Journal of Building Engineering 41 (2021) 102751

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

A data driven performance assessment strategy for centralized chiller systems using data mining techniques and domain knowledge

Muhammad Bilal Awan [a], Kehua Li [a], Zhixiong Li [b], Zhenjun Ma [a,*]

[a] Sustainable Buildings Research Centre, University of Wollongong, 2522, Australia
[b] Yonsei Frontier Lab, Yonsei University, Seoul, 03722, Republic of Korea

ARTICLE INFO

Keywords:
Chillers
Data mining
Performance assessment
Conditional inference tree
Association rule mining

ABSTRACT

Chillers are among the major energy consumers in building heating, ventilation and air conditioning systems and appropriate performance assessment of chiller systems is essential to ensuring their operational optimality while delivering satisfactory indoor thermal comfort. This paper presents a data driven performance assessment strategy for centralized chiller systems using multiple data mining and advanced visualization techniques. The energy consumption patterns of the chiller system were quantitatively and qualitatively analyzed by using the Conditional Inference Tree (CIT) and Agglomerative Hierarchical Clustering (AHC), and Association Rule Mining (ARM), respectively. A performance indicator of Coefficient of Performance (COP) Destruction (%) was introduced to represent the quality of the achieved COP. The performance of this strategy was evaluated using one-year operating data of a centralized chiller system installed in a commercial building. The results showed that the data mining techniques can be effectively used for performance assessment of chiller systems. The results from the quantitative and qualitative analysis showed that the chiller performance was strongly influenced by the temperature difference across the evaporator. The system studied generally showed good performance when the part load ratio was above 45% and the chiller power ratio was above 50%, and it showed relatively poor performance when the temperature difference across the evaporator was below 3.1 °C.

1. Introduction

Buildings and their related construction consume one-third of the global energy and generate nearly 40% of the total CO_2 emissions [1]. According to the Department of the Environment and Energy, Australia, HVAC systems are the major contributor to building energy consumption [2]. Energy consumption of HVAC systems is considered sensitive to operational, maintenance, environmental, and load conditions [3]. Ji et al. [4] reported that HVAC performance improvement is crucial for minimizing building energy consumption and operational cost, and a successful performance assessment by using fault detection and diagnosis (FDD) can save 10–40% of HVAC energy.

Performance assessment is considered as a process of measuring, estimating, and verifying the performance and energy consumption of a system at actual load and design conditions [3]. In current practice, the performance assessment of building energy systems is often achieved through fault detection and diagnosis (FDD) and energy benchmarking [4]. FDD methods can be classified into model-based, hardware-based, and history-based approaches, and are used to discover the errors in physical systems while attempting to identify the source of the problem [5]. Benchmarking techniques are often used to identify and implement the best practices that lead to exceptional performance [6]. COP, integrated part load value (IPLV), kW/ton, and energy efficiency ratio (EER) or European seasonal energy efficiency ratio (ESEER) are often used as performance indicators. Yu et al. [7] reviewed the standard performance indicators of chillers in 9 countries, and it was found that IPLV and ESEER cannot truly reflect the part-load operation of chiller systems. It was recommended that HVAC performance standards should be further developed to allow operators and designers to analyze the performance under actual operating conditions with more transparency and details.

With the improvement in sensing and data collection technologies [8], large sets of building energy usage data are now readily available. Data mining techniques could extract useful information from this time-series data and provide additional insights on the performance improvement of HVAC systems. Data mining methods have been extensively used for building energy profiling and benchmarking. Li et al. [9], for instance, developed a data driven methodology to benchmark and evaluate the electricity usage patterns of various buildings. Cluster analysis was used to group similar behaving buildings

* Corresponding author.
E-mail address: zhenjun@uow.edu.au (Z. Ma).

https://doi.org/10.1016/j.jobe.2021.102751
Received 11 November 2020; Received in revised form 27 March 2021; Accepted 16 April 2021
Available online 24 May 2021

Nomenclature

DBSCAN	Density-based spatial clustering of applications with noise
CIT	conditional inference tree
PLR	part load ratio
AHC	agglomerative hierarchical clustering
COP	coefficient of performance
COPD	COP Destruction
ARM	association rule mining
AHU	air handling unit
FCU	fan coil unit
C	central zone
P	perimeter zone
W	west zone
E	east zone
BG	building B ground floor
K	kitchen
R	reception
F	fitness centre
$\dot{m}$	mass flow rate
T	temperature
ε	specified radius
CPR	chiller power ratio
CHWP	chilled water pump
CWP	condenser water pump
CTF	cooling tower fan
DoW	day of week
TDE	water temperature difference across evaporator
TDC	water temperature difference across condenser

Subscripts

e	evaporator
c	condenser
p	pump
i	inlet
o	outlet

based on their yearly electricity usage patterns. Multivariate adaptive regression splines (MARS) was used to describe the complex non-linear relationships between the explanatory variables and building electricity usage per square meter. Similarly, Panapakidis et al. [10] developed a data-driven model to investigate the electricity consumption patterns of a building. The performance of various algorithms including Fuzzy C-means, *k*-means ++, self-organized map, and minimum variance criterion with 8, 12, and 16 clusters, was evaluated. Other data-driven models such as Partitioning around medoids (PAM) [11], Symbolic aggregate approximation (SAX) [12], Agglomerative hierarchical clustering (AHC) [13], and Gaussian mixture model clustering (GMMC) [14], have also been used for building energy performance assessment.

Although data mining techniques have been widely used for building energy profiling, the use of data mining techniques to assess the operational performance of building HVAC systems is still in its early stage. Li and Ju [15] assessed the operating parameters of a chiller system using hierarchical clustering. The analysis was performed based on the simulated datasets. It was concluded that the clustering technology offered fast performance assessment of the chiller system. In a series of studies presented by Yu and Chan [16–21], the performance of chiller systems was evaluated using various statistical and unsupervised learning techniques. In Ref. [16], the relationships between the chiller system COP and operating variables were established for each cluster by using regression correlation and Pearson correlation. However, the established relationships had a low R^2 value. In Ref. [17], energy conservation opportunities in the chiller system were identified by using the data envelopment analysis (DEA) based benchmarking. It was concluded that the high performance of the chiller system can be achieved via fine tuning of the operating variables. Under the perfect control conditions, the system COP could be enhanced from 3.87 to 4.56. In Ref. [18], a correlation was built between the chiller system COP and the external (climate) variables by using the multivariate analysis. Part load ratio (PLR), and temperature differences at the evaporator and condenser were identified as the most significant variables. Three system effectiveness-based parameters, i.e. overall efficiency, scale efficiency and technical efficiency, were calculated. Scale efficiency was linked with the climate variables, whereas technical efficiency was independent of the climate variables. The results showed that fine tuning of the operating variables can achieve electricity savings of around 5.34%. In Ref. [21], cluster analysis was performed to assess the performance of the chiller system and it was concluded that clustering is an efficient and rapid choice to assess the performance of the chiller system.

These studies prove the effectiveness of data-driven methods for the performance assessment of HVAC systems. COP was mostly considered as the performance assessment indicator. However, COP alone may not be able to fully reflect the operational performance of chiller systems as it cannot determine how far the system performs from the respective ideal performance. Therefore, another performance indicator, named as COP destruction (COPD), was introduced in this study. Furthermore, these studies were mainly focused on investigating the quantitative effect of the variables on the performance of the chiller system without developing a rational qualitative model, and also lacked the use of advanced visualization techniques for identifying the distinctive energy usage patterns of the chiller system. This paper aims to develop a more reliable data-driven performance assessment strategy for centralized chiller systems using data mining techniques and domain knowledge. Energy usage patterns of the chiller system were visualized by using calendar view heat maps. CIT model was used to temporally classify the chiller power ratio (CPR), which was defined as the ratio of chiller operating power to the chiller rated power. AHC and CIT techniques were used to provide a quantitative assessment of the effects of selected operating variables on the system performance, whereas an ARM model was used to provide a detailed qualitative analysis about the collective effect of selected operating variables on the system performance. The proposed strategy was tested and evaluated using the one-year operational data of an HVAC system implemented in a commercial building.

