Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings XX (2016) XXX–XXX

materialstoday: PROCEEDINGS

www.materialstoday.com/proceedings

PMME 2016

Outlier Detection and Missing Value in Seasonal ARIMA Model Using Rainfall Data☆

Arumugam. P[a], Saranya. R[b*]

[a]*Associate Professor, Department of Statistics, Manonmanium Sundarnar University, Tirunelveli-627012,Tamilnadu, India*
[b]*Research Scholar,Department of Statistics, Manonmanium Sundarnar University, Tirunelveli-627012,Tamilnadu, India*

Abstract

Forecasting the trend of rainfall is a difficult task in meteorology and environmental sciences. Statistical approaches from time series analysis provide another way for rainfall prediction. The ARIMA model incorporating seasonal characteristics, which is devoted to as seasonal ARIMA model was presented. The time series data are the monthly rainfall data from 2006 to 2016. The model was denoted as Seasonal ARIMA $(1, 1, 1)\ (0, 1, 1)_{12}$ in this study. A serious problem in analyzing rainfall data is what to do when missing or extreme values occur, perhaps as a result of a breakdown in automatic counting equipment. We can analyze the stability and the correlation of the time series. The aim of this paper is to attempt look at ways of explaining this problem by using the residuals from a fitted SARIMA model. The most successful method in finding outliers and unique them from other events, being less expensive than case deletion. In our result, the model fitted the data well and the stochastic seasonal variation was successfully model. Seasonal ARIMA model was a proper method for modeling and forecasting the time series of monthly rainfall data.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Seasonal ARIMA Model, Missing value, Innovative Outlier, Seasonal Additive Outlier, Forecasting

* Corresponding author. Tel.: +91-9942253610;
E-mail address: mahasaranya18@gmail.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

1. Introduction

Water is important in everyday human life. Water is the most important part of our human body and not including which, one cannot survive. Drinking 3 to 4 liters of water daily, makes healthy body and healthy life. Thus need of water not only fulfill human body, but also makes use of the various purposes for all human beings. Water is usually shared in many places like rivers, lakes, ponds, wells and etc, from the stable resources on earth. India is one of the fastest rising countries in the world whose major occupation is agriculture. For good agriculture yield, rainfall is considered as a key factor. Rain flows through North-East monsoon in Tamilnadu and the South-West monsoon in India. Commonly, the rainfall occurring during to these monsoons provides a sufficient and energetic from of agriculture. Rain is the primary source of fresh water for most areas of the world, providing suitable conditions for several ecosystems, as well as water for hydroelectric power paints and crop irrigation. Rainfall is measured through the use of rainfall. Rainfall amounts are estimated active weather by radar and inactively by weather satellites. Rain is a water rainfall, as opposed to other kinds of precipitation such as snow, hail and sleet. Rain requires the presence of a thick layer of the Earth's surface. Rainfall forms via collision with other rain drops or ice crystals within a cloud. Rain drops sort in size from oblate, pancake-like shapes for larger drops, two small spheres for smaller drops. Convective rain or showery rainfall, from convective clouds for example, cumulonimbus or cumulus congests. It falls as showers with quickly changing intensity. Convective rainfall falls over a certain area for a relatively short time, as convective gases have limited horizontal extent .Most rainfall in the tropic appears to be convective however, it's been suggested that stratiform rainfall also occurs. In mid-latitudes, convective rainfall is intermittent and often associated with bar clinic boundaries such as cold fronts, squall lines, and warm fronts.

1.1 Handling Missing Values and Mean substitution

Data cleaning is often the first step that data scientists and predictors take to ensure the statistical Modelling is supported by good data. Missing data is one of the big issues in everywhere. Missing data are a part of very nearly all research, and we all have to make a decision how to deal with it from time to time. There are a number of different methods of dealing with missing data. Kohn and Ansley (1986) have defined likelihood for an ARIMA model under any shape of missing data and to define predictors and interpolators for the missing data.

