Air Quality in India

Forecasting and Analytics

Anit Mathew and Ritwik Katiyar

Summary

Our project aim is to analyze air pollution in India. To predict what the situation will be in the next year with reference to time. It turns out that winter is the most polluted time of the year. Pollution levels decrease during the summer and monsoons. The lowest contamination levels were recorded in August and September. Air pollution in June 2022 was the highest compared to June in the last five years. Our forecast model shows neither an increase nor a decrease in pollution levels in the coming year.

Introduction

Air Pollution is the contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere.PM2.5 are tiny particles in the air that reduce visibility and cause the air to appear hazy when levels are elevated. It is important to understand what level of Air pollution is surrounding us. India has been at the top of the index for the last few years. We have gathered aggregated data on Air pollution in India over 5 years from 2017 to 2022. In this project we will try to find in which month of the year air pollution is the highest in the country, is there any relationship between time of the year? Further, we will try to answer the question of whether air pollution will increase or decrease in India through machine learning.

The dataset used was uploaded by *'fedesoriano'* as Kaggle-Dataset: Air Quality Data in India The dataset contains the following variables:

- Timestamp Sort: Date and time when the data was collected.
- Year-sort: Year when the data was collected.
- Month-sort: Month when the data was collected.
- Day-sort: The day when the data was collected.
- Hour-sort: Hour when the data was collected.
- PM2.5: PM2.5 level of the country when the data was collected.

The first five rows of our dataset are as follows:

	Timestamp	Year	Month	Day	Hour	PM2.5
0	2017-11-07 12:00:00	2017	11	7	12	64.51
1	2017-11-07 13:00:00	2017	11	7	13	69.95
2	2017-11-07 14:00:00	2017	11	7	14	92.79
3	2017-11-07 15:00:00	2017	11	7	15	109.66
4	2017-11-07 16:00:00	2017	11	7	16	116.5

The Queries for our project are as follows:

1. Analyzing the air pollution dataset year-wise and visualizing the current scenario. Is the air pollution situation in India every year or is there any difference?
2. Identifying if there is any relationship between Air pollution and months of the year. Is the air pollution in India the same throughout the year or does it fluctuate on a month to month basis.?
3. Predicting Air pollution based on PM 2.5 levels. What will be the air pollution for the year 2023, will it increase or decrease?

Methodology

The data was collected was Kaggle and was easy to download from the website. Upon looking at the website, it was hard to understand the data. The data consisted hourly data of Air pollution in India for the past years.

1. Data Cleaning: The data contained more than 5000 records. To start with the research, it was important to clean the data.

2. For analyzing the year wise scenario, we combined the data in python for every year using groupby and calculated the average air pollution for that particular year. We pulled statiscal data and plugged the data into bar plots to better the visualize the scenario.

3. For identfying the relationship between hour, month and year, we segregated the data in python for every month wise, hourly wise, for each year separate using groupby and calculated the average air pollution for that particular year. We pulled statistical data and plugged the data into bar plots to better the visualize the scenario. Different approach helps us to gain insight on the actual situation.

4. For Machine learning, before we even start to run any models we need to make sure that our data is ready for machine learning. To accomplish that, we will be altering our dataset to be organized by day instead of an hour, for easy processing. We will be accomplishing that by taking the mean pollution level for every recorded day. We will also be removing any extra columns within our data set such as “day”, “month”, and “year” variables. Finally, we will be creating testing and a training split for our dataset. This would be useful to test our model against pre-existing values. We will be splitting the data into 70% training and 30% testing.

