



Contents lists available at ScienceDirect

Journal of King Saud University – Science

journal homepage: [www.sciencedirect.com](http://www.sciencedirect.com)

Original article

# An algorithm for outlier detection in a time series model using backpropagation neural network

Gajendra K. Vishwakarma <sup>a,\*</sup>, Chinmoy Paul <sup>a,b</sup>, A.M. Elsayah <sup>c,d</sup><sup>a</sup> Department of Mathematics & Computing, Indian Institute of Technology Dhanbad, Dhanbad 826004, India<sup>b</sup> Department of Statistics, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya, Eralgool, Karimganj 788723, India<sup>c</sup> Division of Science and Technology, Beijing Normal University–Hong Kong Baptist University United International College, Zhuhai 519085, China<sup>d</sup> Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

## ARTICLE INFO

### Article history:

Received 13 March 2020

Revised 18 August 2020

Accepted 11 September 2020

Available online 22 September 2020

**Mathematics Subject Classification:**  
68Wxx

### Keywords:

Multivariate outliers

Detection

Neural network

Robust estimate

Time series

Backpropagation algorithm

## ABSTRACT

Outliers are commonplace in many real-life experiments. The presence of even a few anomalous data can lead to model misspecification, biased parameter estimation, and poor forecasts. Outliers in a time series are usually generated by dynamic intervention models at unknown points of time. Therefore, detecting outliers is the cornerstone before implementing any statistical analysis. In this paper, a multivariate outlier detection algorithm is given to detect outliers in time series models. A univariate time series is transformed to bivariate data based on the estimate of robust lag. The proposed algorithm is designed by using robust measures of location and dispersion matrix. Feed forward neural network is used for designing time series models. Number of hidden units in the network is determined based on the standard error of the forecasting error. A comparison study between the proposed algorithm and the widely used algorithms is given based on three real-data sets. The results demonstrated that the proposed algorithm outperformed the existing algorithms due to its non-requirement of a priori knowledge of the time series and its control of both masking and swamping effects. We also discussed an efficient method to deal with unexpected jumps or drops on share prices due to stock split and commodity prices near contract expiry dates.

© 2020 Published by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

The detection of outliers or unusual data structures is one of the important tasks in the statistical analysis of time series data as outliers may have a substantial influence on the outcome of an analysis. Appropriate definition of an outlier usually depends on the assumptions about the structure of data and the applied detection method. Hawkins (1980) defined the outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Barnett and Lewis (1994) indicated that an outlying observation, or outlier, is

one that appears to deviate markedly from other members of the sample in which it occurs. Similarly, Johnson (1992) viewed that, an outlier is an observation in a data set which appears to be inconsistent with the remainder of that set of data. There are many definitions of outlier proposed in the literature of time series. Outlier observations in some situations are also referred as anomalies, discordant observations, or contaminants Carreno et al. (2019).

The presence of outliers in a time series has a significant effect on the results of standard procedures of analysis. The consequences may lead to improper model specification, faulty parameter estimation and substandard forecasting. A crucial point here is that any outlier detection technique can at most detect a set of data points having different behavior than the rest of the data and hence, it can be termed as a probable set of outliers. However, it is up an analyst to take various itineraries to come up with a final decision to justify these detected points as outliers. It is probable that a point detected as an outlier has some real facts behind it, e.g., the price of a stock just after the date of stock split with split ratio of 2-for-1 or 3-for-1, which means a stockholder gets two or three shares, respectively, for every share held. In a reverse stock

\* Corresponding author.

E-mail addresses: [vishwagk@rediffmail.com](mailto:vishwagk@rediffmail.com) (G.K. Vishwakarma), [chinmoy.gcc@gmail.com](mailto:chinmoy.gcc@gmail.com) (C. Paul), [amelsawah@uic.edu.cn](mailto:amelsawah@uic.edu.cn) (A.M. Elsayah).

Peer review under responsibility of King Saud University.



Production and hosting by Elsevier

<https://doi.org/10.1016/j.jksus.2020.09.018>

1018-3647/© 2020 Published by Elsevier B.V. on behalf of King Saud University.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

split, a company divides the number of shares that stockholders own, raising the market price accordingly. In such a scenario, most of the outlier detection algorithms will detect the prices after the date of stock split as outliers. The impact of outliers on parameter estimation has been studied by Pena (1990), Fox (1972). Deutsch et al. (1990) explored the effects of a single outlier on autoregressive moving-average (ARMA) identification. In their study, Barreyre et al. (2019) used statistical outlier detection methods to detect anomaly in space telemetries. The suggested methods addressed the issue of outlier limited to the nature and number of outliers. Moreover, the some of the method of parameter estimation is based on maximum likelihood estimation or on the least square approach.

Outliers in a time series are usually generated by dynamic intervention models at unknown points of time. A common practice to deal with the outliers in a time series is to identify the locations of outliers and then to use intervention models to analyze the outlier effects. This is an iterative method that requires iterations between outlier detection and estimation of model parameters. Tsay (1988) discussed the significance of outliers in level shift and its dynamics that leads to change in variance of the series. In their findings, Chang et al. (1988) introduced two types of outliers, namely additive outliers (AO) and innovative outliers (IO). However, Chen and Liu (1993) later introduced two more types of outliers in time series such as temporary change (TC), and level shift (LS), addressing their effect in modeling and estimating the parameters of time series. They further demonstrated that the sensitivity of the forecast intervals are mainly due to AO and discussed the issue of forecasting when outliers occur near, or at the forecast origin. The consequence of additive outliers on forecasts was addressed by Ledolter (1989) in the case of ARMA model. In their study, Battaglia and Orfei (2005) discussed the problem of identifying the location of outliers and estimation of the amplitude in nonlinear time series. An alternative semi-parametric estimator for the fractional differencing parameter in the autoregressive fractionally integrated moving average (ARFIMA) model was introduced by Molinarena et al. (2009) which is robust against additive outliers. In their study, Leduca et al. (2011) considered the implementation of auto-covariance function that is robust to the additive outliers. Loperfido (2020) discussed a method based on achieving maximal kurtosis for outlier detection in multivariate and univariate time series models. However, the estimation procedure is based on the assumption that the model parameters are known, which may not be the case always, especially in case of real data.

Modeling in time series depends on finding the functional relationship of the time series with its lagged variables. Artificial neural networks (ANN) are well known for their ability to find functional relationship between input and sets of output variables. Moreover, modeling with ANN may not require a priori knowledge of parameters of the time series. It has been observed that a network with properly designed architecture can approximate any function to its desired accuracy (cf. Hornik et al. (1989), Hornik (1991)). In their study, Shaheed (2005) used the approach of feed-forward neural network with an external input and resilient propagation algorithm to model non-linear autoregressive process. Several studies have been carried out to address the application of ANN for estimation and forecasting of linear or nonlinear autoregressive moving average (NARMA) process. Farayay and Chatfield, 1998, Zhang et al., 1998, Khashei and Bijari, 2010 are some of the studies where ANN has been used in parameter estimation and forecasting of the ARMA process. However, the presence of outlier may affect both model identification and forecasting in time series data.

