Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2015, Article ID 708948, 9 pages
http://dx.doi.org/10.1155/2015/708948

Research Article

Data Transformation Technique to Improve the Outlier Detection Power of Grubbs' Test for Data Expected to Follow Linear Relation

K. K. L. B. Adikaram,[1,2,3] M. A. Hussein,[1] M. Effenberger,[2] and T. Becker[1]

[1]*Group Bio-Process Analysis Technology, Technische Universität München, Weihenstephaner Steig 20, 85354 Freising, Germany*
[2]*Institut für Landtechnik und Tierhaltung, Vöttinger Straße 36, 85354 Freising, Germany*
[3]*Computer Unit, Faculty of Agriculture, University of Ruhuna, Mapalana, 81100 Kamburupitiya, Sri Lanka*

Correspondence should be addressed to K. K. L. B. Adikaram; lasantha@daad-alumni.de

Received 9 September 2014; Revised 9 December 2014; Accepted 10 December 2014

Academic Editor: Carlos Conca

Grubbs test (extreme studentized deviate test, maximum normed residual test) is used in various fields to identify outliers in a data set, which are ranked in the order of $x_1 \le x_2 \le x_3 \le \cdots \le x_n$ ($i = 1, 2, 3, \ldots, n$). However, ranking of data eliminates the actual sequence of a data series, which is an important factor for determining outliers in some cases (e.g., time series). Thus in such a data set, Grubbs test will not identify outliers correctly. This paper introduces a technique for transforming data from sequence bound linear form to sequence unbound form ($y = c$). Applying Grubbs test to the new transformed data set detects outliers more accurately. In addition, the new technique improves the outlier detection capability of Grubbs test. Results show that, Grubbs test was capable of identifing outliers at significance level 0.01 after transformation, while it was unable to identify those prior to transforming at significance level 0.05.

1. Introduction

Grubbs test [1] is a statistical test used to detect outliers which was introduced in 1950 and extended in 1969 [2] and 1972 [3] by the same author. Grubbs test locates outliers that exist in a univariate data set using mean, standard deviation, and tabulated criterion. Grubbs test is also known as maximum normed residual test or "extreme studentized deviate" (ESD) test, and the data set is assumed to be normally distributed. The test is defined as

$$G = \frac{\max_{i=1,\ldots,N} \left| Y_i - \overline{Y} \right|}{s}, \tag{1}$$

where s is standard deviation and $\overline{Y}$ is the sample mean. If the maximum G related to the ith element is greater than the relevant tabulated criterion, then the element is considered an outlier. The testing procedure is continuing until no more outliers are detected. However Grubbs test is not recommended for detecting outliers for sample size of six or less. When the sample size is six or less, most of the times Grubbs test identified nonoutliers as outliers [4].

During the last decades Grubbs test was used to identify outliers in different disciplines [5–9]. Also, during the last decades pros and cons of Grubbs test were identified and were improved as well. In 1975 Rosner showed that Grubbs test (ESD) performs much better than studentized range methods and performs equally as Kurtosis and R-statistic methods [10]. In 1983 Rosner introduced an improved version of ESD as generalized extreme studentized deviate (GESD) test [11]. However, GESD does not work well when the sample size is less than 25 [11]. Brant in 1990 stated that the combination ESD rules and boxplot provide comparable performance [12]. On the other hand, it was shown that the standard deviation and mean are affected by two or more outliers; Grubbs test does not detect outliers [13] correctly. Also, if the standard deviation of the data set is too large or too small, the test will tend to detect false outliers and vice versa. This was overcome by setting a threshold value for

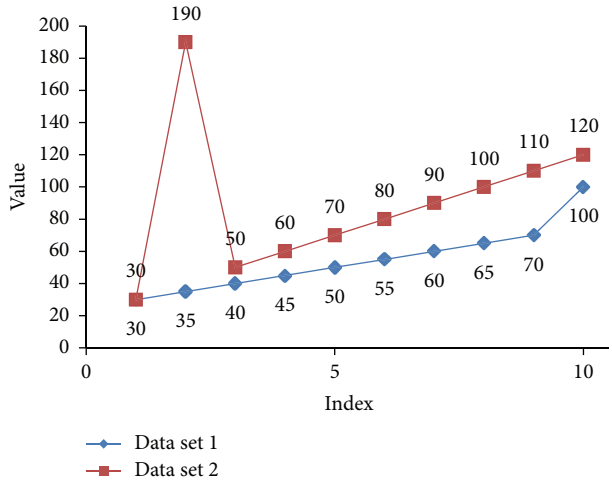

Figure 1: Two data sets, each containing one outlier which cannot be located by Grubbs test. The outlier in data set 1 is the last element of the series and considerably deviates from the expected value. The outlier in data set 2 is the maximum of the series, but it is not the first or last element.

standard deviations for the specific considered data domain [13]. Meanwhile, some publications show that Grubbs test is robust against the effect of intraclass correlation structure [14] and data that have Baldessari's structure [15, 16].

