Detecting and Revamping of X-Outliers in Time Series Database

S.Sridevi
Assistant Professor
Department of CSE
Thiagarajar College of Engg, Madurai

S.Abirami
PG Student
Department of CSE
Thiagarajar College of Engg, Madurai

S.Rajaram, PhD.
Associate Professor
Department of ECE
Thiagarajar College of Engg, Madurai

ABSTRACT
Dataset with Outliers causes poor accuracy in future analysis of data mining tasks. To improve the performance of mining task, it is necessary to detect and revamp of outliers which are there in the dataset. Existing techniques like ARMA (Auto-Regressive Moving Average), ARIMA (Auto- Regressive Integrated Moving Average) and Multivariate Linear Gaussian state space model don't consider the periodicity for outlier detection. The above methods are used to find out only Y Outliers which are present in Y axis. These methods are not applicable to detect the time at which the peculiar data occurs (so called X-Outliers). This paper focuses different methods for detecting and revamping of X-Outliers that have abnormal data according to a known periodicity. These are practically applied in fraud detection, Market-basket analysis and medical applications to detect certain abnormal diseases. First the data is modeled to get the trend of the data and to remove noises by means of kernel smoothing. Next the outliers are detected by similarity measurements. If the dataset has outliers it can be replaced by considering periodic indices from the historical dataset. The performance of system is measured by precision, recall and F Score. The proposed method is tested with three different time series datasets namely, Electricity power consumption dataset, Weather dataset and Electricity price market dataset. Experimental results have demonstrated that the proposed method is effective and accurate than the earlier methods.

Keywords
Outlier, Time series database, Smoothing methods, Revamping

1. INTRODUCTION
Temporal data mining deals with the harvesting of useful information from temporal or time series database. Temporal database is a database which contains time related information like year, month, week, day, minute and second. A Time series database [2] is a special type of temporal database with built-in time aspects. A time series [4] is a sequence of real numbers, each number representing a value of an attribute of interest, observed at different point of time. Typical examples include stock prices, currency exchange rates, the volume of product sales, biomedical measurements, weather data, etc, collected over monotonically increasing time.

Gathering all data accurately in fine granularity is a challenging task. There is often missing and corrupted data in the process of information collection and transfer, due to various causes including wrong data entry, communication failures, outages, lost data, unexpected shutdown, un-scheduled maintenance, temporary weather changes due to cyclone and some other unknown factors. When these set of data are used for future analysis such as association rule mining, clustering and prediction which results in poor accuracy. To improve the accuracy it is necessary to replace the missing data, corrupted data and abnormal data. This paper focuses on how to handle the abnormal data (called as outliers).

Many data mining algorithms try to minimize the influence of outliers or eliminate them all together. Hence the task is to identify and repair such abnormal time series data by considering periodicity. Thus, outlier detection [5] and analysis is an interesting data mining task, referred to as outlier mining. There are three fundamental approaches of outlier detection: Unsupervised, Semi-Supervised [8] and Supervised detection. An Unsupervised method is used to determine the outliers with no prior knowledge of the data. This approach processes the data as a static distribution, pinpoints the most remote points, and indicates them as potential outliers. Semi-supervised approach needs pre-classified data. This approach is suitable for static or dynamic data as it only teaches one class which provides the model of normality. It can learn the model incrementally as new data arrives, tuning the model to improve the fitness of the function .It aims to define a boundary of normality. The third one is the supervised detection which models both normality and abnormality. This approach is analogous to supervised classification and requires pre- labeled data, tagged as normal or abnormal.

The above three methods are further classified into Y-Outlier [9] or X-Outlier detection [12]. Y Outliers are used to detect abnormal data which are present in Y axis and it won't consider periodicity. X-outliers are used to detect Y-outliers as well as the time at which the abnormal data occurs. This paper presents supervised X-Outlier Detection[1] because it needs some prelabeled data for performance measurements.

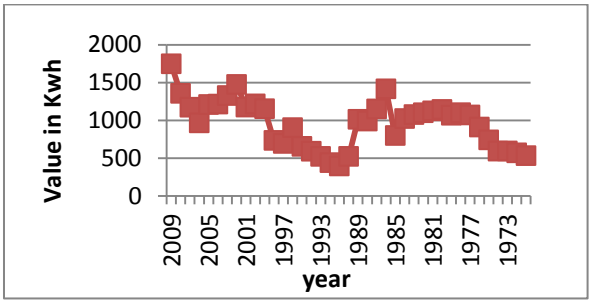

Fig1: Electricity power consumption dataset with Outliers (before smoothing)

.

