MODELING MONTHLY RAINFALL IN KATUNAYAKE REGION USING SEASONAL ARIMA MODEL

P.A.H.R.Ariyarathna[1], M.C.Alibuhtto[2]

[1]Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sri Lanka
[2]Department of Mathematical Sciences, Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sri Lanka
himashrashmika@gmail.com, mcabuhtto@seu.ac.lk

Abstract:

The accurate forecast of rainfall is much important as the rainfall is one of the factors which is bound to human beings in routine life. The prediction of rainfall on a seasonal time scale has been attempted by various research groups using different techniques. In the present study, a univariate time series seasonal autoregressive integrated moving average (SARIMA) model has been developed for monthly rainfall data from a period of January, 2001 to January, 2016 (181 observations) in Katunayake region, Sri Lanka. For the model selection it was used 157 observations while the rest 24 observations were used to validate the developed model. The Johnson transformation was used to transform observations in order to correct the non-normality of the residuals. Based on the results, the SARIMA (2,0,2) $(2,0,1)_{12}$ model was found to be most suitable for forecasting the mean rainfall. The Akaike Information Criterion (AIC), Schwarz Information Criterion (SIC) and Durbin Watson statistics were used to test the validity of the developed model in different stages. This model is appropriate to forecast the monthly rainfall for the future months to assist decision and policy makers to establish priorities for water demand, storage and disaster management.

Keywords: Akaike Information Criterion, Johnson transformation, Rainfall, SARIMA, Schwarz Information Criterion

Introduction

Sri Lanka can be considered as a tropical country as it is located near the equator. Usually Western and Southern region of country receives higher rainfall from May to September and Northern and Eastern region receives high rainfall from October to January. But Western and Southern region receives significantly high rainfall than the other region. The mean annual rainfall in Sri Lanka varies from 900mm to 5000mm.

Many researchers have been made in the recent past to model and forecast rainfall using various time series techniques proving to be the most common (Brath et al., 2002; Rabenja et al., 2009; Helman, 2011; Mahsin et al., 2012). In time series analysis it is assumed that the series consist of a systematic pattern and random noise which usually makes the pattern difficult to identify.

The purpose of this article is to identify appropriate Box-Jenkins time series model for forecasting monthly rainfall in Katunayake region, Sri Lanka, using the observations of monthly rain fall data from the period January 2001 to January 2016, representing 181 observations. These data were obtained from the Metrological Department, Colombo. The first 157 observations were used for developing time series forecast model and rest of 24 observations were used to validate the developed model.

Methodology

ACF and PACF

Autocorrelation function and partial autocorrelation function are a type of graphs which containing correlations of different time lags. ACF and PACF can be used to identify the behaviour of the series whether stationary or not and to identify the number of components in ARMA model. The exponentially decaying spikes in ACF and PACF indicate the stationary series. The number of significant spikes in ACF indicate the number of MA terms in the model and the number of significant spikes in PACF indicate the number of AR terms in the model.

The autocorrelation function,

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{n}(x_t \cdot x_{t-k}) - (n-k-1)\bar{x}_t^2}{\sum_{t=k+1}^{n} x_t^2 - (n-k-1)\bar{x}_t^2} \quad (2.1)$$

The partial autocorrelation function,

$$\phi_k = \frac{|\rho_k^*|}{|\rho_k|} \quad (2.2)$$

Where; $$\rho_k = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_k \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_2 & \rho_1 & 1 & \ldots & \rho_{k-2} \end{bmatrix} \quad (2.3)$$

$$\rho_k^* = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_2 \\ \rho_2 & \rho_1 & 1 & \ldots & \rho_3 \end{bmatrix} \quad (2.4)$$

Autoregressive moving average model

If current x_t value can be expressed as a linear combination of both past p number of x_t values and past q number of e_t values, then it is an ARMA time series model of p and q. This model is an addition and extension of both AR and MA models. The ARMA model can be identified from the number of significant spikes of autocorrelation and partial autocorrelation graphs. The ARMA(p,q) model can be written as,

$$x_t = \alpha_0 + \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q} + e_t \quad (2.5)$$

The characteristic equation of autoregressive moving average model can be written as,

$$\phi_p(B)x_t = \alpha_0 + \theta_q(B)e_t \quad (2.6)$$

Where; $e_t \sim WN(0, \sigma^2)$

Stationary Time Series

Stationary time series is a stochastic process whose joint probability distribution does not change with the time. There are several conditions that must be satisfy for the stationarity of a series. They are the mean and the variance of the series should not be time dependent and the covariance of time series and lagged time series should depend only on the difference of the time. There are several methods available to transform a non-stationary series in to stationary time series. If the series is non-stationary due to non-constant mean, then the lag differencing can be used. If the series is non-stationary due to non-constant

variance, then the log transformation can be used. The most applicable method to stabilize the variance is Box-Cox transformation.

There are several tests such as Augmented Dickey Fuller (ADF) test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test that can be used to check whether the series is stationary or not.

Autoregressive iIntegrated moving average model

The generalization of ARMA model can be identified as ARIMA model. The ARIMA model can be used in the presence of non-stationary series. If current dx_t value can be expressed as a linear combination of both past p number of dx_t values and past q number ofe$_t$ values where dx_t is 1st or 2nd order difference of x_t due to non-stationary, then it is an autoregressive integrated moving average time series model of p, d and q.
The ARIMA(p,d,q) model can be expressed as,

$$dx_t = \alpha_0 + \alpha_1 dx_{t-1} + \cdots + \alpha_p dx_{t-p} + \beta_1 e_{t-1} + \cdots + \beta_q e_{t-q} + e_t \quad (2.7)$$

