Application of ARIMA Models in Forecasting Monthly Total Rainfall of Rangamati, Bangladesh

Fuhad Ahmed

Department of Natural Science (Statistics), Port City International University, Chittagong, BANGLADESH

ARTICLE INFO

*Corresponding Contact
Email: fuhad.ahmed055@gmail.com

Article URL:
https://journals.abc.us.org/index.php/ajase/article/view/1098

Source of Support: None

No Conflict of Interest: Declared

ABSTRACT

In this paper, the time series data of monthly total rainfall of Rangamati (Bangladesh) from January 1969 to December 2011 is taken for studying the trend and fitting the best Autoregressive Integrated Moving Average (ARIMA) model to forecast. Seasonal Autoregressive Integrated Moving Average model SARIMA (1,0,0)(2,0,0)[12] has been fitted to the data and then the forecast of the monthly total rainfall for the next 24 months, from January 2012 to December 2013, has been made using statistical software R. The best model forecasts for the year 2012 that the highest amount of rainfall (495.97 mm) occurs in June; whereas, in January the amount is the lowest (46.28 mm). For the year 2013, the model forecasts that August is the month having maximum (438.26 mm) and January with the lowest (68.77 mm) amount of rainfall than the other months.

Key Words: ARIMA model, SARIMA model, R software, Forecast


This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. Attribution-NonCommercial (CC BY-NC) license lets others remix, tweak, and build upon work non-commercially, and although the new works must also acknowledge & be non-commercial.

INTRODUCTION

Rangamati is a district in south-eastern Bangladesh which is a part of Chittagong Division. It is the only district in Bangladesh having international borders with two countries, which are India and Myanmar. The area of the Rangamati district is 6116 km² of which 1292 km² is riverine and, 4825 km² is under forest vegetation. The region is not well protected against flooding caused by heavy rainfall. (French et al., 1992) noted that the complex nature of the atmospheric process that produced rain made rain modeling and forecasting an arduous task. (Gwangseob and Ana, 2001) and (Hung et al., 2008) emphasized that exact forecasting...
of rain is extremely challenging in operational hydrology although there might be the
improvement of weather forecasting.

The purpose of the study is to fit the best time series model that will forecast the trend of
rainfall. The forecast will be useful for flood forecasting consequently. Early warning of
harsh weather and heavy rain can save many lives and properties.

For the experiment, the data of monthly total Rainfall of Rangamati for each year from 1969
to 2011 are collected from Bangladesh Meteorological Department (BMD). R statistical
programming software has been used to analyze the data for finding the best model and
then applying the model to forecast future rainfall.

**Methodology**

ARIMA models are capable of modeling a wide range of seasonal data. A seasonal ARIMA
model is formed by including additional seasonal terms in the ARIMA models we have seen
so far. It can be written as follows:

ARIMA (p, d, q) (P, D, Q)m: the first parenthesis represents the non-seasonal part of the
model and second represents the seasonal portion of the model, where the letter m= number
of periods per season. We use the uppercase notation for the seasonal parts of the model
and the lowercase notation for the non-seasonal parts of the model. The additional seasonal
terms are multiplied with the non-seasonal terms.

**Autoregressive (AR) Model**

The autoregressive process \{Y_t\} with order p satisfies the following equation,

\[ Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \]  \hspace{1cm} (1)

or

\[ \phi_p(x) Y_t = e_t \]

with AR characteristic polynomial

\[ \phi_p(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p \]  \hspace{1cm} (2)

The current value of the series \(Y_t\) is the linear combination of the p most recent past values
of itself plus an “innovation” term \(e_t\) that incorporates everything new in the series at time
t that is not explained by the past values. Thus, for every \(t\), we assume that \(e_t\) is independent
of \(Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots\).

