https://dx.doi.org/10.22161/ijcmp.3.3.2 ISSN: 2456-866X

A Method for Detection of Outliers in Time Series Data

Evan Abdulmajeed Hasan

Erbil, Kurdistan region of Iraq
Evan.hasah@outlook.com

***Abstract*—** An outlier is a data value that is an unusually small or large, or that deviates from the pattern of the rest of the data. Outliers are usually removed from the data set before fitting a forecasting model, or not removed but the forecasting model adjusted in presence of outliers. There are four types of OUTLIERS are as follows: Additive outlier (AO), Innovational outlier (IO), Level shift (LS) and Temporary change (TC). There is more than one method for the detection of outlier; the study considers the detection of outlier in two cases: first, at the time when the parameters are known. Second, when the parameters are unknown. There are several reasons for outlier detection and adjustment in time series analysis and forecasting which are mentioned in this study. The study has used the volume of water inflow in the reservoir of Dokan dam in Sulaymaniah city as a time series for the purpose of the study. The study came to conclude that throughout the research, the following conclusions: first, every time increasing the critical value, the value of residual standard error (with outlier adjustment) increased. Second, every time increasing the critical value, the number of outlier values decreased. Third, in the case of presence of outliers the forecasts with adjustment of outliers better than the forecasts without adjusting outliers.

Keywords*— *ARMA model, Innovational Outlier, Temporary Change, Time Series.

I. INTRODUCTION

The study of outliers is not a new phenomenon. It has in fact a long history dating back to the earliest statistical analysis. Outlier methods have developed hand in hand with other statistical methods. Unfortunately, in time series analysis this expansion of outlier methods has not been as rapid and widespread. One reason for this must be that methods of time series outliers were first considered explicitly. However, since then the amount of papers dealing with the issue has grown steadily (Rousseeuw & Bossche, 2018).

Outliers and structure changes are commonly encountered in time series data analysis. The presence of those extraordinary events could easily mislead the conventional time series analysis procedure resulting in erroneous conclusions. The impact of those events is often overlooked, however, for the lack of simple yet useful methods available to deal with the dynamic behaviour of those events in the underlying series. The primary goal of this paper, therefore, is to consider unified methods for detecting and handling outliers and structure changes in a univariate time series. The outliers treated are the additive outlier (AO) and the innovational outlier. The structure changes allowed for are level shift (LS) and variance change (VC). Level shift is further classified as permanent level change (LC) and transient level change (TC) (Rousseeuw, et al. 2019).

The literature study forms the first stage of a research project aiming to establish the applicability of time series and other techniques in estimating missing values and outlier detection/replacement in a variety of transport data. Missing data and outliers can occur for a variety of reasons, for example the breakdown of automatic counters (Cabrieto, et al. 2017). Initial enquiries suggest that methods for patching such data can be crude. Local authorities are to be approached individually using a short questionnaire enquiry form to attempt to ascertain their current practices. Having reviewed current practices, the project aims to transfer recently developed methods for dealing with outliers in general time series into a transport context. It is anticipated that comparisons between possible methods could highlight an alternative and more analytical approach to current practices (Staal, et al. 2019).

Several approaches have been considered in the literature for handling outliers in a time series. Abraham and Box (1979) used a Bayesian method, Martin and Yohai (1986) treated outliers as contamination generated from a given probability distribution, and Fox (1972) proposed two parametric models for studying outliers. Chang (1982) adopted Fox's models and proposed an iterative procedure to detect multiple outliers. In recent years, this iterative procedure has been widely used with encouraging results (Liu, et al. 2018). The methods mentioned above may be regarded as batch-type procedures for detecting outliers, because the full data set is used in detecting the existence of outliers. On the other hand, Harrison and Stevens (1976), Smith and West (1983), West, Harrison and Migon (1985) and West (1986) have considered sequential detecting methods for handling outliers. These sequential methods assume probabilistic models forbyDenby and Martin (1979). This approach is summarized in Martin and Yohai (1985). However, the study of Chang and Tiao (1983) shows that Denby and Martin's robust procedure is not powerful in handling innovational outliers. (Note that the effect of a single I0 on estimation is usually negligible provided that the I0 is not close to the end of the observational period. The effect of multiple IOs, however, could be serious. There is no comparison available between the batch-type and the sequential procedures in handling outliers. The probabilistic treatment has its appeal but may not be easy to implement as it requires prior information of the underlying model to begin with. Since level shifts and variance changes are also considered, the approach of Chang and Tiao (1983) and Tsay (1986a) is adopted and generalized in this study (Arumugam & Saranya, 2018).

Outliers can take several forms in time series. There are additive and innovational outliers. An additive outlier affects a single observation, which is smaller or larger in value than expected. In contrast an innovational outlier affects several observations. Three other types of outliers can be defined, namely level shifts, transient changes and variance changes (Aminikhanghahi& Cook, 2017). A level shift simply changes the level or mean of the series by a certain magnitude from a certain observation onwards. A transient change is a generalization of the additive outlier and level shift in the sense that it causes an initial impact like an additive outlier, but the effect is passed on to the observations that come after it. A variance change simply changes the variance of the observed data by a certain magnitude (Wang & Mao, 2018).

Outliers have some effects on the forecasts from ARMA models, and especially outliers near the beginning of the forecast period can have serious consequences. Point forecasts may suffer only a little from additive outliers, but the prediction intervals can become severely misleading, as outliers can inflate the estimated variance of the series. Level shifts and transient changes can have more serious effects also on point forecasts even when outliers are not close to the forecast region. Attempts have been made to construct forecasting intervals in the presence of outliers (Liu, et al. 2018).

II. LITERATURE REVIEW

Types of outliers in a time series

Temporary Change (TC):

An additive outlier (AO) and a level shift (LS) represent two distinct patterns in which an event affects a series. For LS, the level of the underlying process is affected for all future time, while an AO affects the series for only one time period. It is useful to consider an event that has some initial impacts on a series, and then the impact eventually disappears (Hermosilla, et al. 2015). A temporary (or transient change) (TC) is an event having such an initial impact and whose effect decays exponentially according to some dampening factor, say δ. We can represent the observed series as:

$$Y_t = Z_t + \frac{1}{1-\delta B} W_c \, P_t^{(T)} \qquad 0 < \delta < 1$$

Innovational Outlier (IO):

An innovational outlier is characterized by an initial impact with effects lingering over subsequent observations. The influence of the outliers may increase as time proceeds.

We consider integer-valued autoregressive models of order one contaminated with innovational outliers. Assuming that the time points of the outliers are known but their sizes are unknown, we prove that Conditional Least Squares (CLS) estimators of the offspring and innovation means are strongly consistent. In contrast, CLS estimators of the outliers' sizes are not strongly consistent. We also prove that the joint CLS estimator of the offspring and innovation means is asymptotically normal. Conditionally on the values of the process at time points preceding the outliers' occurrences, the joint CLS estimator of the sizes of the outliers is asymptotically normal (Capozzoli, et al. 2015).