Missing value estimation methods for substituting missing values by Series mean replaces missing values with the mean for the entire series.

1.2 Outliers

Outlier is an observation, which so many moves away from other observations as to arouse suspicions that it was generated by a different tool by Hawkins (1980). In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered irregular. Before irregular observations can be singled out, it is necessary to illustrate normal observations. Fundamentally outlier tests represent some simple statement, when we have a technical error do not said that point is incorrect. Mohammad, et.al. (2010) have considered Additive and Innovative outliers in time series data.

1.2.1 Innovative Outlier (IO)

An innovation outlier is the type of outlier that affects the subsequent observation starting from its position in other words that occurs as a result of ordinary randomness. The model, defined as "randomness outlier" in the literature, is given by,

$$y_t = \frac{\theta(B)}{\phi(B)}(e_t + \delta x_t)$$

Where y_t is the observed value, δ is the magnitude of outlier and $x_t = \begin{cases} 1 & t = T \\ 0 & \text{Otherwise} \end{cases}$

1.2.2 Seasonal Additive Outlier (SAO)

An outlier that have emotional impact a particular observation and all subsequent observations separated from it by one or more seasonal period. All such observations are affected equally. A seasonal additive outlier might occur if, beginning in a certain year, sales are higher every January.

The aim of this paper, to attempt look at ways of explaining this problem by using the residuals from a fitted seasonal ARIMA model and the most successful method in finding outliers and unique them from other events, being less expensive than case deletion.

2. Review of Related Work

There are two kinds of models are considered for outliers and their effects in time series. Likelihood ratio and estimated likelihood ratio criteria are derived from these models and the power functions are compared with that of the method generally applied in the past Fox (1972), Abraham (1979) & (1989). The effects of outliers using the influence function for the estimation of the autocorrelation function (ACF) of a time series by Cher nick (1982), have investigated. Outliers in time series can be regarded as being generated by dynamic intervention models at unidentified time points with two special cases, innovation outlier (IO) and additive outlier (AO). The likelihood ratio criteria for analysis the existence of outliers of both types, and the criteria for distinguishing between them are derived. An iterative method for detecting IO and SAO in practice and for estimating the time series parameters in ARIMA models in the incidence of outliers was suggested by Chang (1988).

There are several authors established the difficulty that traditional outlier detection methods Balke (1993), Tight(1993), Ljung(1993) have in detecting level shifts in time series. Initializing the outlier or level-shift search with an estimated autoregressive moving average method lowers the power of the level-shift detection statistics. Furthermore, the rule working by these methods for distinguishing between level shifts and innovation outliers does not work well in the incidence of level shifts. A simple modification to Tsay's procedure is proposed that improves the ability to correctly identify level shifts. This alteration is relatively easy to implement and appears to be quite effective in practice.

Classifying outliers in autoregressive models were used to identify the innovative and additive outliers. Investigate the impact of outliers on parameter estimates and model selection and find that Innovative Outliers (IO) had less effect on limitation estimators than Seasonal Additive Outliers (SAO). Monte Carlo simulation studies were performed well by McQuarrie and Tsai (2003), once large outliers occurred. Seasonal Additive and innovative outliers detected in time series Generalized Autoregressive Conditional Heteroskedasticity (GARCH) methods developed the procedure for the presence of outliers in statistical test. The resulting outlier statistic has been used in GARCH (1, 1) model by Mohammad (2010).

3. Materials and Methods

Monthly observations of rainfall data for the period January 2006 to April 2016 were obtained from iari.res.in. In this study we are using Modelling in the Presence of Outliers, Seasonal ARIMA Model, Model Identification procedure, Parameters Estimation, Diagnostic Checking, and Forecasting with Seasonal ARIMA Model.

