ANALES de Edificación

Received: 08/11/2022
Accepted:14/11/2022

Anales de Edificación
Vol. 8, Nº3, 8-13 (2022)
ISSN: 2444-1309
Doi: 10.20868/ade.2022.5090

Aplicación de una herramienta avanzada de análisis de datos en el entorno de la construcción. Application of advanced data analysis tool in building environment.

Arrate Hernández-Arizaga[a]; Ana Picallo-Pérez[a]; José María Sala-Lizarraga[a]

[a] Research group ENEDI, Department of Energy Engineering, University of the Basque Country (UPV/EHU); Alameda Urquijo, S/N, 48013 Bilbao, Vizcaya, Spain, e-mail: arratehernandez@ehu.eus

Resumen-- **Los sistemas de monitorización de edificios proporcionan grandes volúmenes de información y existen herramientas avanzadas de análisis de datos. Un problema de detección y diagnóstico de fallos (FDD) en los sistemas energéticos de los edificios también puede considerarse un problema de aprendizaje automático puro. El objetivo de este trabajo es promover la FDD con aplicaciones de aprendizaje automático en el entorno de los edificios. Como contribución, en este trabajo se procesan series de datos temporales brutos, obtenidos de un SCADA, para la posterior construcción de patrones de una instalación térmica de un edificio. La instalación térmica abastece las demandas de ACS y calefacción de un edificio residencial, compuesto por 26 viviendas sociales y situado en Durango (norte de España). Los datos registrados cada 24 horas en valores acumulados se incluyen en el software R para el cálculo de gráficos estadísticos. Para los valores de los contadores de consumo de ACS y calefacción se obtienen 229 puntos de datos válidos y los rangos de consumo diario están comprendidos entre 1,94 - 5,90 m3 y 0 - 547,63 kWh.**

Palabras clave— Detección y diagnóstico de fallos; Instalaciones térmicas; Aprendizaje automático; SCADA; Software R..

Abstract— **Building monitoring systems deliver large volumes of information and advanced data analysis tools are available. A fault detection and diagnosis (FDD) problem in building energy systems can also be regarded as a pure machine learning problem. The aim of this work is to promote FDD with machine learning applications in building environment. As a contribution, in this work raw time data series, obtained from a SCADA, are processed for further pattern construction of a building thermal facility. The thermal facility supplies the DHW, and heating demands of a residential building, consisting of 26 social dwelling units and located at Durango (northern Spain). Data recorded every 24 hours in cumulative values is included in the R software for computing statistical graphs. For DHW and heating consumption meter values, 229 valid data points are obtained, and the daily consumption ranges are between 1.94 - 5.90 m3 and 0 - 547.63 kWh respectively.**

Index Terms— Fault detection and diagnosis; Thermal facilities; Machine learning; SCADA; R software.

I. Introduction

IN an attempt to deal with the climate change, the European Union is committed to decarbonisation by 2050 (European Union, 2016), reducing CO2 emissions by 80% and energy consumption by 50%. Saving energy by retrofitting existing buildings is one of the most attractive and low-cost options for reducing CO2 emissions (Kylili et al., 2016). Residential and service sector buildings are responsible for 29.5% of the final energy consumption (IDAE, 2018), to maintain thermal comfort and indoor air quality (IAQ), as well as the need to supply the required domestic hot water (DHW), which are supplied through thermal facilities. Building energy systems proper behaviour is quite complex since these systems consist

A.H., A.P., and J.M.S. are researches and assistant professor at ENEDI, Department of Energy Engineering, University of the Basque Country (UPV/EHU); Alameda Urquijo, S/N, 48013 Bilbao, Vizcaya, Spain.

of sensors, actuators, controllers, and devices simultaneously interacting in a very dynamic mode. Poor maintenance, improper performance of components, installation faults, and control errors significantly affect the efficiency of energy systems.

In operation systems, fault detection and diagnosis (FDD) are vital to reduce the energy consumption (Ahmad t al., 2016). A FDD problem in building energy systems can also be regarded as a pure machine learning problem (Kalogirou, 2003). If there is sufficient training data, the task of fault detection is to distinguish whether the patterns of monitoring data are like those of the normal training data; and if not, it means that there is, at least, one fault.