Anomalies in a time series are considered as the observations that deviate from some usual or standard patterns. Anomaly detection in time series is a growing area of research, where different

techniques have been developed in the field of machine learning as well as statistics. Omar et al. (2013) used machine learning techniques for anomaly detection. In the field of statistics, some significant developments in terms of anomaly detection techniques in time series are Jeng et al. (2013), Bardwell and Fearnhead (2017), Ahmad et al. (2017). The statistical procedures are generally direction dependent and may be based on some prior assumptions, be it distribution of the data or prior knowledge of the parameters. Whereas, optimization techniques are the basis for machine learning algorithms and selection of an improper algorithm may result in misleading outcome.

It is therefore necessary to look for robust methods which do not require a priori knowledge of time series and may not dependent on number, nature of outliers. Robust statistics deals with the theory of stability of statistical procedures. It methodically departs from the modeling assumptions of well-known procedures and eventually tries to develop efficient procedures. Motivated by these facts, an algorithm is proposed to detect outliers in time series. The proposed algorithm uses robust measures of location and dispersion matrix. Outlier free data has been modeled with feed forward ANN. The architecture of the ANN is determined experimentally. A 'R' package (Otsad) developed by Iturria et al. (2020) which can detect point outliers in a univariate time series along with BARD technique by Bardwell and Fearnhead (2017), a technique based on bayesian approach to detect abnormal regions has been used for comparative study and results thereof are presented.

## 2. Preliminaries and notations

### 2.1. Autoregressive (AR) model

In general, an autoregressive process of order  $p$  can be defined as

$$\sum_{i=1}^p a_i Y_{t-i} = e_t, \quad t = 0, \pm 1, \pm 2, \dots, \pm n, \quad (1)$$

where  $a_0 \neq 0$ ,  $a_p \neq 0$  and  $\{e_t\}$  are uncorrelated random variables with zero mean and constant variance.

### 2.2. Autoregressive moving average (ARMA) model

In an ARMA process, both autoregressive and moving average processes are considered. Let  $\{Y_t, t \in (0, \pm 1, \pm 2, \dots, \pm n)\}$  be a stochastic process. An ARMA process of order  $(p, q)$  is defined as

$$\begin{aligned} Y_t + a_1 Y_{t-1} + a_2 Y_{t-2} + \dots + a_p Y_{t-p} \\ = e_t + b_1 e_{t-1} + b_2 e_{t-2} + \dots + b_q e_{t-q}. \end{aligned} \quad (2)$$

where  $a_p \neq 0$ ,  $b_q \neq 0$  and  $\{e_t\}$  is the sequence of random variables which are uncorrelated in nature.

### 2.3. Types of outliers in a time series

Let  $\{Y_t\}$  be a time series that follows a general ARMA process, then  $\{Y_t\}$  can also be presented as follows

$$Y_t = \frac{\theta(B)}{\alpha(B)\phi(B)} u_t, \quad t = 1, 2, \dots, n, \quad (3)$$

where  $n$  denotes the number of observation in the time series,  $\theta(B)$ ,  $\alpha(B)$  and  $\phi(B)$  are polynomials of  $B$ . The outliers in a time series may generally be represented by  $L(B)I(t_j)$ , where  $L(B) = \frac{A(B)}{G(B)H(B)}$  is polynomial function of lag operators and  $I(t_j)$  in an indicator function which takes value 1 when  $t = j$  else take the value 0. Here  $A(B)$ ,  $G(B)$ , and  $H(B)$  represent the polynomials

$\theta(B)$ ,  $\alpha(B)$ , and  $\phi(B)$ , respectively and  $B$  is a backshift operator. The types of outliers in a time series are (i) innovational outlier (IO), (ii) additive outliers (iii), level shift, and (iv) temporary change, which can be defined as follows

$$\left. \begin{aligned} IO : L(B) &= \frac{\theta(B)}{\alpha(B)\phi(B)}, & AO : L(B) &= 1, \\ TC : L(B) &= \frac{1}{(1-\delta B)} \text{ and } LS : L(B) &= \frac{1}{(1-B)} \end{aligned} \right\} \quad (4)$$

Further, details about outliers can also be found in [Chen and Liu \(1993\)](#).

### 3. Methodology

In this study, back propagation neural network (BPNN) has been used to model time series. Brief descriptions of some methodologies are as follows.

#### 3.1. Backpropagation neural network

The BPNN is one of the popular neural network methods. It is a feed forward, multilayer perceptron (MLP) supervised learning network. The backpropagation algorithm looks for the minimum of the error function in the weight space using the method of gradient descent. The combination of weights which minimizes the error function is considered to be a solution of the learning problem. Since this method requires computation of the gradient of the error function at each iteration step, continuity and differentiability of the error function need to be guaranteed. Obviously, one has to use an activation function other than the step function. For MLP, the output of one layer becomes the input of the subsequent layers. The neurons in the first layer receive external inputs, and the neurons in the last layer present the output of the network. The following equation describes this operation

$$a^{m+1} = f^{m+1}(W^{m+1}a^m + b^{m+1}), \text{ for } m = 1, 2, \dots, M - 1,$$

where  $a$  is the input vector to the layer,  $b$  is the noise vector,  $W$  is the weight matrix to each neuron,  $M$  is the number of layers in the network, and  $f$  is the activation function. The widely used type of activation function used in the hidden layer is the tangent hyperbolic, i.e.,

$$f(x) = \frac{\sin(hx)}{\cos(hx)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \quad (6)$$

#### 3.2. Forecasting using ANN

For forecasting in time series, a training pattern is constructed using lagged time series. Let us suppose, there are  $N$  observations  $y_1, y_2, \dots, y_N$  in the time series and we need one-step ahead forecasting using ANN. The first training pattern will consist of  $y_1, y_2, \dots, y_n$  as inputs and  $y_{n+1}$  as an output. If there are  $n$  nodes at the input layer, the total number of training patterns will be  $(N-n)$ . The cost function that is used during the training process is given as follows

$$E = \frac{1}{2} \sum_{i=n+1}^N (y_i - a_i)^2, \quad (7)$$

where  $a_i$  denotes the actual output of the network and  $1/2$  is included for the simplification of derivative computed during the training algorithm.

### 4. Outlier identification and forecasting with proposed method

The measures of location and dispersion are two of the most useful alternatives for describing data mean and variation. Usually,

sample mean  $\bar{x}$  and variance  $s^2$  of a sample  $X_N = \{x_i\}_{i=1}^N$  provide good estimate of location and dispersion, if data is free from outliers. When data is contaminated, even a single observation with large deviation may affect sample mean as well as dispersion matrix significantly. Therefore, in case of contaminated data, robust estimation of the model such as M-estimator by [Huber \(1981\)](#), generalized M-estimator by [Denby and Martin \(1979\)](#) are useful. However, the efficiency of these estimators decreases when the order of AR ( $p$ ) model is high.

