Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 22 | No. 1 | pp. 221-228 | 2023

Original Research Paper

https://doi.org/10.46488/NEPT.2023.v22i01.020

Open Access Journal

Forecasting Particulate Matter Emissions Using Time Series Models

S. Suresh†, M. R. Sindhumol, M. Ramadurai, D. Kalvinithi and M. Sangeetha
Department of Statistics, University of Madras, Chepauk, Chennai-600005, Tamil Nadu, India
†Corresponding author: S. Suresh; sureshstat22@gmail.com

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 25-05-2022
Revised: 17-07-2022
Accepted: 19-07-2022

Key Words:
Particulate matter
$PM_{2.5}$
Forecast
ARIMA
LSTM
Prophet

ABSTRACT

Environmental pollution is a serious concern nowadays with its disastrous impact on living organisms. In several types of pollution, Air pollution takes on a crucial role by directly affecting the respiratory system and causing fatal diseases in humans. Air pollution is a mixture of gaseous and particulate matter interweaved by different sources and emanating into the atmosphere. In particular, particle pollutants are critical in growing air pollution in India's main cities. Forecasting the particulate matter could mitigate the complications caused by it. The employment of a model to predict future values based on previously observed values is known as time series forecasting. In this paper, the $PM_{2.5}$ pollutant emission data recorded at the Kodungaiyur region of Chennai city were forecasted using three-time series models. The standard ARIMA model is compared with the deep learning-based LSTM model and Facebook's developed Prophet algorithm. This comparison helps to identify an appropriate forecasting model for $PM_{2.5}$ pollutant emission. The Root Mean Squared Error (RMSE) acquired from experimental findings is used to compare model performances.

INTRODUCTION

Air is an inhalable elixir and without it, life is out of the question. The contamination of air affects living beings significantly termed as air pollution. The combination of Gaseous and Particle elements are the primary sources of air pollution. Particulate matters are substances with a diameter of fewer than 10 microns that are categorized as highly prioritized pollutants. Particulate Matters generated from interconnected sources such as protracted road constructions, infrastructural activities, big dumpsites, and those particles from automobile exhaust have been a major cause of pollution in cities in recent years. Substantially, particulate matter is a complex mixture of metals, nitrates, sulfates, dust, water, and tire rubber. Domingo & Rovira (2020) mention sulfur dioxide, particulate matter, and oxides of nitrogen have an immediate and profound effect on human health. These pollutants can be directly emitted from different sources and have different chemical compositions. Chen et al. (2019) discussed Fugitive road dust (FRD) particles which are discharged from vehicular traffic, combustion of gasoline, oil, diesel fuel, and wood produce most of the particulate matter pollution found in outdoor air. Lo et al. (2016) and Fang et al. (2019) brought up the association between daily exposure to particulate matter and respiratory mortality. Buoli et al. (2018) stated that the actual risk of detrimental effects depends on one's state of health. But on the contrary, Sivarethinamohan et al. (2020) argued that polluted air can cause critical problems in healthy people, including respiratory irritation or breathing difficulties during exercise or outdoor activities. Therefore, it is inevitable to combat air pollution.

The research conducted by Pavlos et al. (2005) suggests that setting up continuous monitoring stations at excess pollution-emitting areas would help to restrain air pollution. The Monitoring stations give the air pollution data recorded at different time stamps which are known as time series data. The Methods for studying time series data to extract useful statistics and other aspects of the data are referred to as time series analysis. The Time series analysis carried out by Bai et al. (2018) overviewed different forecasting models to predict the future values of air pollution. The forecasting also gives insights to track the pattern which helps us to appropriate

ORCID details of the authors:

S. Suresh
https://orcid.org/0000-0003-1894-8548
D. Kalvinithi
https://orcid.org/0000-0003-2928-5057
M. Sangeetha
https://orcid.org/0000-0003-0450-5287
M. Ramadurai
https://orcid.org/0000-0001-7067-8827