It is the type of outliers that affects the subsequent observations starting from its position, in other words that occurs as a result of natural randomness. The model, defined as "randomness outlier" in the literature, is shown as follows:

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

Thus, the AO case may be called a gross error model, since only the level of the T'th observation is affected. On the other hand, an IO represents an extraordinary shock at time point T influencing, ,... T T+1 z z through the dynamic system described by ψ(B) =ϕ(B)/ϕ(B) (Chang, Tiao and Chen, 1988).

Unlike an additive outlier, an innovational outlier (IO) is an event whose effect is propagated according to the ARIMA model of the process. In this manner, an IO affects all values observed after its occurrence. In practice, an IO often represents the onset of an external cause. The model for the observed series can be expressed as

$$Y_t = Z_t + \frac{\theta(B)}{\varphi(B)} w_I\, p_t^{(T)}$$

The above model can also be written as

$$Y_t = \frac{\theta(B)}{\varphi(B)} (a_t + w_I\, p_t^{(T)})$$

As a result, an AO only affects one observation, T Y , while an IO affects all values of T Y for t ≥ T according to the ψ-weights {where ψ(B)= () () B B□□ } of the model. The terminology IO arises because of the representation given in (2.6) as ta is also referred to as innovation. The contaminated series t Y is identical to the original series t Z until t=T ; then t Y , shift up (if I W >0) or down (if I W < 0) by I W units at t=T; after t=T ,this effect fades exponentially at a rate determined by the decay coefficient φ(B).

For t ≥ T, t Y is higher than t Z by tT□□ IW units .The effect of the IO fades until eventually the contaminated series t Y is indistinguishable from the original series t Z .

Level Shift (LS):

A level shift (LS) (sometime known as a level change LC) is an event that affects a series at a given time, and whose effect becomes permanent. A level shift could reflect the change of a process mechanism, the change in a recording device, or a change in the definition of the variable itself. The model for the series the study observes may be represented by

$$Y_t = Z_t + \frac{1}{(1-B)} W_L\, p_t^{(T)}$$

The above representation can also be written as

$$Y_t = Z_t + w_L\, S_t^{(T)}$$

Auto Regressive Moving Average

An ARMA model, or Autoregressive Moving Average model, is used to describe weakly stationary stochastic time series in terms of two polynomials. The first of these polynomials is for autoregression, the second for the moving average (Chen, et al. 2017). The autoregressive-moving average (ARMA) process is the basic model for analyzing a stationary time series. First, though, stationarity has to be defined formally in terms of the behavior of the autocorrelation function (ACF) through World's decomposition. Several simple cases of the ARMA model are then introduced and analyzed, with the partial autocorrelation function (PACF) also being defined, before the general model is introduced. ARMA modelbuilding and estimation may then be developed, and this is done via a sequence of examples designed to demonstrate some of the intricacies of selecting an appropriate model to explain the evolution of an observed time series (Johansen & Nielsen, 2016).

Often this model is referred to as the ARMA(p,q) model; where:

- p is the order of the autoregressive polynomial,
- q is the order of the moving average polynomial.

The equation is given by:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}.$$

Where:

- φ = the autoregressive model's parameters,
- θ = the moving average model's parameters.
- c = a constant,
- ε = error terms (white noise).

As we have remarked, dependence is very common in time series observations. To model this time series dependence, we start with univariate ARMA models. To motivate the model, basically we can track two lines of thinking. First, for a series xt , we can model that the level of its current observations depends on the level of its lagged observations (Li, et al. 2015). For example, if we observe a high GDP realization this quarter, we would expect that the GDP in the next few quarters are good as well. This way of thinking can be represented by an AR model. The AR(1) (autoregressive of order one) can be written as:

$$x_t = \phi x_{t-1} + \epsilon_t$$

We introduced the ARMA model that may be written as:

$$\varphi(B) Z_t = C + \theta(B) a_t$$

The model of equation above can be directly extended to include differencing operators to induce stationarity and to encompass seasonal terms (as multiplicative AR or MA operators). To facilitate our understanding of outliers, we will concentrate our discussions to nonseasonal models. Moreover, we will assume C=0 so that we may re-write as:

$$Z_t = \frac{\theta(B)}{\varphi(B)} a_t$$

In the above model, t z represents a series that is not contaminated with outliers. We will use t Y to represent the values observed for t z in the presence of an outlier. As we will see, our representation for an outlier will take the form of the intervention model. The AR operator (and the differencing operator if exists) is placed in the denominator of the ARIMA model. Therefore the effect of an outlier is relative to t Y , rather than relative to the AR filtered t Y (Reiche, et al. 2015).

We now define and illustrate the types of outliers. These are additive outlier (AO), innovational outlier (IO), level shift (LS), and temporary (or transient) change (TC), and to illustrate the effect of each type of outlier, and how it affects the values of a time series, we assume that we have AR(1), then the following simple AR process is employed:

$$Z_t = \frac{1}{1-\varphi_1 B} a_t$$

(2-1-1) Additive Outlier (AO)

An additive outlier (AO) is an event that affects a series for one time period only. One illustration of an AO is a recording error. For this reason, an additive outlier is sometimes called a gross error. If we assume that an outlier occurs at time t=T, we can represent the series we observe by the model

$$Y_t = Z_t + W_A p_t^{(T)}$$

where () T tp is a pulse function (that is, assumes the value 1 when t=T and is 0 otherwise). The value A W represents the amount of deviation from the “true” value of T Z. Such additive outlier (AO's) affect observations in isolation due to some nonrepetitive events and may occur as a result of measurement errors of economic, political and financial events such as oil shocks, wars, financial crashes and changes in policy regimes.

Outliers detection in time series

1-Likelihood ratio tests:

In practice we don’t know if an AO, LS or IO event has occurred at any time t. We use a hypo study testing procedure to decide if such events have occurred.

$$(\text{AO}) \quad e_t = W_A \pi(B) X_t + a_t$$

$$(\text{LS}) \quad e_t = W_S c(B) X_t + a_t$$

$$(\text{IO}) \quad e_t = W_I X_t + a_t$$

Let HAdenote the alternate hypostudy, A W ≠ 0; Let HS denote the alternate, S W ≠0; and let HI denote the alternate, I W ≠ 0. Tests may be performed with the following likelihood ratio statistical (denoted as L):

$$H_0 \text{ vs. } H_A: L_{A,i} = W_A^* k_A^{-1/2} / \sigma_a$$

$$H_0 \text{ vs. } H_S: L_{S,i} = W_S^* k_S^{-1/2} / \sigma_a$$

$$H_0 \text{ vs. } H_I: L_{I,i} = W_I^* / \sigma_a$$

we are just dividing each estimated * w coefficient by its corresponding standard error [the square root of the variance] given by:

$$\text{var}(W_A^*) = k_A \sigma_a^2$$

$$\text{var}(W_S^*) = k_S \sigma_a^2$$

$$\text{var}(W_I^*) = \sigma_a^2$$

Under the null hypostudy H0, and assuming that both time i and the parameters of the ARIMA model, the statistics LA, LS, and LIare normally distribution with mean zero and variance. In practice, we don’t know the parameters of the ARIMA model in

$$Z_t = \frac{\theta(B)\theta(B^s)}{\varphi(B)\varphi(B^S)\nabla^d \nabla_S^D} a_t$$

