EXISTING OUTLIER VALUES IN FINANCIAL DATA VIA WAVELET TRANSFORM

Dr. Sadam Alwadi
Department of Risk Management and Insurance,
Faculty of Management and finance, University of Jordan/Aqaba, Jordan

Abstract

Outlier detection is one of the major problems of large datasets. Outliers have been detected using several methods such as the use of asymmetric winsorized mean. Al-Khazaleh et al. (2015) has proposed new methods of detecting the outlier values. This is achieved by combining the asymmetric winsorized mean with the famous spectral analysis function which is the Wavelet Transform (WT). Thus, this method is regarded as MTAWM. In this article, we will expand this work using the modern Wavelet function known as the Maximum Overlapping Wavelet Transform (MODWT). The results of the study shows that after comparing the new technique with the previous mentioned techniques using financial data from Amman Stock Exchange (ASE), the Maximum overlapping wavelet transform- asymmetric winsorized mean (MWAW) was considered the best method in outlier detections.

Keywords: Detecting outliers, asymmetric winsorized mean, maximum overlapping wavelet transform, financial time series data

Introduction

Outliers present helpful information about various processes. Hence, it can be generated by the location of the mean or variance. Usually, the value given to be outlier is the right tail of the skewed distribution. Researchers are not being able to ignore any observation without reasons. Therefore, the outlier values were prevented from been removed randomly. However, the outlier values should be solved (detected) from any dataset that has outliers' values. Also, the financial time series data does not have any outlier values. Nevertheless, in some cases, the outlier values can be found.

Furthermore, outlier detection is very important in many fields of study, since an outlier indicates the bad behavior of the dataset. Such field involves data analysis tasks where a huge number of observations are being sampled. One step towards getting a logical analysis is the detection of

outlier observations. Although some researchers considered outliers to be noise, because these researchers possess important and reliable information. Detected outliers are assigned as abnormal data that may otherwise adversely lead to model misspecification, biased parameter estimation, and wrong results. It is therefore significant to recognize them prior to modeling and analysis (Liu et al., 2004; Williams et al., 2002; Hodge and Austin, 2004).

Consequently, many methods of detecting outliers have been created. This method include: Z-score, box plot method, statistical measures, asymmetric winsorized mean, and Wavelet Transform Asymmetric Winsorized Mean (WTAWM) (Al-Khazaleh et al., 2015).

As has been critically reviewed, many researchers (e.g. Hodge and Austin, 2004) have carried out research in this area from the middle of the last century. Thus, their aim is to cover outlier detection in machine learning and statistical field. Also, symbolic data approaches were discussed (Motohiro, 2008), and statistical approximations in the outlier field were also discussed (Markou & Singh, 2003). Furthermore, for other examples about books and research articles in all fields of sciences, see Agyemang et al. (2006), Patcha and Park (2007), Rousseeuw and Leroy (1987), Barnett Lewis (1994), and Bakar et al. (2006). However, very rare contributions have been made on outlier detections based on wavelet transform.

More specifically, significant amount of work in the time series domain has been done (e.g. Fox, 1972). Thus, this represents the first work on outlier detection for time series data. Many models were proposed in the statistics literature such as autoregressive integrated moving average, vector auto regression, and exponentially weighted moving average (e.g. Barnett & Lewis, 1978; Hawkins, 1980; Rousseeuw and Leroy, 1987; and Gupta et al. (2014)).

