ARIMA based daily weather forecasting tool: A case study for Varanasi

NIKITA SHIVHARE, ATUL KUMAR RAHUL, SHYAM BHARI DWIVEDI and PRABHAT KUMAR SINGH DIKSHIT

Indian Institute of Technology (IIT), Banaras Hindu University (BHU), Varanasi, India

(Received 15 June 2018, Accepted 28 November 2018)

email: nikitas.rs.civ15@iitbhu.ac.in

ABSTRACT. Autoregressive integrated moving average (ARIMA) is a data mining technique that is generally used for time series analysis and future forecasting. Climate change forecasting is essential for preventing the world from unexpected natural hazards like floods, forest fires and droughts. It is a challenging task to forecast weather data accurately. In this paper, the ARIMA based weather forecasting tool has been developed by implementing the ARIMA algorithm in R. Sixty-five years of daily meteorological data (1951-2015) was procured from the Indian Meteorological Department. The data were then divided into three datasets- (i)1951 to 1975 was used as the training set for analysis and forecasting, (ii)1975 to 1995 was used as monitoring set and (iii)1995 to 2015 data was used as validating set. As the ARIMA model works only on stationary data, therefore the data should be trend and seasonality free. Hence as the first step of R analysis, the acquired data sets were checked for trend and seasonality. For removing the identified trend and seasonality, the data sets were transformed and the removal of irregularities was done using the Simple Moving Average (SMA) filter and Exponential Moving Average (EMA) filter. ARIMA is based on method ARIMA (p,d,q) where p is a value of partial autocorrelation, d is lagged difference between current and previous values and q is a value from autocorrelation. In the present study, we worked on ARIMA (2,0,2) for rainfall data and ARIMA (2,1,3) for temperature data. As a result, it estimated the future values for the next fifteen years. The root means square error values were 0.0484 and 0.085 for rainfall data and temperature data respectively which show that the algorithm worked accurately. The resulted data can be further utilized for the management of solar cell station, agriculture, natural resources and tourism.

Keywords – ARIMA, Weather forecasting, Time series analysis.

1. Introduction

In today's environment, climate plays a vital role in the sustainability of life on earth. However, the climate is continuously changing causing a severe issue for the planet. Rainfall is an essential element of the hydrological cycle. The pattern of rainfall is continuously changing due to the effects of changing climatic conditions. This change...
causes many severe problems like flooding, landslide and drought (Shivhare et al., 2017). These problems affect the agriculture and farming. In a country like India where agriculture and farming are its backbones, the most significant concern is the success and failure of the crop every year. A minute change in the seasonal rainfall and temperature may lead to a devastating effect on crops (Shivhare et al., 2018). Temperature data is also essential for the sustainability of agriculture, vegetation, water resources and tourism (Brath et al., 1999). Also, the temperature has a direct impact on evaporation & melting of snow or ice and an indirect impact on precipitation condition and atmospheric stability. Therefore accurate prediction of future rainfall and temperature are essential for preventing the country from natural disasters and managing natural resources (Shobha and Shobha, 2014; Sawale and Gupta, 2013; Haviluddin et al., 2015).

It is a challenging task to predict future climate data accurately (Nikam and Meshram, 2013). Although many algorithms have been proposed and developed but still, accurate forecasting is robust. The weather forecasting can be done using two ways: First way is to study the weather for processes to model the underlying physical laws however it may not be feasible because the processes like rainfall are the result of various complex processes which may vary both in space and time. The second way is to forecast the data based on pattern recognition algorithms like data mining techniques and machine learning (Luk et al., 2001).

Data mining is a study of how to determine underlying patterns in the data. Experts utilize data mining techniques like the autoregressive integrated moving average (ARIMA), artificial neural network (ANN) and machine learning for weather forecasting (Shobha & Shobha, 2014; John and Marohasy, 2017). The preprocessing of the data for forecasting using data mining is very important. Good preprocess data will create good knowledge and will give a better forecast result (Nhita et al., 2015).

