**Mumbai Air Quality Prediction.**

**ABSTRACT**

With the development of industry and the consumption of fossil fuel, air quality is worsening. In recent years, pollution has increased significantly in many parts of Maharashtra. Prediction of air quality is a topic of great interest in air quality research due to direct association with health effect. The prediction provides pre-information to the overall population of the area about the status of pollution on which they can take precautionary measures and can protect their health. The problem arises when the level of SO2, NO2 and residual suspended particulate matters in the air increases than that of theirs restricted level. These are the major air pollutants caused by the dominant usage of diesel, petrol CNG fuels by vehicles and by industries. The trend of these three pollutants is quantitatively predicted based on comprehensive predicting model. In this paper, the Prophet Algorithm, open source software, is applied to predict the trend of air pollution in the city of Mumbai, Maharashtra. The Prophet is machine learning algorithm to forecast and also to predict time series data. It is based on additive model where non linear trends are fit with yearly and weekly seasonality. The graphical results are generated after using this algorithm which shows the trending pattern of the pollutants in the air of Mumbai.

**INTRODUCTION**

The alarming growth in atmospheric pollution has led many countries in the world to establish severe laws and regulations defining air quality and the required standard of emission levels. Nowadays, the continuous and strict monitoring of air pollutants is has also been adopted. Particular attention is being devoted to urban area where the problem of atmospheric air pollution are especially worrying, because the pollutant level in these areas are high and many people exposed to health hazard are growing constantly. Therefore the administrations of these regions are installing and actively monitoring the air pollutant matters in order to keep them under control. Sulphur dioxide (SO2), Nitrogen dioxide(NO2) , carbon dioxide(CO2) and respirable suspended particulate matter(RSPM) are the major pollutants caused by vehicles and industries. The forecasting method analyzes the sequence of historical data in a period of time to establish the forecasting model. Air quality reports for the various region of the country are published regularly. As a result, data are getting accumulated and this results in generating various reports, including statistical one in order to find different pattern among those data.

In our paper, two hazardous gas, sulphur dioxide and nitrogen dioxide are considered because both of these gases are most harmful and we have collected the dataset containing these two gases. The main source of SO2 in the air is industrial activities and burning of coal and mineral that contain sulphur, and the source of NO2 in the air is burning of fossil fuel and from exhaust of vehicles. Scientific research has proven that these gases have many negative effects on human health. Sulphur dioxide is significantly a toxic gas, when breathed in it irritates nose, throat and can cause coughing wheezing and shortness of breathe. It can also cause inflammation and irritation of the respiratory system, whereas Nitrogen dioxide, another toxic gas for human beings, can form nitric acid with water in the eyes, lungs, mucus membranes and skin. Exposure to high concentrations of NO2 can cause lung irritation and potentially lung damage. In this paper, the Prophet algorithm, a open source software which is just recently launched by the Facebook is implemented to predict the trend of pollutant in the air of Mumbai, by taking dataset from the month January 2018 to March 2018. This algorithm is developed by Sean J. Taylor and Ben Letham from Facebook. It was launched recently released as open source software with an implementation in R software.

**LITERATURE SURVEY**

The prediction of air quality is becoming essential for minimizing the environmental imbalance further effectively known as air pollution. There are various available tools and techniques for prediction and analysis of air pollution. The emergence of advance technique from traditional computing method to readdress the problem is known air quality prediction.

The traditional approach for prediction of the quality of the air uses mathematical and statistical techniques. In these techniques, at first a physical model was designed and then data is coded with mathematical differential equations. But such methods suffers from disadvantages like they provide limited accuracy as they were unable to predict the extreme points i.e. the pollution maximum and minimum cut-offs cannot be determined using such approach. Also, such methods were lengthy and inefficient approach for better output prediction. But with the advancement in technology and research, an alternative to traditional methods has been proposed i.e. Artificial Intelligence (AI) techniques can be used for prediction purposes.