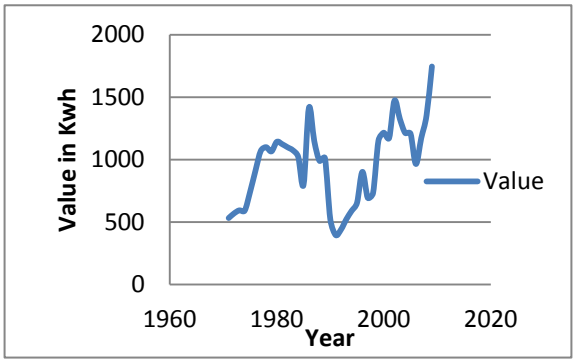

Fig 2: Power consumption data after smoothing

The aim of this paper is to detect and revamping of X-Outliers by means of Kernel smoothing since X-Outliers can't be detected by neighborhood based techniques such as moving average and other smoothing methods. To illustrate these points consider the electricity power consumption input data which was shown in figure 1.The above figure contains Yearly periodicity in X axis and the corresponding power consumption value (in Kwh) is denoted in Y axis. After smoothing, figure 1 can be shown as like in figure 2. Smoothing is used to obtain the trend of the data and to remove noises.

The contributions of the paper are as follows: i)To get the trend of the data the proposed work uses three probability density functions such as normal or Gaussian, Poisson and exponential and the result was compared with one another. ii) Detection of X-Outliers using LCSS(Longest Common Subsequence) according to the known periodicity iii) Revamping of X-Outliers iv) The proposed method was tested with three different time series datasets namely Electric power consumption dataset, weather dataset and electricity price market dataset and its performance was measured.

The rest of this paper is organized as follows. Section 2 describes related work, section 3 defines proposed method. Experimental result and performances of the proposed method are reported in section 4 and section V covers conclusion and future work.

2. RELATED WORK

Peculiarity Oriented Multidatabase Mining proposed by Ning Zhong, Yiyu Yao, and Muneaki Ohshima discussed in [1], presents a method of Peculiarity mining for a single source and it is extended to multidatabases. This method doesn't consider X axis value. Outlier detection by considering only the Y-axis was discussed in [2]. B-Spline smoothing method is used in this paper to clean the corrupted and missing data. This method does not consider the periodicity hence the existence of outliers still exists. Various methodologies for outlier detection were proposed by V. J. Hodge and J. Austin [3]. Different methods discussed in this paper for outlier detection are Non- parametric methods and parametric methods. Here the survey convey that the developer should select an algorithm that is suitable for their data set in terms of the correct distribution model, the correct attribute types, and the scalability for outlier detection.

Auto- Regressive Moving Average Model(ARMA) assumes that the time series is stationary. This assumption cannot handle relatively large amount of data. This ARMA method was discussed by A. J. Fox in [4]. The Auto- Regressive integrated Moving Average Model (ARIMA) model treats each outlier as a single observation and detects multiple point outliers as a sequence of observations. If multiple outliers exist in a close proximity, these outliers may mask each other so that no points are identified as outliers. Besides, the ARIMA method requires considerable computer time and memory for a long time series. This method was explained by G. M. Ljung in [5]. To detect the outliers, the regression relationship of the data can be modeled by a continuous function. The estimated value in the regression function at given time is modeled in a form of nonparametric regression. This was explained by W. Hardle in [6].It also doesn't consider the periodicity.

In[9] I. Chang and G. C. Tiao explained the Multivariate Linear Gaussian state space model which provides a more general modeling technique for time series and it also allows for non-stationary models. The state space model has primarily been used for forecasting. The drawbacks of Multivariate Linear Gaussian state space model is it cannot handle a relatively large portion of missing data. In [10] the existing methods for time series in statistics explained by D. Gasgupta and S. Forres gives a common idea to slide a limited window across the time series data. This approach has to predefine the window size for searching anomalous subsequences. The length of abnormal data can vary considerably from a few hours to several months, and therefore, the idea of determining a proper window size in advance does not work. Many statistical tests exist [11] for outliers include Z-Value, box plot, Dixon test and Grubbs test. The common assumption for these test method is the normal distribution of the data, which is invalid for the power consumption data. The above problem can be overcome by detecting and revamping X-Outliers by kernel smoothing and similarity match which was introduced in [12].

