A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings [α]

Marco Savino Piscitelli [1,2], Silvio Brandi [1], Alfonso Capozzoli [1,], Fu Xiao [2]*

1. Department of Energy "Galileo Ferraris", TEBE research group, Politecnico di Torino, Italy

2. Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Abstract

In this paper, a tool for the detection and diagnosis of anomalous electrical daily energy patterns relative to a transformer substation of a university campus was developed and tested. Through an innovative pattern recognition analysis consisting in a multi-step clustering process, six clusters of anomalous daily load profiles were identified and isolated in two-year historical data of total electrical energy consumption. The infrequent electrical load profiles were found to be strongly affected, in terms of both shape and magnitude, by the energy consumption behaviour related to the heating/cooling mechanical room. Then, a fault-free predictive model, which uses Artificial Neural Network (ANN) in combination with a Regression Tree, was developed to detect anomalous trends of the electrical energy consumption. The model was able to detect the 93.7% of the anomalous profiles and only the 5% of fault-free days were wrongly predicted as anomalous. Eventually, a diagnosis phase was conceived and validated with a testing data set. A number of daily abnormal load profiles were detected and compared with the centroids of the anomalous clusters identified in the pattern-recognition stage. The work led to the development of a flexible intelligent tool useful for operating a continuous commissioning of the campus facilities.

[α] Published in

Piscitelli, M. S., Brandi, S., Capozzoli, A., & Xiao, F. A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings. In Building Simulation (pp. 1-17). Tsinghua University Press.

Keywords: anomaly detection; data analytics; energy management; pattern recognition; prediction models

1. Introduction

The importance of enhancing energy management in buildings has been recognised as a crucial aspect for optimising their energy performance during operation. The attention given to the automation, control and monitoring systems is growing also at European level, since the Energy Performance of Building Directive (EPBD) 2010/31/EU encourages the installation of intelligent metering systems to achieve significant energy savings. In this context the process of Fault Detection and Diagnosis (FDD) allows energy inefficiencies to be promptly detected during building operation.

The growing penetration of Building Management Systems (BMS) in buildings offers the opportunity to exploit analytical methods (Yu et al., 2013) to determine whether a fault occurs under specific boundary conditions and identify the most probable cause. In the last few years, the data-driven approach gained more and more interest, thanks to its applicability even in the case where engineering models of the building and systems are inadequate or difficult to develop. In this context, particularly promising is the application of data mining techniques which include both supervised and unsupervised algorithms. In literature many examples are reported on the application of data analytics procedures for performing FDD at different scales of investigation, from whole building to component level (Katipamula et al., 2017).

The FDD approach, based on the analysis of historical monitoring data, generally consists in extracting typical patterns representative of the relation between disturbances and energy demand and then detecting deficiencies when those energy use patterns are violated over time (Capozzoli et al. 2018). The implementation of prediction models of different nature or their combination makes it possible to compare the expected and the

* corresponding author: Alfonso Capozzoli, e-mail address: alfonso.capozzoli@polito.it

actual energy consumption over time and to detect abnormal energy patterns in quasi real-time (Ku & Jeong, 2018; Fan et al., 2018). Data analytics-based methodologies proved to be particularly effective for the automatic detection and identification of operation strategies (Qui et al. 2018) and faults in building systems (Miller et al., 2018).

Most of the FDD frameworks conceived at system/component level focused on the operation of Heating Ventilation and Air Conditioning (HVAC) systems. The optimal management and the on-line Automated Fault Detection and Diagnosis (AFDD) of HVAC systems are promising applications to investigate (Ahmad et al. 2016), considering that such systems account for 50% of the energy demand in commercial buildings (DOE, 2012). Li et al. (2016) investigated the performance of a tree-structured learning method of a building cooling system, with the aim of detecting both fault type and its severity. In (Du et al., 2014) an effective diagnostics tool was introduced to improve the energy efficiency and thermal comfort of buildings by detecting faults. Clustering analysis was used to create a library of faults, while Back Propagation Neural Networks were exploited to determine if a fault was occurring by measuring the relative error with respect to the predicted data. Liang et al. (2007) proposed a hybrid method that merges a model-based approach with SVM (Support Vector Machine), to exploit the accuracy of the system model and the classification capability of faults through SVM. During the on-line phase a number of faults were detected by means of the residual analysis (i.e., comparing residuals with pre-set thresholds). The selection of thresholds represents an issue to face, with the aim of searching a trade-off between the increase of the probability to detect faults and the occurrence of false alarms.

In (Han et al., 2011a, Han et al., 2011b) an AFDD methodology for a centrifugal chiller was proposed. In detail five types of faults at different levels were considered. An Artificial Neural Network (ANN) was employed for estimating the system reference

behaviour. The faults were diagnosed with a multi-layer SVM classifier based on residual rules by comparing actual data of the chiller system with the reference model estimations.

Data-driven based fault detection was also extensively addressed to detect AHU faults (Yan et al., 2016). Dehestani et al. proposed a methodology to detect faults related to the failure of fans and the air dampers of an Air Handling Unit (AHU). The methodology used a Multi-Class Support Vector Machine (MC-SVM) for detecting both pre-labelled and new anomalies during system operation (Dehestani et al., 2013). In (Zhao et al., 2015; Zhao et al., 2017) a Bayesian Network (BN) was adopted for the diagnosis of faults related to the positions of air dampers, cooling coil valve and the failure of the return fan. The BN diagnoses faults using as input the residuals obtained from a set of regression models capable to estimating air temperature, water flow rate, air flow rate and fan power energy consumption.

Despite most of the FDD literature is focused on system/component level, research activities were also conducted at a larger scale following a top-down approach (e.g., analysis of whole building energy consumption to discover component or system faults).

In fact, in most of real cases, just few and aggregate variables related to energy consumption of a building are monitored and collected (i.e., high level metering). Improving the building energy performance by analysing aggregate data is challenging, especially if several factors such as occupant behaviour, comfort levels, and operation schedules of systems can generate the existence of different energy consumption patterns, not always easily inferable.

The automated detection of anomalous energy patterns is currently a topic of great importance. In fact, the development of a predictive-based management analysis methodology considering a daily time scale is particularly desirable (Capozzoli et al.

2017). The development of innovative robust FDD methodologies at whole building level to automatically detect anomalous energy consumption (profiles with shape/magnitude significantly different from the typical operation patterns) (Capozzoli et al., 2015) makes a sort of continuous commissioning of the building possible, also supporting the definition of easy and interpretable recommendations to be considered in the optimisation of high-level control logic.

The definition of a robust framework, based on pattern recognition techniques to handle inconsistent, infrequent, abnormal energy use in buildings/systems is a crucial aim to pursue (Piscitelli et al. 2018). In this perspective, (Capozzoli et al. 2018) introduced a novel framework based on an effective transformation process of whole building energy consumption time series. Data reduction, transformation and machine learning methods (i.e. decision trees, symbolic aggregate approximation) were coupled and applied to discover unexpected patterns in electrical energy consumption data. The identification of infrequent/anomalous patterns was performed by comparing the expected and the actual daily energy consumption profile. The prediction task was accomplished for five sub-daily time windows, considering as input parameters the indoor/outdoor climatic conditions, occupants' presence and temporal variables.

