Time Series Analysis for Air Quality Forecasting

[1]Dixita Dessai, [2]Shaba Dessai, [3] Sufola Das Chagas Silva e Araujo

[1]Student, M.E in Information Technology, Padre Conceicao College of Engineering, Goa University, Verna-Goa, India,

[2,3]Assistant Professor, Department of Information Technology, Padre Conceicao College of Engineering, Goa University, Verna-Goa, India.

Abstract: This study conducts a Time Series Analysis of air quality indicators with an aim to forecast the values of air pollutant concentrations like SO_2, NO_2 and PM_{10}, using Auto Regressive Integrated Moving Average (ARIMA) and Long Short Term Memory (LSTM), for the cities of Margao and Sanguem in the Indian state of Goa. On comparing the results from the two methods, based on the performance measures of Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) it is observed that LSTM model achieves a fair prediction of the air pollutant concentration values.

***Keywords*- Time Series Analysis, Forecasting, ARIMA, LSTM, MAPE, RMSE, Air Quality Index (AQI)**.

I.INTRODUCTION

Air is a very important resource that supports life on earth. Good quality air is essential for the sustainable growth of all living beings, but nowadays, the air quality is decreasing, making it difficult for some people to breathe while causing respiratory diseases in others. An increase in the number of vehicles, burning of plastics, cutting of trees, emission of harmful gases due to industrial activities are some of the reasons for air quality degradation [12]. Therefore, it is not only necessary to monitor the air quality regularly but for proper planning of the appropriate control actions, it is of utmost importance that we are able to predict the air quality indicators for the future. Time series analysis allows us to identify the underlying forces that lead to a particular trend and to forecast the future air pollutant values.

Having an idea of the future air pollutant values gives us a broader picture of the air quality and also helps us to take the necessary preventive and control actions. Accurate forecasting helps individuals or organizations to plan and to decrease the ill effects of the air pollutants on our health. Such awareness can help us to create a cleaner environment and to live a healthy life.

Air quality can be analyzed in terms of air pollutants by using time series models [12]. A time series is a set of observations x_t, each one being recorded at a specific time t [1]. The four components of the time series [2] are as follows:

- Trend – Variations that go up and down in a predictable manner.
- Seasonal – Variations that repeat in a regular or periodic manner which includes a day, week, month, or season.
- Cyclic – These are long-term oscillations over a span of more than a year such as a Business cycle.
- Random variations – These variations occur at an irregular interval and are mostly unpredictable.

An air quality index (AQI) is used to describe the quality of air [3]. It tells us how clean or polluted the air is and what is the associated health effects. AQI converts air pollutant concentrations into a number lying between 0 and 500. Two steps that are involved in calculating an AQI are: formation of sub-indices (for each pollutant: SO_2, NO_2, PM_{10}) and aggregation of sub-indices to get an overall AQI.

II. LITERATURE SURVEY

Nurul Nnadiah Zakaria, Mahmod Othman, Rajalingam Sokkalingam, Hanita Daud, Lazim Abdullah, Evizal Abdul Kadir [5] proposes a simple forecasting method to forecast the air quality with a Markov chain model. The proposed model starts by defining states of the Markov model and then the State transition matrix (N) and the Probability transition matrix (P) is constructed. Next, confirm if the developed Markov chain model is the Ergodic Markov chain. Stationary probability distribution and Mean return time are calculated for Markov process probability values. Finally, the forecasting values are obtained by multiplying the initial probability and the state transition probability. The model is validated using the Chi-square test.

C.L.Karmaker, P.K.Halder, and E.Sarker [6], uses different forecasting time series models like Simple Moving Average Method (SMA), Single Exponential Smoothing Method (SES), Double Exponential Smoothing(Holt's Method), Winters Method, Linear Trend Analysis to forecast the jute yarn demand from the year 2010 to the year 2013. The raw data is collected from the jute product manufacturer in Bangladesh, namely, Akij Jute Mills, Akij Group Ltd., in Noapara, Jessore. Winters Additive Model is the most suitable one to forecast jute yarn demand.