Data mining techniques are getting utilized more and more for future weather forecasting. Scientists have reviewed different data mining techniques which could be used for data forecasting. These techniques were artificial feedforward neural network, a fuzzy interference system, decision tree method, time series analysis, learning vector quantization and biclustering techniques. Experts are even trying to use the data mining technique for predicting rainfall of an area by dependent features (Mohapatra et al., 2017). The most utilized technology of data mining is ANN and ARIMA. Scientists are utilizing ANN techniques like backpropagation neural network (BPNN) for predicting monthly rainfall data (Haviluddin et al., 2015; Sachan, 2014). Mean, maximum and minimum temperature data is predicted using BPNN for solar cell management (Routh et al., 2012; Ustaoglu et al., 2008). Sawale and Gupta, 2013 have employed the BPNN & Hopfield network model for weather forecasting. Luk et al., 2001 has applied three types of techniques-feed forward neural network (FDNN), partial recurrent neural network and time delay neural network for rainfall forecasting. Nong (2012) employed SSA for rainfall pattern forecasting, Wang and Wu (2012) proposed a novel hybrid radial basis function neural network based on wavelet support vector machine for rainfall forecasting. The scientist also forecasted one-day advanced rainfall intensity to assess the risk of the landslide (Devi et al., 2014). Some experts have used ARIMA model for weather forecasting (Pratiher et al., 2016; Toth et al., 2000). Experts have compared the ARIMA model with different models like single input and multi-input and transfer function (Saikhu et al., 2017) and ARMA (Valipour et al., 2012). They concluded that ARIMA is better than ARMA.

Weather forecasting using data mining technique can be done into methodology by (i) Statistical Method (ii) Numerical Weather Prediction Models. The weather forecasting can be done using a statistical method like autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), Autoregressive integrated moving average (ARIMA) multiple regression. Each method has its limitations. AR model regresses against the past value of the series. MA model uses past errors as an explanatory variable. The AR model is suitable only for linear correlated data and is not appropriate for nonlinear data. AR and MA can be combined to form a general and useful class of time series model known as the ARMA model; ARIMA gives better results than ARMA. The details of this methodology are explained in further sections. It is usually in statistical technique when the data is responsibly extended and the correlation between past observations is stable (Darji et al., 2015).

In this study, the ARIMA algorithm is used for future climate data forecasting. Daily climate data from 1951 to 2015 was used for the forecasting. The algorithm was implemented using the R language. The data was divided into three sets. The first set is the training set from 1951- 75 used for analysis and forecasting; the second set is the monitoring set, the forecasted data from 1975 to 1995 was used for monitoring and testing; and the third set is the validating set, the data from 1995 to 2015 was used for validation. The primary objectives of this paper are as follows:

(i) Plotting the data as a time series plot,

(ii) Checking the data, if it has any trend or seasonality,
This technique works on the assumption- “The fact that has impacted in the past will continue to impact in future.” This can be used for predicting revenue, earthquakes, new job vacancy and for predicting student enrollment in college.

Time series is a set of numerical data obtained at the regular period. There are four components of time series data - (i) Trend (ii) Seasonality (iii) Irregular Component (iv) Cyclic Component.

These components are important because if we want to predict future data, we need to remove these components to make the data stationary. In case any seasonal or random fluctuations are present, we need to change it by doing the log transformation of the time series data. In R we use log() command for log transformation. For example:

```r
>data_ts_log<- -log(data_ts)
```

To check the details of randomness, seasonality and trend of the data, the decompose () function is utilized in R. After getting these details we can remove trend and seasonality using log transformation and for smoothing the irregularity is present in the data we use filters. In R we use two types of filter for smoothing of the signals- Simple Moving Average (SMA) and Exponential Moving Average (EMA) filter. Comparison to SMA, EMA is better, as it gives better results because it gives weight to most recent observations and is a more aggressive function than SMA.

2. Methodology and data used

Varanasi is the most ancient city of India. Situated at the bank of river Ganga, this mystic city receives the footfall of lakhs of tourists every year. The latitudinal and longitudinal position of Varanasi is 25.3176° N, 82.9739° E. The population of the city is around 12,00,000. The exact location of Varanasi in India is shown in Fig. 1. Due to climate change and rapidly changing rainfall pattern, this city faces massive flood problems, like the flood of 2016, disrupting the lives of ordinary people. So it is essential to focus on the future climate data of this area for preventing the city from natural hazards.

