Application of data mining in understanding the charging patterns of the hot water tank in a residential building: a case study

To cite this article: M Miraghi et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 609 052038

View the article online for updates and enhancements.
Application of data mining in understanding the charging patterns of the hot water tank in a residential building: a case study

M Mirnaghi¹, K Panchabikesan¹ and F Haghighat¹,*

¹ Department of Building, Civil & Environmental Engineering, Concordia University, Montreal, Canada
* fariborz.haghighat@concordia.ca

Abstract. Today’s buildings are adequately informative because of the wide implementation of the building automation system (BAS). In this regard, data mining (DM) tools have an excellent ability to interpret and discover interesting, unknown information from raw data collected from BAS. The present work is a case study, in which DM tools (such as clustering and decision tree) are applied to understand the operational pattern of a hot water storage tank (sanicube) used in a residential building, located in Scotland. The main objective of this study is to explore the correlation among the important features (storage tank temperature, solar collector, the operation of the sanicube heating element, etc.) in the chosen building. Interesting correlations and patterns between the solar collector, storage tank, sanicube-heating element are explored. The clustering results show the existence of different charging patterns of the storage tank and it highly influences the space heating (SH) and domestic hot water (DHW) temperature. The main inference from the result is that the charging of the storage tank is not automatic, and it is influenced by the occupancy behaviour. To maintain the temperature of the storage tank and subsequently to meet the SH and DHW demand, the operation of both solar collector and sanicube heating element (during night time) is recommended rather than operating only the solar collector.

1. Introduction
In a building, heating ventilation and air conditioning (HVAC) systems are responsible for approximately half of the building’s total energy consumption. In recent years, renewable energy resources (RES) are utilized to cover partial or entire energy usage in buildings and in some cases, the surplus amount of energy is generated by these sources. However, most of the RES is intermittent and hence there exists a mismatch between energy demand and supply. Meanwhile, thermal energy storage (TES) is a promising solution to solve this problem. Studies conducted on solar assisted TES system exemplified how solar collectors coupled with electrical elements are capable of reducing the heating energy consumption in buildings [1]. In this case, optimizing the trade-off schedule between solar energy and grid electricity for charging the tank is very important. In recent years, data are monitored and collected from different parts of the building’s heating and electrical systems by employing several sensors. To systematically analyze the collected raw data and to extract interesting, useful, previously unknown rules, application of data mining (DM) techniques would be beneficial [2]. In general, DM-based methods are grouped into two categories: supervised and unsupervised methods [3]. Supervised methods are used for predictive modelling and can interpret the relationship between the input and output variables. Domain expertise is crucial for the successful application of supervised methods. In contrast, unsupervised learning does not involve the use of sample data, but it nonetheless offers a more promising method to discover the intrinsic structure, associations and useful inherent rules in the data. Through the application of unsupervised (e.g., clustering) or supervised (e.g., decision tree (DT)) methods, similar objects in the dataset can be grouped and the unseen relation between the considered
building components can be explored, respectively. Yu et al. [4] applied DT to classify and predict the energy demand levels in a set of residential buildings. The authors stated that by using DT, both the numerical and categorical variables can be processed to perform classification and prediction tasks rapidly without requiring much computational knowledge. On the other hand, though DM tools are widely applied in buildings energy related analysis, studies reporting the application of DM tools toward assessing the performance of TES system are scarcely found in the literature. The present work is an illustration on the application of DM tools over the raw data collected from the residential hot water storage tank system to investigate the operational patterns of the storage system and to discover the previously unknown, useful knowledge from the collected dataset. The main objectives of the present work are 1) to explore the impact of solar collector and sanicube heating element operation in charging the hot water storage tank; and 2) to find the significant factors that influence the room temperature; using DM tools.

2. System description and methodology

2.1. System description

The dataset used in this study belongs to a residential house from Centini district, located in Scotland. Note that the data collected between August to December 2014 was considered for the analysis as most of the features were largely missing between January to July 2014. The data were recorded with five minutes interval and the collected data includes sanicube heating element energy consumption (kWh), outdoor temperature (°C), room temperature (°C), solar collector volumetric flow rate (m³/hr) and temperature (°C), space heating (SH) volumetric flow rate (m³/hr) and temperature (°C), domestic hot water (DHW) volumetric flow rate (m³/hr) and temperature (°C). Figure 1 shows the simplified schematic of the hot water heating system of the considered building. The water temperature is measured at different heights of the storage tank. The location of the temperature sensors in the hot water storage tank is mentioned as Tank 1 to Tank 6 in Figure 1. In the following sections, the tank temperature refers to the average of the water temperature measured at different heights of the tank. The hot water storage tank is a sanicube tank with heating coils surrounded. Note that, the tank is charged by a solar collector, the heating coils (mostly during the night time) and the hot water from the tank are used for both DHW and SH applications.

