Introduction

Delhi, the capital city of India, ranks 4th on the list of the World’s Most Polluted Cities, and has been consistent with it for the past few years. Considering the adverse health effects these pollutants have on the residents, there is a pressing need to understand their patterns in order to forecast the upcoming levels and come up with remedial measures to mitigate their effects.

With this in mind, the objective of the project was to conduct time series analysis of a major (as described later) pollutant, PM2.5, to identify its pattern and trend. This analysis majorly focuses on observing the pattern, and drawing insights which can later be used to control the relevant factors.   
Fortunately, a suitable dataset for this topic was available on Kaggle. The said dataset had Hourly Measurements of Concentration (in varying units) of Air Pollutants across various cities, for the period of January1, 2015 to June 16, 2020. Originally, these observations were taken across multiple stations in every city, but their values were later averaged per city, leading to a single record per city. The pollutants in the dataset were: PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, and Xylene. Along with these pollutants, there were two additional columns for AQI and AQI Bucket. The AQI or Air Quality Index is a metric devised to quantify the Air Pollution of a certain place at that specific time, based on the threshold concentrations of various pollutants. The following is a formula for Subindex of each pollutant:

Ip = ((IHi −ILo) \* (Cp −BPLo ) / BPHi−BPLo) + ILo

Where Ip = the index for pollutant p  
Cp = the truncated concentration of pollutant p  
BPHi = the concentration breakpoint that is greater than or equal to Cp BPLo = the concentration breakpoint that is less than or equal to Cp  
IHi = the AQI value corresponding to BPHi  
ILo = the AQI value corresponding to BPLo

Based on the above formula, subindices were calculated per pollutant for each available time frame, and the maximum of these subindices was selected as the AQI for the area. To identify the most significant pollutant, the highest subindex per time frame along with its corresponding pollutant must be noted. This information can be used to determine the most detrimental pollutant overall by aggregating the data. The pollutants with the maximum index for each time frame in Delhi are presented below, indicating the primary cause of pollution.

As seen, PM2.5 was responsible for more than half of the instances of AQI and is thus considered the focus of this project. Going forward, it is assumed that the data consists of PM2.5 for Delhi alone.

Data Preparation

As mentioned earlier, the data considered had recorded concentrations of various pollutants for each hour. While trying to establish the overall trend throughout the year, an hourly data would be cumbersome to model, so they were averaged per day to condense it to daily data records. Doing so revealed 2 instances of missing values, which happened to be consecutive. Thus, in order to deal with these, we replaced the null values with the most recent non-null value, and the following time series plot was obtained for Daily PM2.5 Concentrations in Delhi.

As seen in Figure 2, the data seems to be seasonal in nature, indicating that need of Seasonal Models. The range of values from 0 to 700, being inconvenient to deal with, necessitates a transformation to scale down the values to comparable levels. Thus, a log transform was applied, and the following plot was obtained.

Modeling Methodology

Seasonal Decomposition

The team started off with the most basic Seasonal Decomposition Model using Moving Average with a period of 365, since the period of seasonality is ought to be 365. The decomposition plots obtained are shown in Figure 3

Based on the above figure, significant lags were noticed up to the order of 4 in both the ACF and PACF Plots. Not only that, but as demonstrated in Figure 4., these lags become quite significant near the order of 365 (the assumed seasonality), thus making this model unfit.

Auto SARIMA

The next approach was to capture the Seasonality while factoring for the observed ARIMA nature from the previous method. Much like the previous one, the seasonality was assumed to be that of 365, and a SARIMA model was fit using the Auto ARMA function, passing the seasonality as one of the arguments. The best model chosen was SARIMA (5,1,1) (0,1,0) [365], and the following is the analysis of its residuals.

From the get-go, one can notice that while the residuals exhibit no pattern of seasonality, there is a very significant lag at the order of 365. Not only that, the QQ Plot also exhibits heavier tails, indicating that the residuals may not be normally distributed, violating one of the assumptions about errors. Thus, this model is deemed to be unfit as well.