**Moving Average (MA) Model**

A moving average process of order q, denoted by MA(q), is given by

\[ Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \]  \hspace{1cm} (3)

or

\[ Y_t = \theta_q(x) e_t \]

\[ \theta_q(x) = 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_p x^p \]  \hspace{1cm} (4)

The terminology moving average arises from the fact that \(Y_t\) is obtained by applying the
weights \(1, -\theta_1, -\theta_2, \ldots, -\theta_q\) to the variables \(e_t, e_{t-1}, e_{t-2}, \ldots, e_{t-q}\) and then moving the
weights and applying them to \(e_{t+1}, e_t, e_{t-1}, \ldots, e_{t-q+1}\) to obtain \(Y_{t+1}\) and so on.
**Autoregressive Moving Average ARMA (p, q) Model**

The combination of AR and MA models by adding them together as a mixed autoregressive moving average (ARMA) model of order \((p,q)\), where we have \(p\) AR terms and \(q\) MA terms.

\[
Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad \ldots \quad (5)
\]

where  
\[
\phi_p(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p \quad \ldots \quad (6)
\]

and  
\[
\theta_q(x) = 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_q x^p \quad \ldots \quad (7)
\]

**The General Autoregressive Integrated Moving Average (ARIMA) Model**

The stationary process resulting from a properly differenced homogeneous nonstationary series is not necessarily white noise. More generally, the differenced series \((1 - x)^d Y_t\) follows the general stationary ARMA \((p,q)\) process discussed earlier. Thus, we have

\[
\phi_p(x)(1 - x)^d Y_t = \theta_0 + \theta_q(x)\epsilon_t \quad \ldots \quad \ldots \quad \ldots \quad (8)
\]

Where the stationary AR operator
\[
\phi_p(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p \quad \ldots \quad (9)
\]

And the invertible MA operator,
\[
\theta_q(x) = 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_p x^p \quad \ldots \quad (10)
\]

Share no common factors. The parameter \(\theta_0\) plays very different roles for \(d = 0\) and \(d > 0\). When \(d = 0\), the original process is stationary, and \(\theta_0\) is related to the mean of the process, i.e., \(\theta_0 = \mu (1 - \phi_1 - \cdots - \phi_p)\). However, when \(d \geq 1\), \(\theta_0\) is called the deterministic trend term and is often omitted from the model unless it is needed.

The resulting homogeneous nonstationary model has been referred to as the Autoregressive Integrated Moving Average model of order \((p,d,q)\) denoted as the ARIMA \((p,d,q)\).

**Seasonal ARIMA Models**

A process \(\{Y_t\}\) is said to be multiplicative seasonal ARIMA model with nonseasonal (regular) orders \(p, d,\) and \(q\), seasonal orders \(P, D,\) and \(Q\), and seasonal period \(s\) if the differenced series

\[
W_t = \nabla^d \nabla^d_s Y_t \quad \ldots \quad \ldots \quad (11)
\]

Satisfies an ARIMA\((p,q)\times(P,Q)\) model with seasonal period \(s\). We say that \(\{Y_t\}\) is an ARIMA\((p,d,q)\times(P,D,Q)\) model with seasonal period \(s\). Using the backshift operator notation, we may write the general ARIMA\((p,d,q)\times(P,D,Q)\) model as

\[
\phi(x) \Phi(x) \nabla^d \nabla^d_s Y_t = \theta(x) \Theta(x) \epsilon_t \quad \ldots \quad \ldots \quad (12)
\]

where,

\[
\phi(x) = 1 - \phi_1 x - \phi_2 x^2 - \cdots - \phi_p x^p \quad \ldots \quad (13)
\]

and

\[
\Phi(x) = 1 - \Phi_1 x^s - \Phi_2 x^{2s} - \cdots - \Phi_P x^{Ps} \quad \ldots \quad (13)
\]

\[
\theta(x) = 1 - \theta_1 x - \theta_2 x^2 - \cdots - \theta_q x^q \quad \ldots \quad (14)
\]

\[
\Theta(x) = 1 - \Theta_1 x^s - \Theta_2 x^{2s} - \cdots - \Theta_Q x^{Qs} \quad \ldots \quad (14)
\]

