers in the standardized residuals and their normal Q-Q plot. A small amount of autocorrelation that still remains but otherwise, the model fits well. The Kernel Density Estimate shows

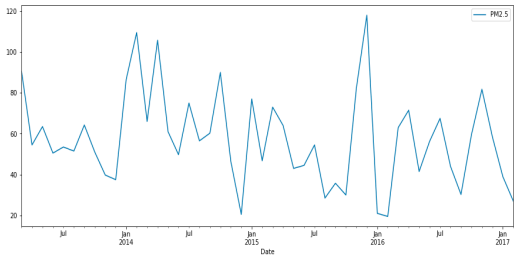


Fig. 21 PM2.5 emission over the years

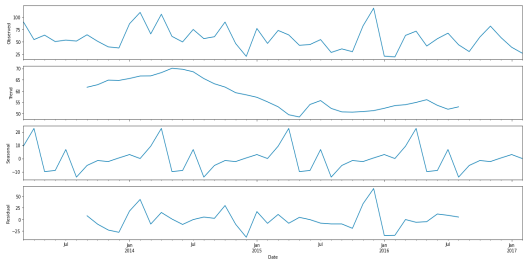


Fig. 22 Seasonal Decomposition

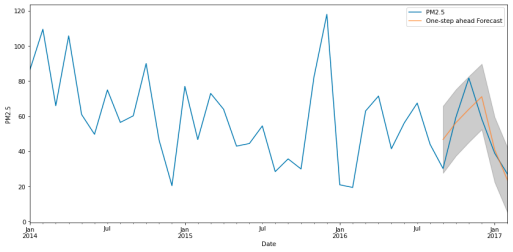


Fig. 23 Fitting a SARIMA model

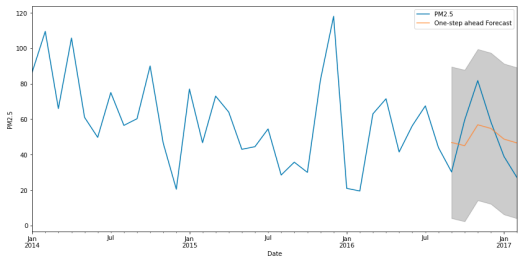


Fig. 24 Fitting an ARIMA model

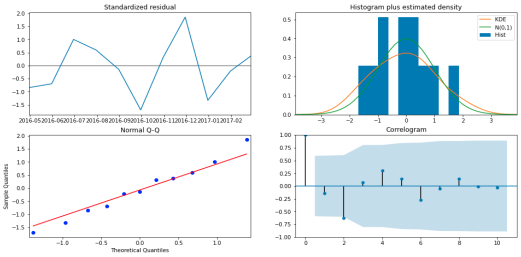


Fig. 25 Plotting the diagonistics

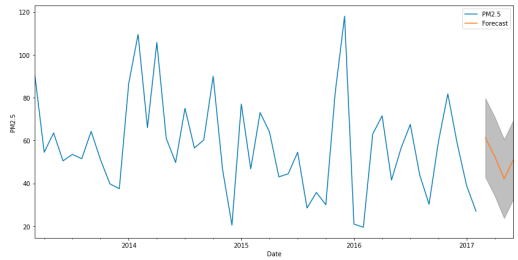


Fig. 26 Forecasting on unseen data

that the distribution of the dataset is a bit platykurtic in comparison to a normal (0,1) distribution. Finally we use the SARIMA model to forecast 4 months of unseen data in Fig. 26. A SARIMA model has 2 parameters. The trend parameters $(p, d, q)$ and the seasonal parameters $(P, D, Q, m)$. Where p = trend autoregression order, d = trend difference order and q = trend moving average order, P = seasonal autoregressive order, D = seasonal difference order, Q = seasonal moving average order and m = periodicity. To forecast for 4 months into 2017, we use SARIMA model (0,1,1)x(3,0,0,12). As we can see, the forecasted pollutant level of PM2.5 is lower in the mid months of 2017 as compared to previous years.