3.1. Modelling in the Presence of Outliers

In practice, Modelling and forecasting time series data in the presence of outliers are a testing problem for several reasons. Outliers and structure changes are commonly encountered in time series data analysis by Tsay (1988). The incidence of outliers be able to adversely affect the model identification and estimation steps. Their incidences close to the end of the observation period can have a serious impact on the forecasting presentation of the model. In some cases, level shifts are related with changes in the mechanism that drives the observation process, and separate models might be proper to different sections of the data. In view of all these difficulties, diagnostic tools for instance outlier detection and residual analysis are essential in any modelling process.

3.2 Seasonal ARIMA Model

The general form of seasonal ARIMA (p, d, q) (P, D, Q)s ,
Where, *p* -is the order of the autoregressive process
d -is the order of the differencing
q -is the order of the moving-average process
P -is the order of the seasonal autoregressive process
D -is the order of the seasonal differencing
Q -is the order of the seasonal moving-average process.
Mathematically the pure Seasonal ARIMA model is written by

$$\phi_p(B^s)\varphi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t \quad (1)$$

where, $\{w_t\}$ is the non stationary time series, {wt} is the usual Gaussian white noise process and s is the period of the time series. The ordinary autoregressive and moving average components are denoted by polynomials φ(B) and θ (B) of orders p and q. The seasonal autoregressive and moving average components are $\phi_p(B^s)$ and $\Theta_Q(B^s)$, where P and Q are their orders. ∇^d and ∇_s^D are ordinary and seasonal difference components. B is the backshift operator. The expression given by,

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p$$

$$\phi_p(B^s) = 1 - \phi_1 B^s - \phi_2 B^{2s} - \ldots - \phi_p B^{ps}$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$$

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{Qs}$$

$$\nabla^d = (1 - B)^d$$

$$\nabla_s^D = (1 - B^s)^d$$

$$B^k x_t = x_{t-k}$$

In this study, we concentrate on monthly rainfall time series. If the seasonal period of the series s = 12. It is clear that we may then rewrite the equation (1) is given by,

$$\phi_p(B^{12})\varphi(B)\nabla_{12}^D\nabla^d x_t = \Theta_Q(B^{12})\theta(B)w_t \quad (2)$$

3.3. Model Identification procedure

In the tentative specification phase, namely model identification, the goal is to employ computationally simple methods to narrow down the range of parsimonious models. The B-J method is only suitable for stationary time series data. In such case, we ought to possibly observe the time series graph and transform the data appropriately.

First, we should concept a time plot of the data and inspect the graph for any anomalies (Cryer and Chan, 2008). If the variance grows with time, it will be necessary to stabilize the variance. The next step is to check for additive and lasting level shifts unaccounted for by the model using the outlier statement. Augment the original dataset with the regression variables that match up to the detected outliers. If the data are identified preliminary values of autoregressive order p, the order of differencing d, the moving average order q and their corresponding seasonal parameters P, D and Q. Here, the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) are the most important elements (Stoffer and Dhumway, 2010). The ACF methods the amount of linear dependence between explanations in a time series that are separated by a lag q. The PACF benefits to determine how many autoregressive terms p is essential. The parameter d is the order of difference incidence from non-stationary time series to stationary series. Furthermore, a time series plot and ACF of data will typically suggest whether any differencing is needed. If differencing is called for, the time plot will illustration some kind of linear trend.

When initial values of D and d have been fixed, the next step is to check the ACF and PACF of $\nabla_{12}^{D}\nabla^{d}x_t$ to determine the values of P, Q, p and q. We can further select parameters using Akaike's Information Criterion (AIC) to determine the values of parameters (Stoffer and Dhumway, 2010).

Form the residuals from your chosen model by plotting the ACF of the residuals, and doing a multiple test of the residuals. If they do not look like white noise, try a modified model. Once the residuals aspect like white noise,calculate forecasts.

3.4. Parameters Estimation

Once the model is tentatively established, the parameters and the corresponding standard errors can be estimated using statistical methods, such as Maximum Likelihood (ML), least square estimation method and Yule-Walker.