The characteristic equation of ARIMA model can be expressed as,

$$\phi_p(B)\nabla^d x_t = \alpha_0 + \theta_q(B)e_t \quad (2.8)$$

Seasonal Autoregressive Integrated Moving Average Model

The data series which contains seasonal component has to be modelled as a Seasonal ARIMA model in order to fit a Box-Jenkins model. Seasonal ARIMA model is a combination of non-seasonal ARIMA model and seasonal ARIMA model. If the data series differenced into two parts such as seasonally adjusted component and seasonal component due to the existence of seasonality, then the ARIMA model used in both components can be identified as SARIMA model. The SARIMA model denoted by SARIMA(p,d,q)(P,D,Q)s, where first bracket and second bracket denote the seasonally adjusted and seasonal factor series respectively. The SARIMA model can be written as;

$$dx_t = \alpha_0 + \alpha_1 dx_{t-1} + \cdots + \alpha_p dx_{t-p} + \beta_1 Dx_{t-1s} + \cdots + \beta_P Dx_{t-Ps} + \gamma_1 e_{t-1} + \cdots + \gamma_q e_{t-q} + \delta_1 e_{t-1s} + \cdots + \delta_Q e_{t-Qs} + e_t \quad (2.9)$$

The characteristic equation of SARIMA model can be expressed as,

$$\phi_p(B) \cdot \Phi_P\left(B^{(S)}\right) \cdot (1-B)^d \cdot \left(1-B^{(S)}\right)^D x_t = \alpha_0 + \theta_q(B) \cdot \Theta_Q\left(B^{(S)}\right) \cdot e_t \quad (2.10)$$

Results and Discussion

Preliminary Analysis

The Preliminary analysis was done for the complete data set which was from 2001 January to 2016 January. The Descriptive measures are shown in Table 1.

Table 1: Descriptive measures of rainfall data.

Variable	Mean	SD	Minimum	Maximum
Values	169.8	146.9	0.0	790.1

According to the table 1 the minimum and the maximum rainfall received for the Katunayake region from 2001 to 2016 were 0.0mm and 790.1mm respectively. The mean

rainfall received for the period was 169.8mm and the standard deviation for the rainfall was 146.9.The time series plot and the seasonal graph and were obtained.

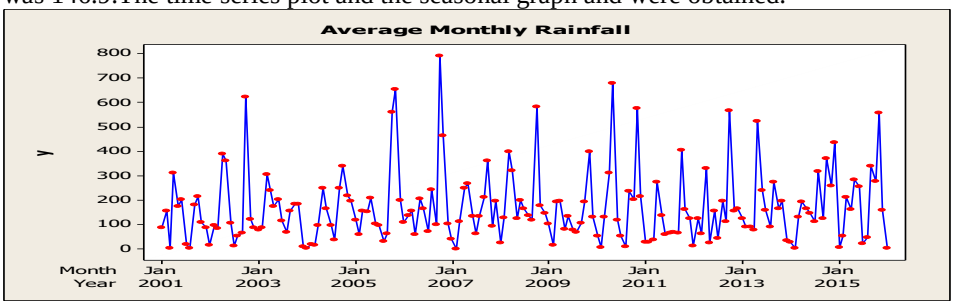

Figure 1: Time series plot of the rainfall data

From the Figure 1 it can be seen that the data set is highly fluctuating and there is no visible trend.

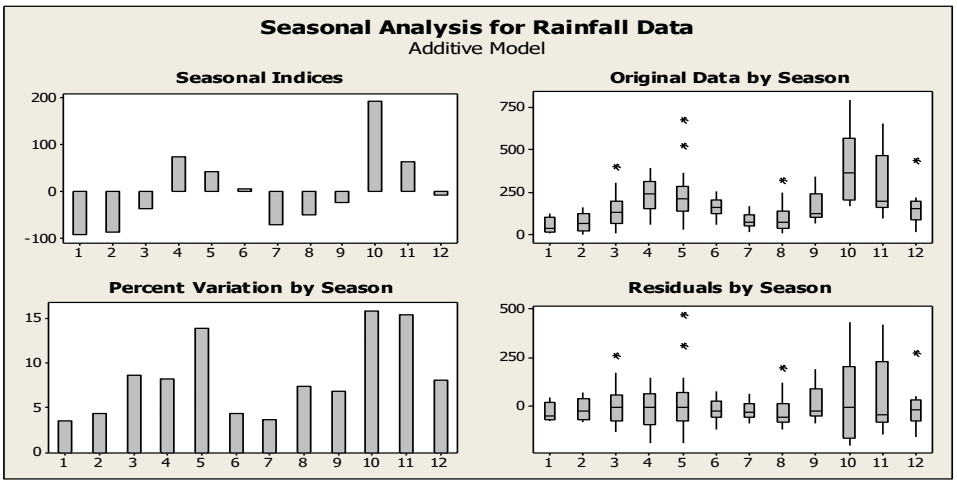

Figure 2: Seasonal graph of the rainfall data

From the seasonal indices graph in figure 2 it can be seen that there exist a seasonality. Some outliers of the observations can be seen in figure 2.The descriptive measures of monthly wise observations were obtained and given in table 2.

Table 2: Descriptive measures of monthly wise data.

Month	Mean	SD	Minimum	Maximum
January	51.5	44.1	2.0	122.9
February	66.3	53.7	0.0	153.9
March	138.7	106.7	1.7	397.8
April	223.5	100.8	58.1	389.0
May	243.2	169.9	24.3	680.1
June	158.4	53.9	57.5	254.6
July	77.2	46.4	13.1	163.0
August	99.6	90.4	3.4	317.7
September	159.4	85.1	61.1	340.4
October	388.3	194.2	164.6	790.1

November	287.6	189.2	92.7	652.6
December	152.2	99.1	9.1	436.3

According to the table 2 the maximum and the minimum rainfall received for the Katunayake region were in October and January respectively.

ARIMA (Box-Jenkins) Modelling

The Box-Jenkins modelling was done with 157 data which was 2001 January to 2014 January. Rest 24 observations were kept for measuring the accuracy of the forecasted model which was from 2014 February to 2016 January.

The data series was checked for the seasonality by using Kruskal Wallis test and the seasonality was extracted by using Census X-12 seasonal adjustment method and the line chart of both seasonally adjusted and seasonal factor series were obtained.