Methods of Outlier Detection

Outlier detection when ARMA parameters are

It is natural to consider the residuals of a fitted model for use in detecting outliers in a time series, since most diagnostic checks of a model are based on residuals. However, outliers in a time series can affect both the model we may identify for the series as well as the parameter estimates of the identified model. As a result, it is unclear how useful the residuals may be for outlier detection in certain situations. To better understand how a single outlier manifests itself in the residual series, consider the filtered series (Zhang, et al. 2016).

$$e_t = \pi(B)Y_t$$

where () B □ is the polynomial operator in the π-weights of the ARIMA model. The weights in π(B) may be obtained by equating coefficients in the backshift operator in an expression involving π(B) and the polynomial operators of the model. In the case of the non-seasonal stationary model.

$$\theta(B)\,\pi(B) = \varphi(B)$$

The values of t e become the residuals of the fitted model if the πweights are computed from the estimated parameters of the ARIMA model rather than from the known parameters of the “true” ARIMA model.

We may be able to use the analytic representation of t e to test for the effect of an outlier. If only one outlier occurs in a time series, then a least squares estimate for the effect of the outlier at time t=T , ˆi W (i=1,2,3,4), and the statistic that may be used for testing its significance can be easily derived. An adjusted series (i.e., one with the outlier effect removed) can also be obtained. However, some problems remain since:

1. In the event there is an outlier, we do not know its type;
2. We do not know whether an outlier occurs, and if it occurs, the time of its occurrence;
3. There may be more than one outlier present in the series; and
4. We do not know precisely what the “true” underlying model is, nor are we sure of the accuracy of the estimates of a correct model.

Procedures to account for (1) - (3) above have been developed during the past few years. Most of these outlier detection procedures are based on the residuals from fitted models. In this way, we can diagnostically check a fitted model for the presence of outliers.

An iterative detection procedure

Suppose there is unknown number of AO, LS and IO events in a time series t Y , occurring at unknown times t= 12 , ii,… . A detection procedure is as follows:

1. Identify and estimate an ARIMA model (or DR model) forecast t Y assuming that no AO, LS, or IO events are present.
2. Compute the model residuals (ˆte) and estimate 2 a □ as:

$$\hat{\sigma}_a^2 = m^{-1} \sum_{i=1+n_1}^{n} \hat{e}_t^2$$

where m is the number of residuals available (m=n- 1 n and 1n = p+ S P + d+ S D)

3. Compute the likelihood ratios. Set 0,tˆ L equal to the largest of these statistics; that is, 0,t ˆ L = max { A,t ˆ L ,s,tˆ L , I,t ˆ L }for the m time periods t=1+ 1 n , 2+ 1 n ,….,n. 4. Find ˆ L =max { 0,t ˆ L }. Compare ˆ L with a predetermined critical value dc (discussed later). If ˆ L ≤ d c , stop the procedure. If ˆ L > d c ,then a possible AO, LS, or IO is detected. At the time (t = i), type (AO, LS, or IO), and estimated w coefficient of the identified possible event are those associated with ˆ L .
4. Find ˆ L =max { 0,t ˆ L }. Compare ˆ L with a predetermined critical value dc (discussed later). If ˆ L ≤ d c , stop the procedure. If ˆ L > d c ,then a possible AO, LS, or IO is detected. At the time (t = i), type (AO, LS, or IO), and estimated w coefficient of the identified possible event are those associated with ˆ L .

a- If a possible LS is detected, its size is estimated by ˆ SW in * SW = s k () C F i e = s k remove this LS effect from the residual series by replacing each ˆ te with ˆ te - ˆS W () ˆ t BXc for t ≥ i. Reestimate 2 a□ using the new ˆ te series; use this new estimate to recomputed S,t ˆ L.

c- If a possible IO is detected, its effect is estimated by ˆI W according to * IW = Remove this IO effect from the residual series by replacing ˆ teattime t = i with ˆ te - ˆI W =0. Re estimate 2 a using the new ˆ te series; use this new estimate to recompute I,t ˆ L.

5. Suppose T possible AO,LS or IO effects are found at times i1,i2,…., Ti . Treat these times as known and estimate the w coefficients for each effect simultaneously within a DR model. For example, suppose we find T= 3 effects, with a possible AO detected at time t = i3 . Then we estimate the model

$$Y_t = \omega_A X_{1,t} + \omega_S X_{2,t} + \frac{\theta(B)\theta(B^s)}{\varphi(B)\varphi(B^s)\nabla^d\nabla_s^D}(\omega_I X_{3,t} + a_t)$$

Where 1,t X =1 at t = i1 and 1,t X =0 otherwise; 2,t X =0 for t <i2 and 2,tX =1 for t ≥ i2 ; 3,t X =1 at t=i3 and 3,t X =0 otherwise. The model may also call for a constant term. Diagnostic checking may lead to us to modify.

III. METHODS AND FINDINGS

Collection of data

The researcher gathered data for the application of a research from the Dokan dam in Sulaimaniah city, where the data are the volume of water inputting the reservoir of Dokan dam(daily rates cubic meters) late 2018 and early 2019,the very large volume of data has been converted to monthly averages (cubic meters) time series.

Building ARIMA model

Model identification

A time series plot of volume, the study is certain that the series does not have a fixed mean level and not stable in the variance. First to stabling the variance, we transform the data by using the natural logarithmic. We will store the transformed data under the name Lvolume, by using SCA paragraph.A time series plot of Lvolume, furthermore the new series still exhibits a trend and seasonality, but we seem to have stabilized the variability over the length of the series (as seen in figure 1).