Form the definition of Grubbs test, it locates outliers in a data set which are ranked in the order of $x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n$. This implies that Grubbs test considers only the value of data points but not the real order of the data. In other words, Grubbs test treats sorted series and unsorted series in the same manner. Thus, Grubbs test is valid only for those data domains where the occurrence order is of no importance. However, with respect to outliers, the order of the data points is a very important factor for data series that are expected to have gradual increment or decrement over time. Thus, applying Grubbs test to data which has a relation with the occurrence order will not give a correct output. This is particularly important in the area of process control where the order of the data points has a very high impact on data interpretation. Therefore, Grubbs test is not a reliable method for detecting outliers in time series, because in time series occurrence order is a critical factor.

Grubbs test is capable of checking whether a certain suspected data point is an outlier. By default suspected points are the minimum and the maximum of the data set. If the most suspected data points are not outliers, Grubbs test does not identify other data points as outliers. Figure 1 shows two artificial data sets (data sets 1 and 2) with one outlier in each data set. Table 1 shows results of Grubbs test for two significance levels (α) of 0.05 and 0.01. In both data sets, the test does not detect outliers for any considered significance level. In data set 1, the outlier is not significant enough to be identified by Grubbs test. Although the outlier in data set 2 (190) deviated significantly in relation to its position, after ranking it moves to the end of the series and becomes an insignificant outlier.

The aim of this paper is to introduce a method for transforming "sequence bound" data into a "sequence unbound" form. Since the transformed series is totally independent of the sequence, applying Grubbs test could produce more robust results. Furthermore, the transformation increases the outlier detection capability of Grubbs test for data which are expected to have linear or nearly linear relation.

2. Methodology

Data transformation techniques are used to convert data status that is closer to the requirements of the technique or method to be applied [17]. The transformation process converts each data point of x_i into the transformed value y_i by means of a function f, where $y_i = f(x_i)$. Since Grubbs test is not suitable for detecting outliers in sequence bound series such as time series, one solution is to transform the sequence bound series into a sequence unbound series. In the domain of linear regression, any curve with the form of $y = c$, value of any data point is always constant. Therefore, any curve with the form of $y = c$ is a curve that is independent of the sequence.

Lemma 1. *If it is possible to find a proper reference curve f_R for any curve f_A which has the same domain as f_R, it is possible to transfer f_A into a constant.*

Proof. If f_A represents the curve of actual data and d is a constant, the function $f_{A'} = f_A + d$ has the same domain. However, $f_{A'}$ has a different range than f_A. If the curve of $f_A - f_{A'} = f_D$ then $f_D = d$. Since $f_{A'}$ is $f(f_A)$, f_D is also $f(f_A)$. Then, f_D can be considered as a transformation form of f_A, which is equal to a constant, and $f_{A'}$ can be considered as the f_R, which is the reference curve.

The curve f_D, which is the transformed form of f_A, has a simpler form than f_A. Also, f_D can be used to describe the behaviour of f_A. Because $f_D = d$, then f_D is independent of the sequence of the data. Therefore, because Grubbs test gives correct detections with the data sets that are independent of the occurrence sequence, f_D is a suitable data set, which can be tested with Grubbs test. Figures 2 and 3 illustrate usage of the above-mentioned concept for outlier detection. □

In the real world, it is not always possible to find the exact f_R for a certain data set in advance. Thus, if f_R has an approximate relation to the behaviour of the real data, then $f_D \approx c$ for all data elements (Figure 2). If the actual curve has abnormal data (outlier), f_D shows higher deviation from c (Figure 3). Applying Grubbs test to f_D the suspected element can be checked for an outlier.

When f_R is known in advance it is possible to apply this method for any data set of any form. f_R can be known in advance theoretically or by means of preknowledge of data. If these two options are not available, one possibility is to derive f_R from existing data of original data (f_A). This paper shows a method of deriving f_R for a data set that is expected to have linear ($y = mx + c$) form, using the original data (f_A).

Table 1: Results of Grubbs' test (two-sided) for data set 1 and data set 2 for significance levels (α) of 0.05 (critical value of $G = 2.29$) and $\alpha = 0.01$ ($G = 2.48$).

	Data set 1	G	Outlier?		Data set 2	G	Outlier?	
			$\alpha = 0.05$	$\alpha = 0.01$			$\alpha = 0.05$	$\alpha = 0.01$
1	30	1.22	No	No	30	1.34	No	No
2	35	0.98	No	No	190	2.24	No	No
3	40	0.73	No	No	50	0.89	No	No
4	45	0.49	No	No	60	0.67	No	No
5	50	0.24	No	No	70	0.45	No	No
6	55	0	No	No	80	0.22	No	No
7	60	0.24	No	No	90	0	No	No
8	65	0.49	No	No	100	0.22	No	No
9	70	0.73	No	No	110	0.45	No	No
10	100	2.20	No	No	120	0.67	No	No

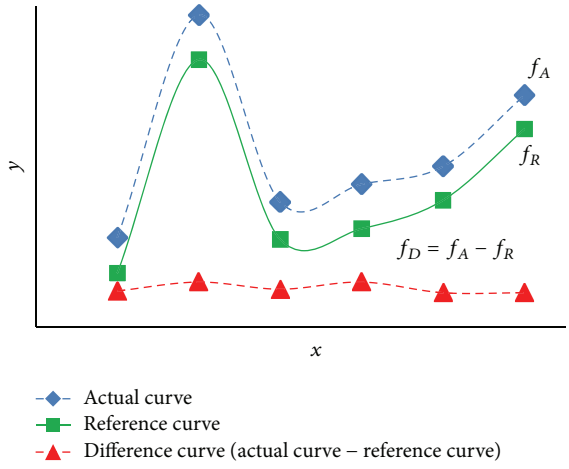

Figure 2: If f_A represents the curve of actual data and if it is possible to find a curve $f_D = f_A - f_R \approx c$ where f_R is a reference curve that represents actual behaviour of the real data, f_D reflects the existing outliers of f_A. When there are no outliers, difference curve (f_D) is always nearly a constant.

