EARTH OBSERVATION AND GEOMATICS ENGINEERING
website: https://eoge.ut.ac.ir

A modified wavelet-based method for detection of outliers in time series

Amirreza Moradi[1]*, Sajjad Asiaei Mojarad[1]

[1] Department of Surveying Engineering, Arak University of Technology, Arak, Iran

Article history:
Received: 30 October 2018, Received in revised form: 12 April 2019, Accepted: 18 April 2019

ABSTRACT

As a multi-resolution analysis, wavelet transformation tool has been used to detect contingent outliers in time series data with no need to specify a model for the data. The objective of this article is to design an orthonormal wavelet system that optimizes the wavelet-based outlier detection procedure. In addition, we show that regardless of the selected base functions, the existing wavelet-based methods extract two adjacent suspicious observations so that probably one of them is an outlier. Therefore, we modify the wavelet-based outlier detection scheme by introducing a transformation matrix consisting of our designed wavelet filters that can be used to detect outlying observations without the above-mentioned ambiguity.

In a numerical example, a sample observation vector is analyzed using our scheme. At the same time, a robust statistical approach- modified z-score method- has been used to evaluate the capability of our desired wavelet-based procedure. The results were completely reliable and comparable.

KEYWORDS
Wavelet Transform
Outlier Detection
Time Series

1. Introduction

Physical quantities belong to infinite-dimensional spaces; this means, in order to determine the true value of any given quantity, one should measure it indefinitely. However, in practice, only a sample of a population corresponding to the desired quantity will be used to estimate the most probable approximation to the truth. One of the first and important tasks to accomplish when processing observations or time-series data is the detection of outlying observations. According to a general definition, an outlier is an observation in a dataset which appears to be inconsistent with the remainder of that set of data (Wichern & Johnson, 1992).

So far, several methods have been proposed to detect outlying observations. Malik et al. (2014), comparatively analyzed the main outlier detection techniques. In general terms, these methods can be grouped into two categories: parametric (statistical) and nonparametric approaches. Some of these schemes are robust techniques since they are minimally affected by outliers.

Wavelet transformation, as an efficient tool for decomposing time series into fine and coarse parts, was primarily used to detect outliers in time series by Bilen & Huzurbazar (2002), with no need to fit an initial model to the observations. They proposed a wavelet based outlier detection procedure, which used the wavelet coefficients resulting from the Discrete Wavelet Transform (DWT) of the time series, based on Haar wavelet and scaling functions.

The powerful properties of wavelet transformation for detection of outliers such as robustness, simplicity, fastness, and automatability have led to the use and development of the wavelet-based methods for this purpose (see for example, Barua & Alhajj, 2007; Grané & Veiga, 2010; Silva, I. & Silva, M.E., 2016; Ranta et al., 2005).

One of the powerful points of the wavelet transform is that the wavelets, as the base functions, are adjustable and adaptable. In other words, the wavelet transform is an infinite set of various transforms, depending on the wavelet used for its computation (Burrus et al., 1997). Unlike many other signal expansion systems, the wavelets can be designed to fit individual applications. Burrus et al. (1997) introduced the

* Corresponding author
E-mail addresses: a-moradi@arakut.ac.ir (A. Moradi); sajjad74asiaie@gmail.com (S. Asiaei.M)

DOI: 10.22059/eoge.2019.285487.1054

most commonly-used wavelet families, such as Daubechies, Coiflets and, so on, each giving the signal representation, decomposition, or compression from different points of view. All aforementioned wavelet-based outlier detection studies made use of Daubechies wavelets because of yielding more significant wavelet coefficients at times corresponding to outliers in the original series. In this research, we design an orthonormal wavelet system that improves the ability to maximize the wavelet coefficients at times where there are outliers. Moreover, the existing wavelet-based methods extract two adjacent suspicious observations so that probably one of them is an outlier. Typically, this ambiguity needs to be resolved in a decision-making process. Accordingly, we modify the wavelet-based outlier detection scheme, by introducing a transformation matrix consisting of our designed wavelet filters that can be used to detect outlying observations without the above-mentioned inexactness.

