International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 378-382
Published Online August 2019 in IJEAST (http://www.ijeast.com)

FORECASTING OZONE CONCENTRATION DATA: ARIMA V/S LSTM

Harguna Sood
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala, India

Devanshu Narula
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala, India

Prashant Singh Rana
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala, India

Abstract—Forecasting time series data is an important subject in climate monitoring, weather forecasting and pollution level estimation. Traditional techniques used are univariate Moving Average (MA) and Autoregressive Integrated Moving Average (ARIMA). ARIMA models have proven their superiority in precision and accuracy of predicting the next lags of time series. Due to the recent advancements in computational power of computers we are able to use data intensive techniques such as deep learning. The question explored in this paper is that whether or not the deep learning-based algorithms, like "Long Short-Term Memory (LSTM)", is better or not when compared to the traditional algorithms when used for time series forecasting of weather data. The study conducted here shows that ARIMA outperformed algorithms such as LSTM for short-term weather- related data prediction. The ARIMA model provided a decent reduction in error rate when compared to the LSTM approach. Also, there was noticeable difference in the overall processing time for both the algorithms with the ARIMA model finishing first, thereby providing reduction in the running time required for such type of operations.

I. INTRODUCTION

In the last 30-40 years, air pollution has become one of the major environmental problems faced by dozens of countries around the world. Several developed countries now also provide forecasts of air quality along with the general weather forecasts ie. temperatures, humidity, day time hours etc. Surface ozone (O3) has been recognized as one of the key air pollutants. Studies have shown that there exists a relation between ozone exposure and cardiovascular and respiratory mortality (Bell et al. 2005). Therefore, predicting daily surface ozone concentrations has become an important part of the daily weather forecasting. Such forecasts of air quality are usually made with the help of predictive air pollution models. Earlier we were using models based on Gaussian approach for predicting air pollution concentration (Cheng et al. 2011). But now due to vast improvements in storage and computational technologies various other techniques are being employed for this work. The main objective of this paper is to assess which methods offer the best forecast with respect to lower errors and higher accuracy. There are various methods for time series forecasting available to us nowadays. The most used method is univariate "Auto-Regressive Moving Average (ARMA)" for time series data in which Auto-Regressive (AR) and Moving Average (MA) models are combined. Univariate "Auto-Regressive Integrated Moving Average (ARIMA)" is a particular type of ARMA in which differencing is considered. Due to the research efforts of many, newer and more advanced algorithms and techniques are being developed in the field of machine learning and particularly in deep learning. Support Vector Machines, Random Forests and Neural Networks have gained popularity due to their better results. Deep learning- based approaches like Convolutional Neural Networks and Recurrent Neural Networks have also become the favourites of the researchers and the industry due to their exceptional ability to deal with the non-linearities in data (Gomez et al. 2003). (Solaiman et al. 2008). In particular, LSTM (a type of RNN) has been used in many application domains such as natural language processing, handwriting recognition, speech recognition, time-series prediction as well as its applications in estimating economic and financial trends. An interesting and important research question is the accuracy and precision of traditional forecasting techniques when compared to deep learning-based forecasting algorithms. This paper compares the performance of traditional methods and the more recently developed deep learning-based methods. The methods representing these classes are the ARIMA model for the traditional approach and LSTM model for the deep learning- based approach. Here ARIMA was chosen because of the non- stationarity in the data and LSTM is chosen due to its ability to retain memory of previous data which is particularly useful in forecasting.

Auto Regressive Integrated Moving Average Model (ARIMA) ARIMA combines two processes - Autoregressive (AR) process and Moving Average (MA) processes and builds a composite model of the time series.

As acronym indicates, ARIMA, p,d,q captures the key elements of the model: - AR: Auto Regression. A regression model that uses the dependencies between an observation and a number of lagged observations (p). - I: Integrated. To stabilize the series by subtracting its current and previous values d times (d). - MA: Moving Average. Moving average parameters consider the relation between the observations in period t to the errors from the previous time periods (q).

ARIMA models are applied where data has non-stationarity. This is particularly useful for time series data where integration and memory are very important. Since we are using hourly collected Ozone data, LSTM can play a good role and can give the ARIMA model a run for its money and might even win in some case.