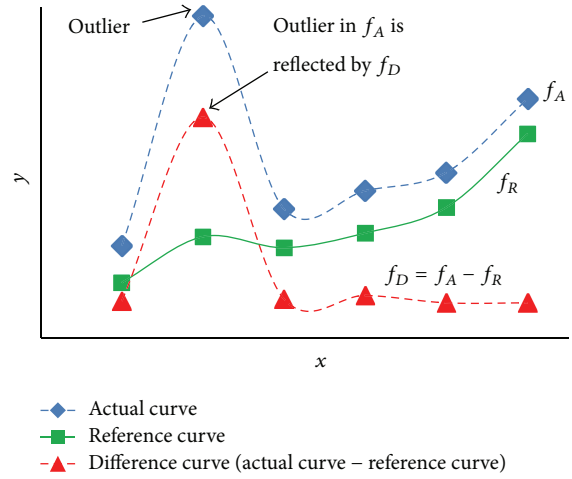

Figure 3: For an outlier, difference curve (f_D) shows significant deviation in relation to the outlier.

For any $f_A = mx + c$, the curve $f_{A'} = mx$ is a curve which has the same gradient as f_A, where m is the gradient of f_A. Then,

$$f_A - f_{A'} = mx + c - mx = c. \quad (2)$$

According to (2) the curve $f_A - f_{A'}$ is a constant. Therefore, for any linear function $f_A = mx + c$ the function $y = mx$ can be considered as the reference function f_R. In other words f_R is the curve, which goes through the origin with the same gradient as f_A. As shown in Figure 4, $y = mx$ form (f_R) can be considered as the transformation of $y = mx + c$ form (f_A). Because $f_A - f_R = c$ (or $y \approx c$), then $f_D = f_A - f_R$. Since the gradient of $f_R : m = \tan\theta$, it can be calculated either by using known theoretical and/or practical information or by deriving it from existing data. We focus on deriving the gradient of f_R using a part of original data (f_A).

When deriving any information from existing original data (f_A), the influence of outliers introduces distortions to the derived value. Outlier detections methods are used to remove such data points. However, when detecting outliers this is not a feasible solution. Therefore, when detecting outliers, the best solution is to exclude all suspected data points to minimize the influence of outliers to identify outliers.

Unlike most of outlier detection methods, Grubbs test always considers the maximum and the minimum as most suspected data points. Thus, we excluded the maximum and the minimum from the calculations. After removing the maximum and the minimum, the original series splits into a maximum of three small series (Figure 4: Segment 1, Segment 2, and Segment 3 of f_A). If there are equal maximum values and minimum values, the value with low index is considered as the maximum and the value with high index is considered as the minimum for a data series with increment. For a data series with decrement the value with high index is considered

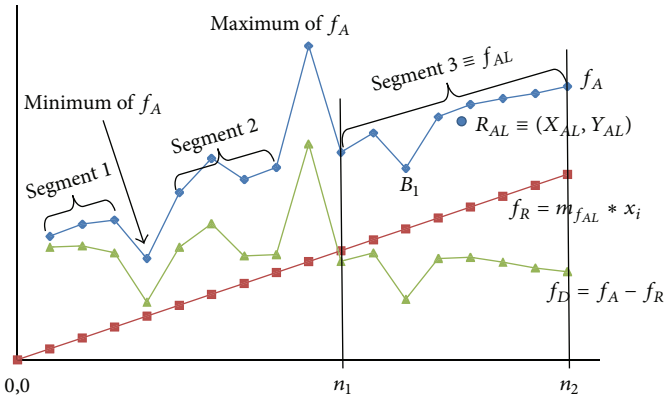

FIGURE 4: Transformation technique: if it is possible to find f_R which has the form of $y = mx$ in relation to f_A, then it is possible to find f_D as $f_A - f_R$.

as the maximum and the value with low index is considered as the minimum.

Then the series with the highest number of consecutive items (longest series) was considered for calculating the gradient of f_R. If the longest series of f_A is f_{AL} and $G_{AL} \equiv (X_{AL}, Y_{AL})$ then

$$Y_{AL} = \frac{\left(\sum_{i=n_1}^{n_2} y_i\right)}{(n_2 - n_1 + 1)}, \qquad X_{AL} = \frac{(n_1 + n_2)}{2}, \tag{3}$$

where n_1 is the starting index of the f_{AL} and n_2 is the end index of the f_{AL}.

