Aberrant Observation Detection in Frequency Domain: The Wavelet Approach

Aideyan Donald Osaro[1], Kogi state University, Department of Mathematical Sciences, Anyigba, Kogi State, Nigeria.

Shittu Olarewanju Ismail[2] University of Ibada, Department of Statistics, Ibadan, Oyo State, Nigeria

ABSTRACT: Spectral method has been used in detection of aberrant observations in the frequency domain with promising results. The method is useful only when the data is stationary, without jumps or discontinuities, otherwise decisions on suspected aberrant observations might be incorrect. This study proposed the use of wavelet shrinkage algorithm capable of ameliorating the limitations of spectral method in the non-parametric setting. It will be extended to the parametric setting and the result compared with previous studies in outlier detection. The Mallat algorithm was used to reduce the size of the data into smaller resolutions while preserving the desired statistics. In the non-parametric setting, Turkey's and modified Turkey's method initially developed for thresholding was adapted for the wavelet shrinkage approach by examining the presence of aberrant observations at different resolutions.

Index Term: Wavelet, nonparametric, Parametric, Data analysis, Statistics, Findings, Interpretation

1 INTRODUCTION

Wavelet analysis as a Statistical tool that can be used to extract information from any kind of data and are generally needed to analyze data fully at different resolution (scale) and location (time) whether the data is stationary or not. It is also used to analyze localized variation and allows us to partition (decompose) the information in a time series into pieces that are associated with different resolution (scale) and location (time). Wavelet transform is known as non-parametric orthogonal series estimators which are capable of providing the necessary time and frequency information on time series data simultaneously in a highly flexible fashion. In statistics, aberrant observation is an observation that is numerically distinct from the rest of the data. They occur by chance in any distribution but are often indicative either of measurement error or that the population is heavy tailed. Section 2 will dwell on the key features of wavelet analysis and the aim and objectives of this paper. Section 3 describes the methods used in detecting these aberrant observations in both the parametric and non-parametric settings which are the main goal of this paper. Section 4 shows the tables of the analysis and detection of these aberrant observations while Section 5 includes summary of findings, interpretation of results, conclusion, recommendations and suggested area for further work.

Section 2

2.1 The Key Features of Wavelet Methods

- sparsity of representation for a wide range of data as resolution decreases including those with discontinuities guaranties the presence of required statistics.
- The ability to analyze data at a number of resolutions and also to work with information at such resolutions.
- Ability to detect aberrant observations and represent neighbourhood features and also to create localized features on synthesis (the process of combining differences into a new whole)

- Efficiency in terms of compilation speed and storage.

2.2 AIM AND OBJECTVES

The major goal of this paper is to compare the efficiency of wavelet coefficients as a tool for detecting aberrant observations in the non-parametric and parametric approaches. Below are the underlying objectives.

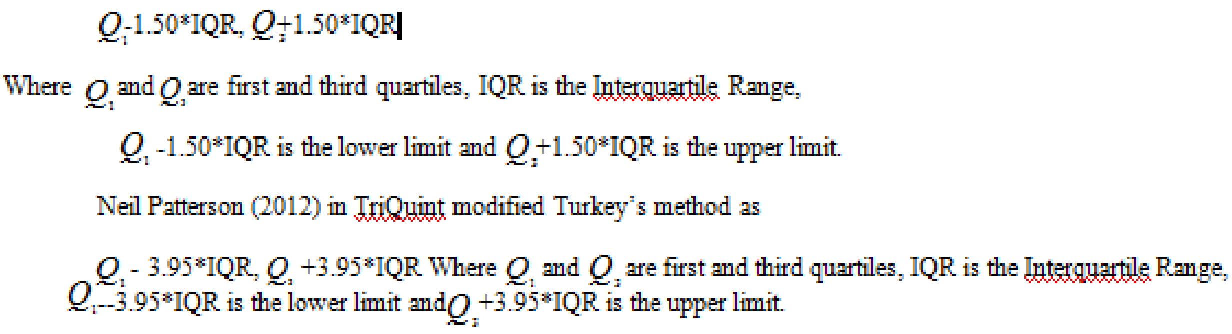
- To derive wavelet coefficients for detection of aberrant observations using Turkey's and Modified Turkey's methods.
- To develop test statistic in the parametric setting that can be used to achieve bullet 1.
- To compare what is obtained in bullet 1 with what was obtained using the parametric approach

