A GENERIC APPROACH FOR OUTLIER DETECTION IN TIME-SERIES DATA

SOURABH SUMAN[a1], B. RAJATHILAGAM[b] AND KARTHIK VAIDHYANATHAN[c]

[ab]Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, India

[c]Product Lead, Knowledge Lens, Bengaluru, India

ABSTRACT

This paper proposes a novel method to detect any outlier in a time-series data by analyzing the data in frequency domain and the chunk mean taken in the time domain. The combination of these two techniques increases the accuracy in detecting an outlier. The idea behind this approach is that whenever there is an outlier in the time series data, information would shift in higher frequency with greater magnitude. Thus, detecting this information will help to detect an outlier. This technique is not specific to a particular source generating time-series data. Instead, it is a generic approach to detect an outlier in any type of time-series data.

KEYWORDS: Chunk Mean, FFT, Outlier and Time-Series Data.

Data points indexed with respect to time in a sequence is a time-series data. These data contain very useful information. One of the features that a data encase is an outlier. A sudden change in behavior of a data which is not related to general behavior of the data is an outlier [1] [2]. An outlier constitutes information relating to the abnormality in the data. The information conveyed by an outlier is then analyzed to mitigate the abnormality created. Hence, detecting an outlier in a data becomes very significant. Detecting outliers in time series data in any industry is very important as they contain very useful information. This information could be any defect in the component of the machine which needs to be replaced or it can be due to human failures. In nuclear reactors, outliers are very sensitive as little ignorance can cause a catastrophe. As soon as any outlier is detected, nuclear fission is brought down to stay under the critical limit. In border areas with sensors installed, outliers help to detect any trespassing. In medical field, outliers help to keep check on the health of the patient. Outliers are also helpful for cyber security purpose as any kind of hacking can be detected by keeping a check on web traffic. Any fraudulent activity in banks can be checked by presence of any outlier in the money flow from an account. Sudden increase in the TRP of any program on television is also an example of outlier which helps to decide the cost of advertisement. In chemical industries, outliers can be very effective to analyze the sudden increment in the concentration of a chemical. Hence, detecting an outlier becomes very important.

In this paper, a new approach has been proposed to detect any outlier by analyzing the signal in the frequency domain and taking the chunk mean in time domain. This approach helps to observe the randomness in the data and the amount of information in the higher frequency range.

LITERATURE SURVEY

A lot of methods have been introduced to detect an outlier. Sabyasachi Basu et al. [2] proposed the method of one-sided and two-sided median methods. But, the author proposes that this method is inaccurate when there are consecutive outliers spanning longer than window width as it is difficult to differentiate between same value and actual signal. Kaouther Nouira et al. [2] gave the method of graphical approaches and Gibb's sampling approach. There is no experimentation performed using these two methods but, it has been promised that this information may be used to improve outlier detection. Chris E. Zwilling et al. [4] used the information from covariance of time series data to detect an outlier. This method becomes efficient because it provides the user to choose any number and type of features and the algorithm will correctly identify the outlier. The author also proposed Multivariate Voronoi Outlier Detection (MVOD) [5] method. It is also proposed that this method is accurate, sensitive and robust in multivariate time series data. Hermine N. Akouemo et al. [6] adduced an autoregressive integrated moving average with exogenous inputs (ARIMAX) model. But, the ARIMAX model needs to be trained on cleaner data at each step. Hui-xin Tian et al. [7] came up with a new method that combines density-based clustering algorithm with soft sensor modeling process. But, this method directs to detect outlier of soft sensor modeling in complex industrial processes.

METHOD PROPOSED

A time series data is being analyzed in the frequency domain to detect an outlier. Step by step procedure is as follows:

[1]Corresponding author

1. A time series data can have null points at some time stamps. So, these null values are removed by replacing null values with the mean of adjacent points (mean of one value previous to the null value and other value next to the null value).
2. This data is divided into equal chunk values.
3. For each chunk, the magnitude of FFT and chunk mean is calculated.
4. The threshold is set for both, the magnitude of FFT of chunk and the mean of the chunk.
5. For the magnitude of FFT, value at first index is not taken into consideration. This is done to neglect any information in the lower frequency range.
6. If any of the thresholds is violated, then the chunk is considered to have an outlier.