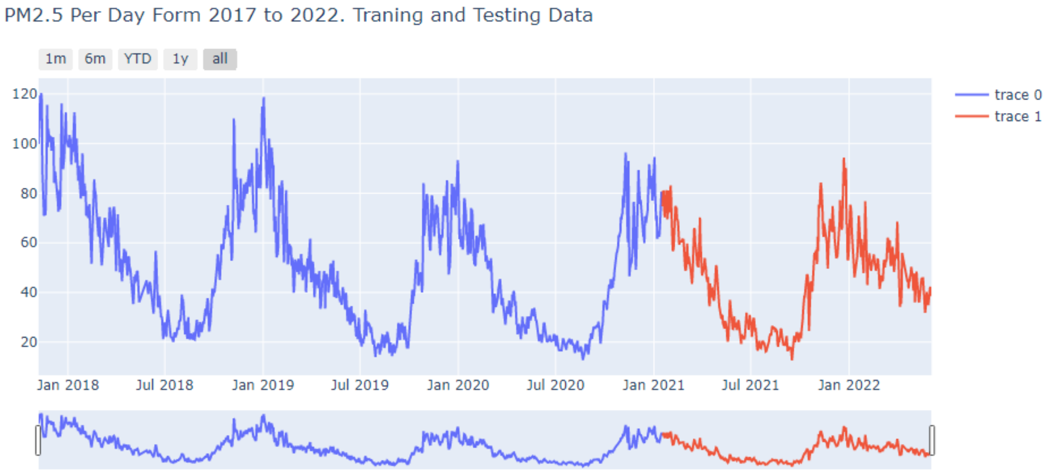
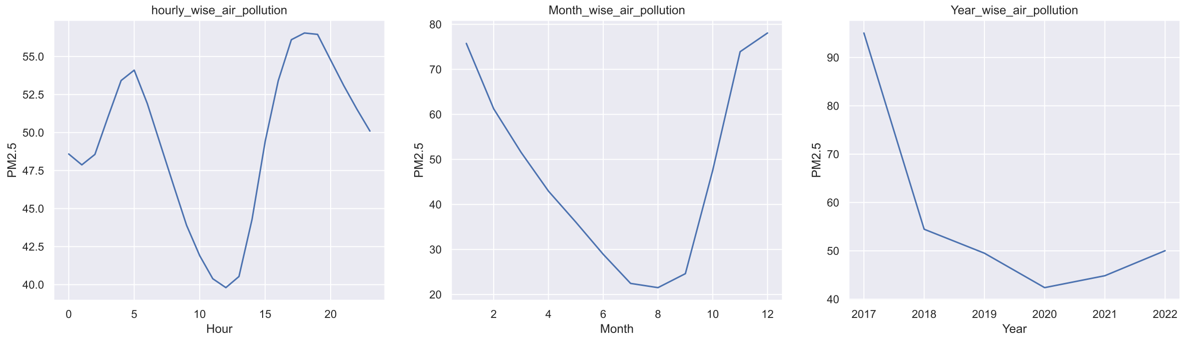


Figure 1: Training and Testing Split

Data Analysis and Visualizations

What is the situation in India in terms of air pollution? The data is vast and confusing to read. To better analyze the situation, we plotted the data on bar plots. First, we took a broader approach, we cumulated 5-year data into an hour, month and year. The grading of air quality is as follows:

Air Quality Category	PM2.5 µg/m3 Averaged Over an Hour	PM2.5 µg/m3 Averaged Over 24 Hours
Good	Less than 25	Less than 12.5
Fair	25-50	12.5-25
Poor	50-100	25-30
Very poor	100-300	50-150
Extremely Poor	More than 300	More than 150