3.5. Diagnostic Checking

Although we selected model may appear to be the best among those models considered, it is also essential to do diagnostic checking to verify that the model is adequate.

Residuals is a good forecasting model, the residuals left over next fitting the model should be simply white noise. Therefore, if the ACF and PACF of the residuals are obtained, we would confidence to find no significant autocorrelation and no significant partial autocorrelation.

Outliers are normal in such plots to standardize the residuals so they have a variance equal to one. This makes it easier to spot outliers. In the least residual smaller than -3 or greater than 3 is an outlier and maybe worth investigating. None of the ACF or PACF spikes are outside the limits, also suggesting the residual series is white noise.

3.6. Forecasting with Seasonal ARIMA Model

Once a model has been identified and all the parameters have been estimated, we can forecast future values of a time series with this model.

4. Results and Discussion

In this study, we have proposed the outlier detection and mean substitution technique is used for missing observation. The results are obtained by using SPSS and R-software package. The data set has continued some missing year of the observation and some outlying observations are shown in **Table 1**. When we have forecasted with missing observation, it will be affecting the original results are shown in **Fig.1 and 2.**

The ACF and PACF of the original data $\{x_t\}$, t =1, 2… 115 are shown in **Fig.3**. The ACF and **Fig.2** shown seasonal fluctuation occur every 12 months, resulting in s = 12 (Wang, 2008; Momani and Naill, 2009). Concentrating on the ACF of unique data, we note a measured decreasing trend in the ACF peaks in seasonal lags, h= 1s, 2s, 3s, 4s, where s = 12. It indicates a non stationary behaviour and suggests a seasonal difference. **Fig.4** shows the ACF and PACF of the seasonal rainfall data. The ACF decreases to zero exponentially indicating a stationary behaviour (Stoffer and Dhumway, 2010; Han *et al*.,2008). Then the Seasonal ARIMA (p, 0, q)(P, 1, Q)$_{12}$ model could be fitted to the seasonal data. From ACF of the stationary series, we can notice the ACF peak at h =1s; while for PACF, it peaks at h =1s, 2s… 6s. This phenomenon means that the ACF is cutting off following lag 1s and the PACF is tailing off in the seasonal lags. So we be able to build two models: (i) an SAR model of order Q = 1, or (ii) an Seasonal ARMA of orders P = 1, 2,…,6 and Q = 1. The characteristic of graph turns out the model (i) is much better. Inspecting the ACF and PACF at lags h = 1, 2,…,11, it appears that either (a) ACF and PACF are both investigation off (b) PACF cuts off at lag 1, ACF tails off (c) ACF cuts off at lag 1, PACF tails off.

The result indicates that we should consider the following models and choose a better model based on BIC criteria. The optional model and the correlation values are shown in **Table 2**. Obviously, Seasonal ARIMA (1,1,1) (0,1,1)$_{12}$ has the smallest value of BIC and then we temporarily have Seasonal ARIMA

$(1,1,1)\ (0,1,1)_{12}$. As a rule of thumb in Seasonal ARIMA modelling, we need to minimize the sum squared of residuals (RSS) and the number of model parameters. We have considered this message when calculating the related values (Stoffer and Dhumway, 2010).

The model parameters are estimated by using the Maximum Likelihood Estimation. The related parameters are shown in **Table 3.** It can be observed the parameters of model Seasonal ARIMA $(1, 1, 1)\ (0, 1, 1)_{12}$ are all significant. Then we plug the related parameter into the equation (2) and (3) the fitted model in this case is given by

$$\phi_1(B^{12})\varphi(B)\nabla_{12}^{1}\nabla^{1}x_t = \Theta_1(B^{12})\theta(B)w_t \qquad (3)$$