Kostandina Veljanovska and Angel Dimoski [7], uses one unsupervised algorithm Neural Network (NN) and three supervised learning algorithms k-nearest neighbor (k-NN), Support Vector Machines (SVM) and Decision Tree (DT) to predict the air quality index. SVM algorithm has 80.0% of accuracy using the linear kernel function.

Songting Xing and Yuansheng Lou [8], uses ARIMA model and neural network to predict the hydrological time series. They combine these two models ARIMA-RBF and then proposed a particle swarm optimization algorithm to optimize the RBF neural network. Purwanto, Chikkannan Eswaran, and Rajasvaran Logeswaran [10], uses ARIMA, Neural Network, and Linear Regression to predict the Infant Mortality Rate in Indonesia. The data used for this study is obtained from the Indonesian Health

Profile, Ministry of Health Republic of Indonesia from the year 1995 to 2008. Among the three models used, the Neural Network model with 6 input neurons, 10 hidden layer neurons, and hyperbolic tangent activation functions was found to be the best.

Lintao Yang and Honggeng Yang [9], proposed a combined ARIMA-PPR model for short term load forecasting. The power load time series data was obtained from the Sichuan Electric Power Company of China. The combined model is constructed by assigning weight coefficients (determined by the root mean square error) to the individual models. The combined model achieves better prediction results as it captures both linear and non-linear modes. Huixiang Liu, Qing Li, Dongbing Yu, and Yu Gu [19], uses SVR and RFR to predict AQI in Beijing and nitrogen oxides concentration in an Italian city. The SVR model performed better in predicting AQI whereas the RFR model makes better nitrogen oxide concentration prediction. Wei Kexin and Du Mingxing [20], uses the Autoregressive moving average model to predict the temperature of Insulated Gate Bipolar Transistor (IGBT**).**

Adil Moghar and Mhamed Hamiche [21] investigated on how epochs can improve the prediction of LSTM machine learning algorithm. The data used in this paper is from the New York Stock Exchange NYSE (GOOGL and NKE). The model is trained using different epochs (12 epochs, 25 epochs, 50 epochs and 100 epochs).The result show that the LSTM model is capable of tracing the opening prices of GOOGL and NKE. Siddharth Banyal, Pushkar Goel, and Deepank Grover [22], uses stacked LSTM and Multi-layered perceptron (MLP) to make NSE-Stick market prediction. LSTM and MLP models are compared using MSE under different optimizers such ad Adam and RMSProp, activation functions like sigmoid and ReLU and different timestamps. Stacked LSTM with 60 Time-stamp, RMS Optimizer and ReLU activation functions performs better.

T. Sujjaviriyasup, K. Pitiruek [23] compares the forecasting results of ARIMA and MODWT-SVM-DE using MAE, MAPE, and sMAPE. Ten datasets from International Energy Agency (IEA) from 1980 to 2014 were used for this study. In MODWT-SVM-DE model, MODWT is used to decompose the initial input series resulting in scaling coefficients. The new dataset is formulated by combining the original and the transformed data and then rearranging it into m columns. This new dataset is then passed through an SVM model whose parameters are found using the DE technique. The results show that the MODWT-SVM-DE model is more accurate than ARIMA model at significance level $\propto$= 0.10.

III. TIME SERIES FORECAST MODELS

3.1 ARIMA

Auto Regressive Integrated Moving Average (ARIMA) is a time series model that predicts a value y_t based on its past values and past forecast errors. The ARIMA model is characterized by three parameters: p (order of Auto Regressive term), d (order of differencing), q (order of Moving Average term).

An ARIMA model is defined by the following expression:

$$Y_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \quad (1)$$

Where, c is the constant term, ε_t is the error term, φ_i is the AR model parameter and θ_j is the MA model parameter.

Steps involved in ARIMA model:

1. First check if the time series is stationary. A time series is stationary if all its statistical properties are consistent over time [11]. If the series is not stationary, then difference the series to make it stationary. The number of times the data points are differenced to obtain stationarity is known as the order of differencing and is denoted by d.
2. Find the order of AR and MA terms (i.e. value of p and q respectively) by using the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) graphs (also known as correlogram). ACF is a correlation between the time series and the lagged time series whereas PACF is the correlation between the time series and the lagged time series after removing the effect of the intermediate lagged terms.
3. This step involves estimating the AR (φ) and MA (θ) model parameters.
4. Once the value of p, d, and q are identified, find the most adequate model which has the minimum AIC (Akaike Information Criteria) value.
5. Using the ARIMA model predict the future values and then validate the model by comparing the predicted values to the actual values.

