An improved framework for periodic outlier patern detection in time series sequences

Dhwani Dave[#1], Tanvi Varma[#2]

#1 Parul Institute of Technology, LIMDA, WAGHODIA, GUJARAT, INDIA.
#2 Vadodara, 8410298007

ABSTRACT

Patterns which repeat over a period of time are known as periodic patterns. This Periodic pattern detection is an important data mining task, which discovers in the time series all patterns that exhibit temporal regularities. Periodic pattern mining has a large number of applications in real life as well; it helps understanding the regular trend of the data along time, and enables the forecast and prediction of future events. An interesting related problem that has not received enough attention is to find out outlier periodic patterns in a time series .Outlier patterns are defined as those which are different from the rest of the patterns; outliers are different from noise. While noise does not belong to the data and it is mostly eliminated by pre-processing, while outliers are actual instances in the data but have exceptional characteristics compared with the majority of the other instances. Outliers are unusual patterns that rarely occur, and, thus, have lesser support (frequency of appearance) in the data. Outlier patterns may hint toward discrepancy in the data such as fraudulent transactions, network intrusion, change in customer behavior, epidemic and disease biomarkers, severe weather conditions, recession in the economy, etc. so here the argument is that detecting the periodicity of outlier patterns might be more important in many sequences than the periodicity of regular, more frequent patterns. In this paper, a suffix tree-based algorithm capable of detecting the periodicity of outlier patterns in a time series using MAD (Median Absolute Deviation) is presented. An existing algorithm makes use of mean values, which is inefficient. Use of MAD increases the output of these algorithms and gives more accurate information. Several experiments have been conducted using real data; all aspects of the proposed approach are compared with the existing algorithm which makes use of the mean value; the reported results demonstrate the effectiveness and applicability of the proposed approach.

Key words: Periodic patterns, outliers, Mean, MAD, confidence, periodicity

Corresponding Author: Dhwani Dave[#1]

INTRODUCTION

A TIME series records features captured regularly at uniform interval of time. Real life has several examples of time series such as weather record, stock price movement, road or network traffic density pattern, sensory data, transactions record, etc. Time-series database records such time-ordered features, and data mining aims to discover in the database hidden information. sometimes the information is in the form of patterns, which may lead toward some sort of rules. For Periodicity detection in time-series databases is a data mining problem where periodically repeating patterns are discovered. Periodic patterns are found in weather data, transactions history, stock price movement, road and computer network traffic density, gene expression, etc. [1], [2]. In time-series analysis, periodicity detection identifies the periodic functions to capture seasonality. Periodic pattern mining is an important task, as it facilitates data analysis leading to prediction or forecast of future events and patterns [3]. It also helps in identifying abnormal activities (or anomalies) in the data, i.e., events which occur at unexpected time. For example, a large number of power consumption patterns of a housing community are repeated weekly. An interesting extension to this problem is discovering the periodicity of outlier or surprising patterns. For example, a number of credit card transactions completed over time and involving small amounts may hint toward fraudulent activity. Further, by considering the recent economy crises, it should be possible to realize that similar economic situations have been observed in the past shall the historical data be available and well analyzed. It would be interesting to mine out whether the turn down in economy has a periodic pattern. Discovering how periodically repeating other features (e.g., oil price) are aligned with the turn down in economy might help in identifying the relationship between different features.