Miller et al. (2015) proposed a process based on Simbolic Aggragate approXimation (SAX) of time series to reduce and transform daily load profiles of buildings. SAX was used for discovering motifs and a clustering analysis was carried out to group together infrequent load profiles (i.e., discords). Also Fan et al. (2015) used SAX for conducting analyses on load profiles. In that case, the motif discovery was coupled with the extraction of Temporal Association Rules in order to mine temporal relations in energy consumption time series. The extracted knowledge, supported by domain expertise, was helpful in detecting typical patterns and at the same time anomalies,

making the identification of effective energy conservation opportunities possible. Such advisory tools can be easily translated in a set of decision rules and embedded in a DSS (Decision Support System), helping thus stakeholders in the early detection of anomalous energy patterns and in the avoidance of further energy waste over time.

In this paper an FDD methodology at whole building level is introduced by analysing the energy consumption of an electrical substation of Politecnico di Torino, a university campus located in the northern Italy. A tool capable to detect anomalies and rare events was developed to improve the energy management and to increase users' awareness about the behaviour of the campus. The proposed methodology relies on the application of data mining-based algorithms for a robust estimation of the expected whole building energy consumption, and the consequent detection of anomalous load profiles and diagnosis of the most probable associated causes during operation.

On the basis of the literature concerning the detection and diagnosis of anomalous energy daily patterns at whole-building level, the main innovative aspects introduced by the present paper are the following:

• An innovative and scalable process for the identification of fault-free dataset was employed to robustly develop a reference estimation model. A multi-step clustering process was implemented on three different portions of each daily load profile in order to perform an advanced temporal abstraction of the time series and to enable an effective identification of infrequent patterns. The adopted unsupervised algorithm for grouping similar load profiles belongs to the family of partitional clustering techniques, but it requires as input parameter a distance threshold, instead of the number K of desired clusters. The methodology introduced in this paper sets the optimal value of this threshold by minimising a validity metric, ensuring a very flexible and automatic clustering process. Moreover, differently from other papers, the anomalous load profiles were identified by

analysing the frequency of text string obtained by concatenating the three cluster labels of the sub-profiles which belong to the same day.

• A fault-free reference estimation model consisting in an ANN combined with a regression tree was tested and validated. The model uses easy-to-collect variables that can be gathered from external web services (i.e., weather data) and high-level sensors. More in detail, one of the inputs of an ANN was provided as response of a regression tree model that can be translated in a set of interpretable patterns in form of IF-THEN rules. Such approach made the output of the reference model more understandable by increasing the interpretability of the building energy behaviour.

• The tuning of the parameters of the anomaly-detection tool was completely automatic and conceived for minimising the rate of potential false alarms. A process aimed at searching the optimal number of hourly anomalies per day was introduced in order to alert the detection of an anomalous daily energy trend, minimising the occurrence of false-positive values. In addition, a sensitivity analysis was performed for searching the optimal hyperparameter configuration of ANN model. This analysis allows the configuration with the minimum Mean Absolute Percentage Error (MAPE) to be identified automatically.

• The diagnosis process was conceived to be open and upgradable over time. In particular, when a new anomalous event is detected, the historical-based fault library is automatically updated. This opportunity allows the developed tool to evolve during building operation significantly increasing its generalizability.

In that perspective, this paper proposes and discusses an anomaly detection and diagnosis framework conceived for whole-building level applications.

The rest of the paper is organised as follows. Section 2 presents the case study under analysis. Section 3 describes the methodology adopted. Section 4 presents a brief

theoretical description of the methods used to perform the analysis. The last two sections present and discuss the results obtained.

2. Case study

The analysed dataset refers to the energy consumption of a part of the main campus of Politecnico di Torino (PoliTo), an Italian university located in the Piedmont region (Figure 2 (a)). The headquarter of Politecnico di Torino has about 200.000 m^2 of floor area available for lectures, research activities and services. The campus is equipped with an electrical power station that is composed by a loop of ten transformer substations. In this paper, the electrical energy consumption data related to one substation (i.e., substation C) is considered for the analysis. Anomalous daily energy patterns were detected by analysing the total energy consumption of the whole substation also exploring the relation between unexpected trends of the total electrical load and the energy behaviour of the services served by the substation.

Figure 1(a) shows the hourly boxplots of the electrical load of the substation C across different day types (i.e., Weekdays, Saturdays, Sundays and Holidays) for the period included between 1 January 2016 and 31 December 2018. As expected, the boxplots referred to unoccupied hours (i.e., night hours, Sundays and holidays) are smaller than the boxplots of working hours, suggesting the presence of building systems that have different energy patterns during the year. In particular, Figure 1(b) shows the daily load profiles of weekdays, Saturdays, Sundays and holidays (with their average profiles) for 3 months belonging to winter, summer and mid-season respectively. From Figure 1(b) it can be inferred that the shape of the daily load profiles of the substation C does not significantly change during the year, while its magnitude is affected by a seasonal effect with an increase during the summer season. This behaviour is typically due to thermal

sensitive electrical loads that in this case study are related to the operation of a heating/cooling mechanical room served by the Substation C.

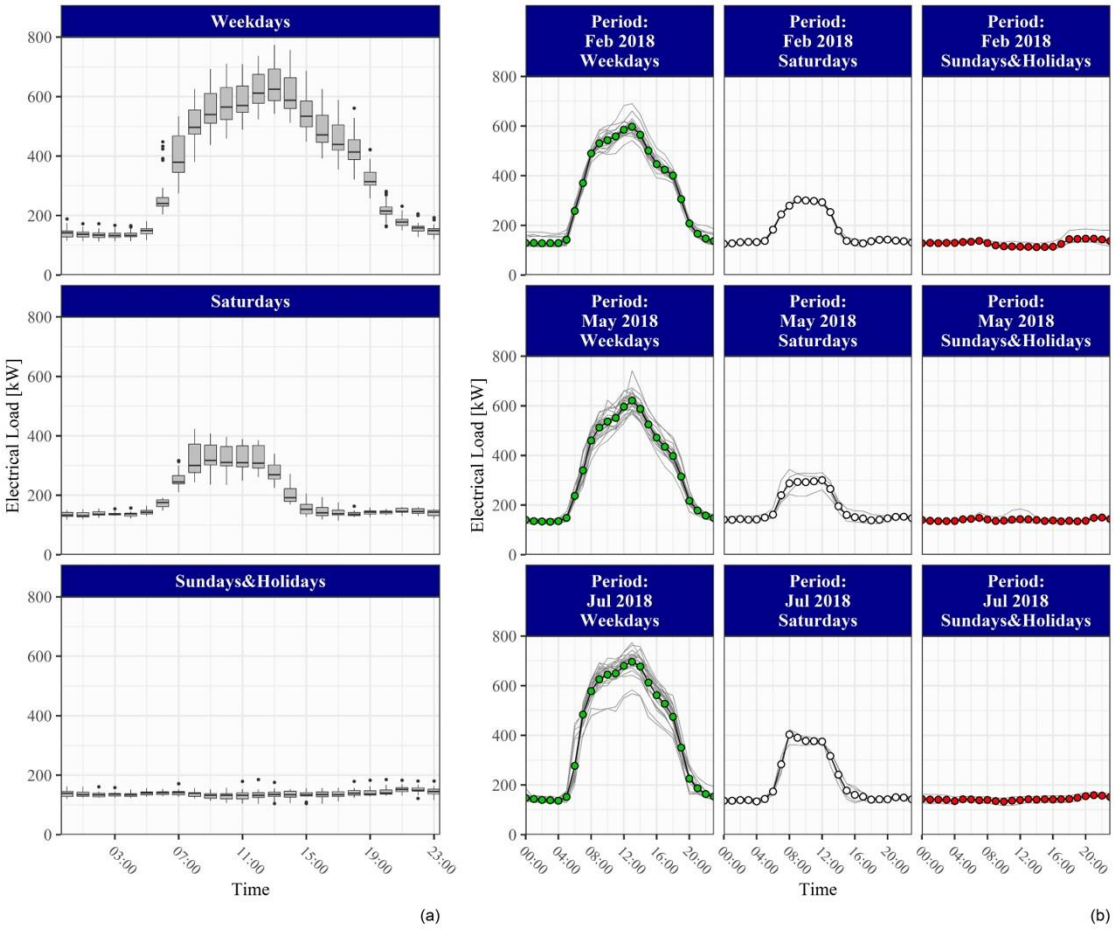

Figure 1. *Hourly boxplots of the electrical load of the substation C across different day types (i.e., Weekdays, Saturdays, Sundays and Holidays) (a) Daily load profiles of weekdays, Saturdays, Sundays and holidays (with their average profiles) for 3 months belonging to winter, summer and mid-season respectively (b)*

