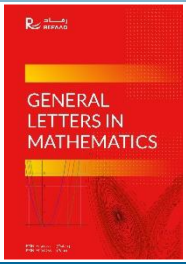
Available online at www.refaad.com
Gen. Lett. Math., 13(2) (2023), 36- 48
Research Article

General Letters in Mathematics (GLM)

Journal Homepage: https://www.refaad.com/Journal/Index/1

ISSN: 2519-9277 (Online) 2519-9269 (Print)



Detection of Outlier in Time Series with Application to Dohuk Dam Using the SCA Statistical System

Shelan Saied Ismaeel[a,*], Kurdistan M.Taher Omar[a], Sameera Abdulsalam Othman[b]

[a] Department of Mathematics, Faculty of Science, University of Zakho, Kurdistan Region, Iraq.
[b] Department of Mathematics, College of Basic Education, University of Duhok, Kurdistan Region, Iraq.

Email: [a] shelan.ismaeel@uoz.edu.krd, [a] kurdistan.taher@uoz.edu.krd, [b] sameera.othman@uod.ac

Abstract

Outliers are data points or observations that stand out significantly from the rest of the group in terms of size or frequency. They are also referred to as "abnormal data". Before fitting a forecasting model, outliers are often eliminated from the data set, or if not removed, the forecasting model is altered to account for the presence of outliers. The first scenario covered in the study is the detection of outliers when the parameters have been established. Second, where there are unidentified parameters. This article mentions a number of causes for outlier correction and detection in time series analysis and forecasting. For the objective of the study, a time series of the volume of water entering the Dohuk dam reservoir in Dohuk city was used. The study arrived at the following conclusions after conducting their research: first, whenever the critical value increased, the value of residual standard error (with outlier adjustment) increased. Second, the quantity of outlier values dropped each time the critical value was raised. Third, forecasts with outlier correction perform better than forecasts without outlier adjustment when outliers are present.

Keywords: Time series, Outlier, ARIMA model, Forecasting.
2010 MSC: 62F40, 62J02, 62J05.

1. Introduction

Outlier research is not a recent development. Actually, it has a long history going all the way down to the first analysis of statistics. Along with other statistical techniques, outlier methods have grown. Sadly, the adoption of outlier methods has not been as quick and broad in time series analysis. This must be due, in part, to the fact that time series outlier approaches were first expressly taken into consideration. However, since then, the number of studies addressing the subject has progressively increased [23].

Time series data analysis frequently involves dealing with outliers and structural changes. The existence of such unusual events could easily lead to incorrect results when using the traditional time series analysis approach. Yet, because there aren't many straightforward but effective ways to deal with those events' dynamic behavior in the underlying series, the significance of those occurrences is frequently missed. Considering unified approaches

* Corresponding author
Email addresses: shelan.ismaeel@uoz.edu.krd (Shelan Saied Ismaeel).
doi: https://doi.org/10.31559/glm2023.13.2.1
Received: 5 Feb 2023 Revised: 21 Mar 2023 Accepted: 27 May 2023

for identifying and managing outliers and structural changes in a univariate time series is the main objective of this research. Level shift (LS) and variance change are allowed structural changes (VC), and the outliers that are dealt with are the additive outlier (AO) and the innovative outlier. Transient level change (TC) and permanent level change (LC) are additional categories for level shift [15] and [24].

The purpose of the study project in the first phase of the literature review is to determine the applicability of time series and other methods for estimating missing values and outlier detection/replacement in various transport data.

For example, the failure of automatic counters can result in missing data and outliers [6] and [5]. Early research indicates that such data patching techniques can be simple. A brief questionnaire inquiry form is to be used to contact each local authority separately in an effort to learn more about their current procedures. The research intends to translate recently discovered approaches for handling time series outliers in a transport environment after reviewing existing procedures. A different, more analytical approach to current practices is intended to be shown through comparisons of prospective alternatives [14] and [26].