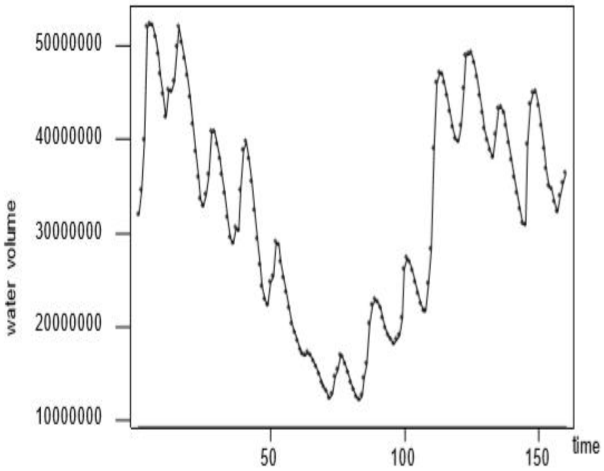


Fig.1-Plot of water volume series of Dokan dam

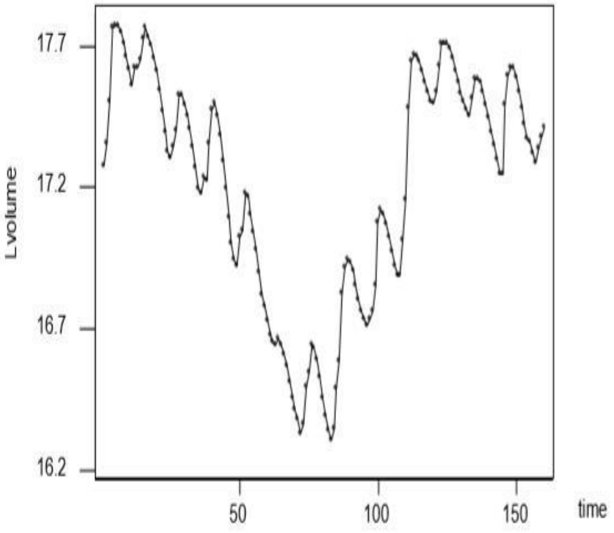
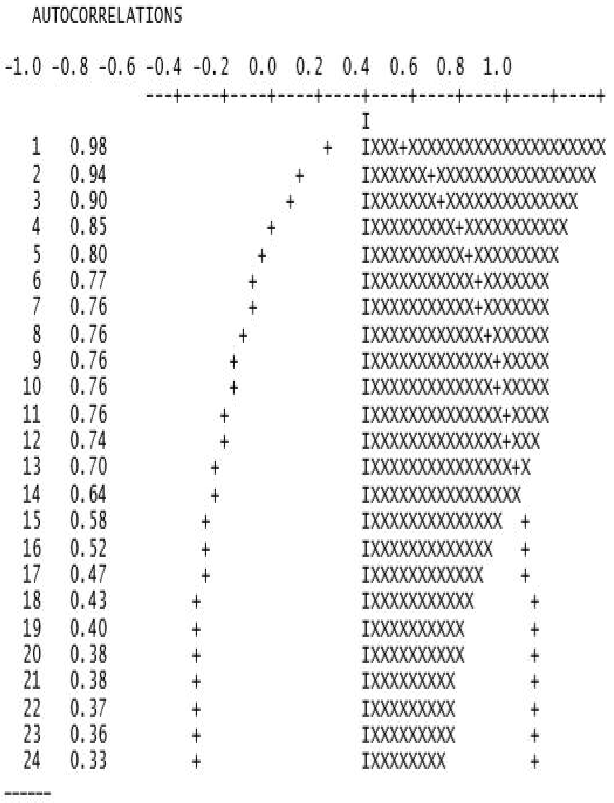


Fig.2-Plot of log of water volume (Lvolume) ofDokan dam

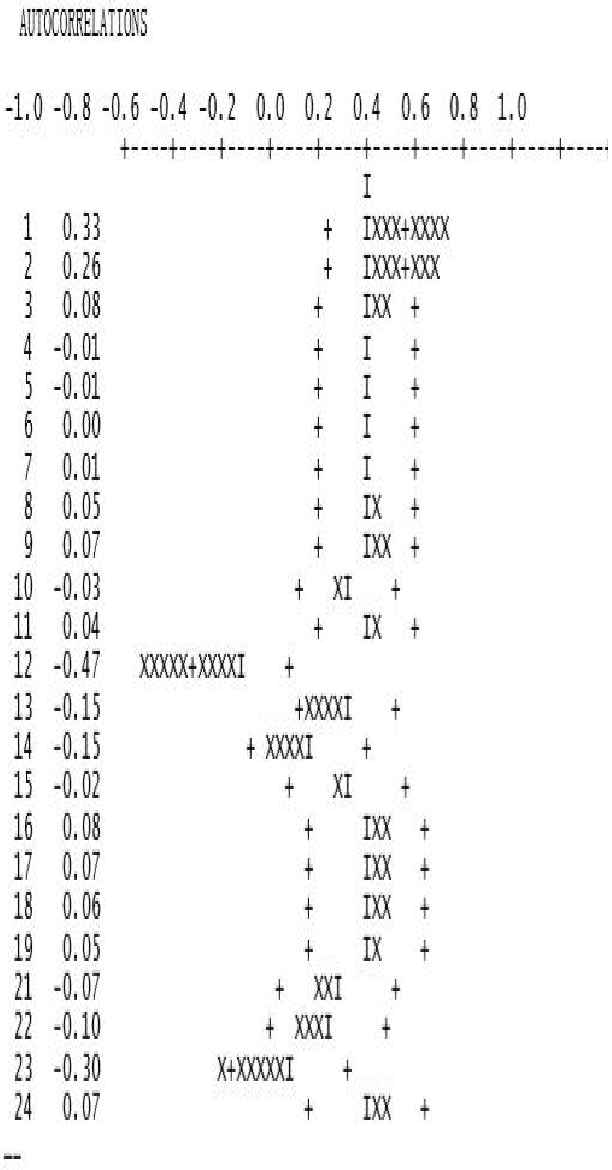
We expect that the Lvolume is not stationary. This is confirmed when we compute and display the sample ACF of the series, by using ACF paragraph of SCA system.



```
   AUTOCORRELATIONS

-1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
              ---+----+----+----+----+----+----+----+----+----+
                                     I
  1   0.98                       +   IXXX+XXXXXXXXXXXXXXXXXXXXX
  2   0.94                    +      IXXXXXX+XXXXXXXXXXXXXXXXX
  3   0.90                   +       IXXXXXXX+XXXXXXXXXXXXXX
  4   0.85                 +         IXXXXXXXXX+XXXXXXXXXXX
  5   0.80                +          IXXXXXXXXXX+XXXXXXXXX
  6   0.77               +           IXXXXXXXXXXX+XXXXXXX
  7   0.76               +           IXXXXXXXXXXX+XXXXXXX
  8   0.76              +            IXXXXXXXXXXXX+XXXXXX
  9   0.76             +             IXXXXXXXXXXXXX+XXXXX
 10   0.76             +             IXXXXXXXXXXXXX+XXXXX
 11   0.76            +              IXXXXXXXXXXXXXX+XXXX
 12   0.74            +              IXXXXXXXXXXXXXX+XXX
 13   0.70           +               IXXXXXXXXXXXXXXX+X
 14   0.64           +               IXXXXXXXXXXXXXXXX
 15   0.58          +                IXXXXXXXXXXXXXX  +
 16   0.52          +                IXXXXXXXXXXXXX   +
 17   0.47          +                IXXXXXXXXXXXX    +
 18   0.43         +                 IXXXXXXXXXXX      +
 19   0.40         +                 IXXXXXXXXXX       +
 20   0.38         +                 IXXXXXXXXXX       +
 21   0.38         +                 IXXXXXXXXX        +
 22   0.37         +                 IXXXXXXXXX        +
 23   0.36         +                 IXXXXXXXXX        +
 24   0.33         +                 IXXXXXXXX         +
------
```

Fig.3-Estimate of ACF for the Lvolume .