Various data mining techniques have been used by various researchers in the air pollution to study the behaviour of air pollutants in the atmosphere. Sheng-Tun Li et al (2006) has assessed one of the spatiotemporal data mining techniques, the cluster analysis, on air pollution. Multi-scale input data was used as an input to the SOM neural networks and the spatio-temporal change was studied. The study showed that the number of clusters reduced as the scale of data input increases. This indicates that the scale is an important factor in the cluster analysis. The study showed that Self Organized Maps (SOM) is an excellent visual tool for studying the inner structure of the data transformed and provides the capability for hard clustering.

Nikov et al. (2005) proposed AirPolTool; a web-based tool for predicting air pollution in Istanbul. The tool models the association between local meteorological data and concentrations of air pollution indicators like SO2, PM10 and CO with the help of neural networks. The tool is very easy to use giving a three-days prediction of air pollutant and necessary warning signals and appropriate actions to be taken by the administration to reduce the level of the particular air pollutant to a level so that it is no longer harmful.

Athanasiadis et al. (2007) used classification techniques, in an environmental management system to help the decision making. The system supervises the ambient air quality and generates warnings in case of occurrence of events. They used an extensive data set from the real world which has ensured reliable prediction procedures that are capable of taking decisions at the operational level. These methods have added value to the traditional approaches in air quality assessment, thus improving the performance of the forecasting methods. A C4.5 algorithm for decision tree generation was applied in WEKA and the study produced reliable models that had a forecasting accuracy exceeding 93%. This classification approach made it possible to handle data uncertainties involved in an air quality and thus support decision makings in an operating timeframe.

**OBJECTIVE**

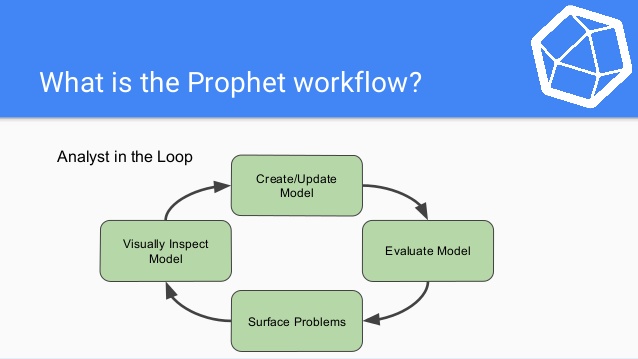
1. The dataset, which contains the value of pollutant in the air of Mumbai, is collected from the month of January 2018 to March 2018 and is pre-processed.

2. Pre-processing is done to remove unwanted attribute in the dataset. Real world data is often inconsistent, incomplete and is likely to contain many errors. Data pre-processing is proven method to resolve such issues.

3. The algorithm is then applied to predict the future pollutant level based on the given dataset.

4. Statistical reports are then generated of the output, which are produced by the applied algorithm.

**INNOVATION COMPONENT**



**METHODOLOGY**

In this paper, the Prophet Algorithm is applied to predict the trend of air pollution in the city of Mumbai, Maharashtra by taking the dataset from the month of January 2018 to March 2018. The Prophet is a procedure for forecasting time series data. It is based on additive model where non linear trends are fit with yearly and seasonality. It is also designed to have the intuitive parameters that can be regulated without knowing the underlying details of the model. Our implementation is done in the open source software R by using this algorithm. This algorithm can be used to predict the trend with hourly, daily or weekly observations with at least few months of history. With this software the result predicted are often accurate as those produced by skilled forecaster, with minimal efforts. The prophet has four main components:

i) The prophet algorithm automatically detects changes in trend by selecting change points from the data.

(ii) A yearly seasonal component is model using Fourier series.

(iii) A weekly seasonal component uses dummy variables and,

(iv)The user provide important holiday in dataset.