After intensive research in the literature, we found very crawling movement of the Wavelet transform area in order to detect the outlier values. Struzik and Siebes (2002) have discussed a method of detecting outliers in time series data by checking the interior stability of the scaling spectrum of the process within the paradigm of the multifractal spectrum. Thus, they make use of the continuous wavelet transform in the detection processes. Zhao et al. (2003) detected outliers using wavelet transform in the content of Meteorological Data. In Hazan et al. (2012), the issue of detecting irregular vibrations from spectrum were been discussed. Moreover, network traffic which is based on discrete wavelet Transform has been studied (Salagean and Timisoara, 2010). Consequently, there is a cap in the wavelet transform literature that no research articles have been conducted in outlier detecting using maximum overlapping discrete wavelet transform. Therefore, our contribution in this article is on outlier detection focus on a particular research area. We presented a large general idea of the detailed research on

outlier detection techniques and its applications, and have emphasized the prosperity associated with solving real world problem. Therefore, this method will combine the traditional technique of Winsorized Mean with the new technique in the field of Maximum Overlapping Wavelet Transform (MODWT). This is aimed to get a new significant method in outlier detections known as MWAW. The following section will present the mathematical review for the previous methods used before the contributions was stated.

Mathematical and Literature Review of WT Models

Wavelet Transform

WT is a mathematical model employed to convert the original observations into a different domain (Chang and Moretin, 1998; Gencay et al., 2002; Daubechies I. (1992);Al Wadi et al., 2010). This model is very appropriate with the non-stationary data since most of the financial data are non-stationary. In addition, Al-Khazaleh et al. (2015) has used DWT in the detection of the outlier. However, we will make use of the expanded WT which is called MODWT. The DWT and MODWT are models of transformation which is usually used to transform from time-scale domains into smooth data set. Thus, these dataset can be used for many purposes. These procedures are known as signal decomposition since a specified signal is decomposed into several other signals with different resolution. These processes allow the recovering of the original time domain signal without losing any information. WT and MODWT have reverse process which is called the inverse WT or the signal reconstruction. Also, WT and MODWT are applied using a multiresolution pyramidal decomposition technique. Actually, a recorded digitized time signal $S(n)$ can be decomposed into its detailed $D1(n)$ and smoothed (approximations) $A1(n)$ signals using a filters. Consequently, the filtered signal $D1(n)$ is known by a detailed coefficient of $S(n)$, while $A1(n)$ is known as the approximation signal. $A1\ (n)$ and $D1(n)$ is the first scale decomposition in the WT processes. Usually, the researchers can reconstruct the original dataset: the approximations and details coefficients (Barnett and Lewis, 1994; Ming-Cai, 2005; Ababneh et al., 2013; Al-Khazaleh et al., 2015; Daubechies I., (1992)).

It is well known that WT has more benefits than Fourier Transform which motivate us to focus on its application. For instance, WT is highly redundant, and the analysis of variance can be simply applied for the transform coefficients based on WT. DWT can be defined naturally for specific samples sizes (i.e., “N” needs not to be a multiple of the power of two), while MODWT can be used as the sample size. Therefore, in this paper, we will focus on the most suitable function of DWT which is

MODWT. This is achieved using a closed price data in the content of ASE for detecting the outlier value.

Moreover, outlier data problem has been assigned in many scientific fields and researches. Thus, Ben-Gal I. (2005) defines the outlier values. Also, Karanjit Singh and Dr. Shuchita Upadhyaya (2012) have collected many reviews and results about outlets detections methods. Moreover, missing data problem has considered a special case of outlier. Therefore, Jiří Kaiser (2014) widely studied the missing data problem. Joseph Graham (2002) reviewed the methods and explained their strengths and limitations with the missing data issues in order to fill the gaps of the rapidly changing field.

Mathematical Review

Definition: WT in general can be defined by the following function (Gencay et al., 2002):

$$\psi_{j,k}(t) = 2^{\frac{j}{2}}\psi(2^j t - k),\ j,k \in Z;\ z = \{0,1,2,....\}. \quad (1)$$

