IECON2015-Yokohama
November 9-12, 2015

# Outliers Detection Method Using Clustering in Buildings Data

Usman Habib*, Gerhard Zucker*, Max Blöchle*, Florian Judex*, Jan Haase†
*Energy Department, AIT Austrian Institute of Technology, Vienna, Austria
Email: {usman.habib, gerhard.zucker, max.bloechle, florian.judex}@ait.ac.at
†Helmut-Schmidt-University Hamburg, Germany
Email: janhaase@ieee.org

***Abstract*— To achieve energy efficiency in buildings, a lot of raw data is recorded, during the operation of buildings. This recorded raw data is further used for the analysis of the performance of buildings and its different components e.g. Heating, Ventilation and Air-Conditioning (HVAC). To save time and energy it is required to ensure resilience of the data by detecting and replacing outliers (i.e. data samples that are not plausible) in the data before detailed analysis. This paper discusses the steps involved for detecting outliers in the data obtained from absorption chiller using their On/Off state information. It also proposes a method for automatic detection of On/Off and/or Missing Data status of the chiller. The technique uses two layer K-Means clustering for detecting On/Off as well as Missing Data state of the chiller. After automatic detection of the chiller On/Off cycle, a method for outlier detection is proposed using Z-Score normalization based on the On/Off cycle state of chillers and clustering outliers by Expectation Maximization clustering algorithm. Moreover, the results of filling the missing values with regression and linear interpolation for short and long periods are elaborated. All proposed methods are applied to real building data and the results are discussed.**

***Keywords—adsorption chillers; physical rules; K-Means Clustering Algorithm; Outliers; Z-Score Normalization; Expectation Maximization Clustering Algorithm (EM); Heating, Ventilation and Air-Conditioning(HVAC); Fault detection and diagnosis (FDD); Machine learning; regression; linear interpolation.***

## I. INTRODUCTION

Nowadays, a lot of raw data is recorded by the monitoring of different buildings using sensors. These sensors record some physical properties of a real life phenomenon, e.g. temperature, pressure, flow rate. There are a lot of external factors that may affect the measurements by the sensors, thus leading to various corruptions in the data. Some of the main factors are errors in instrumentation, changes in the environment (e.g. temperature), and human errors [1]. Hence, before processing the data to obtain some useful information, it is necessary to validate it. One important aspect of this validation is to detect the outliers that do make sense under the circumstances. The detection of outliers may well help getting rid of the data that is not expected/usual behavior of the machine.

The Outlier detection can be defined as "the problem of finding patterns in data that do not conform to expected normal behavior" [1]. The data is normally visualized using some graphic tool, then validated and analyzed by experts from the field. However, this method can be applied only to a limited amount of data [2]. It is also difficult to define the abnormal behavior due to the different types of malfunctioning of systems, making it unreliable to depend on the human judgment [3]. At the same time validation of the data is labor extensive and thus not feasible.

The aim of this research is to develop methods that can automatically find anomalies in data. As a proof the methods are applied to the operation data of an adsorption chiller. Since the behavior of a chiller can strongly vary in the two different states (On/Off) of the chiller. Therefore an outlier detection based on these states, which is also called as On/Off cycle has been proposed. In first step automatic detection of On/Off and Missing Data state, using two layer K-Means has been proposed. The Missing Data state is representing the duration when there is no data recorded for any sensor of the chiller. After detection of the On/Off state, the cycle based Z-Score normalization is used with expectation maximization clustering algorithm to find the outliers in the data. Furthermore, the filling of missing gaps for a sensor with regression and linear interpolation is also discussed.

The rest of the paper is arranged as follows. Section II takes a look at the existing relevant research. Section III describes the details of the proposed model. Section IV explains the methodology adapted for this research. In Section V, the proposed methods are applied to the real buildings data and the results are examined. Finally, Section VI concludes the results achieved along with giving a future outlook.

## II. STATE OF THE ART

The advancement in sensor technology and the wider spread of building automation systems has made it easy to record different physical real life phenomena such as temperature, pressure and electricity consumption of HVAC components in buildings. As these sensors are in the field, thus there is always a chance of receiving corrupted data. Some of the factors that add outliers to the data can be malicious activities, faults in the instruments, environmental features like temperature, and human errors [1].This lead to the need to validate the data before starting, detailed analysis in order to save energy, as the time spent on wrong data is of course lost. There are many methods available for data validation

This work was partly funded by the Austrian Funding Agency in the funding programme e!MISSION within the project "extrACT", project number 838688.

depending on the fields in which they are used [1], [4]. Some of the common methods for detecting outliers can be based on the status of sensor, physical range check, detection of gaps and constant values, tolerance band method, material redundancy detection, signal gradient test, extreme value check, physical and mathematical based models and data mining techniques [3], [5], [6]. The confidence parameter can be involved with data by using different labels e.g. data can be labeled as A, B and C on the basis of data validation techniques where A represents correct, B represents doubtful and C represents wrong data [6]. The confidence parameter can also be represented with adding a possibility of assigning a confidence value between 0 and 100, in order to provide refined information about data, where 0 is representing the minimum confidence and 100 represents the maximum confidence in data [7].