## Particulate Matter 10 (PM10)

Similar to the last pollutant, the trend graph of Fig. 28 shows us that there is a decreasing trend of PM10 in the air over Beijing over the years.
We have a seasonal component where in during early and mid year we have spikes on PM10 pollution. First we test a SARIMA model taking the first 44 months as training data and the last 4 months as test data(Fig. 29).(We take the model with the lowest AIC). Next we test the best ARIMA model on the same dataset(Fig. 30). We move on ahead with the SARIMA model since it was giving a better result(RMSE(SARIMA) = $14.85\mu gm^{-3}$ against RMSE(ARIMA) = $30.3\mu gm^{-3}$). In Fig. 31 we can see a few

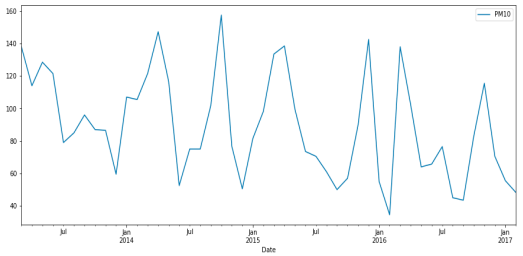


Fig. 27 PM10 emission over the years

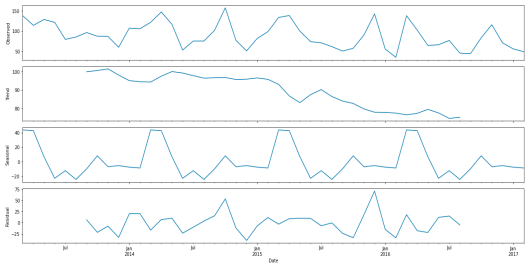


Fig. 28 Seasonal Decomposition

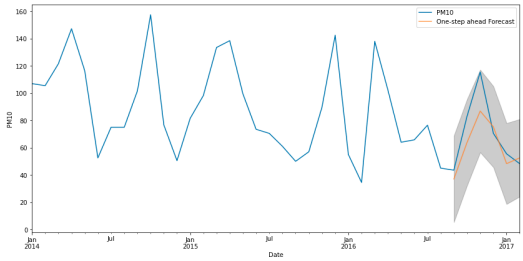


Fig. 29 Fitting a SARIMA model

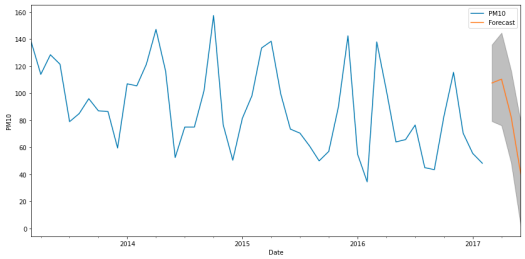


Fig. 30 Fitting an ARIMA model

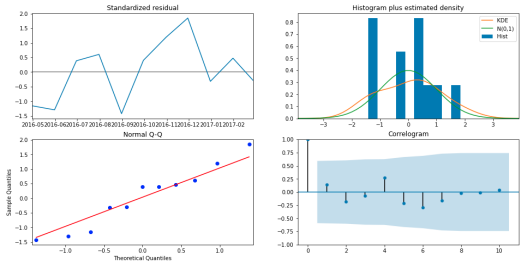


Fig. 31 Plotting the diagonistics

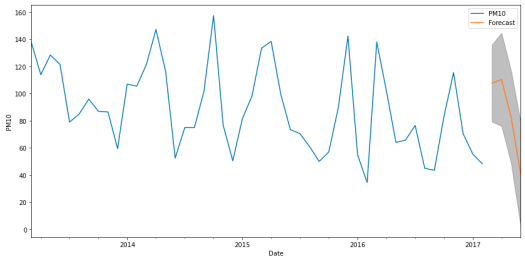


Fig. 32 Forecasting on unseen data

outliers in the standardized residuals and their normal Q-Q plot. A small amount of autocorrelation that still remains but otherwise, the model fits well. The Kernel Density Estimate shows that the distribution of the dataset is a bit platykurtic in comparison to a normal (0,1) distribution. Finally we use the SARIMA model to forecast 4 months of unseen data in Fig. 32. To forecast for 4 months into 2017, we use SARIMA model (0,1,4)x(2,1,0,12). As we can see, the forecasted pollutant level of SO2 is lower in the mid months of 2017 as compared to previous years.