Where ψ is a real valued function which is compactly supported, and $\int_{-\infty}^{\infty} \psi(t)dt = 0$. Generally, the wavelet transforms were evaluated using dilation equations, given as:

$$\phi(t) = \sqrt{2}\sum_k l_k \phi(2t - k), \quad (2)$$

$$\psi(t) = \sqrt{2}\sum_k h_k \phi(2t - k). \quad (3)$$

Father and mother wavelets were defined by the last two equations where $\phi(2t-k)$ represents the father wavelet, and $\psi(t)$ represents the mother wavelet. The father wavelet gives the high scale approximation components of the signal, while the mother wavelet shows the deviations from the approximation components. This is because the father wavelet generates the scaling coefficients, while the mother wavelet evaluates the differencing coefficients. Also, the father wavelet defines the lower pass filter coefficients (h_k). High pass filters coefficients (l_k) are defined as (Al-Khazaleh et al, 2015):

$$l_k = \sqrt{2}\int_{-\infty}^{\infty} \phi(t)\phi(2t-k)dt,$$

$$h_k = \sqrt{2}\int_{-\infty}^{\infty} \psi(t)\psi(2t-k)dt.$$

HWT is the simplest and oldest wavelet transform; and was improved by DWT in 1992 (Igelewicz & Hoaglin, 1993). The researchers developed

the frequency domain characteristics of the HWT. However, we do not have a specific formula for this method of wavelet transform. So for more details about the mentioned models (DWT and MODWT), please refer to the study of Salagean and Timisoara (2010) and Motohiro (2008).

Asymmetrical Winsorized mean

Winsorized mean is one of the measures of central tendency. Winsorized mean consists of the calculation of the mean after replacing the given parts of a probability distribution or sample at the beginning and low end with the most extreme remaining values. The objective of this method is to manage the variability due to the r lowest sample values $x_{(1)}, x_{(2)}, ..., x_{(r)}$ and the s highest ones $x_{(n-s+1)}, x_{(n-s+2)}, ..., x_{(n)}$ (Al-Khazaleh et al., 2015).

Moreover, the r lowest sample values are each restored by the value of the closest to observation to be retained unchanged. Similarly, the s is highest by $x_{(n-s)}$, such that we work with a transformed sample of size n. Therefore, we obtained the (r,s)-fold winsorized mean as:

$$\mathrm{W(r,s)}=\frac{1}{n}\Big[rX_{(r+1)} + \sum_{k=r+1}^{n-s} x_{(k)} + sX_{(n-s)} \Big] \quad (7)$$

When the amounts of lower-tail and upper-tail winsorizing are the same, i.e. $r=s$, we have the r-fold symmetrically winsorized means and Eq(7) becomes (Al-Khazaleh et al., 2015):

$$\mathrm{W(2r)}=\frac{1}{n}\Big[rX_{(r+1)} + \sum_{k=r+1}^{n-r} x_{(k)} + rX_{(n-r)} \Big] \quad (8)$$

Moreover, the sum of the square deviation can be computed using the following equation for the asymmetric winsorized mean:

$$s_{W(r,s)}{}^2 = \frac{1}{n}((r+1)(x_{r+1}-W_{(r,s)})^2 + \sum_{i=r+2}^{n-s-1}(x_i - W_{(r,s)})^2 + (s+1)(x_{n-s}-W_{(r,s)})^2) \quad (9)$$

Consequently, a robust estimate of the variance can be based on the Winsorized sum of squared deviations (Al-Khazaleh et al., 2015). Thus, the result of the Winsorized *t* test is given by:

$$T_{w(r,s)} = \frac{W_{(r,s)}}{STDERR(W_{(r,s)})} \quad (10)$$

where $STDERR(\bar{x}_{w(r,s)})$ is the standard error of $W_{(r,s)}$

$$STDERR_{w(r,s)} = \frac{(n-1)S_{w(r,s)}}{(n-r-s-1)(\sqrt{n(n-1)}} \quad (11)$$

The data from the symmetric distribution i.e. the distribution of the $t_{w(r,s)}$ is approximately from the student distribution with $n-r-s-1$ degree of freedom. The confidence interval $100(1-\frac{\alpha}{2})$ can be calculated for the location parameter as upper and lower limit by:

$$W(r,s) \pm t_{(1-\frac{\alpha}{2},n-r-s-1)} STDERR_{w(r,s)} \quad (12)$$

Dataset

In order to find the power of WT in detection, two functions have to be used i.e. DWT and MODWT. Also, in the content of daily closed price, data taken from Amman Stock Exchange (ASE) will be used for the period from 1993 to 2009 as a case study. The analysis of the behavior of the dataset using the mentioned models and the raw dataset can be presented as shown in Fig. 1 (Al-Khazaleh et al, 2015).