The diagnostics for the model Seasonal ARIMA $(1, 1, 1)(0, 1, 1)_{12}$ is shown in **Fig.5 and 6.** The standardized residual shows no obvious patterns, although there are a few suspicious values and unusual values (Kantz and Schreiber, 2004). The model fits well, although a small amount of autocorrelation still remains. Moreover, we use the Ljung-Box test to examine the independence of the residuals. The p-values of Q-statistic for the first 12 lags of the model are shown in **Fig 5.** The model Seasonal ARIMA $(1, 1, 1)(0, 1, 1)_{12}$ Equation (4)

$$(1-\varphi_1 B)(1-\phi_1 B^{12})\nabla_{12}^{D}\nabla^{d}x_t = (1+\Theta_1 B^{12})(1+\theta_1 B)w_t \qquad (4)$$

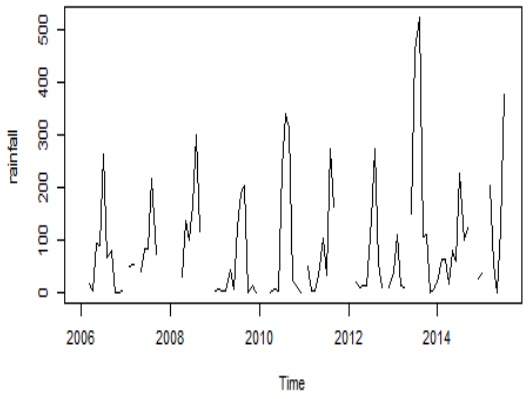

Fig.1: Time series plot for rainfall data

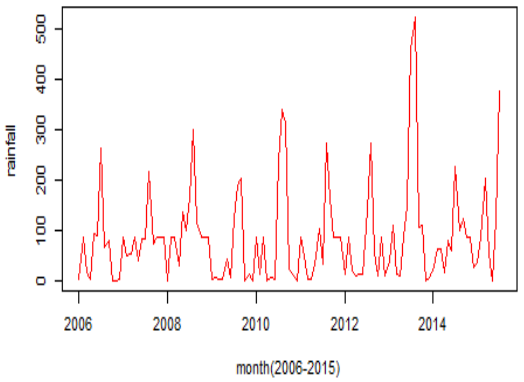

Fig.2: Plot for missing value replaced by mean

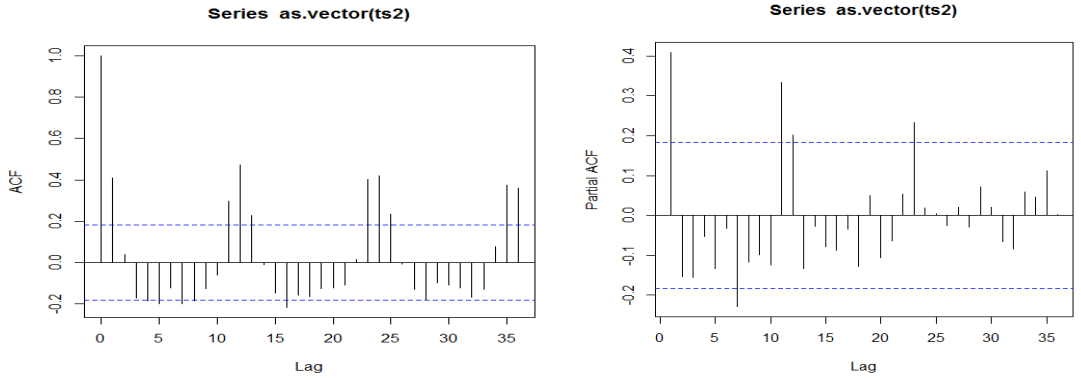

Fig. 3: (a) Autocorrelation (ACF) and (b) Partial Autocorrelation (PACF) for original time series of monthly rainfall data