FDD methods are classified into data-driven, grey box and prior knowledge-based methods (Yang et al., 2014) or into history-based, qualitative model-based and quantitative model-based methods (Katipamula et al., 2005). According to (Mirnaghi et al., 2020), precise modelling is vital for FDD, since the characterization of the real dynamic behaviour is base for diagnosis and fault detection. In this sense, data-driven models seem to be promising (Yang et al., Kim et al., 2016, 2018).

Data-driven or black-box models need data extracted from the monitoring system of the facility and advanced data analysis tools. They are constructed simply by measuring the input and output data of each component (or box) and fitting a specific mathematical function that corresponds to the extracted recorded data. Therefore, black-box models, which have been shown to have high accuracy, do not require the understanding of the system physics; however, they have poor generalization capabilities (Afram et al., 2015).

The monitoring system gathers relevant data over time to evaluate equipment or system performance (ASHRAE, 2014). Data measured, among others, are the energy consumption, temperatures, mass flow rates, etc., at a specific time frequency. Since databases are large, data-mining techniques are used to extract usable data that are the basis for further data-driven models' construction. The most used data-mining techniques are clustering, classification, and regression (Lumbreras et al., 2020). Nevertheless, before applying data-mining techniques, raw data collected by the monitoring system must be processed, since it usually contains missing values and noise that should be removed. Besides, the identification of outliers is a key step for all applications related to data mining, as they can disturb the real nature of the data.

Considering the relevance of the raw-data treatment, in this work, 9 raw time data series are taken from the Supervisory Control and Data Acquisition (SCADA) (Boyer S.A., 2009) system of a thermal facility that supplies the DHW and heating demands of a residential building. Data series are processed with a script generated in R software (R Software, 2022). The objective is to obtain valid data series, removing missing values and identifying the outliers as a contribution for further data-mining application and data-driven patterns of the facility.

After the introduction and the definition of the objective in Section 1, the case study and the methodology are explained in Section 2. Section 3 contains the graphical and numerical results. Finally, the conclusions are presented in Section 4.

II. Case study

The thermal facility is in a building with 26 social housing units at Durango (Basque Country, northern Spain), Fig. 1. A more extensive description of the building can be found in Ref.(BEST, 2012).

Fig. 1. Case study building.

A. Description of the thermal facility and the meter devices

The thermal facility supplies DHW and heating to the whole building. The generation system has a 68-kW water-water reversible ground-source heat pump (GSHP) and a 120-kW natural gas condensing boiler. The storage system has three buffer tanks with an accumulation capacity of 2,000 l each. Two of them are used to store DHW and are connected in series; the third tank is used for heating storage.

The priority of the control of the thermal system is to cover the DHW demand and afterwards the heating demand. The GSHP covers either the DHW or the heating demands and provides exclusively the energy to preheat the water coming from the supply network. The boiler can cover DHW, and the heating demands simultaneously and provides the energy until the preheated water reaches the DHW consumption temperature. The set-point temperature values for the preheating tank and the DHW final consumption tank are 40°C and 55°C, respectively. Over the year, the non-heating season (May 15-October 14) and the heating season (October 15-May 14) are distinguished.

The meter devices coded as C7, C8, C16, C17, C26, C28, C31, C32 and C34 belong to the monitoring system of the thermal facility.

- Meters C7, C8, C16, C17 and C34 account for the thermal energy in kWh.
- Meters C26 and C28 account for the consumption of

electrical energy in kWh in the GSHP and gas consumption in m3 in the boiler respectively.

- Finally, meters C31 and C32 account for the DHW consumption in m3 and heating in kWh at the consumption points respectively.

A scheme of the facility with the numbering of the main components and the location of the meter devices C7, C8, C16, C17, C26, C28, C31, C32 and C34 can be found in Figure 2 and Table 1.