3. PROPOSED WORK

In the proposed work a novel class of X-Outliers is presented under the assumption that periodicity is known. Because without knowing periodicity we can't able to predict at which time the Y-Outlier occurs. In this paper we assume that the curve follows the loose periodicity and it should be known to the user. The length of the periodicity can be found by regularity of the underlying patterns in the dataset. For example, for a dataset D of monthly periodicity, the length is 30 or 31 days. Due to cyclone, strike or some other factors the company shutdowns 5 days means then the length of X-Outlier is 5days and the remaining 25 or 26 days are considered as a normal or casual events. In this paper the curve T and Sub curve A is used for smoothing and to detect Outliers. The curve T=$\{_{i=1}^{n}(t_i, y_i)\}$ *is an* ordered sequence of n real valued observations where y_i is the observed data at time t_i. The sub curve A= =$\{_{i=j}^{k}(t_i, y_i)\}$ *of T is a continuous part of* T where i≤j and k≤n. The block diagram of the proposed work is shown in figure 3. Each section of the block diagram is explained as follows:

3.1 Kernel Smoothing

It is the first and important step to detect outliers. To get the trend of the data and to remove background noises kernel smoothing is used. It is a statistical technique for estimating a real valued function when no parametric model for this function is known. The estimated function is smooth, and the

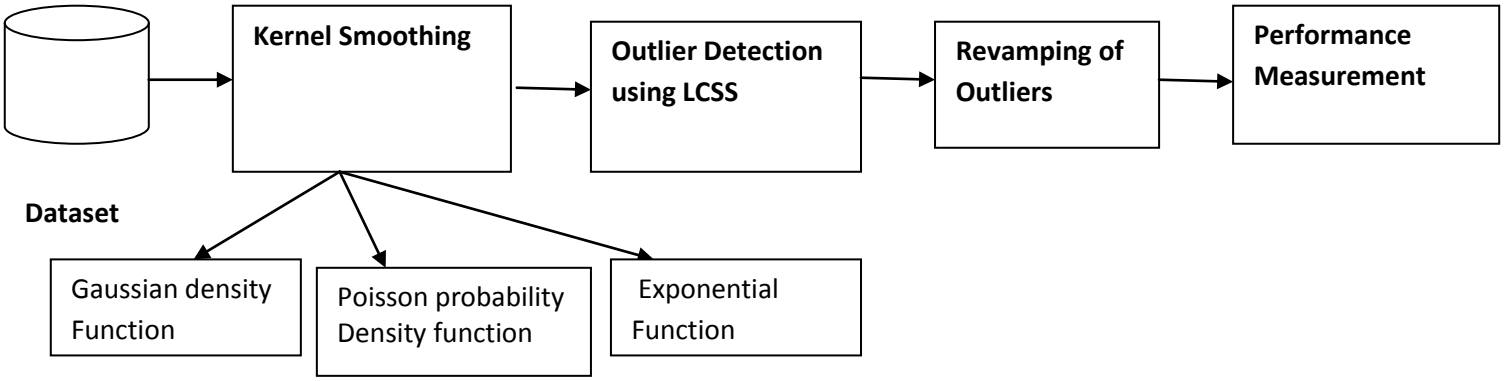

Fig 3: Block diagram of the Proposed Method

level of smoothness is set by a single parameter called smoothing parameter h. In the existing work [12] kernel is chosen as the normal probability density function.

$$\ker n(t) = \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}t^2} \quad (1)$$

Equation (1) can be obtained from (2). The general normal probability density function is given by the formula

$$\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(t-\mu)^2}{2\sigma^2}) \quad (2)$$

Where μ *is the mean* ,$\sigma$the standard deviation and t is is the periodicity of the data. The standard normal distribution has $\mu = 0$ *and* σ=1. Hence equation (2) can be written as in (1).