actions against the worsened environment. Several methods were experimented with in a wide range of studies to identify suitable forecasting models to predict air pollution. Gourav et al. (2020) utilized the monitoring stations data obtained from one of the highly polluted capital cities, Delhi, and forecasted it using the ARIMA model to make recurrent decisions makings. Jai Shankar et al. (2010) discussed the model selection for a given problem using the ARIMA process which can be supported by diagnostic checking and error analysis. Abhilash et al. (2018) and Claudio et al. (2018) also used the ARIMA model to forecast air pollution. The state-of-art deep learning technique is also used to forecast pollutant emissions by considering it as sequential data. Chang et al. (2020) Liu et al. (2020) and Alghieth et al. (2021) took existing pollutant data as sequential data and forecasted it using the Long Short Term Memory model, a type of Recurrent Neural Network. Shen et al. (2020) and Topping et al. (2020) conducted experiments on the non–linear pollutant data using the viable prophet algorithm works on the Generalized additive regressive model. Siami-Namini et al. (2018) and Peter et al. (2019) compared ARIMA and LSTM models in their research works. The ARIMA model and Prophet algorithm were also compared in the experiments conducted by Samal Krishnan et al. (2019) and Ziyuan (2019). The findings of Nath et al. (2021) reflect that the statistical models outperform the deep learning methods.

In the event of health vulnerability due to $PM_{2.5}$ at Kodungaiyur in recent times, as evidence found and discussed by Krishnan et al. (2020), the major dumpsite of Chennai city is considered as the study area. Owing to numerous small-scale industries, the locality also holds commuters from outside the area on a daily basis which causes more traffic congestion. Nadeem et al. (2020) argued how the poorest quality of air in a single locality affects an entire city. These pieces of literature show, the Kodungaiyur region has a substandard environment to breathe due to its highest exposure to particulate matter of diameter less than 2.5 microns. Moreover, it is complex to understand which forecasting method predicts the future $PM_{2.5}$ with high accuracy. Thus, the ARIMA model, Prophet algorithm, and LSTM model were adopted in this study for forecasting $PM_{2.5}$ emissions. This work aims at forecasting $PM_{2.5}$ emissions using a suitable model. The performance of the models was estimated using root mean squared error.

In this article, three established forecasting methods models were applied to the time series data of $PM_{2.5}$ emissions and the results were compared. This experimentation shows, the standard ARIMA model of order (3,1,1) gives the best fit for predicting the observations with a low error rate when compared to other models taken for study. Using the ARIMA model, the future values of 6 months are forecasted and the comparative results were shown.

This paper is organized as follows: The second section of the paper describes the data utilized. The third section has three sub-sections that explain the methods adopted in this study. The fourth section put forth the outcomes of the experimental results carried out on three different methods. The final and concluding section discusses the recommendation of a suitable model for air pollutant forecasting in a given locality and highlights several causes that led to worsening air quality which can be treated carefully in the near future.

MATERIALS AND METHODS

Kodungaiyur is an industrial neighborhood located in the northern part of Chennai city. Day-wise data on $PM_{2.5}$ for the Kodungaiyur area was collected from Tamil Nadu Pollution Control Board for the period from 1st January 2019 to 31st December 2022. The data contain the variables date and $PM_{2.5}$ with 1096 observations.

Forecasting Methods

The study employs three forecasting methods, the ARIMA model, Recurrent Neural Network's Long Short Term Memory algorithm, and Facebook's Prophet algorithm.

ARIMA Model

The Autoregressive Integrated Moving Average is a model of regression type where the predictors contain lags of the dependent variable and/or forecast errors. The linear equation for the ARIMA model is as follows

$$y_t = C + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t \quad \ldots(1)$$

$$y_t = C + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad \ldots(2)$$

$$y_t = C + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad \ldots(3)$$

whereare $\emptyset_1, \emptyset_2, \ldots \emptyset_p, \theta_1, \theta_2, \ldots, \theta_q$ AR and MA coefficients respectively and C is the intercept.

Equation (1) is for a p^{th} order autoregressive (AR) model, equation (2) is a q^{th} order moving average (MA) model, and equation (3) is the equation for an ARIMA (p, d, q) model.

Several combinations of AR(p), I(d), and MA(q) were applied to the obtained time series data and finally, the model with comparatively low AIC value is considered for forecasting.

Prophet Model

The Prophet model relies on Generalised Additive Regression Model which is highly suited for non–linear repressors. The Prophet is a decomposable time series model with four main components,

- The trend of a piecewise linear or logistic growth curve is represented in equation (4). Prophet detects changes in trends automatically by selecting change points from the data.
- A yearly seasonal component represented by a Fourier series represented in equation (5)
- Dummy variables are used to create a weekly seasonal component.
- A list of important holidays submitted by users is given in equation (6)

$$g(t) = \frac{C_t}{1+exp(-k+a(t)^T\delta)(t-(m+a(t)^T\gamma)))} \quad \ldots(4)$$

$$s(t) = \sum_{n=1}^{n}\left(a_n cos\left(\frac{2\pi nt}{P}\right) + b_n sin\left(\frac{2\pi nt}{P}\right)\right) \ldots(5)$$

$$h(t) = Z(t)k \quad \ldots(6)$$

The combined equation turns out to be,

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad \ldots(7)$$

where ϵ_t in equation (7) is the error term.