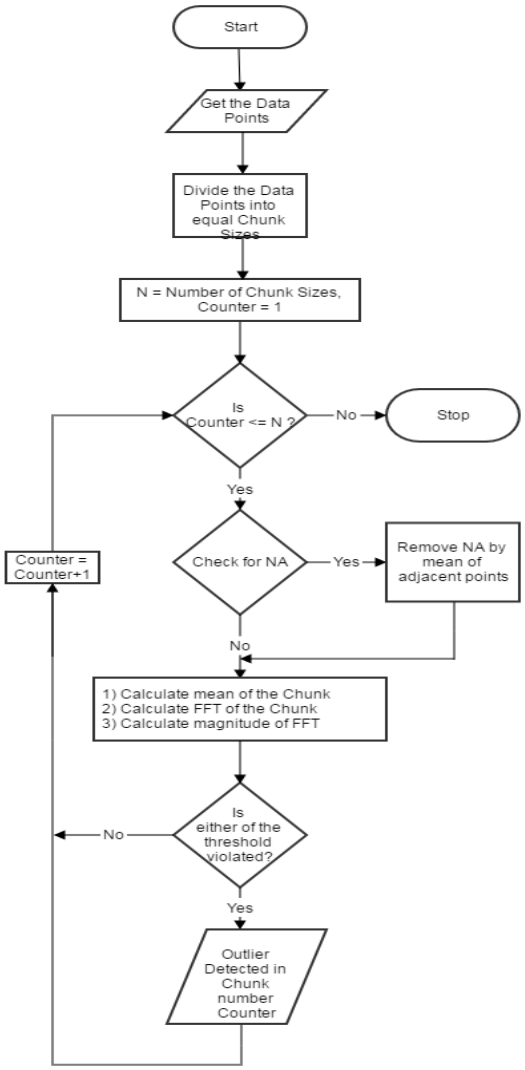

Figure 1: Algorithm of outlier detection

Figure 1 describes the whole procedure algorithmically.

EXPERIMENTATION AND RESULTS

Experimentation has been performed on time series data set as shown in Fig. 2:

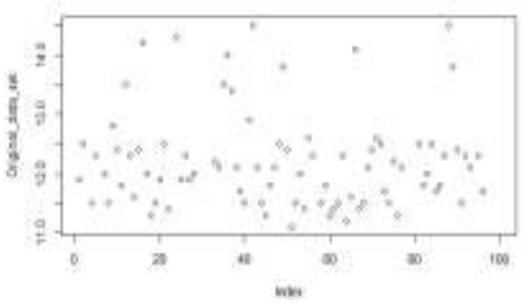

Figure 2: Original Data Set

This dataset also contains NA (Not Available) values i.e. there is no value present at following Index numbers: 3, 6, 23, 29, 30, 31, 32, 57, 64, 67, 78, 79, 80, 94, 97, 98, 99 and 100. These NA points are removed appropriately as described in section III. The revised set of data points obtained is shown in Fig. 3:

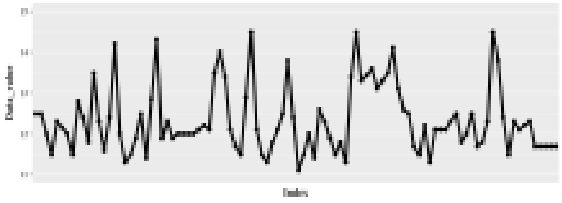

Figure 3: Data set after removing NA

This data set is divided into ten equal sized chunks and then each chunk is analyzed for detecting outlier. Chunk number 1, 6, 8 and 10 doesn't contain outlier as the magnitude of FFT performed on each chunk lies below the threshold value which is 2.5 here. Magnitude value at Index 1 is not taken into consideration for checking the threshold as it is believed that information would certainly lie in the lower frequency range. Fig. 4(a), 5(a), 6(a) and 7(a) are chunks 1, 6, 8 and 10, respectively. Fig. 4(b), 5(b), 6(b) and 7(b) are the plots of the magnitude of FFT of the chunks 1, 6, 8 and 10, respectively. Fig. 4(c), 5(c), 6(c) and 7(c) are the magnified view of the magnitude leaving the first Index of chunks 1, 6, 8 and 10, respectively.