2. Methodology

2.1. Outline of the proposed strategy

The outline of the proposed strategy is shown in Fig. 1. This strategy consists of four steps, which are data collection, data cleaning, energy profiling of the chiller system and data analysis. All steps in the proposed strategy are performed by incorporating the domain knowledge to provide better insights into the data analysis. The performance of chiller systems is often affected by the operating parameters such as inlet and outlet temperatures of the chilled water across the chiller evaporator, inlet and outlet temperatures of the cooling water across the chiller condenser, mass flow rates of the chilled water and cooling water, and the power consumption of chillers, air handling units, water pumps, and cooling towers. A substantial amount of the operational data of the chiller system should be first collected, which are generally available from the building management system (BMS). DBSCAN algorithm is then applied to the collected data to detect and remove the outliers from the raw data, which is further explained in Section 2.2. In the third step, the energy profiling of the chiller system is carried out to identify the distinctive patterns in the operation of the chiller system. In this study,

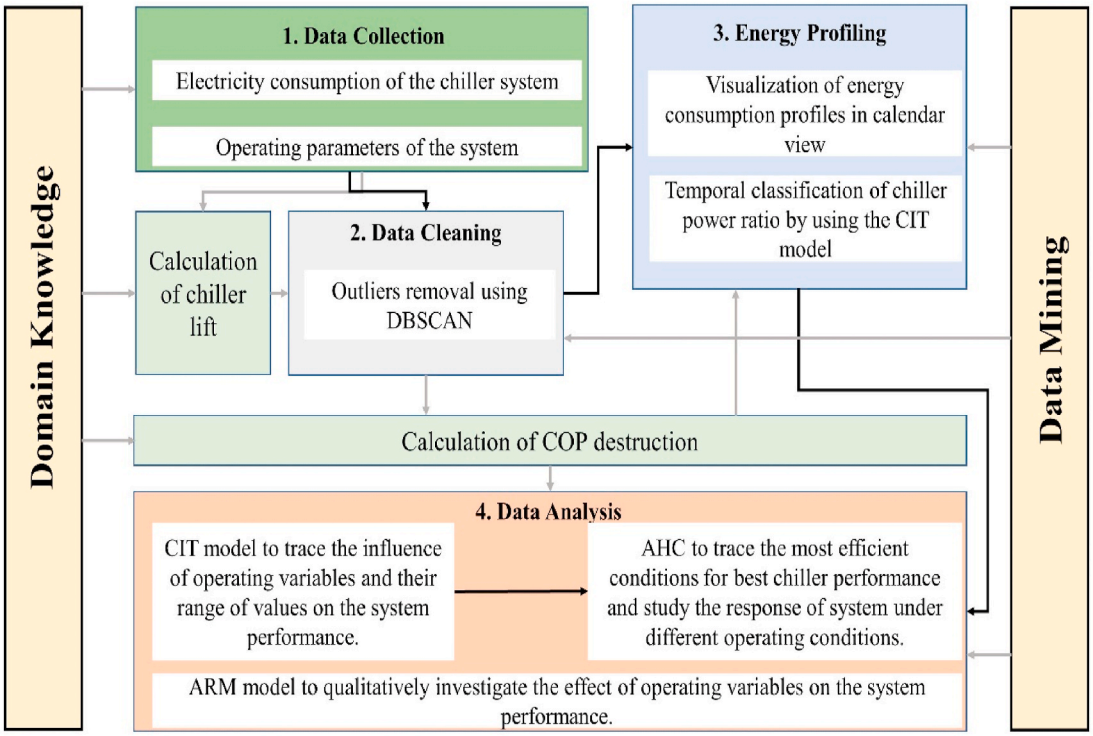


Fig. 1. Outline of the strategy.

the energy consumption patterns of the chiller system were first visualized in a calendar view heat map for preliminary analysis. CIT was then used for CPR classification based on the temporal variations. The last step is the data analysis that involves qualitative and quantitative analysis. In this step, the processed data were first quantitatively analyzed using the AHC and CIT model. Quantitative analysis established the relationships between the selected chiller operating variables and chiller performance in terms of the coefficient of performance (COP) and COP Destruction (COPD), respectively. Secondly, the ARM model was applied to qualitatively analyze the relationships among the chiller performance indicators and selected operating variables. Association rules were established to demonstrate the collective effect of the variables on the system performance. It is worthwhile to note that the method developed in this study could be used for performance assessment of any chiller system with limited changes while different domain knowledge may be required depending on the complexity of the chiller system and potential operating issues involved.

In the study, the coefficient of performance (COP) of the system is considered as the ratio of the heat removed from the chilled water across the chiller evaporator and the total electricity input of the chiller compressors, and chilled water and cooling water pumps that are dedicated to the chillers. The COP can be calculated using Eq. (1).

$$COP = \frac{\dot{m}_e \times c_P \times (T_{i,e} - T_{o,e})}{E_c + E_p}\# \quad (1)$$

where $\dot{m}$ stands for mass flow rate, T and E stand for temperature and input power respectively, the subscripts e, c, p, i and o stand for evaporator water loop, compressor, pump, inlet, and outlet respectively, and c_P represents specific heat of the water.

To enhance the accuracy of the chiller performance evaluation, COP Destruction (COPD), as expressed in Eq. (2), was introduced as a new performance indicator. It represents the quality of COP and determines how far the system performs from the respective ideal performance i.e. Carnot COP. For instance, 50% COP Destruction shows that ideally, the system has the capacity to improve its performance by 50%. This indicator in combination with COP could provide more details to identify the effectiveness of a system as compared to the use of COP alone. Higher values of COPD indicate an inefficient operation. Negative values of COPD and the values close to 100% may be resulted in some cases which were due to sudden change of the operating conditions and these values in principle should be discarded.

$$COPD = \frac{\left[\left(\frac{T_{o,e}}{T_{i,c} - T_{o,e}}\right) - COP\right]}{\left(\frac{T_{o,e}}{T_{i,c} - T_{o,e}}\right)} \times 100\% \quad (2)$$

2.2. Data cleaning using DBSCAN

Data cleaning, as a substantial part of the data processing, can exclude the outliers from the collected data before further analysis in order to enhance overall performance assessment. In this study, the outliers were defined as the observations in which the CPR and chiller lift (i.e. the temperature difference between the cooling water leaving the condenser and chilled water leaving the evaporator) were considerably different from most of the other observations. DBSCAN algorithm was used in this study for outlier detection. As a density-based algorithm, DBSCAN can identify the data points which have a low density of neighbors as the outliers [22]. Compared to other commonly used statistics-based outlier detection methods (e.g. generalized extreme studentized deviate test and interquartile range test), DBSCAN performs better if the data do not obey the specific probability distributions [23]. To conduct the outlier detection using DBSCAN, the density of each data point in the dataset was first estimated by counting the number of neighbors around the estimated data points in a specified radius (ε). The data points were then classified into three types based on the specified radius (ε) and the threshold of the density (k). The three types of data points include core points, border points and noise points (outliers). A core point is a data point whose density exceeds the threshold (i.e. k). A border point is a data point whose density is lower than k while the distance between it and a core point is smaller than ε. A noise point is a data point which is neither a core point nor a border point and is

identified as an outlier.

In this study, CPR and chiller lift data were first rescaled to 0 mean and 1 deviation. All the data points were then plotted on a two-dimensional scatter plot. Each point in the scatter plot of DBSCAN stands for an observation at a specific time interval. The value of ε was chosen by using a *k*-distance graph [22]. After the estimation of the density for each data point, the noise points were identified and removed from the dataset for further analysis.

2.3. Conditional Inference Tree (CIT)

Decision tree is a well-known classification technique which can generate a tree-like model consisting of a set of if-then rules. Compared to the widely known decision tree generation algorithms such as Classification and Regression Tree (CART), the main advantage of CIT is that the bias in the classification result can be avoided due to significant tests in the variable selection process [24]. Additionally, the CIT algorithm has the advantage of controlling the size of the tree that avoids the processes of cross-validation and tree pruning [24]. The first step in the development of a decision tree using the CIT model is to test the global null hypothesis of independence between the explanatory variables and response variables. If explanatory variables have no significant influence on the response variables, the process will be terminated, i.e. null hypothesis will not be rejected. Otherwise, a partial null hypothesis will be used to test the independence between each response variable and the explanatory variables. Explanatory variables with the strongest association with the response variable are then added to the model. A binary split in the selected explanatory variable is then implemented in the next step. The goodness of a split is evaluated using Eq. (3) [24].

$$A_{opt} = \underset{A}{\text{argmax}} c\left(x_j^A, \partial_j^A, \vartheta_j^A\right)\# \tag{3}$$

where x^A is a metric to measure the incongruity among the two split samples, ∂^A represents the conditional expectation and ϑ^A represents the covariance.

In this study, a decision tree was used to identify the chiller operating routine, and to model the relationships between the system performance in terms of COP and COPD, and multiple operating parameters. CIT was employed for tree generation that was dependent on the temporal variations and system variables. CIT helped identify the temporal variations-based operational behavior of the chiller system. Specific months, days of week and hours were detected, when the system was operating above a specific CPR. Such assessment helps identify the building energy consumption patterns that further lead to improve the performance of the system. In the data analysis step, CIT also provided a detailed insight into the performance variation of the chiller system based on the different operating conditions.

