Water Level Forecasting in Reservoirs Using Time Series Analysis – Auto ARIMA Model

Avinash Reddy Kovvuri, Padma Jyothi Uppalapati(✉), Sridevi Bonthu, and Narasimha Rao Kandula

Vishnu Institute of Technology, Bhimavaram, A.P, India
20pa1a0585@vishnu.edu.in, padmajyothi64@gmail.com

Abstract. Forecasting the upcoming water level of a dam or reservoir is the goal of water level forecasting in reservoirs. In order to predict the water level of the dam or reservoir for the subsequent consecutive time interval, this paper proposes a method based on the ARIMA (Auto Regressive Integrated Moving Averages) machine learning model, which fed on historical data of water levels with respect to consecutive time intervals. Additionally, the anticipated output, whether it be in TMC or MFTC units, is depending on the data that is given. The model's performance is further examined in the study using certain machine learning metrics.

Keywords: Water Level Forecasting · ARIMA · Time Series Analysis · Auto ARIMA

1 Introduction

Water Level Forecasting in Reservoirs is to forecast the future outflow of city or country reservoirs. The purpose of predictions or forecasts is to make ourselves ready to meet the future needs and to make the ruling bodies aware of the future trends of any given commodity. This solution or analysis is used as a basis for any government bodies to make any substitutions to the insufficient levels of water for the people. Now a days the techniques of machine learning are making a huge impact in society or the market by getting an analysis on the future needs. And the machine learning model used in this is ARIMA which is a Time Series Analysis model based on the single variable data which varies with respect to the time [1].

The Word Time Series Analysis means by analysing a sequence of data collected over an interval of time [2]. ARIMA or Auto Regressive Integrated Moving averages which is a combination of AR model (Auto Regressive) and MA (Moving Averages) model with respect to differencing (Stationarizing data) [3].

The incident laid foundation for the idea - Forecasting water levels was due to the vast growth of the capital city of Tamil Nadu state in India named Chennai

N. Gupta et al. (Eds.): IC4S 2022, LNICST 472, pp. 192–200, 2023.
https://doi.org/10.1007/978-3-031-28975-0_16

from 1893 to 2017, areas of the surrounding floodplain, along with its lakes and ponds had disappeared. This leads to the decrease of Chennai's water bodies from 12.6 km^2 to about 3.2. And, finally the water crisis in 2019 was declared as "Day Zero" by city officials on 19th June. The people along with the government face a huge problem with the sudden lack of water for the entire city. In order to solve the above, forecasting the water levels through Time Series Analysis Technique in machine learning will –

- The forecasted data used by the government to fulfil the needs of the people in Chennai city regarding water whether there is any chance of occurrence of water crises by comparing it with the population.
- With forecasted data, the officials can also evacuate the people of Chennai city.
- Prevention measures for any issues related to floods were taken into consideration.

2 Chennai City's Reservoirs Data

Actually the city Chennai has a source of 4 main reservoirs named Poondi, Cholavaram, Redhills, Chembarambakkam. The source for these 4 dams was rainfall water. These 4 dams together with addition of extra rainfall water adding up to another reservoir as a source for people of Chennai city. The water in the final reservoir plays a key role for the people of Chennai city in order to make decisions on the usage of water. According to the statistics on the day zero in chennai in 2019, Due to the drain of reservoirs named poondi from 3,231 to 22, cholavaram from 1,081 to 0 and redhills, chembarambakkam from 3300 and 3645 to 0, being india's fourth largest city, needs about 800 million litres of water daily but the public water board has been able to supply 525 million litres which impacts the people along with government officials of the state.

3 Time Series Analysis

Time Series Analysis can be applied on the time-series data which is a sequence of data noted or stored with respect to specific intervals of time in chronological order [4]. This level of forecasting has a huge impact on economic, commercial and other financial aspects of business and life. This type of analysis plays a major role in situations of natural calamities.