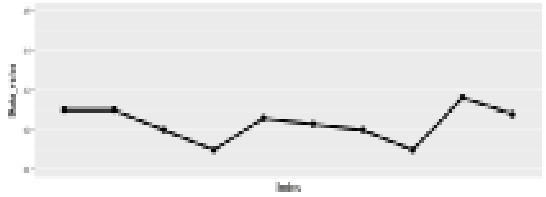

Figure 4(a): Chunk 1

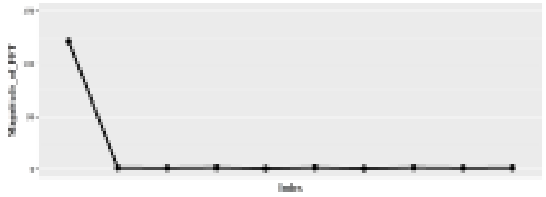

Figure 4(b): Magnitude of FFT of Chunk 1

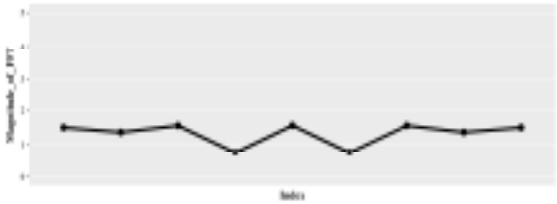

Figure 4(c): Magnified view of Fig. 4(b) after leaving the value at first Index

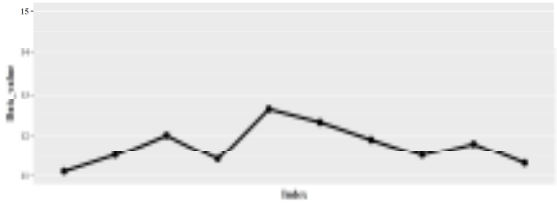

Figure 5(a): Chunk 6

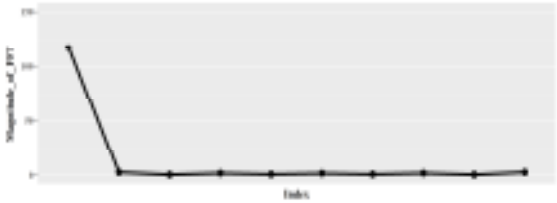

Figure 5(b): Magnitude of FFT of Chunk 6

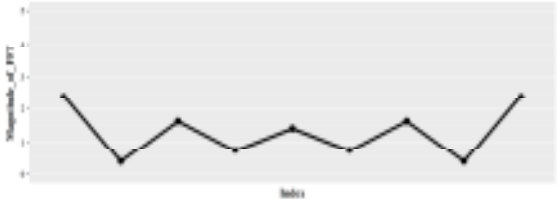

Figure 5(c): Magnified view of Fig. 5(b) after leaving the value at first Index

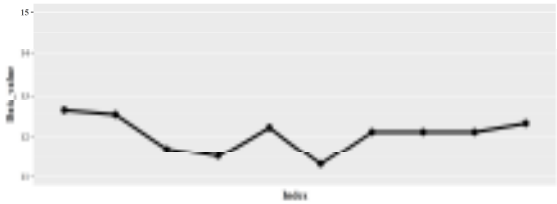

Figure 6(a): Chunk 8

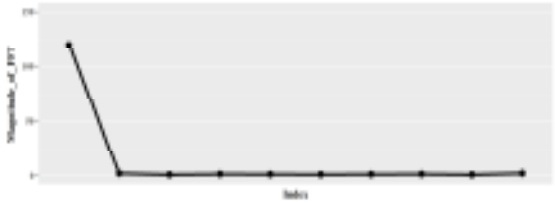

Figure 6(b): Magnitude of FFT of Chunk 8

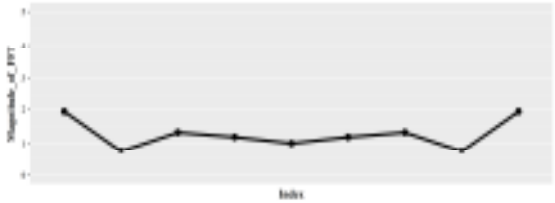

Figure 6(c): Magnified view of Fig. 6(b) after leaving the value at first Index

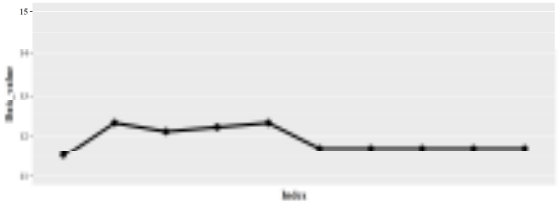