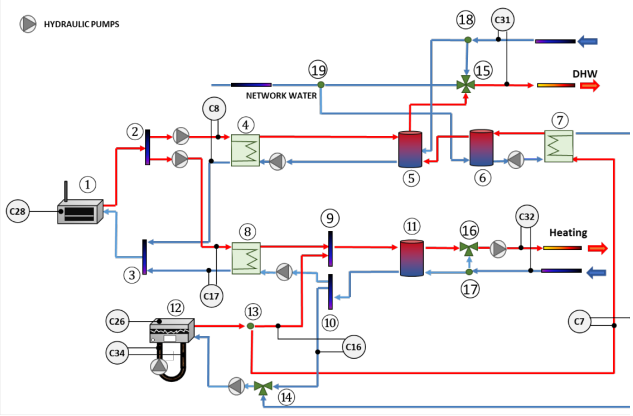

Fig. 2. Scheme of the facility.

TABLE I
NUMBERING AND BRIEF DESCRIPTION OF THE MAIN ELEMENTS.

No	DESCRIPTION
1	Boiler
2	DHW/heating diverter from the boiler
3	DHW/heating mixer from the boiler
4	HX for DHW from the boiler
5	DHW final temperature storage
6	DHW preparation storage
7	HX for DHW from the GSHP
8	HX for heating from the boiler
9	Heating mixer from the GSHP and the boiler
10	Heating diverter from the boiler and the GSHP
11	Heating storage
12	Ground Source Heat Pump (GSHP)
13	DHW/heating diverter from GSHP
14	DHW and heating 3-way valve
15	DHW four-way valve
16	Heating preparation 3-way valve
17	Diverter
18	Diverter
19	Net water 3 –way valve

B. *Methodology for raw data treatment*

As said, raw time data series for one year, from June 1st, 2019, to May 31st, 2020, of meters C7, C8, C16, C17, C26, C28, C31, C32 and C34 are obtained from SCADA. The main handicap of these data is that the values are accumulated every day and may include errors that produce inconsistencies for further data analysis. Therefore, the first step to be made is to create a methodology that cleans the errors and extracts the appropriate data for each of the sensors.

For that, a method that combines Excel with R software is created for data processing. To begin with, Excel is used to display the raw data series from SCADA and to pre-process them. After that, pre-processed data series are included in a R-script which plots the data in the form of graphs, according to the following procedure:

- Firstly, each time data series with cumulative values are displayed in scatter plots according to the time record order, where gaps refer to the missing data. These plots enable us to evaluate the quality of each raw time data series and to calculate the percentage of valid data.
- Secondly, the valid data of each time data series that contain daily values are displayed in boxplots using the interquartile range (IQR), which accounts the distance between the Q_1 first and Q_3 third quartile as follows:

$$IQR = Q_1 - Q_3 \quad (1)$$

- Data points outside the boundary of the boxplot's whiskers are taken as outliers, with the following limit values:

$$Outlier_{lower} < Q_1 - 1.5 \cdot IQR \quad (2)$$

$$Outlier_{upper} > Q_3 + 1.5 \cdot IQR \quad (3)$$

- Finally, the median value, the Q_1 and Q_3 values, and the range of each daily data series are calculated.

III. RESULTS

In this section, the graphical and numerical results are showed.

A. *Results for cumulative values*

As mentioned in the previous section, scatter plots are used to depict the raw time data series for the meters: C7, C8, C16, C17, C26, C28, C31, C32 and C34, with 366 data points each. The first data point belongs to June 1st, 2019, and the last one to May 31st, 2020. Besides, heating season data (from October 15th, 2019, to May 14th, 2020) is represented inside the shadow area; and missing data, due to loss of connection between the meters of thermal facility and the monitoring system, is observed as gaps in the plots (Fig. 3, 4 and 5). The following conclusions are obtained from the plots:

- Fig. 3 shows the scatter plots obtained for the meters C7 (DHW distribution from GSHP), C8 (DHW distribution from boiler), C16 (heating distribution from GSHP) and C17 (heating distribution from boiler) located in the distribution circuit. The valid data points for each one is 84.4% of the total data points and the following trends are distinguished in the data distribution according to the sensor:
  - Since C16 and C17 are the sensors used to measure the heating demand, their slope in the data distribution increase in the shaded heating season.
  - It is also observed that heating is mainly produced with the GSHP (C16) compared to the boiler heating production (C17). The difference between the maximum and the minimum values for C16 and C17 data series are 31,436 kWh and 3,596 kWh

respectively as showed in Fig. 3.