The ACF has a slow die-out pattern that is indicative of a nonstationary series. Differencing is required. However, because the data is seasonal, the study may wonder if the "proper" differencing operator is (1-B) or (1- 12 B). We can examine the sample ACF by using both of these differencing operators. The output is edited for presentation purposes as shown below. -- >ACF LVOLUME. DFORDERS 1 12.



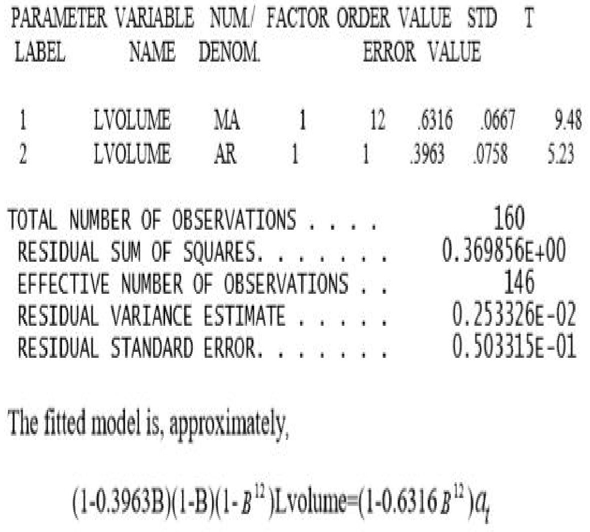
```
AUTOCORRELATIONS

-1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
          +----+----+----+----+----+----+----+----+----+-----
                                   I
 1   0.33                      +   IXXX+XXXX
 2   0.26                      +   IXXX+XXX
 3   0.08                     +    IXX  +
 4  -0.01                     +    I    +
 5  -0.01                     +    I    +
 6   0.00                     +    I    +
 7   0.01                     +    I    +
 8   0.05                     +    IX   +
 9   0.07                     +    IXX  +
10  -0.03                   +   XI    +
11   0.04                     +    IX   +
12  -0.47     XXXXX+XXXXI    +
13  -0.15                   +XXXXI    +
14  -0.15              + XXXXI     +
15  -0.02                  +   XI     +
16   0.08                    +     IXX   +
17   0.07                    +     IXX   +
18   0.06                    +     IXX   +
19   0.05                    +     IX    +
21  -0.07                 +   XXI     +
22  -0.10                + XXXI     +
23  -0.30           X+XXXXXI     +
24   0.07                    +     IXX   +
--
```

Fig.4-Estimate of ACF for differenced Lvolume (d=1,D=1).

Model estimation

The study estimates the volume model by using ESTIM paragraph as:

Table 1-Summary of estimate time series for Lvolume



```
PARAMETER VARIABLE  NUM/  FACTOR ORDER VALUE  STD    T
LABEL      NAME    DENOM.              ERROR  VALUE

1        LVOLUME    MA      1    12   .6316  .0667   9.48
2        LVOLUME    AR      1     1   .3963  .0758   5.23

TOTAL NUMBER OF OBSERVATIONS . . . .          160
 RESIDUAL SUM OF SQUARES. . . . . . .     0.369856E+00
 EFFECTIVE NUMBER OF OBSERVATIONS . .         146
 RESIDUAL VARIANCE ESTIMATE . . . . .     0.253326E-02
 RESIDUAL STANDARD ERROR. . . . . . .     0.503315E-01
```

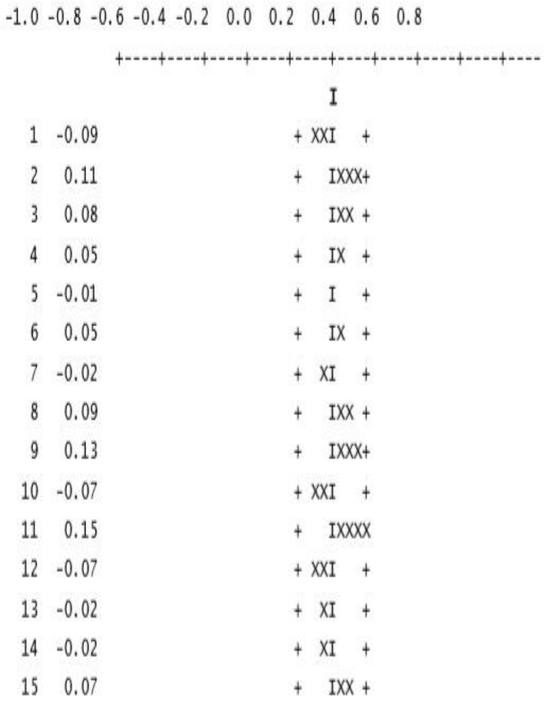
The fitted model is, approximately,

$$(1-0.3963B)(1-B)(1-B^{12})\text{Lvolume}=(1-0.6316B^{12})a_t$$

Parameters estimates are significant based on their t-values.

Diagnostic check of model

A time plot of the residual series does not reveal any gross abnormalities, although some unusual points appear to be present. We can compute and display 24 lags of the sample ACF of the residuals. We see the sample ACF of the residuals is "clean". The output is edited for presentation purposes.



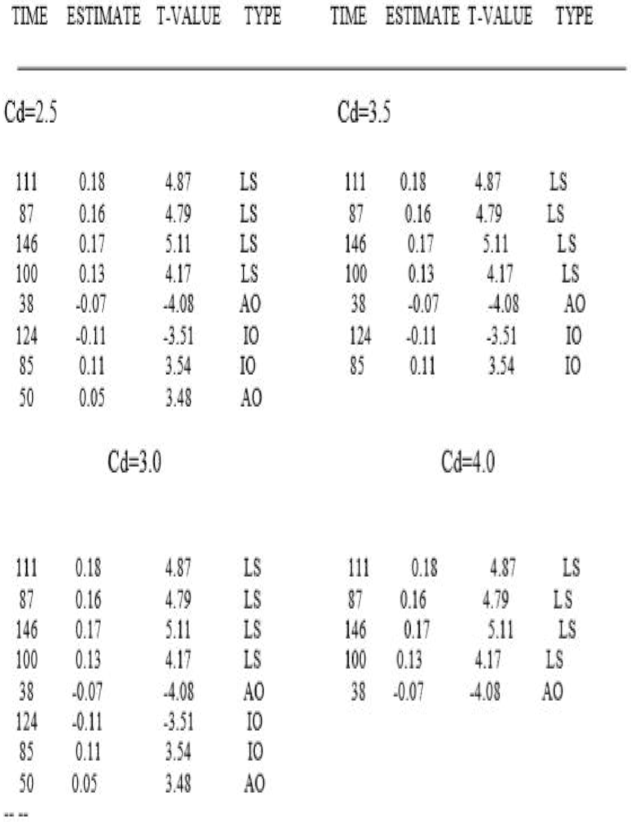
```
-1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8
          +----+----+----+----+----+----+----+----+----+----
                                   I
 1  -0.09                      + XXI    +
 2   0.11                      +   IXXX+
 3   0.08                      +   IXX +
 4   0.05                      +   IX  +
 5  -0.01                      +   I   +
 6   0.05                      +   IX  +
 7  -0.02                      +  XI   +
 8   0.09                      +   IXX +
 9   0.13                      +   IXXX+
10  -0.07                      + XXI   +
11   0.15                      +   IXXXX
12  -0.07                      + XXI   +
13  -0.02                      +  XI   +
14  -0.02                      +  XI   +
15   0.07                      +   IXX +
```

Fig.5-The ACF plot for the residuals of suggested model.