Table 3: Kruskal Wallis test

Kruskal Wallis Statistic	Degrees of freedom	Probability value
73.9850	11	0.000

Table4: Moving Seasonality test

	Sum of Sq.	Degrees of freedom	Mean Square	F-Value
Between	104254.8080	12	8687.9006	1.042
Error	1100726.2624	132	8338.8353	

According to the Kruskal Wallis test it indicated that there exist a significant seasonality and the moving seasonality test indicated that there is no any moving seasonality exist.

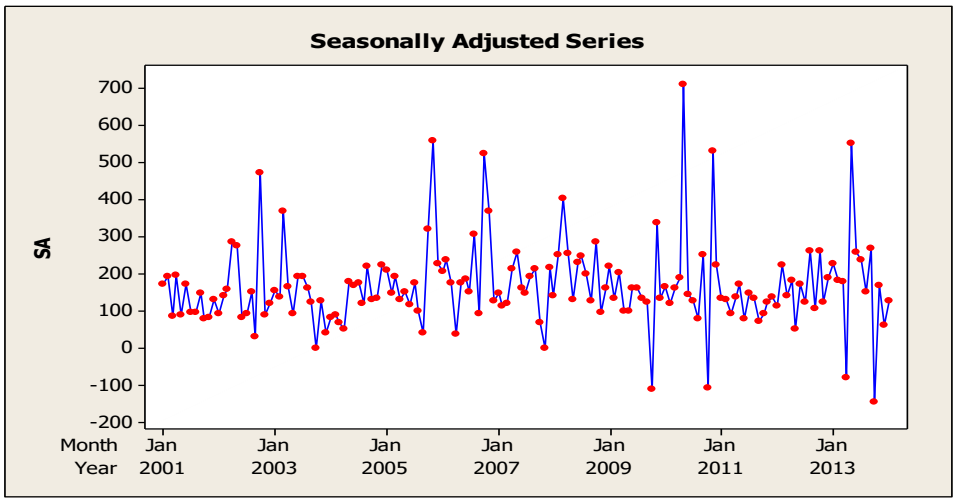

Figure 3: Time series plot of seasonally adjusted series

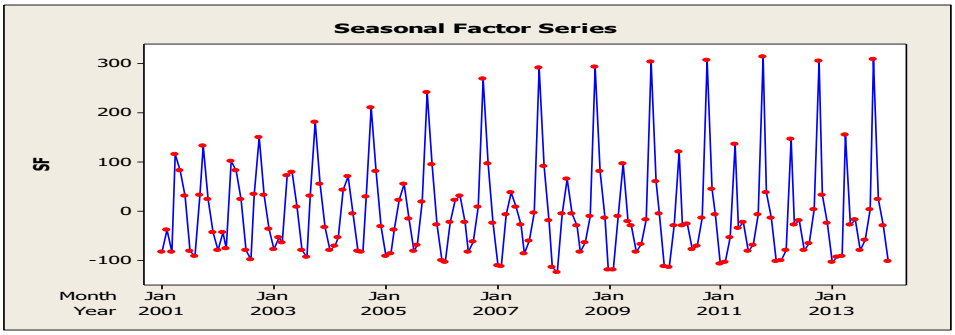

Figure 4: Time series plot of seasonal factor series

The seasonally adjusted series does not indicate any trend or any other pattern. The data series fluctuating around the mean.The seasonal factor series indicates a seasonality. It shows an increasing pattern over the years.Then the ACF and PACF of Seasonally adjusted and seasonally facture series were obtained.

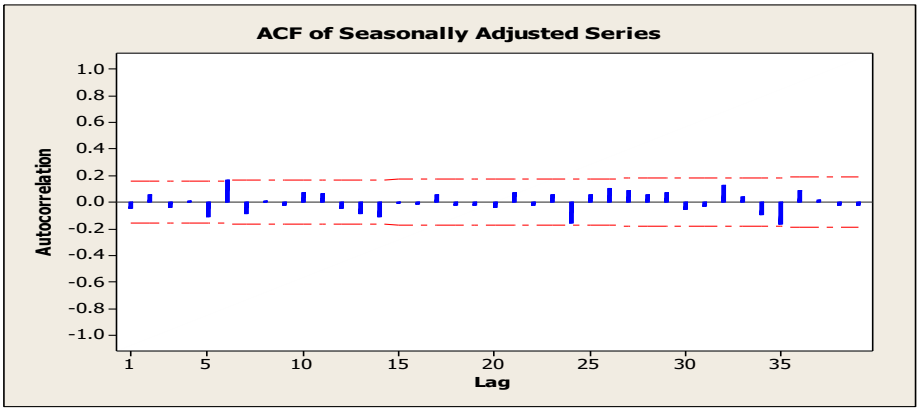

Figure 5: ACF plot of seasonally adjusted series

ACF of seasonally adjusted series does not indicate any significant spikes.

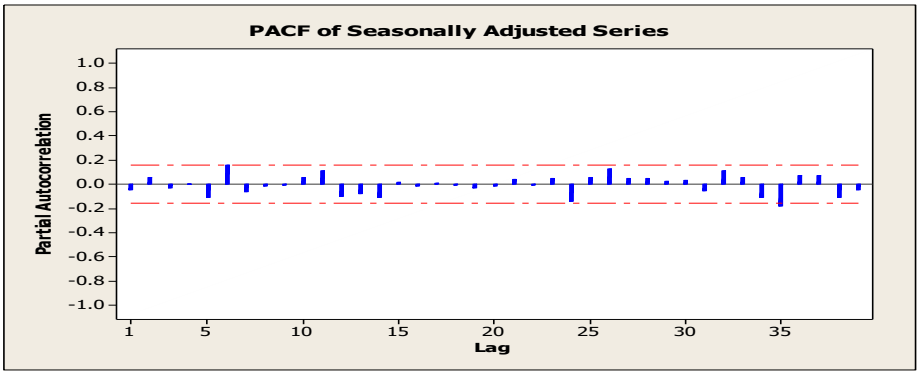

Figure 6: PACF plot of seasonally adjusted series

PACF of seasonally adjusted series does not indicate any significant spikes.