The rest of this paper is organized as follows. In section 2, we briefly review some foundations of the wavelet transform. In section 3, we describe how to design an appropriate wavelet system for detection of outlying observations, and then present our modification approach to detect outliers without ambiguity. In Section 4, a numerical evaluation based on the proposed method is given. At the same time, a robust statistical approach- box plot method- is used to evaluate the capability of our modified wavelet-based scheme that shows the effectiveness of our desired scheme. Finally, we conclude the paper in Section 5.

2. A brief background on wavelets and wavelet transformation

According to wavelet terminology (Burrus et al., 1997), the wavelet expansion of a function *f(t)* will be

$$f(t) = \sum_{k=-\infty}^{+\infty} \sum_{j=-\infty}^{+\infty} d_{j,k}\, \psi_{j,k}(t) \qquad (1)$$

Where $d_{j,k}$ s are the expansion coefficients at scale j and time translation k and the wavelets $\psi_{j,k}(t)$, as the base functions are constructed by translating and stretching the main producer function, known as mother wavelet $\psi(t)$

$$\psi_{j,k}(t) = \psi\left(2^{j} t - k\right) \qquad (2)$$

The coefficients $d_{j,k}$ in the wavelet expansion (1) are called the discrete wavelet transform (DWT) of the signal $f(t)$.

In order to separate the coarse and fine parts of the analyzed signal, one may use some other base functions called scaling functions$\varphi_{k}(t)$, along with wavelets, as follows

$$f(t) = \sum_{k=-\infty}^{+\infty} c_k \varphi_k(t) + \sum_{k=-\infty}^{+\infty} \sum_{j=0}^{+\infty} d_{j,k} \psi\left(2^{j} t - k\right) \qquad (3)$$

where, the scaling functions $\varphi_{k}(t)$ are translated versions of a basic scaling function $\varphi(t)$ as

$$\varphi_{k}(t) = \varphi(t-k) \qquad (4)$$

In expression (3), the first part consists of the coarse information of the function and the details are reflected in the second part; such that the bigger j, s the more obtained details.

The coefficients in expressions (1) and (3) are called the discrete wavelet transform (DWT) of the signal. If the base functions construct an orthonormal basis, the expansion coefficients are computed using the inner products of $f(t)$ with the corresponding wavelet or scaling functions.

In practice, the fast and simple discrete wavelet transforms and their inverts can be implemented based on the multiresolution analysis proposed by Mallat (1989) and improved by Beylkin et al. (1991) in which, one does not need to deal directly with the scaling functions and wavelets. Instead, the wavelet-based computed fine and coarse parts of a time series can be computed using highpass (wavelet) and lowpass (scaling) filters, with impulse responses $h(n)$ and $g(n)$,respectively. Li (1996) briefly explained these procedures for a given sequence of a signal. The relation between these scaling filters can be obtained as follows (Li, 1996)

$$g(n) = (-1)^{n-1} h(-n-1) \qquad (5)$$

One of the attractive and unique features of the wavelet transform is the possibility of designing the scaling and wavelet filters to meet some desired properties, after satisfying the necessary conditions for the existence and orthogonality of the base functions. The following equations will guarantee the existence and orthogonality conditions, respectively (Burrus et al., 1997)

$$\sum_{n=1}^{N} h(n) = \sqrt{2} \qquad (6)$$

$$\sum_{n=1}^{N} h(n) h(n-2k) = \begin{cases} 1 & \text{if } k=0 \\ 0 & \text{Otherwise} \end{cases} \qquad (7)$$

where *N* is the support or length of $h(n)$, which must be an even number (Burrus et al., 1997).

Expressions (6) and (7) state that if the existence and orthogonality of the base functions is fulfilled, $\frac{N}{2}+1$ equations must be satisfied, which leaves $\frac{N}{2}-1$ degrees of freedom to design the N values of scaling filters.

3. System design and modification for detection of outliers

The first wavelet system with the possibility of designing the filtering coefficients has a length of N=4, because in this

case, there is one degree of freedom after satisfying the constraints of equations (6) and (7). Let angle α be the parameter corresponding to the remaining degree of freedom. The following scaling filters will be achieved as (Burrus et al., 1997)

$$\begin{cases} h(0) = \dfrac{1 - \cos(\alpha) + \sin(\alpha)}{2\sqrt{2}} \\ h(1) = \dfrac{1 + \cos(\alpha) + \sin(\alpha)}{2\sqrt{2}} \\ h(2) = \dfrac{1 + \cos(\alpha) - \sin(\alpha)}{2\sqrt{2}} \\ h(3) = \dfrac{1 - \cos(\alpha) - \sin(\alpha)}{2\sqrt{2}} \end{cases} \quad (8)$$

The famous Daubechies-4 wavelet system was obtained for $\alpha=\pi/3$, in this category. This system belongs to a common family of wavelets designed by Daubechies (1992) to obtain maximum regularity for a given N. In the present research, we use the remaining degree of freedom to design a length-4 wavelet system to effectively detect the outlying observations in time series datasets.