Such models represent a broad, flexible class from which to select an appropriate model for a particular time series. It is observed that many series can adequately fit by these models, usually, with some parameters, say three or four.
There are four segments in Box-Jenkins methodology (Box et al., 1994). The steps are identification, estimation, diagnostic checking and, forecasting. Firstly, for becoming stationary around its mean and variance, the original series should be transformed. Secondly, applying autocorrelation and partial autocorrelation functions, the proper order of p and q should be identified. Thirdly, some non-linear optimization procedure can estimate the magnitude of the parameters, which can minimize the sum of squares of the errors. The last step is about searching for the accurate model by diagnostic checking. In this way, we get a model which is adequate and adequate. After getting the appropriate model, using minimum mean square error method (Box et al. 1994), we can generate future forecasts. To make the comparison of the performance of the other models from the same data set, we can use SARIMA models as benchmark models. Several researchers and scientist have used these models for several technical and scientific studies. (Brinskiene and Rudzkiene, 2005) used ARIMA models to model and forecast tourism development in Lithuania while (Brooks, 2002) among several others.

AIC and BIC

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are based on the maximum likelihood estimates of the model parameters. The idea is that we should estimate the parameters, for what, the probability of the observed data could be as greater as possible under the model. To choose best predictor subsets in regression and compare non-nested models, we can use AIC and BIC, as it is not possible for the other ordinary statistical tests. They measure a model’s generalizability. The general form of the Akaike Information Criterion or Bayesian Information Criterion for a model is \[-2\log L + kp\]. Here, k is 2 for AIC and \log(n)\ for BIC, and, L is the likelihood function, whereas, p is the number of parameters in the model. The value of the likelihood must be between 0 and 1, as the function here means the probability. And, that’s why, log-likelihoods are always negative numbers.

A lower Akaike Information Criterion implies a model which is counted to be closer to the truth (Akaike, 1974), as it is an estimate where we add a constant to the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model. It is always AIC, regardless of the value of n, has a chance of electing too big a model. On the other hand, for selecting too small a model, the chance of BIC is larger. Although, if n is sufficient, then the possibility of choosing too big model is very low in the case of BIC than AIC. To ensure perfection in model selection, it is best to consider both the criteria.

Augmented Dickey-Fuller test (ADF)

An augmented Dickey-Fuller test (ADF) is used to test the null hypothesis that a unit root is present in a time series sample (Fuller, 1976). The alternative hypothesis is usually stationarity or trend-stationarity depending on the version of the test used. Actually, for a more complicated set of time series models, it is an augmented version of the Dickey-Fuller test. The Augmented Dickey-Fuller (ADF) statistic is a negative number used in the test. The more negative the statistic is, the stronger the rejection of the null hypothesis that there is a unit root will be at some level of confidence.

RESULTS AND DISCUSSION

The data has been taken to study the rainfall of Rangamati from the year 1969 to 2011, and then to forecast the amount of rain for the coming years. The descriptive statistics of the data are found as below.

Table 1: Summary details of the Total Monthly Rainfall from 1969 to 2011

<table>
<thead>
<tr>
<th>Rangamati</th>
<th>Min</th>
<th>1st Quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>11.75</td>
<td>133.00</td>
<td>210.41</td>
<td>339.50</td>
<td>1276.00</td>
<td></td>
</tr>
</tbody>
</table>
The minimum and maximum rainfall are 0.00 mm and 1276.00 mm, first and third quartiles are 11.75 mm and 339.50 mm. Based on the quartile, the deviation is 163.875 mm. The average rainfall of the region is 210.41 mm, whereas, the median rainfall is 133.00 mm. Depending on the above summary, it’s not possible to decide the rain of Rangamati. That is why farther analysis has been carried out.

To check the stationary property of the data, Augmented Dickey-Fuller test is applied. The result is as following:

\[
\text{Dickey-Fuller} = -13.254, \quad \text{Lag order} = 8, \quad p\text{-value} = 0.01
\]

alternative hypothesis: stationary

The test is significant (\(p<0.05\*\)), that is the data is stationary at 5% level of significance.