In more detail, the analysed dataset includes the electrical power demand of the whole substation and part of its main services such as heating/cooling mechanical room, a canteen, the department of mathematics, the data centre and the administration building. However, the energy consumption related to some of these sub-loads (e.g., canteen, data centre) show flat or very repetitive patterns over the year and are not thermal sensitive.

On the other hand, particular attention was paid on the heating/cooling mechanical room, considering that it accounts on average for about the 14% of the annual electrical energy demand of the substation (Figure 2 (b)) and it is characterised by an energy consumption patterns with strong seasonality (justifying the trend of daily load profiles in Figure 1 (b)). The equipment located in the mechanical room include parts of the hot and chilled water circuits of the campus with the corresponding auxiliary pumping systems. The hot water is provided through a district-heating heat exchanger located in a separate area of the campus. The chilled water, instead, is produced by two chillers with nominal electrical power of 220 kW and a rated cooling capacity of 1120 kW, and a reversible water-water heat pump, with nominal power and cooling capacity of 165 kW and 590 kW, respectively.

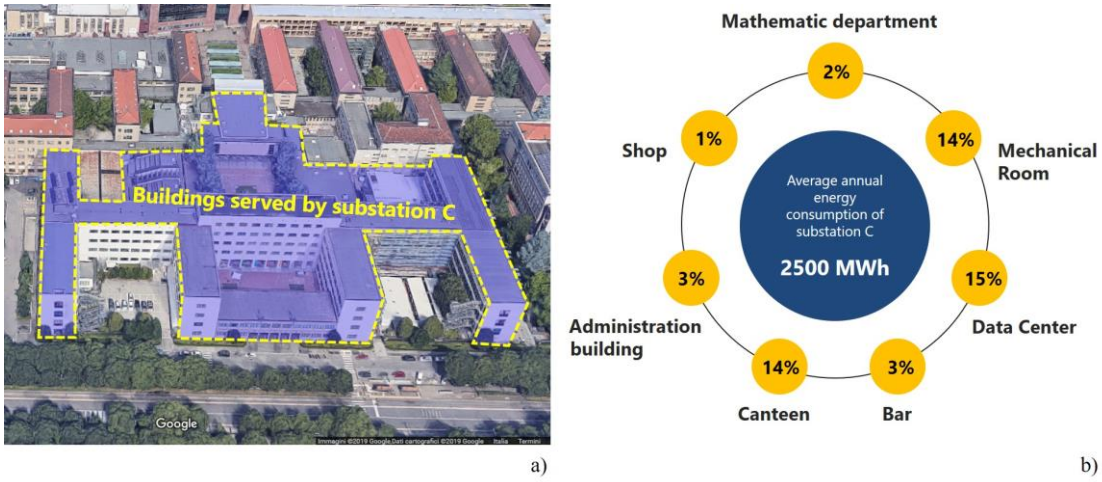

Figure 2. *Geospatial identification of the buildings served by the substation under analysis (a) breakdown of the annual energy consumption of the substation C (b)*

The electrical energy consumption data of the substation are available with a timestamp of 15 minutes from January 1, 2016 to July 25, 2019. In addition, also data referred outdoor air temperature measurements were analysed for better explaining the variation of the energy consumption patterns due to the presence of thermal sensitive electrical loads (i.e., energy consumption of chillers).

3. Methodology

The process is primarily aimed at developing a robust fault-free estimation model of the hourly energy consumption of the substation C. The FDD process was performed by analysing the hourly difference between the energy consumption estimated by a fault-free inverse model and the actual energy consumption during building operation. The methodology process was conceived as multistep data analytics procedure and unfolds over different stages, as shown in Figure 3.

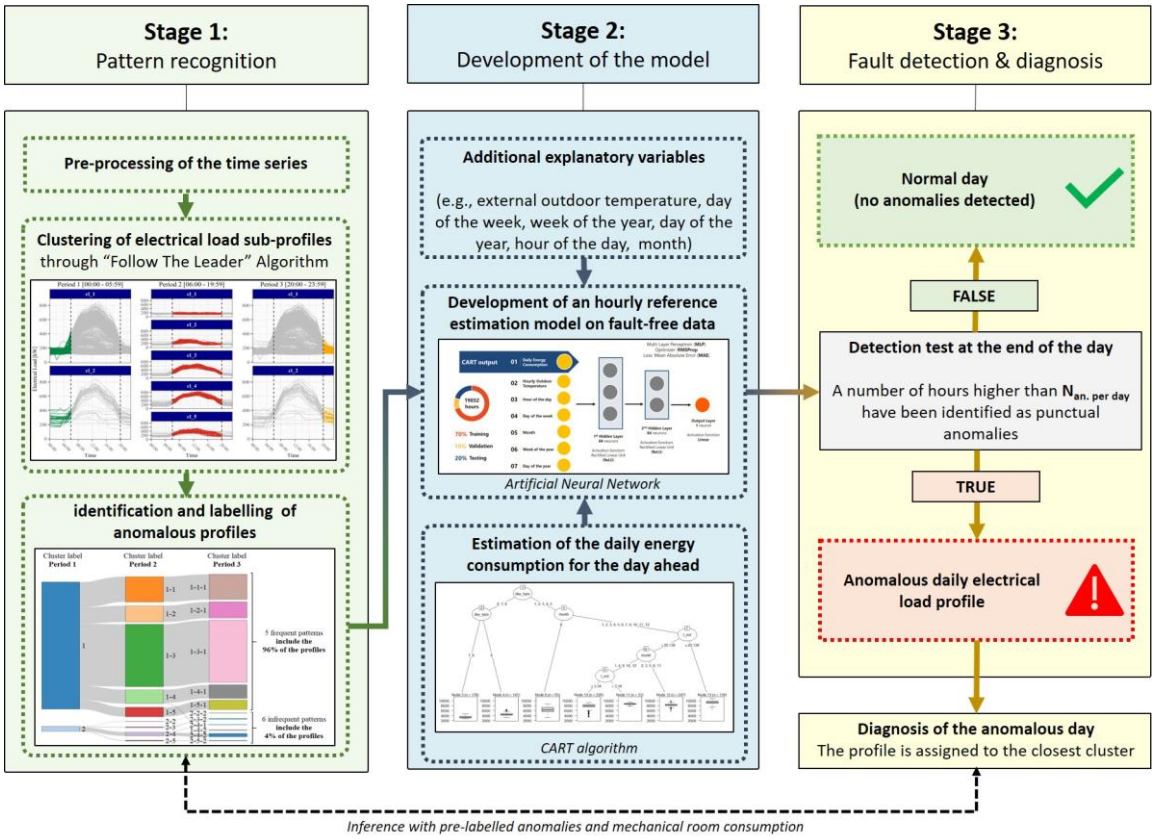

Figure 3. *Methodological framework of the analysis*

The stage 1 of the process is aimed at improving the quality of the data and at obtaining a dataset free of infrequent and anomalous load profiles. To this purpose, a pre-processing task was preliminary accomplished for detecting and replacing punctual