LSTM Model

A long short-term memory is a type of recurrent neural network. The output of the previous step is used as input in the current step in RNN. Long-term dependency is the issue of RNN addressed, in which the RNN is unable to predict words stored in long-term memory but can make more accurate predictions based on current data. RNN does not provide an efficient performance as the gap length rises. By default, the LSTM can keep the information for a long time. It is used for time-series data processing, prediction, and classification. The LSTM features (Fig. 1) a chain structure with four neural networks and various memory blocks known as cells.

The cells store information, whereas the gates manipulate memory. The model has three gates.

(a) **Forget Gate:** The forget gate deletes information that is no longer useful in the cell state. Two inputs, x_t (at-the-time input) and h_{t-1} (previous cell output) are fed into the gate and multiplied with weight matrices before bias is added. The result is fed into an activation function, which produces a binary output. If the output for a specific cell state is 0, the information is lost; if the output is 1, the information is saved for future use.

(b) **Input Gate:** The input gate is responsible for adding useful information to the cell state. First, the information is regulated using the sigmoid function, and the values to be remembered are filtered using the h_{t-1} and x_t inputs, similar to the forget gate. The tanh function is then used to generate a vector with values ranging from -1 to +1 that contains all of the possible values from h_{t-1} and x_t. Finally, the vector and regulated values are multiplied to obtain useful information.

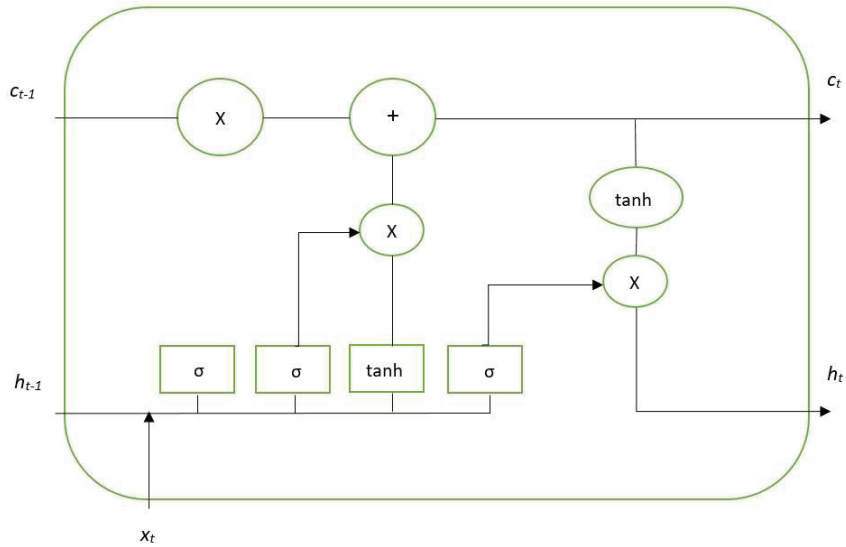

Fig. 1 Structure of an LSTM Cell.

(c) **Output Gate:** The output gate is in charge of extracting useful information from the current cell state and presenting it as output. To begin, a vector is created by applying the tanh function to the cell. The information is then regulated using the sigmoid function and filtered by the values to be remembered using h_{t-1} and x_t inputs. Finally, the vector values and the regulated values are multiplied and sent as output and input to the next cell.

The three methods are implemented and their empirical out-turns are stated in the following section.

RESULTS AND DISCUSSION

In the experimental phase, several combinations of parameters attributed to each time series model were evaluated. The parameter estimation for individual methods gives the best models and is implemented to predict the observations.

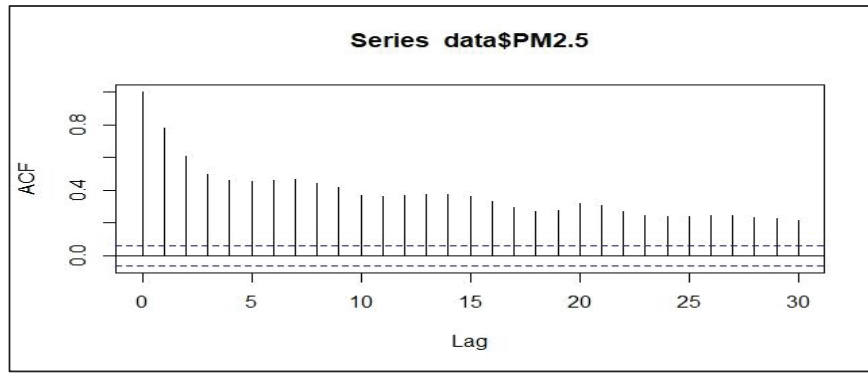

Fig. 2: ACF correlogram for $PM_{2.5}$ emission data.