This highlights a serious issue with previous modeling techniques: the assumption of a large seasonality. These models require the declaration of a fixed seasonality, and then model each term in the seasonal bracket separately. For example, in a Seasonal Decomposition Model with a period of 365, every term occurring at an interval of 365 is averaged and assigned a coefficient, which appears in the seasonal plot. Similarly, in SARIMA, the model looks back at the previous 365th term, along with a few other terms around it based on the model order and coefficients. By doing so, the model becomes overly granular and hyper-focused on terms that occurred a year apart, failing to capture the bigger picture of seasonality.

This calls for a more general technique to capture seasonality; one which doesn’t need rely solely on past values for given seasonality.

Dynamic Harmonic Regression

The Dynamic Harmonic Regression Technique (DHR) extracts seasonality using Fourier Transform by converting each peak or trough of seasonality into a signal with a fixed frequency. Thus, for modeling this technique, we assume a Linear Regression based ARIMA Model, where the regression terms are the sine and cosine Fourier Components, and the remaining error is modeled using an ARIMA process. The required number of components (K) for Fourier analysis is a hyperparameter to be defined by the user. DHR can model multiple and non-integer seasonality, making it advantageous.

The above shows the forecast of DHR models fitted for various values of K. As shown above, increasing the number of components, increases the resolution or specificity of the seasonality, with 1 showing the most general model, while 6 is the most specific.

Based on the above few plots, we decide on 4 as the appropriate number of Fourier Components. Thus, a DHR model with K=4 was fit, and the residuals obtained were as follows.

As seen, the ACF and PACF Plots show no significant lags, indicating stationarity, so far.

TBATS

TBATS (Trigonometric Seasonal Exponential Smoothing State Space Model) is one such time series forecasting technique that can handle multiple seasonalities and changing trends. It also applies Box Cox Transformation to handle non-linear trend, and then uses ARIMA to model the error. Unlike the previously mentioned DHR, which uses Fourier Transform to extract seasonality, TBATS uses a combination of trigonometric functions and exponential smoothing to capture multiple seasonalities with varying frequencies. Furthermore, TBATS incorporates a trend damping parameter that allows the model to adapt to changing trends, which DHR was incapable of.

Fitting our data in the TBATS function with a pre-specified order of 365.25, results in a TBATS Model of the order (1, {2,3}, -, {<365.25,4>}). Here, the first ‘1” stands for Box-Cox Transformed, the {2,3} is the ARMA Error, and {365.25,4} means that 4 pairs of sine and cosine terms were needed to model the seasonality. The forecast wherein the seasonal pattern looks similar to DHR is shown below.

The residual plot along with their corresponding ACF and PACF Plots are shown above. From the above plots, it is evident that the resulting series is indeed stationary, but further analysis is done before confirming it.

Results

The following table displays the result of various tests conducted on the residuals and their squared values.

As seen, there is some ARCH effects in both the models which is handled easily by fitting a ARCH (1) Model. The table below shares the results again.

Thus, the TBATS model is selected as the final selected, although it is very similar to DHR for the sole reason that it is a more elaborate version which takes the previously known information of approximate 365 seasonality, and then models the data. The overall Trend and Seasonality of the model are shown below.

Firstly, an overall decline in PM2.5 trend is observed. This dip is especially evident in the last few months, corresponding to the months of lockdown following the COVID-19 pandemic. One interpretation of this could be the inverse relationship of human activities and PM2.5 concentrations. As for the seasonal pattern, the winter months are shown in blue, the summer ones in red and the monsoon ones in orange yellow. From this, two things are implied – 1) Inverse relationship of Temperature and PM2.5 concentration. 2) Inverse Relationship of Rainfall and PM2.5 concentrations

Conclusion

* TBATS and DHR Models are preferred for Long Term Seasonality
* PM2.5 has been showing a decreasing downward trend.
* Weather/ Seasons a major factor affecting the Pollutant Concentration

Future Scope

* Look into Spatial-Temporal Distribution of Pollutants
* Factor in other variables like Temperature, Wind, Traffic for better analysis
* Define a similar pattern for Hourly Distribution of Data

References

1. Dataset Link: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india?datasetId=630055&sortBy=voteCount&select=city\_hour.csv
2. Indian AQI Formula Specific: https://www.pranaair.com/us/blog/what-is-air-quality-index-aqi-and-its- calculation/
3. TBATS Model Research Paper: https://robjhyndman.com/papers/ComplexSeasonality.pdf
4. Textbook for DHR: https://otexts.com/fpp2/dhr.html