Section 3

3.1 WAVELET REVIEW

3.1.1 (NON - PARAMETRIC APPROACH)

- Wavelet analysis appears best suited to exploratory data analysis of complex, non-stationary data which summarizes their main characteristics in easy-to-understand form of visual graphs without using a statistical model or having formulated a hypothesis proposed by John Turkey to encourage statisticians visually to examine their data set e.g. Box-plot. (Bruce et.al. 1996).
- wavelets can be viewed as non-parametric orthogonal series estimators with new elegant statistical results and efficient computational algorithms, (Fan and Gij-bels, 1996, pp 26-39) that can effectively handle the discontinuities caused by different regime shift (characteristic conditions) that typically plague the economic and financial data. Donald B. Percival et al (2000).
- Their (non-stationary) are especially suitable to the comprehensive multi-decision analysis of disaggregated (scaled) series; the process of data aggregation and concept of equispaced series do not play any fundamental role in the context of wavelet analysis Abraham Maslow (1998)
- In the non-parametric setting, Turkey's method was used in the detection of aberrant observations in the two series. Turkey's method is defined as:



Q_1-1.50*IQR, Q_3+1.50*IQR

Where Q_1 and Q_3 are first and third quartiles, IQR is the Interquartile Range,

Q_1 -1.50*IQR is the lower limit and Q_3+1.50*IQR is the upper limit.

Neil Patterson (2012) in TriQuint modified Turkey's method as

Q_1 - 3.95*IQR, Q_3 +3.95*IQR Where Q_1 and Q_3 are first and third quartiles, IQR is the Interquartile Range, Q_1--3.95*IQR is the lower limit and Q_3 +3.95*IQR is the upper limit.

3.1.2 The parametric approach

In the parametric setting, a test statistic was developed to analyse these data. The test statistic is

$$U = \frac{y_j - \bar{y}_j}{\sigma_j \sqrt{2 \log n}_j}$$

Where:

y_j Are the observations in the resolution being analysed

$\bar{y}_j$ Is the mean at the resolution

σ_j Is the standard deviation at the resolution

n Is the number of observation at the resolution

Section 4: Data Analysis

Table I: Wavelet analysis of UCH Diabetic Data using Turkey's method

Resolutions level (No 0f observations)	Location (L) of aberrant observations	Aberrant observation values	T_L	T_U
7 (128)	47	58	0.50	36.50
6 (64)	24	30.41	-17.68	17.68
5 (32)	12	-21.00	-17.75	17.25
4 (16)	6	11.67	-21.55	8.22

Table II: Wavelet analysis of UCH Diabetic Data using modified Turkey's method by Neil Patterson (2012)

Resolutions level (No 0f observations)	Location (L) of aberrant observations	Aberrant observation values	T_L	T_U
7 (128)	-	-	-21.55	58.55
6 (64)	-	-	-39.33	39.33

5 (32)	-	-	-39.19	38.69
4 (16)	-	-	-29.41	22.51

Table III: Wavelet analysis of Zadakat Data using Turkey's method

Resolutions level **(No of observations)**	**Location (L) of aberrant observations**	**Aberrant observation values**	T_L	T_U
7 (128)	23,110,120	111.5,135,130	0.06	107.44
6 (64)	46,54,59	48.79,45.96,	-44.85	44.59
5 (32)	-	-	-54.31	47.69
4 (16)	-	-	-43.20	20.62

Table IV: Wavelet analysis of UCH Diabetic Data (parametric analysis)

For N=128, =5%, 7% and 10% where U= 1.96, 1.50 and 1.28 respectively.

Resolution Level (no of observations)	Location (L) of aberrant observations	Aberrant observation values	Location (L) of U value s	U Values
7 (128)	47	58	47	2.4491
6 (64)	24	30.4056	24	2.2379
5 (32)	12	-21.0000	12	-1.7845
4 (16)	6	11.6673	6	1.3781

Table V: Wavelet analysis of Zadakat Data (parametric analysis)

For N=128, = 5%. 7% and 10% where U= 1.96, 1.50 and 1.28 respectively.