The detection of outliers when the parameters Unknown

To demonstrate outlier detection, the study used the OUTLIER paragraph of SCA system for Lvolume time series. The study obtained the following estimates for model parameters and outliers at different critical values (2.5, 3.0, 3.5 ,4.0) for outlier detection, as seen in table(2).

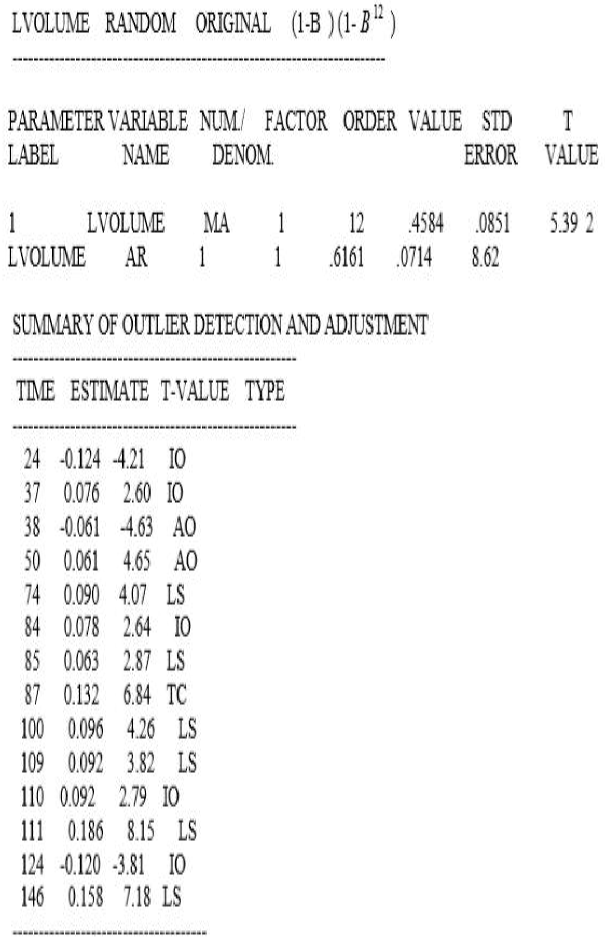
Table 2-The estimates of outliers and it types with different critical values



TIME	ESTIMATE	T-VALUE	TYPE	TIME	ESTIMATE	T-VALUE	TYPE
Cd=2.5				Cd=3.5			
111	0.18	4.87	LS	111	0.18	4.87	LS
87	0.16	4.79	LS	87	0.16	4.79	LS
146	0.17	5.11	LS	146	0.17	5.11	LS
100	0.13	4.17	LS	100	0.13	4.17	LS
38	-0.07	-4.08	AO	38	-0.07	-4.08	AO
124	-0.11	-3.51	IO	124	-0.11	-3.51	IO
85	0.11	3.54	IO	85	0.11	3.54	IO
50	0.05	3.48	AO				
Cd=3.0				Cd=4.0			
111	0.18	4.87	LS	111	0.18	4.87	LS
87	0.16	4.79	LS	87	0.16	4.79	LS
146	0.17	5.11	LS	146	0.17	5.11	LS
100	0.13	4.17	LS	100	0.13	4.17	LS
38	-0.07	-4.08	AO	38	-0.07	-4.08	AO
124	-0.11	-3.51	IO				
85	0.11	3.54	IO				
50	0.05	3.48	AO				

The study illustrates that the number of outliers decrease whenever critical values increase. Alternatively,we could have estimated model Lvolume using the OESTIM paragraph. In this way the SCA System will simultaneously detect outliers and jointly estimate their effects with the parameter.When critical value equal to 2.5 as seen in table (3).

Table 3-Summary of estimate time series model for volume (cd=2.5)



LVOLUME RANDOM ORIGINAL (1-B)(1- B^{12})

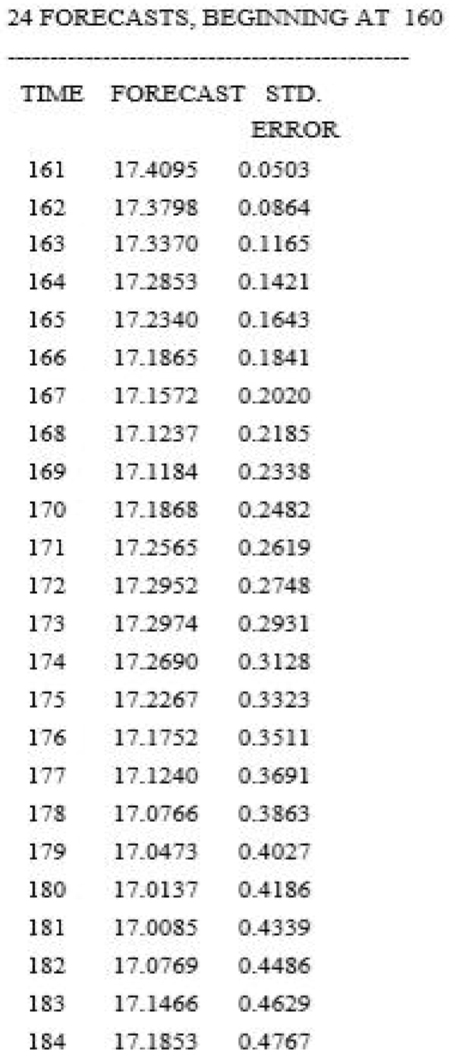
PARAMETER LABEL	VARIABLE NAME	NUM./ DENOM.	FACTOR	ORDER	VALUE	STD ERROR	T VALUE
1	LVOLUME	MA	1	12	.4584	.0851	5.39 2
LVOLUME	AR	1	1	.6161	.0714	8.62	

SUMMARY OF OUTLIER DETECTION AND ADJUSTMENT

TIME	ESTIMATE	T-VALUE	TYPE
24	-0.124	-4.21	IO
37	0.076	2.60	IO
38	-0.061	-4.63	AO
50	0.061	4.65	AO
74	0.090	4.07	LS
84	0.078	2.64	IO
85	0.063	2.87	LS
87	0.132	6.84	TC
100	0.096	4.26	LS
109	0.092	3.82	LS
110	0.092	2.79	IO
111	0.186	8.15	LS
124	-0.120	-3.81	IO
146	0.158	7.18	LS