Before running this algorithm in R software, we need to import the libraries files “prophet”, an essential file for executing all the command of the algorithm, “gplot”, a file for plotting the graph in the algorithm. After successfully importing the required the csv file of our dataset is loaded into the software. The prophet library expects input as a dataframe with one column containing the time information and other column containing the metric value that we need to forecast. This requires the algorithm to generate the “ds” column and “y” column. The “ds” column contains the dates and time whereas the “y” column contains the values. The algorithm uses only these two values to predict the weather and all other irrelevant data in the dataset are ignored. As the dataframe is ready, we are ready to use the Prophet library to produce prediction. The “prophet” command is use for prediction. Conveniently, we do not have to concern ourselves with manually creating this dataframe, as Prophet provides the “make\_future\_dataframe” helper function. Along with this command we need to specify the filename and the frequency and accordingly the algorithm will generate that. The dataframe of future dates is then used as inputs to the “predict” method. The algorithm return a large dataframe with many interesting columns but we minimised our output with the column which are relevant for predicting which are:

(i) ds – the datestamp of the predicted value

(ii) yhat – the predicted value of y.

(iii) yhat\_lower – the lower bound of the prediction and,

(iv) yhat\_upper – the upper bound of the prediction.

Prophet also provide the convenient way of plotting the graph by using “plot” function. Given below is the dataset attributes which we have used in our study.

|  |  |  |
| --- | --- | --- |
| **VARIABLE NAME** | **VARIABLE DESCRIPTION** | **DATATYPE** |
| Sr No | Serial number. | Integer |
| Date | Gives the date and time of recording the observation. | Date |
| So2 | Sulphur  Dioxide | Integer |
| No2 | Nitrogen Dioxide | Integer |
| Rspm | Respirable Suspended Particulate Matter | Integer |

**RESULTS**

* Initial Data (Raw Data)

In our paper, we have imported the dataset of Mumbai, Maharashtra from the month of January 2018 to March 2018, which contains the information of contamination of hazardous pollutants in the air of Mumbai. R studio application software is chosen in our study in which an open Prophet algorithm is applied to extract the trending pattern from the dataset.

The process is begun by importing the required package for the algorithm and the dataset which we have used in our study. The graph of raw data is then generated with giving the commands in the application. The raw graphical representation of the data is shown below:

**A picture containing outdoor, sky, wall

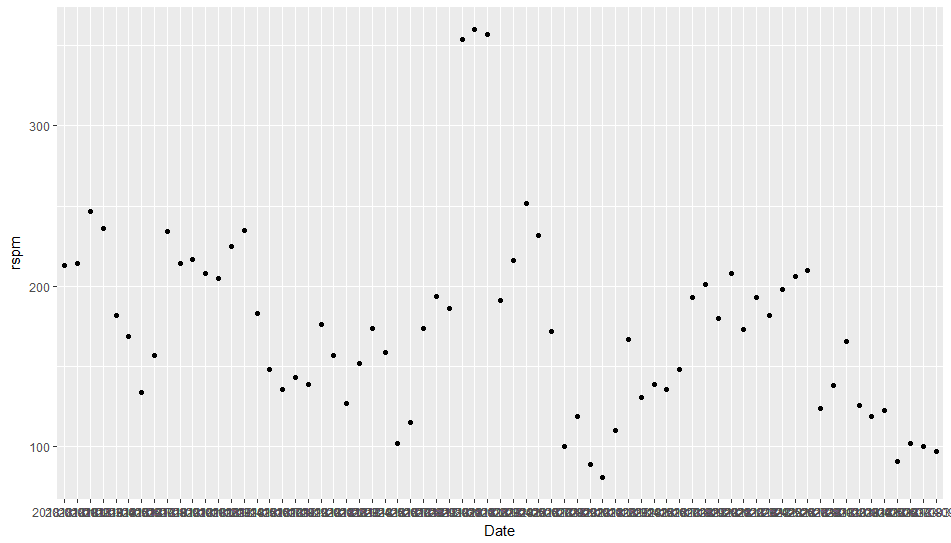
Description generated with high confidence**