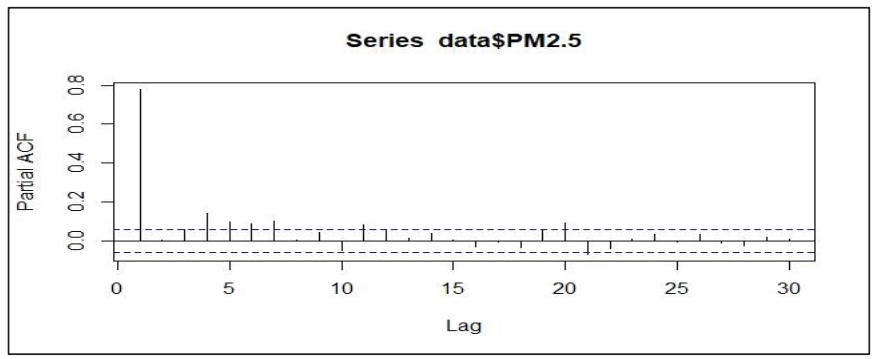

Fig. 3: PACF correlogram for $PM_{2.5}$ emission data.

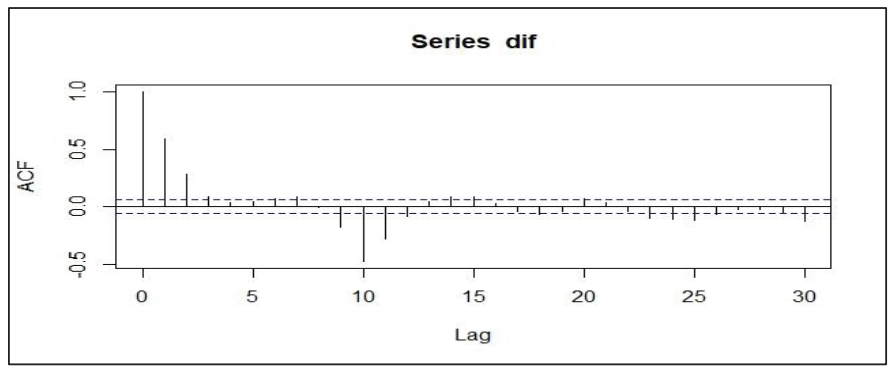

Fig. 4: ACF Correlogram for first differenced $PM_{2.5}$ emission data.

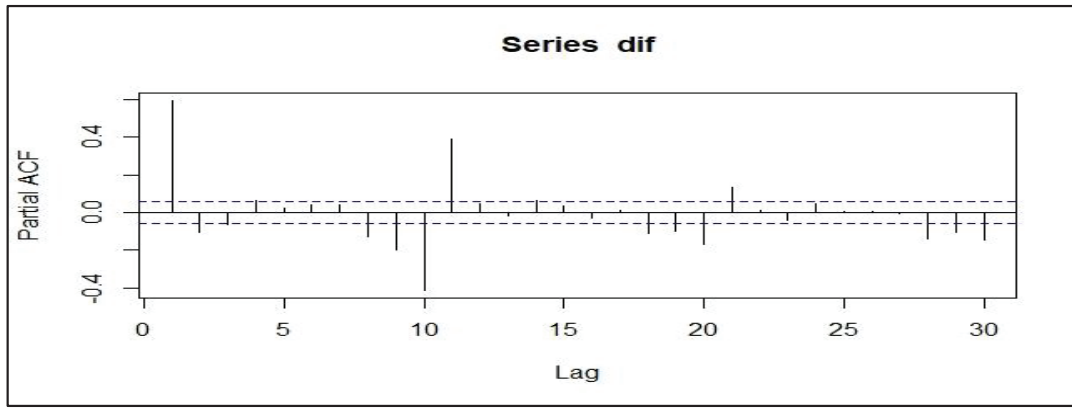

Fig. 5: PACF Correlogram for first differenced $PM_{2.5}$ emission data.