## Sulphur Dioxide ($SO_2$)

Similar to the earlier pollutants, we can see from the trend graph in Fig. 34 that SO2 pollution is massively decreasing over the years. We also can see heavy seasonality in the graph with regular spikes during the winter months. Firstly we fit a SARIMA model to the data taking the first 44 months as training data and the next 4 months as test data (Fig.35). We do the same with a ARIMA model (Fig. 36). Due to the glaring seasonality component, our SARIMA model works better than our ARIMA model (RMSE(SARIMA) $= 1.08\mu gm^{-3}$ against RMSE(ARIMA) $= 3.22\mu gm^{-3}$) and hence we move forward with it.

In Fig. 37 we can see a few outliers in the standardized residuals and their normal Q-Q plot. A small amount of autocorrelation that still remains but otherwise, the model fits

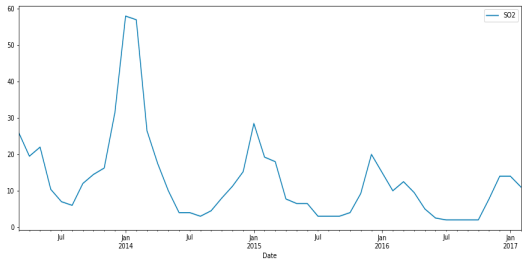


Fig. 33 SO2 emission over the years

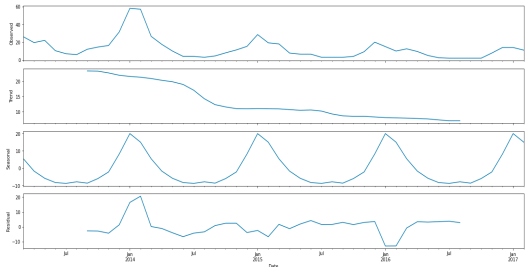


Fig. 34 Seasonal Decomposition

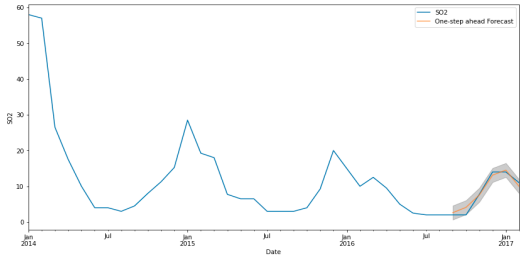


Fig. 35 Fitting a SARIMA model

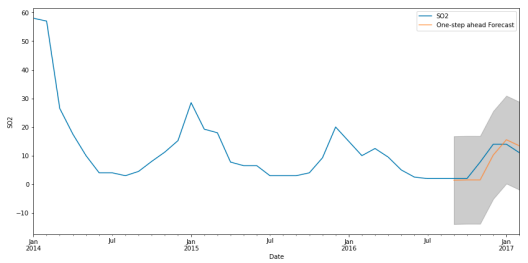


Fig. 36 Fitting an ARIMA model

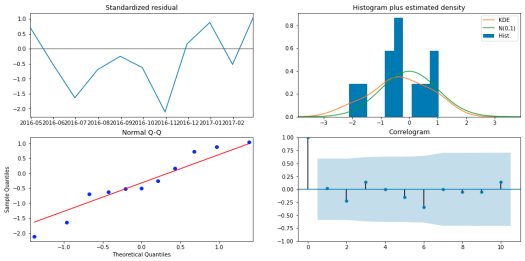


Fig. 37 Plotting the diagonistics

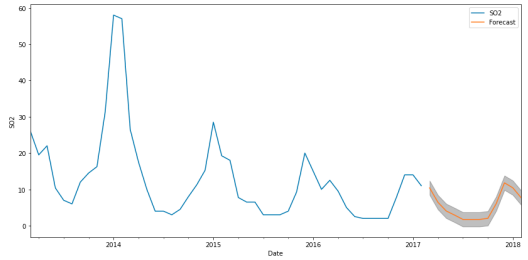


Fig. 38 Forecasting on unseen data

well. The Kernel Density Estimate shows that the distribution of the dataset is a bit platykurtic in comparison to a normal (0,1) distribution. Finally we use the SARIMA model to forecast 12 months (gives better visualization than 4 months) of unseen data in Fig. 38. To forecast for 12 