Raw SO2 Data

A picture containing outdoor, wall, sky, bird

Description generated with very high confidence

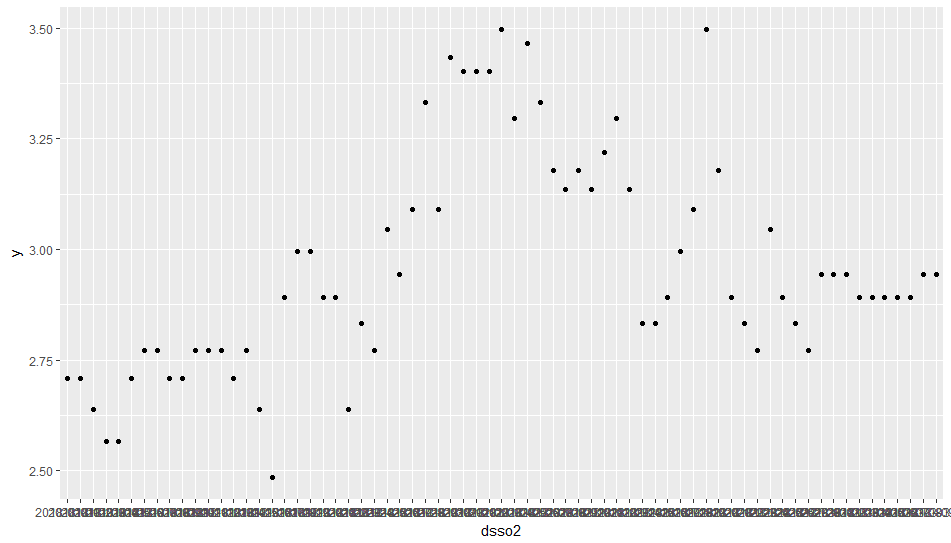
Raw NO2 Data



Raw RSPM Data

* **Pollutant SO2**: After displaying the raw data of the pollutant SO2, a column named ‘dsso2’ is created as Prophet algorithm requires one column named dsso2 which contains dates and another column named ‘y’ storing the numeric values. This two column are treated as input for prediction. The two required column are created by duplicating the two existing columns ‘time’ and ‘so2 values’.

The ‘dsso2’ graph is shown below:



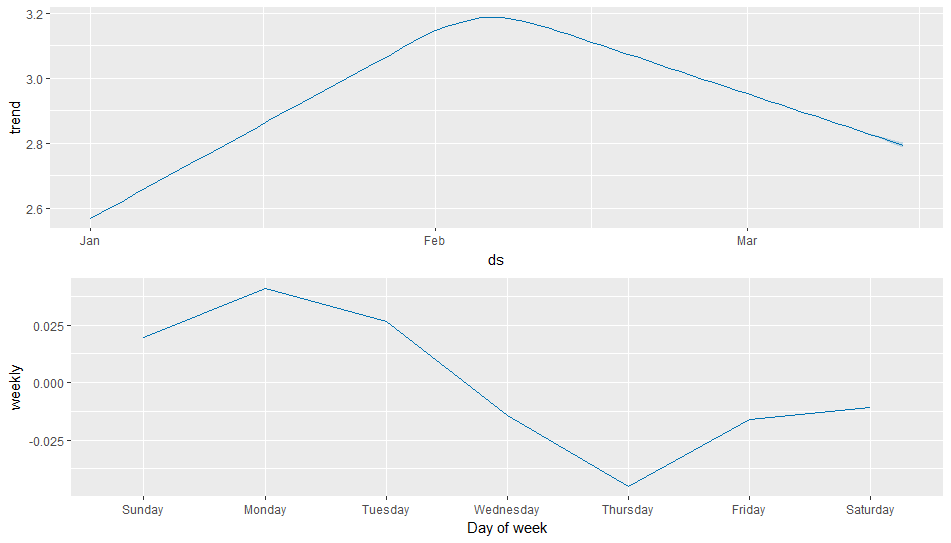
dsso2 graph

The dataframe is then created and to predict the weather, all you need is to call “prophet” function and based on the model and dataframe, Prophet predicts the value of future on monthly, trend and weekly basis. The outcome of study on the pollutant SO2 in graphical representation is shown below:

A close up of a map

Description generated with very high confidence

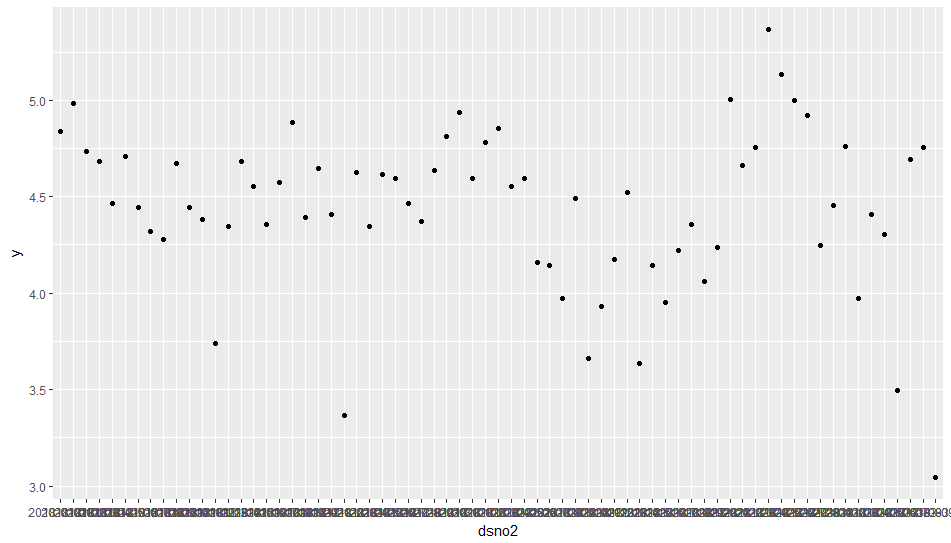
SO2 prediction on monthly basis



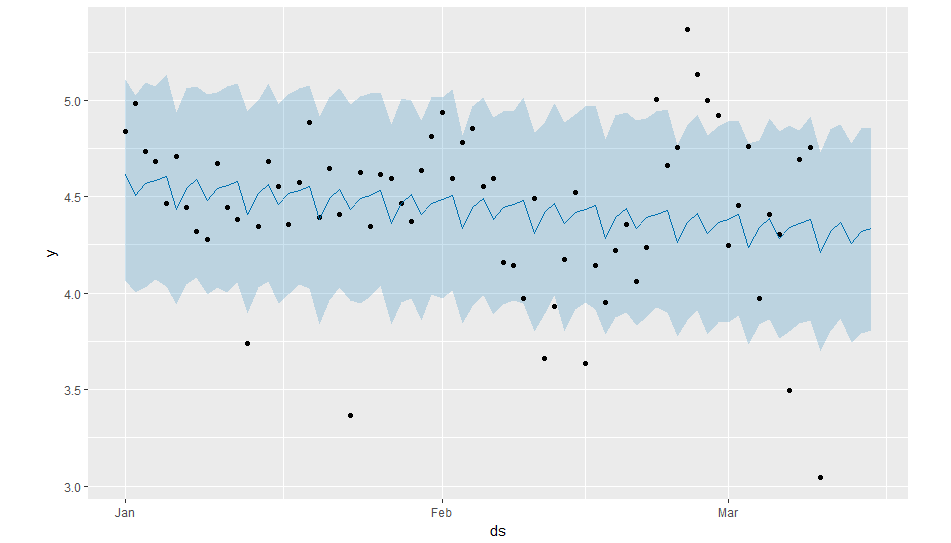
SO2 prediction on trend and weekly basis

* **Pollutant NO2**: After plotting the raw data of the pollutant NO2, another column named ‘dsno2’ is created with dates and another column named ‘y’ with numeric values. The two required column are created by duplicating the two existing columns ‘time’ and ‘no2 values’ from the dataset. The dataframe is then created and stored ‘dfno2’ which is used as input to Prophet.