To improve the robustness of the model, [Liu et al. \(2004\)](#) suggested the transformation of the original univariate series  $\{y_t\}_{t=1}^N$  into a bivariate  $\{Y_t = (y_t, y_{t-k})\}_{t=k+1}^N$  series. One of the multivariate robust estimation methods, such as the minimum covariance determinant (MCD) estimator developed by [Rousseeuw \(1984\)](#), [Rousseeuw and Driessen \(1999\)](#) can be applied instead of the univariate M or GM-estimators for the robust estimation of the model. However, when a time series is transformed from univariate into bivariate, application of MCD will detect outliers as pairs. If the original time series has outlier at  $m^{th}$  and  $l^{th}$  position,  $(m, l) < k$ , then the application of the MCD will detect both the pairs, i.e.,  $(m, m+k)$  and  $(l, l+k)$  as outliers. Hence, together with the original outliers, same number of additional observations will also be detected by the MCD. Further, the MCD detects outlier based on some fixed threshold value (e.g.,  $\chi^2_{(p, 0.98)}$ ) that is subjective as suggested by [Filzmoser et al. \(2005\)](#) because of the following reasons:

- If the data is drawn from a single multivariate normal distribution, the threshold is most likely to be infinity as there are no observations from different distributions.
- A fixed threshold may not be always appropriate for every data set.

To deal with this problem, the subsequent algorithm is proposed as follows.

#### 4.1. Proposed algorithm for outlier detection

##### Step I. Estimation of Robust Lag

In this step, autocorrelation coefficient for a time series is estimated by a multivariate location and scatter estimator. The original univariate series  $\{y_t\}_{t=1}^N$  is transformed into a bivariate  $\{Y_t = (y_t, y_{t-k})\}_{t=k+1}^N$  series using the estimated lag.

##### Step II.1. Initial Step to Identify Outliers (as pairs)

In this step, a multivariate clustering algorithm is used to identify the outliers in the transformed bivariate data. [Chatterjee and Roy \(2014\)](#) used Mohalanobis distance to define radius of clustering algorithm in their study. Mohalanobis distance is computed using sample mean and covariance matrix. It is well-known that both the estimators are very sensitive to extreme observations. To increase robustness, [Paul and Vishwakarma \(2017\)](#) proposed an algorithm based on distance measures of [Hadi \(1994\)](#). This algorithm has been used in identifying the outliers in transformed bivariate data from Step I.

##### Step II.2. Final Step to Identify Outliers (as pairs)

Let us consider that, data has been drawn from a  $p$  ( $p = 2$ ) variate multivariate normal distribution. Let  $G_n(u)$  and  $G$  denote empirical distribution function of  $MD^2$  (Mohalanobis distance) and distribution function of  $\chi^2_p$ , respectively. By strong law of large numbers,  $G_n(u)$  converges to  $G$  almost surely i.e.  $G_n(u) \xrightarrow{a.s.} G$ . The tail values of  $G_n(u)$  and  $G$  can be used to decide outliers. Tails in this case can be defined by  $\delta = \chi^2_{p, 1-\alpha}$  for certain value of  $\alpha$  and  $p_n(\delta) = \sup_{u \geq \delta} (G(u) - G_n(u))^+$ . Here '+' denotes positive difference  $p_n(\delta)$  and is the measure of departure of empirical distribution

from theoretical distribution and can be considered as measure of outliers in the sample. However,  $p_n(\delta)$  is not directly used as a measure of outliers rather its critical value, i.e.,  $p_{crit}(\delta, n, p) = (0.24 - 0.003p)/\sqrt{n}$  is used in general.

**Step III.** Identification of Actual Outlier

This step identifies the actual outlier from each pair of selected outliers. The observed residuals can be written as follows

$$\eta_t = g(y_{(t-1)}; \eta_{(t-1)}), \quad t = r + 1, \dots, n. \tag{8}$$

If there is no outlier in the data then  $\eta_t$  will approach  $\varepsilon_t$ . Suppose an outlier with amplitude  $\omega_q$  occurs at time  $t$ , then we get

$$\eta_t = \varepsilon_t, \quad t < q; \quad \eta_q = \varepsilon_q + \omega_q, \quad t = q; \quad \text{and } \eta_{q+j} = \varepsilon_{q+j}, \quad j = 1, 2, \dots, n - q.$$

Thus, we have

$$\sum_{t=r+1}^n \varepsilon_t^2 = \sum_{t=r+1}^{q-1} \eta_t^2 + (\eta_q - \omega_q)^2 + \sum_{t=q+1}^n \eta_t^2. \tag{10}$$

Minimizing (10) with respect to  $\omega_q$  gives  $\hat{\omega}_q = \eta_q$ .

Test statistics for the likelihood ratio test  $H_0$ : no outlier at  $t = q$  against  $H_1$ : there is outlier at  $t = q$  (as defined by [Chang et al., 1988](#)) is  $\hat{\omega} / \hat{\sigma}_{l,q}$ , where

$$\hat{\sigma}_{l,q} = \frac{\eta_{r+1}^2 + \dots + \eta_{q-1}^2 + \eta_{q+1}^2 + \dots + \eta_n^2}{n - r}. \tag{11}$$

Under  $H_0$ , the test statistic asymptotically follows  $N(0, 1)$ . Thus,  $\hat{\sigma}_{l,q}$  can be considered as an estimate of the standard error of the outlier at time  $q$ .

**4.2. Issues related to the proposed algorithm**

In Step II.1, as described above, Mohalanobis distance is calculated using robust measures of location and dispersion matrix. If the dispersion matrix becomes singular, then the algorithm will end up with the issue of infinite loop. In the present scenario, possibility of such issues arises because a time series  $y_t$  is, in general, highly correlated with its lagged series  $y_{t-k}$  and the co-variance matrix corresponding to  $\{Y_t = (y_t, y_{t-k})\}_{t=k+1}^N$  may tend to be singular. The problem of singularity of covariance matrix is avoided by considering the nearest positive definite covariance matrix as suggested by [Higham \(2002\)](#).

The algorithm starts with transforming the time series to a bivariate data. The initial step, i.e., Step II.1 of the algorithm will generate clusters based on the cluster radius. The radius of the algorithm increases gradually as the algorithm proceeds. Hence, the possibility of the outliers or suspicious observations will be at the end clusters. The Step II.2 will help in identifying the clusters containing outliers amongst the end clusters. Further, Step III helps in identifying actual outliers which are identified in Step II.2 as pairs in end clusters.