The OFORECAST paragraph extends the outlier detection and adjustment capabilities of the SCA System to the forecasting of a time series in the presence of outliers. Unlike other forecasting capabilities that simply utilize the current parameter estimates and the data on hand to compute forecasts, the OFORECAST paragraph also performs its own outlier detection and adjustment. As a result, it provides us with:

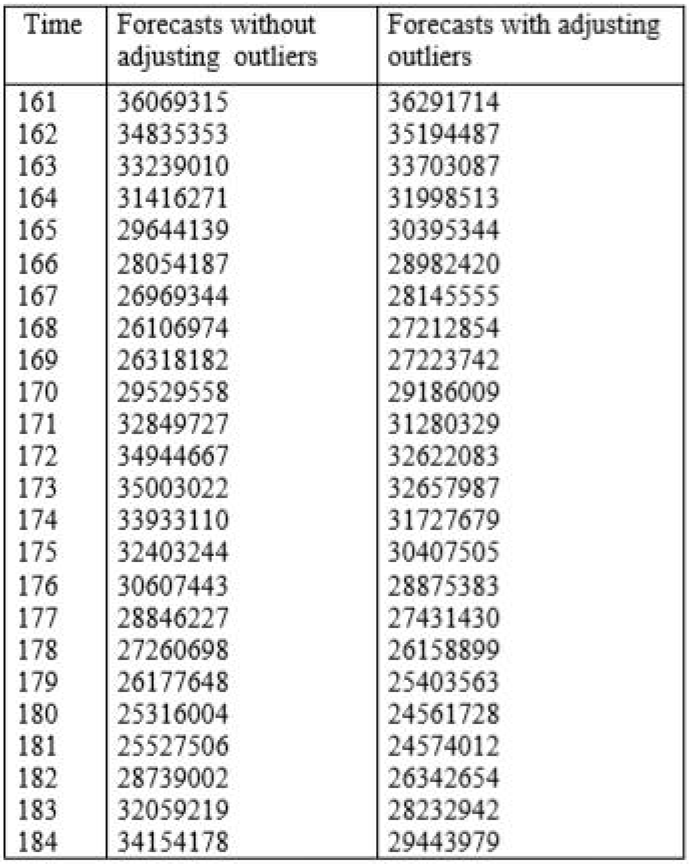
Table 4-Forecasts for Volume after adjusting the outliers (Cd=2.5).



```
24 FORECASTS, BEGINNING AT  160
---------------------------------------------
 TIME   FORECAST   STD.
                  ERROR
  161   17.4095   0.0503
  162   17.3798   0.0864
  163   17.3370   0.1165
  164   17.2853   0.1421
  165   17.2340   0.1643
  166   17.1865   0.1841
  167   17.1572   0.2020
  168   17.1237   0.2185
  169   17.1184   0.2338
  170   17.1868   0.2482
  171   17.2565   0.2619
  172   17.2952   0.2748
  173   17.2974   0.2931
  174   17.2690   0.3128
  175   17.2267   0.3323
  176   17.1752   0.3511
  177   17.1240   0.3691
  178   17.0766   0.3863
  179   17.0473   0.4027
  180   17.0137   0.4186
  181   17.0085   0.4339
  182   17.0769   0.4486
  183   17.1466   0.4629
  184   17.1853   0.4767
```

We can converse the forecasting value of Lvolume to the origin values before taking the natural logarithmic. Then compare these values with the forecasts by using the fitted ARIMA model, assuming that the outliers are not presence, the results shown as in the table (3.16) below by using critical value (4.0).

Table 5-Forecasts for the volume data with and without adjusting the outliers



Time	Forecasts without adjusting outliers	Forecasts with adjusting outliers
161	36069315	36291714
162	34835353	35194487
163	33239010	33703087
164	31416271	31998513
165	29644139	30395344
166	28054187	28982420
167	26969344	28145555
168	26106974	27212854
169	26318182	27223742
170	29529558	29186009
171	32849727	31280329
172	34944667	32622083
173	35003022	32657987
174	33933110	31727679
175	32403244	30407505
176	30607443	28875383
177	28846227	27431430
178	27260698	26158899
179	26177648	25403563
180	25316004	24561728
181	25527506	24574012
182	28739002	26342654
183	32059219	28232942
184	34154178	29443979

And we note that the Mse(0.052905) forecast without adjusting the outlier is greater than the Mse (0.043100) of forecasting with adjusting the outlier. This means that when analyzing the data of time series, first we must detect and adjust the outliers.

IV. CONCLUSIONS

The study came to conclude that throughout the research, the following conclusions: first, every time increasing the critical value, the value of residual standard error (with outlier adjustment) increased. Second, every time increasing the critical value, the number of outlier values decreased. Third, in the case of presence of outliers the forecasts with adjustment of outliers better than the forecasts without adjusting outliers.

Since the procedures are based on simple techniques, they are widely applicable. For instance, they can be used as data screening device in spectral density estimation and in robust time series analysis. They can also be used in biological study where exogenous disturbances are unavoidable. For example, Greenhouse, Kass and Tsay (1987) analysed body temperature of an individual involved in a psychiatric study where the observations clearly depended on the individual physical activities. A variance change from day to night seems highly plausible. A third application of the procedures is that they can be used to identify the time point of an intervention in the intervention analysis of Box and Tiao (1975). In the traditional intervention analysis, the time point of an intervention is assumed to be known. Finally, two remarks are made on the procedures. First, in Section 4 the adjusted series was used in the detection process to demonstrate the usefulness of the suggested procedure. This, however, does not imply that one can rely on the adjusted series to make inferences. A more appropriate strategy would be (a) to search for the causes of the identified outliers, level changes and variance changes, (b) to specify a general model in the form of (2) based on causes of the exogenous disturbances, and (c) to estimate jointly the impact of disturbances and the time series parameters. This strategy allows for the use of prior information of the disturbances. It can also reduce the possibility of over parameterization that arises from the abuse of the detection procedure. Readers are referred to Tsay (1986a) for further discussion. Second, to detect the transient level change, 6 = 0.8 was used in Section 4. In fact, other values of 6 can also be used. As an example, 6 = 0.6 was used to the air-passenger-miles data of Example 1. The procedure still identified the same time points as significant disturbances even though some of the classifications between permanent and transient level changes are different. Similarly, to detect the variance changes, h = 30 was used to compute residual variances at both ends of a series. The choice of h is not critical as long as it is reasonable. For instance, the same detection results were obtained in Example 2 when h = 20 was used. In general, a h between 20 and 30 appears to be useful.