Features of Time-series are:

- Trend
- Seasonality
- Cyclicality
- White noise

It's mandatory that the data must be stationary if we are using time-series forecasting. Stationarity in the data indicated that the distribution of the data doesn't change with the time. We explain it with the statistical measure such as trend, variance, autocorrelation to remain constant. In general, stationary data is said to have the following three properties:

- Trend = zero
- Variance = constant
- Autocorrelation = constant

The trend may be upward, downward or constant. For a stationary time-series the trend must be constant. In Simple words, Variance is of average distance of the specified data with respect to certain time interval from the zero line in the graph if we draw the varying time quantity with respect to time.

The Variance must be constant for a time-series data to be Stationary. Autocorrelation is nothing but how each value or observation in time-series data relates to its neighbours and it must be constant for a data to be stationary.

If the data is non-stationary and if fed to the model, the final results may be highly inaccurate. The conversion of non-stationary to stationary by using techniques of Differencing or log of the series. Most commonly used method is differencing.

4 ARIMA

ARIMA or Auto Regressive Integrated Moving Averages is a statistical analysis model there on top of all other time-series models which feed on time series data to either get clear insight about the data or forecast future values. It is a combination of AR and MA with respect to differencing (Stationarizing data) [5].

4.1 Auto Regression (AR)

In simple words, it is a model which regresses on its own lagged, or previous values by changing variable [6]. The Eq. 1 shows the equation represents the AR model. Epsilon-t is white noise or term that represents shock-term at that particular time. White noise is a series of measurements in which each value is uncorrelated with previous values.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t \tag{1}$$

4.2 Integrated (I)

It refers to some techniques where the actual data was differenced to convert it to stationary and finally new data is replaced with old non-stationary one.

4.3 Moving Averages (MA)

It represents an equation which regresses values of time series against previous shock values of the same time series as of Eq. 2.

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_p \epsilon_{t-p} \tag{2}$$

The Eqs. – 1, 2 represents the equation which shows the combination of rela tion of specified data to its previous with respect to time with an order of q. The three components involved in ARIMA which are of integers whose functionality defined as:

- p: Integer value, which decides the number of values to be regressed is used in the AR model.
- d: Integer value, which is the number of times the differencing tech-nique applied on data in order to convert to stationary.
- q: Integer value, which decides the number of shock terms with respect to time, is used in the MA model.

$$y_t = \mu + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=1}^{q} b_i \epsilon_{t-i} + \epsilon_t \tag{3}$$

$$\epsilon_t = \sqrt{\sigma_t Z_t}, \sigma^2 = w + \sum_{i=1}^{p} \alpha_i \epsilon_{t^2-i} + \sum_{i=1}^{q} \beta_i \sigma_{t^2-i} \tag{4}$$

The Eq.-3, 4 shows the equation of the ARMA model which is a time-series model, regressed on previous values and previous shock terms]cite7. After forecasting through the ARIMA model, the values are stationary but we need non-stationary (based on input). For that we actually convert by using the values of differencing and the technique used in it.

For **differencing**, we use cumulative sum. The count of differencing applied on the data to achieve stationarity before it is fed to the model will be the number of times the cumulative sum must be applied in order to achieve non-stationarity in the result.

For the **log method**, we use an exponential function. The number of times the log method is applied on the data to achieve stationarity before it is fed into the model will be the number of times the exponential function will be used to achieve nonstationarity in the result.

It's important to note that when we process the data, we tend to remove the seasonality (patterns repeated at regular intervals) in the data. Even after we remove the seasonality in the data, but still the data holds the seasonal properties, then such data can't be fed to the ARIMA model, such data can be satisfied using the SARIMA or the Seasonal ARIMA model.

5 Data Preparation

After collecting the data with respect to time i.e., the final reservoir which is a combi-nation of all the 4 dams in Chennai city. The data is stored in colaboratory notebook of Google in order to execute. The data is dated from 2014 and to the year 2019 and the sample of top 5 observations. First step needed to check before going to feed data to the model is stationarity. Can check stationarity through visually and various tests. Through Fig. 1 and the concept of Stationarity in ARIMA in above, the data is non - stationary. And a test named AD- Fuller used to verify the data used is stationary or non-stationary.