The literature has addressed a number of strategies for dealing with outliers in a time series. Fox [10] developed two parametric models for examining outliers,[1] used a Bayesian approach, [21] considered outliers as pollution produced from a certain probability distribution. Fox's models were adopted by [7], who then suggested an iterative process for identifying several outliers. This iterative process has been regularly used in recent years with positive outcomes [10]and [16] .

In time series, outliers can appear in a variety of ways. Both additive and innovative outliers exist. An observation that is lower or bigger in value than anticipated is impacted by an additive outlier. An innovational outlier, on the other hand, has an impact on many observations. Level shifts, transitory changes, and variance changes are three further categories of outliers that might be identified [3]. A level shift simply modifies the series' level or mean after a particular observation. In that it has an impact on later observations in addition to having an initial impact similar to that of an additive outlier, a transitory change is a generalization of an additive outlier and a level shift. Simply put, a variance change increases or decreases the observed data's variation [20].

ARMA model forecasts are somewhat impacted by outliers, and those that happen right before the forecast period begins might have very negative effects. Point predictions may only slightly suffer from additive outliers, but because outliers can overestimate the series' estimated variance, the prediction intervals may become seriously deceptive. Impacts even when outliers are far from the forecast zone have an impact on point projections. There have been attempts to create forecasting intervals with outliers present [17].

2. Methodology

Outliers in a time series and their types

- **Additive Outlier**

An additive outlier (AO) is a single-period incident that has an impact on a series. One example of an AO is a recording mistake (e.g., the actual value 2.1 may be recorded as 21.0, 0.1, or the like). A huge mistake is another name for an additive outlier [4] and [5]. The series we see can be represented by the model if we suppose that an outlier happens at time t=T.

$$Y_t = Z_t + W_A p_t^{(T)} \tag{1}$$

Where the pulse function $p_t^{(T)}$ is assumed to be 1 when t=T and 0 otherwise. The extent of departure from the "true" value of Z_T is indicated by the value Z_T.

- **Innovational Outlier**

An innovational outlier (IO), as opposed to an additive outlier, is an occurrence whose impact is propagated in accordance with the ARIMA method framework. This is how an IO has an impact on all variables seen after it happens [8] and [9]. In actuality, an IO frequently denotes the beginning of an external cause. The following is how to express the model for the observed series:

$$Y_t = Z_t + \frac{\theta(B)}{\phi(B)} W_I p_t^{(T)} \tag{2}$$

The above model can also be written as

$$Y_t = \frac{\theta(B)}{\phi(B)} (a_t + W_I \, p_t^{(T)}) \tag{3}$$

By contrasting (3) with (1), we can more effectively understand the distinction between an IO and an AO. As seen in (1), an AO only modifies the observation Z_T, whereas an IO only modifies the shock a_T.

A result of this is that an IO impacts all values of Y_T for t ≥ T in accordance with the model's ψ-weights {where $\psi(B) = \frac{\theta(B)}{\phi(B)}$ }, whereas an AO merely impacts one observation, Y_T.

The depiction in (3) that a_t is also known as innovation [14] and [18] gives rise to the term IO.

- **Level Shift (LS)**

A level shift (LS), sometimes known as a level change (LC), is an occurrence that has an immediate and lasting impact on a series. A level shift could be caused by a modification to a process mechanism, a modification to a recording system, or a modification to the variable's definition [2] and [11]. A possible representation of the model for the series we observe is:

$$Y_t = Z_t + \frac{1}{(1-B)} W_L p_t^{(T)} \tag{4}$$

The above representation can also be written as

$$Y_t = Z_t + W_L s_t^{(T)} \tag{5}$$

Assuming a value of 0 before to t = T and a value of 1 thereafter, $s_t^{(T)}$ is a step function.

It is clear that the models for an AO and LS provided by (1) are equivalent, with the exception that an AO affects Z_t only at t = T (due to the pulse function P_t((T))) while an LS impacts Z_t permanently from t = T onwards (due to the step function $s_t^{(T)}$).

Up until t=T, the new polluted series Y_t is the same as the original series Z_t; after that, Y_t is shifted up (if W_L > 0) or down (if W_L <0) by W_L units for all t ≥T.