The proposed algorithm is designed in a way that it does not depend on any assumption regarding the distribution or the nature of the time series. The algorithm can perform well with small sample size. However, presence of very high co-linearity in the data may affect the performance of the algorithm. The performance of the proposed algorithm considering various types of simulated data and real data are discussed in the subsequent section.

**5. Simulation and data analysis**

Simulation helps in comparison of analytical techniques even when techniques under study deviate from the standard assumptions. The different methodologies that have been adopted

to identify outliers in time series are [Chen and Liu \(1993\)](#), the MCD by [Rousseeuw and Zomerren \(1990\)](#), Bayesian approach to detect abnormal regions (BARD), OTSAD and finally the proposed method. Outlier free data is used as an input to a single layered feed-forward neural network for training. Architecture for training network is determined experimentally based on the standard error (SE) of the prediction. An architecture corresponding to which SE of predicted error is least is considered as the suitable architecture for that particular time series. For the simulated time series, initial 90% of the data is used for training the network and rest of the 10% is used for prediction. The following simulated and real data sets have been considered for carrying out comparative study.

**Data set -I:** A time series data from an ARMA (2, 2) process has been simulated with AR and MA coefficients (0.8897, -0.4858) and (-0.2279, 0.2488) respectively. Length of the time series is 100 and every 10<sup>th</sup> observation is contaminated by adding an observation of magnitude  $5\sigma$ . When an ARMA (2, 2) model was fitted to the contaminated data, the resulting coefficient for AR and MA component are found to be (1.37267, -0.58397) and (-1.46634, 0.52976), respectively.

The plot of the time series and the residuals as presented in [Fig. 1](#), clearly suggest the presence of outliers in the time series data. The performance of the proposed method is compared with other methods via correlation of the predicted and original values, and standard error of predicted residuals ( $e_i$ ) which is given as below

$$SE(e_i) = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y - \hat{Y}_i)^2}. \tag{12}$$

For comparing the performance of the proposed method, some more time series with different set of parameters are simulated. To each simulated time series, outliers are added after a regular interval. The magnitude of the outliers is kept at  $4\sigma$  and  $5\sigma$ . Here  $\sigma$  denotes the standard deviation of the time series. [Table 1](#) shows the performance of the proposed algorithm along with other methods.

From [Table 1](#), it can be observed that the performance of the proposed method is better compared to others methods. Performances are compared in terms of correlation of the predicted values with that of originals and SE of predicted errors. The method which is effective in identifying outliers is expected to have lesser SE and higher correlation between predicted and original values.

**Data set II:** In this case, the following system that represents biological characteristics of single neurons has been considered to simulate nonlinear time series. The set of differential equation discussed below represents the Morris-Lecar (M-L) neural system

$$\begin{aligned} \dot{X} = & -0.5g_{Ca} \left\{ 1 + \tanh\left(\frac{X - V_1}{V_2}\right) \right\} (X - 1) \\ & - g_K Y (X - V^K) - g_L (X - V^L) + Z + \alpha \xi(t), \end{aligned} \tag{13}$$

$$\begin{aligned} \dot{Y} = & (1/3) \cosh\left(\frac{X - V_3}{2V_4}\right) \\ & \left\{ 0.5 \left( 1 + \tanh\left(\frac{X - V_3}{V_4}\right) \right) - Y \right\}, \end{aligned} \tag{14}$$

and

$$\dot{Z} = -\mu(V_0 + X). \tag{15}$$

In the above model,  $X$  represents the membrane potential of the neuronal cell,  $Y$  is the activation variable, and  $Z$  is the applied input current. The parameter values  $-g_{Ca}$ ,  $g_K$  and  $g_L$  represent the maximum conductances corresponding to the  $Ca^{+2}$ ,  $K^+$  and leak currents, respectively.  $V^K$  and  $V^L$  represent the reversal potentials



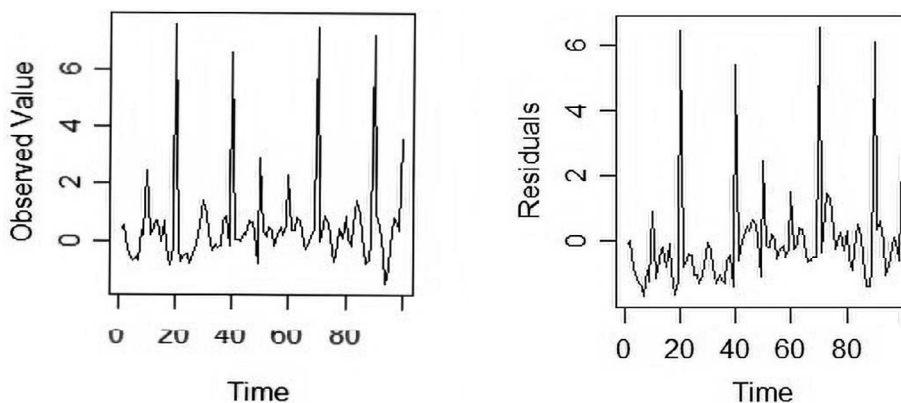


Fig 1. ARMA (2, 2) data contaminated with outliers and its estimated residuals.

Table 1

Correlation of predicted values with the original values and SE of predicted error for the time series with outliers (Original TS) and outlier free time series by different methods i.e., Chen and Liu, MHD, OTSAD, BARD and the proposed method.

Series	AR	MA	OUT Size	Method	Actual Mean	Predicted Mean	SE	Correlation
1	(0.25, 0.5)	(0.15, 0.25)	5	Original TS	-0.98428	2.33188	1.40399	0.42110
				TS out		0.67281	1.14356	0.67140
				MHD		5.09674	2.33637	0.73592
				OTSAD		-1.02115	2.18269	-0.02344
				BARD		-0.97036	2.17000	-0.01677
				Proposed		-0.12080	<b>1.36421</b>	<b>0.92216</b>
2	(0.25, 0.5)	(0.15, 0.25)	4	Original TS	-0.98428	2.33188	2.14380	-0.56041
				TS out		0.10008	1.59520	-0.50953
				MHD		2.14803	1.80944	0.40594
				OTSAD		-1.08540	2.20467	-0.10353
				BARD		-0.94229	2.17188	-0.09783
				Proposed		-0.78951	<b>1.42248</b>	<b>0.98135</b>
3	(0.35, 0.4)	(0.25, 0.35)		Original TS	-0.81950	1.52126	0.92348	-0.29183
				TS out		1.24810	1.00461	-0.14333
				MHD		1.22683	0.92051	0.12738
				OTSAD		0.52043	1.85552	0.07658
				BARD		0.21498	1.48411	0.60338
				Proposed		1.73246	<b>0.90988</b>	<b>0.71946</b>
4	(0.35, 0.4)	(0.25, 0.35)		Original TS	1.87942	0.62172	1.32867	0.01295
				TS out		0.07351	1.27462	-0.38708
				MHD		-0.07543	1.19450	-0.24250
				OTSAD		0.79915	1.59565	0.01211
				BARD		0.55382	1.69728	-0.23965
				Proposed		0.04587	<b>1.18582</b>	<b>0.02165</b>
5	(0.45, 0.3)	(0.35, 0.5)		Original TS	0.73290	-1.52951	1.08669	0.41153
				TS out		-2.12211	1.35197	-0.81606
				MHD		-1.79043	1.23342	-0.71067
				OTSAD		-0.55577	1.09098	0.72645
				BARD		-0.54276	1.08638	0.65570
				Proposed		-0.05579	<b>1.04609</b>	<b>0.82315</b>
6	(0.45, 0.3)	(0.35, 0.5)		Original TS	-0.30581	4.46680	1.81264	-0.52627
				TS out		1.61485	1.75415	-0.37026
				MHD		3.32455	1.56829	-0.23019
				OTSAD		-0.52036	3.24570	-0.41797
				BARD		0.04587	1.74190	0.32706
				Proposed		3.32059	<b>1.54702</b>	<b>0.85747</b>