The logic of wavelet based outlier detection approaches is that the estimated fine part of the signal, as a result of the discrete wavelet transform, yields the coefficients that are expected to be large in magnitude at times where there are jumps or outliers in the time series. Our procedure is based on designing a wavelet system that provides the best performance in magnifying outliers in the fine part of the time series being studied. Our process of detecting outliers uses the only remaining degree of freedom-the angle α- to achieve the optimal results.

Considering the forward and inverse one-dimensional discrete wavelet transforms converting any given signal from the time domain to the time-frequency domain and vice versa, the images of coarse and fine parts of the observation vector in the time domain are computed by multiplying the observation vector by two products H2×H1 and G2×G1, respectively; where H1, H2, G1 and G2 are the forward and inverse wavelet transforming matrices consisting of wavelet and scaling filter coefficients. Our criterion to distinguish the probable outliers is based on the quantity of the elements of the image of the fine part. Therefore, the structure of the matrix of transformation used for calculating the fine part of the time series in the time-space is the key point of the optimum wavelet system identification for outlier detection. If we consider the vector of observations as an n-by-1 vector, the mentioned operator will be a square matrix of order n and every element of the detail vector in the time domain is the result of inner production of the corresponding row of the transformation matrix and the observation vector; accordingly, the best wavelet system is the one producing the largest peaks located at the positions corresponding to the row number of the vector of observations. The biggest element of the jth row of the transformation matrix should be the jth member of that row and the others should be close to zero.

Considering the structure of forward and inverse transform matrices (Keller, 2004), the fine part of the time series in the time-space can be directly computed using the transformation matrix G2×G1 as follows:

$$\begin{bmatrix} h_1^2+h_3^2 & -h_0h_1-h_2h_3 & h_1h_3 & -h_0h_3 & 0 & 0 & \cdots & & & & & h_1h_3 & -h_1h_2 \\ -h_0h_1-h_2h_3 & h_0^2+h_2^2 & -h_1h_2 & h_0h_2 & 0 & 0 & & & & & & -h_0h_3 & h_0h_2 \\ h_1h_3 & -h_1h_2 & h_1^2+h_3^2 & -h_0h_1-h_2h_3 & h_1h_3 & -h_0h_3 & \cdots & & & & & & \\ -h_0h_3 & h_0h_2 & -h_0h_1-h_2h_3 & h_0^2+h_2^2 & -h_1h_2 & h_0h_2 & & & & & & & \\ \vdots & & & & & & \ddots & & & & & & \vdots \\ & & & & & & & h_1h_3 & -h_1h_2 & h_1^2+h_3^2 & -h_0h_1-h_2h_3 & h_1h_3 & -h_0h_3 \\ & & & & & & \cdots & -h_0h_3 & h_0h_2 & -h_0h_1-h_2h_3 & h_0^2+h_2^2 & -h_1h_2 & h_0h_2 \\ h_1h_3 & -h_0h_3 & & & & & & 0 & 0 & h_1h_3 & -h_1h_2 & h_1^2+h_3^2 & -h_0h_1-h_2h_3 \\ -h_1h_2 & h_0h_2 & & & & & \cdots & 0 & 0 & -h_0h_3 & h_0h_2 & -h_0h_1-h_2h_3 & h_0^2+h_2^2 \end{bmatrix} \quad (9)$$

The optimum wavelet system is obtained for $\alpha= -\pi/4$, that maximizes the diagonal and minimize the off-diagonal elements of the transformation matrix. Four sample consecutive rows of the above transformation matrix are quantitatively compared in Figure 1.