In the proposed work we used two more probability density function such as poission and exponential. It is given by the formula

Poission : $\text{kern}(t) = \frac{e^{-\lambda}\lambda^t}{t!}$ (3)

Exponential: : $\text{kern}(t) = \lambda e^{-\lambda t}$ (4)

Where λ is the smoothing parameter and t is the periodicity of the data. It is given by

$\lambda = 10^{(i-1)/2-9}$ where i=(1,2,…..,10) (5)

The regression relationship of the data can be given by the formula.

$$y_i = m(t_i) + \varepsilon_i \quad (6)$$

where m(t_i) is regression function and ε *is* an observed error. The estimated value at time t is modeled in a form of nonparametric regression is given by

$$\hat{m}(t) = \frac{1}{n}\sum_{i=1}^{n} w_i(t) y_i \quad (7)$$

This equation is called as Smoothing curve. Here y_i is the consumption value at time t , n is the size of dataset ,$w_i(t)$is the weight of point at time t. In smoothing $w_i(t)$ is given by

$$w_i(t) = \frac{\ker n_h(t-t_i)}{\hat{f}_h(t)} \quad (8)$$

Where kernel is given by

$$\ker n_h(t) = \frac{1}{h}\ker n\frac{t}{h} \quad (9)$$

h is bandwidth or smoothing parameter. The parameter h is given by

h= $\frac{1}{480-45xi}$ where i=(1,2,…..,10). (10)

Level 1 yields the roughest smoothing curve and the higher level give the soft smoothing curve. These levels are not equally spaced it can be chosen randomly.

Kernel density estimator is given by

$$\hat{f}_h(t) = n^{-1}\sum_{i=1}^{n} \ker n_h(t-t_i) \quad (11)$$

Nadaraya-Watson estimator for density estimator is given by

$$\hat{m}_h(t) = \frac{n^{-1}\sum_{i=1}^{n} \ker n_h(t-t_i) y_i}{n^{-1}\sum_{i=1}^{n} \ker n_h(t-t_i)} \quad (12)$$

In the proposed work three density functions are used which are given in equations (1), (3), and (4) to obtain the smoothness of the curve.

3.2 Outlier Detection

Outliers can be detected by comparing two sub curves say A, B of T.First of all the sub-curves are extracted from kernel smoothing. The length of the sub-curve should be matching with one another.There are two reasons to find the similarity between two sub curves. First reason is ,it is less sensitive to time shifting, scaling and stretching and the second reason is to detect the outliers based on the similarity measurement. In this paper we used LCSS (Longest Common Subsequences) to find similarity match. Popular Distance functions such as Euclidean Distance, Minkowski Distance and Manhattan distance can't be used here because it is insensitive to time shifting and stretching and moreover it doesn't consider periodicity when calculating the similarity measurements.

LCSS can be explained as follows: Consider two sub-curves say A= (a_1, a_2,…….., a_m) and =(b_1,b_2,…….,b_n) from the

curve T to find out the similarity match. LCSS is given by the formula

$$D_{i,j} = \begin{cases} 0, if \quad i = 0 \quad \vee j \quad = 0 \\ 1 + D_{i-1,j-1}, if \quad |a_i - b_j| \quad \leq \quad \varepsilon \wedge |i - j| \leq \delta \\ \max(D_{i,j-1}, D_{i-1,j} Otherwise \end{cases} \quad (13)$$

in the above equation, ε is y axis value stretching threshold which represents tolerance of noises and δ is time stretching threshold which represents tolerance of time shifting and stretching. In the above equation the first statement is used for initialization and the second statement calls the similarity function recursively until the condition fails. If the condition is false it will move to the next statement. From the above equation dissimilar values in one or or both curves can be identified and treated as a outlier The LCSS similarity can be measured by

$$\gamma(\delta, \varepsilon, A, B) = \frac{D_{|A|,|B|}}{\min(|A|, |B|)} \quad (14)$$

Where $D_{A,B}$ is the length of the common subsequence to both A and B which can be computed from the equation (13). IAI and IBI are the length of the two sub curves say A and B.

The sub curves A and B are similar if

$$\gamma(\delta, \varepsilon, A, B) \geq \theta \quad (15)$$

θ is the user defined threshold. If suppose the equation (15) is not satisfied then the user has to identify the outliers and revamped them.