The ‘dsno2’ graph is represented below:



To predict the contamination level for the future, “prophet” function is called in the R software and the algorithm will generate the output based on the model and the dataframe ‘dfno2’. A subplot is created of the dataframe by calling the function “predict” and this will return a dataframe with details about all the prediction with the timestamp. For each timestamp, ‘yhat’ value will generated along with ‘yhat\_lower’ and ‘yhat\_upper’ which shows the range of uncertainty level. The Prophet algorithm predicts the value of future on monthly, on trend and on weekly basis. The future prediction for the pollutant NO2 in graphical representation is shown below:



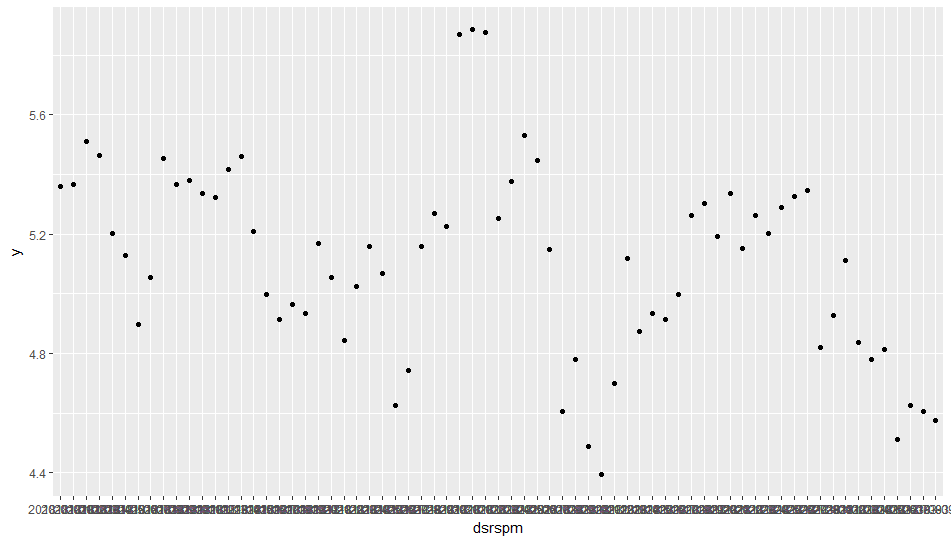
NO2 prediction on monthly basis

A close up of a map

Description generated with high confidence

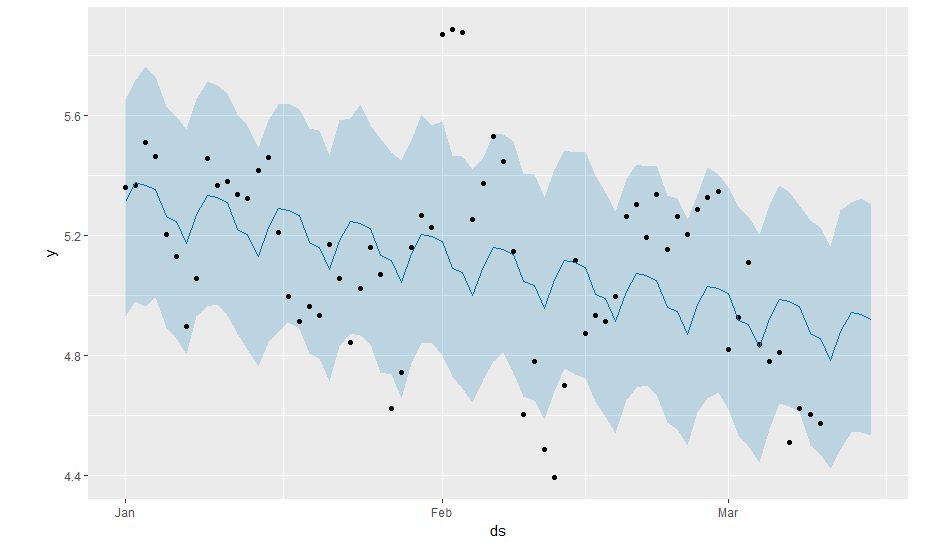
NO2 prediction on trend and weekly basis