2.4. Agglomerative hierarchical clustering (AHC)

Ward's method based AHC algorithm was used to analyze the performance of the chiller based on the operating conditions. AHC is a bottom-up clustering strategy. In this study, a total of 6 variables were analyzed including the water temperature differences across evaporator and condenser, CPR, chiller PLR and chiller performance in terms of COP and COPD. All these 6 variables at a specific time were termed as a single observation. Initially, each observation was treated as a separate cluster, the observations (clusters) close to each other were then identified and merged into larger clusters. The process continued until all clusters were merged into a single cluster. The merging process is represented using a dendrogram (i.e. a tree-like structure). The dendrogram can then assist in determining the optimal number of clusters that can be achieved by cutting the dendrogram at a specific height. Dissimilarity measure is an important component in a hierarchical clustering algorithm. In this study, Euclidean distance (ED) was used as the dissimilarity measure as it has the advantage of measuring the dissimilarity in terms of magnitude.

2.5. Association rule mining (ARM)

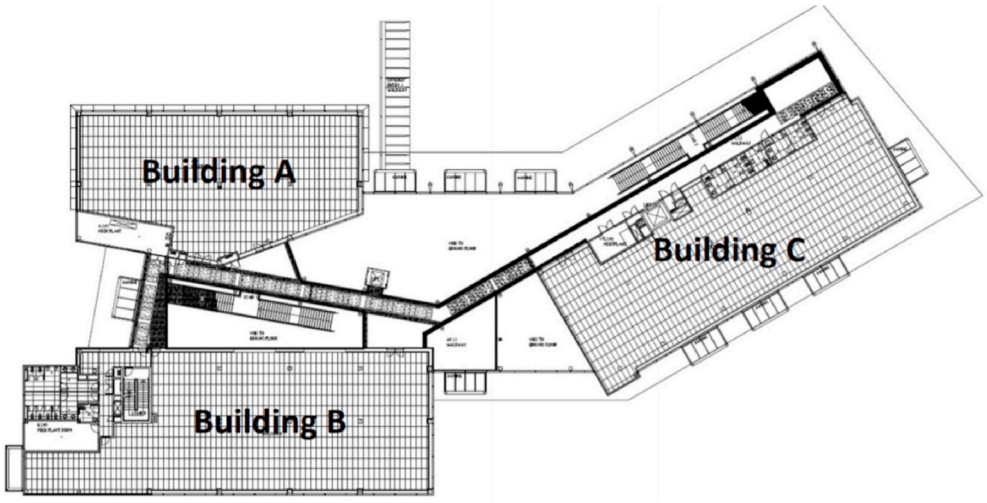
ARM discovers the relationships in large datasets [25]. It has been effectively used in the fields of bioinformatics [26], medical diagnosis [27] and marketing [28] and has also been proved to be effective in analyzing building operational data [29]. The representative setup of an association rule is like A→B, where A and B represent the separate sets of items or events. For example, a rule {high PLR of the HVAC system}→{efficient performance of the HVAC system} advocates that a relationship exists between the variation in the PLR and the performance of the system. Two indicators, i.e. support and confidence, are used to measure the strength of an association rule. Support represents how frequently a rule is applicable within the constraints of a dataset, whereas confidence determines how frequently A will appear if B appears. Apriori method was used in this study to discover the association rules. The Apriori algorithm consists of two steps, including frequent item set generation and rule generation. In the frequent item set generation step, all the item sets that satisfy the threshold of support are identified. In the rule generation step, all the high-confidence rules are extracted from the frequent item sets identified in the previous step. In this study, the thresholds of support and confidence were set as 0.2 and 0.8 respectively, so that the identified high-frequency association rules are within a controllable number for further analysis and interpretation using domain knowledge.

3. Performance test and evaluation of the developed strategy

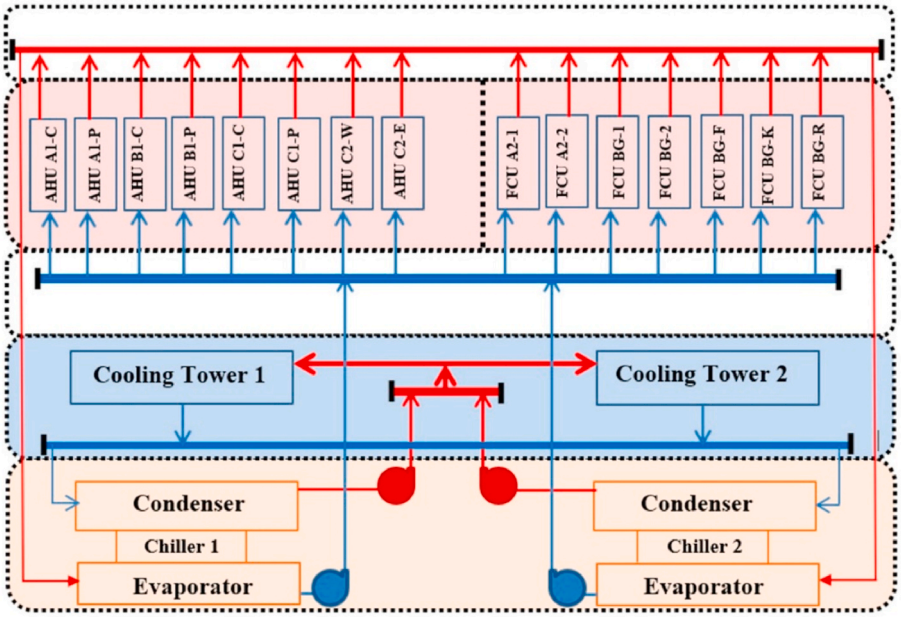
3.1. Description of the case study chiller system

In this study, a centralized chiller system (Fig. 2) installed in a multi-functional university building cluster was used to test and evaluate the performance of the developed strategy. The building cluster consists of three independent buildings (A, B and C) with three floors each (Fig. 2a). The total floor areas of buildings A, B and C are 1,420 m^2, 2,265 m^2 and 2,190 m^2, respectively. AHU-A1-P, AHU-B1-P and AHU-C1-P are used to serve the perimeter zones. AHU-A1-C, AHU-B1-C and AHU-C1-C are used to serve the central zones in the ground level and Level 1 of building A, Levels 1 & 2 of building B and Level 1 of building C, respectively (Fig. 2b). AHU-C2-E and AHU-C2-W are used to condition the east and west zones of level 2 of building C, respectively. The fitness centre in building B, kitchen in building C and the reception are conditioned by the fan coil units FCU-F, FCU-K and FCU-R, respectively. FCU A2-1 and FCU A2-2 are used to condition Level 2 of building A, and FCU BG-1 and FCU BG-2 serve the ground floor of building B. The ground floor of building C is conditioned by the packaged units. The building occupancy pattern is nearly identical for most of the year, as the building concerned mainly consists of offices, cafeteria, and gym that are operational almost all the year. Two identical scroll chillers with a cooling capacity of 220 kW each were used. When the cooling load was less than 220 kW, one chiller was in operation. Otherwise, two chillers were in operation. R407c is used as the refrigerant. The system is equipped with two chilled water pumps with a rated power of 2.7 kW each and two cooling water pumps with a rated power of 3.8 kW each. The supply air is conditioned through eight air handling units (AHUs) and seven fan coil units (FCUs) (Fig. 2b). There are two cooling towers used for heat rejection and each has a heat removal capacity of 357 kW and an input power of 2.2 kW. Variable speed drive (VSD) pumps and fans are employed in the system to enhance energy performance. Chilled water and cooling water pumps were operated based on the operation of their dedicated chillers. Both cooling towers were put into operation if any chiller is operating.

The operation data of the chiller system were collected from the Building Management System (BMS) and the main data were recorded every 30 min.



a) Building layout.