- Besides, a light slope is observed in the data distribution for C16 out of the heating season that could be due to faults in the meter, or even faults in the distribution elements of the primary circuit (see numbers 13 and 14 in Fig. 2).
- The sensors that measure DHW demand produced by GSHP (C7) and boiler (C6) follow a linear trend over the year, showing that both equipment participate in supplying DHW.After performing the energy calculations explained in the previous section, the result of the energy demands and consumptions are shown together with the CO2 emissions of the building in its initial state, comparing them with the same indicators of the case relative to the improvement proposed in the building's thermal installations. This comparison can be seen in Table 3.
- Fig. 4 shows the scatter plots obtained for C26, C28 and C34 located in the generation circuit. Accordingly, the valid data points for C26 (GSHP) and C28 (boiler) are 84.4% of the total data points and for C34 (underground heat exchanger) 62.6%. In addition, and because of evident reasons, a higher slope in the data distribution is observed in the heating season.
- Finally, Fig. 5 shows the scatter plots obtained for the meters C31 (DHW consumption) and C32 (heating consumption) located in the final consumption circuit. The valid data points for each one is 84.4% of the total data points. As it is common, the slope in the data distribution is almost constant for C31 and increases only during the heating season for C32.

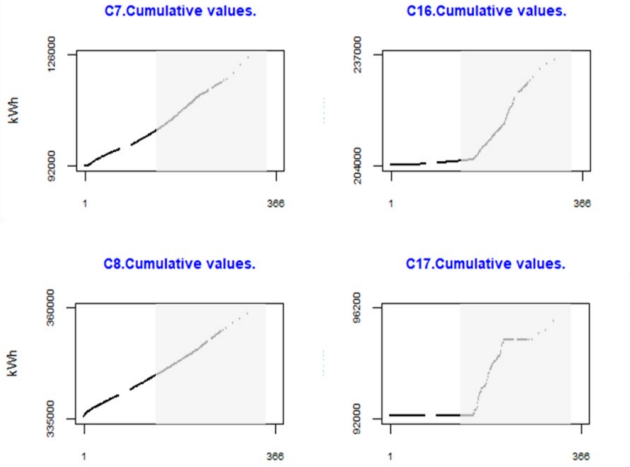

Fig. 3. Scatter plots for C7, C8, C16 and C17.

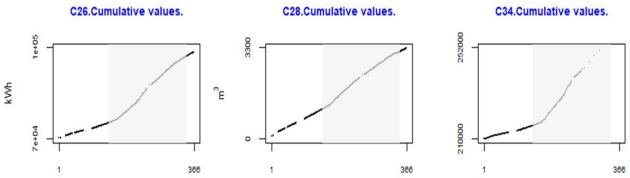

Fig. 4. Scatter plots for C26, C28 and C34.

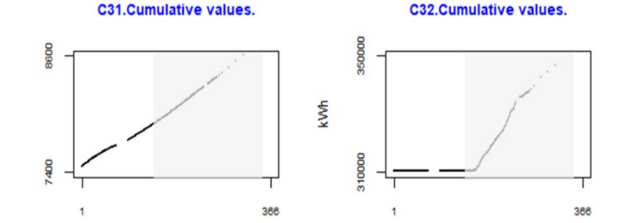

Fig. 5. Scatter plots for C31 and C32.

B. *Results for daily values*

As mentioned in the methodology section, the valid daily data for the meters, C7, C8, C16, C17, C26, C28, C31, C32 and C34 are depicted in boxplots, where outliers are represented with circumferences (Fig. 6, 7, 8). The median value, the Q1 and Q3 values and the range for each daily data series are also included (Table 2, Table 4, Table 5). The following trends are observed:

- Fig. 6 and Table 2 show the results of C7, C8, C16 and C17 meters of the distribution circuit; all datasets have outliers.
- Fig. 7 and Table 3 show the results for the C26, C28 and C34 meters of the generation circuit as can be seen only C28 and C34 datasets have outliers.
- Fig. 8 and Table 4 show the results for the C31 and C32 meters of the consumption circuit; it is noted that only C31 dataset has outliers.