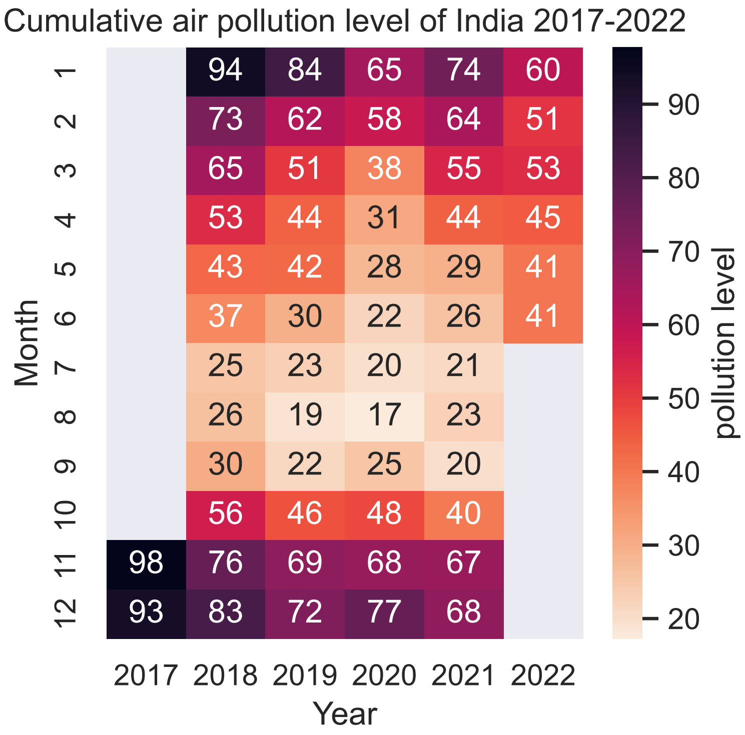


- In the first plot, we cumulated the hour-wise plot. We can see that there is wave-like formation. Overall the average air pollution is above 40 which brings somewhere between poor and very poor category. The pollution level peaks two times during the day. One is from 4 to 5 and the other one is from 5 to 7. The lowest it drops during noon.
- In the second plot, we cumulated month-wise data plot. We can see that the pollution level is at its peak during the start and end of the year. By looking at the plot, we see that July and August has the lowest level of air pollution which is less than 30 which brings them under the fair category.
- In the third plot, we cumulated year-wise data to plot. We can see that except for 2017, each year has average air pollution between 40 to 60. In 2017, air pollution was above 80. To better understand the results, we tried to dig a little deeper.

Bar plot may not provide accurate visualization which in turn hamper our analysis. To accurately summarize the data, we cumulated it and plotted it on a heat map.

The plot below provides an intense, deeper analysis of the whole situation of air pollution in India. Winters are the most polluted time of the month. The Pollution level has been declining with every coming year. During the summer and monsoon, pollution level declines.

Now, the question arises, will pollution increase or decrease in 2023. To understand the same, we focused on running a model through the data, so that we can understand what type of relationship does the data posses.



Forecasting

Linear Regression Model

As our first attempt at forecasting air pollution levels, we began by using a simple linear regression model. The linear regression model is used to find a relationship between two variables by fitting a linear equation to the observed data. This is a simple model that should be able to utilize the relationship between time and air quality to generate predictions on what air quality would be like.

The equation for the linear regression line would be as follows for our model:

$$AirQuality = \beta_0 + \beta_{Time} + \epsilon$$

Since dates aren't something a linear regression model can handle we will be adding a column that sets the first day of our data set as 0 and integrates for every day. This is done so that the model has a numerical value to compare the pollution levels to. We can then construct our model and see if our model is useful for making any predictions.

From the summary, of our model, we notice that our R-squared value is just 0.063 meaning that our model is only able to explain a mere 6% of the variation in our dataset. We can't use a model that can only explain such a small proportion of variance. Using a simple linear line to explain such a complex relationship is just isn't enough. To obtain better results and eventually forecast, we need to use a different model.

ARIMA Modeling

ARIMA stands for autoregression integration of a moving average. This model is specifically used for forecasting time series data. The equation for the ARIMA model consists of three parts.
The first part is autoregression: Auto regression refers to the changing variable within the model that regresses on its own and has prior values.
The second part is integrated: This refers to the differencing between observations.
The third part is the Moving Average: This takes into account the dependency between an observation and a residual error from the moving average model.

The ARIMA parameters for the model are as follows:
P = liner combination lags.
d = Number of times the raw observations difference.
q = Liner combination of lagged forecast errors. Simply, the size of the moving average window.
Before we even run the model we need to first determine the values for p,q, and d.

Determining the d-value

D-value refers to differencing. Which in turn refers to making the time series stationary. To test if our dataset is stationary we can utilize what is known as the Augmented Dickey-Fuller test (ADF). ADF is a unit root test, in other words, the test looks for the presence of a 'unit root' (A unit root essentially means a systematic pattern within a data set that is unpredictable.) To determine if the series is non-stationary. we can set up our hypothesis test as follow by taking an α = 0.05.
H_o = The Data set is not stationary
H_a = The Data set is stationary

```
ADF Statistic: -2.5720220909338756
p-value: 0.09891869282208948
```

We reject the Null hypothesis if the p-value < 0.05
From the summary above we can observe that the p-value is greater than 0.05. Which means we do not have significant evidence to reject the null hypothesis. Hence this means our data is non-stationary.