It is necessary to detect the missing gaps between the data acquired as it is also an important factor to indicate the reliability of the data. These missing gaps are always undeniable and lowering the amount of meaningful calculations. It is crucial to make the missing gaps fewer as data with fewer gaps is considered as good quality data. There are different methods available for handling the missing values e.g. regression [8]–[10], depending on the nature of the data and other parameters like computations, precision, robustness, accuracy etc. [11]. Moreover, it is also advised that a constant value issue can be diagnosed by analysis of the variance, as the sensors always have some small variation [7].

In order to detect unrealistic gradients, it is required to use a time continuity check. The requirement of the time continuity check gives the opportunity to define the maximum allowed difference between the measurements in a given interval of time. Thus helps in detecting the unrealistic gradients [12]

There are different diagnostic techniques available for detecting and diagnosing faults in HVAC (Heating, Ventilation and Air-Conditioning). The faults can be detected by using prior knowledge, keeping the main focus on the first hand principles. Similarly, there are other techniques available that are heavily dependent on the behavior or process of the system, which is usually captured from the historical data. Black box models come under the category of such kind of techniques [3], [13]. Moreover, the machine learning algorithms can be used to detect faults in buildings using the installed electricity consumption meters [14], [15]. Furthermore there are certain parameters that can help to predict the electricity consumption of each of the HVAC component. These parameters can be derived by using multivariate analysis [16].

For achieving energy efficiency in buildings, it is necessary to intelligently monitor and analyze the data of HVAC components. There are different machine learning algorithms available, which can be used for analysis purpose of buildings energy performance. To detect the state of machine for On/Off cycle, clustering algorithms can be used as data vary in these two different states. The X-Means clustering algorithm can be used to automatically detect the system states (On/Off) in order to examine the operational data of adsorption [17].

The method for finding the maximum likelihood estimates of the data distribution in the case when data is unknown or hidden (unsupervised), is called Expectation Maximization (EM).The EM clustering algorithm consists of finite Gaussian mixtures. It tries to estimate a set of parameters until the anticipated convergence value is attained. After convergence the mixture model has K probability distributions and each distribution represents a unique cluster. The maximum probability is used for assigning the membership of cluster to each instance of the data [18]. The EM clustering algorithm can also decide the number of clusters automatically by using the cross validation method [19].

There are methods available that can help to detect outliers with the help of machine learning techniques like clustering. Depending on the nature of data different types of clustering e.g. K-Means (distance based clustering) [20], [21], DBSCAN (density based clustering) [22] can be used for outlier detection [1], [4].

## III. System Design

### A. Adsorption Chillers

The data used for this paper is taken from a solar cooling system that uses an adsorption chiller as described in Figure 1, with naming conventions from Task 38 of IEA solar heating and cooling program [23].

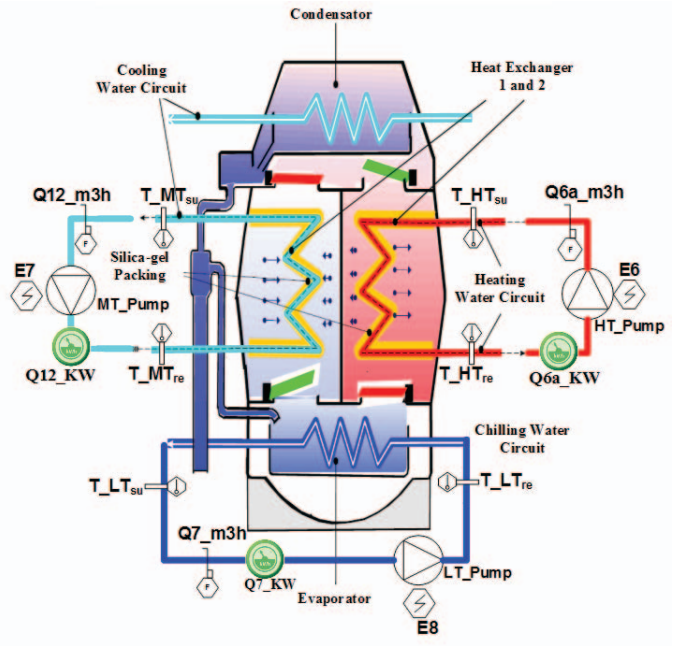


Figure 1: Adsorption chiller

The working of adsorption chiller can be defined in the following steps (for details, see [24])

1. The water is evaporated in the lower chamber called evaporator which makes the water cool in Low Temperature (LT) cycle.