Data used in this study is the 65 years daily meteorological data including rainfall data, minimum temperature and maximum temperature are procured from Indian Meteorological Department (IMD).

2.1. Time series forecasting

Time series forecasting is a technique of data science for predicting future events based on historical activities.
irregularities, we use simple moving average and exponential moving average. Example:

```r
> data_ts_sma <- sma(data_ts, 300)
> data_ts_ema <- ema(data_ts, 300, ratio = .25)
```

(iii) **Identify the model technique that is best suited for forecasting:** Any of the following models can be used for forecasting, like auto-regression, moving average, least square regression, former quadratic regression, exponential regression and ARIMA models. In this study, we have used the ARIMA model.

(iv) **Build the model:** This is the primary step of forecasting. Here in ARIMA for building model p, d, q values are identified (explained in section 4.3).

(v) **Do the forecasting:** In R we used Holtwinter and ARIMA functions for forecasting using following codes:

```r
> data1_forecast <- HoltWinters(data1_ts)
> plot(data1_forecast)
> data1_arima <- auto.arima(data1_ts)
> data1_arima
> data1_arima_forecast <- forecast(data1_arima, h=100)
> plot(data1_arima_forecast)
```

2.2. **Autoregressive integrated moving average (ARIMA)**

The working principle behind autoregressive (AR) model is that there is a relationship between the present value and the past values. It means that the present value is equal to past values adding with some random value. Moving average (MA) model says that present value is related to the residuals of the past. AR is not capable of forecasting nonlinear data; it can be utilized for data which are linearly related. So utilizing AR and MA models together can give a better result for nonlinear data. As the weather is a nonlinear data we need to use AR and MA together, ARMA and ARIMA are two methods in which AR and MA can be combined. However, it can be only used for stationary time series data and forecasting short-term weather data. Therefore ARIMA is developed to forecast long-term data. ARIMA gives a better result than the ARMA model.

The generalized equation used behind the ARIMA model is:

\[ \phi_p (1 - B)^d Z_t = \theta_q (B) Q_t \]  

(1)

where,

- \( Z_t \) is the present value, moreover, \( Q_t \) is the random value.
Steps followed forecasting in the ARIMA model is shown in Fig. 2.

3. Results and discussion

Based on the flowchart given in Fig. 2, firstly the climate data were plotted according to time series. The time series plots of rainfall and temperature data are shown in Figs. 3(a&b).

Secondly, to check whether the signals have trend and seasonality or not, the signals were decomposed. The decomposed plot of rainfall, minimum temperature and maximum temperature are given in Figs. 4(a-c). In each plot first row shows the signal, the second row shows the trend of the data, the third row shows the seasonality of the data and the last row shows the randomness of the data.

According to the Figs. 4(a-c), by looking at the trend, seasonality and randomness of the signal and then ACF and PACF of the data were calculated after which the p,d,q values were calculated for each signal, the values of ARIMA(p,d,q) are ARIMA(2,0,2) for Rainfall data and ARIMA(2,1,3) for minimum and maximum temperature data. Using the above values, the data was forecasted for each signal, which is shown in Fig. 5 (a-c). The blue colored lines show the exact values of the future forecasting and the grey portion and light grey colored portion shows the probable values of data for 95% and 85% precision respectively.

3.1. Skills of the ARIMA model

After determining the three parameters p,d,q evaluating the model with fit statistics is required to quantify the performance of forecast with its acceptable limits. Some of the statistical measures are RMSE (Root Mean Square Error), MAPE (Mean Absolute percentage errors), MAE (Mean Absolute Error). The values of these errors should be minimum for better performance of the model. Apart from this measure, two more statistical measures are Bayesian information criterion (BIC) and Lungs Box Q statistics. BIC can be calculated using the expression as follows:

$$BIC = \log \left( \frac{RSS}{n} \right) + \frac{k}{n} \log n$$
Figs. 4(a-c). Decomposed data to check trend and seasonality. Decomposition of additive time series data of
(a) Rainfall (b) Maximum temperature and (c) Minimum temperature