If $P_i \equiv (x_{Ai}, y_{Ai})$ is any point of f_{AL}, then

$$m_i = \frac{(y_{Ai} - Y_{AL})}{(x_{Ai} - X_{AL})}, \tag{4}$$

where m_i is the gradient of f_{AL} at point i.

All P_i's on f_{AL} are not suspected data points and candidates for calculating the gradient of f_{AL}. However, among all the data points of f_{AL}, still it is not possible to determine the most suitable point for calculating m_i. If the selected P_i is a bias data point (e.g., point B_1 in Figure 4), it may introduce distortions to the calculated m_i even though it is not suspected. Therefore, it is necessary to have a more reliable method for calculating the gradient of f_{AL}. If the average of all gradients at all P_i's is considered, it will provide much better approximation for gradient instead of a gradient derived by referring to a certain single point. Therefore, if the resultant gradient of f_{AL} is $m_{f_{AL}}$, then $m_{f_{AL}}$ can be defined as the mean of all m_i's. Then,

$$m_{f_{AL}} = \frac{\sum_{i=n_1}^{n_2} m_i}{(n_2 - n_1 + 1)}. \tag{5}$$

From (4),

$$m_{f_{AL}} = \frac{\sum_{i=n_1}^{n_2} \left((y_{Ai} - Y_{AL})/(x_{Ai} - X_{AL})\right)}{(n_2 - n_1 + 1)}. \tag{6}$$

Because f_{AL} is the longest segment of f_A, $m_{f_{AL}}$(gradient of f_{AL}) is considered as the gradient of f_A. For the linear relation $y = mx$ form is the reference function (f_R). Therefore, gradient of f_R is the same as the gradient of f_{AL}. Then $f_R = m_{f_{AL}} * x_i$ for all i and $f_D = f_A - f_R$. According to Lemma 1 f_D is a constant and has the form of $y = d$, where d is a constant. Then the function f_D is the final transformation form, which is suitable for applying Grubbs test.

Finally, we applied Grubbs test on f_D and checked for outliers in the f_D. The existence of outliers in f_D confirms the existence of outliers in f_A. If the kth item of f_D is identified as an outlier, the kth item of f_A is considered as an outlier. Since f_R depends on x which is the index of the data point, f_D is also a function of x. This modification establishes a relation between data points and their index and eliminates the major problem identified for Grubbs test. After transformation Grubbs test can be applied repeatedly on f_D until no outliers were detected.

2.1. Evaluation Using Artificial Data. Four artificial data sets with one outlier in each data set (which cannot be identified by Grubbs test) were tested with the new method. Each data set consists of 10 elements with an outlier of different type, as mentioned in Table 2. Data sets 1 and 2 are the same data sets as in Table 1.

2.2. Evaluation Using Real Data. Real data sets collected from a biogas plant over a period of 60 days with a frequency of one data point per day were tested using both our transformation technique and standard Grubbs test. Among the different parameters, the counter reading of the electricity generator and the volumetric percentage of methane (CH4) in the biogas were selected for testing. During the stable situation, the counter reading of the electricity generator (operating hours) is continuously increasing, while the percentage of CH4 is fluctuating around a certain value. Both data sets were tested with the new technique and standard Grubbs test for the significance level of 0.05 and window sizes of 4, 5, 6, and 10 without overlapping. Also, Grubbs test was repeatedly applied until there were no outliers detected.

Table 2: Type of outlier contained in the data sets.

Data set name	Number of elements/outliers	Type of outlier
Data set 1	10/1	Last element of data set, considerably deviated
Data set 2	10/1	Maximum element of data set, not the last element
Data set 3	10/1	Neither the maximum nor the minimum; deviation is very small
Data set 4	10/1	Neither the first nor the last; data set is not continuously incrementing or decrementing