inconsistencies (e.g., outliers, missing values). Missing values in time series were replaced by means of a linear interpolation; moreover, when more than 3 consecutive hours included missing values, the entire load profile of the day was filtered out. The detection of outliers and their replacement were performed by means of the Hampel filter method (Pearson 1999). For each data point in the time series, the algorithm computes the median of a window that includes the considered data point and its k surrounding samples. If a data point differs from the median by more than a standard deviation, it is tagged as a statistical outlier and, for that case study, replaced through a linear interpolation. Successively, a pattern recognition process was developed for isolating the anomalous daily profiles of energy consumption in the dataset. To this purpose, the daily load profiles included in the training period (between 01-01-2016 and 31-12-2018), were analysed. Firstly, the 15-minutes measurements were aggregated in order to obtain an hourly dataset; successively data were chunked in hourly daily electrical load profiles. Then a pattern recognition analysis was performed for the identification of infrequent daily load profiles. To that purpose, a clustering analysis was performed by using a partitive algorithm based on the "Follow the Leader" approach (Chicco et al. 2006). The pattern recognition stage was carried out for three representative time periods of a day. For each period, a clustering analysis was performed separately. At the end of this process, for each load profile, three cluster labels (one for each period) were available. In order to identify the infrequent load profiles, a frequency analysis was performed among the sequences of cluster labels during each day. All the load profiles filtered out from the database were further analysed with the aim of creating a library of labelled anomalous events, to be used in the successive diagnostic phase. The identification of the causes related to the anomalies detected at whole substation level are associated to the sub-load

that has been found as the major contributor to the occurrence of infrequent energy patterns.

In the second stage, a prediction model was trained on the fault-free dataset obtained in the first stage. The model consists in a non-autoregressive Multi-Layer Perceptron ANN, combined with a regression tree.

In the last stage of the process an anomaly detection strategy was developed. The anomaly detection is enabled by comparing the predicted and the actual hourly electrical energy consumption of the substation, in order to identify the presence of anomalous events. A punctual anomaly is detected if the hourly difference between the estimated and actual energy consumption is higher than three times the standard deviation (Chou et al. 2014). At the end of each day, if at least a certain number of hours ($N_{an.\text{-}per\text{-}day}$) have been identified as punctual anomalies the entire day is supposed to have an anomalous energy trend. A sensitivity analysis to evaluate the optimal value of $N_{an.\text{-}per\text{-}day}$ per day was performed.

Furthermore, the entire procedure was applied on an independent set of data (days from 01-01-2019 to 25-07-2019) for evaluating its capability in diagnosing fault causes. In the diagnosis phase, the actual load profiles detected as abnormal were compared with the centroids of the anomalous clusters previously identified in the pre-processing stage (i.e., first stage of the framework). The detected anomalous daily load profiles were then assigned to clusters with the closest centroids and the corresponding anomalous events (a-priori labelled) were assumed as their most probable causes.

In order to make the process as flexible as possible, the profile assignment to a pre-identified cluster was performed considering a proximity distance threshold. Specifically, the detected anomalous profile is assigned to the cluster of its closest centroid only if the distance between these two profiles is lower than a fixed threshold

otherwise a new cluster is generated. In the latter case, the new cluster is labelled by the analyst and represents a new possible fault cause, to be included in the diagnosis library.

4. Methods of the analysis

In this section, a brief theoretical description is given for the aforementioned methods (e.g., FTL clustering algorithm, classification and regression tree, multilayer perceptron neural network), used to perform the anomaly detection process. The analytics methods are also discussed with the aim of better specifying the advantages they offer in relation to the objectives of this work and for introducing the main hyperparameters which were set during the training phase.

4.1 The "Follow the leader" clustering algorithm

In this study an automatic clustering procedure based on the "Follow The Leader" (FTL) approach was employed for grouping similar profiles during the pattern recognition stage (Chicco et al. 2005; Piscitelli et al. 2019). Differently from other partitive clustering algorithms (e.g., K-means) the FTL does not require the a-priori definition of the number of clusters K, but it is initialized selecting a maximum distance threshold ρ^* between the cluster centroid and the objects in the same cluster. The algorithm consists in a sequential scan of object database (i.e., daily load profiles) over a number n of iterations, large enough to ensure the stabilization of the clustering results. In the first iteration, the FTL approach defines, as a first attempt, the total number of clusters K and the number of load profiles that are assigned to each cluster. The process is then iterated and if the distance between a load profile and the cluster centres computed until that iteration is lower than the threshold ρ^* the load profile is assigned to the cluster of the closest centroid, otherwise a new cluster composed by a single element is

generated. As a consequence, during the n iterations the number of clusters and the load profiles belonging to them may change until a stable configuration is reached.

The algorithm was implemented from scratch in the statistical software R (R core team, 2017).

Given that the parameter ρ is a-priori set by the analyst, a sensitivity analysis is required for supervising its tuning. In this study, the identification of the optimal value of ρ is conducted though a "trial-and-error" process: different values of ρ were tested and the results in terms of number of clusters, separation between clusters and cohesion within clusters were computed by analysing the Davies-Bouldin index was (Panapakidis et al. 2018). The Davies-Bouldin Index (DBI) (Davies & Bouldin, 1979) is a cluster validity metric based on the concept that for a good partition, inter cluster separation as well as intra cluster cohesion should be as high as possible. For each value of ρ the clustering results were assessed by calculating the DBI as follows (eq. 1):

$$DBI(\rho) = \frac{1}{N} \sum_{n=1}^{N} \underset{n \neq m}{max} \left(\frac{\sigma_n + \sigma_m}{d_{n,m}} \right) \quad (1)$$

Where:

- N is the number of clusters obtained by fixing a certain value of ρ.
- $d_{n,m}$ is the Euclidean distance between centroids of the clusters C_n and C_m.
- σ_n, σ_m are the standard deviations of the distances of objects in clusters C_n and C_m.

The value of ρ* which minimises DBI was considered as the optimal value of the distance threshold for initialising the FTL algorithm.

4.2 Classification and Regression Tree

Classification and regression trees are machine-learning algorithms that are used to develop descriptive and/or predictive models from a collection of records. Each record

can be expressed as a tuple (**x**,y), where **x** represents the input attribute set while y is the target variable (Tan et al., 2006). The type of target variable distinguishes classification from regression trees (i.e. categorical in the first case, while numerical in the second). In this work, the CART algorithm has been used to conduct a regressive modelling task, using as explanatory attributes both numerical and categorical variables.

The learning approach of the CART algorithm is based on a recursive binary splitting of the records into purer subsets called nodes (Breiman et al., 1984).

The decision trees can be then represented by a hierarchical structure composed by nodes and directed edges (i.e., branches). In particular, for regression trees, the terminal nodes represent the predicted numerical values of the target attribute, while the branches represent the conjunctions of the input attributes that lead to those predictions. The development of a regression tree unfolds over two steps: training and testing of the model. The k-fold cross-validation procedure is used for that purpose. In the present work, a regression tree was developed for providing a robust and interpretable estimation of the daily energy consumption for the day ahead of the substation. The output of the regression tree was then used as input attribute of the neural network for predicting the hourly energy consumption of the substation.