Since our data is non-stationary we can try to differentiate our data set. Differencing simply refers to calculating the difference between each variable within our dataset.
$y_t = Air_{pollution}$
$y'_t = y_t - y_{t-1}$
After differencing we can then run the ADF test again to see if differencing helped make our data stationary. If our data is still not stationary we would need to differentiate again. Running the test on the differenced data we get.

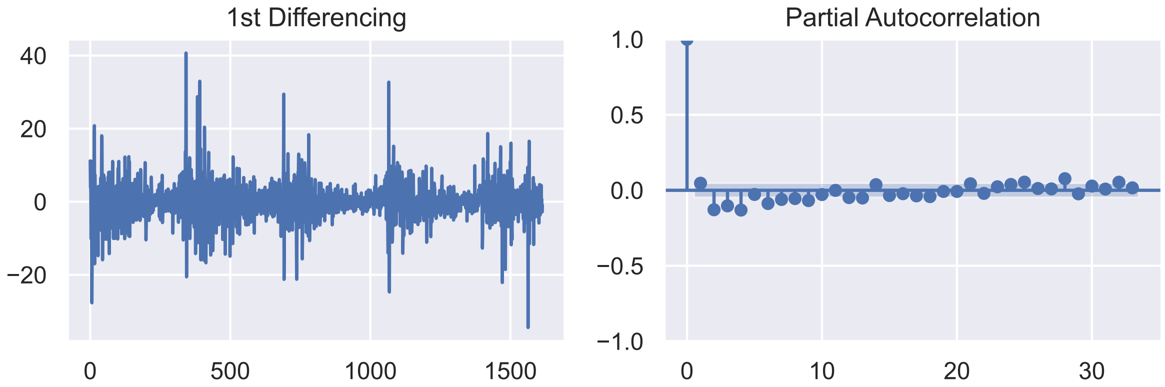
Setting up a new hypothesis test with an α of 0.05
H_o = The once differenced data set is not stationary
H_a = The once differenced data set is stationary
We reject the Null hypothesis if the p-value < 0.05

```
ADF Statistic: -13.175156123883427
p-value: 1.2275992099643855e-24
```

After differencing our p-value has decreased significantly meaning we now have enough evidence to reject the null and conclude that our data is now stationary and we were able to accomplish that via only one differencing once. Hence meaning that d = 1.

Determining p

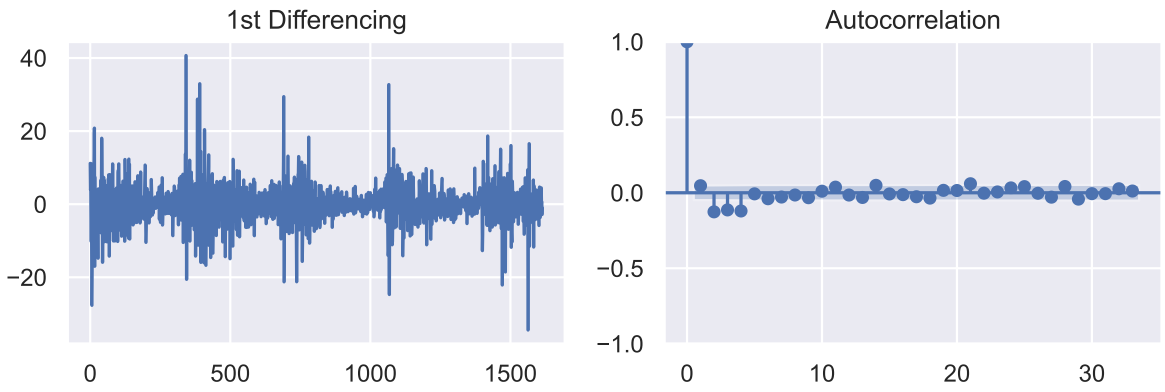
We can determine the value of the Auto regressive term p by taking a look at the Partial Autocorrelation plot (PACF). PACF is a plot that displays the correlation between the series and its lag. Mathematically speaking:
The PACF for our 1^{st}-order differenced data set looks as follows:



We can observe that the PACF lag 2 is quite significant since it is above the significant line (in blue). There seem to be other significant lags however, we will take the lowest number of lags.

