Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 75(1) 2021 1–7

Detection of Multiple Outliers in Time Series: An Application to Rice Yield Data

Gopal Saha, Ranjit Kumar Paul and L.M. Bhar

ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 25 April 2018; Revised 23 November 2020; Accepted 24 November 2020

SUMMARY

Detection of outliers in time series data is a key component of data analysis. As the presence of outlier have a serious effect on model identification statistics, therefore conclusions drawn through analyzing the data series contaminated with outliers may be erroneous. It is, therefore, important to identify the time points where outliers are present and then remove the effect of the outliers from the corresponding series. The present paper considers the detection of outliers in time series data. An iterative method based on the procedure proposed by Chang and Tiao (1983) along with use of robust estimate of error variance is discussed. The power of this iterative procedure in detecting outliers is also investigated. The methodology is illustrated using rice yield data for all India during 1950-2013. The result of the study clearly indicates outlier detection technique using the robust estimate of error variance can successfully detect all the outliers present in the data series.

Keywords: Autoregressive moving average model, Intervention model, Iterative estimation, Outlier.

1. INTRODUCTION

Most of the time series are observational in nature. Besides the possible gross error, time series data are frequently subject to the influence of some non-repetitive errors, for example, major changes in economic policies or political scenario, implementation of new regulation, occurrence of natural disaster *etc.* Consequently, occurrence of outliers or discordant observations is very common in time series data. The basic objective behind designing of time series model is to grasp the homogeneous memory pattern of the series. But the presence of outliers raises the question of efficiency and adequacy in fitting a general auto-regressive moving average (ARMA) model.

Fox (1972) seemed to be the first person to deal with outliers in time series. He defined two types of outliers in time series namely: Additive outliers (AO) and Innovative Outliers (IO). The study of Chang (1982) showed that even if the order of the time series is known, the existence of outliers may cause serious bias in estimation of the autoregressive (AR) and moving average (MA) parameters. If we know the time points where the outliers are occurring, then the biases in estimation of model parameters can be reduced by using the intervention technique of Box and Tiao (1975). But in practical situation, we hardly have any knowledge regarding the timing and type of outliers present in a given data series. For detecting outliers in time series, several scientists proposed different approaches. Among them Chang and Tiao (1983) used an iterative procedure for outlier detection.

This study's main goal is to identify the time points at which outliers are occurring for any given data series. We use the iterative procedure as described by Chang and Tiao (1983) but in place of using the usual estimate of error variance, we try to use some robust statistics. The reason behind it is that the presence of outliers makes the usual estimate of error variance unstable and it will be highly biased, which ultimately affect the value of the test statistics used to detect the outliers. But if we use the robust estimate of error variance then it will be least affected by the presence of outliers and it will consistent also. Besides outlier detection, we will

Corresponding author: Ranjit Kumar Paul
E-mail address: ranjit.paul@icar.gov.in

try to obtain an adjusted residual series after removing effects of all the detected outliers.

The paper is divided as follows. Section 1 deals with the introduction, the models and the test statistics for detecting outliers are discussed in section 2. Then, the outlier detection criteria is given in section 3. In section 4, illustrations are given for the discussed outlier detection procedure. Finally, in section 5, the conclusions are provided.

2. MATERIALS AND METHODS

2.1 Description of Models

Consider an outlier-free time series, the series is assumed to follow an autoregressive integrated moving average [ARIMA (p, d, q)] model,

$$\phi(\mathrm{B})\alpha(\mathrm{B})x_t = \theta(\mathrm{B})a_t \tag{2.1}$$

where, $\phi(\mathrm{B}) = 1 - \phi_1\mathrm{B} - ... - \phi_p\mathrm{B}^p$ and $\theta(\mathrm{B}) = 1 - \theta_1\mathrm{B} - ... - \theta_q\mathrm{B}^q$ are the 'autoregressive' and 'moving average' polynomials in B of order p and q respectively. B is defined as a backshift operator, *i.e.*, $\mathrm{B}x_t = x_{t-1}$. $\{a_t\}$ is the white noise which is independently and identically normally distributed with zero mean and constant variance σ_a^2. For seasonal data considering period s (for monthly data s=12), one can assume $\alpha(\mathrm{B})=(1\text{-}\mathrm{B})^{d_1}(1\text{-}\mathrm{B}^s)^{d_2}; \mathrm{d} = \mathrm{d}_1 + \mathrm{sd}_2$ which is helpful to allow seasonal non-centrality.