Figure 1: Time series plot for Total Monthly Rainfall from 1969 to 2011..

Figure 2: Decomposed Time Series plot for Total Monthly Rainfall from 1969 to 2011

Figure 3: ACF plot for Total Monthly Rainfall from 1969 to 2011
Figure 4: PACF plot for Total Monthly Rainfall from 1969 to 2011

In Figure 1, the time series plot of 516 observations of rainfall data from January 1969 to December 2011 is shown. According to the graph, the monthly total rainfall shows seasonal variation over time. The four graphs in Figure 2 represent the original data, seasonal component, trend component and the remainder and this display the periodic seasonal pattern extracted out from the original data and the trend. Then both the plots of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) have also been drawn in Figure 3 and Figure 4 respectively.

Table 2: All possible ARIMA models along with the best model

Table 3: Estimate of the coefficients of the best model

From Table 2, it is seen that a total of 13 Seasonal ARIMA models are generated. Out of all those models, Seasonal ARIMA model ARIMA (1,0,0)(2,0,0)[12] with non-zero mean is
found as the best model based on the smallest AICc (6743.899). The best SARIMA model found here is a combination of the Non-seasonal Autoregressive coefficient of the model of order 1 and Seasonal Autoregressive model of order 2. Table 3 gives the value of both the Non-seasonal Autoregressive and Seasonal Autoregressive coefficients along with the mean and standard error. The result of $\sigma^2$ is estimated as 26950 with the Log likelihood -3366.89. The outcomes of AIC, AICc, and BIC are 6743.78, 6743.9, and 6765.01 respectively, which are the lowest values compared to other SARIMA models.

Table 4: Point and Interval estimation of the forecast for Total Monthly Rainfall from 2012 to 2013