3.2 LSTM

An LSTM is a type of recurrent neural network that passes the necessary information as it moves forward. LSTM solves the vanishing gradient problem of the recurrent neural network by using a gating mechanism and the memory cell. A standard LSTM cell consists of the forget gate (f_t), input gate (i_t), cell state (c_t), and the output gate (o_t). The forget gate decides what information should be thrown away or passed through the cell and the cell state represents the internal memory of the LSTM cell.

The equations involved in the LSTM unit are as follows:

$$f_t = \sigma(w_f[\, h_{t-1}, X_t\,]) \quad (2)$$
$$i_t = \sigma(w_i[\, h_{t-1}, X_t\,]) \quad (3)$$
$$\tilde{c}_t = tanh(w_c[\, h_{t-1}, X_t\,]) \quad (4)$$
$$c_t = (c_{t-1} * f_t) + (i_t * \tilde{c}_t) \quad (5)$$
$$o_t = \sigma(w_o[\, h_{t-1}, X_t\,]) \quad (6)$$
$$h_t = o_t * tanh(c_t) \quad (7)$$

Where, f_t is the forget gate, i_t is the input gate, $\tilde{c}_t$ is the intermediate cell state, c_t is the cell state, o_t is the output gate, h_t is the hidden state and X_t is the input vector.

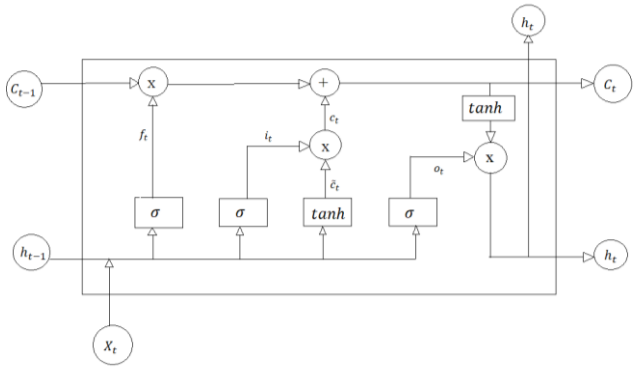

Figure 1: LSTM cell unit

Steps to build an LSTM network:

1. Load the dataset and split it into a training set and a testing set. Then normalize the data using a MinMaxScalar.
2. Converting the data into a supervised learning problem such that the LSTM network can learn how to predict the output pattern based on the previous input patterns.
3. Building the LSTM network by adding LSTM, Dropout, and Dense layer.
4. Compile the LSTM network once it is built. Mean squared error is used to compute the loss function and adam optimizer is used to minimize the loss function.
5. Fit the data to the network and make predictions

IV. RESULTS

The dataset used in this study was collected from the Goa State Pollution Control Board. The dataset contains air pollutant concentration values of SO2, NO2 and PM10 collected from April 2014 to August 2018 across cities of Goa. All the air pollutants are measured in µg/m3. The raw data contained some missing values and special characters. These values are replaced by the taking average of the readings of that month. In this study, data from April 2014 to March 2018 is used as a training set and data from April 2018 to August 2018 is used as a testing set.