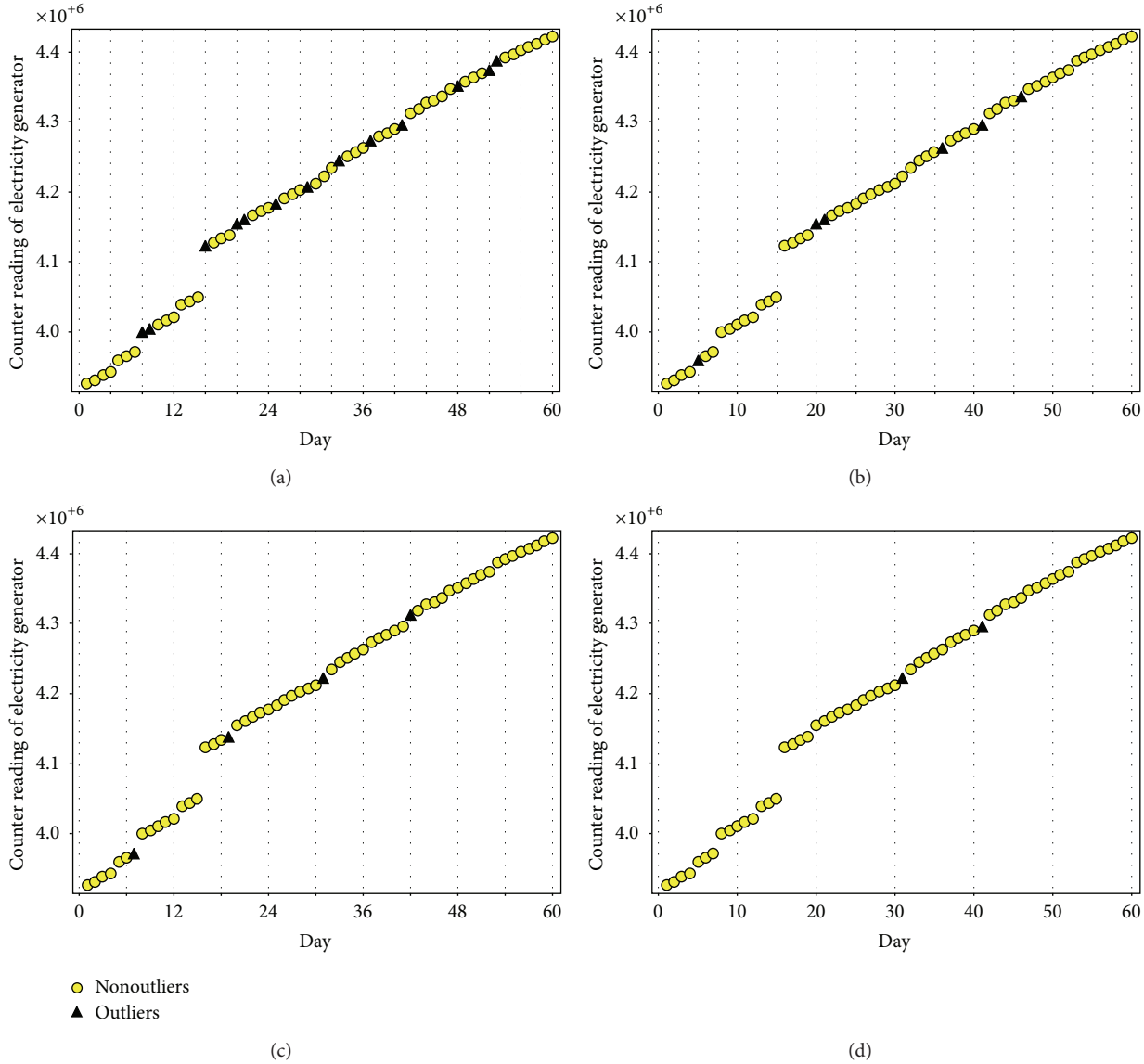

Figure 5: Outlier detection of the counter reading of electricity generator (operating hours) of the biogas plant with new transformation technique for significance level of 0.05 for different window sizes without overlap. Plots (a), (b), (c), and (d) correspond to window sizes 4, 5, 6, and 10, respectively.

3. Results and Discussion

The results from the test with artificial data show that applying Grubbs test on the transformed data set using our proposed method is capable of locating outliers at a significance level of 0.01 (Tables 3, 4, 5, and 6). When applied on the original data set, Grubbs test was unable to locate the outliers even with significance level of 0.05. The outlier in Table 5 deviates very little and is also neither the maximum nor the minimum, which is the worst case situation for single outlier domain. However, after transformation, Grubbs test identifies the outlier with a high level of confidence.