corresponding to the above ionic currents. The parameters  $V_1, V_2, V_3$  and  $V_4$  are appropriately chosen for the hyperbolic functions so that they can attain their equilibrium points instantaneously.

Fourth order Runge-Kutta (RK) method has been used to simulate data for  $(X, Y, Z)$ . Time interval considered in the simulation process is  $\Delta t = 0.001$  seconds and simulation is carried out over a time period of 20 s. Simulated time series are presented in Fig 2. Data for these simulated variables are tested for nonlinearity as suggested by Terasvirta et al. (1993) and the below table presents the result of nonlinearity test.

Results in Table 2 confirm that  $X$  and  $Y$  are non-linear in nature. The discussed methods of identifying outliers in a time series are

applied to the simulated time series of  $(X, Y)$ . While 90% of the simulated initial data is used for training, the rest 10% is used for prediction. Correlation of the predicted and actual values along with the SE of the predicted errors is presented in Table 3. The noticeable point here is that, Chen and Liu method does not find any outlier in simulated time series of  $X$ . However, BARD and OTSAD identifies outliers in both the time series of  $X$  and  $Y$ . The outlier-free data is used for prediction and the result obtained by application of various methods is presented in Table 3.

**Data set III:** In this case, two time series of TCS stock price and Aluminum trading prices for the period January 2, 2006 to December 31, 2010 and December 24, 2014 to April 12, 2015 are considered respectively. The reason for considering these two time

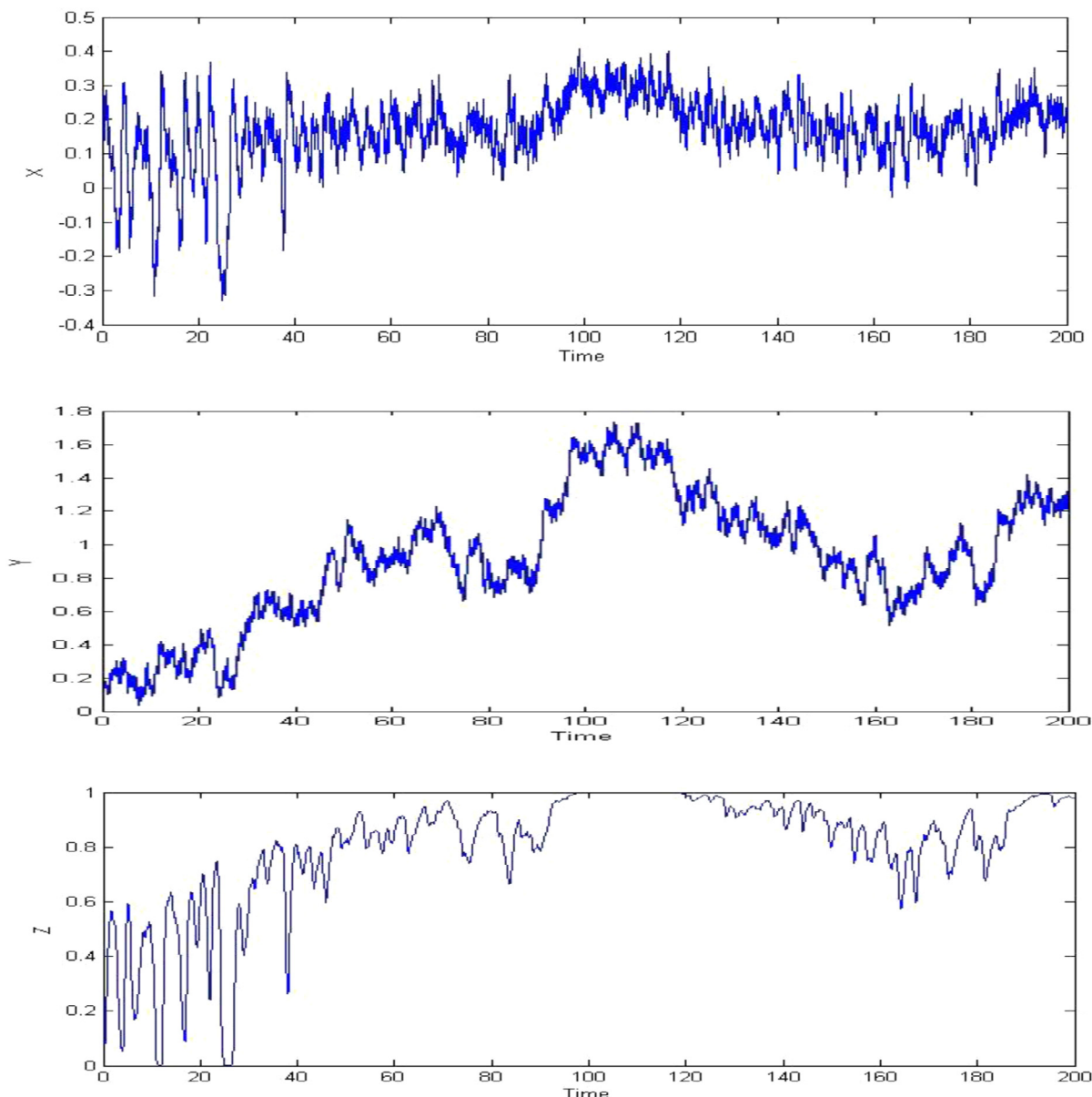


Fig 2. Plot of the time series (X, Y, Z) from the model described in equations (13), (14) and (15).