2. The evaporated water is adsorbed on the receiver side, which is the middle chamber, using silica-gel.
3. The adsorbed water is then de-adsorbed with the heat provided from the Hot water cycle.
4. The de-adsorbed water is condensed and brought back to the evaporator.

### B. *Data Description*

This section describes the data used in this research. The description of the 15 parameters which were taken under consideration is given in Table I given below with naming convention taken from IEA Task 38 [23].

TABLE I : PARAMETERS DESCRIPTION

| Sensors | Description |
|---|---|
| E6 | Electricity consumption meter reading at high temperature cycle. |
| E7 | Electricity consumption meter reading at medium temperature cycle. |
| E8 | Electricity consumption meter reading at low temperature cycle. |
| Q6a_m3h | Flow of water readings in high temperature cycle. |
| Q12_m3h | Flow of water readings in medium temperature cycle. |
| Q7_m3h | Flow of water readings in low temperature cycle. |
| $T\_HT_{re}$ | Temperature reading at high temperature cycle on return side. |
| $T\_HT_{su}$ | Temperature reading at high temperature cycle on supply side. |
| $T\_MT_{re}$ | Temperature reading at medium temperature cycle on return side. |
| $T\_MT_{su}$ | Temperature reading at medium temperature cycle on supply side. |
| $T\_LT_{re}$ | Temperature reading at low temperature cycle on return side. |
| $T\_LT_{su}$ | Temperature reading at low temperature cycle on supply side. |
| Q6a_KW | Energy consumption reading of high temperature cycle |
| Q12_KW | Energy consumption reading of medium temperature cycle |
| Q7_KW | Energy consumption reading of low temperature cycle |

The data of each sensor is recorded at an interval of 4 minutes. The duration from 2010 to 2011 of the recorded data is considered for this research. The data availability for 2010 can be seen in Figure 2. It is clear from Figure 2 that the cooling operation recorded data is mainly available during the summer season every year, which is understandable. It is also important to note that even during the summer season; there are days where little or no recorded data is available due to communication line failures or some other unknown faults nature causing missing gaps in the data.

## IV. METHODOLOGY

In this section different methods that are used in this paper are discussed.

### A. *Using K-Means clustering to detect On/Off and Missing (Incomplete) Data status:*

Normally the solar chillers are in operational state during summers except in industries, where the chillers are On for the whole year. Due to communication disconnection and other invisible issues the data is not stored in the database thus causing data gaps. To fill these gaps we can either use historic gaps we can either use historic data for replacing it or other techniques, thus it will not add any information which can further help in diagnosis of faults. In order to detect the On/Off and missing data status automatically, it has been suggested to use K-Mean clustering algorithm in 2 levels.

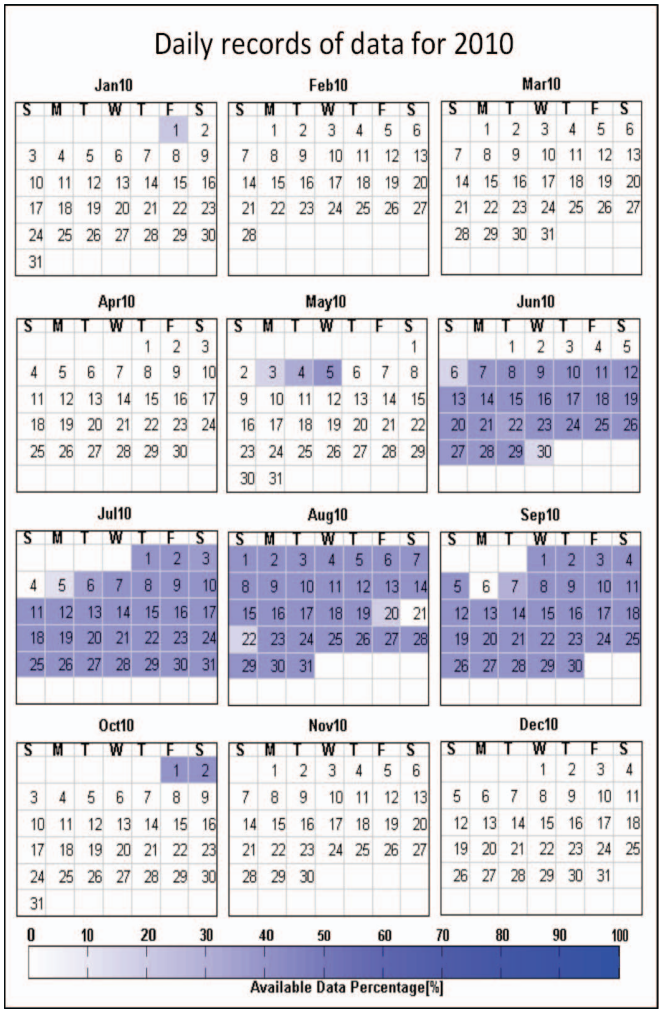


Figure 2: Data availability

Before using the data some pre-processing steps have been applied. The missing values are filled with 0 value as that will represent it, and min-max normalization is applied to scale the values while preserving the same relationship between the data values.