Artificial Neural Network (ANN) Artificial Neural networks are a computing system inspired by the natural neural networks that are present inside the brains of animals. These are made up of connections of nodes which as a unit form a neuron, similar to the ones present in the biological ones. A common implementation of the ANNs consists of three layers: an input layer, hidden layer(s) and an output layer. The number of nodes in each layer depend on the dimensions of the dataset being taken into consideration. The nodes are connected through edges which carry weights. These weights are multiplied to the values when going from one layer to the next. And sometimes there are some activation functions like sigmoid, hyperbolic tangent, rectified linear unit etc. which are also applied once the values are sent from one layer to the next. The weights play the most important role in the decision-making process as it is up to them to decide which inputs will pass from one layer to the next. Therefore, it is the weights that are adjusted at every step during the learning stage of the neural networks. There are two processes going on during this stage- feed forward and back propagation. During the feedforward process the weights multiply with the inputs and present an output. During the backpropagation process, the error generated during the feedforward process is propagated backwards so as to adjust the weights accordingly to match the targets given in the training dataset. When the output layer with the least amount or error or lowest cost is generated, the training stops and the neural networks is ready to make predictions on new data sets (Sharma et al. 2012).

Recurrent Neural Network (RNN) Recurrent Neural Networks are a particular class of artificial neural networks that show temporal dynamic behavior which makes them particularly useful for tasks like speech and hand-writing recognition.

RNNs use their state memory to process input sequences. They are useful in forecasting and trend prediction. The state memory helps them to learn from previously observed steps and then make predictions for continuing steps. And because of this the previously observed data need to be stored and RNNs use their hidden states as the stores for this data.

Long Short-Term Memory (LSTM) The LSTM is a particular kind of RNN with the ability to memorize the sequence of data. This ability is due to the use of special gates inside an LSTM cell. The inner components of the LSTM cell are shown in Diagram 1.

What separates the LSTMs from the rest is their ability to add or subtract information from the cell state, using carefully controlled gates. The sigmoid activation layer outputs numbers between zero and one, describing how much of each component should be let through (Olah et al. 2015). An output value of zero- “let nothing through,” and a value of one - “let everything through!” An LSTM cell has three gates to control the cell state:

- **Forget Gate**: outputs 1 means “completely keep this” and 0 means “completely get rid of this”.
- **Memory Gate**: decides which new data should be retained. First, a sigmoid layer, called the “input layer” or “input gate” decides which values will be modified. Then there is a tanh activation layer that makes a new vector that could be added to the state.
- **Output Gate**: decides what will be the yield out of each cell. The yielded value will be based on the cell state along with the filtered and newly added data.

The main questions answered through this research is that which of ARIMA or LSTM, performs better with more accurate prediction of weather (specifically O3 related) time series data?

II. DATASETS

The dataset was provided by the Weather department at Thapar Institute of Engineering and Technology, Patiala, India. It consists of hourly measurements of Ozone levels in Patiala City (areas around Thapar Institute of Engineering and Technology), from 1st January 2013 to 31st December 2015. The O3 levels are measured in Dobson Units.

III. DATA PREPARATION

The Dataset had a few missing values and some outliers. The data set was split in two: training dataset and test dataset where 70% of the dataset was used for training and the remaining 30% for testing the accuracy of models.

IV. ASSESSMENT METRIC

The Root Mean Squared Error (RMSE) is a commonly used metric for checking the prediction accuracy of the results of a model. It measures the differences between actual and predicated values. RMSE is a measure of how spread out the data points are. It tells us how concentrated the data is around the line of best fit. The formula for computing RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_i - x_i')^2}$$

The greatest benefit of using Root Mean Square Error is that by squaring the difference term, it is able to penalize larger errors, more than what they would penalize the smaller errors. This is of great help when dealing with a large number of features (variables).

V. METHODOLOGY

A. ARIMA

ARIMA works best on stationary data. For this, Dickey Fuller test was used to verify stationarity. ARIMA has one important parameter that is order of the model.