3.3 Revamping Of Outliers

After Outliers are detected, it should be replaced by either method 1 or method 2 because it will not happen again and it won't affect the future analysis. The replacing data is selected from normal data for the corresponding time interval in other period.

3.3.1 Method1

In this method the outliers can be replaced by mean or standard deviation of the particular periodicity. For example If the periodicity follows monthly periodicity and the april month has two outlier means it can be replaced by considering mean or standard deviation of all the observed value (except the outlier data) of April month .

3.3.2 Method2

In this method Outliers can be revamped by considering the previous and next periodicity and its corresponding observed value. The periodic index at time t_i is computed as follows:

$$P(t_i) = \frac{y_i}{T(t_i)} \quad (16)$$

Where y_i is the observed value at time t_i and $T(t_i)$ represents long-term trends.The periodic index $P(t_i)$ at time t_i belonging to an outlier is estimated by the average of periodic indexes at the corresponding time of its previous and next period. It is given by ,

$$P(t_i) = \frac{1}{2}(P(t_i - l) + P(t_i + l)) \quad (17)$$

where l is the length of the periodicity. Finally the outlier data can be replaced by the formula

$$Y(t_i) = T(t_i) * P(t_i) \quad (18)$$

3.4 Performance Measurement

The performance of system is measured by Precision, Recall and F-measure. Precision (P) is the percentage of detected observations that were pre-labeled. Here, D denote the set of observations that belong to some outliers and E denote the set of observations that belong to some pre-labeled outliers. Precision is given as

$$P = \frac{|E \cap D|}{|D|} \quad (19)$$

Recall (R) is the percentage of pre-labeled observations that are detected. It is given as $R = \frac{|E \cap D|}{|L|}$ (20)

F measure is given by

$$F = \frac{2 * P * R}{P + R} \quad (21)$$

4. EXPERIMENTAL RESULT

4.1 Data Selection

Three datasets are tested with the proposed work. The first dataset is the electricity power consumption which was collected from www.data.un.org. It has three attributes namely country, year, and value in Kwh.This dataset is yearly electricity consumption for different countries. In the proposed work the fields are selected for albania country from the year 1971 to 2009, with 50 observations. The same input is used for the traditional smoothing methods also to compare the efficiency. The second dataset is the weather dataset and it has following fields: Date, StationId, maxtemp, mintemp, maxwet bulb and min wet bulb which was collected from www.nysio.com. The proposed work supports only univariate data along the periodicity. Hence in the above dataset only three fields are consider namely Date, StationId, and maxtemp. In the proposed work the fields are selected for the particular station from the January 1, 2012 to till date.

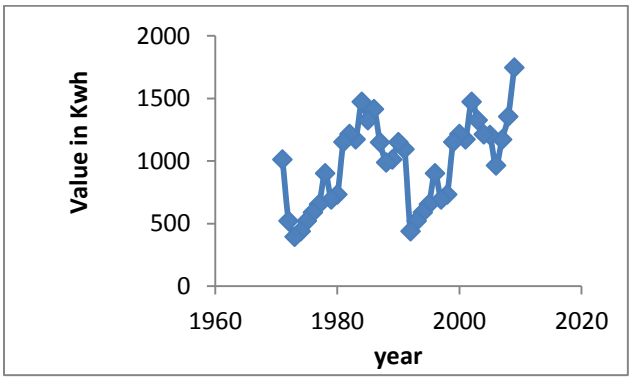

Fig 4: Smoothing of power consumption data using Exponential density function

The third dataset is the electricity price market which was collected from www.aemo.co..au. It has the following attributes: Region, date, price and demand .This dataset contains hourly electricity price rate for different regions. The dataset is identified to have some outliers and were pre-labeled. We use such pre-labeled outliers as the class label to evaluate the accuracy of the proposed method. The output of kernel smoothing using Poisson, exponential and normal density function was shown in figure 4, 5 and 6.