**Table 2**  
Test for nonlinearity of time series.

Hypothesis	Terasvirta's Test		White NN test	
	Linearity in mean		Linearity in mean	
	Statistic	P-Value	Statistic	P-Value
X	26.1623	2.084E-06	26.3139	1.932E-06
Y	10.3352	0.005698	11.0802	0.003926
Z	0.610	0.7369	0.522	0.7703

series is to discuss the issue of unusual jumps and drops observed in a financial time series. Generally, in a corporate action such as stock split, a company divides its existing shares into multiple shares. This causes increase in the number of shares by a specific multiple but at the same time the whole value of the shares remains the same. The price of the resulting shares will suddenly drop to half or one third based on split ration. Similar fluctuations are also observed in case of nearing expiry dates of commodity

contracts. A contract that will expire after several years from now will remain relatively illiquid until it gets closer to its expiry. As the liquidity increases near the expiry date, sudden change in price of the commodity can be observed. The consequences of these corporate actions may result in the following:

- a) The return distribution may change.
- b) Most of the outlier detection algorithm may detect the changes in price brought in due to stock split or nearing maturity dates as outlier.

Data for commodity (Aluminum) has been taken from the website <http://www.mcxindia.com>. Contract expiry dates for the corresponding metal can also be obtained from the same website. The change in price of a financial instrument due to its nearing maturity of stock is not market-driven and brings in unexpected volatility. Leveling such volatility is to smoothen the historical prices of a time series based on the changes brought in due to recent corporate events. The following method of adjustment

**Table 3**

Correlation of predicted values with the original values and SE of predicted error for the time series with outliers (Original TS) and outlier free time series by different methods i.e., Chen and Liu, MHD, OTSAD, BARD and proposed on Nonlinear time series from a dynamical system.

Series	Method	Actual Mean	Predicted Mean	SE	Correlation
X	Original TS	-0.13628	0.05940	0.12221	-0.82892
	Chen and Liu		0.05242	0.11792	-0.72504
	MHD		0.02549	0.10536	-0.78706
	OTSAD		-0.24971	0.02310	-0.09882
	BARD		-0.25023	<b>0.02313</b>	0.02313
	Proposed		0.07258	0.11783	<b>0.96080</b>
Y	Original TS	0.25878	0.31182	0.08124	0.94024
	Chen and Liu		0.18787	0.08096	-0.90088
	MHD		0.43872	0.10395	0.93376
	OTSAD		0.07274	0.15993	0.88120
	BARD		0.32472	<b>0.03009</b>	0.03009
	Proposed		0.20747	0.05746	<b>0.94633</b>

**Table 4**

Change in volatility by the proposed adjustment method.

Variance	Actual Variance	Adjusted Variance	Change
TCS	376.8456651	245.3078093	-53.62%
Aluminum (AL)	6.969504787	5.86304104	-18.87%

has been applied to reduce the effect of corporate events on a financial time series.

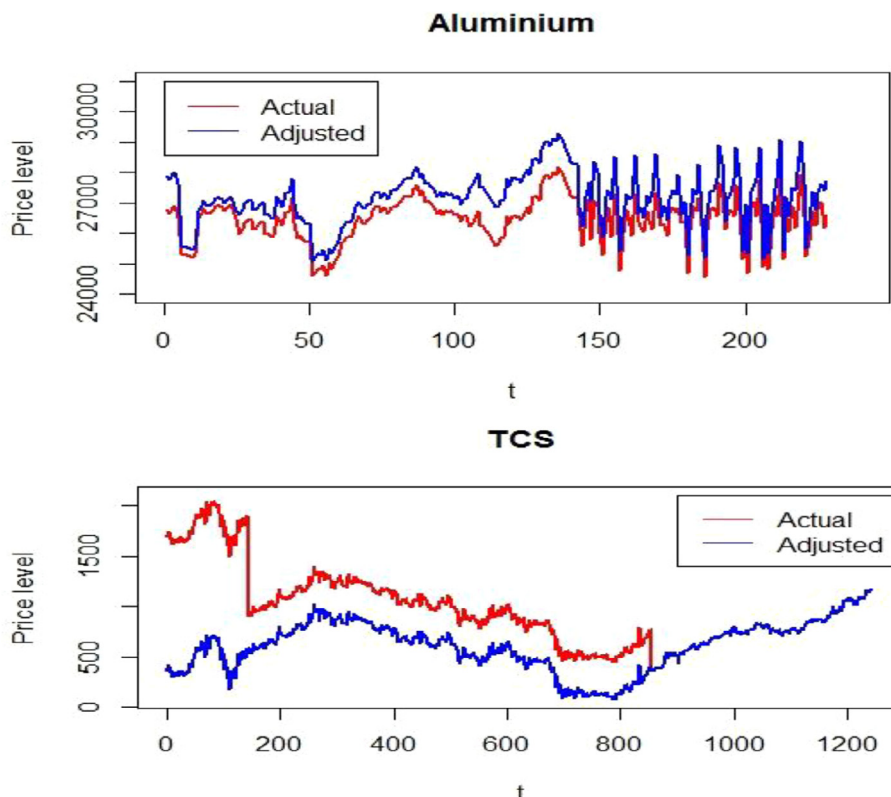
Let  $t_1, t_2, \dots, t_k$  are the dates of  $k$  corporate events of a financial instrument.  $\Delta p_i = \{p_{i+1} - p_i\}$ ,  $i = 1, 2, \dots, k - 1$  is the change in price due to the  $i^{th}$  corporate event. Then the method of adjusting historical prices for a financial instrument will be as below

- i) Calculate the price change factor at 1st date of corporate event  $\Delta p_{t_1} = p_{(t_1+1)} - p_{t_1}$

- ii) Adjust the historical prices till  $(t_2 + 1)$  with change factor  $\Delta p_{t_1}$ . Denote adjusted price on  $(t_2 + 1)$  as  $p'_{(t_2+1)}$  where  $p'_{(t_2+1)} = p_{(t_2+1)} + \Delta p_{t_1}$ .
- iii) The change factor on date  $t_2$  is calculated based on the adjusted price on  $(t_2 + 1)$  day (i.e.,  $p'_{(t_2+1)}$ ) and actual historical price on  $t_2^{th}$  day as  $\Delta p_{t_2} = p'_{(t_2+1)} - p_{t_2}$
- iv) Continue the process till the last date of corporate event.

The process of such an adjustment reduces the unwanted volatility in a financial instrument. Table 4 shows the change in volatility level due to the application of the method of the adjustment discussed above.