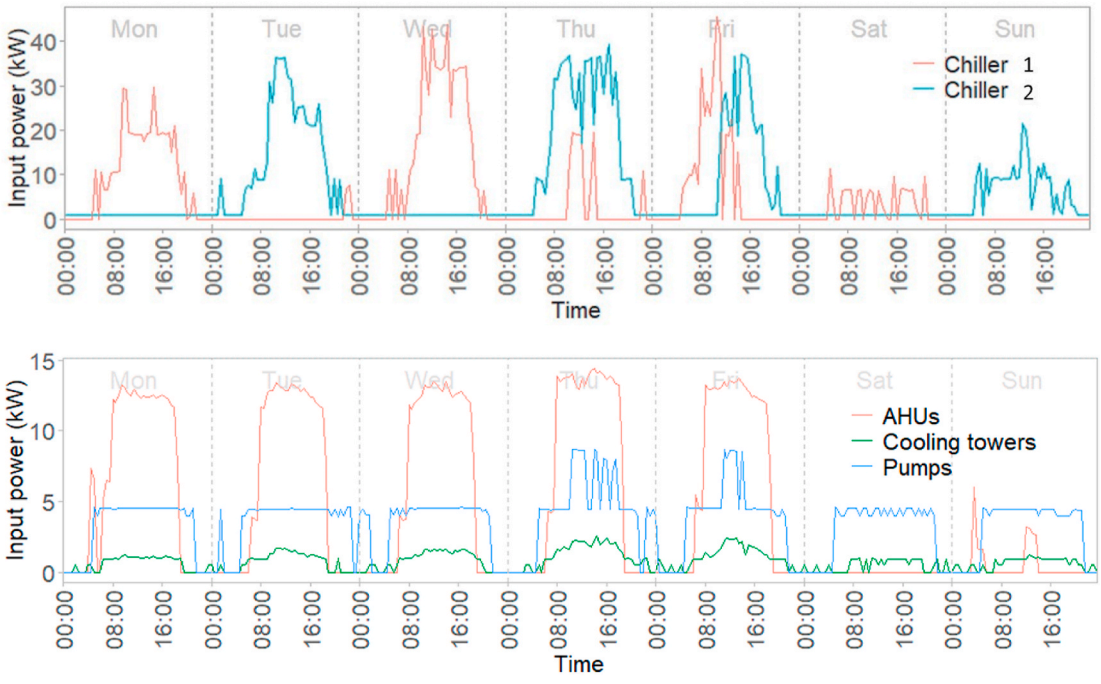
b) Line diagram of the chiller system.

Fig. 2. Illustration of the centralized chiller system.

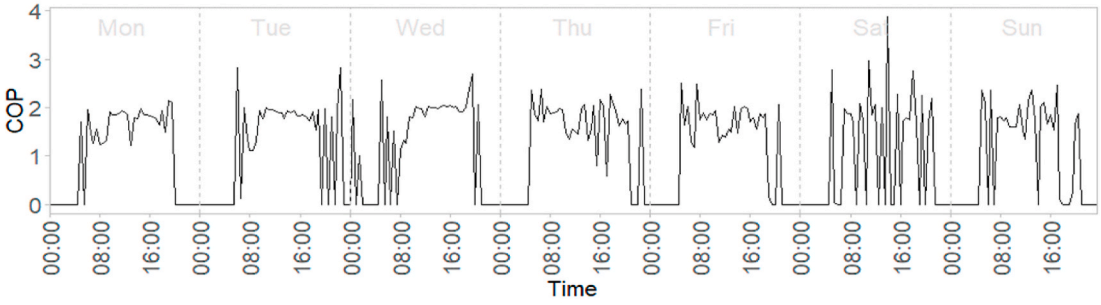
The operating variables of the chiller system, including the inlet and outlet temperatures of the chilled water across the chiller evaporator, inlet and outlet temperatures of the cooling water across the chiller condenser, mass flow rates of the chilled water and cooling water, power consumption of chillers, air handling units, water pumps, and cooling towers were collected from the Building Management System. The data were collected from November 2017 to October 2018 with ½ hour intervals and a total of 17,520 observations were used in this study. Fig. 3a) shows an example of the retrieved data in a week from December 4 to December 10, 2017. It can be seen that the highest power consumption of the system generally occurred from 9:00 to 16:00 during weekdays. The power consumption of the AHUs remained similar irrespective of the chiller operation. The power consumption of the water pumps was around 4.8 kW when one chiller was operating while it was increased to 8.0 kW when both chillers were operating. Although the pumps are equipped with variable speed drives, they were mostly operated at a constant speed. The COP of the chiller system was then calculated using Eq. (1) and is illustrated in Fig. 3b.

3.2. Data pre-processing

DBSCAN was first used for data cleaning. As introduced in Section 2.2, the parameter ε in the DBSCAN was determined using a k-distance graph (Fig. 4a). A k-distance graph showed the dissimilarity from each data point to its kth nearest neighbor in the increasing order and a sharp variation at the value of the dissimilarity corresponded to a suitable value of ε (i.e. 0.075 in this study). The outliers identified via the DBSCAN are illustrated in Fig. 4b with the blue dots. The observations that corresponded to these outliers were removed in the following analysis. After the data cleaning, 17,314 observations were retained.



a) Power consumption of different components.



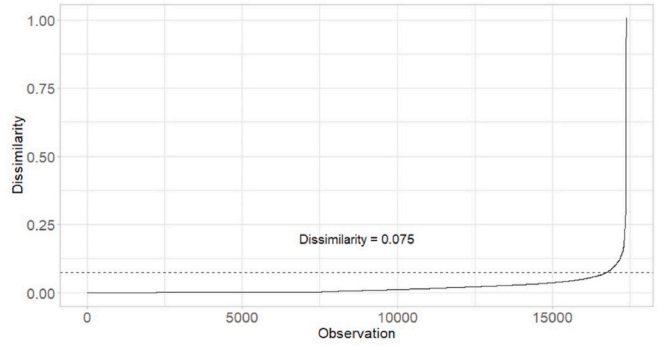
b) COP of the system.

Fig. 3. Illustration of the retrieved and calculated variables in a week.

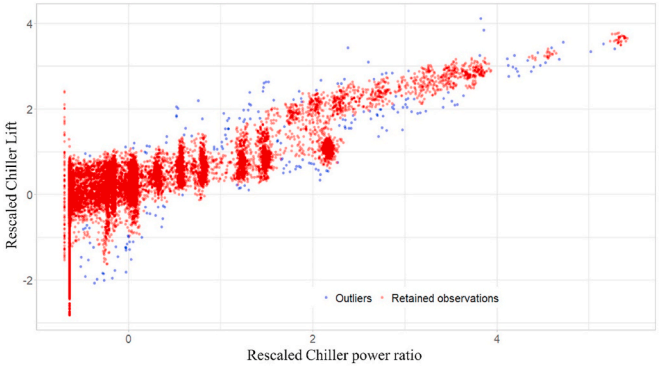
3.3. *Energy profiling of the chiller system*

Fig. 5 shows the daily power ratio of each component in the chiller system for 12 consecutive months. All these components are variable speed drives. The dark grey blocks in Fig. 5 indicated the days with no data recorded or the data with outliers which have been removed from the raw data, whereas the blue shaded blocks showed that each component in the chiller system was operated below 50% of its rated power and the dark yellow shaded blocks showed each component in the chiller system operated above its 50% rated power. In the city where the case study system is located, the cooling demand generally starts rising in October and remains high till April next year. Therefore, the chillers were frequently used during these months for space cooling. Fig. 5a and b indicate the power usage patterns of the two chillers for 12 consecutive months, from November to October of the next year. A comparison between both figures indicated that the chillers were rarely operated at their full rated power. Moreover, the cooling demand below 220 kW was met by operating one chiller only. For instance, only chiller 1 was operating at near 75% of its rated power on 30th October (Fig. 5a and b). It can be observed from Fig. 5d that the cooling tower operation was directly linked with the chiller operation and the power consumption patterns of the condenser water pumps and chilled water pumps were nearly identical throughout the year. As shown in Fig. 5c, AHUs were generally in operation in the working days throughout the year. It is noted that the power consumption of AHU-C2-E and AHU-C2-W was not included in Fig. 5, because these two AHUs are constant volume and their separate power consumption data were not available. In the cold season, building heating was achieved by using a boiler system. As Wollongong, Australia, has mild weather and sometimes cooling is still needed during winter months, the chillers were therefore also operated during the cold season when needed.

The energy profiles of the chiller system were further explored by using the CIT model. The CIT models illustrated the variations in the CPR with respect to the months, days of a week (DoW) and hours. The input data used for the CIT model generation consisted of the response variables and explanatory variables. The response variable used was whether the CPR of the chiller exceeds its 25%. The explanatory variables included the month, DoW and hour. The generated CIT model is presented in Fig. 6, in which the green bars indicated the percentage of the data observations with the CPR equal to or higher than 25%, whereas the red bar showed the CPR below 25%. The chiller system was most frequently operated from 8:00 to 18:00 during weekdays of the last



a) *k*-distance diagram.



b) Result of the outlier detection.

Fig. 4. Outlier detection using DBSCAN.

three and the first four months of the next year. The presence of the green bars in the winter months showed that the chillers were also occasionally operated in the winter months. Clear segregation of the operating conditions achieved from this model can be used to develop optimal control strategies for chiller systems as per temporal variations in the load demand.

3.4. Quantitative analysis

To increase the reliability of the analysis, only observations recorded during the office hours of the cooling months were further quantitatively analyzed by using the AHC and CIT. As the system performance is affected by multiple parameters concurrently, both AHC and CIT were adopted to examine the effect of four variables (i.e. CPR, chiller PLR, and water temperature differences across the evaporator and condenser) on the two performance indicators of COP and COPD.