Note: Unit of temperature in degree Celsius, unit of Rainfall is mm and Time is in years

where,

\[ rss = \text{residual sum of squares,} \]

\[ k = \text{no. of coefficients estimated and} \]

\[ n = \text{no. of observations.} \]

The expression for Lungs Box Q statistics is as follows:

\[ n(n + 2) \sum_{k=1}^{h} \frac{\rho^2}{n-k} \]
SHIVHARE et al.: ARIMA BASED DAILY WEATHER FORECASTING TOOL 139

Fig. 6. The graph of forecasted and observed values

Note: Unit of temperature in degree Celsius, unit of Rainfall is mm and Time is in years

<table>
<thead>
<tr>
<th>ARIMA(p,d,q)</th>
<th>RMSE</th>
<th>MAPE</th>
<th>MAE</th>
<th>Normalized BIC</th>
<th>Lungs Box Q statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA(2,0,2)</td>
<td>0.094</td>
<td>0.32</td>
<td>0.053</td>
<td>13.3</td>
<td>0.8</td>
</tr>
<tr>
<td>ARIMA(2,1,3)</td>
<td>0.085</td>
<td>0.31</td>
<td>0.049</td>
<td>13.3</td>
<td>0.6</td>
</tr>
</tbody>
</table>

where,

\( n \) = the number of residuals,

\( h \) = number of time lags includes in the test,

\( \rho^2 \), \( k \) = the residual autocorrelation at lag \( k \).

The values of all these Fit parameters for both the ARIMA models are given in Table 1. These values show the skills of the ARIMA models for forecasting future climate data.

3.2. Advantages and limitations of the ARIMA model

ARIMA modeling has frequently been highlighted as a useful forecasting approach. The main advantage of the ARIMA model is for short-run forecasts with high-frequency data the results may be hard to beat. They also have the advantage of being less sensitive to the underlying assumptions of the nature of the data fluctuations than many other systems. However, while the general form will handle many functional forms, the specific form identified must match the actual data closely.

Limitations of ARIMA model are the ARIMA method is appropriate only for a stationary time series (i.e., its mean, variance and autocorrelation should be approximately constant through time) and it is recommended that there are at least 50 observations in the input data. Because of the extensive data requirements, the lack of available updating procedures and the fact that they must be estimated using nonlinear estimation procedures, the ARIMA models tend to be high cost. It is also assumed that the values of the estimated parameters are constant throughout the series.

4. Conclusions

In this study, the ARIMA model was used to forecast future climate data. The accuracy of the model was calculated according to the root mean square error (RMSE) estimated for each forecasting. It was found that the RMSE estimated for forecasting rainfall data using ARIMA(2,0,2) was 0.094 and the RMSE estimated for forecasting minimum and maximum temperature using ARIMA(2,1,3) was 0.085. By looking at the values of the RMSE, it can be concluded that since the error is minimal, the ARIMA model has forecasted the data accurately. Apart from these values, the forecasted value of the ARIMA model was compared with the observed values. The graphs of forecasted and observed values of minimum temperature, maximum temperature, rainfall data are shown in Fig. 6, which concludes that more than 90 percent of values forecasted by the ARIMA model are accurate. These values can be further used for hydrological and sediment yield modeling, solar cell management, managing agriculture and tourism, etc. This study showed that data mining techniques are straightforward and helpful for forecasting future weather data.

Disclaimer

The contents and views expressed in this research paper/article are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References


Mohapatra, Sandeep Kumar, Upadhyay, Anamika and Gola, Channabasava, 2017, “Rainfall prediction based on 100 years of meteorological data”, Computing and Communication Technologies for Smart Nation (IC3TSN), 2017 International Conference on, 162-166, IEEE.


Nong, Jifu, 2012, “An Ensemble Technique to Daily Rainfall Forecasting Based on SSA”, Computational Sciences and Optimization (CSO), 2012 Fifth International Joint Conference on, 5-9, IEEE.


Saikhu, Ahmad, Ariffin, Agu Zainal and Fatichah, Chastine, 2017, “Rainfall forecasting by using autoregressive integrated moving average, single input and multi input transfer function”, Information & Communication Technology and System (ICTS), 2017 11th International Conference on, 85-90, IEEE.