In order to detect the three different status of a chiller a two level of K-Means clustering technique has been proposed as seen in the Figure 3 below. The number of clusters is set to two and Euclidean distance is selected for finding the given number of clusters for both levels. At first level the K-Means decide whether the chiller is On or Off. Whereas, at the second level the data with Off status is forwarded to K-Means for predicting whether the information is enough to decide the status of the chiller is Off. Otherwise it will classify it as missing data (incomplete data). The benefit of adding the third status of incomplete data will be in analysis phase as it will give indication that the data at this stage was incomplete during the operation of the system and can be ignored.

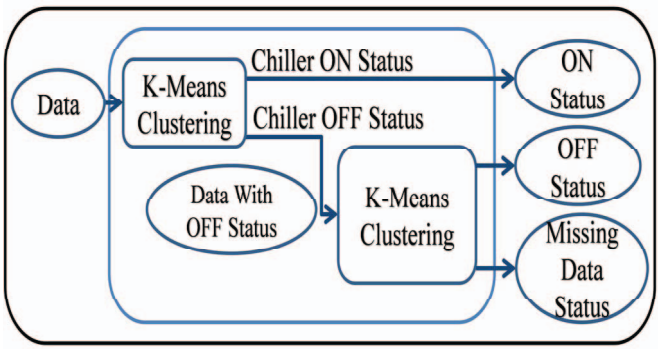


Figure 3 : Chiller On/Off status two level K-Means clustering algorithm

### B. *Outliers Detection Using Z-Score Normalization and EM Clustering*

In this section a method has been proposed for outlier's detection using the Expectation Maximization (EM) clustering algorithm on an On/Off cycle based on Z-Score normalized data.

The problem with taking the normal Z-Score is that it considers the mean and standard deviation of the whole data for normalization, which might not make the outliers away from the normal behavior, thus makes it difficult for clustering algorithm to identify the outliers. It is very difficult to find these kinds of faults with normal normalization process as these points need to be more highlighted than the normal behavior of the data in that specific machine cycle. The benefit of highlighting these points will help the clustering algorithm to cluster these points away from the normal behavior of the data. Thus help in finding the invalid data points in the data.

Furthermore, as we are dealing with the solar cooling system, where the behavior of the system will be different during the chiller's On and Off cycle. At the same time it may differ on a rainy day than the sunny day, therefore the mean of the sunny day will be higher than the rainy day as well the standard deviation. If we are considering the z-score for the whole data, there is a higher chance that we neglect or consider the correct data points as erroneous data points as the dominated data e.g. sunny days (high temperature) behavior will overcome the behavior of rainy days (low temperature).

To overcome this problem, it has been suggested in this research to use the Z-score as following in Equation 1

$$\mathbf{Z-Score_{cycle}} = \frac{(X-\mu_{cycle})}{\sigma_{cycle}} \qquad \text{Equation 1}$$

Where $Z-Score_{cycle}$ is the z-score for each Cycle when the system is either On or Off, X is the value of the sensor, $\mu_{cycle}$ is the population mean of each Cycle, and $\sigma_{cycle}$ is the standard deviation of each Cycle.

### C. *Steps for Outliers Detection:*

In order to sanitize the data the following steps are followed as can be seen in Figure 4

1. There will be three states of chiller states i.e. On (the chiller has cooling load), Off (the chiller does not have cooling load but data is recorded) and Missing data state (there is no data recorded for any sensor of the chiller).
2. Insert 0 values in all the missing gaps of sensors. This will help K-means clustering algorithm to gather all the missing data in a cluster, and that cluster can be entitled as zero cluster or missing data cluster.
3. Use 2 level of K-Means to get the three states i.e. On, Off and Missing data state.
4. Ignore the Missing data state data. The benefit of adding the third status of Missing data (incomplete) data state will be in analysis phase as it will give indication that the data at this stage is not available.
5. Normalize the data using Z-Score based on the mean and standard deviation of the each on the On/Off cycle of the chiller.
6. Find the outliers in the data using cycle based Z-Score with EM clustering algorithm.
7. Fill the missing gaps of each sensor with regression or linear interpolation.