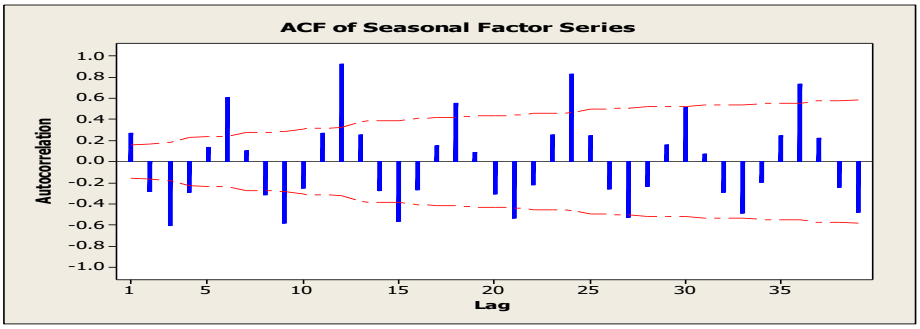

Figure 7: ACF plot of seasonal factor series

The ACF of Seasonal factor series indicates some significant spikes which shows a seasonal pattern. Therefore there are seasonal moving average terms in the model.

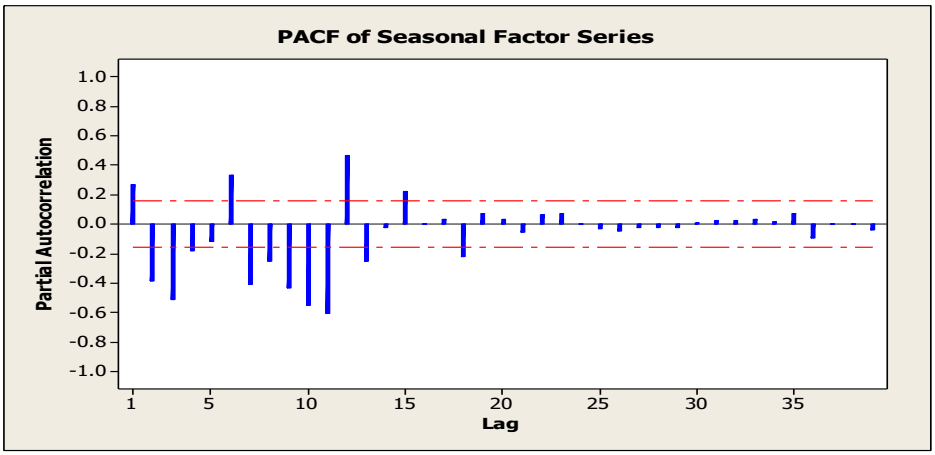

Figure 8: PACF plot of seasonal factor series

The PACF of Seasonal factor series indicates some significant spikes. Therefore there are seasonal autoregressive terms in the model.

The model selection criterion was conducted to select the best model from tentative models. The log likelihood, AIC, SIC and DW values used to select the best model. The minimum values of log likelihood, AIC, SIC and the DW value close to 2 indicated the best model. Therefore from the table 5 the SARIMA$(0,0,0)(0,0,2)_{12}$ was selected as the best model.

Table 5: Parameter estimation

Model	Log likelihood	AIC	SIC	DW
SARIMA(0,0,0)(0,0,0)	-945.8934	12.0623	12.0818	1.4545
SARIMA(0,0,0)(0,0,1)	-937.2729	11.9780	12.0364	1.9625
SARIMA(0,0,0)(0,0,2)	-919.9508	11.7700	11.8479	1.5404
SARIMA(0,0,0)(1,0,0)	-940.0465	12.0133	12.0717	1.7828
SARIMA(0,0,0)(1,0,1)	-928.4793	11.8787	11.9565	1.2948
SARIMA(0,0,0)(1,0,2)	-912.2277	11.6844	11.7817	1.8027

Therefore the fitted SARIMA model is,
$y_t = 164.84 - 0.4071\, e_{t-12} - 0.0648\, e_{t-24} + e_t$

Model Diagnostics
Then the model diagnostics was conducted for the fitted model. The normal probability plot for the residuals were obtained.

The normal probability plot indicates that the residuals are highly not normally distributed. The residuals of the fitted SARIMA model was not-normally due to the high not-normality of the data. Therefore the data should be transformed.

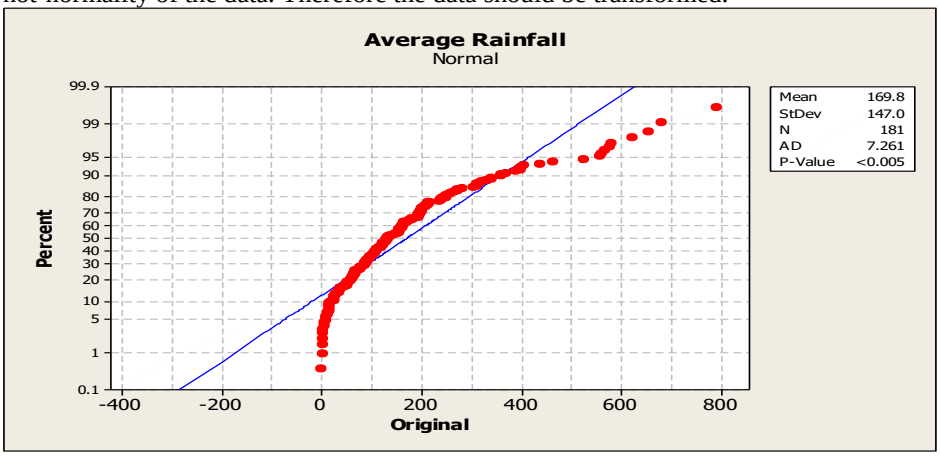

Figure .9: Normal probability plot of the data

Johnson Transformation
Therefore the data transformed by using Johnson transformation. The function used for the Johnson transformation was$\ln(x_t + 62.7918) * 1.81931 - 9.57815$. The Box-Cox transformation cannot be used in this analysis due to existence of non-positive data.