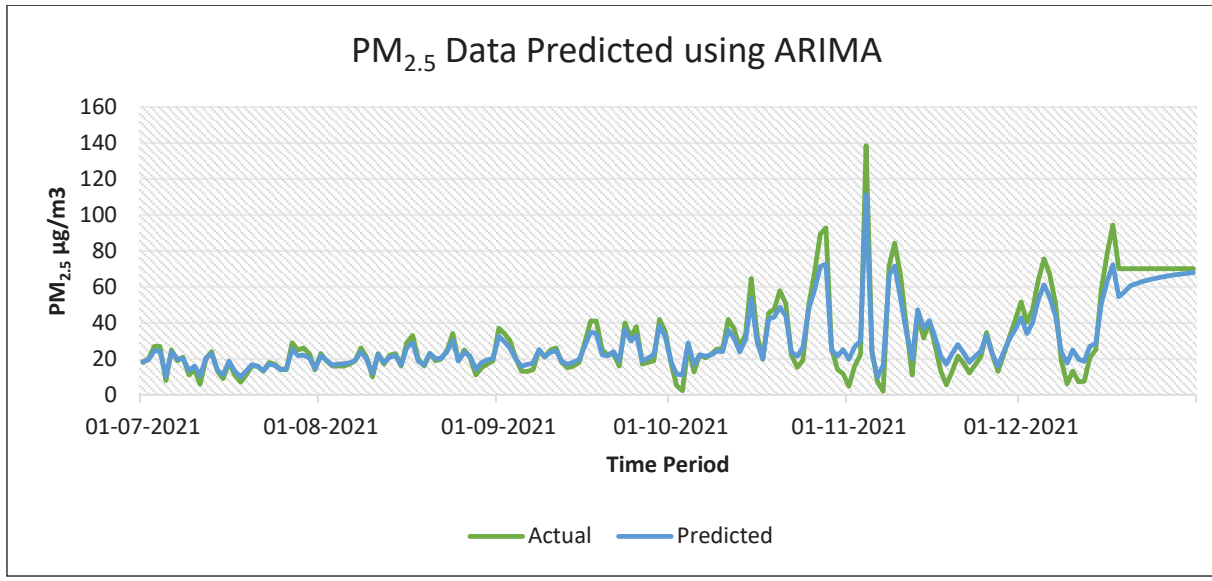

Fig. 6: $PM_{2.5}$ pollutant data predicted using ARIMA (3,1,1).

Based on the residual difference between the observed and predicted value, the error value for each model is calculated. Later, the $PM_{2.5}$ data is forecasted for 6 months (from 01-01-2022 to 30-06-2022). The identified models were stated in this section.

Fitted ARIMA Model

ACF and PACF Correlograms were plotted to identify the best ARIMA model for forecasting. The ACF plot gives the order of Moving Average (MA) and the PACF gives the order of Auto-Regressive (AR). In addition to that, the model is also tested for low Akaike's Information Criteria (AIC). The parameters *p, d, q,* and corresponding optimum AIC value give the tentative ARIMA model.

Fig. 2 and Fig. 3 show the ACF and PACF Correlograms plotted for $PM_{2.5}$ emission data, which indicates the non – stationarity of data. Hence the first differences in the data are taken for analysis.

Fig. 4 and Fig. 5 show the ACF and PACF plots for different $PM_{2.5}$ emission data. Fig. 4 indicates the MA (1) order. Fig. 5 shows that there are significant spikes at lags 1, 10, and 11 in the PACF indicating the order of AR to be 3. Therefore, for the obtained data, the model ARIMA (3,1,1) is performed for predicting the observations.

Fitted Prophet Model

The Prophet algorithm accounts for the change points, which influence the trend of the time series

data. The trend of the fitted values becomes flexible when the value for the range of the change point gets increased. The fitted Prophet model is displayed below.

```
# Initializing the Model
model=Prophet(interval_width=0.95, yearly_seasonality=True, weekly_sea
sonality=True, changepoint_prior_scale=2, changepoint_range = 0.8)
```