II. RELATED WORK

There are several algorithms that discover the frequent periodic patterns having (user specified) minimum number of repetitions or with minimum confidence (ratio between number of occurrences found and maximum possible occurrences), e.g., [4]–[5], [6], and [7]. However, not much work has been done for periodicity detection of outlier patterns. It is important to note that surprising or outlier patterns are different from outlier (values) in the data [8]. There are many techniques to find outliers in the data, but outlier or surprising patterns are different from others patterns. There are few algorithms, e.g., [9] and [10], which discover the surprising patterns in time series. Keogh *et al.* [11] presented their suffix tree-based algorithm to mine surprising patterns. Their algorithm requires the user to supply a “regular” series, which is used for training purpose. Patterns in the test data are compared with the training data and those having different expected values are qualified as “surprising” patterns. Since the algorithm requires the training data, it might not be possible in many cases to define the “regular” data; secondly, the algorithm only discovers surprising patterns which are not necessarily the periodic patterns. Yang *et al.* have presented their so-called InfoMiner algorithm [12]. They define the measure of “surprise” using their notion of information gain which gives more significance to patterns involving lesser frequent events and having more support (matching repetition). The Faraz Rasheed and Reda Alhajj’s suffix tree based algorithm considers the relative frequency of a pattern and the area of its coverage to measure the surprise of a pattern, which is expected to capture outlier or unusual patterns more effectively. In addition, their algorithm, being based on STNR, is flexible to work with noisy data where periodic repetitions are not strict; it can also detect periodic patterns in a section of time series. Another way to classify the existing algorithms is based on the type of periodicity they detect; some detect only symbol periodicity [4], and some detect only sequence or partial periodicity [6]. Their single algorithm can detect all the three types of periodicity. Earlier algorithms, e.g., [5], [7]–[9] require the user to provide the expected period value, and then check the time series for the patterns that are periodic with that period value. For example, in power consumption time series, a user may test for weekly, biweekly, or monthly periods. However, it is usually difficult to provide expected period value; and this approach prohibits finding unexpected periods, which might be more useful than the expected period. Sheng *et al.* [13], [14] presented their algorithm which is based on Han's [15] partial periodic patterns algorithm, which can detect periodic patterns in a section of time series, and utilizes the optimization steps to find the dense periodic areas in the time series. However, their algorithm, being based on Han's algorithm, requires the user to provide the maximum period value. They argue that the maximum period value is difficult to be defined by the user and which may lead toward missing some interesting periodic patterns. The time performance of Chang's algorithm deteriorates when the maximum period value is changed. The maximum period value is more difficult to provide while discovering outlier patterns as these type of patterns may have large period value, for example, the economic turn down pattern might have the periodicity of five to ten years. Recently, Huang and Chang [16] presented their algorithm for finding asynchronous periodic patterns, where the periodic occurrences can be shifted in an allowable range within the time axis.

III. PERIODIC OUTLIER PATTERNS—BACKGROUND[17]

Here the basic terminology of periodic patterns is explained.

A. Periodic Pattern[17]

A pattern X is periodic starting from index position i_{st} and ending at position i_{end} with period p in string s if

$$s[i_{st} + j * p, /X/] = X; j = 0, 1, 2, . . . ,i_{end} + 1 - /X/ - i_{st} \qquad (1)$$

p and $s[y, z]$ is a substring of s of length z starting at y.

If f is the actual number of repetitions (frequency) in the given range, then the confidence of the period is calculated as

$$\mathrm{conf}(X, i_{st}, i_{end}, p) = \mathrm{f}/\mathrm{f}_{max} \qquad (2)$$

For example, consider the following string s:

```
s = abce abcd abcd abcd abcd abbb abcb
    0123 4567 8901 2345 6789 0123 4567.
```

Pattern $X = abc$ is repeating in s with period: $p = 4$, starting at position $i_{st} = 0$, ending at position $i_{end} = 26$, with pattern length: $/X/ = 3$. The maximum number of repetitions $f_{max} = i_{end}$ d $+1 - /X/ -$is t $p + 1 = 7$; $X = abc$ is repeating six times $f = 6$; and the confidence of this period X is conf(abc, 0, 26, 4) = 6/7 .Similarly, we have another period $X\_ = abcd$, $/X\_/ = 4$, $i_{st} = 4$, $i_{end} = 19$, $p = 4$ with conf($abcd$, 4, 19, 4) = 4/4= 1.

B. Outlier Periodic Patterns[17]

A less frequent pattern with larger coverage area (having repetitions in larger subsection of the sequence) is more interesting than those with smaller coverage area (repeating in smaller subsection of the sequence). Hence, the pattern $X = ab$ with period $p = 7$ is a better candidate for outlier pattern in the sequence

S = acbacba **ab**acbac **ab**acbac **ab**acbacbacb
0123456 7890123 4567890 12345678901

than the pattern $X = ab$ with period $p = 2$ in the sequence

S_ = acbacba **ababab** acbacbacbacbacba
0123456 789012 3456789012345678 90

If $f(X)$ represents the frequency (repetition count) of the pattern X, and $segLen(X)$ represents the segment length of the repetitions of X, then X is *candidate outlier pattern* if

$$f(X) < \mu(f(Xi)) \text{ AND } segLen(X) > minSegLen;\ \forall i \qquad (3)$$

$$\text{such that } /Xi/ = /X/$$

where $\mu(f(Xi))$ is the mean of the frequency of all patterns of length exactly the same as that of pattern X.
The measure of *surprise* of a pattern X is defined as one minus the ratio of the frequency of X over the average frequency of all patterns with same length as X

$$\text{surprise}(X) = 1 - f(X)/\mu(f(Xi));\ \forall i \text{ such that } /Xi/ = /X/ \qquad (4)$$