4.3 Multilayer perceptron neural network

Artificial Neural Networks (ANN) are a class of nested mathematical functions that belong to supervised learning algorithms. The main advantage of ANN, also for energy consumption forecasting, consists in their ability of modeling multivariate problems by capturing complex and non-linear relationships between the variables (Gonzales & Zamarreno 2005). Neuron, also referenced as node or unit, is the core computational element of an ANN. It receives a series of inputs (**x**) from other neurons or from external sources and calculates an output. To each input is associated a weight

(**w**) and the weighted sum of the inputs is fed into an activation function (**g**), the output of which correspond to the output of the neuron. In addition, another unitary input called bias with an associated weight (**b**) is fed into the neuron. The output value of each neuron can be consequently calculated as follows (eq. 2):

$$z = g(wx + b) \tag{2}$$

The role of the activation function is to introduce a non-linear mapping between inputs and output of the neuron. Several functions have been introduced in the literature (e.g., hyperbolic tangent, sigmoid). In the present study the REctified Linear Unit (RELU) (Nair & Hinton, 2010) was used as activation function of a feedforward neural network model. In such kind of network, the neurons are grouped into connected layers. All outputs of one layer are connected to each input of the succeeding layer. This architecture is called fully connected. There are three type of layers in a feedforward neural networks: input layer (i), hidden layers (ii) and output layer (iii). The configuration of the output layer distinguishes network models for conducting regression or classification analysis.

The learning process aims at finding a set of neuron weights capable to map the target variable in an accurate way as possible. The performances of the network are evaluated through a loss function. The loss function computes the distance between the output of the network and the actual target output providing a distance score. This score is used as a feedback signal to the network in order to adjust neuron weights. This process was carried out through an optimizer implementing the back-propagation algorithm. In the present study, the RMSProp optimizer was employed (Hinton et al., 2012). Backpropagation is a well-known procedure, and in this paper the tuning of hyperparameters of the Neural Network (i.e. batch size, number of epochs, number of neurons and learning rate) was performed assessing the prediction performances for a set

of their combinations. The combination of parameters leading to the best solution in terms of accuracy for the validation set was used as the configuration of the final model.

5. Results

Following the methodological framework discussed in section 3, which makes use of the methods discussed in section 4, the results of the analyses performed for each stage of the FDD process are presented in this section.

5.1 Stage 1: Pattern recognition

The first stage of the analysis aims at recognising infrequent daily load profiles in the available data set depurated from inconsistencies (i.e., missing values and outliers).

The detection of infrequent daily load profile was performed through an iterative clustering analysis using data from 01-01-2016 to 31-12-2018. In more detail, the "Follow The Leader" algorithm was implemented for three sub-profiles of each daily pattern according to the following time periods selected through the domain experience:

- Period 1: unoccupied period from 00:00 to 05:59 a.m. during which the building energy systems are started up;
- Period 2: occupied period from 06:00 a.m. to 07:59 p.m. during which the building energy systems are operated;
- Period 3: unoccupied period from 08:00 p.m. to 11:59 p.m. during which building energy systems are turned OFF.

For each period, the optimal number of clusters was identified by performing a sensitivity analysis on the distance threshold ρ. In particular, for each period and for each value of ρ, the quality of clustering process was assessed by calculating the DBI index. Considering that the sub-profiles were not normalised, the threshold distance ρ is expressed in terms of kWh. As a consequence, the searching space of ρ can be identified

in an easier and feasible way by the analyst. For this case study the sensitivity analysis was performed considering the minimum value of ρ equal to 20 kWh per hour and the maximum value equal to 100kWh per hour. As a reference, for a sub-profile of four hours the searching space of ρ is included between 80 kWh (i.e., 20 kWh · 4) and 400 kWh (i.e. 100 kWh · 4).

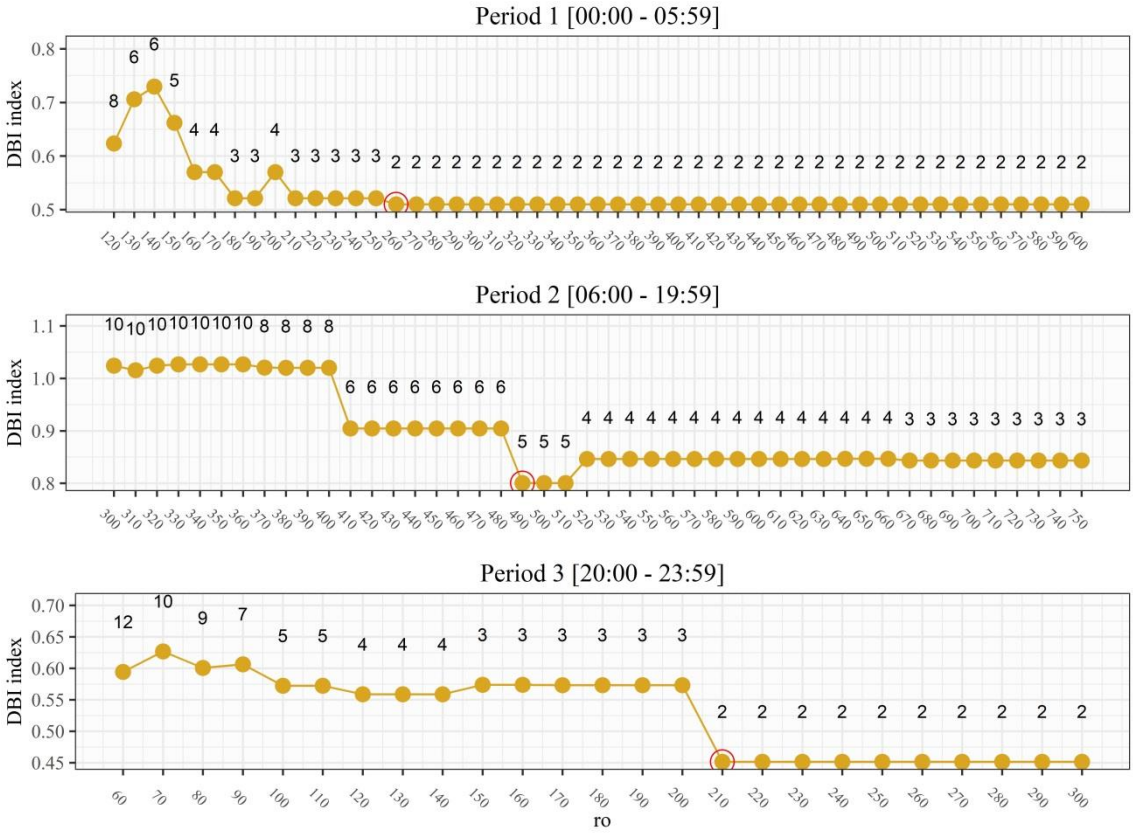

Figure 4. Identification of the optimal number of clusters for each period

The identified searching spaces were mapped incrementing ρ with a constant step of 10 kWh. A further constraint was considered for setting the minimum number of clusters in each period. For period 1 and 3 (i.e., night periods) the minimum number of clusters was set to 2, while for period 2 (i.e., period with occupancy) the minimum number of clusters was set to 3. This choice was driven by the a-priori knowledge on the existence of at least one dominant pattern for Sundays, one pattern for Saturdays and one

pattern for weekdays. The results of the sensitivity analysis for the three periods considered are shown in *Figure 4*.

Figure 4 shows the optimal values ρ* of the parameter ρ (that minimize the DBI) for each period with the corresponding number of clusters. It means that for ρ = ρ* the resulting clusters exhibited optimal inter cluster separation and intra cluster cohesion.