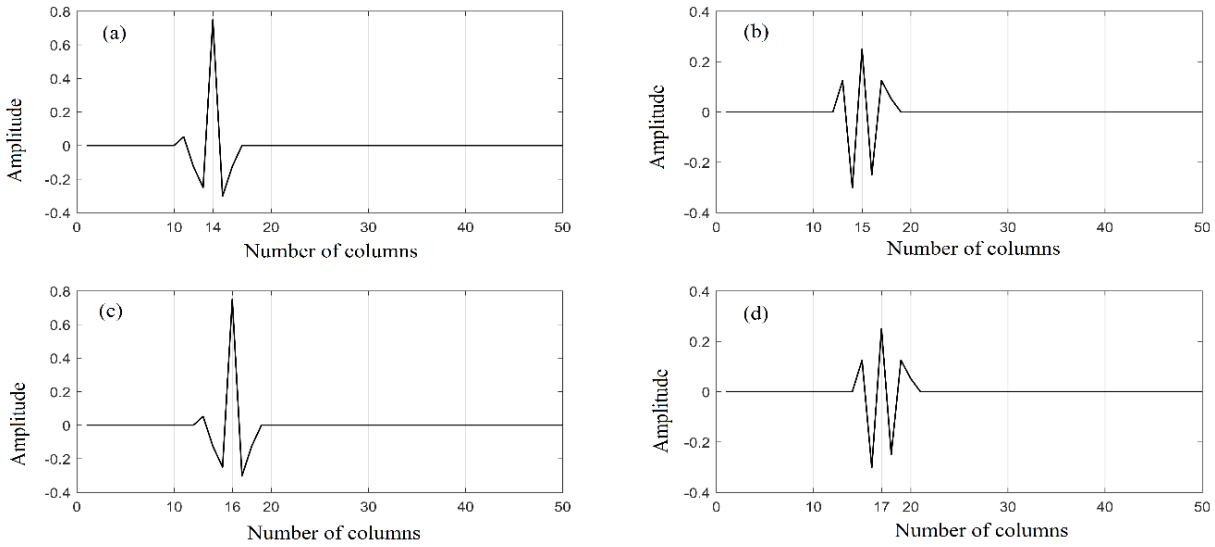

Figure 1. Four sample rows of the transformation matrix G2ˣG1, for the designed length-4 wavelets with α=-π/4, (a) the 14th row, (b) the 15th row, (c) the 16th row and (d) the 17th row

It can be noted that the format of these rows, with the largest peaks located at the positions corresponding to the number of the row, is suitable for analyzing the vector of observation for detecting the outliers. However, the presence of relatively large elements next to the peaks in even rows leads to an ambiguity in detecting probable outliers in some cases. This ambiguity can be resolved by comparing the suspicious observation with the sample mean of the original series (Bilen and Huzurbazar, 2002). Since the focus of the present research is on the transformation matrix structure, we are expected to fix this issue by modifying the transformation matrix. Since this problem is caused by the non-zero off-diagonal elements of the transformation matrix, it can be seen that by choosing α=π/4, these off-diagonal elements are exactly the negative value of their corresponding elements in the initial transformation matrix. Therefore, the summation of these two operators, one obtained for α=-π/4 and the other for α=π/4, reproduces a modified operator.

Figure 2 shows four consecutive rows of the new modified matrix in the same order as represented in Figure 1. Again, the largest peaks at the positions corresponding to the number of rows maintain the capability of detecting outliers; at the same time, there is no considerable quantity in the neighborhood, making this task uncertain.