![Figure 1. Simplified schematic diagram of the hot water storage tank.](image)

2.2. Methodology

Note that the raw dataset collected from the considered building should be processed before using the data for the DM. As the first step of data pre-processing, outliers and the missing values should be processed. In this study, outlier detection was carried out by the interquartile range (IQR) method. After removing the outliers and missing values, in total, 103 days were considered for the analysis. Successively, data aggregation and normalization were done. The five minutes data was aggregated to hourly data, and after data aggregation, 2472 observations for each attribute were considered. Data normalization was done prior to the clustering using the min-max normalization method, and
accordingly, the data were transformed into a range between 0 to 1. After data processing, k-means clustering, and DT were used as DM tools to explore the significant and unseen correlations between the considered variables. The k-means clustering is most commonly used clustering method which separates ‘n’ number of observations into ‘k’ number of clusters, in which the observations in the same cluster have similar mean/centroid values. In this study, the optimum number of clusters was determined using the silhouette and Dunn indices. The silhouette index measures the similarity of an object to its own cluster compared to other clusters. It ranges between -1 to +1, where a higher value indicates that the object is very similar to its own cluster and different from the other clusters [5]. Dunn index is another validating index widely used to identify the clusters that are compact and well separated. The main objective of Dunn’s index is to maximize the intercluster distance while minimizing the intracluster distances. The number of clusters with higher Dunn’s index is generally be considered as the optimal number of clusters [6]. Once the dataset was grouped into several clusters, DT analysis was performed on each cluster separately to find out the factors that highly influence the variations in room temperature. To perform the k-means clustering and to validate the number of clusters, the open source software ‘R’ [7] was used. Thus, the ‘RapidMiner’ software [8] was applied to perform the DT analysis.

3. Results and Discussion
3.1. Clustering results
Figures 2 (a) and (b) show the clustering validation results for Dunn index and Silhouette index, respectively. Both validation indices show that k = 2 and 3 as the optimal number of clusters. To get better insights, Table 1 and 2 show the centroid values for the considered attributes for k = 2 and 3, respectively. Though several parameters were considered in the cluster analysis, the inference from the centroid results is that the clusters were formed mainly based on the tank temperature and outdoor temperature. This is because these two parameters had more variations in the dataset than others. The centroid values presented in table 1 indicates that cluster_0 had a higher room temperature (0.61), tank temperature (0.64), lower outdoor temperature (0.46), and higher sanicube heating element consumption (0.04) compared to the cluster_1. Note that, the values mentioned inside the parenthesis denotes the centroid values of the respective attributes. Since the tank temperature (0.31) is lower in cluster_1, subsequently, SH temperature (0.26), DHW temperature (0.27) and room temperature (0.56) is also lesser in cluster_1 compared to cluster_0. On the other hand, in the case of k = 3, cluster_0 had higher centroid values for room temperature (0.66), tank temperature (0.61), outdoor temperature (0.54), solar flow temperature (0.47), and lesser sanicube heating element consumption (0.017), whereas cluster_2 had relatively lower room temperature (0.54), tank temperature (0.30), outdoor temperature (0.53), solar flow temperature (0.45), and higher sanicube heating element energy consumption (0.021) than cluster_0. As ‘k’ was set to 3, an extra cluster (cluster_1) was formed for the days when the outdoor temperature is extremely low (0.155) compared to cluster_0 and cluster_2 and the heating element energy consumption (0.10) was relatively very high compared to the other two clusters. During the days that are grouped in cluster_1, charging of the hot water tank was done completely by sanicube heating element and hence a higher sanicube heating element energy consumption. In specific, data belonged to all the days in December and a few days in November were grouped in cluster_1. Since the more detailed interpretation of the dataset can be made when ‘k’ is set to 3, the dataset was segregated into 3 clusters, in which 40%, 9% and 51% of the dataset belonged to cluster_0, 1 and, 2 respectively.
Table 1. Centroid values of the considered attributes (for $k = 2$).