A candidate outlier pattern X is an *outlier periodic pattern* iff

$$\text{surprise}(X) > \text{surprise}_{min} \text{ AND conf}(X, i_{st}, i_{end}, p) > \text{conf}_{min} \qquad (5)$$

IV. USE OF MEDIAN ABSOLUTE DEVIATION [18]

In existing algorithm mean value is used to find out outliers. But for various reasons this method is not efficient, this reasons are explained in the paper

1. Mean

Mean of any given data set is derived as follows:

$$\mu = \sum_{i=1}^{n} y_i / n \qquad (6)$$

It is the average value of any given data set. The reasons why it is considered a nonrobust estimator are as follows:

1) Mean value is highly biased even if there is a single outlier and 2) in a large data sets a mean value can be changed even though an outlier is removed . So, while using a mean value for detecting an outlier an outlier can be considered as a normal data point. This reduces the efficiency of the method and makes it a nonrobust estimator.

2. *Standard Deviation*

Standard deviation is another estimator used to find out outliers. It is also not very efficient estimate ,the reason for this is given below:

The **standard deviation** (represented by the Greek letter sigma, **σ**) shows how much variation or dispersion from the average/mean exist [19].

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (7)$$

A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value); a high standard deviation indicates that the data points are spread out over a large range of values [20].

For this outlier detection method, the mean and standard deviation of the residuals are calculated and compared. If a value is a certain number of standard deviations away from the mean, that data point is identified as an outlier. The specified number of standard deviations is called the threshold. The default value is 3.

This method can fail to detect outliers because the outliers increase the standard deviation. The more extreme the outlier, the more the standard deviation is affected [21].

There are three problems to face using mean. 1) It assumes that the distribution is normal including outliers. 2) The mean and standard deviation are strongly influenced by outliers. 3) As stated by Cousineau and Chartier (2010), this method is very unlikely to detect outliers in small samples [22].

Reference [23] illustrates this problem with example given below:

Accordingly, this indicator is fundamentally problematic:

It is supposed to guide our outlier detection but, at the same time, the indicator itself is altered by the presence of outlying values. In order to appreciate this fact, consider a small set of n = 8 observations with values 1, 3, 3, 6, 8, 10, 10, and 1000. Obviously, one observation is an outlier (and we made it particularly salient for the argument). The mean is 130.13 and the uncorrected standard deviation is 328.80.Therefore, using the criterion of 3 standard deviations to be conservative, we could remove the values between − 856.27 and 1116.52. The distribution is clearly not normal (Kurtosis = 8.00; Skewness = 2.83), and the mean is inconsistent with the 7 first values. Nevertheless, the value 1000 is not identified as an outlier, which clearly demonstrates the limitations of the mean plus/minus three standard deviations method [23].

3. The median absolute deviation (MAD)

The MAD overcomes these problems. In [23], authors have illustrated the efficiency of MAD over mean and standard deviation with example. Which is given here as follows: The median (M) is, like the mean, a measure of central tendency but offers the advantage of being very insensitive to the presence of outliers. One indicator of this insensitivity is the “breakdown point”[24]. The estimator's breakdown point is the maximum proportion of observations that can be contaminated (i.e., set to infinity) without forcing the estimator to result in a false value (infinite or null in the case of an estimator of scale). For example, when a single observation has an infinite value, the mean of all observations becomes infinite; hence the mean's breakdown point is 0. By contrast, the median value remains unchanged. The median becomes absurd only when more than 50% of the observations are infinite. With a breakdown point of 0.5, the median is the location estimator that has the highest breakdown point.

Exactly the same can be said about the Median Absolute Deviation as an estimator of scale (see the formula below for a definition). Moreover, the MAD is totally immune to the sample size. These two properties have led [25] to describe the MAD as the “single most useful ancillary estimate of scale” (p. 107). It is for example more robust than the classical inter quartile range [26], which has a breakdown point of 25% only.

To calculate the median, observation has to be sorted in ascending order. Let us consider the previous statistical series: 1, 3, 3, 6, 8, 10, 10, and 1000. The average rank can be calculated as equal to (n + 1) / 2 (i.e., 4.5 in our example). The median is therefore between the fourth and the fifth value, that is, between six and eight (i.e., seven).