Under the above notations and considering that all the assumptions hold good, one can easily estimate the parameters using 'Box-Jenkins' methodology. But the presence of outliers or exogenous observations in the model lead to introduction of serious bias during estimation of model parameters. The impact of exogenous intervention occurring in the time series can be represented by the dynamic model due to Box and Tiao (1975):

$$Z_t = \frac{\omega(\mathrm{B})}{\beta(\mathrm{B})}\zeta_t^{(T)} + x_t \tag{2.2}$$

where,

$\zeta_t^{(T)}$ represents the impulse variable which takes value 1 at $t = T$ and otherwise 0, *i.e.*,

$\xi_t^T = 1$ for $t = T$

$\xi_t^T = 0$ otherwise.

T signifies the time point at which occurrence of intervention (Outlier) takes place. $\omega(\mathrm{B})/\beta(\mathrm{B})$ describes the dynamic response of the outlier.

2.2 Detection of Outliers

In the time series process as described in (2.1), two possible conditions may arise. Firstly, the ARMA parameters and the error variance are known and secondly, they are unknown. The first criteria is seldom arise in practical situation. In the second situation we cannot go directly for estimation of impact as we need to estimate the model parameter first.

2.2.1 *Estimation of impact of IO and AO*

To estimate the effect of outliers, the impact of AO and IO needs to be estimated. The least square estimate of impact of additive and inovative outliers can be represented as the following way (Chang and Tiao, 1983),

$$\hat{\omega}_{\mathrm{I}} = \hat{e}_t \ \ldots (\mathrm{IO}) \tag{2.3}$$

$$\hat{\omega} = \rho\ \pi(\mathrm{F})\hat{} = \rho^2(1 - \pi_1\mathrm{F} - \pi_2\mathrm{F}^2 - ... - \pi_{n\ T}\mathrm{F}^{n\ T})\hat{}_T \ \ldots (\mathrm{AO}) \tag{2.4}$$

where $\rho^2 = (1 + \pi^2_1 + \pi^2_2 + ... + \pi^2_{n-T})^{-1}$ and F is a forward shift operator like $\mathrm{F}e_T = e_{T+1}$. is the residual at time point T, $\pi(\mathrm{B}) = \phi(\mathrm{B})\alpha(\mathrm{B})/\theta(\mathrm{B}) = 1 - \pi_1\mathrm{B} - \pi_2\mathrm{B}^2 - ...$ is the polynomial in B and are the value of coefficients of different lags of residuals.

Next, we need to find the variance of estimated impact for both models and the variances for the estimators are as follows:

$$\mathrm{var}(\hat{\omega}_{\mathrm{I}}) = \sigma_a^2 \ \ldots (\mathrm{IO}) \tag{2.5}$$

$$\mathrm{var}(\hat{\omega}_{\mathrm{A}}) = \rho^2\sigma_a^2 \ \ldots (\mathrm{AO}) \tag{2.6}$$

2.2.2 When ARMA Parameters and Error Variances are Unknown

In most of the practical situations, both the ARMA parameters and error variance (σ_a^2) are unknown. There we need to find the estimates of model parameters along with the estimates for the impact, *i.e.*, for both IO and AO. The estimates of the model parameters is obtained by MLE technique.