Order of the model has inputs: p, d, and q

p - AR order: number of previous auto-regressive terms considered for next term
d - integration order: level of differencing between current tern and previous term
q – MA order: number of previous deviations from mean taken into account for next deviation from mean

To find P, we use ACF plot and to find q, we use PACF plot. Since, our data is already satationary and we don't have to do any differencing, d will be 0. After ACF and PACF plots, p comes out to be 9 and q to be 5. Thus, order for ARIMA model becomes (9,0,5). Total dataset was divided into testing and training. Testing dataset consisted of 100 observations and the rest was treated as training dataset. After each prediction was made (i.e. after each iteration), a new model was built using the above-mentioned order on training set and first observation from testing was appended to the end of training dataset. Simultaneously, the first observation was removed from testing dataset. Thus, making it a rolling ARIMA. The reason for using rolling ARIMA was to make predictions and check for error rate simultaneously (Adhikari et al. 2013).

B. LSTM

LSTM requires data to be scaled to the range -1 to 1. To do so, MinMaxScaler was used to transform the dataset. Then the data was shifted by one lag to form another column. The empty row in the shifted data is replaced by 0. The original column act as Y and the shifted column becomes X, for the model. X: Actual Y: Predicted

After each prediction was made (i.e. after each iteration), a new model was built on training set and first observation from testing was appended to the end of training dataset. Simultaneously, the first observation was removed from testing dataset. This was done for both, X and Y columns. Thus, making rolling LSTM. Various parameter combinations were tried for the model. Variations included changing number of epochs,

neurons, resetting states, number of layers and batch size. The best combination was selected based on accuracy and training time, which turned out to be:

Batch size = 1, Neurons = 128, Layers = 1, Epochs = 2, Stateful = True (Lipton et al. 2015)

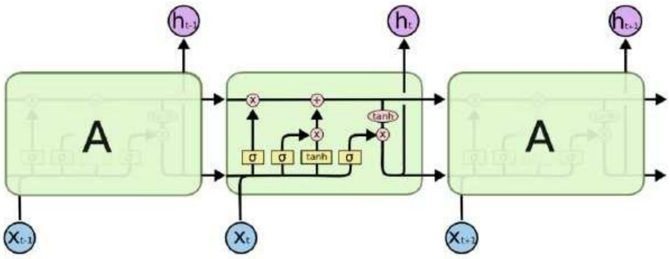


Diagram 1. The internal structure of an LSTM [Colah's Blog, 2015].

VI. PREVIOUS WORK

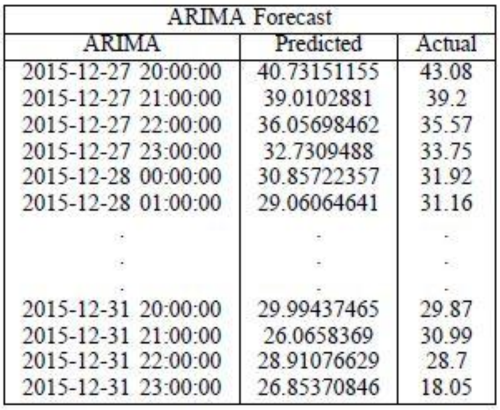
Work has been done in predicting ground ozone levels in several places and using various techniques. The methods that were used in these attempts included - multiple linear regressions, non-linear regressions and fuzzy systems. It was observed that hybrid model outperformed individual linear and nonlinear models. Also, work has been done on comparing the performance of LSTM (or in general about deep learning-based techniques) and ARIMA techniques and there have been varying results for different types of data. A study conducted for comparing the performance of ARIMA and LSTM models on economic data was conducted which showed that for such data LSTMs performed better (Siami et al. 2018).

VII. RESULTS

The results for the predictions by the ARIMA and LSTM models are reported in the table below. Root Mean Squared Error (RMSE) for model using Rolling ARIMA and Rolling LSTM models are 4.79 and 5.48, respectively. Thus, on an average ARIMA provides 13 percent reductions in error rates when compared with LSTM based approach. The RMSE values clearly indicate that ARIMA-based model outperformed the LSTM-based model with a decent margin, (between 12% - 16% reduction in error rates). Also, the running/processing time of the ARIMA model was around 10 hours and for the LSTM model was around 20 hours. Thus, we can see that the ARIMA model takes half the amount of processing than the LSTM model. This shows that the ARIMA model outperformed the LSTM both in terms of time required and accuracy.