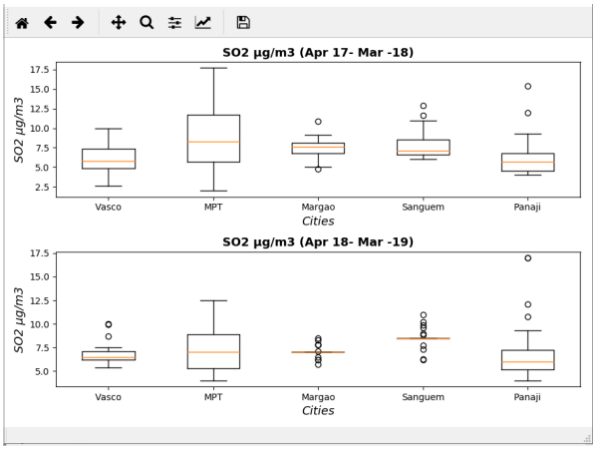

Figure 2: Boxplot for SO_2

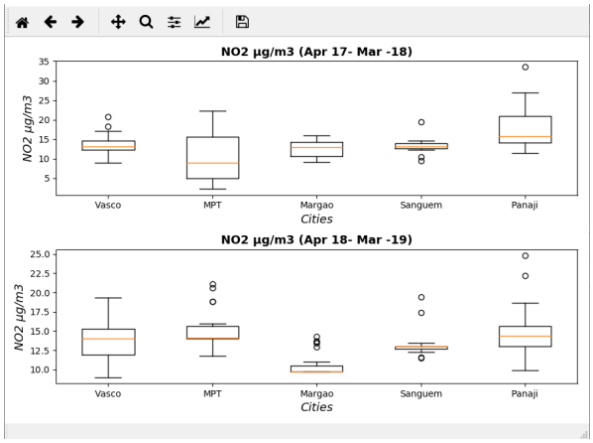

Figure 3: Boxplot for NO_2

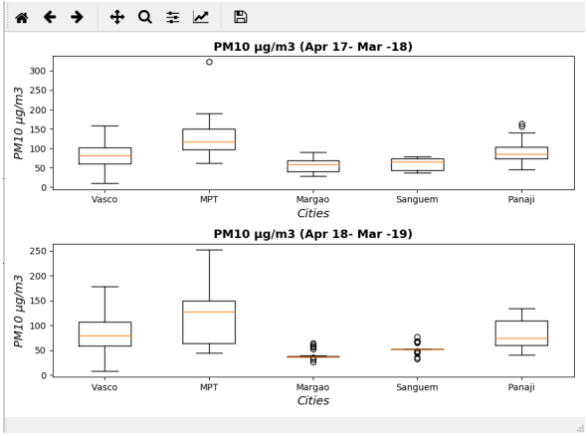

Figure 4: Boxplot for PM_{10}

Boxplots for SO_2, NO_2, and PM_{10} over five cities of Goa namely Vasco, MPT, Margao, Sanguem, and Panjim were constructed. Fig. 3 shows that SO_2 values for MPT and Panaji are almost symmetric during Apr 17-Mar 18 and Apr 18-Mar 19. MPT appears to have larger variability than the other four cities. Fig. 4 shows that in the year Apr 17-Mar 18, the minimum value of MPT is the smallest of the other four cities. Whereas in the year Apr 18-Mar 19, the minimum value of Vasco is the smallest among the other four cities. In Fig. 5, PM_{10} values for Vasco are almost symmetric during Apr 17-Mar 18 and Apr 18-Mar 19. The minimum value of Vasco is the smallest among the other four cities and the maximum value of MPT is the highest among the other four cities.

In this study, we will consider SO_2, NO_2 and PM_{10} concentration in Margao and Sanguem cities of Goa, India. Figure 5 to Figure 10 shows the actual and predicted values of air pollutants in Margao and Sanguem cities of Goa, India using ARIMA, and LSTM time series model.