Calculating the MAD involves finding the median of absolute deviations from the median. the MAD is defined as follows [25]:

$$MAD = b \;\; M_i\left(\left|x_i - M_j\left(x_j\right)\right|\right) \tag{8}$$

where the x_j is the n original observations and M_i is the median of the series. Usually, b = 1.4826, a constant linked to the assumption of normality of the data, disregarding the abnormality induced by outliers (Rousseeuw & Croux, 1993).

Calculating the MAD implies the following steps:

(a) the series in which the median is subtracted of each observation becomes the series of absolute values of (1–7), (3–7), (3–7), (6–7), (8–7), (10–7), (10–7), and (1000–7), that is, 6, 4, 4, 1, 1, 3, 3, and 993; (b) when ranked, we obtain: 1, 1, 3, 3, 4, 4, 6, and 993; (c) and (d) the median equals 3.5 and will be multiplied by 1.4826 to find a MAD of 5.1891.

To calculate MAD all the observations has to be sorted first. This can be a huge overhead in large data set.

It requires preprocessing and it can become time consuming and in highly dynamic data it may become more difficult to hold a correct value.

V. WORKING

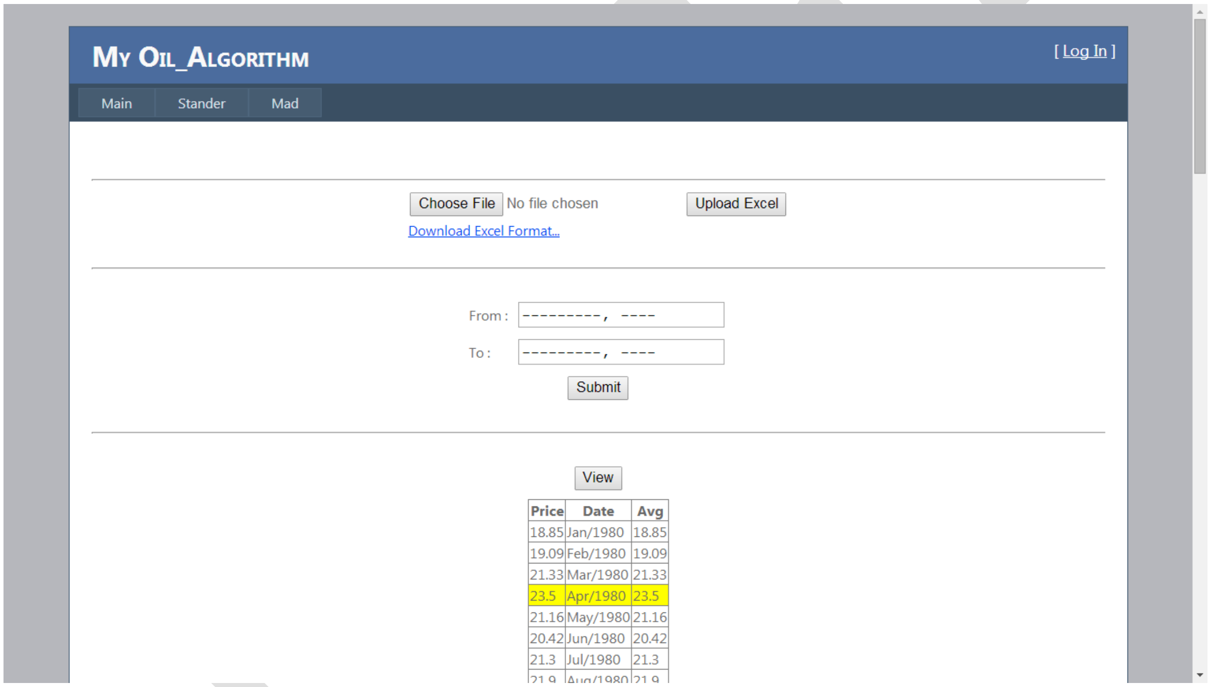
The algorithm is applied on the time series data set. The data values in data set are used to calculate MAD values. This calculated MAD value is used to determine surprising values from the given data set by comparing them with the MAD. Values which are 3 times away from the

Median values are considered as Outliers. Periodicity of detected outliers are calculated by above mentioned Periodicity formula in Section III.

VI. EXPERIMENT EVALUATION

For this experiment time series dataset of monthly oil prices are used. The crude oil price are selected for particular time window. The algorithm is applied to this selected values and by calculating MAD , outliers from this range of values are detected. Once outliers are detected their periodicity is detected for entire dataset.