- **Temporary Change (TC)**

The two different ways that an event influences a series are represented by an additive outlier (AO) and a level shift (LS). LS impacts the level of the underlying process for all future time whereas an AO just has an impact on the series for that particular time period. Consider an event that initially affects a series, but later has no further effects [12] and [22]. An event with such an initial impact and whose influence decays exponentially in accordance with some dampening factor say δ, such as, is considered a temporary (or transient change) (TC). The observed series can be represented as:

$$Y_t = Z_t + \frac{\mathbf{1}}{\mathbf{1-\delta B}} W_c P_t^{(T)} \qquad 0 < \delta < 1 \tag{6}$$

Techniques for Detecting Outliers

Outlier detection with known ARMA parameters

Since residuals are typically used in model diagnostic checks, it makes sense to take the residuals of a fitted model into account while looking for outliers in a time series. Outliers in a time series, however, may have an impact on both the methods we might choose for the series and the estimation of parameter of the chosen method [13], [19] and [25]. The residuals' potential utility for outlier detection in specific circumstances is therefore unknown. Examine the filtered series to get a clearer idea of how a single outlier appears in the residual series.

$$e_t = \pi(B) Y_t \tag{7}$$

When the ARIMA model's -weights' polynomial operator $\pi(B)$ is used. In an equation using $\pi(B)$ and the model's polynomial operators, the weights in $\pi(B)$ can be determined by equating the coefficients in the backshift operator. The ARMA model of $z_t = \frac{\theta(B)}{\phi(B)} a_t$, which is the nonseasonal stationary model, allows for the computation of these -weights from

$$\theta(B)\pi(B) = \phi(B) \tag{8}$$

The values of e_t turn into the residuals of the fitted model if the weights are computed using the estimated ARIMA method parameters as opposed to the known parameters of the "real" ARIMA model.

Let's imagine the series Y_t. has a single outlier, let's call it an AO at time T. By swapping (1) into (1), we can derive an analytical description of e_t (7). For a single IO, LS, or TC, analogous analytic descriptions can be obtained in a similar way.

The analytical representation of e_t may allow us to test for the impact of an outlier. If there is just one outlier in a time series, it is simple to compute least squares estimate for the effect of the outlier at time t=T, $\widehat{W}_i$ (i=1,2,3,4) and the statistic that may be used to test its significance. Moreover, a series that has had the outlier effect eliminated, or corrected, can be obtained. Yet, some issues continue because:
(1) If an outlier occurs, we don't know when it will happen or even if it will occur at all;
(2) We do not know the type of an outlier if one exists;
(3) The series may contain more than one outlier; and
(4) Neither are we certain of the quality of the estimates produced by a valid model, nor do we know with certainty what the "real" underlying model is.

In recent years, procedures to account for (1) through (3) above have been created. The residuals from fitted models form the basis of the majority of these outlier detection techniques. This allows us to diagnostically examine a fitted model for outliers.

3. Application

Data are acquired for the Dohuk Dam, where the data represent the amount of water entering its reservoir on a daily basis between 1 January 1994 and 30 April 2007 (in cubic meters). The extremely high volume of data has been transformed into a time series of monthly averages (in cubic meters).

Figure (1) displays a time series plot of volume. We can say with certainty that the series' mean level and variance are not fixed. We first use the natural logarithmic to modify the data and stagger the variance. With the help of the SCA system, we stored the converted data as Lvolume.
-->Lvolume = ln(volume)

Figure (2) displays a Lvolume time series graphic. The new series still has seasonality and a trend, but we appear to have stabilized the variability over the series' whole duration.