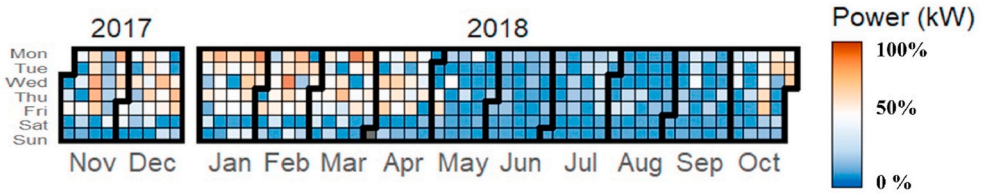
Fig. 7 and Fig. 8 illustrate the effect of the CPR, chiller PLR, and the temperature differences across the evaporator and condenser on the system performance in terms of COP and COPD, respectively. COP and COPD datasets were divided into three categories with equal magnitude. Top, middle, and low 33.3% values were termed as high, medium, and low COP and COPD, respectively. Top 33.3% COP and bottom 33.3% COPD presented the best performance, i.e. the chiller can remove more heat at a relatively lower power input as compared to that under other conditions. Low COP and high COPD indicated poor or low chiller performance. The results from Fig. 7 showed that a high COP could be resulted when the temperature difference across the evaporator was above 5.1 °C at low and medium CPR conditions. At some low CPR conditions, the evaporator temperature difference between 3.1 °C and 5.1 °C also resulted in a high COP. However, lower chiller performance was resulted when the water temperature difference across the evaporator was below 3.1 °C at some low and medium CPR conditions, and that across the evaporator was below 4.1 °C at some medium CPR conditions. At high CPR conditions, the temperature difference at the evaporator below 5.1 °C generally resulted in low system performance, whereas the best performance was achieved when the temperature difference across the evaporator was above 6.4 °C and that across the condenser was above 6.0 °C. In contrast, the water temperature difference across the condenser below 6.0 °C and that across the evaporator below 6.4 °C generally resulted in low performance at high CPR conditions. Overall, the results from Fig. 7 indicated that the system COP was most strongly influenced by the temperature difference across the evaporator among the variables considered.

The effect of the operating parameters on COPD (Fig. 8) showed that the system performed worst when the temperature difference across the evaporator was below 3.1 °C. At low CPR conditions, the combination of the water temperature difference across the evaporator above 3.6 °C and that across the condenser below 4.4 °C provided better performance. At high CPR conditions, the best performance was achieved when the temperature difference across the condenser was above 6.3 °C and that across the evaporator was above 6.5 °C. At low and medium CPR conditions, the system showed high performance when the temperature difference across the condenser was above 4.4 °C and that across the evaporator was above 4.8 °C. The above analysis showed that the temperature difference across the evaporator had the strongest relationship with COPD.

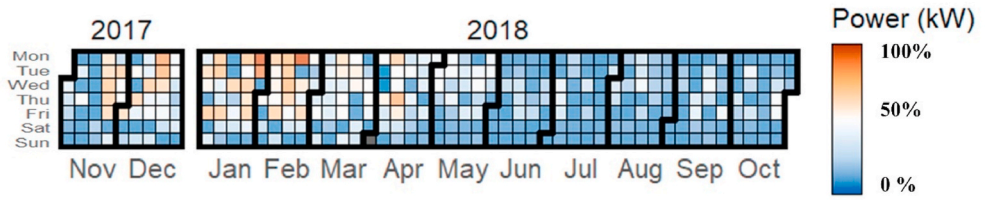
A comparison between Figs. 7 and 8 showed that, to achieve good performance, the system could be operated at high CPR conditions. However, high performance can also be achieved at some low and medium CPR conditions. For instance, the system showed better performance (i.e. high COP and low COPD) when the water temperature difference across the evaporator was above 3.6 °C at some low CPR conditions and above 4.8 °C at some medium CPR conditions. Similarly, for some high CPR conditions, the temperature difference across the evaporator above 6.4 °C resulted in better performance. High COP at low CPR conditions frequently resulted in high COPD. Moreover, the water temperature difference across the evaporator below 3.1 °C generally resulted in low performance. The quantitative numbers from the CIT model can be used as the constraint conditions for performance optimization and facilitating the development of optimal control strategies for this chiller system.

Although CIT model is computationally fast to identify the optimum operating conditions, CIT can just establish links for the most strongly inter-linked variables and ignores the weak connections with the increase in the number of variables. For instance, CIT was unable to establish a link between the chiller PLR and system performance by considering the operational variables. Hence, to further evaluate the interconnection between the chiller operating variables and performance variables, hierarchical clustering was used. Fig. 9a shows the dendrogram of all six observations, including four variables and two performance indicators, which were all merged into a single observation. Every single observation was identified as an individual cluster. Seven main clusters were formed and identified, which were labeled with different colors. The sub-clusters in the same cluster had similar trends but were different from those in the other clusters. Fig. 9b shows the relationship between four variables and two performance indicators with 0 mean and 1 standard deviation. The mean and median values of each variable are summarized in Table 1.

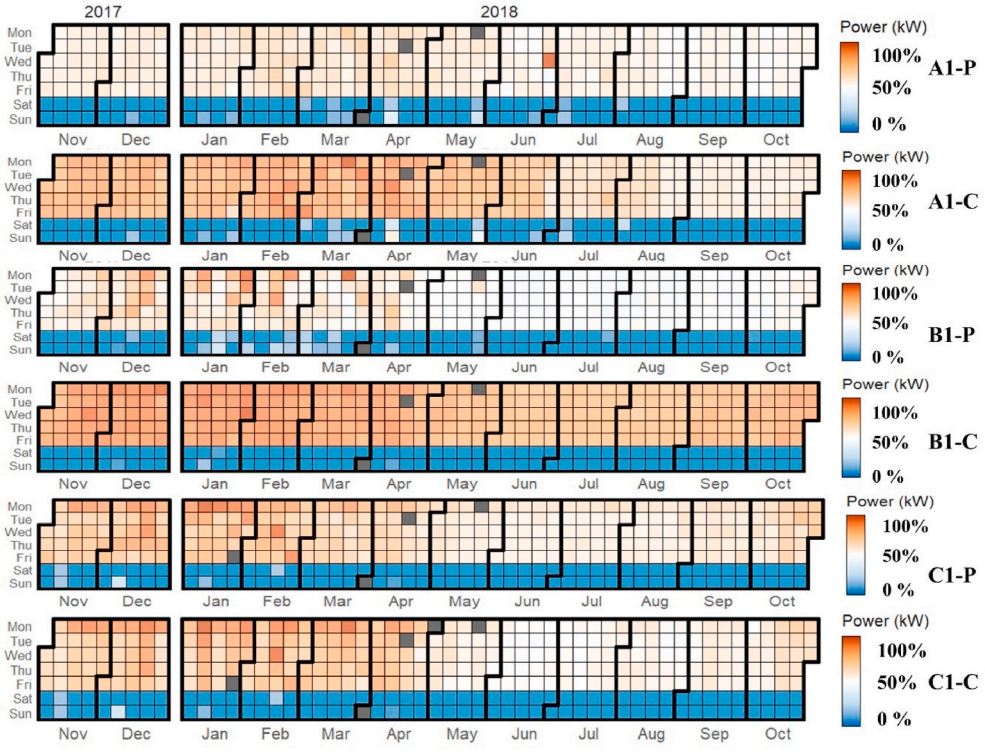
The best performance was reported in cluster 2, in which chiller PLR, CPR and water temperature differences across the evaporator and condenser, and COP were considerably above their respective mean values and COPD was below its mean value. The worst performance was reported in cluster 6 and cluster 7, with each operating variable and COP below their mean values and COPD above its mean value. These two clusters showed that operating the system at low CPR conditions (i.e. less than zero of the rescaled value in Fig. 9) will generally cause high irreversibility in the system. The clustering results indicated that the chiller performance had a strong relationship with the temperature



a) Chiller 1.