The experiment is carried out using mean, Standard deviation and MAD. In the following snap shots, detected outliers are highlighted in yellow colour. First of all we select the range/time window in which we want to detect outliers. Select the method with which we want to detect the outliers i.e. Mean, Standard Deviation or MAD. Snapshot given below is the one using mean method.



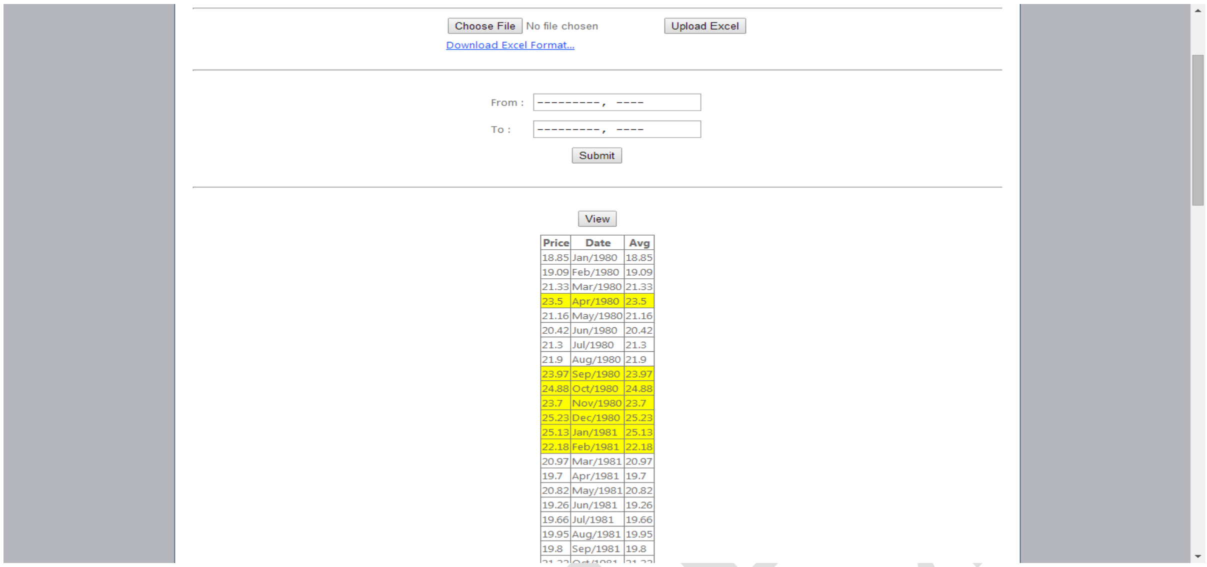
From : ---------, ----

To : ---------, ----

Price	Date	Avg
18.85	Jan/1980	18.85
19.09	Feb/1980	19.09
21.33	Mar/1980	21.33
23.5	Apr/1980	23.5
21.16	May/1980	21.16
20.42	Jun/1980	20.42
21.3	Jul/1980	21.3
21.9	Aug/1980	21.9

Here the detected outlier is of price 23.5 as it deviates greatly from other values.

Same experiment is done with the use of MAD value, the result of it is shown by highlighting the detected outliers with yellow color.



From : ---------, ----

To : ---------, ----

Price	Date	Avg
18.85	Jan/1980	18.85
19.09	Feb/1980	19.09
21.33	Mar/1980	21.33
23.5	Apr/1980	23.5
21.16	May/1980	21.16
20.42	Jun/1980	20.42
21.3	Jul/1980	21.3
21.9	Aug/1980	21.9
23.97	Sep/1980	23.97
24.88	Oct/1980	24.88
23.7	Nov/1980	23.7
25.23	Dec/1980	25.23
25.13	Jan/1981	25.13
22.18	Feb/1981	22.18
20.97	Mar/1981	20.97
19.7	Apr/1981	19.7
20.82	May/1981	20.82
19.26	Jun/1981	19.26
19.66	Jul/1981	19.66
19.95	Aug/1981	19.95
19.8	Sep/1981	19.8

VII. CONCLUSION

Here a novel algorithm for the periodicity detection of outlier, surprising, or unusual patterns is shown. It makes use of the MAD value to compare relative frequency of the outlier pattern instead of mean value which was previously used in the existing algorithm. As the mean method is not robust and do not give the accurate results. It can easily be affected with the presence of outlier. A new measure known as Median Absolute Deviation is used to detect outlier instead of mean, as it is more efficient compare to mean. It increases the accuracy of the existing algorithm. In the carried out experiments outliers detected by MAD are more accurate. The algorithm also takes into account the likelihood of pattern occurrence to classify it as an outlier pattern.