Determining q

Similar to how we determined the value of p we can now look at the Autocorrelation Function plot (ACF) to determine the value of q. The value of ACF is a reference to the number of moving average terms or the moving average error of the lag. The following plot shows us the ACF of our data.



From the plot, we can ascertain that the ACF lag 2 is once again quite significant since it is above the significant line (in blue). Now that we have our p,q, and d values we can finally get ready for modeling. However, before we can model we need to first split our model into

Model 1 - ARIMA Model (p = 2, d = 1, q = 2)

We are now ready to run the ARIMA model. However, before we run any predictions we can take a look at our model to see if there are glaring issues that would need to be corrected.

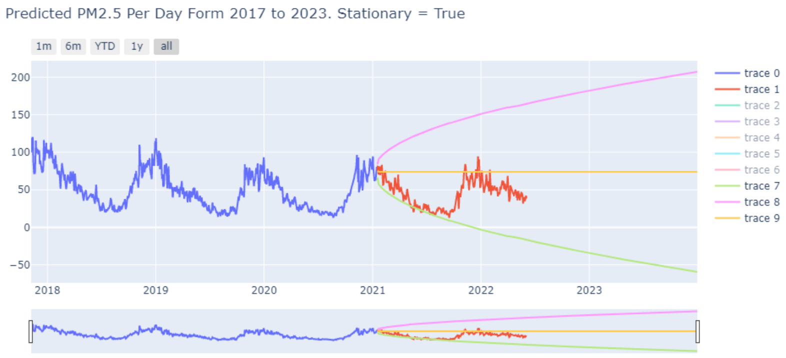


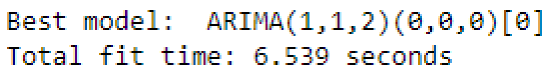
Figure 2: Results From ARIMA Model-1

The plot displays the upper and lower limits as predicted by the model. We can also observe the mean which seems to be a straight line. Meaning that based on our model the air pollution level would essentially remain the same for the coming year and there will be no decrease or increase in overall air pollution.

To test the strength of our model we can also take a look at the mean squared error. For the number of errors, our model had predicting the values. The mean squared error for our model is 34.03, which is relatively high. However, due to the nature of our data set, this is the best our model can produce. Based on manual analysis and determination of the p,q, and d values. Although, there is an ARIMA model that uses an automatic stepwise function to determine the best model based on the lowest Akaike Information Criterion (AIC).

Model 2 - Auto ARIMA

Auto ARIMA is a stepwise ARIMA model that picks the best p, q, and d values based on the lowest AIC score. The AIC score is known as an estimator of the quality of our statistical model for the given data. The AIC estimates the quality of the model, relative to other models for the same data, hence serving as a way to select a model.



```
Best model:  ARIMA(1,1,2)(0,0,0)[0]
Total fit time: 6.539 seconds
```

Figure 3: Results From Auto ARIMA

We can see that model suggests that we use p = 1, d = 1, and q = 2 to better fit our model. This is a little off from what we did in our previous model. However, we will go with the model's output and try to validate if the auto ARIMA did in fact provide us with better results.