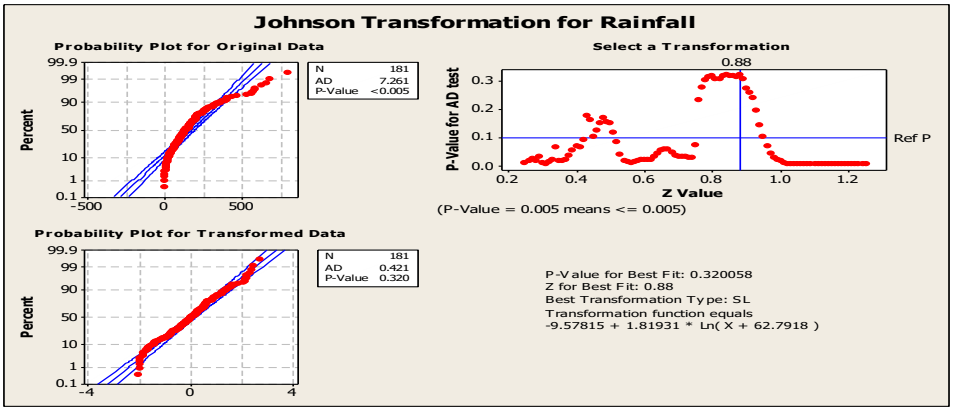

Figure 10: Johnson transformation of the data

These normalized transformed data has used for further analysis.

ARIMA (Box-Jenkins) Modelling

The seasonality of the data series was checked by using Kruskal Wallis test and F test. The moving seasonality was checked by using moving seasonality test.

Table 6: Kruskal Wallis test

Kruskal Wallis Statistic	Deg. of free.	Prob.
79.8947	11	0.000

Table 7: Moving Seasonality test

	Sum of Sq.	Deg. of free.	Mean Square	F-Value
Between	2.5728	12	0.214397	0.704
Error	40.1742	132	0.304350	

The Kruskal-Wallis test has confirmed the existence of significant seasonality in 99% confidence level (P=0.000). The moving seasonality test confirmed that there is no evidence for the moving seasonality in 95% confidence level. Therefore the seasonality of the data has been extracted by using Census X-12 seasonal adjustment method. The line chart of the seasonally adjusted series and seasonal factor series were obtained to identify the behaviour of the both series.

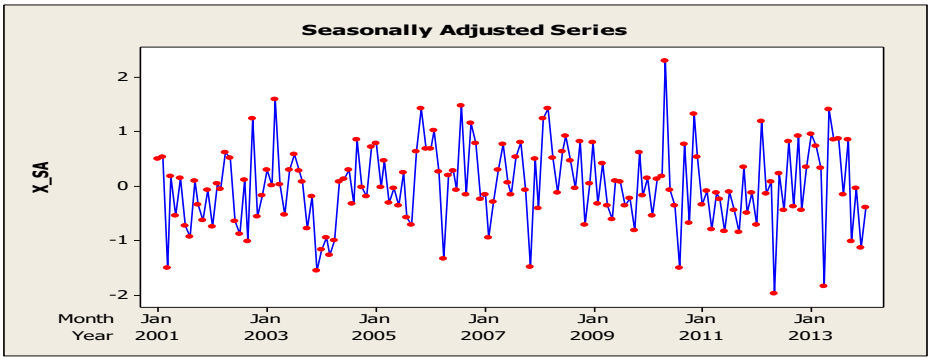

Figure 11: Time series plot of seasonally adjusted series

The seasonally adjusted series does not indicate any trend or unusual fluctuation around the mean.

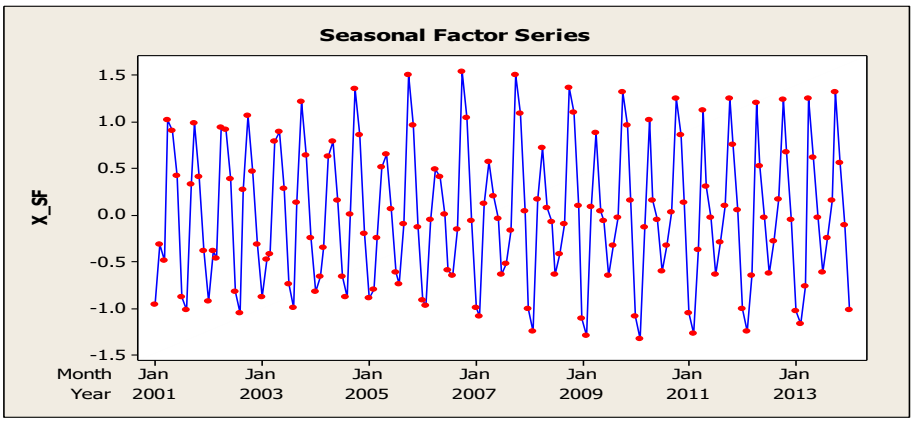

Figure 12: Time series plot of seasonal factor series

The seasonal factor series indicate that there is a seasonality in the data series.
The stationarity of the both series has checked by using ADF unit root test. The unit root test for the seasonally adjusted series (P=0.0000) and seasonal factor series (P=0.0000) has confirmed the both series were stationary in 95% confidence level.
The ACF and PACF were obtained to identify the number of AR and MA terms in both series.