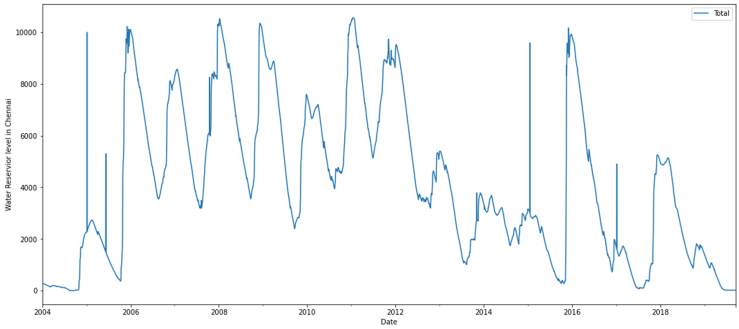

Fig. 1. Plotted Data

5.1 Ad-Fuller Test

Augmented dicky Fuller is one among the tests that are used to verify that the data used in the work is stationary or not. And is the most common test used in the processes which results in some values. If the obtained value is less than threshold then we say that data is stationary. Ad-Fuller test is performed by importing from "statsmodels.tsa.stattools". After performing the AD-FULLER Test also the result same as above stated from data visualization i.e., data is non-stationary. In order to convert data to stationary the technique of differencing or log is applied on the data and continues to repeat the same until the AD-FULLER Test results that data is sta-tionary and counts the number of time applied as the value of the term "d". Figure 2 shows the plotting of the data after converting it into stationary through differencing. And the next step is to train the ARIMA model.

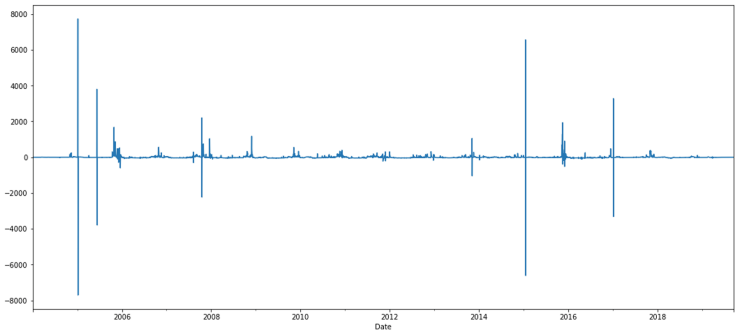

Fig. 2. Stationary Data

6 Training Arima Model

The data is ready to feed the model and the important thing is to find the parameters i.e., p, q and d. The d was also found while converting. Here the actual data i.e., nonstationary is fed along with the hyperparameters. The model itself will do the differencing through the parameter d. i.e., itself converting the data into stationary.

6.1 Experiments

The hyper-parameters were adjusted based on the result obtained through changing the values and the values which give the best scores of AIC and BIC (Minimum) are the best parameters [9]. The identification of p and q terms was done using Partial Autocorrelation and Autocorrelation in Fig. 3. After fixing the terms p, q, d the data is fed to the ARIMA Model along with the hyperparameters. And some consecutive data is left over to test the model. Finally the model is ready to predict the future values of the water levels of the Chennai city's reservoir.