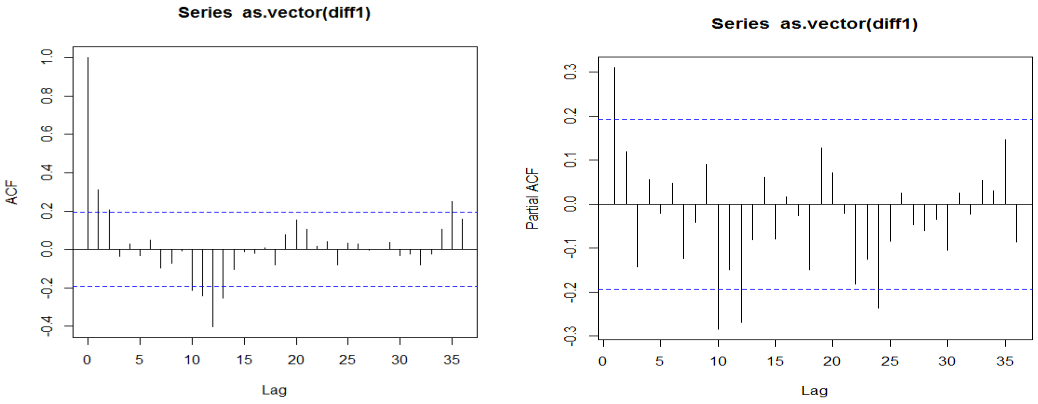

Fig. 4: (a) Autocorrelation (ACF) and (b) Partial Autocorrelation (PACF) for first order seasonal differencing of monthly rainfall data