To compute the test statistics for detection of outliers, the unknown error variance ($\hat{\sigma}_a^2$) is also estimated. The value of test statistics or in other words efficiency of multiple outlier detection is very sensitive to this estimate. As it is obvious that the presence of

outliers makes the residuals contaminated and using the simple variance of residuals as an MLE estimate of error variance is not at all a robust one. We know that median is the robust estimate in presence of outliers. Therefore, here we consider a robust estimate of error variance, i.e., Median Absolute Deviation (MAD).

The test statistics for testing the presence of outliers (Chang and Tiao, 1983) are as follows:

$$\text{Test Statistic1: } H_0 \text{vs} H_1 : \hat{\lambda}_{1,t} = \hat{\omega}_I / \hat{\sigma}_a$$

$$\text{Test Statistic2: } H_0 \text{vs} H_2 : \hat{\lambda}_{2,t} = \hat{\omega}_A / \hat{\rho}\hat{\sigma}_a \ , \qquad (2.7)$$

The MAD estimate of error standard deviation can be defined by:

$$\hat{\sigma}_a = 1.483 \times \text{median}\left\{\left|\hat{e}_t - \tilde{e}\right|\right\}, \qquad (2.8)$$

where $\tilde{e}$ represents the median of the residuals.

After estimating the model parameters and the error standard deviation, one can estimate the impact of outliers using the equations (2.3) for IO and (2.4) for AO respectively.

3. OUTLIER DETECTION CRITERIA

For detecting the presence of AO or IO at different unknown positions, we first need to go through the sequence of $\hat{\lambda}_{2,t}$, $t = 1,2,...,n$ or $\hat{\lambda}_{1,t}$, $t = 1,2,...,n$. On other way, the chances of occurrence of an AO in a series or an IO in a series can be examined by searching for the maximum values of both the statistics. If the maximum value of the statistics exceeds the cut-off value (Fox, 1972) then we can say an outlier is present at that corresponding time point (say, T).

$$\text{(IO) } \hat{\eta}_{IO} = \max_{t=1,2,...,n}\left(\left|\hat{\lambda}_{1.t}\right|\right) \text{ Or}$$

$$\text{(AO) } \hat{\eta}_{AO} = \max_{t=1,2,...,n}\left(\left|\hat{\lambda}_{2.t}\right|\right) \qquad (3.1)$$

As the time of occurrence of outliers remains unknown, using likelihood ratio criteria described in (3.1) one can easily find time points where IO or AO is present. But if for a particular time point both the test criteria for AO and IO are significant, then it is difficult to say which one type of outlier is present in that point. To address this problem, a simple rule is considered as mentioned by Fox (1972) for distinguishing between IO and AO. The rule if at any particular point *T* (say), the possible outlier is called an IO if $\left|\hat{\lambda}_{1.T}\right| > \left|\hat{\lambda}_{2.T}\right|$ and classified as an AO if $\left|\hat{\lambda}_{1.T}\right| \le \left|\hat{\lambda}_{2.T}\right|$.

4. ILLUSTRATION

Data Set

We illustrate the above described procedure of multiple outlier detection by considering the time series data of rice yield (kg/ha) from the period of 1950-51 to 2017-18. The data is collected from *Agricultural Statistics at a Glance 2018*, published by Ministry of Agriculture and Farmers Welfare, Department of Agriculture and Cooperation and Directorate of Economics and Statistics, Government of India. The data series contains 68 observations which are collected yearly basis. The time plot of the above dataset is presented in Fig. 1.