Figure 7(a): Chunk 10

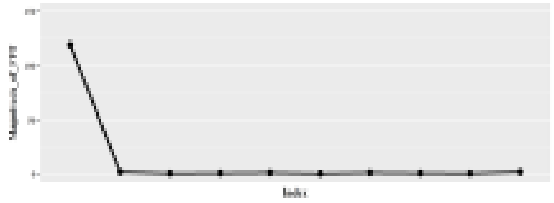

Figure 7(b): Magnitude of FFT of Chunk 10

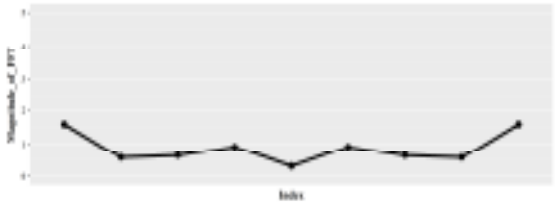

Figure 7(c): Magnified view of Fig. 7(b) after leaving the value at first Index

Chunk numbers 2, 3, 4, 5 and 9 contains outliers. Fig. 8(a), 9(a), 10(a), 11(a) and 12(a) are chunks 2, 3, 4, 5 and 9, respectively. Fig. 8(b), 9(b), 10(b), 11(b) and 12(b) are the plots of the magnitude of FFT of the chunks 2, 3, 4, 5 and 9, respectively. Fig. 8(c), 9(c), 10(c), 11(c) and 12(c) are the magnified view of the magnitude leaving the first Index of chunks 2, 3, 4, 5 and 9, respectively. From Fig. 8(c), 9(c), 10(c), 11(c) and 12(c), it can be seen that the threshold, which is 2.5 in this case, is being violated. Hence, it can be inferred that outlier lies in each case.