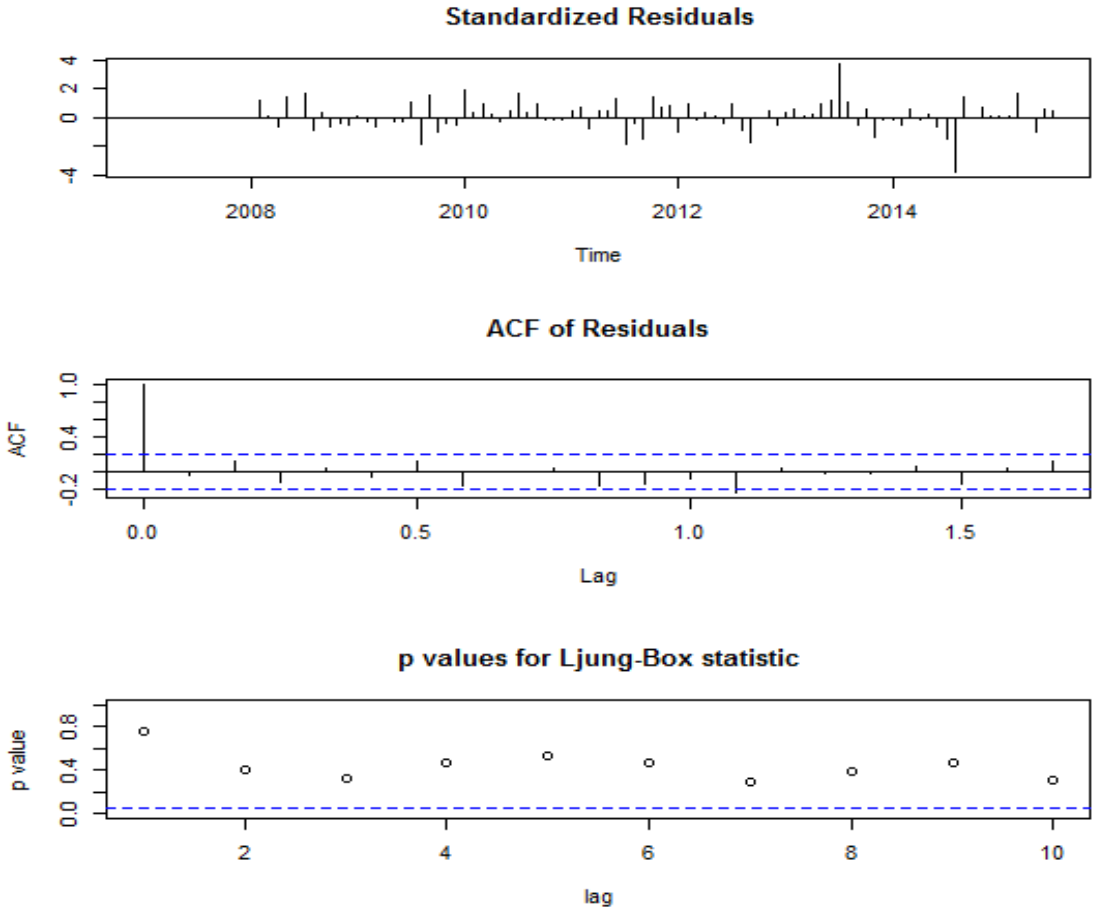

Fig. 5: Diagnostic for the Seasonal ARIMA (1, 1, 1) $(1, 1,1)_{12}$ model

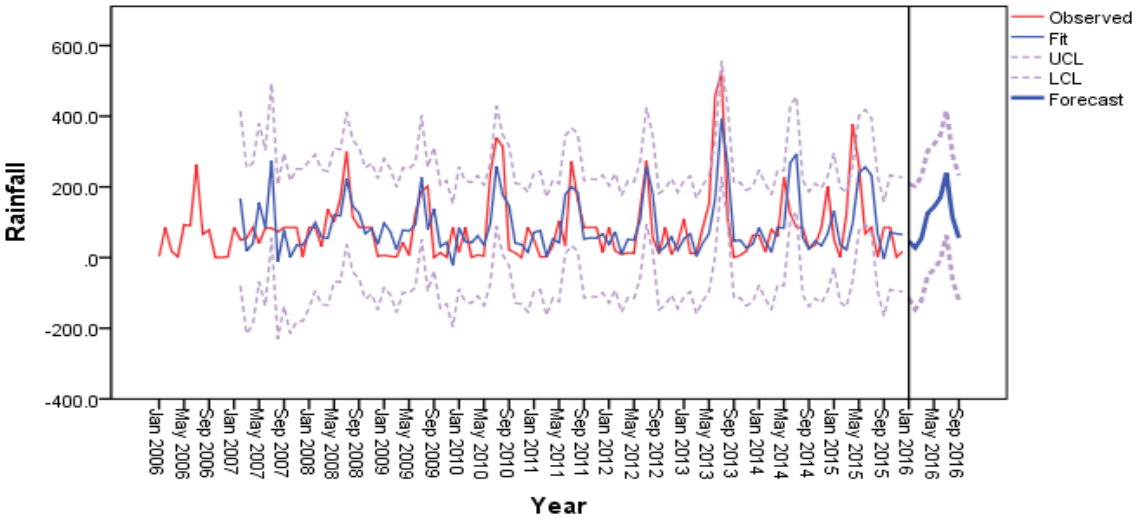

Fig.6: Forecasting and real values of monthly rainfall data

Table 1. Model selection for Mean substituted variables with outlier detection and removing

Model Selection	Outliers		Normalized BIC	
Variable Name SARIMA(P,D,Q) (p,d,q)	**Detected**	**Types**	**With Outlier**	**Without Outliers**
Mean (1,1,1) (0,1,1)	2(Aug 2007, Aug 2014)	SAO	0.433	0.263
	1 (Sep 2013)	IO	0.433	0.263

Table 2. Optional models and the related standard values

Models	Normalized BIC
SARIMA(1, 1, 0) $(0, 1, 1)_{12}$	9.525
SARIMA(0, 1, 0) $(0, 1, 1)_{12}$	10.204
SARIMA(1, 1, 1) $(0, 1, 1)_{12}$	**9.056**
SARIMA(2, 1, 1) $(1, 1, 1)_{12}$	9.363

Table 3. Estimates of the model parameters

	Model parameters		
Model	φ_1	ϕ_1	Θ_1
SARIMA (1, 1, 1) $(0, 1, 1)_{12}$	-0.040	0.996	0.807
Standard Error	0.120	0.834	0.162

The equation can be multiplied and written in the following form that is used in forecasting, the values of the correlation coefficient are shown in Table 3.

$$\hat{Y}_t = c + \varphi_1 x_{t-1} + x_{t-12} + x_{t-13} + w_2 e_{t-12} + \Theta_1 \theta_1 w_{t-13} + w_t$$

Finally, the comparison between the real values and the fitted value is shown in **Figure 6.** The vertical dotted line separates the data from the predictions.

