Applying data mining technique to predict trends in air pollution in Mumbai

To cite this article: Jayakumar Sadhasivam et al 2021 J. Phys.: Conf. Ser. 1964 042055

View the article online for updates and enhancements.
Applying data mining technique to predict trends in air pollution in Mumbai

Jayakumar Sadhasivam1*, V Muthukumaran2, J Thimmia Raja3, V Vinothkumar4, R Deepa5 and V Nivedita6
1School of Information Technology and Engineering (SITE), Vellore Institute of Technology, Vellore, India.
2Department of Mathematics, School of Applied Science, REVA University, Bengaluru – 64.
3School of Computing, Kalasalingam Academy of Research and Education, Tamilnadu-626128, India.
4Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore.
5Chaitanya Bharathi Institute of Technology (A), Hyderabad, Telangana, India.
6Department of Computer Science and Engineering, Dr. MGR Educational and Research Institute, Chennai.
E-mail: *jayakumarsvit@gmail.com

Abstract. 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. 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.

1. Introduction
Nowadays, the continuous and strict monitoring of air pollutants is of great importance in the process of evaluating regulatory control measures related to air quality [1]. Many countries are installing and actively monitoring the air pollutant matters in order to keep them under control. Air quality reports for the various region of the country are published regularly [2][3]. 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 [4][5]. Scientific research has proven that these gases have many negative effects on human health. Sulphur dioxide is significantly a toxic gas that can 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 [6]. Exposure to high concentrations of NO2 can cause lung
irritation and potentially lung damage. In this paper, the Prophet algorithm, an 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 [7]. 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 [8].

2. Proposed Method
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 data frame 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 data frame, 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 timestamp 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.
3. Results
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 [9] [10]. 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 Figure 1, Figure 2 and Figure 3:

![Figure 1. Raw SO2 Data](image)
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 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 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 in Figure 4 and Figure 5:
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. 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 Figure 6 and Figure 7:
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. 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 in Figure 8 and Figure 9:
4. Conclusion
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.

References
[1]. Srinivas DSRK 1999 Spatial patterns of air pollution in Delhi J Environ Prot. 19(3) pp 172-180.
[2]. Chandersekaran G E, Ravichandran C and Chandra Mohan A 1998 J. Environmental Protection.
[3]. Pandey V, Kumar A Pal, Singh N and Yunus M 1999 Indian J. Environmental Protection.
[4]. Ganesh Gopal Deverajan, V Muthukumaran, Ching-Hsien Hsu, Marimuthu Karuppiah, Yeh-Ching Chung and Ying-Huei Chen 2021 Transactions on Emerging Telecommunications Technologies.
[5]. Muthukumaran V and Ezhilmaran D 2020 International Journal of Information Technology and Web Engineering 15(4) pp 18-36.
[6]. Sadhasivam J, Kubendir M, Tomy P, Jeyakumar B, Sathish Kumar M and Anusha R 2017 Int. J. Civ. Eng. Technol. 8(11) pp 61–68.
[7]. Vinoth Kumar V, Ramamoorthy S, Dhilip Kumar V, Prabu M and Balajee J M 2021 International Journal of e-Collaboration 17(1) pp 62-70.
[8]. Jeyakumar, Balajee M A, Saleem Durai and Daphne Lopez 2018 In HCI Challenges and Privacy Preservation in Big Data Security IGI Global pp 159-174.
[9]. Asha J M, Sadhasivam J N S and Rohit R 2017 Int. J. Mech. Eng. Technol. 8(11) pp 964–976.
[10]. Sadhasivam J, Jayavel S, Jeyakumar B and Merchant S 2017 Int. J. Civ. Eng. Technol. 8(11) pp 35–41.