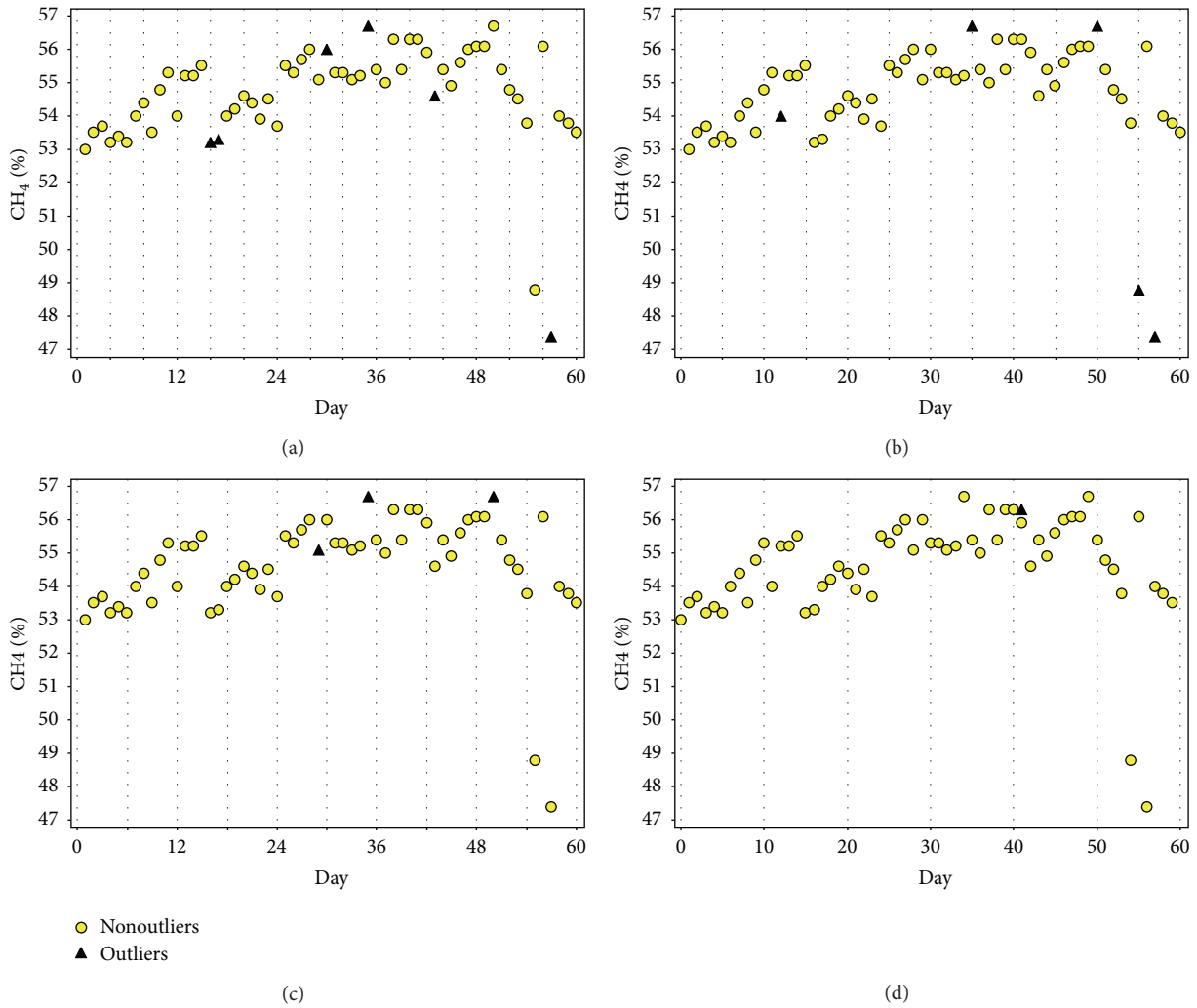

Figure 6: Outlier detection of volumetric percentage of CH4 in biogas with new transformation technique for significance level of 0.05 for different window sizes without overlap. Plots (a), (b), (c), and (d) correspond to window sizes 4, 5, 6, and 10, respectively.

Table 3: Data set 1 ($n = 10$): last element is the outlier and deviates considerably. With new transformation technique, Grubbs' test on f_D is capable of identifying the outlier with significance level of 0.01. However, Grubbs' test on f_A is capable of identifying the outlier with even significance level of 0.05.

x_i	$f_A = y_i$	f_{AL}	$m_i = \frac{y_i - Y_{AL}}{x_i - X_{AL}}$	$f_R = m_{f_{AL}} * x_i$	$f_D = f_A - f_R$	Grubbs' test on f_D ($\alpha = 0.01$)	Grubbs' test on f_A ($\alpha = 0.05$)
1	30	—	—	5	25	No	No
2 (n_1)	35	35	5	10	25	No	No
3	40	40	5	15	25	No	No
4	45	45	5	20	25	No	No
5	50	50	5	25	25	No	No
6	55	55	5	30	25	No	No
7	60	60	5	35	25	No	No
8	65	65	5	40	25	No	No
9 (n_2)	70	70	5	45	25	No	No
10	**100**	—	—	50	**50**	**Yes**	**No**

$X_{AL} = \frac{n_1 + n_2}{2} = 5.5$ $\quad Y_{AL} = \frac{\sum_{i=n_1}^{n_2} y_i}{n_2 - n_1 + 1} = 52.5$ $\quad m_{f_{AL}} = \frac{\sum_{i=n_1}^{n_2} m_i}{n_2 - n_1 + 1} = 5$

Table 4: Data set 2 ($n = 10$): maximum element of the data set is the outlier, but it is not the first or the last element. With new transformation technique, Grubbs' test on f_D is capable of identifying the outlier with significance level of 0.01. However, Grubbs' test on f_A is capable of identifying the outlier with even significance level of 0.05.

x_i	$f_A = y_i$	f_{AL}	$m_i = \frac{y_i - Y_{AL}}{x_i - X_{AL}}$	$f_R = m_{f_{AL}} * x_i$	$f_D = f_A - f_R$	Grubbs' test on f_D ($\alpha = 0.01$)	Grubbs' test on f_A ($\alpha = 0.05$)
1	30	—	—	10	20	No	No
2	20	—	—	20	0	No	No
3	50	—	—	30	20	No	No
4	**190**	—	—	40	**150**	**Yes**	**No**
5 (n_1)	70	70	10	50	20	No	No
6	80	80	10	60	20	No	No
7	90	90	10	70	20	No	No
8	100	100	10	80	20	No	No
9	110	110	10	90	20	No	No
10 (n_2)	120	120	10	100	20	No	No

$X_{AL} = \frac{n_1 + n_2}{2} = 7.5$ $Y_{AL} = \frac{\sum_{i=n_1}^{n_2} y_i}{n_2 - n_1 + 1} = 95$ $m_{f_{AL}} = \frac{\sum_{i=n_1}^{n_2} m_i}{n_2 - n_1 + 1} = 10$