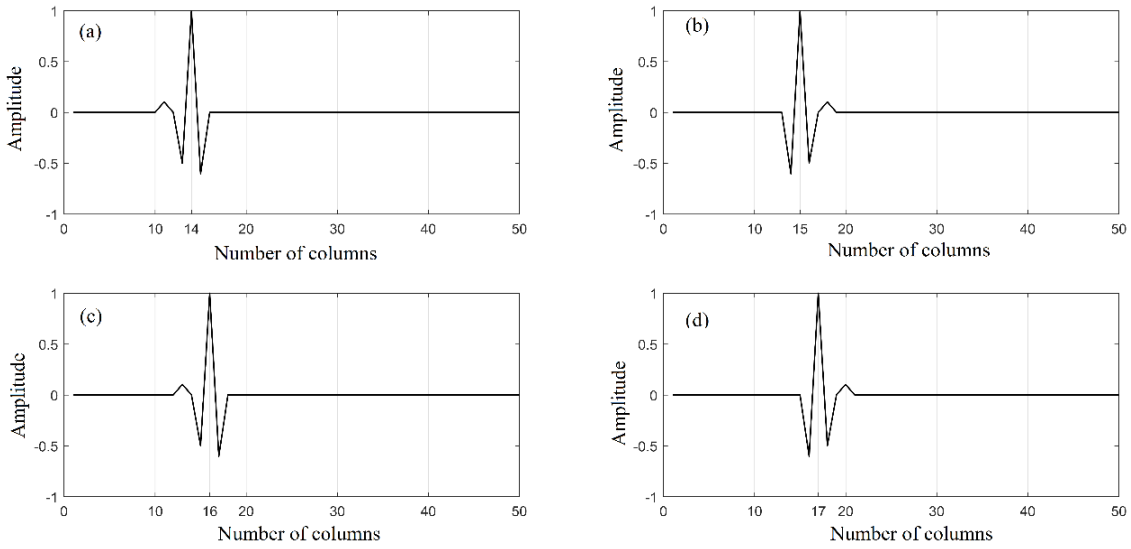

Figure 2. Four sample rows of the transformation matrix G2ˣG1 for the modified length-4 wavelets (a) the 14th row, (b) the 15th row, (c) the 16th row and (d) the 17th row

4. Numerical results and discussion

Illustrating our modified wavelet-based method, two numerical examples are introduced in this section, along with comparing the results of applying the conventional wavelet-based scheme for detecting the outliers as well as with the results of using a robust statistical approach known as the modified z-score method (Iglewicz & Hoaglin, 1993).

As the first example, in order to simulate real outliers in the example signal, two extreme values (outliers) have been inserted to a normally distributed 50 observations at the 32nd and 43rd elements, as follows

L_1= *[48.1717 36.4532 13.1691 31.2274 32.3586 ...*
16.1612 21.3550 37.9409 34.7713 40.9037 12.9577 ...
38.4884 37.4129 28.8269 54.6555 2.9460 17.5721 ...
32.1748 45.2027 44.0752 27.2009 12.2038 47.8418 ...
36.5650 25.0295 36.0776 32.0627 23.8036 28.4913 ...
14.1925 32.9215 110.6737 22.9883 13.4924 13.5561 ...
53.2436 30.0797 25.4254 35.2326 19.5343 18.3316 ...
23.8275 144.7577 57.2185 13.6728 27.3630 21.1187 ...
24.7077 27.8293 34.035]

The second observations are arc-second portion of 50 direction readings from 1" instrument as the vector of observations may contain some outliers.

L_2= *[41.9 49.5 42.6 45.5 46.3 45.5 47.2 43.4 44.6 43.3 47.4...*
45.5 46.1 42.6 44.7 43.1 42.5 44.3 44.2 46.1 45.9 46.1 46.3 ...
43.6 45.0 45.6 49.5 41.8 42.0 52.0 46.0 44.7 47.5 45.5 ...
44.3 46.2 43.2 43.4 42.8 43.2 43.0 42.2 47.1 46.8 45.7 ...
44.3 44.7 47.6 44.1 45.6]

At first, the modified z-score method is used to detect the probable outliers among the elements of the above vectors. In this method, after calculating the median and the median of the absolute deviation of the median (MAD), the modified Z-Score (Mi) is computed as (Iglewicz and Hoaglin, 1993)

$$M_i = \frac{0.6745\left(X_i - \tilde{X}\right)}{MAD} \qquad (10)$$

where, $\tilde{x}$ is the sample median, and according to Iglewicz & Hoaglin (1993), the observations are labeled as outliers when $|M_i| > 3.5$ with a significance level of 0.0005 and when $|M_i| > 2$ with a significance level of 0.05, which in the case of the first sample, the 32nd and 43rd elements of the observation vector are detected as real outliers with significance levels of 0.0005, as real outliers and 0.05.

In the case of the second sample, the 30th element of the observation vector and the 2nd, 27th and 30th ones are detected as outliers with significance levels of 0.0005 and 0.05, respectively.

Now, let us examine the wavelet-based method using Daubechies-4, our designed length-4 and finally, our modified operator. To decide on suspicious observations, we use the threshold limits for wavelet coefficients as proposed by Bilen & Huzurbazar (2002), with the difference that these limits are transformed to the time-space using inverse discrete wavelet transform to match the corresponding fine part of the time series in the time-space.