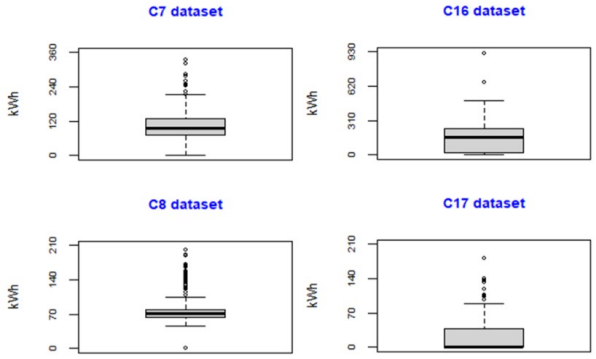

Fig. 6. Boxplots for C7, C8, C16 and C17.

TABLE II
STATISTICAL PROPERTIES FOR C7, C8, C16 AND C17 DATASETS

	UNITS	MEDIAN	Q1	Q3	OUTLIER		RANGE
C7	kWh	97	72	129	>214.5	-	0-334
C8	kWh	71	63	80	>105.5	<37.5	0-201
C16	kWh	163.5	19.5	234.5	>557	-	0-918
C17	kWh	0	0	38	>95	-	0-181

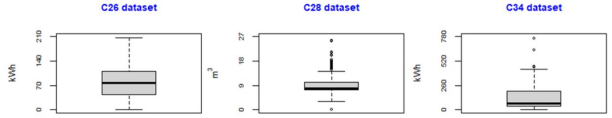

Fig. 7. Boxplots for C26, C28 and C34.

TABLE III
STATISTICAL PROPERTIES FOR C26, C28 AND C34 DATASETS

	UNITS	MEDIAN	Q1	Q3	OUTLIER		RANGE
C26	kWh	77.2	44.3	111.5	-	-	**C26**
C28	m^3	8	7.3	10.2	>14.5	<2.9	**C28**
C34	kWh	68	40	201	>442.5	-	**C34**

C26 and C32 do not have outliers. For these datasets there are no values that exceed the lower and upper limits indicated by the whiskers. Therefore, the values for C26 and C32 datasets are between the ranges indicated in Table 3 and Table 4.

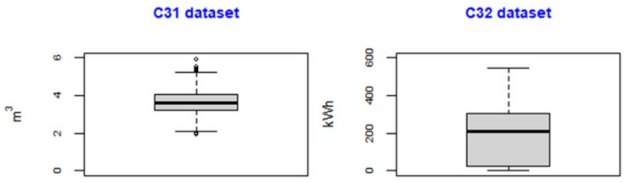

Fig. 8. Boxplots for C31 and C32.

TABLE IV
STATISTICAL PROPERTIES FOR C31 AND C32 DATASETS

	UNITS	MEDIAN	Q1	Q3	OUTLIER		RANGE
C31	m^3	3.6	3.2	4	>5.3	<2	**C31**
C32	kWh	208.6	28.8	303.8			**C32**

C26 and C32 do not have outliers. For these datasets there are no values that exceed the lower and upper limits indicated by the whiskers. Therefore, the values for C26 and C32 datasets are between the ranges indicated in Table 3 and Table 4.

IV. CONCLUSIONS

Obtaining proper data series, free of missing values and outliers is the first step for any application that uses data series. For example, data-mining techniques application and data-driven models construction are used for solving FDD problems in building thermal facilities. Unfortunately, monitoring systems with large amount of sensor are still rare in building thermal facilities but they will become more widespread in the near future, due to the lower price and the versatility of these products. This work describes a methodology to provide suitable raw time data series obtained from a SCADA, using advanced data analysis tools.

According to the results, the time data series of C7, C8, C16, C17, C31, C32 and C34 meters, have a 62.6% valid data and 84.4% for C26 and C28 sensors. Besides, in the generation system, the median for daily values of the GSHP electrical energy and geothermal energy consumption is 77.2 kWh and 68 kWh respectively; meanwhile, the median for boiler daily gas consumption is 8 m3. According to the median of the daily values obtained for the time data series of meters C7, C8, C17 and C16, which correspond to the heating and DHW circuits, the highest values belongs to C7 and C16 (GSHP branches), being 97 kWh and 163.5 kWh respectively, since more energy is supplied with the GSHP than with the boiler. In the consumption points, the median for daily values of the time data series of the meters C31 (DHW) and C32 (heating) are 3.6 m3 and 208.6 kWh respectively. Besides, for each time data series, valid ranges are established and the outliers are identified.