TABLE I: Forecast of ARIMA Model

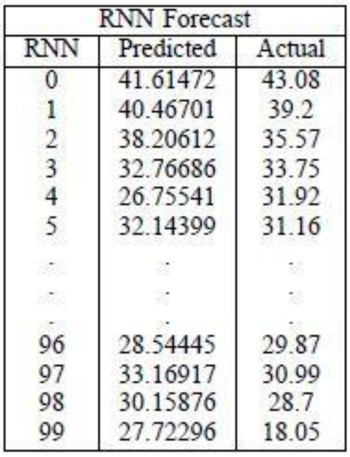
The following table shows the given values of ozone levels and the values predicted by the rolling ARIMA model. We can already see from these few examples that the ARIMA model is performing really well.



ARIMA Forecast		
ARIMA	Predicted	Actual
2015-12-27 20:00:00	40.73151155	43.08
2015-12-27 21:00:00	39.0102881	39.2
2015-12-27 22:00:00	36.05698462	35.57
2015-12-27 23:00:00	32.7309488	33.75
2015-12-28 00:00:00	30.85722357	31.92
2015-12-28 01:00:00	29.06064641	31.16
.	.	.
.	.	.
.	.	.
2015-12-31 20:00:00	29.99437465	29.87
2015-12-31 21:00:00	26.0658369	30.99
2015-12-31 22:00:00	28.91076629	28.7
2015-12-31 23:00:00	26.85370846	18.05

TABLE II: Forecast of RNN(LSTM) Model

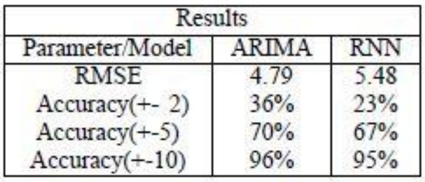
The following table shows the given values of ozone levels and the values predicted by the rolling LSTM model. Seeing just these few results we can say that this model isn't performing as well as the ARIMA.



RNN Forecast		
RNN	Predicted	Actual
0	41.61472	43.08
1	40.46701	39.2
2	38.20612	35.57
3	32.76686	33.75
4	26.75541	31.92
5	32.14399	31.16
.	.	.
.	.	.
.	.	.
96	28.54445	29.87
97	33.16917	30.99
98	30.15876	28.7
99	27.72296	18.05

TABLE III: Comparison of ARIMA and LSTM

The following table shows the results of the comparative study performed between LSTM and ARIMA models for Ozone Level Forecasting. The Results include the Root Mean Square Error which is lower for ARIMA based model and also the trend in accuracy based on the error limit used.



Results		
Parameter/Model	ARIMA	RNN
RMSE	4.79	5.48
Accuracy(+- 2)	36%	23%
Accuracy(+-5)	70%	67%
Accuracy(+-10)	96%	95%

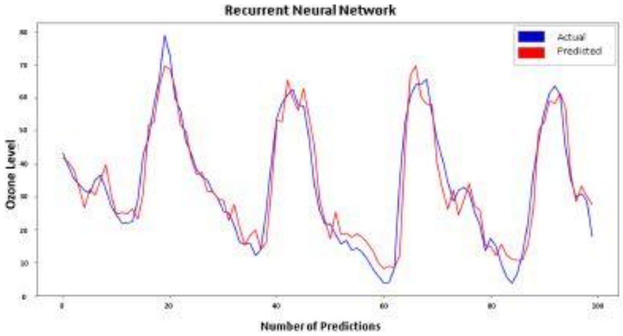


Fig. 1: Plot of the actual predictions and the predictions of the RNN(LSTM)

The graph is the plot of the results of the LSTM model and the actual values of ozone given in the dataset.

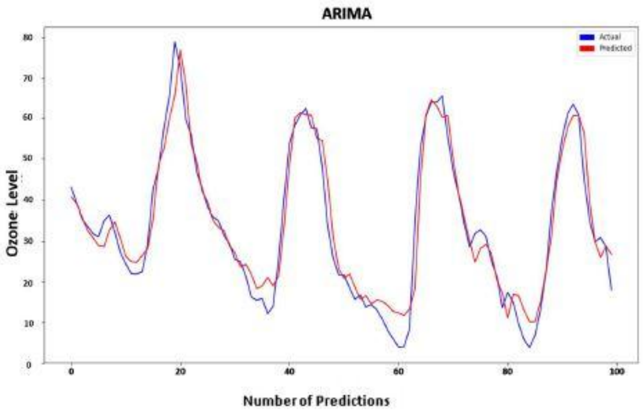


Fig. 2: Plot of the actual predictions and the predictions of the ARIMA Model

The graph is the plot of the results of the ARIMA model and the actual values of ozone given in the dataset.