The changes in price level are also presented graphically in Fig 3. However, question may arise about the need of carrying out such adjustments. The change in price level due to nearing maturity or stock split is such that most of the outlier detecting

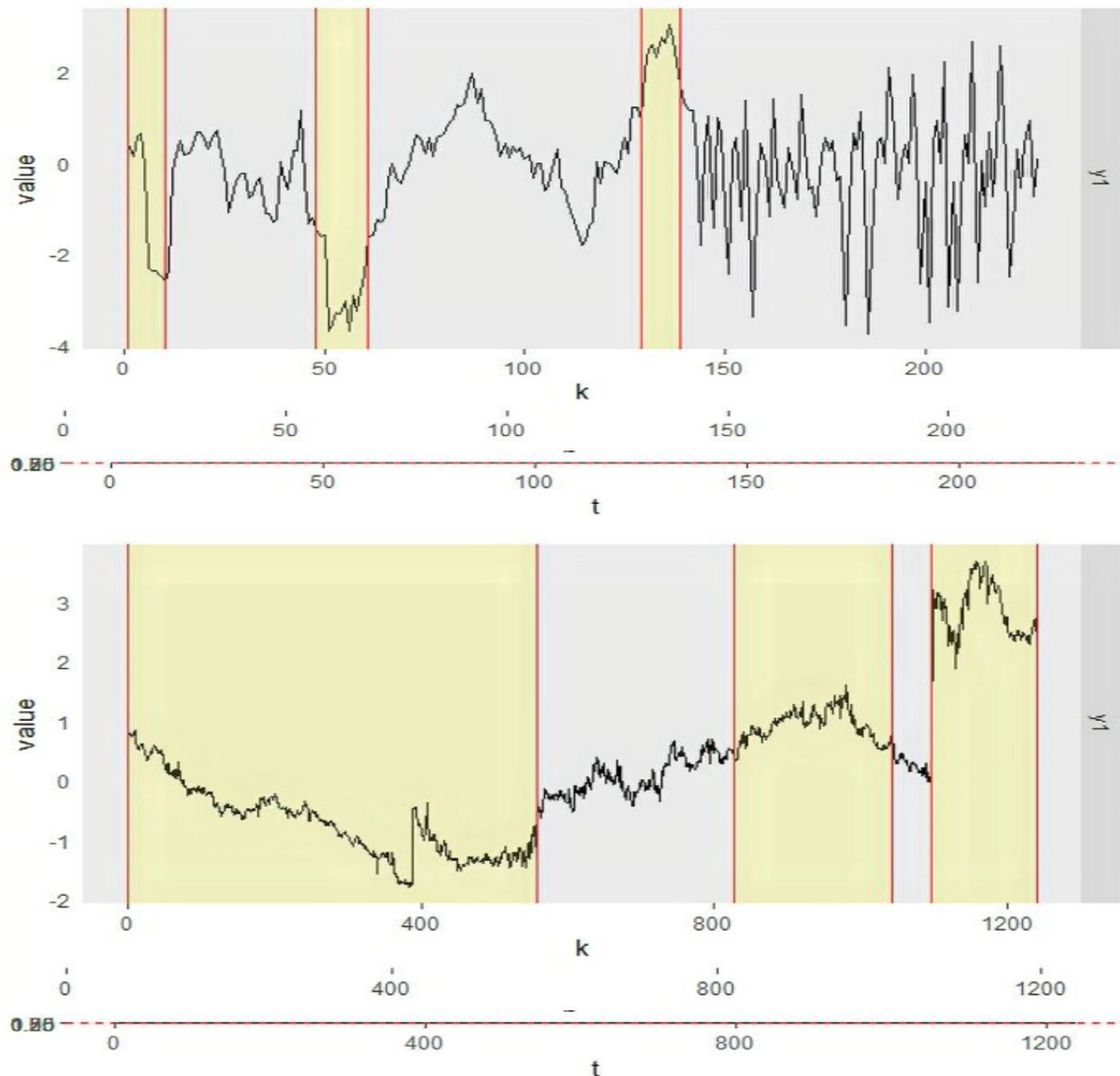


**Fig 3.** Plot showing the effect of adjustment method on the considered time series (TCS and aluminum).

**Table 5**

Correlation of predicted values with the original values and SE of predicted error for the time series with outliers (Original TS) and outlier free time series by different methods i.e., Chen and Liu, MHD, OTSAD, BARD and proposed on TCS share prices and Aluminum prices.

Series	Type	Method	Actual Mean	Predicted Mean	SE	Correlation			
Aluminum	Actual	Original TS	27311.44737	26664.00211	481.61788	-0.14135			
		Chen and Liu		26595.94526	478.51846	0.25158			
		MHD		26592.94132	475.70323	0.32630			
		OTSAD		26352.01000	501.82690	0.50943			
		BARD		26320.58150	711.37989	0.22274			
		Proposed		26716.24368	<b>455.33193</b>	<b>0.57230</b>			
	Adjusted	Original TS	28449.02632	27537.63711	479.11588	0.15110			
		Chen and Liu		27526.30395	476.03141	0.23661			
		MHD		27542.66553	479.93021	0.29479			
		OTSAD		27144.95800	641.88919	0.75588			
		BARD		27134.27400	809.64871	0.56824			
		Proposed		27547.86692	<b>476.20275</b>	<b>0.78854</b>			
		TCS		Actual	Original TS	1789.51413	1336.54828	195.98089	-0.18150
					Chen and Liu		1281.66712	158.14802	0.19871
MHD	888.15584		136.72612		0.07241				
OTSAD	1081.77220		135.56480		-0.82535				
BARD	998.70296		<b>130.85752</b>		-0.67576				
Proposed	1244.22465		143.71735		<b>0.12144</b>				
Adjusted	Original TS		469.81413	535.82003	139.54160	-0.21379			
	Chen and Liu			535.82003	139.54160	-0.21379			
	MHD			578.74642	139.56281	-0.18000			
	OTSAD			1061.27950	141.86397	-0.79451			
	BARD			698.65030	<b>125.66914</b>	0.10552			
	Proposed			587.47307	138.84780	<b>0.37043</b>			



**Fig 4.** Plot showing anomalies region detected by BARD on TCS and aluminum time series.



algorithm will detect such price changes as outliers. Actual outliers may remain in the data for the sake of corporate action and any analysis may come out with fallacious results. Indeed, the proposed adjustment method changes the historical prices relative to the current price level. After adjusting the time series, the proposed method of the outlier detection is applied to both adjusted and original time series, the results of which are presented Table 5. It may be noticed that the efficiency of BARD in some cases is better than the proposed method. The increase in efficiency of BARD is mainly due to the removal of the larger regions from the original data as depicted in Fig 4. The detected regions are presented in a different colour bounded by red lines. From Fig 4 it can be observed that, changes in the level of the data due to stock split and expiry of commodity contract dates are actually identified by BARD as the region of anomalies. However, mere elimination of such region may not be a proper justification. Thus, the method of adjustment discussed above may help in locating actual outliers.