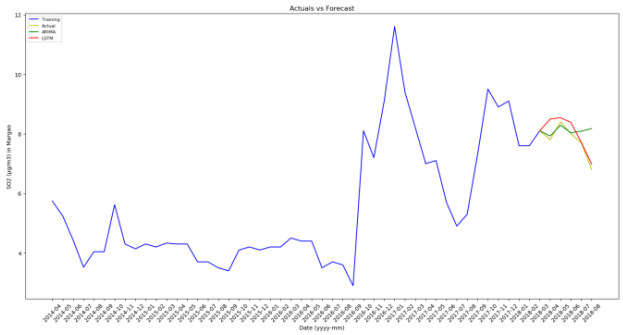

Figure 5: Comparison of model forecast results of SO_2 µg/m^3 in Margao

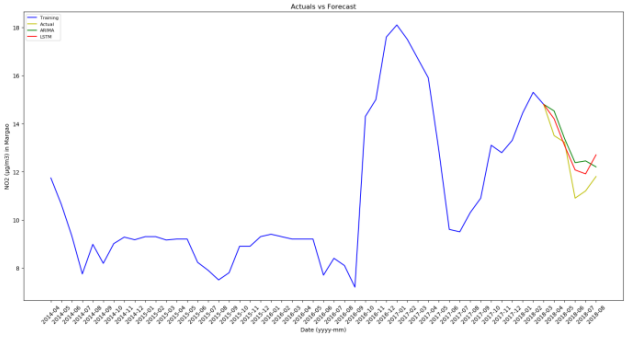

Figure 6: Comparison of model forecast results of NO_2 µg/m^3 in Margao

Table 1: Actual and predicted SO_2 µg/m^3 values in Margao

Date	Actual	ARIMA	LSTM
2018-04	7.8	7.9265	8.4954
2018-05	8.4	8.2867	8.5388
2018-06	8.0	8.0323	8.3834
2018-07	7.7	8.0871	7.7157
2018-08	6.8	8.1771	6.982

Table 2: Actual and predicted NO_2 µg/m^3 values in Margao

Date	Actual	ARIMA	LSTM
2018-04	13.5	14.524	14.1987
2018-05	13.2	13.3719	13.0796
2018-06	10.9	12.3753	12.0725
2018-07	11.2	12.4483	11.9095
2018-08	11.8	12.2004	12.7025

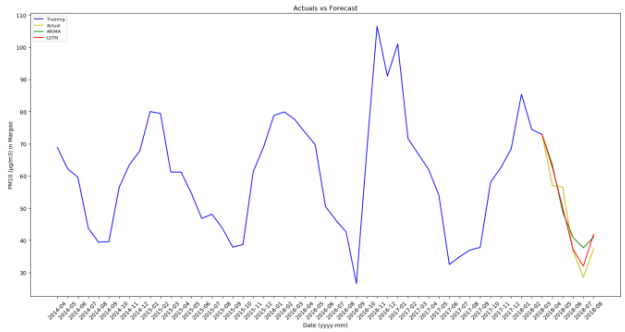

Figure 7: Comparison of model forecast results of PM_{10} µg/m^3 in Margao

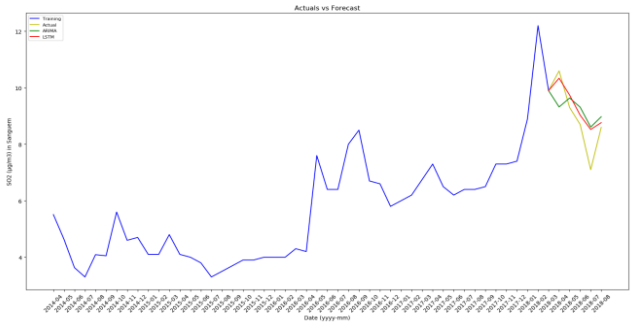

Figure 8: Comparison of model forecast results of SO_2 µg/m^3 in Sanguem

Table 1 to Table 6 shows the actual and predicted values of the air pollutants in Margao and Sanguem from 2018-04 to 2018-08 using ARIMA and LSTM as plotted in Figure 5 to Figure 10.

Table 3: Actual and predicted PM_{10} µg/m^3 values in Margao

Date	Actual	ARIMA	LSTM
2018-04	57.0	63.3612	62.6035
2018-05	56.55	48.7696	50.1157
2018-06	36.7	40.9196	37.191
2018-07	28.5	37.6902	31.9732
2018-08	37.4	41.0084	41.9283

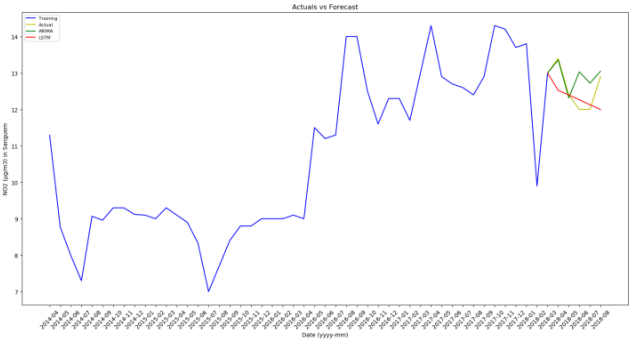

Figure 9: Comparison of model forecast results of NO_2 µg/m^3 in Sanguem

Table 5: Actual and predicted NO_2 µg/m^3 values in Sanguem

Date	Actual	ARIMA	LSTM
2018-04	13.4	13.3559	12.5216
2018-05	12.4	12.3186	12.3976
2018-06	12.0	13.0324	12.2629
2018-07	12.0	12.722	12.1277
2018-08	12.9	13.049	11.9968