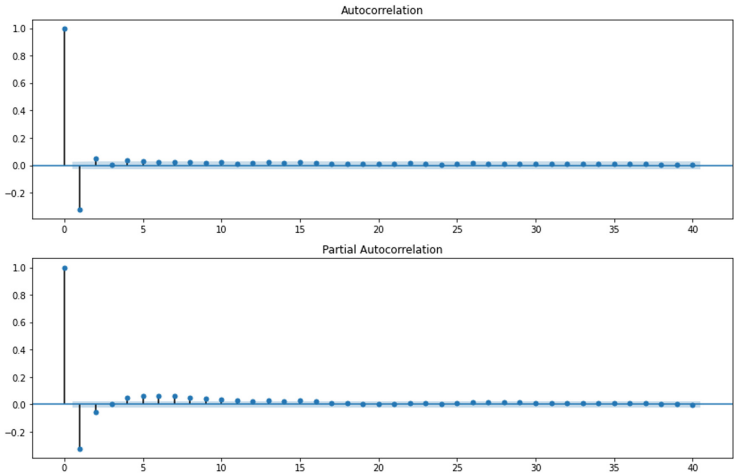

Fig. 3. Autocorrelation and Partial Autocorrelation

7 Auto ARIMA

For a while, we have been going through the process of manually fitting different models and deciding which one is best. So, we are going to automate the process. Usually, in the basic ARIMA model, we need to provide the p, d and q values which are essential. We use statistical techniques as above mentioned

in order to generate these values by performing the difference to eliminate the non-stationary and plotting ACF and PACF graphs. In **Auto ARIMA**, the model itself will generate the optimal p, d and q values which would be suitable for the data set to provide better forecasting [10].

We automated the process of finding p, d and q values by using the Auto ARIMA model and data is fed to the model in order to predict future consecutive values [11].

8 Evaluation

Finally, we need an evaluation metric in order to know the performance of the model. So, we take 20 or 30% of the whole data as test data and compare it with the predicted data from our model in order to get a score of the model. Here the end consecutive data is separated from the training data and is used to test the model.

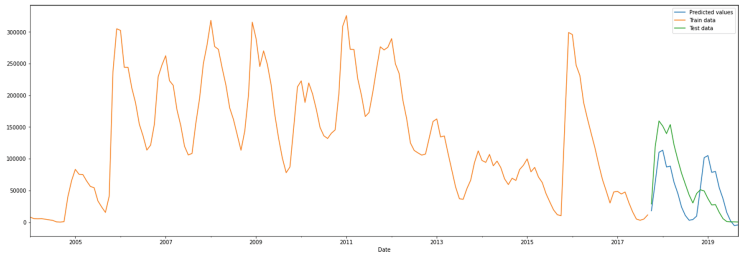

Fig. 4. Plot of Test, Predicted Data and Train Data

In Fig. 4 represents the graphical representation of the data available to us along with the differentiate between train, test data and data predicted by the model. The orange, blue and green colored lines in Fig. 4 indicates the train, predicted and test data.

The Fig. 5 indicates or plotted in order to get clear insight among predicted and actual data. The orange and blue lines in Fig. 5 indicates actual and predicted values. Figure 6 is plotted to showcase the fluctuations of current available and future pre-dicted data by plotting. The orange and blue line in the Fig. 6 indicates predicted fu-ture and total data. The evaluation metric used for the model is r2 score [12] and the obtained r2 score for the above model is 0.3284, which is greater than zero and is nearer to it. So, our model fits well and is able to predict the future data based on past outcomes.