Table 5: Data set 3 ($n = 10$): neither the maximum nor the minimum is the outlier; deviation is very small. With new transformation technique, Grubbs' test on f_D is capable of identifying the outlier with significance level of 0.01. However, Grubbs' test on f_A is capable of identifying the outlier with even significance level of 0.05.

x_i	$f_A = y_i$	f_{AL}	$m_i = \frac{y_i - Y_{AL}}{x_i - X_{AL}}$	$f_R = m_{f_{AL}} * x_i$	$f_D = f_A - f_R$	Grubbs' test on f_D ($\alpha = 0.01$)	Grubbs' test on f_A ($\alpha = 0.05$)
1	30.0	—	—	9.999996	20.00000	No	No
2 (n_1)	**40.0001**	40.0001	9.999975	19.999993	**20.00011**	**Yes**	**No**
3	50.0	50.0	10.000005	29.999989	20.00001	No	No
4	60.0	60.0	10.000008	39.999986	20.00001	No	No
5	70.0	70.0	10.000025	49.999982	20.00002	No	No
6	80.0	80.0	9.999975	59.999979	20.00002	No	No
7	90.0	90.0	9.9999917	69.999975	20.00003	No	No
8	100.0	100.0	9.999995	79.999971	20.00003	No	No
9 (n_2)	110.0	110.0	9.9999964	89.999968	20.00003	No	No
10	120.0	—	—	99.999964	20.00004	No	No

$X_{AL} = \frac{n_1 + n_2}{2} = 5.5$ $Y_{AL} = \frac{\sum_{i=n_1}^{n_2} y_i}{n_2 - n_1 + 1} = 75$ $m_{f_{AL}} = \frac{\sum_{i=n_1}^{n_2} m_i}{n_2 - n_1 + 1} = 9.999996$

Even though the data set in Table 6 has no continuous increment or decrement, Grubbs test located the outlier after transformation.

For the real data sets, the results show that our transformation technique is capable of identifying the outliers depending on the selected window size (Figures 5 and 6). The data points in most of the selected windows of Figure 5 consist of values that slightly deviated from the actual value. After transformation, Grubbs test was able to locate those data points. However, standard Grubbs test was unable to locate any of those points as outliers from both data sets for the same window sizes.

The data series shown in Figure 6 is not a linear series. However, application of windowing technique allowed locating outliers in each window using new transformation method. However, standard Grubbs test was unable to locate the outliers in the same windows. Furthermore, the data points shown in different window sizes in Figure 6 have different forms (increment, decrement, or constant). After transformation Grubbs test located the outliers despite the behaviour of data in the selected window. Another important fact is that the located outliers were outliers in relation to the selected window size and linear relation. Finally, the results show the capability of applying Grubbs test after

Table 6: Data set 4 ($n = 10$): data set is not continuously incrementing or decrementing. Outlier is neither the first nor the last element. With new transformation technique, Grubbs' test on f_D is capable of identifying the outlier with significance level of 0.01.

x_i	$f_A = y_i$	f_{AL}	$m_i = \frac{y_i - Y_{AL}}{x_i - X_{AL}}$	$f_R = m_{f_{AL}} * x_i$	$f_D = f_A - f_R$	Grubbs' test on f_D ($\alpha = 0.01$)	Grubbs' test on f_A ($\alpha = 0.05$)
1	30	—	—	4.667	25.333	No	No
2	28	—	—	9.333	18.667	No	No
3	40	—	—	14	26	No	No
4	**76**	—	—	**18.667**	**57.333**	**Yes**	**No**
5 (n_1)	51	51	4.8	23.333	27.667	No	No
6	54	54	6	28	26	No	No
7	62	62	2	32.667	29.333	No	No
8	66	66	6	37.333	28.667	No	No
9	69	69	4	42	27	No	No
10 (n_2)	76	76	5.2	46.667	29.333	No	No

$X_{AL} = \frac{n_1 + n_2}{2} = 7.5$ $Y_{AL} = \frac{\sum_{i=n_1}^{n_2} y_i}{n_2 - n_1 + 1} = 63$ $m_{f_{AL}} = \frac{\sum_{i=n_1}^{n_2} m_i}{n_2 - n_1 + 1} = 4.667$

transforming the series with new transformation and suitable windowing technique. Thus, the method can be used for locating outliers in time series regardless of the fact that the series is linear or nonlinear. However, still each window is considered as a window containing a linear segment of the curve.

According to the generally accepted idea, Grubbs test is not suitable for locating outliers in a data set with six or fewer terms [4]. However, the results show that after transforming with new method, Grubbs test was capable of locating outliers in the data sets with four and six terms (window sizes four and six). This is in disagreement with the generally accepted idea. In particular, when applying Grubbs test on nonlinear data series it is necessary to apply suitable windowing technique for having data windows which has better approximation for linearity (Figure 6). Therefore, we can state that the new transformation technique eliminates one of the major drawbacks that prevent applying Grubbs test on small windows.