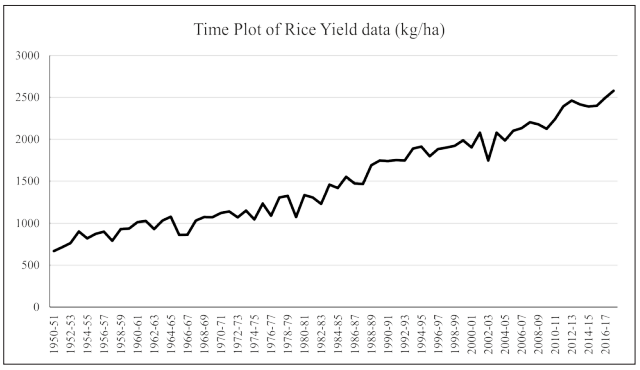

Fig. 1. Time plot of the Rice Yield Data

The flow of analysis is as follows:

(i) The descriptive analysis is done on the time series data to get an evidence of possibility of outlier in the series.

(ii) If there are no sufficient evidence of presence of outliers, the data is fitted as per the usual methodology.

(iii) If there is sufficient evidence of presence of outliers from descriptive analysis of the dataset (which is reflected in our data), at initial stage data is fitted assuming there is no outlier in the dataset. This is done only to obtain the coefficient values of the model which will be used to estimate the impact of the outliers and also to identify the time-points at which outliers are present.

(iv) After obtaining the model coefficients, the robust estimate of error variance (as given in 2.7) is computed. Then the test statistics (mentioned in 2.8) are computed to check whether the impact of the outliers are significant or not.

(v) Further, the iteration procedure (as described in section 3) is carried out to obtain the list of potential outliers in the dataset.

(vi) As we identify the list of outliers in the dataset, further analysis can be done based on the new residual series which have been adjusted after removing the effects of potential outliers.

Descriptive Statistics

The descriptive statistics of the dataset is reported in Table 1. A perusal of table 1 reveals that the difference between the maximum value (2578.00) and minimum value (668.00) is very large which indicates that there may be some chance of presence of outliers in the dataset. From the time plot also, we can visualize that there is possibility of presence of outliers.

Table 1. Descriptive Statistics of Rice Yield Data

Statistics	Value	Statistics	Value
Mean	1513.82	Kurtosis	-1.22
Standard Error	66.33	Skewness	0.29
Median	1437.00	Range	1910.00
Mode	1308.00	Minimum	668.00
Standard Deviation	546.95	Maximum	2578.00
Coefficient of Variation (%)	36.13		

Testing Stationarity and Model Fitting

The stationarity of the series is tested by Augmented-Dickey-Fuller test (ADF) (Dickey and Fuller, 1979) and Phillips-Perron (PP) (Phillips and Perron, 1988) test. The results of both the tests indicate that the series is stationary.

Now we fit the ARIMA model in the data series and it is found that the ARIMA (1, 0, 0) is the best fit for this data with the AIC value 854.32. The model parameter estimates along with their standard errors are given in Table 2. The estimated error variance is 14935.

Table 2. Parameter Estimate of ARIMA Model

Parameters	Value	Standard Error
Intercept	1593.6599	683.6545
AR 1	0.9887	0.0140

After fitting the ARIMA model in the series, the set of residuals are obtained from the fitted model. This series of residual is plotted in the Fig. 2. From this figure it is clear that there are some residuals whose absolute value is very large than the rest of the residuals and hence they might be due to outliers.

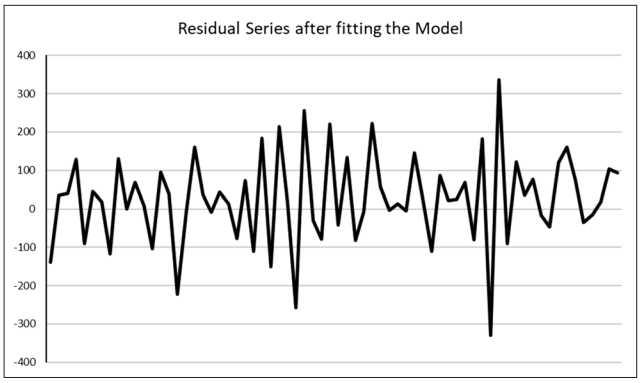

Fig. 2. Residual Series for the Fitted Model

Outlier Detection

As both the descriptive statistics and residual series give indication of presence of outliers in the dataset, we therefore apply the methodology of outlier detection as described in section 3. Here we have used a robust estimate of error variance which appeared in the test statistics. This error variance is robustly estimated in each of the iteration. Finally, we find that there is a presence of total 4 outliers. Among them there are two additive outliers appeared at time point 53, 30 and 27 respectively, and one are innovative outlier which appeared at time point 54. The list of the outliers detected through the iterative procedure using the robust variance is given in the Table 3.