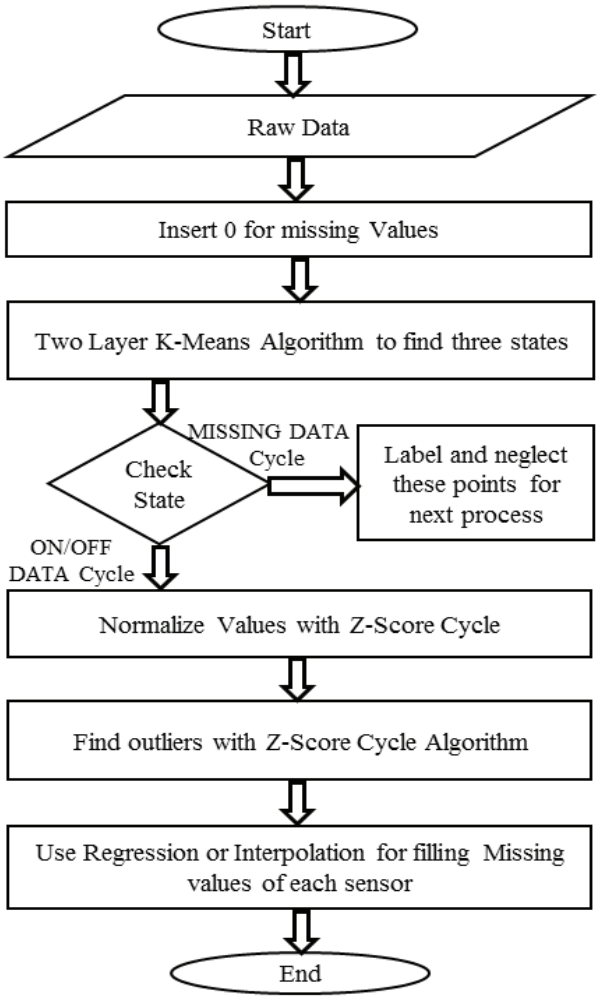


Figure 4 : Cycle based Z-Score outlier detection flow chart

## V. RESUTS AND ANALYSIS

This section discusses the different results that were applied to real buildings data.

### *A. On/Off and Missing Data STATE detection with two layer K-Means*

The results for first level of K-Means clustering can be seen in Figure 5 given below. The Figure 5 demonstrates the On/Off status through K-Means clustering algorithm. The red (dash and dot) line shows the On and Off status. It can be observed from the Figure 5 below the temperatures of the LT, MT and HT cycle are responding according to the detected On/Off state. The dotted green rectangle shows one On cycle of the chiller on 2010-07-08. It is clear from the Figure 5 that when the chiller is detected as On the LT temperatures decreases showing the cooling operation. It can be observed that the temperatures increase in the HT and MT cycle of the chiller. The decrease in temperatures inside LT cycle and increase in HT and MT cycle simultaneously, at the same time is a clear signal that the chiller is in operational mode, which has also been detected by the proposed method of K-Means clustering.

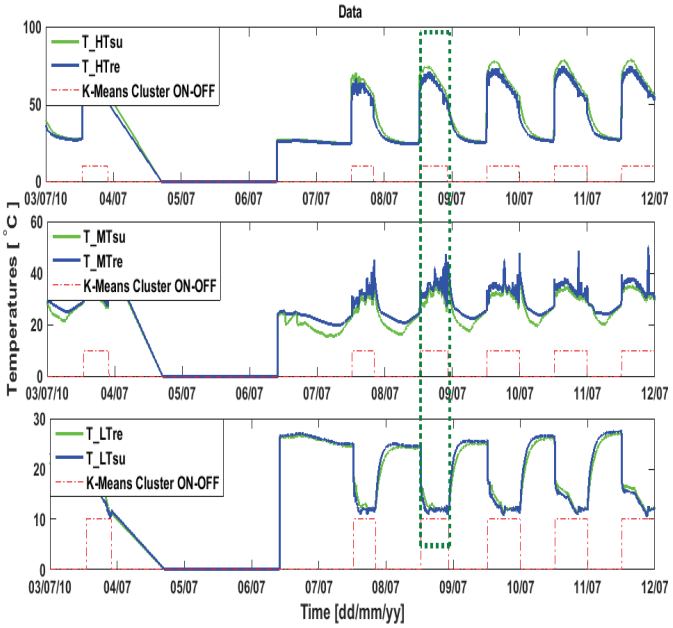


Figure 5 : Chiller On/Off status with two level K-Means clustering algorithm.

Furthermore, Figure 6 below shows the results for the second level of K-Means Algorithm applied. The magenta (dash and dot) line on zero value is denoting where the data is missing, which is also emphasized with red dotted rectangle. The zero value of the missing data status line is representing the data as classified as available data status. The Figure 6 shows that there is no data recorded for some small gap on 2010-06-28 in midnight, and for the whole day on 2010-06-30. The missing data state will help in analysis of performance of the chiller as we can focus on available data only.