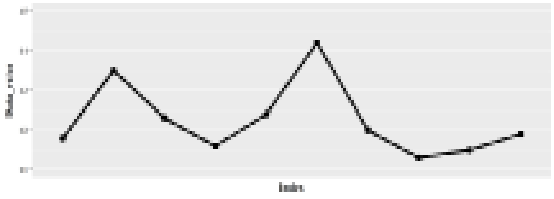

Figure 8(a): Chunk 2

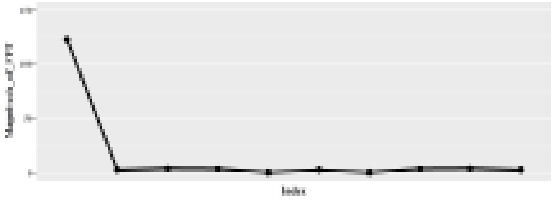

Figure 8(b): Magnitude of FFT of Chunk 2

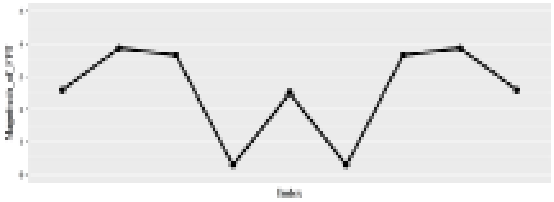

Figure 8(c): Magnified view of Fig. 8(b) after leaving the value at first Index

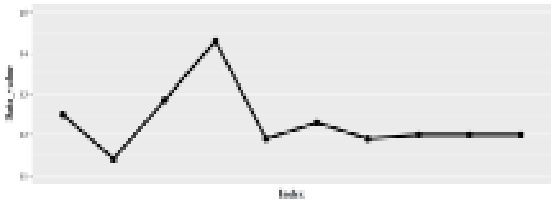

Figure 9(a): Chunk 3

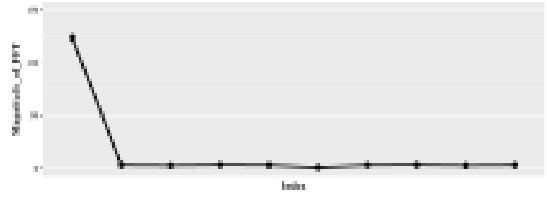

Figure 9(b): Magnitude of FFT of Chunk 3

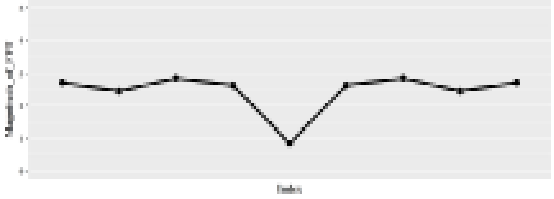

Figure 9(c): Magnified view of Fig. 9(b) after leaving the value at first Index

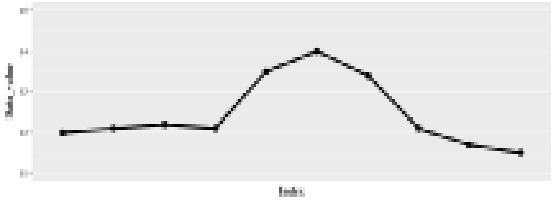

Figure 10(a): Chunk 4

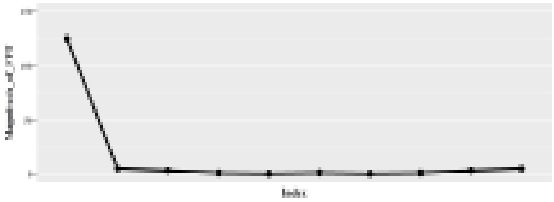

Figure 10(b): Magnitude of FFT of Chunk 4

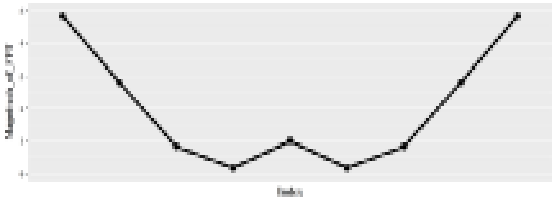

Figure 10(c): Magnified view of Fig. 10(b) after leaving the value at first Index

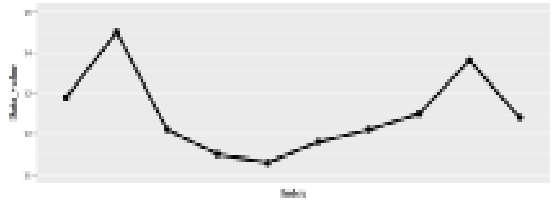

Figure 11(a): Chunk 5

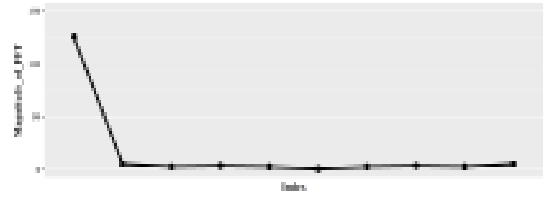

Figure 11(b): Magnitude of FFT of Chunk 5

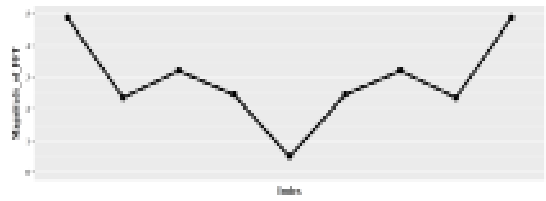

Figure 11(c): Magnified view of Fig. 11(b) after leaving the value at first Index