<table>
<thead>
<tr>
<th>Month</th>
<th>Point Estimate</th>
<th>Low 80</th>
<th>High 80</th>
<th>Low 95</th>
<th>High 95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2012</td>
<td>46.27961</td>
<td>-164.106612</td>
<td>256.6658</td>
<td>-275.47835</td>
<td>368.0376</td>
</tr>
<tr>
<td>Feb 2012</td>
<td>55.18009</td>
<td>-156.304289</td>
<td>266.6645</td>
<td>-268.25735</td>
<td>378.6175</td>
</tr>
<tr>
<td>Mar 2012</td>
<td>117.43281</td>
<td>-94.063024</td>
<td>328.9287</td>
<td>-206.02216</td>
<td>440.8878</td>
</tr>
<tr>
<td>Apr 2012</td>
<td>124.26461</td>
<td>-87.231548</td>
<td>335.7606</td>
<td>-199.19055</td>
<td>447.7198</td>
</tr>
<tr>
<td>May 2012</td>
<td>280.91734</td>
<td>69.421381</td>
<td>492.4133</td>
<td>-42.53782</td>
<td>604.3725</td>
</tr>
<tr>
<td>Jun 2012</td>
<td>495.96849</td>
<td>284.472532</td>
<td>707.4645</td>
<td>172.51333</td>
<td>819.4237</td>
</tr>
<tr>
<td>Jul 2012</td>
<td>338.17060</td>
<td>126.674639</td>
<td>549.6666</td>
<td>14.71544</td>
<td>661.6258</td>
</tr>
<tr>
<td>Aug 2012</td>
<td>451.35471</td>
<td>239.858746</td>
<td>662.8507</td>
<td>127.89955</td>
<td>774.8099</td>
</tr>
<tr>
<td>Sep 2012</td>
<td>274.66666</td>
<td>63.170702</td>
<td>486.1626</td>
<td>-48.78850</td>
<td>598.1218</td>
</tr>
<tr>
<td>Nov 2012</td>
<td>69.75377</td>
<td>-141.742190</td>
<td>281.2497</td>
<td>-253.70139</td>
<td>393.2089</td>
</tr>
<tr>
<td>Dec 2012</td>
<td>58.61241</td>
<td>-152.883554</td>
<td>270.1084</td>
<td>-264.84275</td>
<td>382.0676</td>
</tr>
<tr>
<td>Jan 2013</td>
<td>68.77455</td>
<td>-155.672306</td>
<td>293.2214</td>
<td>-274.48730</td>
<td>412.0364</td>
</tr>
<tr>
<td>Feb 2013</td>
<td>71.95333</td>
<td>-152.825128</td>
<td>296.5318</td>
<td>-271.50978</td>
<td>415.4165</td>
</tr>
<tr>
<td>Mar 2013</td>
<td>113.28621</td>
<td>-111.293627</td>
<td>337.8661</td>
<td>-230.17901</td>
<td>456.7514</td>
</tr>
<tr>
<td>Apr 2013</td>
<td>127.26544</td>
<td>-97.314414</td>
<td>351.8453</td>
<td>-216.19981</td>
<td>470.7307</td>
</tr>
<tr>
<td>May 2013</td>
<td>219.42298</td>
<td>-5.156875</td>
<td>444.0028</td>
<td>-124.04227</td>
<td>562.8982</td>
</tr>
<tr>
<td>Jun 2013</td>
<td>434.30126</td>
<td>209.721403</td>
<td>658.8811</td>
<td>90.83601</td>
<td>777.7665</td>
</tr>
<tr>
<td>Jul 2013</td>
<td>356.85515</td>
<td>132.275295</td>
<td>581.4350</td>
<td>12.38990</td>
<td>700.3204</td>
</tr>
<tr>
<td>Aug 2013</td>
<td>438.26285</td>
<td>213.683001</td>
<td>622.8427</td>
<td>94.79761</td>
<td>781.7281</td>
</tr>
<tr>
<td>Sep 2013</td>
<td>317.06493</td>
<td>92.485081</td>
<td>541.6448</td>
<td>-26.40031</td>
<td>660.5302</td>
</tr>
<tr>
<td>Oct 2013</td>
<td>161.79367</td>
<td>-62.786185</td>
<td>386.3735</td>
<td>-181.67158</td>
<td>505.2589</td>
</tr>
<tr>
<td>Nov 2013</td>
<td>76.76038</td>
<td>-147.819473</td>
<td>301.3402</td>
<td>-266.70487</td>
<td>420.2256</td>
</tr>
<tr>
<td>Dec 2013</td>
<td>72.78127</td>
<td>-151.798581</td>
<td>297.3611</td>
<td>-270.68397</td>
<td>416.2465</td>
</tr>
</tbody>
</table>

Figure 5: Forecast plot for Total Monthly Rainfall from 2012 to 2013
Using the SARIMA model both point and interval estimations of monthly total rainfall are shown in Table 4. The forecasting of twenty-four months from January 2012 to December 2013 is found in the table. In the year of 2012, the highest amount of rainfall forecasted as 495.97 mm in June with 95% Confidence Interval (CI) between 172.51 mm and 819.42 mm; on the other hand, in January, the lowest amount of rainfall forecasted as 46.28 mm. In the year of 2013, the highest amount of rain forecasted as 438.26 mm in August with 95% Confidence Interval (CI) between 94.80 mm and 781.73 mm; on the other hand, in January, the lowest amount of rainfall forecasted as 68.77 mm. In Figure 5, all the rainfall data of all the 516 months from January 1969 to December 2011 of Rangamati are plotted and then the forecasting trends of the next 24 months have also been visualized.

**CONCLUSION**

The data are fitted to the Seasonal ARIMA (1, 0, 0) (2, 0, 0)[12] model for rainfall of Rangamati, Bangladesh. For the stationarity of the data, Augmented Dickey-Fuller Test is conducted. Based on the p-value (0.01), the data has been found stationary. The best forecast model has been obtained by applying the concept of model selection criteria like AIC and BIC. The prediction values and its graphs have also been shown.

**REFERENCES**