In particular, two clusters were identified as the best partition for period 1 and 3 while sub-profiles in period 2 have been grouped in five clusters. The clusters obtained are characterised by different cardinalities as shown in Figure 5.

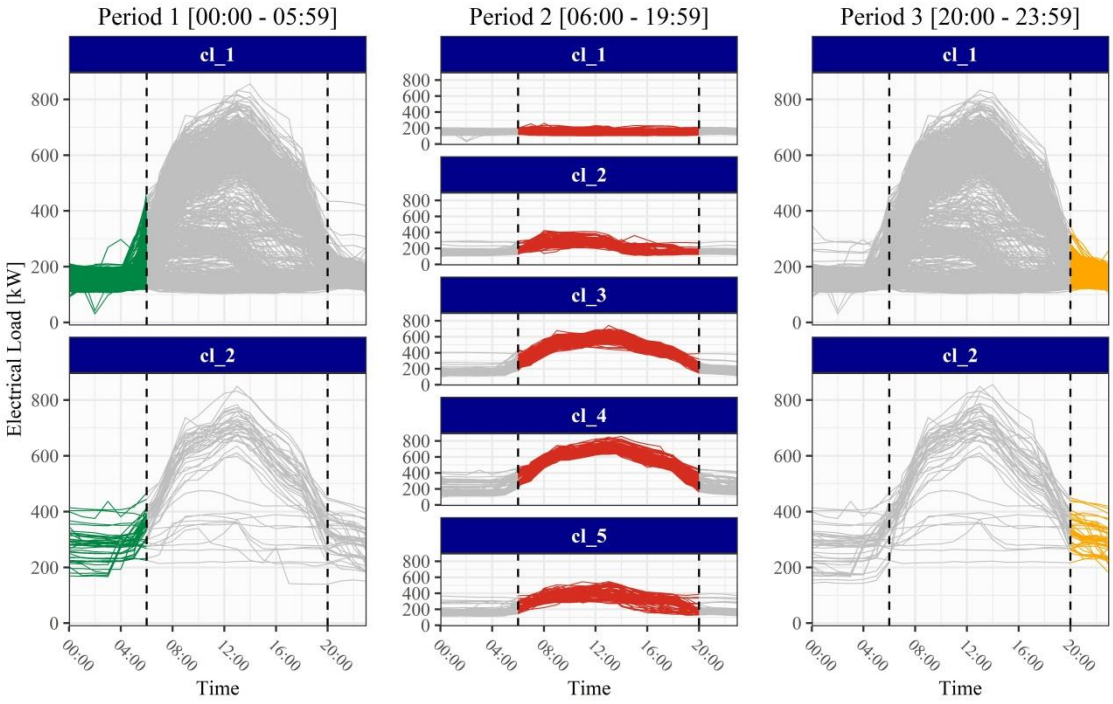

Figure 5. Clusters of sub-profiles obtained for each period

After the clustering phase, a text string obtained by concatenating the three cluster labels of the sub-profiles, belonging to the same day, was associated to each daily load profile. For example, considering a daily load profile whose period 1 is included in the cluster cl_1, period 2 is included in the cluster cl_5 and period 3 is included in the cluster cl_2, the daily sequence was encoded by concatenating the three labels (i.e., cl_1 – cl_5 – cl_2).

In order to identify the infrequent load profiles, a frequency analysis was then performed among the unique sequences identified. As a consequence, even if for period 1 and 3 only two cluster were identified, a cluster label related to a single period cannot explain the existence of an anomalous profile. The same filtering approach was employed by (Miller et al., 2015) for detecting infrequent daily load profiles encoded in symbol strings by means of SAX transformation. In that case the discord candidates were separated according to a threshold of string frequency count equal to 2%.

In this study, eleven unique sequences were found after the concatenation of the three cluster labels. In particular five patterns were identified as frequent (they include about the 96% of load profiles) while six patterns were found as infrequent (each infrequent pattern group includes less than 30 profiles). Figure 6, shows through a Sankey representation, each pattern identified with the evidence of the typical and infrequent ones.

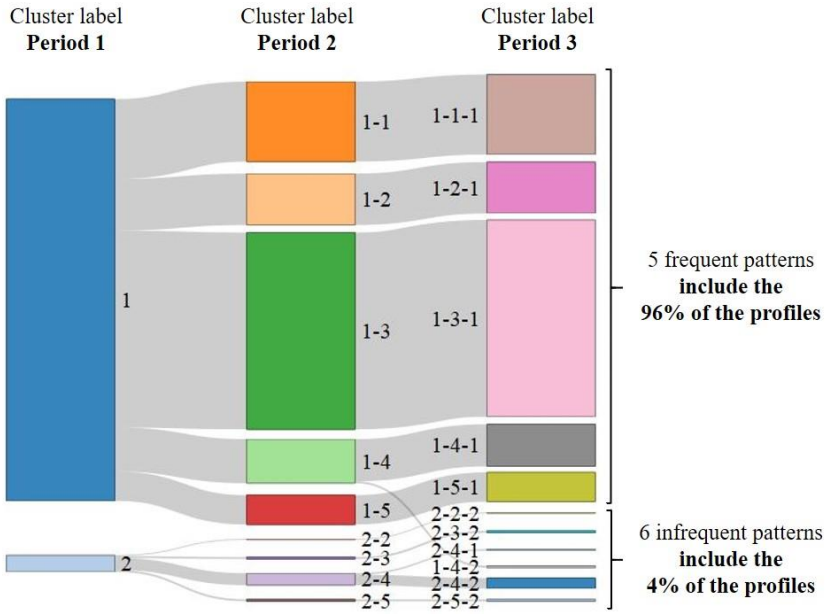

Figure 6.Sankey diagram of patterns identified by concatenating the three cluster labels for each day

The 48 profiles included in the most infrequent patterns were analysed and labelled. In

order to characterise the infrequent profiles, a preliminary analysis was conducted for understanding which sub-load of the substation C majorly contributed to the occurrence of such energy consumption patterns. To this purpose, during days where infrequent load profiles were identified, the Spearman correlation coefficient (Corder, 2014) was calculated for evaluating the strength of the existing associations between the total load of the substation C and each sub-load. From the analysis, the sub-load related to mechanical room resulted to be the highest correlated one ($\rho_{Spearman} = 0.7$). Moreover, for the portion of dataset considered (load profiles detected as anomalous), the mechanical room accounts for about the 28% of the electrical energy consumption of the substation C (the highest between all the sub-loads) almost doubling its average annual impact. In Figure 7 are shown the load profiles related to the electrical demand of the substation C, together with the corresponding average load profile of the mechanical room: all the 6 infrequent patterns were found to be related to an anomalous management of the mechanical room during building operation. Therefore, it can be inferred that the identified infrequent load profiles were strongly affected, in terms of both shape and magnitude, by the energy consumption behaviour of the mechanical room.

The labelling process of the infrequent patterns was then performed according to anomalous events concerning the mechanical room. In particular the labels "*sat_start_stop*" and "*sun_start_stop*" were assigned to the patterns 2-5-2 and 2-2-2 respectively. Such patterns refer to some weekends during which the hourly energy consumption of the substation was always higher than 200 kWh due to a continuative operation of the cooling system. The remaining labels refer to abnormal continuative chiller operation before the start time (i.e., "*wd_start*"), after the stop time (i.e., "*wd_stop*") or both of them during working days (i.e., "*wd_start_stop_1*" and "*wd_start_stop_2"*). In more detail, even though "*wd_start_stop_1*" and

"*wd_start_stop_2*" refer to the same typology of anomalous event, the two groups are characterised by different climatic conditions. In fact, while the pattern 2-3-2 (i.e., "*wd_start_stop_1*") occurred during the middle season (i.e., May, September), the pattern 2-4-2 (i.e., "*wd_start_stop_2*") refers to abnormal weekdays of the warm season.