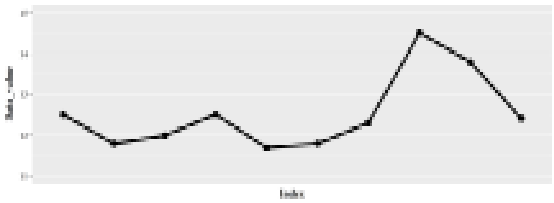

Figure 12(a): Chunk 9

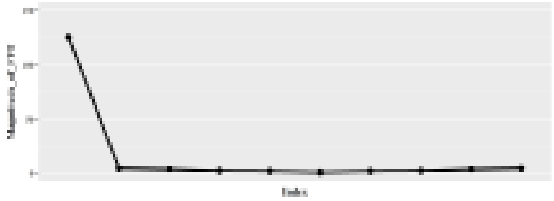

Figure 12(b): Magnitude of FFT of Chunk 9

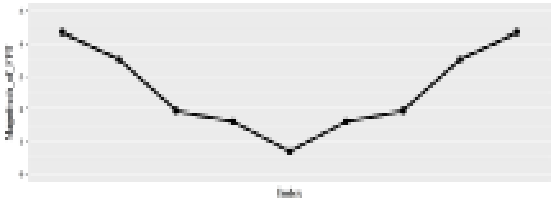

Figure 12(c): Magnified view of Fig. 12(b) after leaving the value at first Index

In chunk number 7, the threshold value (2.5) in the magnitude of FFT is not violated. But, the data value is well above the threshold value in the time domain (Threshold value in the time domain is 13). At such positions, chunk mean would be helpful to detect outlier. Chunk mean for chunk 7 is 13.535. Figure 13(a), 13(b) and 13(c) shows the data value in chunk 7, its magnitude plot and the magnified view of the magnitude plot respectively.

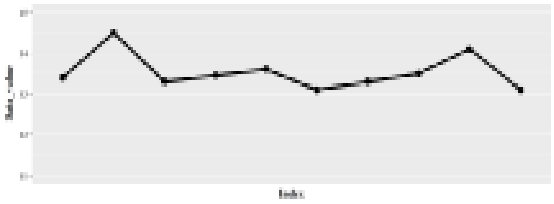

Figure 13(a): Chunk 7

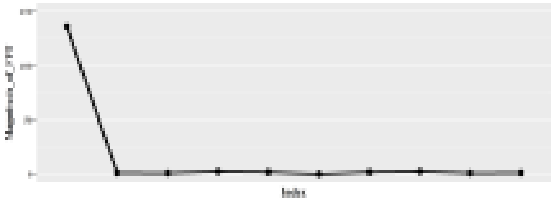

Figure 13(b): Magnitude of FFT of Chunk 7

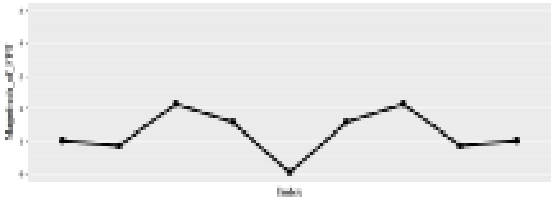

Figure 13(c): Magnified view of Fig. 13(b) after leaving the value at first Index

CONCLUSION

There are a lot of techniques for detecting outliers. But, a particular technique is able to detect a particular form of time-series data. Analyzing time-series data in frequency domain helps to detect outlier for any kind of data. Analyzing the data in frequency domain along with chunk mean improves the accuracy. In future, another technique along with frequency domain analysis can be developed to improve the accuracy for detecting the outlier. Also, the chunk size can be increased and improved equations for the threshold can be developed.

REFERENCES

Ashok, Asha, Smitha S. and Krishna M.H.K., 2016. "Attribute reduction based anomaly detection scheme by clustering dependent oversampling PCA." Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on, IEEE.

Prathibhamol C.P., Amala G.S. and Kapadia M., 2016. "Anomaly detection based multi label classification using Association Rule Mining (ADMLCAR)." Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on, IEEE.

Basu S. and Meckesheimer M., 2007. "Automatic outlier detection for time series: an application to sensor data." Knowledge and Information Systems, **11**(2):137-154.

Kaouther N. and Trabelsi A., 2006. "Time Series Analysis and Outlier Detection in Intensive Care Data." Signal Processing, 2006 8th International Conference on, Vol. **4**, IEEE.

Zwilling C.E. and Wang M.Y., 2016. "Covariance based outlier detection with feature selection." Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the, IEEE.

Zwilling C.E. and Wang M.Y., 2014. "Multivariate voronoi outlier detection for time series." Healthcare Innovation Conference (HIC), IEEE.

Hermine N.A. and Povinelli R.J., 2014. "Time series outlier detection and imputation." PES General Meeting| Conference & Exposition, IEEE.

Tian H.-X., Liu X.-J. and Han M., 2016. "An outliers detection method of time series data for soft sensor modeling." Control and Decision Conference (CCDC), 2016 Chinese. IEEE.