According to Bilen & Huzurbazar (2002), the threshold limit can be estimated based on the results that

$$P\left\{max\left|d_i / \sigma_i\right|_{i=1:n} > \sqrt{2log\ n}\right\} \to 0 \quad as \quad n \to \infty \qquad (11)$$

for wavelet coefficients (details) $d_i \sim N\left(0,\sigma_i^2\right)$. The normality of the wavelet coefficients is guaranteed by the normality of the data used that may be considered the limitation of the proposed method.

The noise level, σ, can be estimated using the mean value of absolute deviations of the wavelet coefficients from the median and then, the threshold limit, $\sigma\sqrt{2log\ n}$ can be computed based on the expressions (11). According to Bilen & Huzurbazar (2002), the false detection rates for the proposed wavelet-based method is comparable to other detection procedures.

Figures 3 and 4 represent the images of fine parts of the sample observation vectors, L1 and L2, along with their corresponding images of the threshold limits in the time domain, based on the three above-mentioned wavelet systems. In each case, some large peaks fall outside the limits as potential outliers. It appears difficult to distinguish actual outliers between some adjacent elements falling outside the corresponding threshold, except for the outliers detected by applying our modified transformation matrix (Figure 3.c and Figure 4.c), in accordance with the results of the modified z-score method.

5. Conclusions

In this study, the wavelet-based outlier detection method, as a robust non-parametric direct procedure for detection of outliers in time series data, has been improved based on the property of designing wavelet and scaling filters to achieve the following advantages:

- Non-ambiguity: Unlike other wavelet-based methods that use conventional base functions, our wavelet system was designed and modified to provide maximum compatibility with outlier detection task. This led to fixing the inherent ambiguity in choosing the outlier between two adjacent suspicious observations without any additional decision making process, which in turn resulted in fastness.
- Perceptibility: The proposed wavelet based procedure was formulated in the same space as the original signal exists that led to more comprehensible results. The image of the fine part of the observation vector was easily and quickly