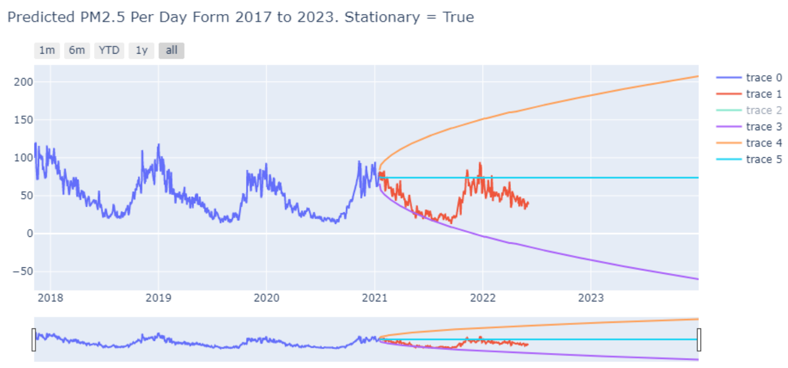


Figure 4: Results From Auto ARIMA Model-2

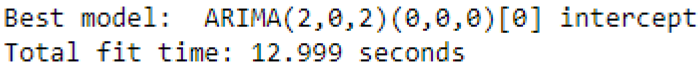
The model's predicted output looks very similar to what we had in our previous model. Since there is only an insignificant change (p = 1 instead of 2) it makes sense that we don't notice a substantial increase or decrease in the overall predictions. The mean squared error for our first model was approximately 34.03. However, the mean squared error for our auto ARIMA model is approximately 34.10. There is a slight decrease in the quality of our model, and overall there doesn't seem to be a major difference between the two variations. However, it was interesting to see the results nonetheless.

Model 3 - Assuming Our data is Stationary

If we go back to the point where we determined the d value. We noticed that the p-value given by the ADF test on our data set was:

```
p-value: 0.09891869282208948
```

If we set α of 0.1 We can reject the null and assume that our data is stationary. Also from looking at our PACF and ACF plots one can notice that we don't seem to be able to see strong collinearity. Which could mean that we are over stationarizing our data. If we run auto-ARIMA again, however, this time we force the model to assume that our data is stationary. We get the following results:



```
Best model:  ARIMA(2,0,2)(0,0,0)[0] intercept
Total fit time: 12.999 seconds
```

Figure 5: Results From Auto ARIMA Assuming Starionary

Based on the auto ARMIA, we need to set our p and q values as 2 for both. Setting these values we can run our model and take a look at the model residual diagnostics one again to see if there are any major issues.

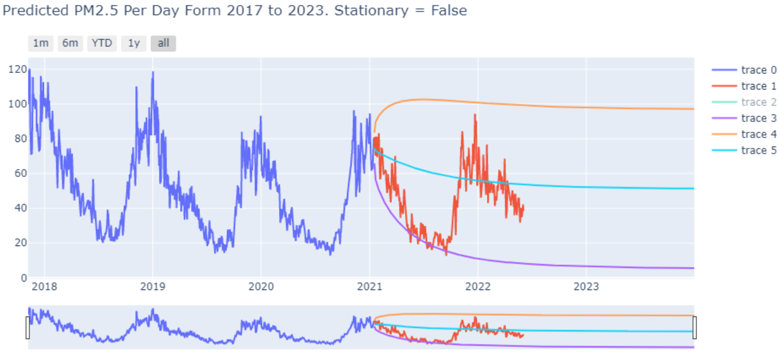


Figure 6: Results From Auto Arima Assuming Starionary

We notice that our mean squared error has gone down substantially to 24.04 from around 34.10 in our outher models. We are not sure why the mean square error has gone down. However, it could be possible that we may be overfitting our model. We can notice from figure 9 that our upper and lower bound predictions have gotten narrow compared to the previous models suggesting possible overfitting. However, despite the shortcomings, it is interesting to note that out of the other two models this is the only model that predicts that the air pollution level would decrease. Although, the decrease in pollution level is rather minor. Finally, to get a different prediction, we decided to run another time series model known as ETS.

Discussion

Forecasting air quality is a complex task due to environmental dynamics, unpredictability, and changes in pollutant status and time. The serious consequences of air pollution on people, animals, plants, monuments, climate and environment require continuous monitoring and analysis of air quality, especially in developing countries like India. The resuts that we can take back from this case study are:

- Winters are the most polluted time of the month.
- Winter of 2017 was the most polluted time in the past 5 year.
- Pollution level has been declining with every coming year.
- During the summer and monsoon, pollution level declines.
- Best month is the month of August and September.
- January 2022 had the lowest air pollution level during winters comparatively and gradually declining.
- But as per the plot, the month of June 2022 had the highest air pollution level if compared with June of the last 5 years.
- Our prediction model suggests that there will be neither an increase nor a decrease in pollution levels for the next year.

Having said that, there are still a lot of factors, which can be a reason for air pollution. There is further scope to this project. This project can be a big research project by extracting more data.

References

https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced#:~:text=In%20plain%20

https://www.statsmodels.org/dev/examples/notebooks/generated/ets.html

https://otexts.com/fpp3/holt.html