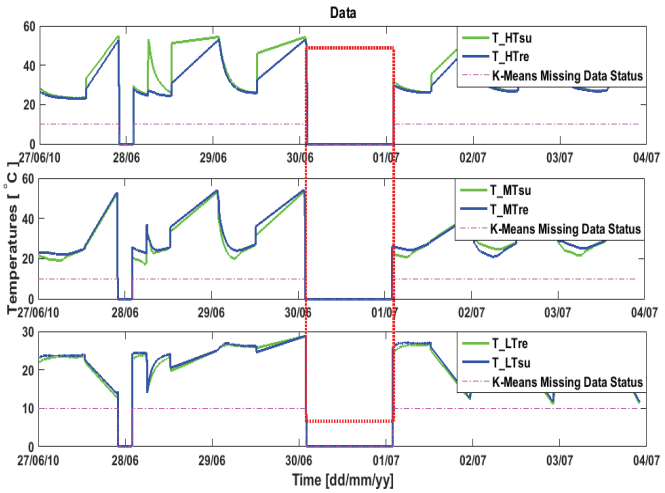


Figure 6 : Chiller missing data status two level K-Means clustering algorithm.

### *B. Outliers Detection Using EM Clustering*

In order to detect outliers in the data the duty cycle based Z-Score is applied as described in Methodology Section. The statistical data for the z-scores (μ and σ) are derived based on single On or Off cycle instead of the whole data. Figure 7 shows the original data (T_LTre) along with the cycle based Z-Score; the detected outliers using expectation maximization clustering (EM) algorithm are marked with red circles both in the corresponding original data (T_LTre) and cycle based z-scored data. These unintended outliers are, however, representing only single samples that can be restored by using linear interpolation for gap filling as described in the next section; furthermore the change of specific On or Off state based Z-Score parameters allows better performance for detection of outliers, as the behavior of a process varies strongly in the two different states.

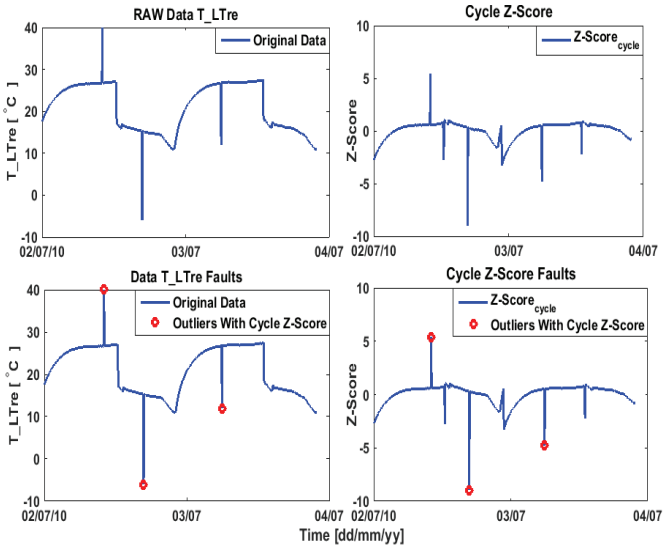


Figure 7 : Outliers detection using cycle Z-Score with EM clustering algorithm.

### C. *Comparsion between outliers detected with Z-Score and Cycle based Z-Score.*

Furthermore, the outliers have been detected with clustering using normal Z-Score and cycle based Z-Score. The normal Z-Score is based on the mean and standard deviation for the whole available data set as compared to cycle based Z-Score where Z-Score is based on each On or Off cycle of the chiller. The results can be seen in Figure 8 given below. The graph has data with outliers detected with both methods represented with red circles. The first graph in Figure 8 which represents data T_LTre shows outlier's detection with normal Z-Score whereas the second graph in the Figure 8, that represents data T_LTre shows outliers detected with cycle based Z-Score. It can be seen clearly from the Figure 8 that normal Z-Score has detected outliers that were away from mean of the whole data and neglected outlier on 2010-07-03 at 6:00 o'clock. Whereas, cycle based Z-Score has detected the other outliers along with the neglected outlier, as the cycle based Z-Score makes the outlier detected on 2010-07-03 at 6:00 away from mean according to the mean and standard deviation of that specific cycle. This helps the clustering algorithm to consider it as outlier.

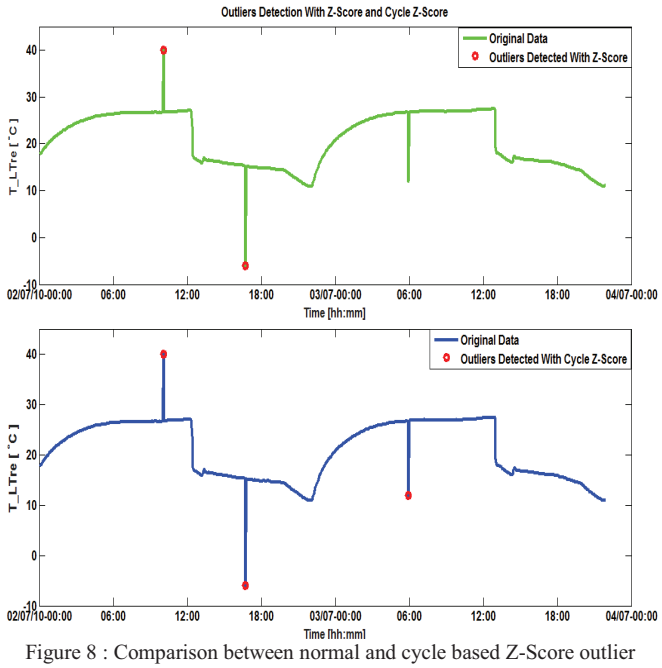


Figure 8 : Comparison between normal and cycle based Z-Score outlier detection.