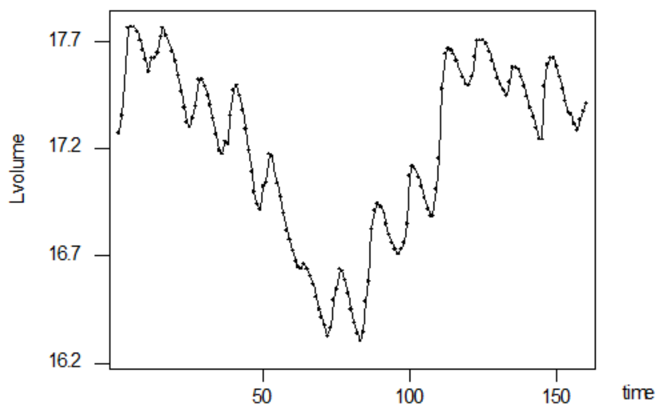


Figure (1): Series of Dohuk dam's water volume plots

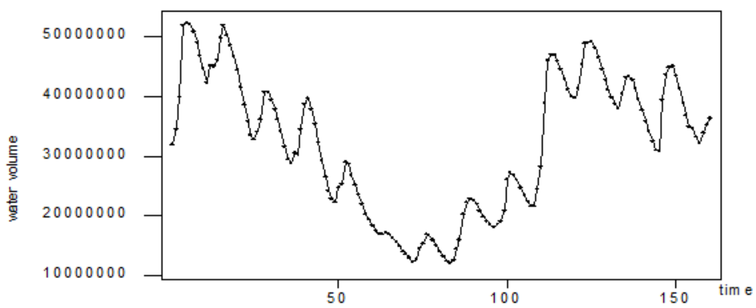


Figure (2): Plot of the Dohuk dam's log of water volume (Lvolume)

The Lvolume is not anticipated to be stationary. This is supported by the example ACF (Autocorrelation Function) of the series that we compute and present using the SCA system's ACF paragraph (shown in Figure (3)).

-->ACF LVOLUME.

NAME OF THE SERIES . . . . LVOLUME
TIME PERIOD ANALYZED . . . . 1 TO 160
STANDARD DEVIATION OF THE SERIES . . . 0.4108

```
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   +
```

Figure (3): Estimate of ACF for the Lvolume

The delayed die-out pattern of the ACF is a sign of a nonstationary series. Differentiating is necessary. However, due to the seasonality of the data, we can question whether (1-B) or (1-B^{12}) is the "correct" differencing operator. By employing both of these differencing operators, we may analyze the sample ACF. For presentational considerations, the result is altered as displayed below.

-->ACF LVOLUME. DFORDERS 1 12.

DIFFERENCE ORDERS. . . . . (1-B) (1-B^{12})
NAME OF THE SERIES . . . . LVOLUME
TIME PERIOD ANALYZED . . . . 1 TO 160
STANDARD DEVIATION OF THE SERIES . . . 0.0645

AUTOCORRELATION

```
-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  +
```

Figure (4): Estimate of ACF for differenced Lvolume (d=1,D=1).

We note from Figure (4) that ACF has a quick die-out pattern, that is indicative of a stationary series. Then to identify the appropriate model, we compute Partial Autocorrelation Function (PACF) for differenced Lvolume, by using PACF paragraph of SCA system as:

-->PACF LVOLUME. DFORDERS 1 12.

DIFFERENCE ORDERS . . . .(1-B)$(1\text{-}B^{12})$
NAME OF THE SERIES . . . . .LVOLUME
STANDARD DEVIATION OF THE SERIES . . . 0.0645

```
PARTIAL 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   0.18                   +  IXXXX
  3  -0.06                   + XI  +
  4  -0.07                   + XXI  +
  5   0.01                   +  I  +
  6   0.02                   +  IX  +
  7   0.01                   +  I  +
  8   0.05                   +  IX  +
  9   0.04                   +  IX  +
 10  -0.09                   + XXI  +
 11   0.05                   +  IX  +
 12  -0.55            XXXXXXX+XXXI  +
 13   0.19                   +  IXXX+X
 14   0.08                   +  IXX +
 15   0.05                   +  IX  +
 16   0.07                   +  IXX +
 17  -0.02                   +  I  +
 18   0.01                   +  I  +
 19  -0.01                   +  I  +
 20   0.03                   +  IX  +
 21  -0.04                   + XI  +
 22  -0.17                   XXXXI  +
 23  -0.16                   XXXXI  +
 24   0.00                   +  I  +
```

Figure (5): Estimated PACF for the differences of the Lvolume (d=1,D=1).