| Cluster number | Sanicube energy | $T_{Room}$ | $T_{Tank}$ | $T_{Solar}$ | Solar vol. flow | $T_{DHW}$ vol. flow | $T_{SH}$ vol. flow | $T_{out}$ |
|----------------|-----------------|------------|-------------|-------------|----------------|---------------------|-------------------|-----------|
| 0              | 0.04            | 0.61       | 0.64        | 0.41        | 0.00           | 0.50                | 0.01              | 0.53      |
| 1              | 0.02            | 0.56       | 0.31        | 0.46        | 0.01           | 0.27                | 0.01              | 0.26      |

Table 2. Centroid values of the considered attributes (for $k = 3$).

| Cluster number | Sanicube energy | $T_{Room}$ | $T_{Tank}$ | $T_{Solar}$ | Solar vol. flow | $T_{DHW}$ vol. flow | $T_{SH}$ vol. flow | $T_{out}$ |
|----------------|-----------------|------------|-------------|-------------|----------------|---------------------|-------------------|-----------|
| 0              | 0.017           | 0.661      | 0.616       | 0.472       | 0.005          | 0.491               | 0.004             | 0.497     |
| 1              | 0.100           | 0.452      | 0.672       | 0.170       | 0.000          | 0.510               | 0.038             | 0.637     |
| 2              | 0.021           | 0.541      | 0.303       | 0.453       | 0.011          | 0.262               | 0.010             | 0.249     |

It is interesting to infer from table 2 that the variation in outdoor temperature ($T_{out}$), solar collector flow temperature ($T_{Solar}$), sanicube element energy consumption (Sanicube energy) is small between cluster_0 and cluster_2. But, the centroid value for the tank temperature ($T_{Tank}$) in cluster_2 was found to be 0.30, whereas, in cluster_0, it is 0.61. The reason for this can be explained by referring to figures 3 and 4. As it can be seen from the figures, the operation of sanicube heating element was more frequent in cluster_0 than the cluster_2. That means, in cluster_0, the hot water storage tank was charged by both solar collector (during the day) and sanicube heating element (during the night). Subsequently, the tank temperature was maintained averagely at 50°C throughout the day, whereas, for most of the days in cluster_2, the hot water tank temperature was charged mainly by the solar collector and the operation of sanicube heating element was less compared to cluster_0. Because of this, the average tank temperature in cluster_2 was found to be 37°C. Interestingly, table 2 shows that the energy consumption of sanicube element in cluster_0 is lesser than cluster_2. Though the sanicube element is frequently operated in cluster_0, for most of the time, energy consumed by the heating element was in the range of 4 to 6 kWh. On the other hand, though the heating element was operated for a few times in cluster_2 (compared to cluster_0), the heating element consumed 8 to 10 kWh per operation. This is the reason for which in cluster_0 and 2, the centroids for sanicube heating element energy consumption was found to be 0.017 and 0.021 respectively. The clustering result reveals the existence of different charging patterns of the hot water storage tank for the same climatic condition. The further inference is that the charging of the hot water tank was not automatic, and it is performed by the occupants. Based on the results obtained from the clustering analysis, charging pattern followed in the cluster_0 is recommended over the charging pattern followed in cluster_2 because, in cluster_0 tank temperature, room temperature, SH and DHW temperature were maintained high and the energy consumed by the sanicube heating element was also lesser.

Figure 3. Profile of the tank, solar collector flow temperature and sanicube element energy consumption in cluster_0.
3.2. Decision tree results