### D. *Filling Missing Values Gaps*

Monitoring samples are deleted randomly to test the algorithms for filling gaps. Then the missing gaps have been filled with regression and interpolation. The Figure 9 consists of four graphs i.e. the original data, data with missing values, missing gap filled with regression and missing gap filled with linear interpolation. The regression method as the blue circle denotes shows that there are some small peaks added to the missing gaps whereas interpolation has produced a smooth data in the case where there are non-consecutive random missing values for short period of time as can be seen in Figure 9.

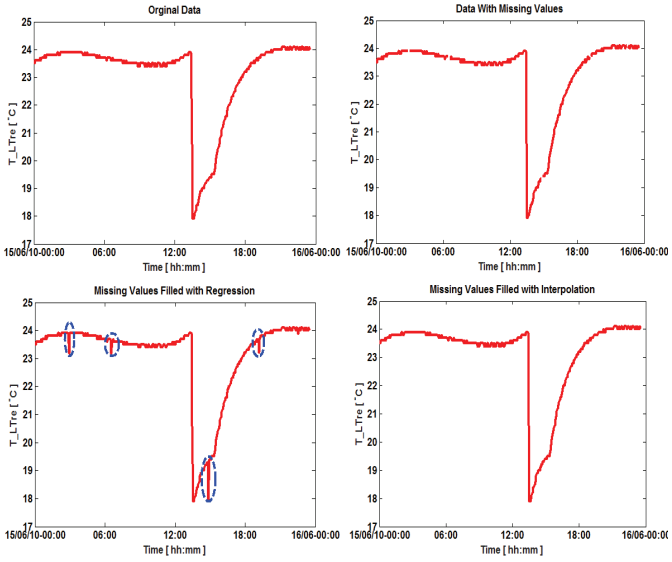


Figure 9 : Filling of short period gaps.

In the case where there are long missing gaps in the data the regression method has shown better results as compared to interpolation which can be noticed in the Figure 10 given below.

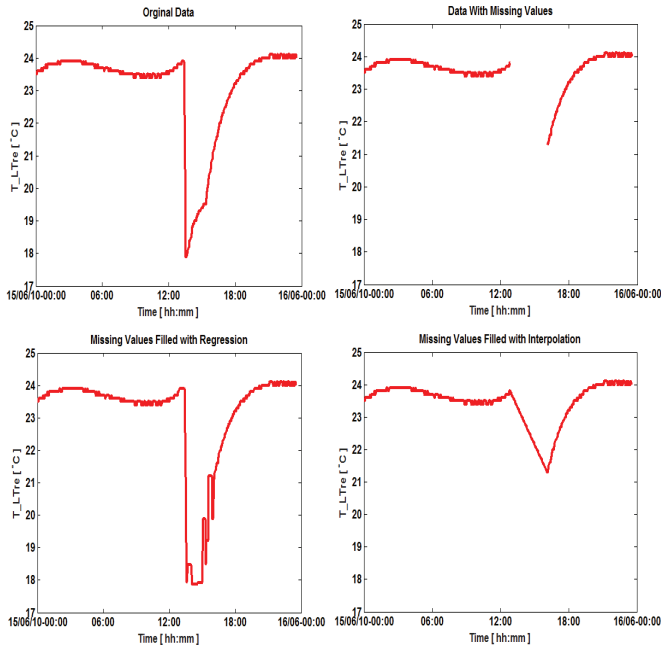


Figure 10 : Filling of longer period gaps.

## VI. CONCLUSION AND FUTURE WORK

In this paper we have used the data collected from an adsorption chiller for outlier's detection. The automatic detection of On/Off state of a chiller has been proposed by employing a two layered K-Means clustering algorithm. The two layered K-Means algorithm gives 3 states of a chiller i.e. On/Off and Missing Data state. The Missing Data state is representing the duration when there is no data recorded for any sensor of the chiller

After having the On/Off cycle information, the Expectation Maximization Clustering Algorithm is used on Z-Score normalized data based on each cycle for outlier detection. Furthermore, the filling of the missing gaps is carried out with the help of interpolation and regression.

There is more space of improvement in the results of handling missing data in the data. Therefore, it is required to further investigate and find better methods for handling the missing gaps in the data. As in this paper we have mainly focused on the first level of fault detection and diagnosis (FDD) of the HVAC components of the building. Therefore, after eliminating the outliers in the data, it can be further used for a detailed analysis of faults in the operation of chiller. Additionally, more research is required to find new methods for automatic detection and diagnosis of faults in the operation state of HVAC components.