Resolution Level (no of observations)	Location (L) of aberrant observations	Aberrant observation values	Location (L) of U value s	U Values
7 (128)	110,120	135,130	110,120	1.7920,1.6693
6 (64)	46,54,59	48.7903,45.9619, -42.0729	46,54,59	1.3946,1.3130, -1.2227
5 (32)	27,30	-34.75,36.25	27,30	1.0902, 1.3570
4 (16)	5,9	15.9806,12.7280	5,9	1.2360,1.0763

Section 5

5.1 Summary of findings

This work is focused on detection of aberrant observations using the mallat algorithm of Wavelet analysis in the frequency domain. The methods were investigated in the non-parametric and parametric settings.

In non-parametric setting, Turkey's and modified Turkey's methods were used in analyzing two real life (University College Hospital Diabetic. Ibadan and Zadakat both in Oyo State) data of 128 observations. T_L and T_U represents the lower and upper limits for the turkeys and modified turkey's methods respectively.

From table I and III, it was observed that the aberrant observations were detected at the same location and at different resolution level only in the UCH diabetic data up to the fourth level while it could only achieve two levels in the Zadakat data. The modified Turkey's method (table II) could not identify any in both data sets.

In the parametric setting, developed test statistic with alpha values of 5%, 7% and 10% respectively .From table IV and V it was observed that these aberrant observations were detected at same location even when the data has been compressed to the fourth resolution level in the two data. This is in agreement with the result obtained by Shittu (2006). It was also observed that the more the data set, the more efficient wavelet method is in detecting these aberrant observations.

5.2 Conclusion

Wavelet technique has been established to be useful in detecting aberrant observations in the same location even when the data is compressed in the non parametric and parametric setting. The proposed test statistic in the parametric approach was found to be more efficient than the existing Turkey's and modified Turkey's methods in the detection of aberrant observations in real life data when the data is to be compressed

5.3 Recommendations

- we recommend that where we have large data and has to be compressed, wavelet analysis should be used and average of fourth resolution is recommended
- It is good when we have high frequency data.

5.4 Suggested Areas for Further Research

This research work focused on aberrant observations detection in the frequency domain using wavelet analysis. After this work, some research areas for feature academic work has now been opened. They are:

- Extend the detection of aberrant observations in the frequency domain using the Non – decimated wavelet transform.
- Comparing our results in Haar wavelet analysis objectively with what may be obtained using the Non – decimated wavelet transform.
- Using multiple wavelet transform for the detection and modeling of aberrant observations in time series data.

References

[1] Abraham Maslow, Histogram Smoothing Via The Wavelet Transform. Journal of Computational and Graphic Statistics, Vol.7, No.4 (Dec., 1998)

[2] A. Dainotti, A. Pescape and G. Viorgio "Wavelet-based Detection of Dos Attack" Proceedings of IEEE Global Telecommunication Conference, San Francisco, 2006

[3]Bruce et al "Country Specific Market Impact of Climate Change" The anals of statistics, 1996

[4] Daniel T.L. Lee and Akio Yamamoto "Wavelet Analysis : Theory and Applications"Hewlett-Packard Journal,1994

[5]Donald B. Percival, Andrew T. Walden, "Wavelets Methods for Time Series Analysis." Cambridge University Press (2000)

[6] Fan and Gij-bels "A Study of Variable bandwidth Selection for Local Polynomial regression" Statistica Sinica 6(1996), pp. 113-127

[7] G.P. Nason, Wavelet Methods In Statistics With R Springer (2008)

[8] Sandrine Anthoine et al "Using neighborhood distributions of wavelet coefficients for on- the- fly, multiscale-based image retrieval" Proceedings of WIAMIS'08, 07-09/05/2008, Klagenfurt, Austria, pp. 38-41

[9] W. Lu and I. Traore "A Novel Unsupervised Anomaly Detection Framework for Detecting Network Attacks in Real Time", Lecture Note in Computer Science, Vol., 3810, pp. 96-109, Springer, 2005, Y.G. Desmedt et al (Eds.)

[10] Wei Lu, Mahbod Tavallaee andAli A. "Ghorbani Detecting Network Anomalous Using Wavelet Basis Functions" CNSR, pg 149-156, IEEE Computer Science (2008)

First Arthur

Aideyan Donald Osaro*

P.D.S.,B.Sc, M.Sc.

Second Arthur

Shittu Olarewanju Ismail**

P.D.S.,B.Sc, M.Sc, M.Phil,Ph.D