Table 3. List of Detected Outliers

S. No.	Year	Time Point	Type of Outlier	Value
1	2002-03	53	AO	1744
2	1979-80	30	AO	1074
3	2003-04	54	IO	2079
4	1976-77	27	AO	1089

The detected outlier points are depicted in the original time series in Fig. 3. Besides outlier detection, we also have a new residual series after adjusting the effect of the detected outliers. The adjusted residual series is computed by removing the effect of the detected outliers. The adjustment is done only for those time points in the residual series where a particular outlier is being detected. The comparison between the original residual series (series 1) and the residual series after adjusting the effect of outliers (series 2) is given in the Fig. 4.

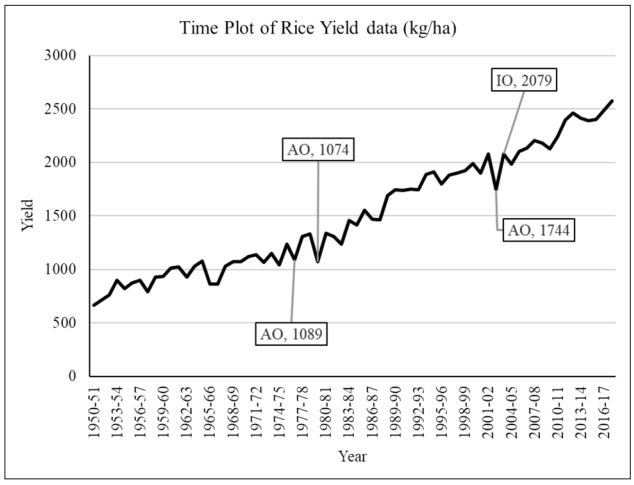

Fig. 3. Outliers Detected in Rice Yield Data

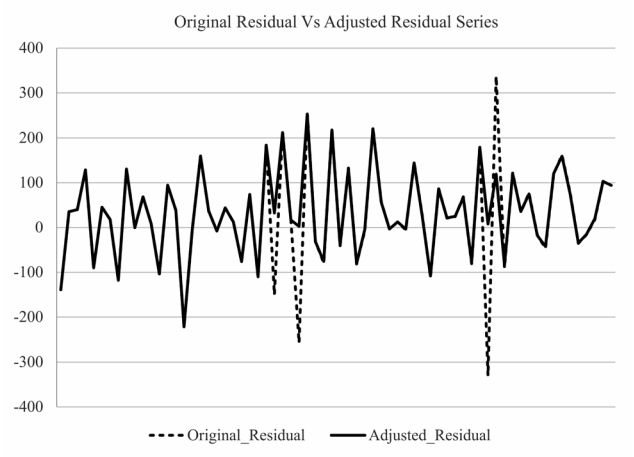

Fig. 4. Comparison between Original Residual (dotted) and Adjusted Residual (line)

From Table 3, we can see that the first outlier detected is an AO at time point 53. From the above Fig. 3, we can see the value of observation at time point 53 is 1744 whereas the value of its two adjacent observations are 2079 (at time point 52) and 2079 (at time point 53) which are relatively higher as compared to 1744.Therefore, there is a sudden decrease in the value of the observation at time point 53 which can be considered as an AO. The same instances occurred at time point 30 (1979-80), the value is 1074 whereas neighboring observation values 1328 (for year 1978-79) and 1336 (for 1980-81) are relatively higher.