From Figure (4) and Figure (5) we can deduce that the appropriate model is ARIMA $(1,1,0)\times(0,1,1)_{12}$. Using the ESTIM model, we estimate the volume method as:

-->ESTIM UTSMODEL.

Table (1): Summary of Lvolume estimation time series

```
 LVOLUME   RANDOM   ORIGINAL   (1-B  ) (1-B^12)
------------------------------------------------------------

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 approximate fit model is,
(1-0.3963B)(1-B)(1-B^{12}) Lvolume=(1-0.6316B^{12})a_t
The t-values of the parameter estimations indicate their significance.
The residual series' time plot does not show any obvious anomalies, but some odd spots do appear to be there. The sample ACF of the residuals can be computed and displayed over 24 delays. We can see that the sample ACF for the residuals is "clean". The product is modified for presentation.

-> ACF RESAD. MAXLAG 24.

```
        .8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8
          +----+----+----+----+----+----+----+----+----
 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 +
```

Figure (6): The ACF plot for the suggested model's residuals

We can conclude that the ARIMA(1,1,0)× $(0,1,1)_{12}$ model is the proper one for Lvolume.

4. The detection of outliers when the parameters Unknown

We use the OUTLIER paragraph of the SCA system for Lvolume time series to demonstrate outlier detection. The following estimates for model parameters and outliers are found at different critical values (2.5, 3.0, 3.5, and 4.0) for outlier detection, as shown in Table (2).

Table (2): Outlier estimates and their types with various 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				

When critical values rise, we observe a decrease in the number of outliers. Alternately, we might have calculated model Lvolume using OESTIM, which is designed to estimate time series models when outliers are present. By doing this, the SCA System is able to identify outliers and estimate their effects in conjunction with the parameter. when the crucial value is 2.5.

- -> OESTIM UTSMODEL. OSTOP CRIT (2.5).

Table (3): OESTAMITE TIME SERIES DESIGN FOR VOLUME (Cd=2.5) classification

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

PARAMETER VARIABLE NUM./ FACTOR ORDER VALUE STD T

LABEL NAME DENOM. ERROR VALUE

1 LVOLUME MA 1 12 .4584 .0851 5.39 2 2

LVOLUME AR 1 1 .6161 .0714 8.62

OUTLIER DETECTION AND ADJUSTMENT SUMMARY

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

TOTAL NUMBER OF OBSERVATIONS. 160

EFFECTIVE NUMBER OF OBSERVATIONS. 146

RESIDUAL STANDARD ERROR (WITHOUT OUTLIER ADJUSTMENT)..0.552113E-01

RESIDUAL STANDARD ERROR (WITH OUTLIER ADJUSTMENT) . . . 0.293978E-01

By using the OFORECAST clause, time series forecasting with outliers is made possible by the SCA System's capabilities for outlier identification and compensation. The OFORECAST section additionally does its own outlier detection and adjustment in contrast to conventional forecasting capabilities that merely use the current parameter estimations and the data at hand to compute forecasts. As a result, it gives us:

- -> OFORECAST UTSMODEL. OSTOP CRIT(2.5).

Table (4): Forecasts for Lvolume 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

Before obtaining the natural logarithmic, Lvolume's forecasted value can be changed into its original values. If there are no outliers, then use the fitted ARIMA model to compare these values to the forecasts with the findings shown in table (5) using the critical value method (4.0).

Table (5): estimates for the volume data, both with and without the outliers being adjusted.

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

Furthermore, we observe that the MSE (0.052905) forecast without adjusting for the outlier is higher than the MSE (0.013100) of the forecast after adjusting for the outlier. Accordingly, detecting and adjusting outliers comes first when studying time series data.