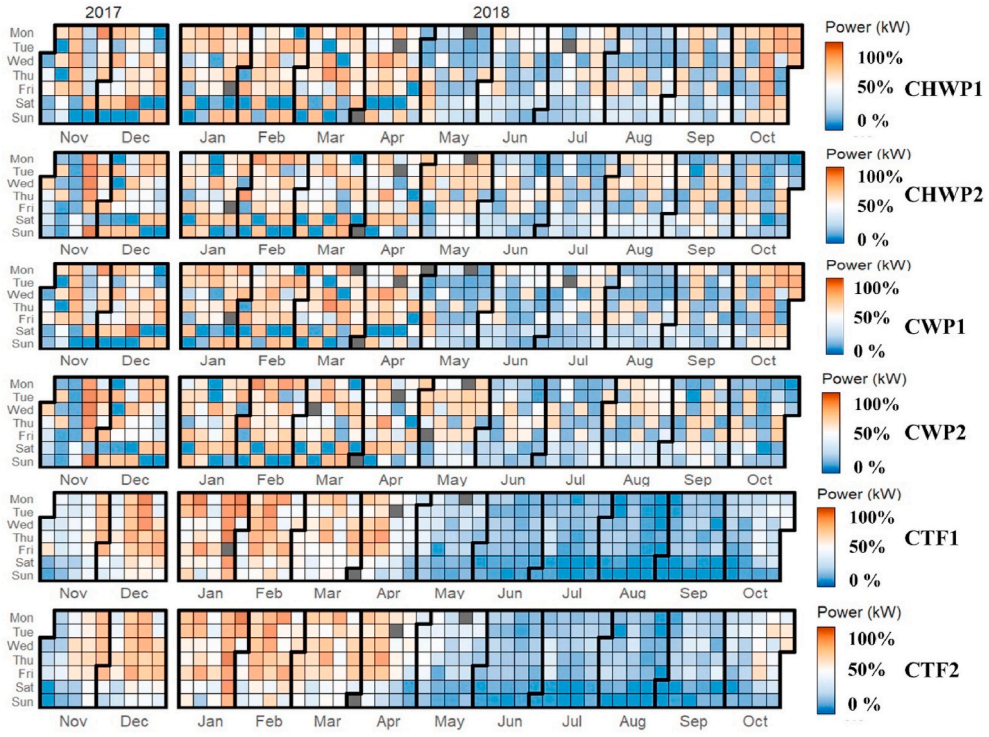
b) Chiller 2.



c) AHUs.

Fig. 5. Daily power ratio of main components in the chiller system in a calendar view.

difference across the evaporator and chiller PLR, and had a relatively weak relationship with the temperature difference across the condenser. For instance, the temperature difference across the condenser in cluster 1 was higher than that of cluster 2, however, the chiller performance in cluster 2 was higher than that in cluster 1. The major reason behind this was, in cluster 2, higher PLR values and higher temperature differences across the evaporator were achieved at relatively low CPR conditions compared to cluster 1. The results of cluster 3, cluster 4 and cluster 5 indicated that at high CPR conditions, a low temperature difference across the evaporator could cause large COP destruction. The overall results indicated that, to achieve the best performance, the chillers should be operated with PLR above 45% and CPR above 50% with the water temperature differences across the condenser and evaporator above their mean values. The best performing cluster can be set as a benchmark to achieve the optimized operation. The clustering results also validated the results of the CIT model but with better insights. The above results showed that the combination of the CIT model and AHC can help achieve optimized operation without opting for extensive conventional optimization models.



d) Water pumps and cooling tower fans.

Fig. 5. (*continued*).

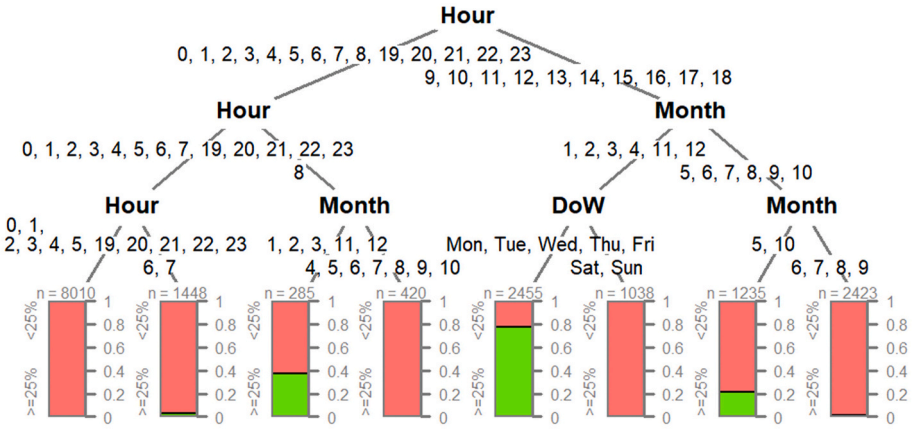


Fig. 6. Temporal classification of the CPR (Green bars indicated the percentage of the data observations with the CPR equal or higher than 25%, whereas the red bar shows the CPR below 25%). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.5. Qualitative analysis

ARM was used for further knowledge discovery from the processed data. By using the ARM model, a total of 57 sets of rules among COP, PLR, CPR, water temperature differences across the evaporator and condenser, and COPD were generated, as shown in Table 2. Antecedent and consequent were the first half and second half of a rule discovered by the association analysis. For example, in a rule {A, B} ⇒ {C}, A and B are antecedents, and C is consequent. "High" means higher than the median value and "Low" means lower than the median value. For example, one rule established in Table 2 with antecedent "CPR=High, COPD=High" and consequent "COP = Low" had support of 0.20 and confidence of 0.88. This indicated that the system COP was low, when the system had high COP destruction at high CPR conditions. 0.20 support indicated that 20% of the data had this rule and confidence of 0.88 indicated that 88% of the time, the dataset followed this rule.

The ARM model identified various combinations of the rules from the dataset. It can be observed that the system showed high performance (e.g. high COP and low COPD) when the water temperature differences across the evaporator and condenser, and PLR were concurrently higher than their respective median values. The overall results from the AHC, CIT and ARM models indicated that the system performed better when PLR was above 45%, CPR was above 50%, and the temperature differences across the evaporator and condenser were above their median values.

The results showed that data mining techniques can help identify both efficient and inefficient operating conditions of chiller systems. Temporal classification obtained by using the heatmaps and CIT model

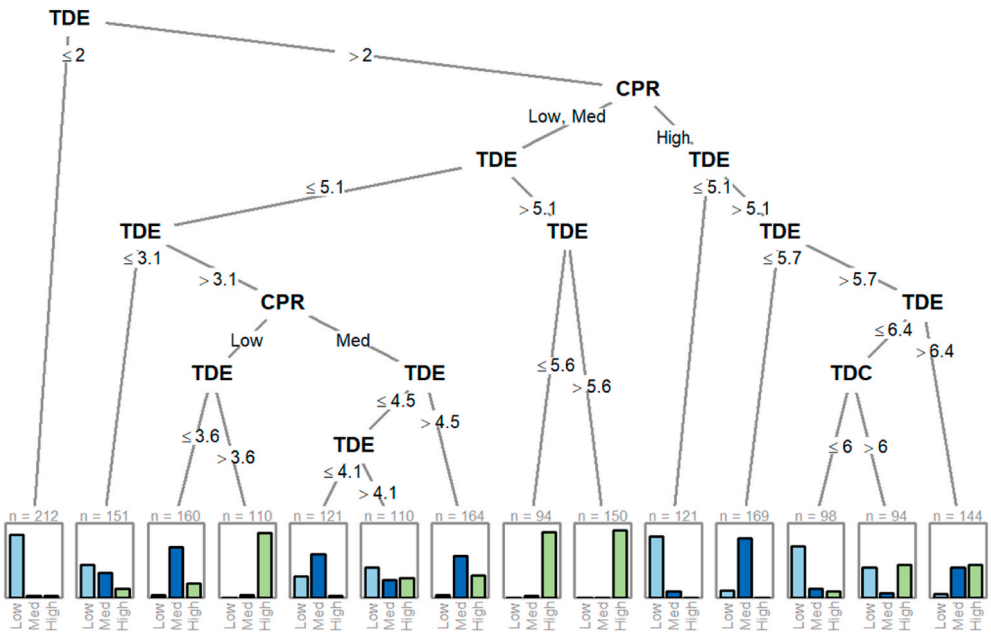


Fig. 7. Relationship between the system COP and the operating parameters.