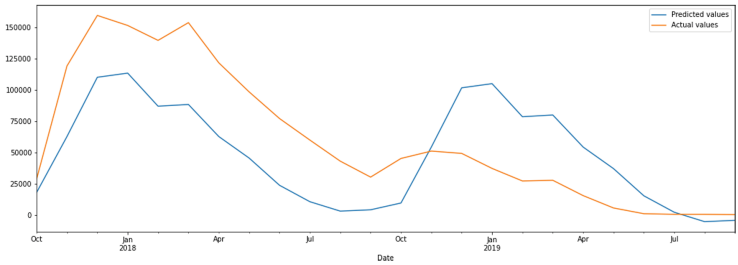

Fig. 5. Plot of Predicted and Actual Data

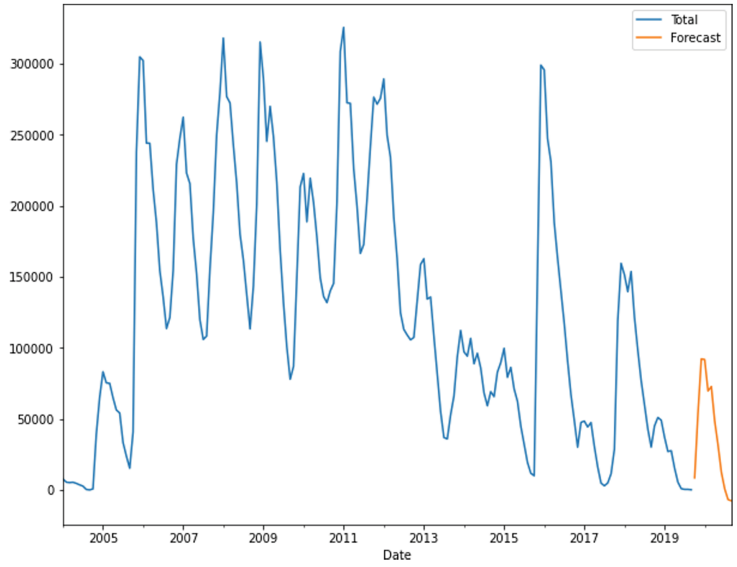

Fig. 6. Plot of Total and Future Data

9 Conclusion and Future Work

ARIMA Model is on top in time-series analysis in order to forecast upcoming data and we automated it with help of Auto ARIMA in order to increase the accuracy of the model. The proposed model is based on Time-series analysis and achieved nota-ble accuracy and good performance. As time-series analysis requires a high amount of data even for predicting for a shorter period of time. So, in order to make the model robust in terms of performance is to train the model on newly updated data at consequent time intervals. Finally the Auto ARIMA model was incorporated to know the future values of availability of water for Chennai capital city of Tamil Nadu in India.

References

1. Nguyen, X.H.: Combining statistical machine learning models with ARIMA for water level forecasting: the case of the Red river. Adv. Water Res. **142**, 103656 (2020)
2. Titolo, A.: Use of time-series NDWI to monitor emerging archaeological sites: case studies from Iraqi artificial reservoirs. Remote Sens. **13**(4), 786 (2021)
3. Wang, J., et al.: Reliable model of reservoir water quality prediction based on improved ARIMA method. Environ. Eng. Sci. **36**(9), 1041–1048 (2019)
4. Arvor, D., et al.: Monitoring thirty years of small water reservoirs proliferation in the southern Brazilian Amazon with Landsat time series. ISPRS J. Photogram. Remote Sens. **145**, 225–237 (2018)
5. Skariah, M., Suriyakala, C.D.: Forecasting reservoir inflow combining exponential smoothing, ARIMA, and LSTM models. Arab. J. Geosci. **15**(14), 1–11 (2022)
6. Huang, L., et al.: Evolutionary optimization assisted delayed deep cycle reservoir modeling method with its application to ship heave motion prediction. ISA Trans. **126**, 638–648 (2022)
7. Üneş, F., et al.: Estimating dam reservoir level fluctuations using datadriven techniques (2019)
8. Paparoditis, E., Politis, D.N.: The asymptotic size and power of the augmented Dickey-Fuller test for a unit root. Econ. Rev. **37**(9), 955–973 (2018)
9. Bai, Z., Choi, K.P., Fujikoshi, Y.: Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. Ann. Stat. **46**(3), 1050–1076 (2018)
10. Yan, B., et al.: Flood risk analysis of reservoirs based on full-series ARIMA model under climate change. J. Hydrol., 127979 (2022)
11. Tegegne, G., Kim, Y.-O.: Representing inflow uncertainty for the development of monthly reservoir operations using genetic algorithms. J. Hydrol. **586**, 124876 (2020)
12. Ali, A., Bello, A.M., Raymond, J.: Machine learning algorithms for predicting reservoir porosity using stratigraphic dependent parameters. Glob. J. Comput. Sci. Technol. (2022)