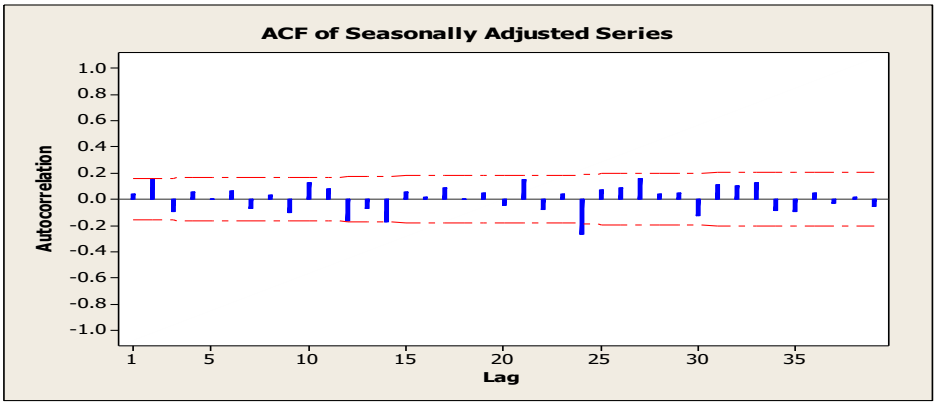

Figure 13: ACF plot of seasonally adjusted series

The ACF of seasonally adjusted series indicate one significant spike near the 25th lag.

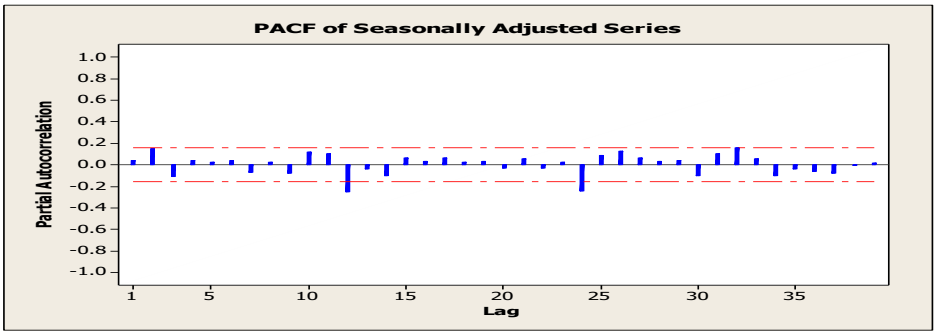

Figure 14: PACF plot of seasonally adjusted series

The PACF of seasonally adjusted series indicate two significant spikes. But both spikes lies near 10th and 25th lags.

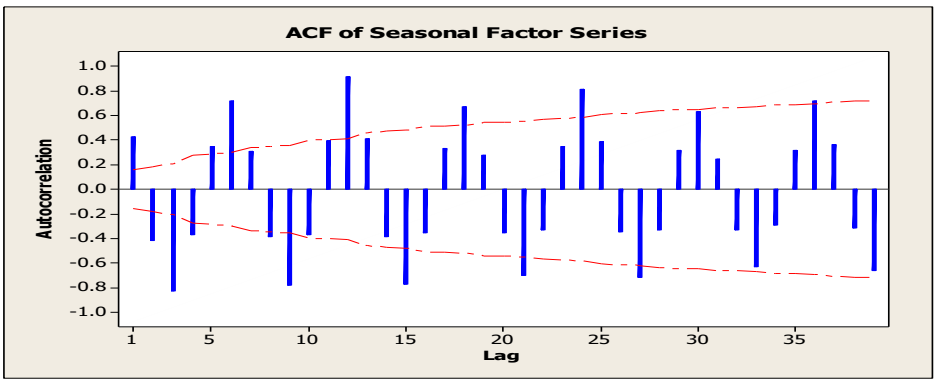

Figure .15: ACF plot of seasonal factor series

The ACF of seasonal factor series indicate a series of significant lags. The lags shows a seasonal pattern.The PACF of seasonal factor series indicate a series of significant lags. The model selection criterion was used to identify the best model.Table 8 used to identify the number of components for the seasonally adjusted series and seasonal factor series. Log likelihood, AIC, SIC and DW values used for the model selection. The best nine models were shown in table 8.

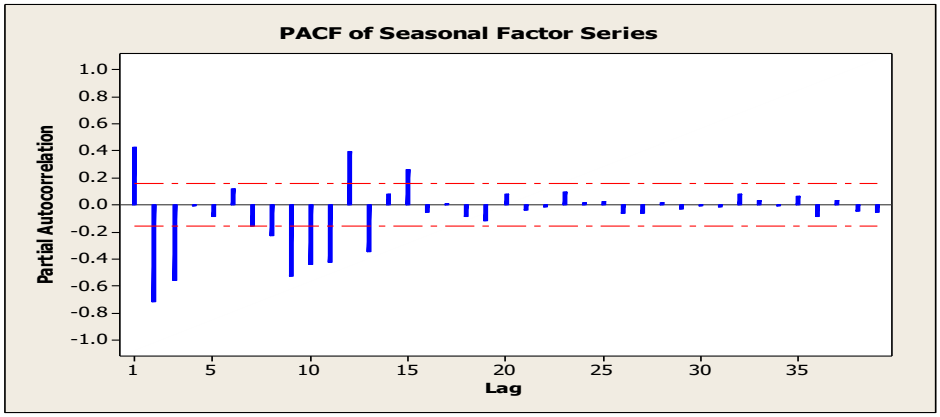

Figure 16: PACF plot of seasonal factor series

Table 8: Parameter estimation.

Model	Log likelihood	AIC	SIC	DW
SARIMA(2,0,2)(0,0,1)	-191.7861	2.5323	2.6685	1.9752
SARIMA(2,0,2)(0,0,2)	-190.0252	2.5226	2.6783	2.0235
SARIMA(2,0,2)(1,0,0)	-191.7170	2.5314	2.6676	2.0055
SARIMA(2,0,2)(1,0,1)	-191.1549	2.5370	2.6927	1.7300
SARIMA(2,0,2)(1,0,2)	-188.8333	2.520	2.6953	1.9863
SARIMA(2,0,2)(2,0,0)	-190.3678	2.5269	2.6827	1.9517
SARIMA(2,0,2)(2,0,1)	-188.7238	2.5187	2.5899	1.9790
SARIMA(2,0,2)(2,0,2)	-188.3957	2.5273	2.7219	1.9670

The best model for the data series can be considered as SARIMA(2,0,2)$(2,0,1)_{12}$.