The accuracy and the reliability of the transformation totally depend on the gradient (m) of f_R. Therefore, applying better method could give much better approximation for m. We considered other statistical properties such as mode and the median of the series as well as the longest segment for deriving m. We excluded the median because it is a single data point. The problem of any single data point is that it is not a reliable data point as a reference data point. Even though the considered single point is neither the maximum nor the minimum, it can be a deviated data point such as B_1 in Figure 4. Not like median the mode of a series represents multiple data points. Therefore, the mode can be considered as a good alternative for a data set expected to follow the form of $y = c$. Unfortunately, the mode cannot be used for a data series expected to follow the form of $y = mx + c$ (increasing or decreasing), because in such a data series it is not possible to expect multiple equal or nearly equal values. Finally, in general we decided to use mean for deriving m. However, if the considered domain ensures the reliability, it is possible to use any other method for deriving m, rather than the method we used. For example, if there is a guarantee of accuracy of a certain data point, even any single data point (such as the first data point of the series) can be used for deriving m.

In this paper we used the longest segment of the data set for deriving m. However, if there are considerable numbers of data points in other segments it is possible to calculate the gradient of other segments and get the average gradient of considered segments as the gradient. On the other hand, if the outliers were clustered and located in the longest segment, the method we mentioned in this paper will not give a better approximation for m due to the influence of outliers. However, if the considered domain is having or expected to have clustered outlier, then excluding the whole cluster before calculating m will give a better approximation for m. One possibility is to remove k nearest neighbours of the maximum and the minimum including the maximum and the minimum. This will provide much better data set for deriving m.

4. Conclusion

The results for artificial and real data show that our new transformation technique improves the outlier detection power of Grubbs test. The transformation is independent of already existing reference data sets and derived reference set from the part of the original data set. This is the main advantage of the new method. After transformation, Grubbs test was capable of detecting outliers at the significance level 0.01 which were not identified without transformation, even at the significance level 0.05. Also, after transformation, Grubbs test was capable of locating outliers in a data set that is not in ranked order, since the new technique transforms data from the form $y = mx + c$ to the form $y = c$ which is independent of the sequence.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The University of Ruhuna, Sri Lanka, provided the article processing charges of this paper. The German Academic Exchange Service (German: Deutscher Akademischer Austauschdienst) financed this work.

References

[1] F. E. Grubbs, "Sample criteria for testing outlying observations," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 27–58, 1950.

[2] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[3] F. E. Grubbs and G. Beck, "Extension of sample sizes and percentage points for significance tests of outlying observations," *Technometrics*, vol. 14, pp. 847–854, 1972.

[4] M. Thompson and P. J. Lowthian, *Notes on Statistics and Data Quality for Analytical Chemists*, Imperial College Press, 2011.

[5] S. Geisser, "Influential observations, diagnostics and discovery tests," *Journal of Applied Statistics*, vol. 14, no. 2, pp. 133–142, 1987.

[6] W.-K. Fung, "A statistical-test-complemented graphical method for detecting multiple outliers in two-way tables," *Journal of Applied Statistics*, vol. 18, no. 2, pp. 265–274, 1991.

[7] B. M. Colosimo, R. Pan, and E. del Castillo, "A sequential Markov chain Monte Carlo approach to set-up adjustment of a process over a set of lots," *Journal of Applied Statistics*, vol. 31, no. 5, pp. 499–520, 2004.

[8] M. K. Solak, "Detection of multiple outliers in univariate data sets," Paper SP06-2009, Schering, 2009.

[9] R. B. Jain, "A recursive version of Grubbs' test for detecting multiple outliers in environmental and chemical data," *Clinical Biochemistry*, vol. 43, no. 12, pp. 1030–1033, 2010.

[10] B. Rosner, "On the detection of many outliers," *Technometrics*, vol. 17, pp. 221–227, 1975.

[11] B. Rosner, "Percentage points for a generalized ESD many-outlier procedure," *Technometrics*, vol. 25, no. 2, pp. 165–172, 1983.

[12] R. Brant, "Comparing classical and resistant outlier rules," *Journal of the American Statistical Association*, vol. 85, no. 412, pp. 1083–1090, 1990.

[13] L. Xu, P. Zhang, J. Xu, S. Wu, G. Han, and D. Xu, "Conflict analysis of multi-source SST distribution," in *High Performance Computing and Applications*, W. Zhang, Z. Chen, C. C. Douglas, and W. Tong, Eds., pp. 479–484, Springer, Berlin, Germany, 2010.

[14] M. S. Srivastava, "Effect of equicorrelation in detecting a spurious observation," *The Canadian Journal of Statistics*, vol. 8, no. 2, pp. 249–251, 1980.

[15] D. M. Young, R. Pavur, and V. R. Marco, "On the effect of correlation and unequal variances in detecting a spurious observation," *The Canadian Journal of Statistics*, vol. 17, no. 1, pp. 103–105, 1989.

[16] J. K. Baksalary and S. Puntanen, "A complete solution to the problem of robustness of Grubbs's test," *The Canadian Journal of Statistics*, vol. 18, no. 3, pp. 285–287, 1990.

[17] O. H. J. Christie and K. H. Alfsen, "Data transformation as a means to obtain reliable consensus values for reference materials," *Geostandards and Geoanalytical Research*, vol. 1, no. 1, pp. 47–49, 1977.