REFERENCES

[1] M. Ahdesmki, H. Lhdesmki, R. Pearson, H. Huttunen, and O. Yli-Harja,“Robust detection of periodic time series measured from biological systems,” *BMC Bioinformatics*, vol. 6, p. 117, 2005.

[2] E. F. Glynn, J. Chen, and A. R. Mushegian, “Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms,” *Bioinformatics*, vol. 22, no. 3, pp. 310–316, Feb.2006.

[3] A. Weigend and N. Gershenfeld, *Time Series Prediction: Forecasting the Future and Understanding the Past*. Reading, MA, USA: Addison-Wesley, 1994.

[4] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid, “Periodicity detection in time series databases,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 7, pp. 875–887, Jul. 2005.

[5] P. Indyk, N. Koudas, and S. Muthukrishnan, “Identifying representative trends in massive time series data sets using sketches,” in *Proc. Int. Conf. Very Large Data*, Sep. 2000, pp. 363–372.

[6] C. Sheng,W. Hsu, andM.-L. Lee, “Mining dense periodic patterns in time series data,” in *Proc. IEEE Int. Conf. Data Eng.*, 2005, p. 115.

[7] C. Sheng, W. Hsu, and M.-L. Lee, “Efficient mining of dense periodic pattern time series,” Nat. Univ. Singapore, Singapore, Tech. Rep. 1,2005.

[8] E. Keogh, S. Lonardi, and B. Y.-C. Chiu, “Finding surprising patterns in a time series database in linear time and space,” in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2002, pp. 550–556.

[9] C. Shahabi, X. Tian, andW. Zhao, “Tsa-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries,” in *Proc. 12th Int. Conf. Sci. Statist. Database Manage.*, 2000, pp. 55–68.

[10] J. Yang, W. Wang, and P. S. Yu, “Infominer: Mining surprising periodic patterns,” in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2001, pp. 395–400.

[11] C. Sheng,W. Hsu, andM.-L. Lee, “Mining dense periodic patterns in time series data,” in *Proc. IEEE Int. Conf. Data Eng.*, 2005, p. 115.

[12] J. Han, W. Gong, and Y. Yin, “Mining segment-wise periodic patterns in time related databases,” in *Proc. ACM Int. Conf. Knowl. Discov. Data Mining*, vol. 8, no. 1, pp. 53–87, Aug. 1998.

[13] S. Ma and J. Hellerstein, “Mining partially periodic event patterns with unknown periods,” in *Proc. IEEE Int. Conf.Data Eng.*,Apr. 2001, pp. 205–214.

[14] C. Berberidis, W. Aref, M. Atallah, I. Vlahavas, and A Elmagarmid,“Multiple and partial periodicity mining in time series databases,” in *Proc. Eur. Conf. Artif. Intell.*, Jul. 2002, pp. 79–94.

[15] J. Han, Y. Yin, and G. Dong, “Efficient mining of partial periodic patterns in time series database,” in *Proc. IEEE Int. Conf.Data Eng.*, 1999, pp. 106–115.

[16] K.-Y. Huang and C.-H. Chang, “SMCA: Ageneral model for mining asynchronous periodic patterns in temporal databases,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 774–785, Jun. 2005.

[17] Faraz Rasheed and Reda Alhajj," A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences" *IEEE TRANSACTIONS ON CYBERNETICS,*April 2013

[18] Dhwani Dave and Tanvi Varma,"A Review of various statestical methods for Outlier Detection" , *International Journal of Computer Science & Engineering Technology (IJCSET),* Vol. 5 No. 02 Feb 2014, , pp. 137–140

[19] Bland, J.M.; Altman, D.G. (1996). "Statistics notes: measurement error.". *Bmj, 312(7047), 1654.*

[20] http://en.wikipedia.org/wiki/Standard_deviation

[21]http://docs.oracle.com/cd/E17236_01/epm.1112/cb_statistical/frameset.htm?ch07s02s10s01.html

[22] Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review.
International Journal of Psychological Research ,3(1),Pages 58–67

[23] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, Laurent Licata, Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median *Journal of Experimental Social Psychology, Volume 49, Issue 4, July 2013, Pages 764-766*

[24] Donoho, D. L., & Huber, P. J. (1983). In Bickel, Doksum, & Hodges (Eds.), The notion of breakdown point. California: Wadsworth

[25] Huber, P. J. (1981). Robust statistics. New York: John Wiley

[26] Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation.Journal of the American Statistical Association, 88(424), 1273–1283