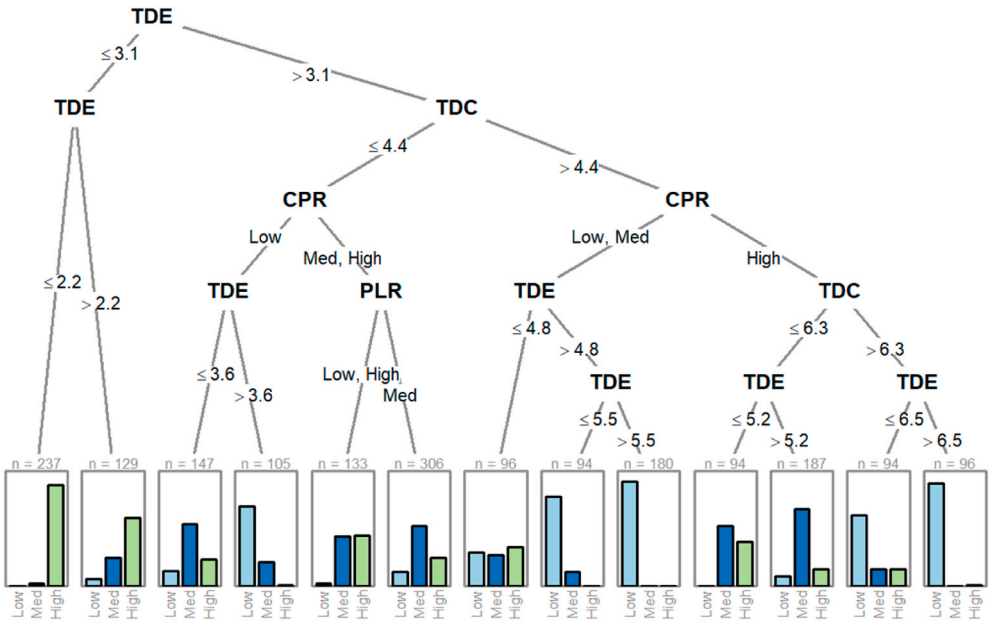


Fig. 8. Relationship between the system COP Destruction and operating parameters.

can be used to assess the operational behavior of the chiller system. The result obtained from the analysis showed several operational alternatives that can achieve a similar level of high performance under the same working conditions and this could provide the controller with certain flexibility in decision making. These results in conjunction with the known operational behavior of the system can help implement better operational strategies for chiller systems. Moreover, the data cleaning process and the introduction of the COPD indicator can increase the reliability of the performance assessment. It is worthwhile to note that the results reported in this study were generated based on the data collected from the case study system under the given climate. The results and the association rules might be different when the strategy developed is applied to a different chiller system at a different climate condition. However, the performance assessment strategy developed in this study can be adapted to evaluate the performance of other chiller systems and building energy systems.

4. Conclusions

This paper presented a data-driven strategy for the performance assessment of a chiller system using data mining and visualization techniques. The chiller performance was evaluated based on the operating variables by using the Agglomerative Hierarchical Clustering (AHC), Conditional inference tree (CIT) and Association rule mining (ARM). The collective effect of the operating variables on the system performance was studied. In this strategy, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was used to detect and remove the outliers in the collected data. CIT was used to classify the

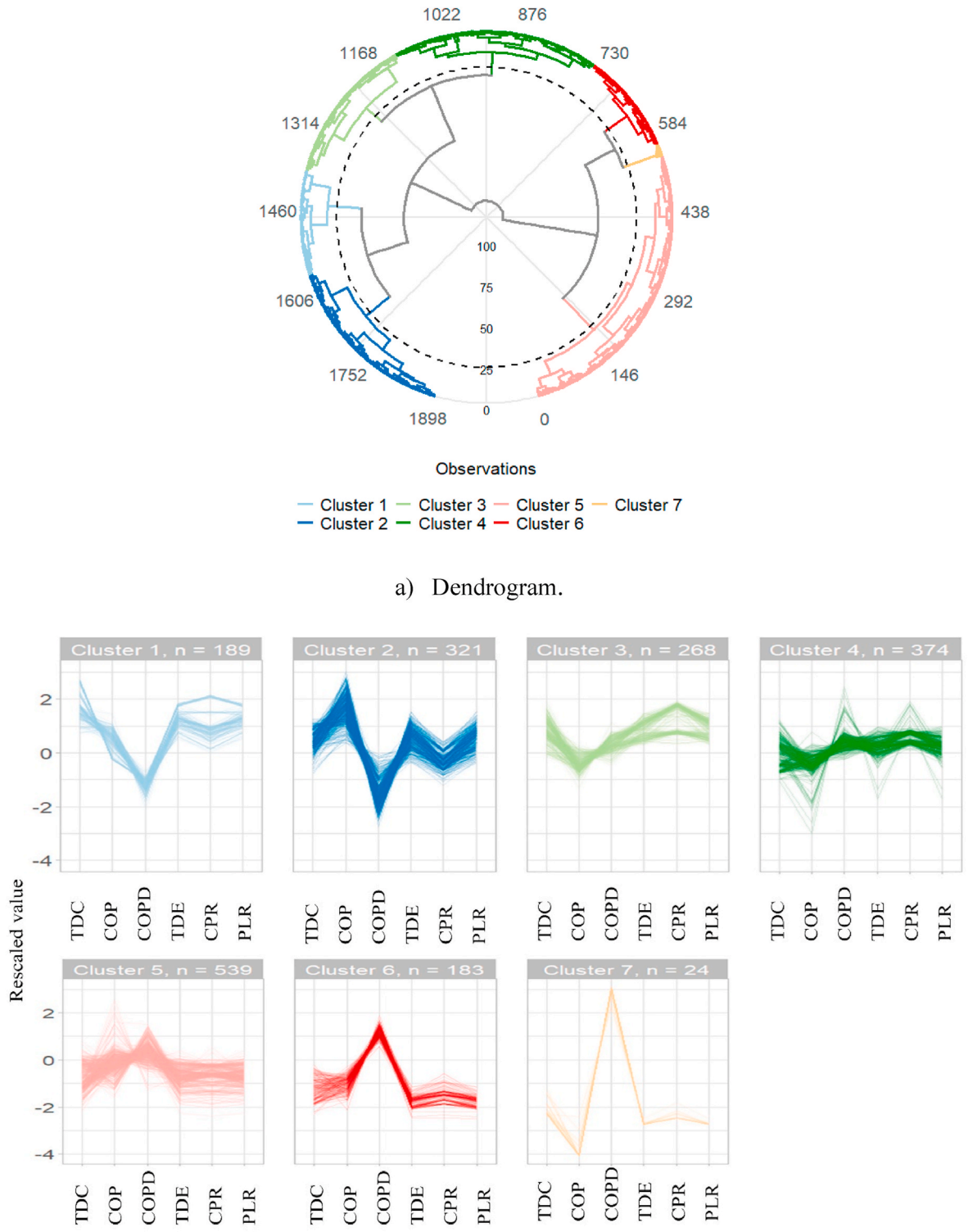


Fig. 9. Relationships between the performance indicators and operating variables in the identified clusters.

CPR based on the temporal variations. The CIT model and AHC were used to quantitatively analyze the effect of the operating variables on the performance of the system. A new performance indicator, COP Destruction was introduced to represent the quality of the COP. The introduction of this performance indicator enhanced the reliability of the performance assessment process. ARM was used for knowledge discovery by interlinking system performance and operational variables. The overall results indicated that to achieve better operating performance, the chiller system studied should be operated with PLR above 45%, CPR above 50%, and the water temperature differences across the evaporator and condenser above their respective median values. A total of 57 sets of rules among COP, PLR, water temperature differences

Table 1
Mean and median of each variable.

	TDE (K)	TDC (K)	PLR (%)	CPR (%)	COPD (%)	COP
Mean	4.4	4.1	44.4	53.8	71.8	3.2
Median	4.7	4.1	47.4	49.8	74.2	3.1

Table 2
Set of the rules generated from the ARM model.

Antecedent	Consequent	Support	Confidence
TDE = Low	PLR = Low	0.5	1
PLR=High	TDE = High	0.5	1
TDC = High	TDE = High	0.42	0.84
TDC = Low	TDE = Low	0.42	0.84
COP = Low	COPD=High	0.42	0.84
COP=High	COPD = Low	0.42	0.84
TDC = Low	PLR = Low	0.42	0.84
PLR=High	TDC = High	0.42	0.84
PLR=High, TDC = High	TDE = High	0.42	1
PLR = Low, TDC = Low	TDE = Low	0.42	1
PLR=High	CPR=High	0.41	0.82
CPR = Low	PLR = Low	0.41	0.82
TDE = High	CPR=High	0.41	0.82
TDE = High, CPR=High	PLR=High	0.41	1
TDE = Low, CPR = Low	PLR = Low	0.41	1
CPR = Low	TDE = Low	0.41	0.82
TDC = Low, CPR = Low	PLR = Low	0.37	0.99
TDC = Low, CPR = Low	TDE = Low	0.37	0.99
PLR = Low, COPD=High	TDE = Low	0.35	1
PLR=High, COPD = Low	TDE = High	0.34	1
TDC = High, CPR=High	PLR=High	0.33	0.89
TDC = High, CPR=High	TDE = High	0.33	0.89
TDC = High, COPD = Low	TDE = High	0.31	0.91
TDC = High, COPD = Low	PLR=High	0.31	0.91
COP=High, PLR=High	TDE = High	0.31	1
COP = Low, PLR = Low	TDE = Low	0.31	1
TDC = Low, COPD=High	TDE = Low	0.29	0.85
TDC = Low, COPD=High	PLR = Low	0.29	0.85
COP = Low, TDE = Low	COPD=High	0.29	0.94
COP = Low, PLR = Low	COPD=High	0.29	0.94
COP=High, TDC = High	COPD = Low	0.29	0.96
COP=High, TDE = High	COPD = Low	0.28	0.92
COP=High, PLR=High	COPD = Low	0.28	0.92
COP = Low, TDC = Low	COPD=High	0.27	0.92
CPR = Low, COPD=High	PLR = Low	0.27	1
CPR = Low, COPD=High	TDE = Low	0.27	1
COP=High, TDC = High	TDE = High	0.27	0.9
COP=High, TDC = High	PLR=High	0.27	0.89
COP=High, TDC = High, TDE = High	COPD = Low	0.26	0.97
COP=High, PLR=High, TDC = High	COPD = Low	0.26	0.97
CPR=High, COPD = Low	PLR=High	0.26	0.95
COP = Low, TDC = Low	TDE = Low	0.26	0.86
CPR = Low, COPD=High	TDC = Low	0.26	0.94
CPR=High, COPD = Low	TDE = High	0.26	0.94
COP = Low, TDC = Low	PLR = Low	0.26	0.85
COP = Low, PLR = Low, TDC = Low	COPD=High	0.24	0.94
CPR=High, COPD = Low	TDC = High	0.23	0.86
TDC = High, CPR=High, COPD = Low	PLR=High	0.23	0.98
TDC = High, CPR=High, COPD = Low	TDE = High	0.23	0.98
COP = Low, CPR = Low	PLR = Low	0.23	1
COP = Low, CPR = Low	TDE = Low	0.23	1
CPR = Low, COPD = Low	COP=High	0.22	0.95
COP=High, CPR=High	PLR=High	0.22	0.96
COP=High, CPR=High	TDE = High	0.22	0.96
COP = Low, CPR = Low	COPD=High	0.22	0.95
COP = Low, CPR = Low	TDC = Low	0.21	0.95
CPR=High, COPD=High	COP = Low	0.2	0.88