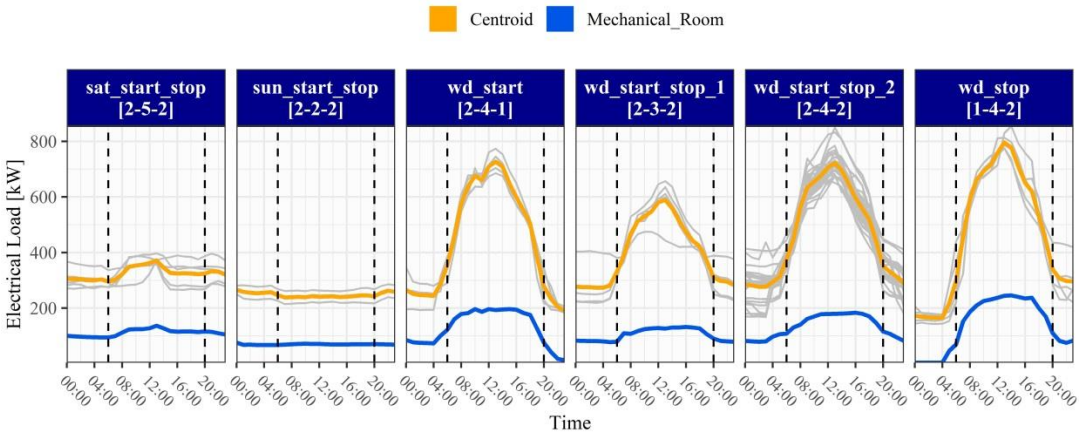

Figure 7. Infrequent profiles labeled according to anomalous management of the mechanical room: in grey the load profiles related to the electrical demand of the substation C, in yellow their centroids and in blue the corresponding average load profile of the mechanical room.

All these profiles were then isolated and removed from the dataset with the aim of obtaining a fault-free data set for the development of a reference prediction model.

5.2 Stage 2: Development of fault-free prediction model

The second stage of the methodological framework aims at developing a robust reference model capable to predict the normal/expected energy consumption of substation C. The prediction model combines a non-autoregressive Multi-Layer Perceptron ANN and a regression tree developed through the CART algorithm.

The considered input variables of the ANN were: i) the hour of the day, ii) the month, iii) the day of the week, vi) the week of the year, vi) the hourly outdoor temperature and vii) the day ahead energy consumption. This latter variable was in turn evaluated through a regression tree model for the day ahead, which uses as input the i)

day of the week, ii) the month and the iii) the daily mean outdoor air temperature (see Figure 8).

The outdoor air temperature can be easily obtained from a weather forecast service with very high accuracy in the prediction. In this case study, actual values of outdoor temperature were used, assuming them as output of a perfect prediction.

The regression tree model was trained and tested, following the same procedure discussed in (Capozzoli et al. 2018), only on the daily energy consumption data from 01-01-2016 to 31-12-2018 that were not labelled as anomalous during the pattern recognition phase (i.e., stage 1 of the framework).

The splitting process of the regression tree is based on the reduction of the variance around the mean value of the numeric target variable in each leaf node until the stopping criteria have been satisfied. In this case study, the selected stopping criterion was based on the minimum number of objects in a child node that was set equal to 30.

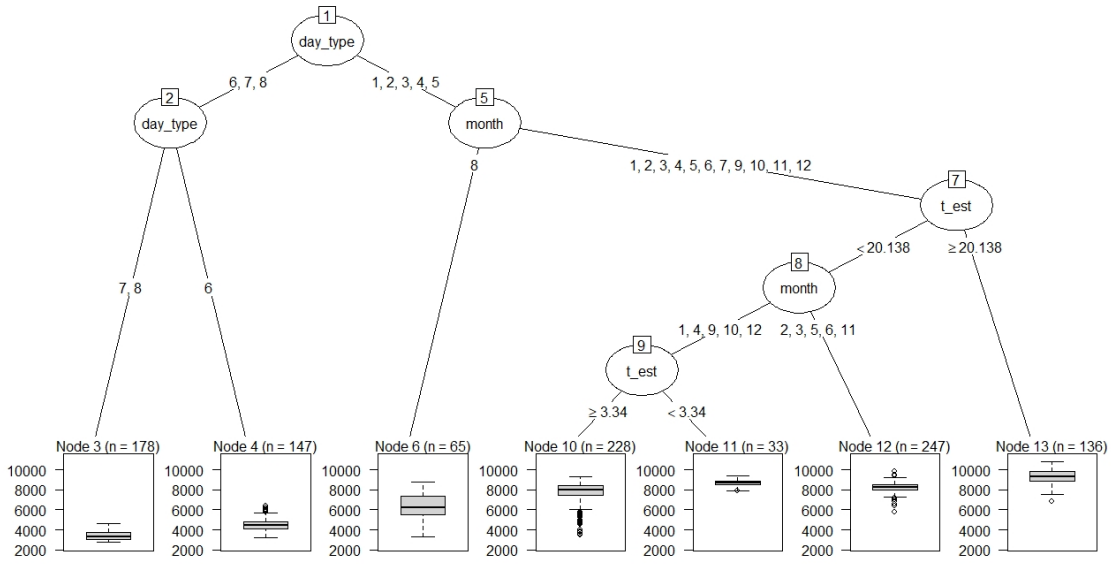

Figure 8. Regression tree developed for predicting the daily energy consumption of the substation C

In addition, a cost-complexity pruning process was performed in order to avoid overfitting problems. Such procedure makes it possible to identify the optimal trade-off

between the regression accuracy (i.e., residuals sum of squares) and the tree complexity (i.e., number of terminal nodes) in order to set a reasonable tree size by reducing the number of branches and terminal nodes.

Table 1. Decision rules extracted from the developed regression tree

Node	N° of objects	Decision Rules	Predicted value
3	178	IF **day_type** = 7 OR 8	3390 kWh
4	147	IF **day_type** = 6	4500 kWh
6	65	IF **day_type** < 6 AND **month** = 8	6210 kWh
10	228	IF **day_type** < 6 AND month = 1, 4, 9, 10, 12 AND **t_est** ≥ 3.34 °C	7650 kWh
11	33	IF **day_type** < 6 AND month = 1, 4, 9, 10, 12 AND **t_est** < 3.34 °C	8700 kWh
12	247	IF **day_type** < 6 AND t**_est** < 20.13 °C AND **month** = 2, 3, 5, 6, 11	8220 kWh
13	136	IF **day_type** < 6 AND **month** ≠ 8 AND **t_est** ≥ 20.13 °C	9230 kWh

The final model (Figure 8) is composed of three splitting levels and six terminal nodes, achieving an accuracy of about 90%. The output of the regression tree consists of a set of decision rules (Table 1) that provide numerical predictions of the total daily electrical energy consumption which were subsequently used as an input for the non-autoregressive Multi-Layer Perceptron (MLP) model.

The use of the predicted daily energy consumption as input of the ANN made it possible to reduce under-fitting problems of the hourly energy consumption during normal days. In addition, the final estimation model results with a higher interpretability considering that the estimation of the regression tree is obtained with a set of IF-THEN rules.