months into 2017, we use SARIMA model (1,0,0)x(3,0,0,12). As we can see, the forecasted pollutant level of SO2 is lower in the mid months of 2017 as compared to previous years showcasing the improvement of pollutant SO2 level in Beijing.

## Nitrogen Dioxide ($NO_2$)

As we can see in similarity to the other pollutants above, we get a decreasing trend (Fig. 42) in this pollutant data over the years. We do have a seasonal component with the pollutant level maximizing after the summer every year.
Firstly we fit a SARIMA model to the data taking the first 44 months as training data and the next 4 months as test data (Fig.41). We do the same with a ARIMA model (Fig. 42). Our SARIMA model works better than our ARIMA model (RMSE(SARIMA) = $6.51\mu gm^{-3}$ against RMSE (ARIMA) = $16.53\mu gm^{-3}$) and hence we move forward with the SARIMA model. In Fig. 43 we can see a few outliers in the standardized residuals. As we can figure from the qq plot, our data does not follow normal distribution. A small amount of autocorrelation that still remains but otherwise, the model fits well. The Kernel Density Estimate shows that the distribution of the dataset is bimodal.

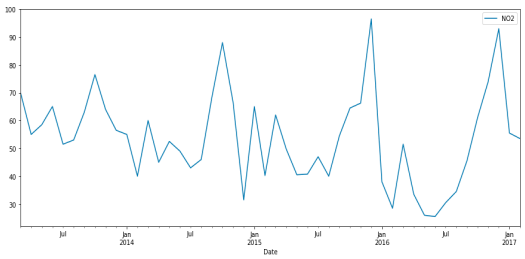


Fig. 39 NO2 emission over the years

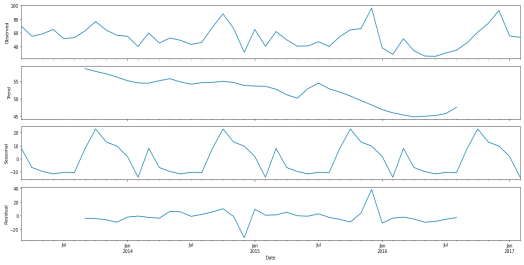


Fig. 40 Seasonal Decomposition

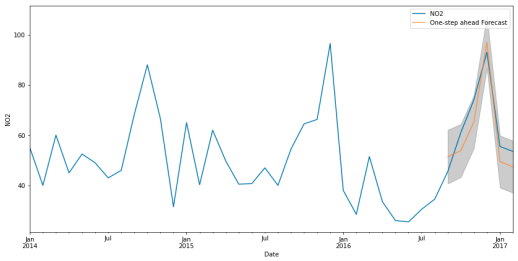


Fig. 41 Fitting a SARIMA model

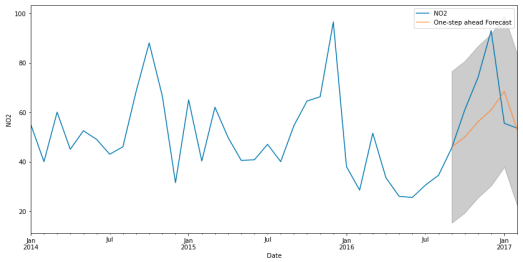


Fig. 42 Fitting an ARIMA model

Finally we use the SARIMA model to forecast 8 months (gives better visualization than