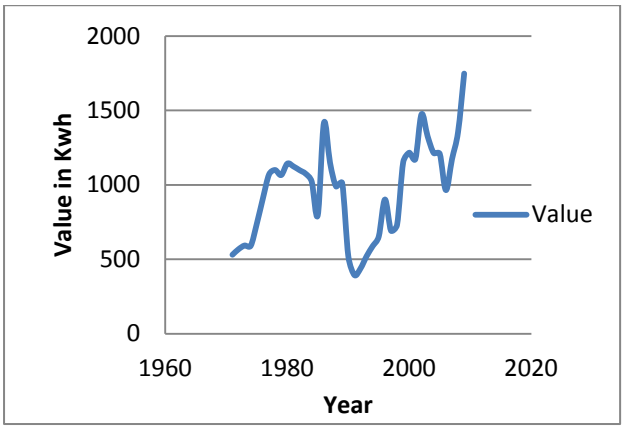

Fig 5: Power consumption data after smoothing (using Normal density function)

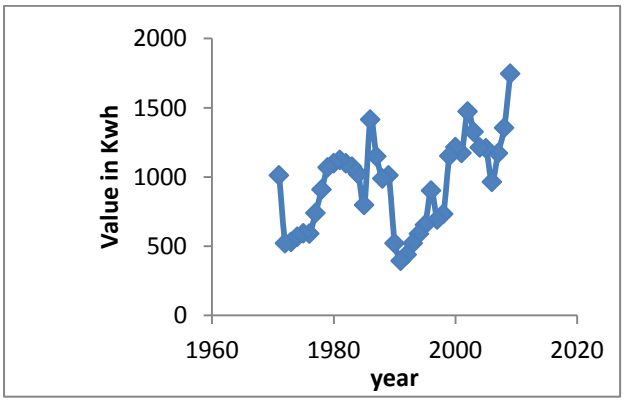

Fig 6: Power consumption data after smoothing (using Poisson density function)

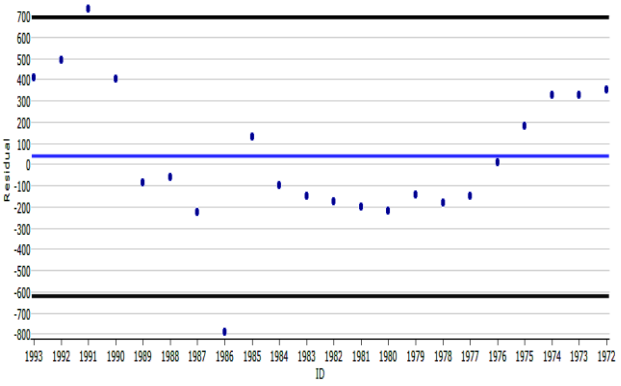

Fig 7: Outlier Detection

From the above figure it was shown that kernel smoothing using normal distribution gives better result when compared with the other distributions such as Poisson and exponential functions. In the above figure the performance was measured in terms of no of outliers that are detected by varying the smoothing parameters h and **λ.** After getting the trend of the data the outlier was detected using LCSS which was shown in fig 7. These outliers can be replaced by using either method or method 2 which was discussed in section 3.3.1 and 3.3.2. If the data values lie below and above the black lines then that are treated as outliers in fig 7.Implementation result shows that method 2 is slightly better than method 1. So here the results are compared based on method2.

4.2. Comparisons

In this section the proposed work is compared with the traditional smoothing method which was introduced in [2]. The accuracy of the proposed work was tested with X-Outliers and without X-Outliers by using the metrics precision, recall and F-measure. Electricity power consumption dataset with no X-Outliers and with X-Outliers using normal density function are shown in table 1 and 2.

Table 1: Electricity Power Consumption dataset with No X-Outliers

Smoothness Level h	P/T	
	Proposed Method	Traditional Smoothing Method
1	1.4%	5.7%
3	1.2%	4.6%
5	0.6%	3.4%
7	0%	3.2%

Table 2: Electricity Power Consumption dataset with X-Outliers

Smoothness Level h	Proposed Method			Traditional Smoothing Method		
	P	R	F	P	R	F
1	86%	97%	91%	48%	74%	58%
3	88%	96%	92%	51%	72%	60%
5	94%	96%	95%	65%	72%	68%
7	98%	95%	96%	75%	68%	71%

In table 1 P denotes no of Outliers that are detected by an algorithm in the Electricity power consumption dataset and T denotes total no of instances in that dataset. P/T is the percentage of points that are incorrectly detected. In table 2 precision, recall and F –measure are calculated using the equations 19, 20 and 21.