Table 9: Parameter estimation of SARIMA(2,0,2)$(2,0,1)_{12}$

Variable	Coefficient	Standard Error	t-Statistics	P-Value
C	0.009582	0.076230	0.125702	0.9001
AR(1)	-0.603442	0.233899	-2.579885	0.0109
AR(2)	0.201462	0.097446	2.067414	0.0404
SAR(1)	1.000690	0.005433	184.1849	0.0000
SAR(2)	-0.999698	0.002538	-393.9135	0.0000
MA(1)	-1.018013	12.46394	-0.081677	0.9350
MA(2)	0.999921	24.48220	0.040843	0.9675
SMA(1)	0.699974	0.220714	3.171407	0.0018

Model Diagnostics

The residual plot, the normal probability plot and the histogram of residuals were obtained to check the behaviour of the residuals of the fitted model.Table 8 used to identify the number of components for the seasonally adjusted series and seasonal factor series. Log likelihood, AIC, SIC and DW values used for the model selection. The best nine models were shown in table 8.

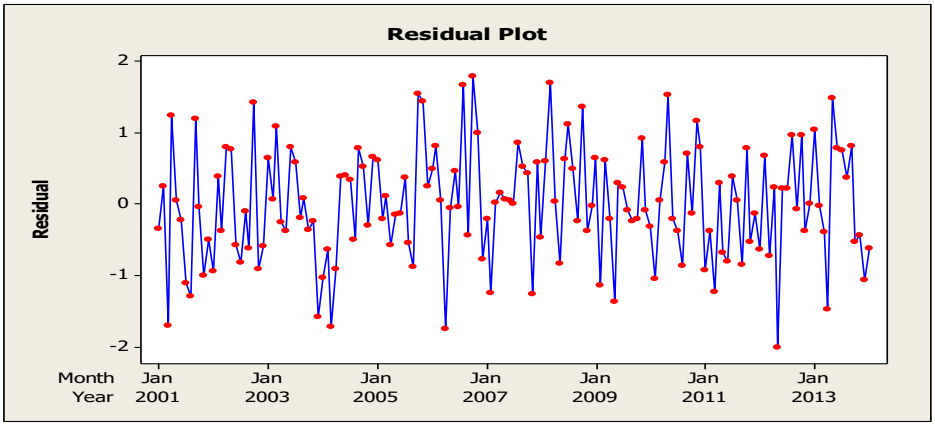

Figure 17: Time series plot of residual

The residuals plot indicated that there is no high deviations and any pattern of residuals. The maximum and minimum deviations of residuals lies between positive and negative two.

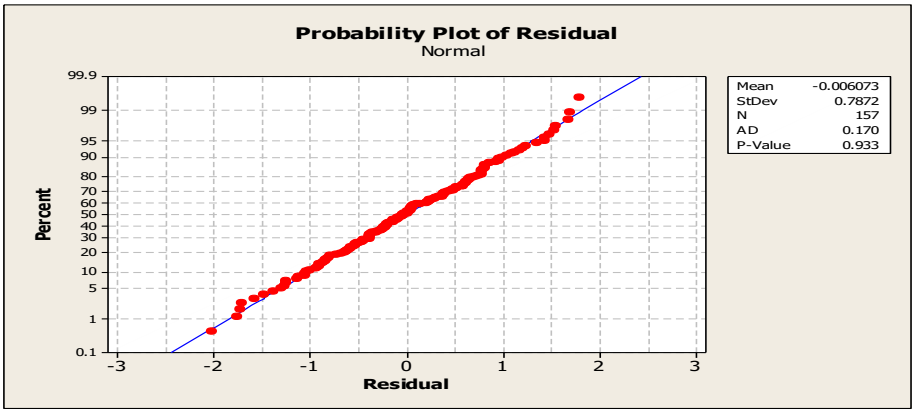

Figure 18: Normal probability plot of residuals

To confirm there was no any serial correlation, the residual ACF and residual PACF were obtained.

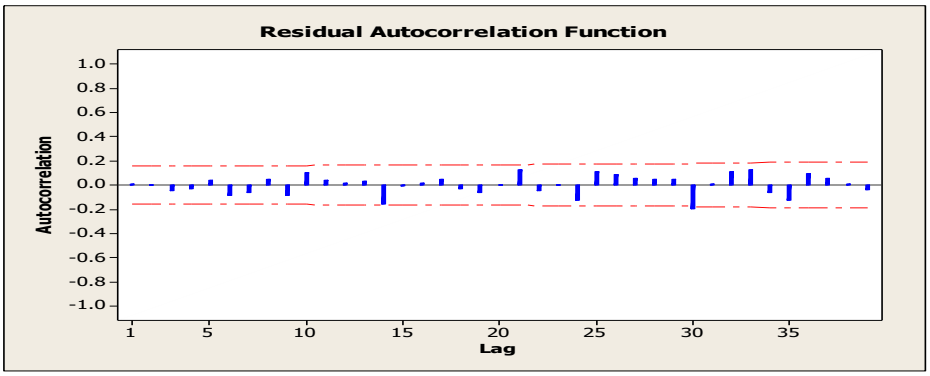

Figure 19: ACF plot of residual

The residual autocorrelation function does not indicated any significant spikes. Therefore there is no any residual serial correlation.