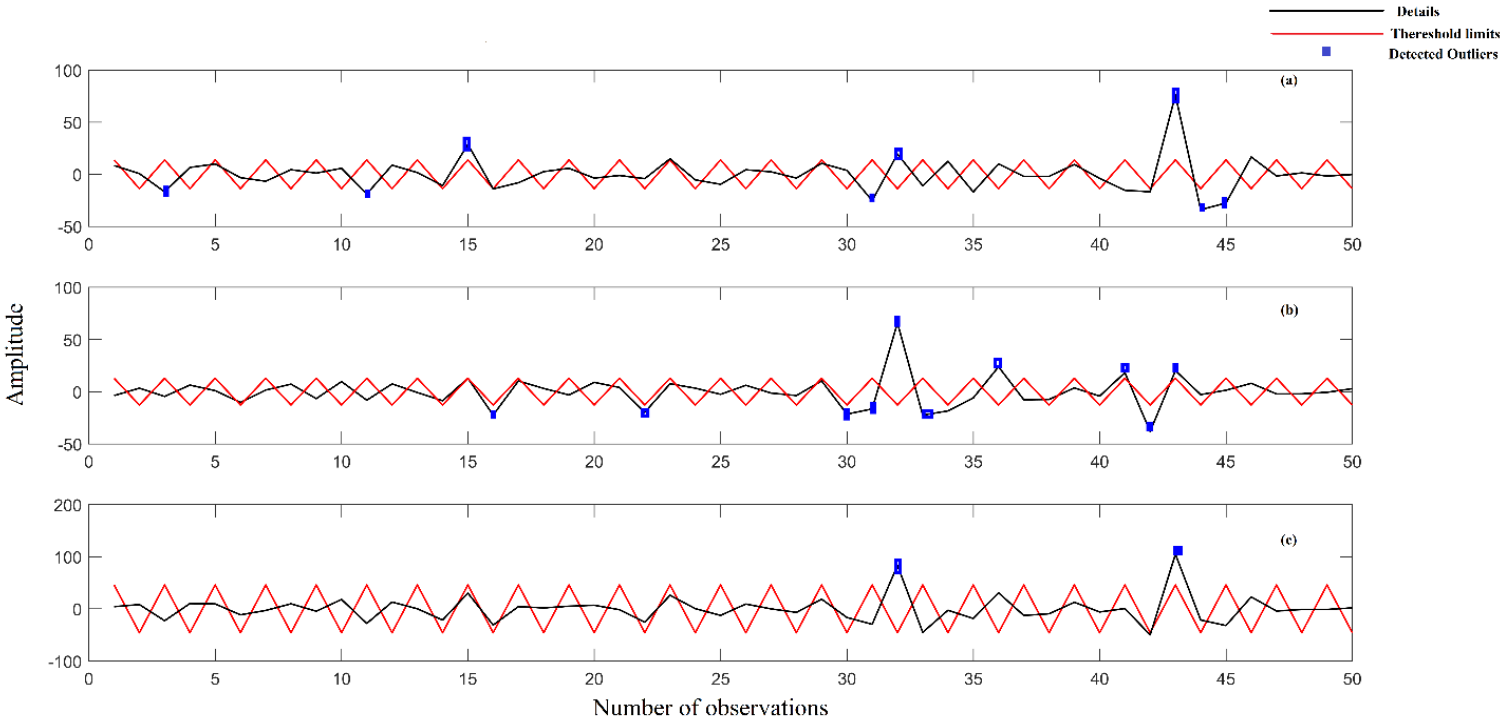

Figure 3. The images of the fine parts of the sample observation vector L_1, along with their corresponding images of threshold limits in the time domain based on (a) Daubechies-4 wavelets, (b) the designed length-4 wavelets with α=-π/4 and (c) the modified length-4 wavelets