From the above table it is noted that the proposed system gives accurate result for the higher values of smoothing

parameter h. In the above table the traditional smoothing method has higher P/T since it has higher false positives. Moreover the result is based on the parameters h, $\varepsilon, \theta, \delta$ which was used in Similarity measurements. The best result of the proposed method for electricity dataset was achieved at smoothness level 7, with P- 98% and R-95% respectively. A large F value indicates more accurate result.

Table 3: Weather dataset with No X-Outliers

Smoothness Level h	P/T	
	Proposed Method	Traditional Smoothing Method
1	2.2%	4.6%
3	1.6%	3.8%
5	0.7%	2.4%
7	0.3%	1.8%

Table 3 and 4 shows that efficiency of weather dataset with and without X-Outliers. From the above, it is concluded that the proposed method is effective and accurate than the earlier methods. The efficiency of the proposed method was also tested with electricity price market dataset. It was achieved the good result at the smoothness level 8 , with P and R are 95% and 84% respectively.

Table 4: Weather dataset with X-Outliers

Smoothness Level h	Proposed Method			Traditional Smoothing Method		
	P	R	F	P	R	F
1	78%	90%	84%	50%	80%	62%
3	83%	88%	85%	58%	78%	6%
5	88%	86%	87%	62%	72%	67%
7	92%	84%	88%	73%	70%	71%
8	93%	84%	88%	81%	67%	73%

5. CONCLUSION AND FUTURE WORK

The time series dataset which contains the abnormal data has been detected and repaired by considering the periodicity. Low-quality data will lead to low-quality mining results. Hence it was preprocessed using Kernel smoothing to get trend of data. For smoothing the different kernel selected are normal distribution, poisson distribution and exponential distribution. Among these, normal distribution gives better result. In the proposed work the outliers were replaced by considering periodicity indices. The proposed work gives better result for higher value of smoothing parameter h.The proposed work can be applied to wide range of datasets and applications. My future work will be focused on outlier detection for multivariate data along the periodicity and to predict the occurrences of future outliers.

6. REFERENCES

[1]. Ning Zhong, Yiyu (Y.Y.) Yao, Muneaki Ohshima, "Peculiarity Oriented Multidatabase Mining", IEEE Trans on Knowledge and Data Engg, Vol. 15, No. 4,pp.613-628, July/August 2003.

[2]. J. Chen, W. Li, A. Lau, J. Cao, and K. Wang, "Automated load curve data cleansing in power systems",*IEEE Trans. Smart Grid*, vol. 1, no.2, pp. 213-221, Sep. 2010.

[3]. V. J. Hodge and J. Austin, "A survey of outlier detection methodologies", *Artif. Intell. Rev.* , vol. 22, no. 2, pp. 5-126, Oct. 2004.

[4]. A.J. Fox, "Outliers in time series",*J. Roy. Stat. Soc. B ,, Methodol*.vol.34, pp. 350-363, 1972.

[5]. G.M. Ljung,"On outlier detection in time series",*J. Roy. Stat. Soc. B, Methodol.*, vol. 55, pp. 559-567, 1993.

[6]. W.Hardle, "Applied Nonparametric Regression", Cambridge,U.K .Cambridge Univ.Prss,1990.

[7]. B. Abraham and A. Chuang,"Outlier detection and time series modeling",*Technometrics* , vol. 31, pp. 241-248, 1989

[8]. W. Schmid,"The multiple outlier problems in time series analysis",*Australian. J. Statist.*, vol. 28, pp. 400-413, 1986

[9]. I. Chang and C. iao,"Estimation of time series parameters in the presence of outliers",*Technometrics*, vol. 30, pp. 193-204, 1988

[10]. D. Gasgupta and S. Forrest,"Novelty detection in time series data using ideas from immunology", *in Proc. Int. Conf. Intelligent Systems*, 1996, pp. 82-87.

[11]. V.Barnett and T.Lewis, " Outliers in Statistical Data",3 rd edition, New York, Wiley 1994,pp 397-415.

[12]. Zhihui Guo, Wenyuan Li, Fellow, Adriel Lau, Tito Inga-Rojas, and Ke Wang, "Detecting X-Outliers in Load Curve Data in Power Systems",*IEEE trans. on power systems*, VOL. 27, NO. 2, MAY 2012.