Because of many stochastic environmental factors, such as temperature, geographic location and climate, the model state of rainfall is a complicated dynamical system. The time series model to study does not model

the extreme values well. Further extensions of study may be undertaken by considering an interference time series analysis, such as an Autoregressive conditional Heteroskedasticity model to model the phenomenon of extremes.

6. Conclusion

In this study, we have presented outlier detection and missing value estimation using Seasonal ARIMA procedure. Seasonal ARIMA models provide a simple and flexible tool to forecast time series data, thus obtaining sufficient data to estimate the characteristic for a standardized lactation length with the same accuracy obtained with more complicated methods of prediction. Precision can be further improved by adopting more selective condition for the construction of more standardized groups of lactations. To improve the model replacing missing values and identifying SAO and IO outliers that values removed by missing value subsequent to it replaced by mean value. The fitted Seasonal ARIMA $(1, 1, 1)(0,1,1)_{12}$is performing improved forecasting results than observed data.

Acknowledgements

The Second author thanks the University Grants Commission, New Delhi for awarding fellowship under the scheme of UGC – Basic Science Research fellowship to carry out this work**.**

References

[1] Abraham B, Box GEP (1979). Bayesian analysis of some outlier problems in time series. Biometrika, 66(2): 229-236.
[2] Abraham B, Chuang A (1989). Outlier detection and time series modeling. Technimetrics, 31(2): 241-248.
[3] Balke NS (1993). Detecting level shifts in time series. Journal of Business and Economic Statistics, 11(1): 81-92.
[4] Box GEP, Jenkins GM (1976). Time Series Analysis: Forecasting and Control, Revised Edition, San Francisco: Holden Day.
[5] Cryer, J. D. and K.S. Chan, 2008. Time Series Analysis with Application in R. 2nd Edn., Springer, New York, ISBN-10: 0387759581, pp: 491.
[6] Gao, M. and X.Y. Hou, 2012. Trends and multiracial analyses of precipitation data from Shandong peninsula, China. Am. J. Environ. Sci., 8: 271-279. DOI: 10.3844/ajessp.2012.271.279
[7] Guo, Z.W., 2009. The adjustment method and research progress based on the ARIMA model. Chinese J.Hosp. Stat., 161: 65-69.
[8] Han, P., P.X. Wang and Y.J. Wang, 2008. Drought forecasting based on the standardized Precipitation index at different temporal Scales using ARIMA models. Agric. Res. Arid Areas, 26: 212-218.
[9] Kantz, H. and T. Schreiber, 2004. Nonlinear Time Series Analysis. 2nd Edn., Cambridge University Press, Cambridge, and ISBN-10: 0521529026, pp: 369.
[10] Mohammad SZ, Siti MZ, Kamarulzamman I, AzamiZ, Sopian K (2010). Additive Outliers (AO) and Innovative Outliers (IO) in GARCH (1, 1) Processes. Recent Advances in Applied Mathematics (pp. 471-479). Cambridge: WSEAS Press.
[11] Senthamarai Kannan.K, Manoj.K, and Arumugam. S.G (2015). Outlier Detection and Missing Value in Time Series Ozone Data. International Journal of Scientific Research in Knowledge, 3(9). pp. 0220-226.
[12] Xinghu Chang, Meng Geo, Yan Wang AND Xiyong Hou (2012). Seasonal Autoregressive Integrated Moving Average Model For Precipitation Time Series. Journal of Mathematics and Statistics 8(4): 500-505.