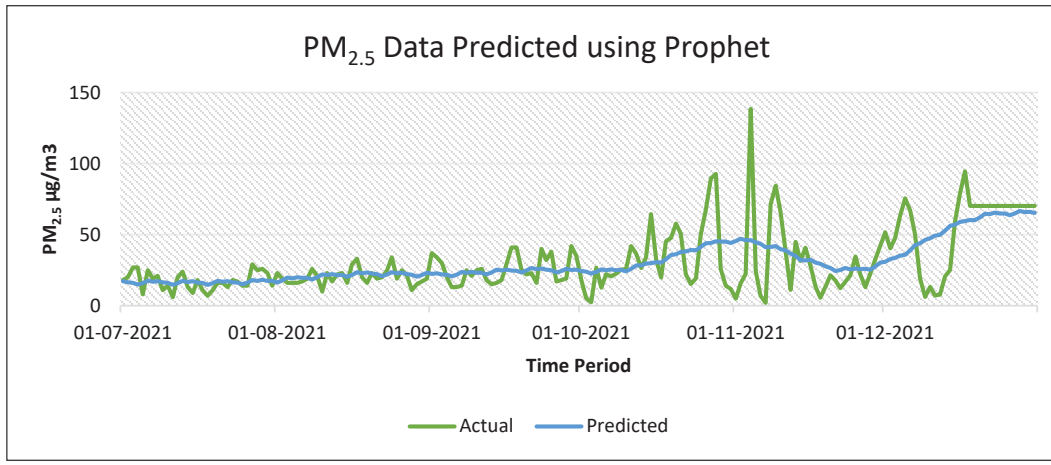

Fig. 7: $PM_{2.5}$ pollutant data predicted using Prophet model.

Fitted LSTM Model

The LSTM model is built on multiple layers comprised of two LSTM layers, one Dropout layer, and one Dense layer. The LSTM layers are counted under input layers and the Dense layer is termed the output layer. The role of the Dropout layer is to make the trivial input values as 0's. This Sequential model is used for predicting the observations present in $PM_{2.5}$ Pollutant data.

Based on the analysis, the actual values are compared with the predicted values of each model and the error rates were determined. Fig. 6, Fig. 7, and Fig. 8 show the line graph of $PM_{2.5}$ pollutant data observations and their predictions made by respective models which are plotted against time period. In this study, the estimated Root Mean Squared Error is used to obtain the error rates of fitted models which are shown in Table 1.

From Table 1, it can be inferred that ARIMA is found to be the best predicting model which has a low RMSE on comparing with the other models. Therefore, the ARIMA (3,1,1) model is used for forecasting the future values.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 7, 64)             16896