VIII. CONCLUSION

With the amount of work and research being done in the field Machine Learning, and in particular deep learning, these techniques are gaining popularity among researchers of all fields. But how powerful and valid are these new techniques when compared with traditional methods. This paper provides a comparison on the basis of accuracy between ARIMA and LSTM for forecasting time series data and in particular weather-related data. These two techniques were applied on hourly collected Ozone level data collected locally at Thapar University, Patiala and the results indicated that ARIMA was superior to LSTM by a decent amount. More specifically, the ARIMA-based algorithm improved the prediction by 13% on average compared to LSTM. Also, we found that the deep learning algorithms being data intensive take a much longer time to complete processing as compared to other algorithms like the ARIMA. These advantages of the ARIMA make it better suitable for these types of tasks. The work described in this paper proves that the deep learning-based algorithms available to us are not always better as compared to the traditional methods. There are several problems in sectors like finance, economics, market trends, customer preferences etc. that can be taken up using deep learning and there deep learning techniques might have an upper hand. But for environmental data, ARIMA triumphs over LSTM.

IX. ACKNOWLEDGEMENT

The authors are grateful to Computer Science and Engineering Department of Thapar Institute of Engineering and Technology for the assistance and support that they provided in helping us complete our study.

X. REFERENCES

[1] Bell ML, Dominici F, (2005), “A meta-analysis of time series studies of ozone and mortality with comparison to the national morbidity, mortality, and air pollution study.” *Epidemiology*, vol. 16, no. 4, pp. 436–445, 2005.

[2] Ching-Hsue Cheng, Sue-Fen Huang, (2011), “Predicting daily ozone concentration maxima using fuzzy time series based on a two-stage linguistic partition method,” *Computers & Mathematics with Applications, Elsevier*, vol. 62, no. 4, pp. 2016–2028, 2011.

[3] Gómez P., Nebot A.,Ribeiro S., Alquézar R.,Mugica F., and Wotawa F.,(2003), “Local maximum ozone concentration prediction using soft computing methodologies in,” *Systems Analysis Modelling Simulation*, vol. 43, no. 8, pp. 1011–1031, 2003.

[4] Solaiman T. A., Coulibaly P., and Kanaroglou P., (2008) “Ground-level ozone forecasting using data-driven methods,” *Air Quality, Atmosphere & Health*, vol. 1, no. 4, pp. 179–193, Dec 2008.

[5] Sharma Vidushi, Sachin Rai, (2012), “A COMPRE-HENSIVE STUDY OF ARTIFICIAL NEURAL NETWORK,” *International Journal of Advanced Research in Computer Science and Software Engineering*, Oct. 2012.

[6] Olah C., (2015) “Understanding lstm networks,” *Co-lah’s Blog*, Aug 2015. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[7] R. S. U. Excel, “Dickey-fuller test.” [Online]. Available: http://www.real-statistics.com/time-series-analysis/stochastic- processes/dickey-fuller-test/

[8] Adhikari R. and Agrawal R. K., (2013) “An Introductory Study on Time Series Modeling and Forecasting,” *ArXiv e- prints*, Feb. 2013.

[9] Lipton Z. C., Berkowitz J., and Elkan C., (2015) “A Critical Review of Recurrent Neural Networks for Sequence Learning,” *ArXiv e-prints*, May 2015.

[10] Siami-Namini S. and Siami Namin A., (2018) “Fore-casting Economics and Financial Time Series: ARIMA vs. LSTM,” *ArXiv e-prints*, Mar. 2018.

[11] Stevenson, Simon, (2007), “A comparison of the fore- casting ability of ARIMA models”*Journal of Property Invest- ment & Finance Health*, vol. 25, no. 3, pp. 223–240, 2007.

[12] Gómez, Pilar and Nebot, Angela and Ribeiro, Sabrine and Alquézar, Renéand Mugica, Francisco and Wotawa, Franz, (2003), “Local maximum ozone concentration predic- tion using soft computing methodologies”*Systems analysis modelling simulation*, vol. 43, no. 8, pp. 1011–1031, 2003.