Table 4: Actual and predicted SO_2 µg/m^3 values in Sanguem

Date	Actual	ARIMA	LSTM
2018-04	10.6	9.3207	10.3387
2018-05	9.3	9.6384	9.7398
2018-06	8.7	9.3129	9.0284
2018-07	7.1	8.6091	8.524
2018-08	8.6	8.9753	8.7652

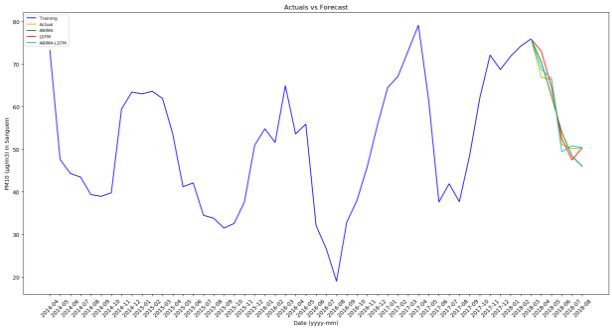

Figure 10: Comparison of model forecast results of PM_{10} µg/m^3 in Sanguem

Table 6: Actual and predicted PM_{10} µg/m^3 values in Sanguem

Date	Actual	ARIMA	LSTM
2018-04	66.9	70.4409	72.987
2018-05	66.3	62.4441	64.5481
2018-06	51.1	54.0051	52.5231
2018-07	50.1	48.3813	47.4711
2018-08	50.3	45.9625	50.3333

Table 3 : MAPE and RMSE values of all the algorithm

City (Pollutant)	Model	MAPE	RMSE
Margao (SO_2)	ARIMA	5.7304	0.6444
	LSTM	3.6485	0.3697
Margao (NO_2)	ARIMA	7.3615	0.9967
	LSTM	6.1957	0.8001
Margao (PM_{10})	ARIMA	15.6621	6.577
	LSTM	9.3682	4.5959
Sanguem (SO_2)	ARIMA	9.6743	0.9534
	LSTM	6.5892	0.6964
Sanguem (NO_2)	ARIMA	3.3521	0.5688
	LSTM	3.3661	0.5784
Sanguem (PM_{10})	ARIMA	5.7695	3.3945
	LSTM	3.9679	3.1323

V. CONCLUSION

In this study, time series forecasting methods like ARIMA, and LSTM are used to predict the concentration of various air pollutants like SO2, NO2, and PM10 in Margao and Sanguem. The experimental data used in this project is obtained from Goa State Pollution Control Board. Performance measures like MAPE and RMSE are used for comparing the performance of the ARIMA, and the LSTM model. From the results obtained, it is observed that LSTM method gives better forecasting results when compared to ARIMA model. Table 7 shows that the highest MAPE of 15% is obtained by using ARIMA in Margao for PM10 air pollutant. It is also seen that in Sanguem for NO2 air pollutant the MAPE obtained from both ARIMA, and LSTM is 3.3%.

REFERENCES

[1] "Introduction to Time Series and Forecasting", Second Edition by Peter J. Brockwell, Richard A. Davis, ISBN 0-387-95351-5, SPIN 10850334.

[2] "An Introductory Study on Time Series Modeling and Forecasting" by Ratnadip Adhikari, R.K. Agarwal, ISBN : 978-3-659-33508-2.

[3] National Air Quality Index:indiaenvironmentportal.org.in/files/file/Air%20Quality%20Index.pdf

[4] Richa Handa, A.K. Shrivas, H.S. Hota, "Financial Time Series Forecasting using Back Propagation Neural Network and Deep Learning Architecture", International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-1, May 2019, ISSN: 2277-3878.