5. Conclusions

We come to the following conclusions: first, the value of residual standard error (with outlier adjustment) rose whenever the critical value was raised. Secondly, the quantity of outlier values reduced each time the critical value was raised. Third, forecasts with outlier correction perform better than forecasts without outlier adjustment when outliers are present.

The procedures are widely applicable because they are based on basic techniques. For example, they can be employed as a data screening tool in robust time series analysis and spectral density estimation. In biological research where external disturbances are inevitable, they can also be exploited.

The study shows that outliers have an effect on the autocorrelation structure of any time series, supporting the assumption that outliers do cause model misspecification. In our case, an example is provided by the fact that the ARIMA(1,1,0)$(0,0,1)_{12}$ was initially the best model that could be fitted to our data. Though the model's parameters were significant, testing the residuals for normality and constant variance revealed that both presumptions were broken. For a decision-maker, using this model for forecasts would have produced false data. The existence of outliers may be to blame for this. After removing the outliers from the series, the best method was discovered to be ARIMA(1,1,2)$\times$ $(0,0,1)_{12}$, and all of the model's parameters were significant. The assumptions of normality and constant variance were not collapsed according to diagnostic tests. This proves that the method is effective in identifying and adjusting for outliers. It works with every invertible ARIMA model. It is also adaptable and simple to understand. For the technique to yield even better results, additional time series diagnostic instruments must be applied.

References

[1] B. Abraham and G. E. P. Box, Bayesian analysis of some outlier problems in time series, *Biometrika*, 66(2), (1979), 229–236. https://doi.org/10.1093/biomet/66.2.229

[2] S. Ahmad and S. Purdy, Real-time anomaly detection for streaming analytics, *arXiv preprint arXiv*:1607.02480, (2016). https://doi.org/10.48550/arXiv.1607.02480

[3] S. Aminikhanghahi and D. J. Cook, A survey of methods for time series change point detection, *Knowledge and information systems*, 51(2), (2017), 339–367. https://doi.org/10.1016/j.matpr.2017.11.277

[4] P. Arumugam and R. Saranya, Outlier detection and missing value in seasonal ARIMA model using rainfall data, *Materials Today: Proceedings*, 5(1), (2018), 1791–1799.

[5] J. Cabrieto, F. Tuerlinckx, P. Kuppens, M. Grassmann, and E. Ceulemans, Detecting correlation changes in multivariate time series: A comparison of four non-parametric change point detection methods, *Behavior research methods,* 49, (2017), 988–1005. https://doi.org/10.3758/s13428-016-0754-9

[6] A. Capozzoli, F. Lauro, and I. Khan, Fault detection analysis using data mining techniques for a cluster of smart office buildings, *Expert Systems with Applications*, 42(9), (2015), 4324–4338. https://doi.org/10.1016/j.eswa.2015.01.010

[7] I. Chang, G. C. Tiao, and C. Chen, Estimation of time series parameters in the presence of outliers, *Technometrics*, 30(2), (1988), 193–204.

[8] W. Chen, K. Zhou, S. Yang, and C. Wu, Data quality of electricity consumption data in a smart grid environment, *Renewable and Sustainable Energy Reviews*, 75, (2017), 98–105. https://doi.org/10.1016/j.rser.2016.10.054

[9] P. Filonov, A. Lavrentyev, and A. Vorontsov, Multivariate industrial time series with cyber-attack simulation: Fault detection using an lstm-based predictive data model, *arXiv preprint arXiv:*1612.06676, (2016). https://doi.org/10.48550/arXiv.1612.06676