 lstm_1 (LSTM)                (None, 32)                12416

 dropout (Dropout)            (None, 32)                0

 dense (Dense)                (None, 1)                 33

=================================================================
```

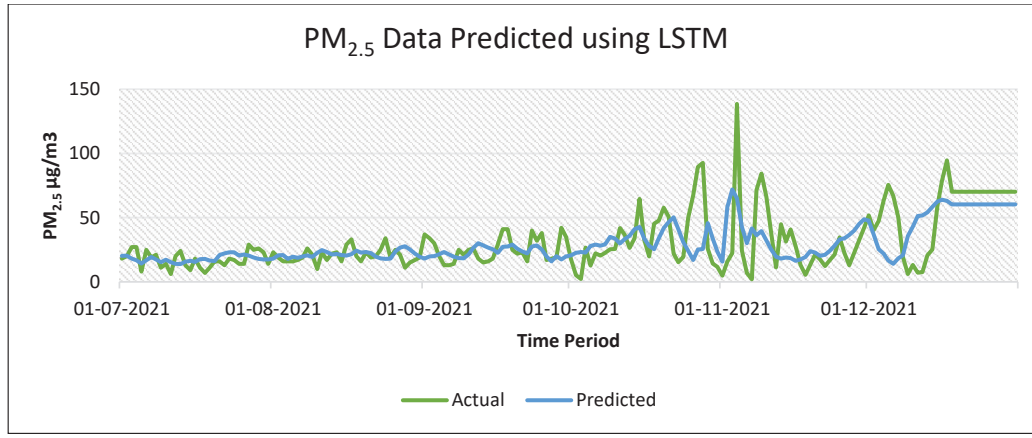

Fig. 8: $PM_{2.5}$ pollutant data predicted using LSTM model.

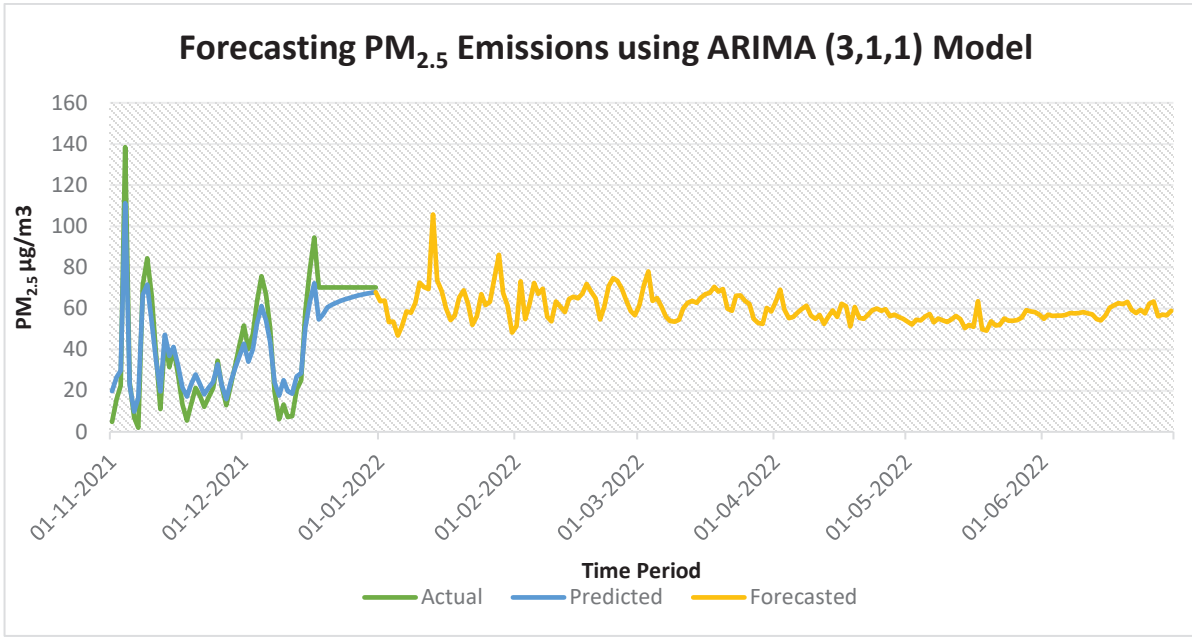

Fig. 9: $PM_{2.5}$ pollutant emissions forecasted using ARIMA(3,1,1) for 6 months.

Table 1: Error rates of the models implemented.

MODELS	ARIMA	PROPHET	LSTM
RMSE	4.33	11.17	12.45

Fig. 9 shows using ARIMA (3,1,1) model the $PM_{2.5}$ pollutant data is forecasted for future values from 01-01-2022 to 30-06-2022.

CONCLUSION

The ARIMA model of order (3,1,1) turns out to be the well-suited model to predict the data with low error rates compared to the other models implemented. Using the suitable model, the forecast of $PM_{2.5}$ is computed shown in Fig 9. The forecast shows the month of January witnessed high emissions of $PM_{2.5}$ over 100 $\mu g.m^{-3}$ due to the Pongal festival. Fig 9 depicts that, the $PM_{2.5}$ emissions at

Kodungaiyur for the first six months of 2022 average around 60 $\mu g.m^{-3}$ which is the 24 h average of $PM_{2.5}$ as prescribed by CPCB. This forecast also manifests the scenario that the sources of air pollution are not only industrial and vehicular emissions but also the particulate matter emitted from unpaved roads, construction sites, and several indoor activities such as domestic burning and natural specks of dust. The performance of every forecasting model highly relies on the data used. For the recorded $PM_{2.5}$ emissions from the Kodungaiyur region, the ARIMA model is viable to use, which will assist policymakers in mitigating air pollution problems caused by particulate matter.

Further, as an extensive study, this research can be focused on the computation of the magnitude of the sources of particulate matter and their contribution to air pollution.

ACKNOWLEDGEMENT

This research work is funded by Rashtriya Uchchatar Shiksha Abhiyan (RUSA2.0), Ministry of Human Resource Development, Government of India.

REFERENCES

Abhilash, M.S.K., Thakur, A., Gupta, D. and Sreevidya, B. 2018. Time series analysis of air pollution in Bengaluru using ARIMA model. Amb. Commun. Comp. Sys. Adv. Intell. Sys. Com., 71: 696.

Alghieth, M., Alawaji, R., Saleh, S.H. and Alharbi, S. 2021. Air pollution forecasting using deep learning, Int. J. Online Biomed. Eng., 17(14): 50-64.

Bai, L., Wang, J., Xuejiao, M. and Lu, H. 2018. Air pollution forecasts: An overview. Int. J. of Env. Res. and Pub. Health, 15(4): 780.

Buoli, M., Grassi, S., Caldiroli, A., Carnevali, G.S., Mucci, M., Lodice, S., Cantone, L., Pergoli, L. and Bollati. V. 2018. Is there a link between air pollution and mental disorders? Environ. Int., 118: 154-168.

Chang, Y.S., Chiao, H., Abimannan, S., Huang, Y. and Tsai, Y. 2020. An LSTM-based aggregated model for air pollution forecasting. Atmos. Poll. Res., 11(8): 1451-1463.

Chen, S., Zhang, X., Lin, J., Huang, J., Zhao, D., Yuan, T., Huang, K., Kuo, Y, Jhia, Z., Zang, Z., Qiu, Y. and Xie, L. 2019. Fugitive road dust $pm_{2.5}$ emissions and their potential health impacts. Environ. Sci. Technol., 53(14): 8455-8465.