[5] Nurul Nnadiah Zakaria, Mahmod Othman, Rajalingam Sokkalingam, Hanita Daud, Lazim Abdullah, Evizal Abdul Kadir, "Markov Chain Model Development for Forecasting Air Pollution Index of Miri, Sarawak", Sustainability 2019, 11, 5190.

[6] C. L. Karmaker, P. K. Halder, E. Sarker, "A Study of Time Series Model for Predicting Jute Yarn Demand: Case Study", Journal of Industrial Engineering, Hindawi, Volume 2017, Article ID 2061260.

[7] Kostandina Veljanovska, Angel Dimoski, "Air Quality Index Prediction Using Simple Machine Learning Algorithms", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 7, Issue 1, January - February 2018, ISSN 2278-6856.

[8] Songting Xing, Yuansheng Lou, "Hydrological time series forecast by ARIMA+PSO RBF combined model based on wavelet transform", 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC 2019), 1711-1715.

[9] Lintao Yang, Honggeng Yang, "A Combined ARIMA-PPR Model for Short-Term Load Forecasting", IEEE PES Innovative Smart Grid Technologies Asia, 2019, 3363-3367.

[10] Purwanto, Chikkannan Eswaran, Rajasvaran Logeswaran, "A Comparison of ARIMA, Neural Network and Linear Regression Models for the Prediction of Infant Mortality Rate", Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation, 2010, 34-39.

[11] Subhra Rani Patra, "Time Series Forecasting of Air Pollutant Concentration Levels using Machine Learning", Advances in Computer Science and Information Technology (ACSIT), Volume 4, Issue 5, October-December, 2017, 280-284

[12] Naveen V, Anu N, "Time Series Analysis to Forecast Air Quality Indices in Thiruvananthapuram District, Kerala, India", Naveen V. Int. Journal of Engineering Research and Application, ISSN : 2248-9622, Volume 7, Issue 6, (Part -3) June 2017, 66-84.

[13] Soawalak Arampongsanuwat, Phayung Meesad, "Prediction of PM10 using Support Vector Regression", International Conference on Information and Electronics Engineering, 2011, IPCSIT Volume 6.

[14] Kunhui Lin, Qiang Lin, Changle Zhou, Junfeng Yao, "Time Series Prediction Based on Linear Regression and SVR*", Third International Conference on Natural Computation (ICNC 2007), 0-7695-2875-9/07.

[15] Soren Kejser Jensen, Torbon Bach Pedersen, "Time Series Management Systems: A Survey", IEEE Transactions on Knowledge and Data Engineering, Volume 29, No.11, November 2017.

[16] Weiqiang Wang, Ying Guo, "Air Pollution PM2.5 Data Analysis In Los Angeles Long Beach With Seasonal ARIMA Model", International Conference on Energy and Environment Technology, 2009.

[17] Rashmi Bhardwaj, Dimple Pruthi, "Time series and Predictability analysis of Air Pollutants in Delhi", 2nd International Conference on Next Generation Computing Technologies (NGCT-2016), October 2016.

[18] Utpal Kumar Das, Kok Soon Tey, Mehdi Seyedmahmoudian, Mohd Yamani Idna Idris, Saad Mekhilef, Ben Horan, Alex Stojcevski, "SVR-Based Model to Forecast PV Power Generation under Different Weather Conditions", Energies 2017, 10, 876.

[19] Huixiang Liu, Qing Li, Dongbing Yu, Yu Gu, "Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms", Applied Sciences Journal, 2019, Volume 9, Issue 19, 4069.

[20] Wei Kexin, Du Mingxing, "A Temperature Prediction Method of IGBT Based on Time Series Analysis", 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010, Singapore.

[21] Adil Moghar, Mhamed Hamiche, "Stock Market Prediction Using LSTM Recurrent Neural Network", International Workshop on Statistical Methods and Artificial Intelligence (IWSMAI 2020) April 6-9, 2020, Warsaw, Poland.

[22] Siddharth Banyal, Pushkar Goel, Deepank Grover, "Indian Stock-Market Prediction using Stacked LSTM AND Multi-Layered Perceptron", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-3, January 2020.

[23] T. Sujjaviriyasup, K. Pitiruek, "A Comparison Between MODWT-SVM-DE Hybrid Model and ARIMA Model in Forecasting Primary Energy Consumptions", IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2017, Singapore.