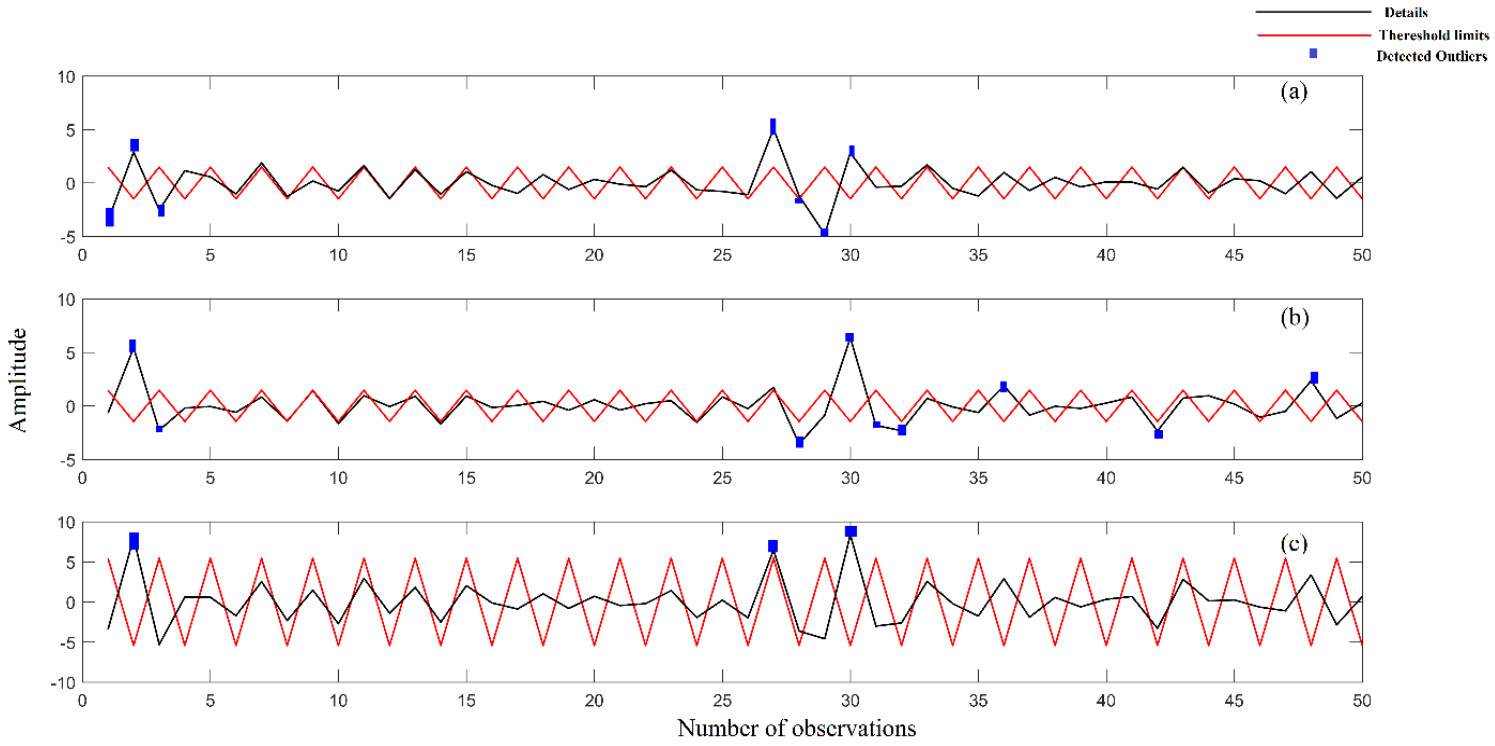

Figure 4. The images of the fine parts of the sample observation vector L_2, along with their corresponding images of the threshold limits in the time domain based on (a) Daubechies-4 wavelets, (b) the designed length-4 wavelets with $\alpha=-\pi/4$ and (c) the modified length-4 wavelets

extracted by applying the designed matrix of transformation to the time series. The probable outliers were magnified due to the characteristic of the designed system, and then compared to the estimated corresponding threshold limits that ultimately resulted to outlier detection task in a visible manner.

- Prioritization: Regarding the logic of wavelet-based outlier detection approaches that was improved in the present study, it is expected that among potential outliers falling outside the estimated threshold limits, the larger the corresponding peak appears in the fine part of the observation vector, the more likely it is an outlier.

The optimally designed transformation operator proposed in this research belongs to the length-4 wavelet systems with one remaining degree of freedom after satisfying the necessary conditions for the existence and orthogonality of the base functions. The remaining degree of freedom was used to achieve our key goal, which was providing the best performance in detecting outliers. However, in order to access some systems with other desired properties, there are more wavelet systems with additional degrees of freedom that can be explored or designed in future.

References

Barua, S., & Alhajj, R. (2007). 'Parallel wavelet transform for spatio-temporal outlier detection in large meteorological data' *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, pp. 684-694.

Beylkin, G., Coifman, R., & Rokhlin, V. (1991). Fast wavelet transforms and numerical algorithms I. *Communications on pure and applied mathematics*, 44(2), 141-183.

Bilen, C., & Huzurbazar, S. (2002). Wavelet-based detection of outliers in time series. *Journal of computational and graphical statistics*, 11(2), 311-327.

Burrus, C.S., Gopinath, R.A. & Guo, H. (1997). Introduction to wavelets and wavelet transforms: a primer.

Daubechies, I. (1992). Ten lectures on wavelets (Vol. 61).

Grané, A., & Veiga, H. (2010). Wavelet-based detection of outliers in financial time series. *Computational Statistics & Data Analysis*, 54(11), 2580-2593.

Iglewicz, B., & Hoaglin, D. C. (1993). How to detect and handle outliers (Vol. 16).

Keller, W. (2004). Wavelets in geodesy and geodynamics. Walter de Gruyter.

Li, Z. (1996). Multiresolution approximation of the gravity field. *Journal of geodesy*, 70(11), 731-739.

Malik, K., Sadawarti, H., & G S, K. (2014). Comparative analysis of outlier detection techniques. *International Journal of Computer Applications*, 97(8), 12-21.

Mallat, S. G. (1989). Multiresolution approximations and wavelet orthonormal bases of L^2 (R). *Transactions of the American mathematical society*, 315(1), 69-87.

Silva, I., & Silva, M. E. (2016). 'Wavelet-based detection of outliers in time series of counts' *Programa e Livro de Resumos-JOCLAD2016-XXIII Jornadas de Classificação e Análise de Dados*.

Ranta, R., Louis-Dorr, V., Heinrich, C., & Wolf, D. (2005). Iterative wavelet-based denoising methods and robust outlier detection. *IEEE Signal Processing Letters*, 12(8), 557-560.

Wichern, D. W., & Johnson, R. A. (1992). Applied multivariate statistical analysis (Vol. 4).