## 6. Conclusion

In the present study, a multivariate outlier detection algorithm has been studied. A univariate time series is transformed to a bivariate data frame based on the robust estimate of lag. Different methods are applied to identify the outliers in time series. Feed forward neural network has been applied to the outlier free time series for model building. Comparison has been carried out based on correlation between predicted and actual values and SE of predicted errors. The discussed method of identifying outliers has been applied to ARMA process as well as to non-linear time series. The efficiency of the proposed algorithm is due to its ability to locate actual outliers unlike BARD and MHD which identify outliers as region and pairs, respectively. The performance of another method OSTAD in identifying outliers has also been compared. The performance of the proposed method is satisfying and may be handy in detecting outliers in time series. In addition, the proposed method does not require a priori knowledge of model parameters. Apart from the simulation examples, time series for TCS share price and commodity price are also considered. However, the proposed method is limited to univariate time series.

## Disclosure of any funding to the study

Elsawah received the partially support from UIC Grants (Nos. R201810, R201912 and R202010) and the Zhuhai Premier Discipline Grant to carry out this research work but not for APC.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Authors are heartily thankful to Editor-in-Chief Professor Omar Al-Dossary and two anonymous learned reviewers for their valuable comments which have made substantial improvement to bring the original manuscript to its present form. Authors are also thankful to Prof. Surajit Ray for his valuable suggestion in grammatical corrections. Elsawah is thankful for the UIC Grants and the Zhuhai Premier Discipline Grant for providing the partial support to carry out this research work.

## References

- Ahmad, S., Lavin, A., Purdy, S., Agha, Z., 2017. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing* 262, 134–147.
- Bardwell, L., Fearnhead, P., 2017. Bayesian detection of abnormal segments in multiple time series. *Bayesian Anal.* 12 (1), 193–218.
- Barnett, V., Lewis, T., 1994. *Outliers in Statistical Data*. John Wiley.
- Barreire, C., Boussouf, L., Cabon, B., Laurent, B., Loubes, J.M., 2019. Statistical methods for outlier detection in space telemetries. In: *Space Operations: Inspiring Humankind's Future*. Springer, Cham, pp. 513–547.
- Battaglia, F., Orfei, L., 2005. Outlier detection and estimation in nonlinear time series. *J. Time Ser. Anal.* 26, 107–121.
- Carreno, A., Inza, I., Lozano, J.A., 2019. Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artif. Intell. Rev.*, 1–20.
- Chang, I., Tiao, G.C., Chen, C., 1988. Estimation of time series parameters in the presence of outliers. *Technometrics* 30, 193–204.
- Chatterjee, S., Roy, A., 2014. Novel algorithms for web software fault prediction. *Qual. Reliab. Eng. Int.* <https://doi.org/10.1002/qre.1687>.
- Chen, C., Liu, L.M., 1993. Joint estimation of model parameters and outlier effects in time series. *J. Am. Stat. Assoc.* 88, 284–297.
- Denby, L., Martin, R.D., 1979. Robust estimation of the first-order autoregressive parameter. *J. Am. Stat. Assoc.* 88, 284–297.
- Deutsch, S.J., Richards, J.E., Swain, J.J., 1990. Effects of a single outlier on ARMA identification. *Commun. Stat.-Theory Methods* 19, 2207–2207.
- Faraway, J., Chatfield, C., 1998. Time series Forecasting with Neural network: a comprehensive study using the airline data. *Appl. Stat.* 47, 231–250.
- Filzmoser, P., Garrett, R.G., Reimann, C., 2005. Multivariate outlier detection in exploration geochemistry. *Comput. Geosci.* 31, 579–587.
- Fox, A.J., 1972. Outliers in time series. *J. Royal Stat. Soc. B* 34, 350–363.
- Hadi, A.S., 1994. A modification of a method for the detection of outliers in multivariate samples. *J. Royal Stat. Soc. Ser. B (Methodol.)* 56, 393–396.
- Hawkins, D., 1980. *Identification of Outliers*. Chapman and Hall.
- Higham, N.J., 2002. Computing the nearest correlation matrix – a problem from finance. *IMA J. Numer. Anal.* 22, 329–343.
- Hornik, K., 1991. Approximation capabilities of multilayer feed forward networks. *Neural Networks* 4, 251–257.
- Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 359–366.
- Huber, P.J., 1981. *Robust Statistics*. John Wiley and Sons, New York.
- Iturria, A., Carrasco, J., Charramendieta, S., Conde, A., Herrera, F., 2020. Otsad: A package for online time-series anomaly detectors. *Neurocomputing* 374, 49–53.
- Jeng, X.J., Cai, T.T., Li, H., 2013. Simultaneous discovery of rare and common segment variants. *Biometrika* 100 (1), 157–172.
- Johnson, R., 1992. *Applied Multivariate Statistical Analysis*. Prentice Hall.
- Khashei, M., Bijari, M., 2010. An artificial neural network (p, d, q) model for time series forecasting. *Expert Syst. Appl.* 37, 479–489.
- Ledolter, J., 1989. The effect of additive outliers on the forecasts from ARMA models. *Int. J. Forecast.* 5, 231–240.
- Leduca, C.L., Boistard, H., Moulines, E., Taqqu, M.S., Reiss, V.A., 2011. Robust estimation of the scale and of the auto covariance function of Gaussian short and long-range dependent processes. *J. Time Ser. Anal.* 32, 135–156.
- Liu, H., Shah, S., Jiang, W., 2004. On-line outlier detection and data cleaning. *Comput. Chem. Eng.* 28, 1635–1647.
- Loperfido, N., 2020. Kurtosis-based projection pursuit for outlier detection in financial time series. *Eur. J. Fin.* 26 (2–3), 142–164.
- Molinarese, F.F., Reiss, V.A., Cribari-Neto, F., 2009. Robust estimation in long-memory processes under additive outliers. *J. Stat. Plann. Inference* 139, 2511–2525.
- Omar, S., Ngadi, A., Jebur, H.H., 2013. Machine learning techniques for anomaly detection: an overview. *Int. J. Comput. Appl.* 79 (2), 33–37.
- Paul, C., Vishwakarma, G.K., 2017. Back propagation neural networks and multiple regressions in the case of heteroscedasticity. *Commun. Stat. – Simul. Comput.* 46 (9), 6772–6789.
- Pena, D., 1990. Influential observations in time series. *J. Bus. Econ. Stat.* 8, 235–241.
- Rousseeuw, P.J., 1984. Least median of square regression. *J. Am. Stat. Assoc.* 79, 871–880.
- Rousseeuw, P.J., Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* 85 (411), 633–651.
- Rousseeuw, P.J., Driessen, K.V., 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41, 212–223.
- Shaheed, M.H., 2005. Feedforward neural network based non-linear dynamic modeling of a TRMS using RPROP algorithm. *Aircraft Eng. Aerosp. Technol.* 77, 13–22.
- Terasvirta, T., Lin, C.F., Granger, C.W.J., 1993. power of the neural network linearity test. *J. Time Ser. Anal.* 14, 209–220.
- Tsay, R.S., 1988. Outliers, level shifts, and variance changes in time series. *J. Forecasting* 7, 1–20.
- Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: the state of the art. *Int. J. Forecast.* 14, 35–62.