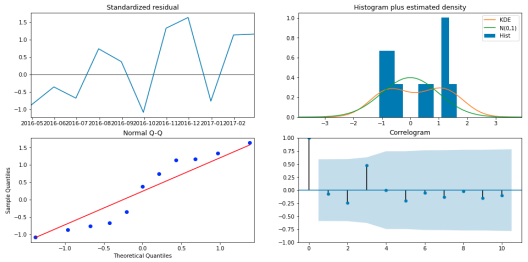


Fig. 43 Plotting the diagonistics

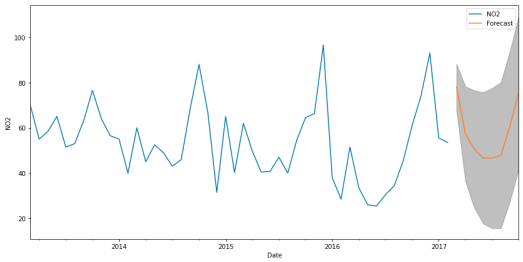


Fig. 44 Forecasting on unseen data

4 months) of unseen data in Fig. 44.To forecast for 8 months into 2017, we use SARIMA model (1,0,1)x(2,1,0,12). As we can see, the forecasted pollutant level of NO2 is peaking near the winter months as is evident from the previous few years.

## Carbon Monoxide (CO)

Carbon Monoxide on the other hand shows no decrease trend (Fig. 46) in the 4 years our dataset is in. Instead it shows an massive spike in CO output in 2016. However the dataset showcases seasonality which shows that carbon monoxide emissions are higher in winter rather than the warmer months.
Firstly we fit a SARIMA model to the data taking the first 44 months as training data and the next 4 months as test data (Fig.47). We do the same with a ARIMA model (Fig. 48). Due to the seasonality component, our SARIMA model works better than our ARIMA model (RMSE(SARIMA) = $157.14\mu gm^{-3}$ against RMSE(ARIMA) = $483.23\mu gm^{-3}$) In Fig. 49 we can see a few outliers in the standardized residuals and their normal Q-Q plot. A small amount of autocorrelation that still remains but otherwise, the model fits well. The Kernel Density Estimate shows that the distribution of the dataset is a bit