* **Pollutant RSPM**: After plotting the raw data, a new column named ‘dsrspm’ containing the time and date and another column named ‘y’ with numeric values to be created.

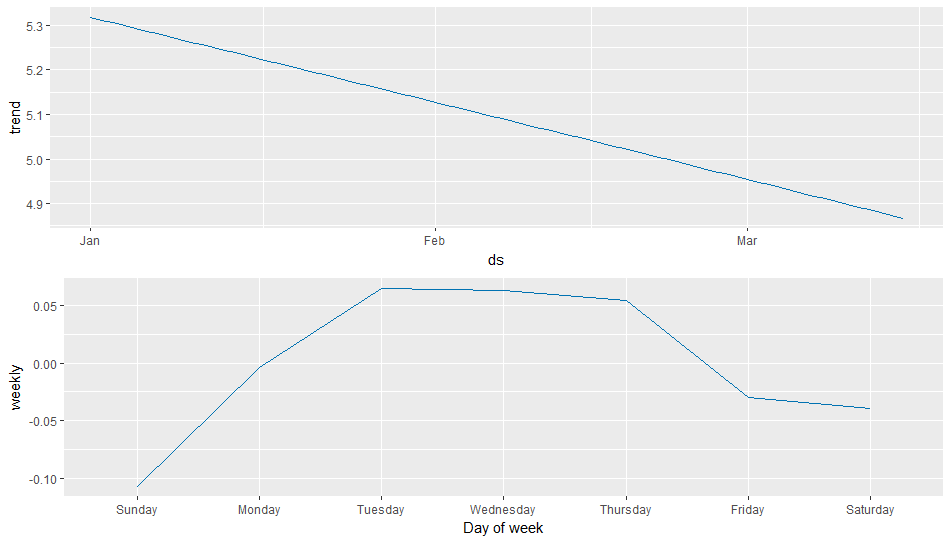


dsrspm graph

The dataframe is then created from the above column and stored ‘dfrspm’ which is used as input to Prophet algorithm. To predict the contamination level for the future ”predict” function is called in the R software and the algorithm will generate the output based on the model and the dataframe ‘dfrspm’. To show uncertainty level, further subplots is created from ‘dfrspm’ value, containing the ‘yhat’ value, ‘yhat\_lower’ value and ‘yhat\_upper’ value for each timestamp. The Prophet algorithm predicts the value of future on monthly, on trend and on weekly basis. The future prediction for the pollutant NO2 in graphical representation is shown below:



RSPM prediction on monthly basis



RSPM prediction on trend and weekly basis

**CONCLUSION**

Prophet is recent and one of the most advanced prediction technique developed by Facebook in 2017. Using the Prophet algorithm to generate prediction trend turns out be easy and interesting. There are several ways to generate the prediction and to inspect the results generated. There are libraries function that have more functionality and flexibility. The prediction trends obtained have high accuracy and can be used for better prediction.

**REFERENCES**

1. <https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-prophet-in-python-3>

2. <https://algobeans.com/2017/04/04/laymans-tutorial-time-series-analysis/>

3. <https://blog.exploratory.io/an-introduction-to-time-series-forecasting-with-prophet-package-in-exploratory-129ed0c12112>

4. <https://arnesund.com/2017/02/26/using-facebook-prophet-forecasting-library-to-predict-the-weather/>