Note that, the variations in the SH and DHW flow temperature are directly proportional to the change in the hot water storage tank temperature. However, the variations in the room temperature are caused by several factors such as outdoor weather conditions, tank temperature, DHW consumption, occupancy behaviour, etc. Hence, to identify the parameters that highly influence the room temperature other than the expected attributes (such as outdoor temperature, hot water tank temperature), DT analysis was performed. Before performing the DT analysis, each attribute was discretized into two ranges. For the explanation purpose, DT results obtained for cluster_2 is shown in figure 5. Note that, the dataset was divided into training (75%) and testing (25%) dataset. Pruning with the confidence of 0.1 and pre-pruning with minimal gain of 0.1 was applied to limit the tree depth. The accuracy of the confusion matrix for cluster_2 was 88.2%. The DT analysis results show that the room temperature is primarily influenced by the outdoor temperature and SH volumetric flow rate, which is obvious. As it can be seen from the left side branch of the tree (in figure 5), if the outdoor temperature is in range 2 (11.4°C - 23.5°C), and the SH volumetric flow rate is in the range 1 (0 m³/hr – 0.11 m³/hr), the room temperature is in the range 2 (18.4°C – 24.4°C). This is because, when the outdoor temperature is high, the SH demand is generally less. Hence, the SH volumetric flowrate is in range 1, and the room temperature is in the range 2. In addition to this, the DT analysis also reveals that the DHW flow rate has an impact on the room temperature. When the SH volumetric flow is in range 2 (0.11 m³/hr - 0.36 m³/hr) and the DHW volumetric flow is also in the range 2 (0.04 m³/hr to 0.19 m³/hr), for most of the data distribution, the room temperature is in the range 1 (13.0°C - 18.4°C). This is because, when there is demand for both SH and DHW usage, more amount of hot water from the storage tank is dedicated to meet the DHW demand, and the priority to SH flow temperature is comparatively lesser. For this reason, the room temperature is in the range 1. However, if the SH flow rate is in range 2 and DHW flow rate is in range 1, the room temperature is in range 2. The right-side branch of the DT shows that, if the outdoor temperature is in the range 1 (-1.47°C - 11.48°C) and if the SH flow rate is in range 2, then the room temperature is in range 1. This condition had occurred for fewer observations, probably when the outdoor temperature is extremely low. Also, if the outdoor temperature and SH flow rate are in range 1, then for most of the data distribution, the room temperature is in range 2. The possible reason for this is, during these periods, the sanicube heating element might be in operation. The overall inference from the DT results is that in cluster_2, during the period when the sanicube heating element was not operated and the usage of DHW is high, priority is given for DHW usage ahead of SH. Subsequently, the room temperature becomes lower (range 1). Similar results are obtained for cluster_0. Hence, whenever the DHW usage is high and the room temperature is in range 1, it is recommended to turn ON the sanicube heating element for the desired time (even during the day) to increase the tank temperature to meet both the SH and DHW demands.
4. Conclusion

In this work, DM tools (clustering and DT) are applied to interpret the charging patterns of the hot water tank and to find the parameters that highly influence the room temperature of the considered building. The main steps of analysis involved: 1) data pre-processing 2) finding the desired number of the clusters 3) k-means clustering, and 4) classification by DT analysis. After data pre-processing, clustering was performed to similar group data into several subsets. The optimal number of clusters was set to be three by using the clustering validation indexes. Comparing three clusters’ centroid, it is observed that cluster_2 has lower tank temperature compared to other clusters. In spite of the similarity of other parameters in cluster_0 and cluster_2, results show that cluster_0 is related to the days when the tank was charged using both solar collector and sanicube heating element operation. However, in cluster_2 tank temperature was mainly relied on the solar collector flow temperature. In cluster_1, the charging of the tank is solely relied on the sanicube heating element operation, because in cluster_1, the days with very low temperature (-1°C to 5°C) was grouped and during these days the contribution of the solar collector to charge the tank was negligible. In addition to the clustering results, the DT results obtained for cluster_2 is explained. The DT analysis results showed that, other than the outdoor temperature and SH volumetric flow rate, room temperature is influenced by the DHW volumetric flow rate. When the DHW demand is high, priority is given for the DHW demand than SH and subsequently, the room temperature is reduced. Therefore, to maintain the room temperature and to meet the DHW demands, operation of the sanicube heating element (irrespective of the time of the day) for the desired time is recommended.

References

[1] Antoniadis C. N and Martinopoulos G 2017 Simulation of Solar Thermal Systems with Seasonal Storage Operation for Residential Scale Applications J.Procedia Environ. Sci 38 405–12

[2] Zhang C, Cao L and Romagnoli A 2018 On the feature engineering of building energy data mining Sustain Cities Soc 39 508–518

[3] Fan C, Xiao F, Li Z and Wang J 2018 Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review J. Energy Build 159 296–30

[4] Yu Z, Haghighat F, Fung B. C. M and Yoshino H 2010 A DT method for building energy demand modelling J. Energy Build 42 1637–46

[5] Lletí R, Ortiz M. C, Sarabia L. A and Sánchez M. S 2004 Selecting variables for k-means cluster analysis by using a genetic algorithm that optimizes the silhouettes J.Anal. Chim. Acta 515 87–100

[6] Dunn J. C 1973 A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters J. Cyberrn 3 32–57

[7] R Development Core Team 2008 R: A language and environment for statistical computing.

[8] Mierswa I and Klinkenberg R 2018 RapidMiner Studio (9.1) Data science, machine learning, predictive analytics.