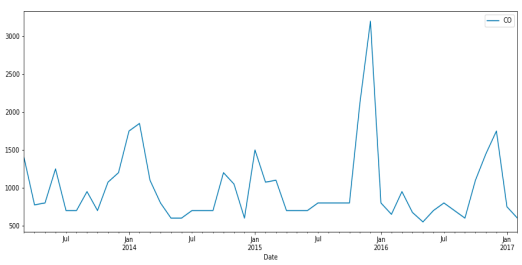


Fig. 45 CO emission over the years

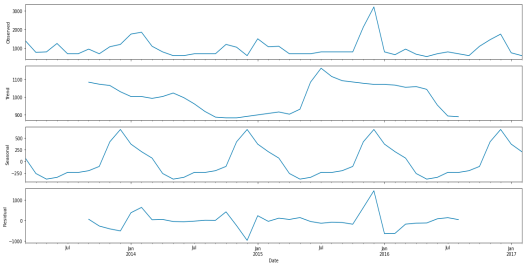

Fig. 46 Seasonal Decomposition

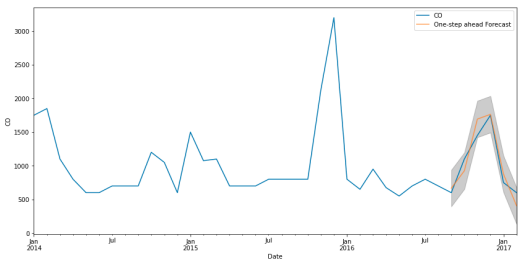


Fig. 47 Fitting a SARIMA model

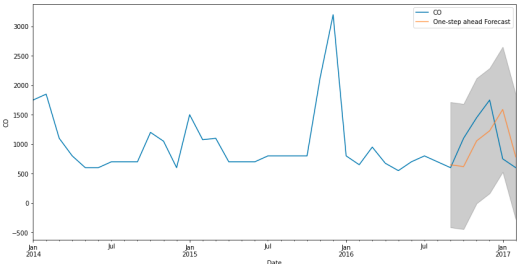


Fig. 48 Fitting an ARIMA model

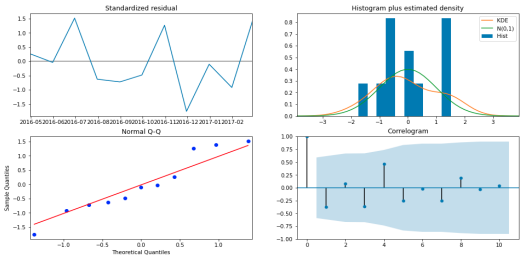


Fig. 49 Plotting the diagonistics

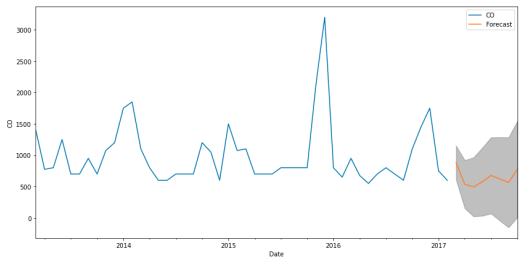


Fig. 50 Forecasting on unseen data

platykurtic in comparison to a normal (0,1) distribution and also shows a bit of a bimodal distribution.
Finally we use the SARIMA model to forecast 8 months (gives better visualization than 4 months) of unseen data in Fig. 50. To forecast for 8 months into 2017, we use SARIMA model (0,1,0)x(2,1,0,12). As we can see, the forecasted pollutant level of CO is actually lower in the mid months of 2017 as compared to previous years showcasing the improvement of pollutant CO level in Beijing despite the trend being higher.

## Ozone ($O_3$)

Like Carbon Monoxide and unlike the other pollutants on this list, Ozone shows an increasing trend over the 4 years (Fig. 52). It also shows a very clear seasonal component with the output of ozone increasing in the summer months (which makes sense because ozone pollution is caused by using refrigerators, air conditioners etc.)
Firstly we fit a SARIMA model to the data taking the first 44 months as training data and the next 4 months as test data (Fig.53). We do the same with a ARIMA model (Fig. 54). Due to the seasonality component, our SARIMA model works better than our ARIMA model (RMSE(SARIMA) = $6.06\mu gm^{-3}$ against RMSE(ARIMA) = $12.39\mu gm^{-3}$) and hence we move forward with the SARIMA model.