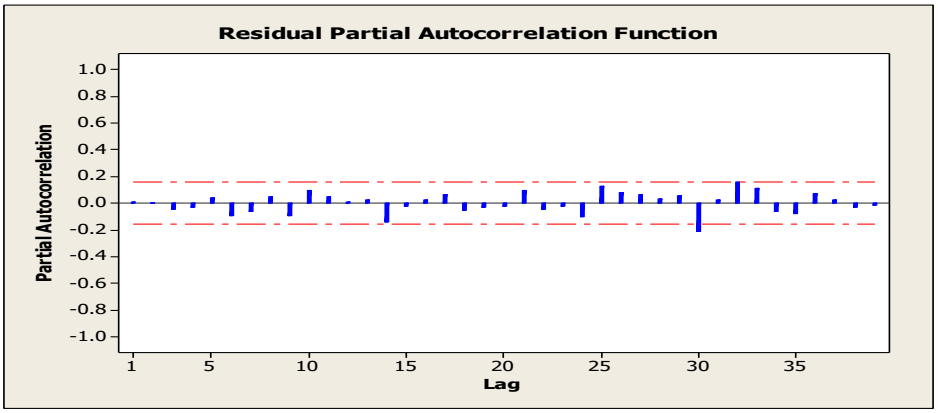

Figure 20: PACF plot of residual

The residual partial autocorrelation function does not indicated any significant spikes. Therefore there is no any residual serial correlation.The ARCH LM test was conducted to observe the heteroscedasticity of the model. According to the test there was no heteroscedasticity in 95% confidence level (P=0.3781).

Forecasting and Validating the Fitted Model

From 2014 February to 2016 January 24 values were forecasted by using the fitted model to check the prediction accuracy of the model.

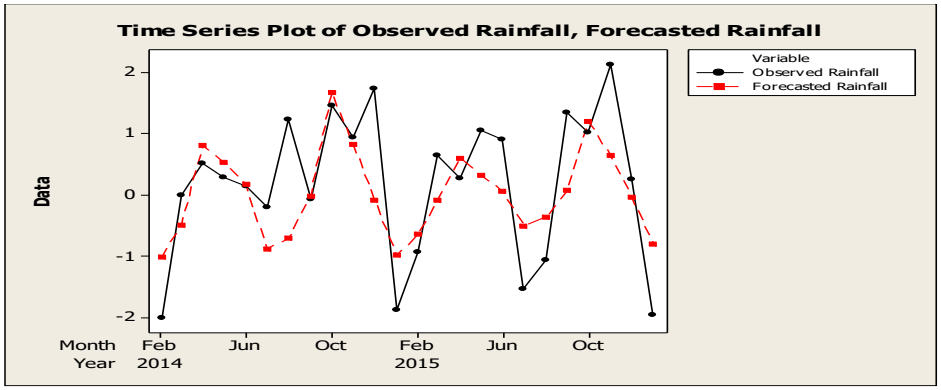

Figure 21: Plot of observed rainfall and predicted rainfall

The forecasted and observed rainfall observations are not much deviated from each other. The t-test showed that there was no significant difference between predicted and observed rainfall in 95% confidence level (P=0.368).

Conclusion

Time series analysis is an important technique in modelling and forecasting rainfall data. In this study we used Seasonal ARIMA model to forecast monthly rainfall obtained rainfalldata Katunayake region.

The best fitting model for the rainfall was identified as SARIMA$(2,0,2)(2,0,1)_{12}$. Model diagnostic checking presented that developed model should have significant result. Therefore, this model could help to determine possible future strategy in the respective field for the Katunayake region.

References

Akpanta, A. C., Okorie, I. E., &Okoye, N. N. (2015). SARIMA Modelling of the Frequency of Monthly Rainfall in Umuahia, Abia State of Nigeria.*American Journal of Mathematics and Statistics*, *5*(2), 82-87.

Ampaw, E. M., Akuffo, B., Larbi, S. O., &Lartey, S. (2013). Time Series Modelling of Rainfall in New Juaben Municipality of the Eastern Region of Ghana. *International Journal of Business and Social Science*, *4*(8).

Brath, A., Montanari, A., &Toth, E. (2002). Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models. *Hydrology and Earth System Sciences Discussions*, *6*(4), 627-639.

Chang, X., Gao, M., Wang, Y., &Hou, X. (2012). Seasonal autoregressive integrated moving average model for precipitation time series. *Journal of Mathematics & Statistics*, *8*(4).

Chonge, M., Nyongesa, K., Mulati, O., Makokha, L., &Tireito, L. (2015). A Time Series Model of Rainfall Pattern of Uasin Gishu Country. IOSR Journal of Mathematics, 11(5), 77-84.

Dong, Y. (2012). *ARMA and GARCH-type modelling electricity prices*(Doctoral dissertation, Masters thesis submitted to Chalmers University of Technology, Sweden).

Eni, D., & Adeyeye, F. J. (2015). Seasonal ARIMA Modeling and Forecasting of Rainfall in Warri Town, Nigeria. *Journal of Geoscience and Environment Protection*, *3*(06), 91.

Etuk, E. H., & Mohamed, T. M. (2014). Time Series Analysis of Monthly Rainfall data for the Gadaref rainfall station, Sudan, by SARIMA Methods.*International Journal of Scientific Research in Knowledge*, *2*(7), 320.

Helman, K. (2011). SARIMA models for temperature and precipitation time series in the Czech Republic for the period 1961–2008. *J. Appl. Mathem*, (3), 4.

Mahsin, M. (2011). Modeling rainfall in Dhaka division of Bangladesh using time series analysis. *Journal of Mathematical Modelling and Application*,*1*(5), 67-73.

Modarres, R., & Ouarda, T. B. M. J. (2013). Generalized autoregressive conditional heteroscedasticity modelling of hydrologic time series.*Hydrological Processes*, *27*(22), 3174-3191.

Nirmala, M., & Sundaram, S. M. (2010). A Seasonal Arima Model for forecasting monthly rainfall in Tamilnadu. *National Journal on Advances in Building Sciences and Mechanics*, *1*(2), 43-47.

Rabenja, A. T., Ratiarison, A., &Rabeharisoa, J. M. (2009). Forecasting of the Rainfall and the Discharge of the Namorona River in Vohiparara and FFT Analyses of These Data. In *Proceedings of 4th International Conference in High-Energy Physics* (pp. 1-12).

Uba, E.S., & Bakari, H.R. (2015). An Application of Time Series Analysis in Modeling Monthly Rainfall Data for Maiduguri, North Eastern Nigeria. Mathematical Theory and Modeling, 5(11).