Then, we can visualize that after the time point (T) 54 (value 2079) there is a continuous increase in the values of observations for at least 5 to 6 years. Therefore, we can say that there is an IO at time point 54 which affects the rest of the series and due to which there is a continuous increase in the observation values from 2004-05 onwards. Similarly, the observations at time point 27 is detected as an AO because the value of the observation at time point 27 is decreased suddenly as compared to its neighboring observations. From the Fig. 4, we can clearly visualize that all the residuals having relatively large absolute values than the rest are adjusted. Therefore, we can conclude that the iterative detection technique is very efficiently handle the outliers and also remove their effects from the original series.

The practical reason of considering these time points as outlier is:

- The year 2002-03 was the first year of tenth five-year plan (2002-07) and in this plan the major target set up by the commission was to revamp the agriculture sector as during ninth five-year plan (1997-2002) the average annual growth rate of value added in agriculture, including allied sectors, declined from 4.7 per cent to 2.1 percent. But, 2002 witness a severe drought and the growth in agricultural sector is declined by -7.0 percent which is mostly attributed by the major drop in rice production (-16% growth in rice yield). Therefore, in our analysis this particular time point is considered as AO. (https://economictimes.indiatimes.com/news/economy/policy/agriculture-production-and-growth-monsoon-04/articleshow/1032957.cms)
- The effect of tenth five-year starts reflecting from 2003-04, where in 2003-04 favorable monsoon facilitated an impressive growth rate of 9.6 per cent in agriculture and allied sector gross value addition whereas in rice production there is a straight jump of 19% as compared to last year. Also, during the last 4 years of tenth five-year plan, the average growth in rice production is around 5.5%. Therefore, the timepoint 2003-04 came out as IO in our analysis.
- Besides, there was a widespread failure of monsoon during 1976-77 and 1979-80. This leads to welting of crops in major producing states including Kerala and Karnataka. (https://indianexpress.com/article/opinion/editorials/july-8-1976-forty-years-ago-2900148/)

5. CONCLUSION

The interpretations drawn from analyzing any data series having outliers is misleading and erroneous. Therefore, detection of outliers in time series dataset is very important. In the above illustrations we detect

outliers of both types. Using the iterative method, we can detect multiple outliers (if present) from any time series dataset. After detecting the outliers, we also find the final residual series which is adjusted for the effect of detected outliers. This final residual series can now be used for further analysis which will be more informative and accurate. R codes are developed for identification of outliers and the same is given in the appendix. Using that code, we can easily check the presence of outliers in any time series dataset. The example discussed also suggests that the outlier having largest effect is detected at the first step of the iteration procedure and subsequently the other outliers are detected based on their magnitude or effect.

ACKNOWLEDGEMENTS

The authors express sincere thanks to the anonymous reviewer for extensive comments which helped in improving the manuscript.

REFERENCES

Box, G.E.P. and Jenkins, G.M. (1970). *Time Series: Forecasting and Control*. San Francisco, Holden-Day.

Box, G.E.P. and Tiao, G.C. (1975). Intervention analysis with application to environmental and economic problems. *J. Amer. Statist. Assoc.*. **70**, 70-79.

Chang, I and Tiao, G.C. (1983). *Estimation of Time Series Parameters in the Presence of Outliners*. Technical report 8, University of Chicago, Statistics Research Centre.

Chang, I (1982). *Outliers in Time Series*. PhD Thesis. Department of Statistics, University of Wisconsin, Madison.

Chen, C. and Liu, L. (1993). Joint estimation of model parameters and outlier effects in time series. *J. Amer. Statist. Assoc.*. **88,** 284-297.

Dickey, D.A. and Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.*. **74,** 427-431.

Fox, A.J. (1972). Outlier in time series. *Journal of Royal Statistical Society*. Series B. **34**, 350-63.

Phillips, P.C.B. and Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika*. **75(2)**, 335-346.