across the evaporator and condenser, CPR, and COP Destruction were established through association analysis. It was found that the system performance decreased when the PLR, and water temperature differences across the condenser and evaporator were concurrently below their median values. The results from the ARM model also validated the results obtained from the AHC and CIT model. It was found that the system performance was most strongly influenced by the temperature difference across the evaporator. The quantitative results showed a range of operating conditions that can be selected for better operation. Similarly, the results generated by the ARM model also showed that under a given condition, there were several alternatives that can achieve the better operational performance of the chiller system. The developed method can be used to assess the system performance, and the generated results can be used to develop energy efficient control strategies.

CRediT authorship contribution statement

Muhammad Bilal Awan: Methodology, Formal analysis, Writing – original draft. **Kehua Li:** Methodology, Formal analysis, Writing – original draft. **Zhixiong Li:** Supervision, Methodology, reviewing. **Zhenjun Ma:** Supervision, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The first author would like to acknowledge the Higher Education Commission (HEC), Pakistan for the provision of PhD scholarship and support.

References

[1] T. Abergel, C. Delmastro, Tracking buildings 2020, Paris, https://www.iea.org/reports/tracking-buildings-2020, 2020.
[2] HVAC, Energy breakdown. https://www.energy.gov.au/publications/hvac-factsheet-energy-breakdown, 2013.
[3] S. Bhawan, R.K. Puram, Energy performance assessment for equipment and utility system: chapter 9 - energy performance assessment of HVAC systems, Bur. Energy Effic. (2006).
[4] Y. Ji, P. Xu, J. Xie, A performance assessment method for main HVAC equipment with electricity submetering data, Procedia Eng., 2017, https://doi.org/10.1016/j.proeng.2017.10.320.
[5] A. Mouzakitis, Classification of fault diagnosis methods for control systems, Meas. Control (United Kingdom) (2013), https://doi.org/10.1177/0020294013510471.
[6] G. Anand, R. Kodali, Benchmarking the Benchmarking Models, Benchmarking, 2008, https://doi.org/10.1108/14635770810876593.
[7] F.W. Yu, K.T. Chan, R.K.Y. Sit, J. Yang, Review of standards for energy performance of chiller systems serving commercial buildings, in: Energy Procedia, 2014, https://doi.org/10.1016/j.egypro.2014.12.308.
[8] S. Voleti, Data collection, Int. Ser. Oper. Res. Manag. Sci. (2019), https://doi.org/10.1007/97833196883742.
[9] K. Li, Y. Sun, D. Robinson, J. Ma, Z. Ma, A new strategy to benchmark and evaluate building electricity usage using multiple data mining technologies, Sustain. Energy Technol. Assessments. (2020), https://doi.org/10.1016/j.seta.2020.100770.
[10] I.P. Panapakidis, T.A. Papadopoulos, G.C. Christoforidis, G.K. Papagiannis, Pattern Recognition Algorithms for Electricity Load Curve Analysis of Buildings, Energy Build, 2014, https://doi.org/10.1016/j.enbuild.2014.01.002.
[11] Z. Ma, R. Yan, N. Nord, A Variation Focused Cluster Analysis Strategy to Identify Typical Daily Heating Load Profiles of Higher Education Buildings, Energy, 2017, https://doi.org/10.1016/j.energy.2017.05.191.
[12] Z. Ma, R. Yan, K. Li, N. Nord, Building Energy Performance Assessment Using Volatility Change Based Symbolic Transformation and Hierarchical Clustering, Energy Build, 2018, https://doi.org/10.1016/j.enbuild.2018.02.015.
[13] K. Li, R.J. Yang, D. Robinson, J. Ma, Z. Ma, An agglomerative hierarchical clustering-based strategy using Shared Nearest Neighbours and multiple dissimilarity measures to identify typical daily electricity usage profiles of university library buildings, Energy (2019), https://doi.org/10.1016/j.energy.2019.03.003.
[14] K. Li, Z. Ma, D. Robinson, J. Ma, Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering, Appl. Energy (2018), https://doi.org/10.1016/j.apenergy.2018.09.050.
[15] M. Li, Y. Ju, The Analysis of the Operating Performance of a Chiller System Based on Hierarchal Cluster Method, Energy Build, 2017, https://doi.org/10.1016/j.enbuild.2016.12.076.
[16] F.W. Yu, K.T. Chan, Using Cluster and Multivariate Analyses to Appraise the Operating Performance of a Chiller System Serving an Institutional Building, Energy Build, 2012, https://doi.org/10.1016/j.enbuild.2011.10.026.
[17] F.W. Yu, K.T. Chan, Chiller system performance benchmark by data envelopment analysis, Int. J. Refrig. (2012), https://doi.org/10.1016/j.ijrefrig.2012.07.003.

[18] F.W. Yu, K.T. Chan, Improved energy management of chiller systems with data envelopment analysis, Appl. Therm. Eng. (2013), https://doi.org/10.1016/j.applthermaleng.2012.08.023.
[19] F.W. Yu, K.T. Chan, Environmental performance and economic analysis of all-variable speed chiller systems with load-based speed control, Appl. Therm. Eng. (2009), https://doi.org/10.1016/j.applthermaleng.2008.08.003.
[20] F.W. Yu, K.T. Chan, Improved energy management of chiller systems by multivariate and data envelopment analyses, Appl. Energy (2012), https://doi.org/10.1016/j.apenergy.2011.11.016.
[21] F.W. Yu, K.T. Chan, Assessment of operating performance of chiller systems using cluster analysis, Int. J. Therm. Sci. (2012), https://doi.org/10.1016/j.ijthermalsci.2011.10.009.
[22] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd Int. Conf. Knowl. Discov. Data Min., 1996.
[23] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to data mining (new international editon). https://doi.org/10.1017/CBO9781107415324.004, 2013.
[24] T. Hothorn, K. Hornik, A. Zeileis, Unbiased recursive partitioning: a conditional inference framework, J. Comput. Graph Stat. (2006), https://doi.org/10.1198/106186006X133933.
[25] P.-N. Tan, M. Steinbach, V. Kumar, Association analysis: basic concepts and algorithms, Introd. to Data Min. (2005), https://doi.org/10.1111/j.1600-0765.2011.01426.x.
[26] K.S. Leung, K.C. Wong, T.M. Chan, M.H. Wong, K.H. Lee, C.K. Lau, S.K.W. Tsui, Discovering protein-DNA binding sequence patterns using association rule mining, Nucleic Acids Res. (2010), https://doi.org/10.1093/nar/gkq500.
[27] A. Almansory, Applying association rules and decision tree algorithms with tumor diagnosis data, SSRN Electron. J. (2018), https://doi.org/10.2139/ssrn.3028893.
[28] M. Kaur, S. Kang, Market basket analysis: identify the changing trends of market data using association rule mining. Procedia Comput. Sci., 2016, https://doi.org/10.1016/j.procs.2016.05.180.
[29] C. Fan, F. Xiao, H. Madsen, D. Wang, Temporal Knowledge Discovery in Big BAS Data for Building Energy Management, Energy Build, 2015, https://doi.org/10.1016/j.enbuild.2015.09.060.