## ACKNOWLEDGEMENT

For implementation of the proposed methods in this paper, the authors are thankful to the teams of Konstanz Information Miner (KNIME) [25] and WEKA [26] software tools.

## REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Comput Surv, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.

[2] M. Mourad and J. Bertrand-Krajewski, "A method for automatic validation of long time series of data in urban hydrology," Water Sci. Technol., vol. 45, no. 4–5, pp. 263–270, Mar. 2002.

[3] S. Katipamula and M. R. Brambley, "Review Article: Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part I," HVACR Res., vol. 11, no. 1, pp. 3–25, 2005.

[4] V. J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies," Artif. Intell. Rev., vol. 22, no. 2, pp. 85–126, Oct. 2004.

[5] S. Sun, J. Bertrand-krajewski, A. Lynggaard-Jensen, J. van den Broeke, F. Edthofer, M. do C. Almeida, À. S. Ribeiro, and J. Menaia, "Literature review for data validation methods," PREPARED 2011.019, 2011.

[6] N. Branisavljević, Z. Kapelan, and D. Prodanović, "Improved real-time data anomaly detection using context classification," J. Hydroinformatics, vol. 13, no. 3, p. 307, Jul. 2011.

[7] G Olsson, M Nielsen, Z Yuan, A Lynggaard-Jensen, and J-P Steyer, "Instrumentation, Control and Automation in Wastewater Systems," Scientific and Technical Report 15, May 2005.

[8] E. Frank and R. R. Bouckaert, "Conditional Density Estimation with Class Probability Estimators," in Advances in Machine Learning, Z.-H. Zhou and T. Washio, Eds. Springer Berlin Heidelberg, 2009, pp. 65–81.

[9] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Third Edition, 3 edition. Burlington, MA: Morgan Kaufmann, 2011.

[10] L. Torgo and J. Gama, "Search-based Class Discretization," in In: Proceedings of the Ninth European Conference on Machine Learning, 1997, pp. 266–273.

[11] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz, "Comparison of missing value imputation methods in time series: the case of Turkish meteorological data," Theor. Appl. Climatol., vol. 112, no. 1–2, pp. 143–167, Jul. 2012.

[12] U.S. Department of Commerce, "Handbook of Automated Data Quality Control Checks and Procedures," National Data Buoy Center, Mississippi 39529-6000, NDBC Technical Document 09-02, Aug. 2009.

[13] S. Katipamula and M. R. Brambley, "Review Article: Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part II," HVACR Res., vol. 11, no. 2, pp. 169–187, Apr. 2005.

[14] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," IEEE Trans. Power Syst., vol. 20, no. 2, pp. 596–602, May 2005.

[15] M. Domínguez, J. J. Fuertes, S. Alonso, M. A. Prada, A. Morán, and P. Barrientos, "Power monitoring system for university buildings: Architecture and advanced analysis tools," Energy Build., vol. 59, pp. 152–160, Apr. 2013.

[16] N. Djuric and V. Novakovic, "Identifying important variables of energy use in low energy office building by using multivariate analysis," Energy Build., vol. 45, pp. 91–98, Feb. 2012.

[17] G. Zucker, J. Malinao, U. Habib, T. Leber, A. Preisler, and F. Judex, "Improving energy efficiency of buildings using data mining technologies," in 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE), 2014, pp. 2664–2669.

[18] X. Jin and J. Han, "Expectation Maximization Clustering," in Encyclopedia of Machine Learning, C. Sammut and G. I. Webb, Eds. Springer US, 2011, pp. 382–383.

[19] P. Smyth, "Model selection for probabilistic clustering using cross-validated likelihood," Stat. Comput., vol. 10, no. 1, pp. 63–72, Jan. 2000.

[20] R. Pamula, J. K. Deka, and S. Nandi, "An Outlier Detection Method Based on Clustering," in 2011 Second International Conference on Emerging Applications of Information Technology (EAIT), 2011, pp. 253–256.

[21] M. F. Jiang, S. S. Tseng, and C. M. Su, "Two-phase clustering process for outliers detection," Pattern Recognit. Lett., vol. 22, no. 6–7, pp. 691–700, May 2001.

[22] M. Celik, F. Dadaser-Celik, and A. S. Dokuz, "Anomaly detection in temperature data using DBSCAN algorithm," in 2011 International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 2011, pp. 91–95.

[23] A. Napolitano, W. Sparber, A. Thür, P. Finocchiaro, and B. Nocke, "Monitoring Procedure for Solar Cooling Systems," International Energy Agency, IEA Task 38, Oct. 2011.

[24] GBU mbH, "Technical Description Adsorption Chiller Nak," Jan. 1999.

[25] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), 2007.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explor Newsl, vol. 11, no. 1, pp. 10–18, Nov. 2009.