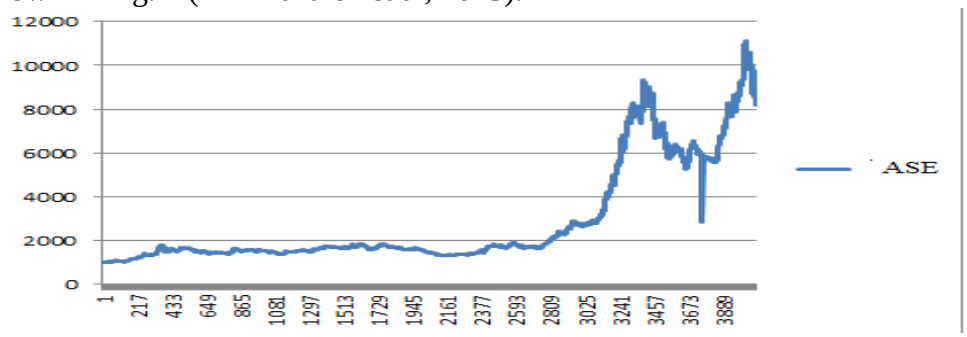

Figure 1: ASE closed price 1993-2009

Methodology

The significant contribution in this paper can be summarized as follows,

1- After the entire data split into three groups, we transformed every group using DWT and MODWT.

2- Obtained number of observation for groups (smoothed) dataset using WT equations.

3- The smooth group of the data will be used for detections. Then, the winsorized mean can be calculated for the data for the three groups before and after transforms using DWT and MODWT. Therefore, the processes and conditions of winsorized mean calculations can be found in Al-Khazaleh et al. (2015). The number of observations inside the first group and the number of observations in the last one was counted. Subsequently, we focused on the number of observations in the first subgroups which is equal to 140 observations. Also, we suggested an assumption of this number as the value of (r). Similarly, we counted the number of observations inside the last

subgroups which is equal to 160 observations; and we assumed this number to be the value of (s). In this way, the asymmetric winsorized mean can be calculated using Eq (7). Table 1 illustrated the statistics for asymmetric winsorized mean for the three groups ASE closed price data for WTAWM (Al-Khazaleh et al., 2015).

Table 1: The statistics for ASE closed price data for WTAWM

Number of groups	Winsorized mean	Standard Error of $W_{(r,s)}$ using WTAWN	Standard error for $W_{(r,s)}$ using MWAW
1	2504	0.507	0.436
2	2578	0.692	0.633
3	2538	0.6077	0.586

In order to explain the proposed method in outliers detections for the three groups for the closed price ASE time series data, since the winsorized mean equal 2504, the winsorized sum of square deviation was evaluated using eq. 11 which is equal to 673. After that, the standard error was calculated using Eq. 12 which is equal to 0.507. In addition, the Winsorized t test is equal 127.9 which is the predicted value of the possible outliers. The t value is the critical value and it can be calculated using mathematical equations (Al-Khazaleh et al., 2015).

In Al-Khazaleh et al. (2015), after the author's evaluation of the t values for WTAWN were computed for the whole data, there are 36 observations of the amount of t value which is greater than the critical value 40.1004. In this case, all the values that are larger are supposed to be outliers; and the same results is for the second and third groups. Furthermore, in this suggested method (MWAW), the results are more significant since the standard error is smaller than WTAWN for all of the groups. Therefore, we concluded that the number of the predicted outliers is the same in the three cases using the three methods which are winsorized mean, WTAWN, and MWAW. Nevertheless, the error is the biggest in winsorized which will be medium in WTAWN, while the error will be very small in MWAW as presented in Table 1. More specifically, Table 2 shows the amount of the t values and the predicted outliers for the three groups from ASE data using MWAW.