[10] A. J. Fox, Outliers in time series, *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(3), (1972), 350–363. https://doi.org/10.1111/J.2517-6161.1972.TB00912.X

[11] F. Ganz, D. Puschmann, P. Barnaghi, and F. Carrez, A practical evaluation of information processing and abstraction techniques for the internet of things, *IEEE Internet of Things journal*, 2(4), (2015), 340–354. https://doi.org/10.1016/j.rse.2014.11.005https://doi.org/10.1109/JIOT.2015.2411227

[12] T. Hermosilla, M. A. Wulder, J. C. White, N. C. Coops, and G. W. Hobart, An integrated Landsat time series protocol for change detection and generation of annual gap-free surface reflectance composites, *Remote Sensing of Environment*, 158, (2015), 220–234. https://doi.org/10.1016/j.rse.2014.11.005

[13] S. Johansen and B. Nielsen, Asymptotic theory of outlier detection algorithms for linear time series regression models, *Scandinavian Journal of Statistics*, 43(2), (2016), 321–348. https://doi.org/10.1111/sjos.12174

[14] M. Kontaki, A. Gounaris, A. N. Papadopoulos, K. Tsichlas, and Y. Manolopoulos, Efficient and flexible algorithms for monitoring distance-based outliers over data streams, *Information systems*, 55, (2016), 37–53. https://doi.org/10.1016/j.is.2015.07.006

[15] L.-M. Liu, G. B. Hudak, G. E. P. Box, M. E. Muller, and G. C. Tiao, *Forecasting and time series analysis using the SCA*

statistical system, 1(2). Scientific Computing Associates DeKalb, IL, (1992).

[16] M. Liu, J. Shi, K. Cao, J. Zhu, and S. Liu, Analyzing the training processes of deep generative models, I*EEE transactions on visualization and computer graphics*, 24(1), (2017), 77–87. https://doi.org/10.1109/TVCG.2017.2744938

[17] S. Liu, A. Wright, and M. Hauskrecht, Change-point detection method for clinical decision support system rule monitoring, *Artificial intelligence in medicine*, 91, (2018), 49–56. https://doi.org/10.1016/j.artmed.2018.06.003

[18] Z. Liu, M. M. Verstraete, and G. De Jager, Handling outliers in model inversion studies: a remote sensing case study using MISR-HR data in South Africa, *South African Geographical Journal*, 100(1), (2018), 122–139. https://doi.org/10.1080/03736245.2017.1339629

[19] D. Loureiro, C. Amado, A. Martins, D. Vitorino, A. Mamade, and S. T. Coelho, Water distribution systems flow monitoring and anomalous event detection: A practical approach, *Urban Water Journal*, 13(3), (2016), 242–252. https://doi.org/10.1080/1573062X.2014.988733

[20] L. Martí, N. Sanchez-Pi, J. M. Molina, and A. C. B. Garcia, Anomaly detection based on sensor data in petroleum industry applications, *Sensors*, 15(2), (2015), 2774–2797. https://doi.org/10.3390/s150202774

[21] R. D. Martin and V. J. Yohai, Influence functionals for time series, *The annals of Statistics*, (1986), 781–818. https://doi.org/10.3390/s150202774

[22] J. Reiche, J. Verbesselt, D. Hoekman, and M. Herold, Fusing Landsat and SAR time series to detect deforestation in the tropics, *Remote Sensing of Environment*, 156, (2015), 276–293. https://doi.org/10.1016/j.rse.2014.10.001

[23] P. J. Rousseeuw and W. Van Den Bossche, Detecting deviating data cells, *Technometrics*, 60(2), (2018), 135–145. https://doi.org/10.1080/00401706.2017.1340909

[24] P. Rousseeuw, D. Perrotta, M. Riani, and M. Hubert, Robust monitoring of time series with application to fraud detection, *Econometrics and statistics*, 9, (2019), 108–121. https://doi.org/10.1016/j.ecosta.2018.05.001

[25] G. Sprint, D. J. Cook, and M. Schmitter-Edgecombe, Unsupervised detection and analysis of changes in everyday physical activity data, *Journal of biomedical informatics*, 63, (2016), 54–65. https://doi.org/10.1016/j.jbi.2016.07.020

[26] O. M. Staal, S. Sælid, A. Fougner, and Ø. Stavdahl, Kalman smoothing for objective and automatic preprocessing of glucose data, *IEEE journal of biomedical and health informatics*, 23(1), (2018), 218–226. https://doi.org/10.1109/JBHI.2018.2811706