APPENDIX:

R code:

```
#Required library function in R#
library("forecast")

#Reading the dataset#
yt<-Rices["Production"]

#Transforming to time series#
yt_ts<-ts(yt)

#Fitting of the original dataset#
fit_yt<-auto.arima(x=yt_ts, d=NA, D=NA, max.p=5, max.q=5,stationary=TRUE,
seasonal=FALSE,ic=c("aic"), allowdrift=TRUE, allowmean=TRUE,stepwise = TRUE)
fit_yt
et<-residuals(fit_yt)
plot(et)
write.csv(et, file = "D:\\Personal\\Paper I\\original_res.csv")

#Obtaining the length of the dataset#
nobs<-length(yt_ts)
nobs

#Putting the value for pai-1 after calculating manually from the estimated values#
p1<-0.9887

#Computing rho^2 with the help of pai-1#
rho<-(1/(1+p1^2))
rho_s<-sqrt(rho)

#Computing the effect of Outlier#
outlier = 1;OutlierSummary <- NULL
while(outlier == 1){
omega<-0
for(i in 1:nobs)
omega[i]=rho*(et[i]-p1*et[i+1])

#Median of residuals: Required to compute the test statistics#
medet<-median(et)

#Deviation of observation from median: Required to compute the test statistics #
dev<-0
for(i in 1:nobs)
dev[i]=(et[i]-medet)
```

```
a_dev<-abs(dev)
md<-median(a_dev)
#Robust estimate of standard deviation needed for computing test statistics value#
sigma1_s<-1.483*md

#Test statistics for AO#
lamda2<-omega/(sigma1_s*rho_s)

#Computing the maximum absolute value of test statistics of AO#
a_lamda2<-abs(lamda2)
max_lamda2<-max(a_lamda2,na.rm=TRUE)

#Computing the time point at which maximum value of test statistics of AO: Identification of time points at which Additive Outlier is occurring#
time_ao<-which.max(a_lamda2)

#Computing the test statistics for IO#
lambda1<-et/(sigma1_s)

#Computing the maximum absolute value of test statistics of IO#
a_lambda1<-abs(lambda1)
max_lambda1<-max(a_lambda1)

#Computing the time point at which maximum value of test statistics of IO: identification of time points at which Innovative Outlier is occurring#
time_io<-which.max(a_lambda1)

#Computing the series of IO and AO at max values: Required to identify the time points at which both types of outliers are occurring#
nt<-0
for(i in 1:(nobs-1))
nt[i]=max(a_lamda2[i],a_lambda1[i])
max_nt=max(nt)
outlier <- ifelse(max_nt>3.0,1,0)
T1<-which.max(nt)
nt[T1]=max(a_lamda2[T1],a_lambda1[T1])
if (T1==time_io)
{
outlierType <- "IO IS present"
temp <- data.frame(T1,outlierType)
}else{
outlierType <- "AO IS present"
temp <- data.frame(T1,outlierType)
}
if (T1==time_ao)
{
pib<-0
zi <- c(rep(0,T1-1),rep(1,(nrow(yt_ts)-1-T1)))
for(i in T1:(nobs-1))
# pib[i]=(1-(p1))
pib[i]=(1-p1*zi[i-1])
et1<-0
for(l in T1:(nobs-1))
et1[l]=et[l]-(pib[l]*omega[l])
et2=et[1:(T1-1)]
zt<-et1[T1:(nobs-1)]
n_et1=c(et2,zt)
}else {
et[T1]
n_et1<-replace(et,et==et[T1],0)
}
n_et1[nobs] <- et[nobs]
et <- n_et1
OutlierSummary <- rbind(OutlierSummary,temp)
}
n_et1
max_nt

# Generating list of Time Points where Outliers are present #
Outlier Summary

#Obtaining the adjusted residuals after removing the effect of outliers#
write.csv(n_et1, file = "D:\\Personal\\Paper I\\adj_res.csv")
```