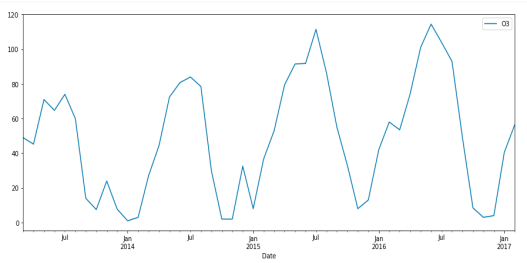


Fig. 51 O3 emission over the years

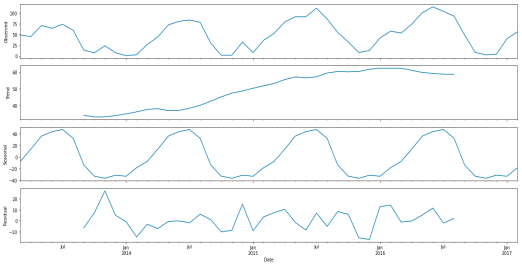


Fig. 52 Seasonal Decomposition

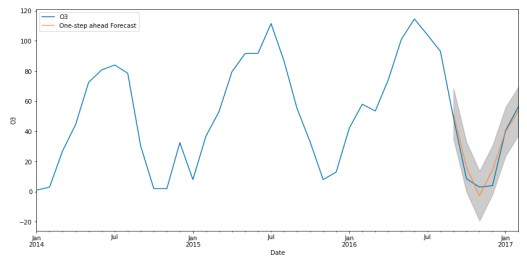


Fig. 53 Fitting a SARIMA model

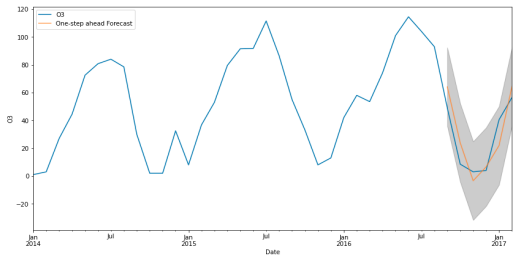


Fig. 54 Fitting an ARIMA model

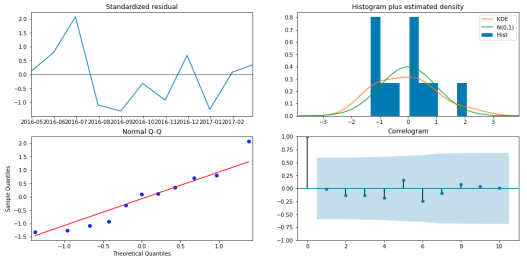


Fig. 55 Plotting the diagonistics

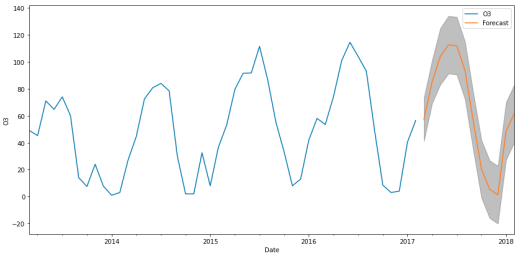


Fig. 56 Forecasting on unseen data

In Fig. 55 we can see a few outliers in the standardized residuals and their normal Q-Q plot. A small amount of autocorrelation that still remains but otherwise, the model fits well. The Kernel Density Estimate shows that the distribution of the dataset is a bit platykurtic in comparison to a normal (0,1) distribution.
Finally we use the SARIMA model to forecast 12 months (gives better visualization than 4 months) of unseen data in Fig. 56. To forecast for 4 months into 2017, we use SARIMA model (1,0,2)x(2,1,0,12). As we can see, the forecasted pollutant level of Ozone is actually higher as compared to previous years showcasing the deterioration of pollutant Ozone level in Beijing.

## Conclusion

Let's start with the conclusion that SARIMA model worked better for every dataset than the best ARIMA model (judging by their AIC). That confirmed our assumption that we made in the EDA that there is seasonality in the data due to which pollutant levels were better predicted by SARIMA.
Apart from that, we see that for most of the pollutants there is a higher output during winter rather than the summer months. This is mainly because of winter inversion. More about winter inversion can be found in these links.

1.) Why pollution levels skyrocket in winter
2.) Winter Inversion
Finally as the trend graph of nearly every pollutant (except Carbon Monoxide and Ozone) showcases that the pollutant levels in Beijing are improving every year since the inception of Beijing Clean Air Action Plan (2013-17). The implementation of a series of measures including energy infrastructure optimization, coal-fired pollution control and vehicle emission controls have successfully reduced air pollution. The forecast shows, the air pollutant levels to be decreasing the forthcoming years which has propelled Beijing from one of the most polluted cities in the world in 2008 to being a model of air quality management over the years. Beijing is currently ranked 122nd in major cities with polluted air quality.

## References

[1] Robert H. Shumway and David S. Stoffer. Time series analysis and its applications, 2017.

[2] Douglas C. Montgomery, Cheryl L. Jennings and Murat Kulahci (2008):"Introduction to Time Series Analysis and Forecasting"

[3]On Scale of 0 to 500, Beijing's Air Quality Tops 'Crazy Bad' at 755

[4]Life in China, Smothered by Smog

[5]Beijing clean air action plan

[6]Beijing ranks 122nd among major cities in air pollution