Table 2: The amount of predicted outliers and t values

Predicted Outliers	Statistics test (T) for			Predicted Outliers	Statistics test (T)		
	1st group	2nd group	3rd group		1st group	2nd group	3rd group
1006.48	40.11	41	42. 051	9219.8801	54	56	56.412
1009.05	40.3	42.1	42.07	9203.5000	56	58	58
1007.57	41.9	43.9	43.040	9353.4779	56	59.1	57.7
1006.92	43.3	44	43.730	9406.8220	57	59.6	59.2
1030.78	43.2	44.6912	44.280	9553.2895	58.4	60.4317	60.4937
1026.42	44.4	45.1418	44.735	9727.9593	59.7	62.4262	61.9900
1029.64	45.6	46	45.613	9737.1382	60	63.8	63.3
1034.07	45	47	46.1	9572.9950	61	65.2	64.2602
1033.2	45.34	47.23	46.279	9608.5530	63	67.2	67.7
1034.04	45.9	47.9	47	9378.1899	64.56	68	68
1028.91	46.1	48	48	9226.4253	65	69.3	68.16
1032	48	49.1	49	9123.5497	66	69.34	69.75
1032.11	49	51	50	8914.4484	66.4	71.52	70.45
9301.963482	49.3	52.	51	8793.7097	69	73.17	74
9177.524779	50.6	53	52	8625.0636	75	90.6	91.7
8842.534749	51	53.2	52	8438.7347	80	101	101
8716.331762	51.12	54	53	8213.237	88	102	101.8
9060.545232	53	54.19	55	8522.207	165.45	191.4	190.3

Conclusion and Recommendations

In this article, we have modified the method used (Al-Khazaleh et al., 2015) by using other WT function which is MODWT in outlier detections. Although Al-Khazaleh et al. (2014) found some outlier values, this method could not give significant solutions and there are some conditions over the dataset. Therefore, we suggested the new technique (MWAW) by dividing the whole data into subgroups using MODWT (this function does not need any conditions over the data set). In addition, the two parameters for the asymmetric winsorized mean were estimated, and the t values were used to determine the critical value. Moreover, the values is greater than what the t value are supposed to be as outliers. In this method, we have 36 values predicted outliers with a smaller standard error than WTAWN. Therefore, the suggested method is better than the traditional methods (WTAWN and winsorized mean directly). As a recommendation, the regression analysis can be used in the calculation of the coefficients of determination to test the best model for the average closed price data after removing outliers.

References:

Chang, C., & Moretin, P., (1998). A Wavelet Analysis for Time Series, Nonparametric Statistics, 10, 1-46.