The procedure applied is suitable to get the goal of this work. Future work should include the increase of the database for further FDD application in the facility.

AGRADECIMIENTOS

The authors acknowledge the support provided by the Laboratory for Quality Control in Buildings of the Basque Government.

REFERENCES

A. Afram and F. Janabi-Sharifi, ‘Black-box modeling of residential HVAC system and comparison of gray-box and black-box modeling methods’, Energy Build., vol. 94, pp. 121–149, 2015, doi: 10.1016/j.enbuild.2015.02.045.

M. W. Ahmad, M. Mourshed, B. Yuce, and Y. Rezgui, ‘Computational intelligence techniques for HVAC systems: A review’, Build. Simul., vol. 9, no. 4, Art. no. 4, 2016, doi: 10.1007/s12273-016-0285-4.

ASHRAE, ‘ASHRAE Guideline 14–2014, Measurement of Energy and Demand Savings’. 2014.

BEST, Bilbao Energy Solutions Trends, ‘Heating installation and centralized DHW production in a building of 26 social housing units in Durango. Execution project.’ 2012.

Boyer, S. A., ‘SCADA: supervisory control and data acquisition. International Society of Automation’. 2009.

European Union, Consolidated version of the Treaty on the Functioning of the European Union. PART THREE - UNION POLICIES AND INTERNAL ACTIONS TITLE XX - ENVIRONMENT Article 191 (ex Article 174 TEC). 2016. [Online]. Available: http://data.europa.eu/eli/treaty/tfeu_2016/art_191/oj

S. A. Kalogirou, ‘Artificial intelligence for the modeling and control of combustion processes: A review’, Prog. Energy Combust. Sci., vol. 29, no. 6, pp. 515–566, 2003, doi: 10.1016/S0360-1285(03)00058-3.

S. Katipamula and M. R. Brambley, ‘Review article: Methods for fault detection, diagnostics, and prognostics for building systems—A review, part I’, HVAC R Res., vol. 11, no. 1, pp. 3–25, 2005, doi: 10.1080/10789669.2005.10391123.

W. Kim and S. Katipamula, ‘A review of fault detection and diagnostics methods for building systems’, Sci. Technol. Built Environ., vol. 24, no. 1, pp. 3–21, 2018, doi: 10.1080/23744731.2017.1318008.

A. Kylili, P. A. Fokaides, and P. A. L. Jimenez, ‘Key Performance Indicators (KPIs) approach in buildings renovation for the sustainability of the built environment: A review’, Renew. Sustain. Energy Rev., vol. 56, pp. 906–915, 2016, doi: https://doi.org/10.1016/j.rser.2015.11.096.

M. Lumbreras, R. Garay, and A. G. Marijuan, ‘Energy meters in district-heating substations for heat consumption characterization and prediction using machine-learning techniques’, in IOP Conference Series: Earth and Environmental Science, 2020, vol. 588, no. 3. doi: 10.1088/1755-1315/588/3/032007.

M. S. Mirnaghi and F. Haghighat, ‘Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review’, Energy Build., vol. 229, 2020, doi: 10.1016/j.enbuild.2020.110492.

R software. 2022. [Online]. Available: https://www.r-project.org/

The Spanish Government´s Institute for Diversification and Saving of Energy (IDAE), ‘Studies, reports, and statistics’, 2018. http://sieeweb.idae.es/consumofinal/bal.asp?txt=2018&tipbal=t

H. Yang, T. Zhang, H. Li, D. Woradechjumroen, and X. Liu, 'HVAC Equipment, Unitary: Fault Detection and Diagnosis', Encyclopedia of Energy Engineering and Technology. Taylor & Francis, Ed. CRC Press, pp. 854–864, 2014.

R. Yang and G. Rizzoni, 'Comparison of model-based vs. data-driven methods for fault detection and isolation in engine idle speed control system', in Proceedings of the Annual Conference of the Prognostics and Health Management Society, PHM, 2016, vol. 2016-October, pp. 25–33.