The study supports the claim that outliers do result in model misspecification as they affect the autocorrelation structure of any time series. In our case it is illustrated by the fact that initially we had the ARIMA (1 1 0) *(0 0 1) 12 as the best model that could be fitted to our data. Testing the residuals for normality and constant variance showed that both assumptions were violated although the parameters in the model were significant. Using this model for forecasts would have given misleading figures for a decision maker. This is possibly attributed to the presence of outliers. The best model was found to be ARIMA (1 1 2) *(0 0 1) 12 after correcting the series for outliers and all the parameters were significant in the model. Diagnostic checks also showed that the assumptions of normality and constant variance were not violated. This therefore demonstrates that the procedure is useful in detecting and correcting for outliers. It can be applied to all invertible ARIMA models. Moreover, it is flexible and easy to interpret. The procedure must be used with other diagnostic tools for time series to produce even better results. Further study is needed to investigate the variances and other sampling properties of the resulting parameter estimates. The message from this study is that when examining economic time series data any potential outliers should be taken seriously, no matter what the ultimate aim or the model used may be. Outliers have already been shown to be potentially harmful, and there is also increasing evidence that the dangers are not only theoretical. Other possible models that might be useful for modelling time series must be explored such as GARCH and ARCH models. These are non-linear forms of time series that might be used to model data that has got a lot of fluctuations in it. Non-linearity tests are normally done on the data before the previous models can be applied. The study suggested for future studies the following: Studying the methods of detection outliers in multivariate time series and application. Studying the detection outlier when occurs at the end of the series, and finally studying the detection outlier when presence the missing data in the series.

REFERENCES

[1] Ahmad, S., & Purdy, S. (2016). Real-time anomaly detection for streaming analytics. *arXiv preprint arXiv:1607.02480*.

[2] Aminikhanghahi, S., & Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, *51*(2), 339-367.

[3] Arumugam, P., & Saranya, R. (2018). Outlier Detection and Missing Value in Seasonal ARIMA Model Using Rainfall Data. *Materials Today: Proceedings*, *5*(1), 1791-1799.

[4] Cabrieto, J., Tuerlinckx, F., Kuppens, P., Grassmann, M., &Ceulemans, E. (2017). Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods. *Behavior*

[5] Capozzoli, A., Lauro, F., & Khan, I. (2015). Fault detection analysis using data mining techniques for a cluster of smart office buildings. *Expert Systems with Applications*, *42*(9), 4324-4338.

[6] Chen, W., Zhou, K., Yang, S., & Wu, C. (2017). Data quality of electricity consumption data in a smart grid environment. *Renewable and Sustainable Energy Reviews*, *75*, 98-105.

[7] Filonov, P., Lavrentyev, A., &Vorontsov, A. (2016). Multivariate industrial time series with cyber-attack simulation: Fault detection using anlstm-based predictive data model. *arXiv preprint arXiv:1612.06676*.

[8] Frantz, D., Röder, A., Udelhoven, T., & Schmidt, M. (2015). Enhancing the detectability of clouds and their shadows in multitemporal dryland Landsat imagery: Extending Fmask. *IEEE Geoscience and Remote Sensing Letters*, *12*(6), 1242-1246.

[9] Ganz, F., Puschmann, D., Barnaghi, P., &Carrez, F. (2015). A practical evaluation of information processing and abstraction techniques for the internet of things. *IEEE Internet of Things journal*, *2*(4), 340-354.

[10] Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., & Hobart, G. W. (2015). An integrated Landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites. *Remote Sensing of Environment*, *158*, 220-234.

[11] Johansen, S., & Nielsen, B. (2016). Asymptotic theory of outlier detection algorithms for linear time series regression models. *Scandinavian Journal of Statistics*, *43*(2), 321-348.

[12] Kontaki, M., Gounaris, A., Papadopoulos, A. N., Tsichlas, K., &Manolopoulos, Y. (2016). Efficient and flexible algorithms for monitoring distance-based outliers over data streams. *Information systems*, *55*, 37-53.

[13] Li, L., Das, S., John Hansman, R., Palacios, R., & Srivastava, A. N. (2015). Analysis of flight data using clustering techniques for detecting abnormal operations. *Journal of Aerospace information systems*, *12*(9), 587-598.

[14] Liu, M., Shi, J., Cao, K., Zhu, J., & Liu, S. (2018). Analyzing the training processes of deep generative models. *IEEE transactions on visualization and computer graphics*, *24*(1), 77-87.

[15] Liu, S., Wright, A., &Hauskrecht, M. (2018). Change-point detection method for clinical decision support system rule monitoring. *Artificial intelligence in medicine*, *91*, 49-56.

[16] Liu, Z., Verstraete, M. M., & de Jager, G. (2018). Handling outliers in model inversion studies: a remote sensing case study using MISR-HR data in South Africa. *South African Geographical Journal*, *100*(1), 122-139.

[17] Loureiro, D., Amado, C., Martins, A., Vitorino, D., Mamade, A., & Coelho, S. T. (2016). Water distribution systems flow monitoring and anomalous event detection: A practical approach. *Urban Water Journal*, *13*(3), 242-252.

[18] Martí, L., Sanchez-Pi, N., Molina, J., & Garcia, A. (2015). Anomaly detection based on sensor data in petroleum industry applications. *Sensors*, *15*(2), 2774-2797.

[19] Reiche, J., Verbesselt, J., Hoekman, D., & Herold, M. (2015). Fusing Landsat and SAR time series to detect deforestation in the tropics. *Remote Sensing of Environment*, *156*, 276-293.

[20] Rousseeuw, P. J., &Bossche, W. V. D. (2018). Detecting deviating data cells. *Technometrics*, *60*(2), 135-145.

[21] Rousseeuw, P., Perrotta, D., Riani, M., & Hubert, M. (2019). Robust monitoring of time series with application to fraud detection. *Econometrics and statistics*, *9*, 108-121.

[22] Sprint, G., Cook, D. J., &Schmitter-Edgecombe, M. (2016). Unsupervised detection and analysis of changes in everyday physical activity data. *Journal of biomedical informatics*, *63*, 54-65.

[23] Sprint, G., Cook, D. J., Fritz, R., &Schmitter-Edgecombe, M. (2016). Using smart homes to detect and analyze health events. *Computer*, *49*(11), 29-37.

[24] Staal, O. M., Sælid, S., Fougner, A., &Stavdahl, Ø. (2019). Kalman smoothing for objective and automatic preprocessing of glucose data. *IEEE journal of biomedical and health informatics*, *23*(1), 218-226.

[25] Stumpf, A., Malet, J. P., &Delacourt, C. (2017). Correlation of satellite image time-series for the detection and monitoring of slow-moving landslides. *Remote sensing of environment*, *189*, 40-55.

[26] Wang, B., & Mao, Z. (2018). Detecting Outliers in Electric Arc Furnace under the Condition of Unlabeled, Imbalanced, Non-stationary and Noisy Data. *Measurement and Control*, *51*(3-4), 83-93.

[27] Zhang, Q., Pandey, B., &Seto, K. C. (2016). A robust method to generate a consistent time series from DMSP/OLS nighttime light data. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(10), 5821-5831.