Barnett and Lewis, (1994). Outliers in statistical data, John Wiley & Sons.
Gencay, R., Seluk, F. & Whitcher B., (2002). An Introduction to Wavelets and Other Filtering Methods in Finance and Economics, Academic Press, New York.
Igelewicz, B., & Hoaglin C., (1993). How to detect and handle outliers. Milwaukee, WI: ASQC Quality press.
Salagean, M., & Timisoara, I., (2010). Romania Anomaly Detection of Network Traffic Based on Analytical Discrete Wavelet Transform. IEEE transaction.
Motohiro Y., (2008). Measure Business Cycle: A Wavelet Analysis of Economic Time Series. Economics Letter, 100, 208-212.
Al Wadi, S., Ismail, M., & Abdul Karim, S. A., (2010). A Comparison between The Daubechies Wavelet Transformation and the Fast Fourier Transformation in Analyzing Insurance Time Series Data. Far East Journal of Applied Mathematics. 45, 53-63.
Daubechies, I., (1992). Ten Lectures on Wavelets, SIAM and Philadelphia.
Ming-Cai (2005). Wavelet Analysis and its Application, Tsinghua University Press, Beijing,
Ababneh, F., Al Wadi, S., & Ismail, M., (2013). Haar and Daubechies Wavelet Methods in Modeling Banking Sector. International Mathematical Forum, 8, 551 – 566.
Hazan, A., Verleysen, M., Cottrell, M., Lacaille J., & Madani, K. (2012). Probabilistic Outlier Detection in Vibration Spectra with Small Learning Dataset.
Liu, H., Shah, S., & Jiang W. (2004). On-line outlier detection and data cleaning, Computer and Chemical Engineering, 28, 1635–1647.
Williams, J., Baxter, A., Hawkins, S., & Gu, L. (2002). A Comparative Study of RNN for Outlier Detection in Data Mining. IEEE International Conference on Data-mining, Japan.
Al-Khazaleh, A., alwadi, S., & Ababneh, F. (2015). Wavelet Transform Asymmetric Winsorized Mean In Detecting Outlier Value. Far East Journal of Mathematical Sciences.
Hodge J., & Austin J. (2004). A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22, 85-126.
Fox, A., (1972). Outliers in Time Series, Journal of the Royal Statistical Society. Series B (Methodological), 34, 350–363.
Cho, H., jin Kim, Y., Jung, H., Lee, S.-W., & Lee, J. (2008). Outlier: An R Package For Outlier Detection Using Quantile Regression On Mass Spectrometry Data, Bioinformatics, 24, 882–884.
Barnett, V.& Lewis, T., (1978). Outliers in Statistical Data, John Wiley & Sons.
Hawkins, D., (1980). Identification of Outliers. Chapman and Hall.

Rousseeuw. P & Leroy A., (1987). Robust Regression and Outlier Detection. John Wiley & Sons.
Gupta, M., Gao, J., Aggarwal, C., Han, J., (2014). Outlier Detection for Temporal Data: A Survey. Ieee Transactions on Knowledge and Data Engineering, 25.
Struzik, R., & Siebes, A. (2002). Wavelet Transform Based Multifractal Formalism in Outlier Detection and Localization for Financial Time Series. Physica A, 309, 388 - 402.
Zhao, J., Lu, C.-T., & Kou, Y., (2003). Detecting Region Outliers in Meteorological Data CIS'03, New Orleans, Louisiana, USA.
Markou M. & Singh, S. (2003). Novelty Detection: A review-part 1: Statistical Approaches. Signal Processing 83, 2481-2497.
Agyemang, M., Barker, K., & Alhajj, R. (2006). A Comprehensive Survey of Numeric and Symbolic Outlier Mining Techniques. Intelligent Data Analysis, 10, 521- 538.
Patcha, A & Park, M., (2007). An Overview of Outlier Detection Techniques: Existing Solutions and Latest Technological Trends. Comput. Networks 51, 3448-3470.
Rousseeuw, J & Leroy, M., (1987). Robust Regression and Outlier Detection. John Wiley & Sons.
Barnett, V & Lewis, T., (1994). Outliers in Statistical Data. John Wiley and sons.
Bakar, Z., Mohemad, R., Ahmad, A. & Deris, M. (2006). A Comparative Study for Outlier Detection Techniques in Data Mining. Cybernetics and Intelligent Systems, 2006 IEEE Conference, 1- 6.
Ben-Gal I. (2005). Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers.
Karanjit Singh & Dr. Shuchita Upadhyaya, (2012). Outlier Detection: Applications and Techniques. International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3.
Jiří Kaiser, (2014). Dealing with Missing Values in Data. Journal of Systems Integration.
Joseph L. Schafer and John W. Graham, (2002). Missing Data: Our View of the State of the Art. Psychological Methods . Vol. 7, No. 2, 147–177.