Claudio, G., Griselda, C.J., Breton, R.M.C. and Tepedion, C. ARIMA models application to air pollution data in Monterrey, Mexico. AIP Conf. Proceed., 82(1): 020041.

Domingo, J.L. and Rovira, J. 2020. Effects of air pollutants on the transmission and severity of respiratory viral infections. Environ. Res., 16: 187.

Fang, X., Fang, B., Wang, C., Xia, T., Boottai, M., Fang, F. and Cao, Y. 2019. Comparison of frequentist and Bayesian generalized additive models for assessing the association between daily exposure to fine particles and respiratory mortality: A simulation study. Int. J. Env. Res. Pub. Health, 16(5): 746.

Gourav, R., Jusleen, N., Preeti, J. and Rachna, J. 2020. Forecasting air quality of Delhi using ARIMA model. Adv. Data Sci., Sec. Appl., 612: 315-325.

Jai Sankar, T., Prabakaran, R., Senthamarai Kannan, K. and Suresh, S. 2010. Stochastic modelling for cattle production forecasting. J. Modern. Math. Stat., 4: 53-57.

Krishnan, M.A., Jawahar, K., Perumal, V., Devraj, T., Thanarasu, A., Kubendran, K. and Sreenivasan, S. 2020. Effects of ambient air pollution on respiratory and eye illness in population living in Kodungaiyur, Chennai. Atmos. Env., 203: -171.

Liu, D., Lee, S., Huang, Y. and Chien‐Ju, C. 2020. Air pollution forecasting based on attention‐based LSTM neural network and ensemble learning. Exp. Sys., 37(8): 444.

Lo, W.C., Shie, R.H., Chan, C.C. and Lin, H.H .2017. Burden of disease attributable to ambient fine particulate matter exposure in Taiwan. J. Formos. Med. Assoc., 111: 4516.

Nadeem, I., Ilyas, A. and Uduman, P.S. 2020. Forecasting ambient air quality of Chennai city in India. Geo. Env. Sust., 13(3): 12-21.

Nath, P., Saha, P. and Middya, A.I. 2021. Long-term time-series pollution forecast using statistical and deep learning methods. Neural Comput. Appl., 33: 12551-12570.

Pavlos, S.K, Jerett, M., Morrison, J.B. and Beckerman, B. 2005. Establishing an air pollution monitoring network for intra-urban population exposure assessment: A location-allocation approach. Atmos. Env., 39(13): 2399-2409.

Peter, T., Yamak, L., Yujian, M. and Pius, K.G. 2019. A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting. Association for Computing Machines, New York, NY, USA, pp. 49-55.

Shen, J., Valagolam, D. and Mccalla, S. 2020. Prophet forecasting model: A machine learning approach to predict the concentration of air pollutants ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) in Seoul, South Korea. Peer J., 8: 11-21.

Siami-Namini, S., Tavakoli, N. and Siami-Namini, A. 2018. A Comparison of ARIMA and LSTM in Forecasting Time Series. Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, December 17-20, 2018, Orlando, Florida, IEEE, Piscataway, NJ, USA, pp. 1394-1401,

Sivarethinamohan, R., Sujatha, S., Priya, S. and Sankaran, M. 2020. Impact of air pollution in health and socio-economic aspects: Review on future approach. Mater. Today Proceed., 16: 45-66.

Topping, D., Watts, D., Coe, H., Evans, J., Bannan, T.J., Lowe, D., Jay, C. and Taylor, J.W. 2020. Evaluating the use of Facebook's Prophet model v0.6 in forecasting concentrations of NO_2 at single sites across the UK and in response to the COVID-19 lockdown in Manchester, England. J. Geosci. Mod. Dev., 56: 919-930.

Ziyuan, Y. 2019. Air pollutants prediction in Shenzhen based on ARIMA and prophet method. E3S Web Conf., 13: 56.