The structure of the neural network consisted in an input layer including 7 nodes, 2 hidden layers and an output layer with 1 node. As previously mentioned the inputs considered are i) the mean hourly outdoor air temperature (T_{ext}) of the hour at which the estimation is carried out, ii) the total daily electrical energy consumption (E_{day}) predicted

with the CART, iii) the day of the week (from 1 to 8, where 8 is the label of holydays), iv) the day of the year (from 1 to 365), v) the week of the year (from 1 to 52), vi) the month of the year (from 1 to 12) and vii) the hour of the day (from 0 to 23). All the inputs were treated as numerical variables and were normalised on their maximum values in the (0,1) range. In order to identify the best configuration of the network different sets of hyperparameters were tested.

Table 2. Configurations of hyper-parameters tested for identify the best architecture of the neural network

Test	Neurons 1st layer	Neurons 2st layer	Batch size	N° of epochs	Learning rate	MAPE training	MAPE testing
1	8	8	720	10000	1E-04	12.0 %	12.2 %
2	8	8	48	1000	5E-03	13.9 %	12.7 %
3	16	16	720	10000	1E-04	11.3 %	11.6 %
4	16	8	48	1000	1E-03	11.6 %	11.5 %
5	16	8	360	10000	1E-04	11.8 %	12.1 %
6	16	16	48	1000	5E-03	12.0 %	11.7 %
7	32	32	360	10000	1E-04	10.2 %	10.8 %
8	32	8	720	10000	1E-04	11.2 %	12.3 %
9	32	8	360	10000	1E-04	12.7 %	12.7 %
10	**64**	**64**	**48**	**1000**	**1E-03**	**8.5 %**	**9.1 %**
11	64	64	360	10000	1E-04	8.9 %	9.6 %
12	64	32	720	10000	1E-04	9.5 %	10.7 %

In Table 2 all the tested configurations are reported, considering the number of hidden layers (i.e., 2 hidden layers), the activation function (i.e., Relu), the optimizer of neuron's weights (i.e., RSMProp) and the loss function (i.e., Mean Absolute Error) as constant parameters. The model was developed in Keras (i.e., a high-level neural networks API) in the statistical R environment (R core team, 2017). The model was

trained, validated and randomly tested by splitting the hourly data from 01-01-2016 to 31-12-2018 in subsets of about 70%, 10% and 20% respectively.

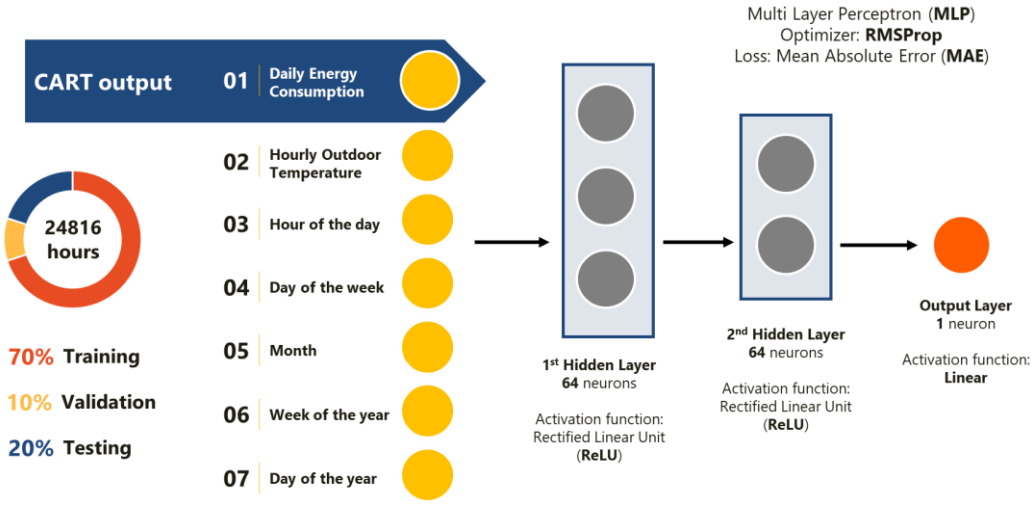

Figure 9. *Architecture of the neural network developed for predicting the hourly energy consumption of substation C*

The best evaluated configuration of the network is the 10th among the tested ones reported in Table 2. The architecture of the model is graphically reported in Figure 9 and consists of 64 neurons for the first hidden layer, 64 for the second one, batch size equal to 48, number of epochs equal to 1000 and learning rate equal to 0.001. The Mean Absolute Percentage Error (MAPE) in training and testing was 8.5% and 9.1% respectively. The results achieved in terms of MAPE were found to be consistent with the findings of the extended benchmarking analysis conducted in (Miller, 2019) about the performance of energy consumption prediction models.

5.3 Stage 3: Fault detection and diagnosis process

After the development of the reference model, the detection of anomalous daily patterns was tested. A punctual anomaly is detected if the hourly difference between the actual and predicted energy consumption is higher than three times the standard deviation of the model residuals.

The standard deviation was evaluated hour by hour separately for working and non-working days on the residuals of training fault-free dataset. Following this process, it was possible to obtain more sensitive thresholds of 3-σ, during days and hours characterised by low deviation in energy consumption (e.g., Sundays, night hours).

Moreover, if the number of punctual anomalies per day overcomes a threshold value the entire daily load profile was supposed to have an anomalous energy trend. A sensitivity analysis was performed in order to select the optimal number of hours ($N_{an.-per-day}$) of out-of-range values per day for robustly detecting anomalous daily energy patterns. The anomalous trend detection rate of the procedure was assessed for different sets of $N_{an.-per-day}$ on a sample composed by both fault-free and faulty days (anomalous days filtered out in the pre-processing phase).

The activation of the anomaly alert was tested on 100 days sampled from the fault-free dataset and the 48 anomalous days filtered out during the pre-processing phase by varying the threshold $N_{an.\ per\ day}$ in the range 1-24 hours, see Figure 10.

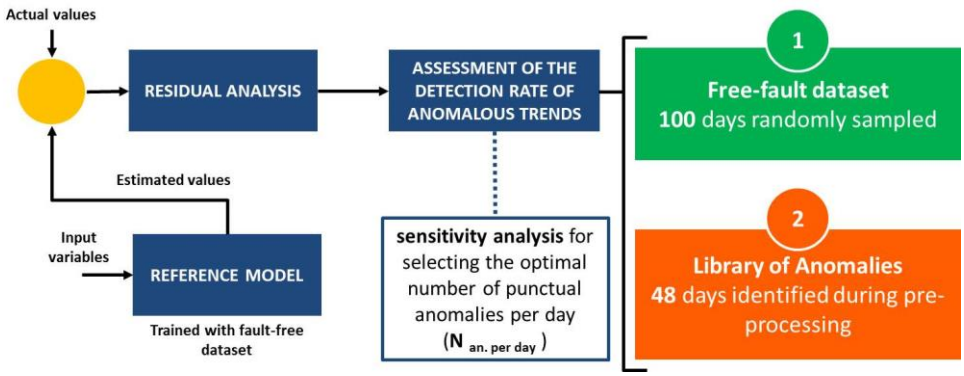

Figure 10. *Process for the identification of the optimal number of hours of out-of-range values per day for robustly detect anomalous daily energy pattern*

For each value of $N_{an.-per-day}$ the detection capability of the process was assessed calculating the number of *true positives* (anomalous days detected as anomalous), *true negatives* (normal days detected as normal), *false positives* (normal days detected as

anomalous) and *false negatives* (anomalous day detected as normal). The main objective of this analysis was to define the minimum number of out-of-range hours for robustly detecting anomalous energy trends and limiting the number of false alarms.

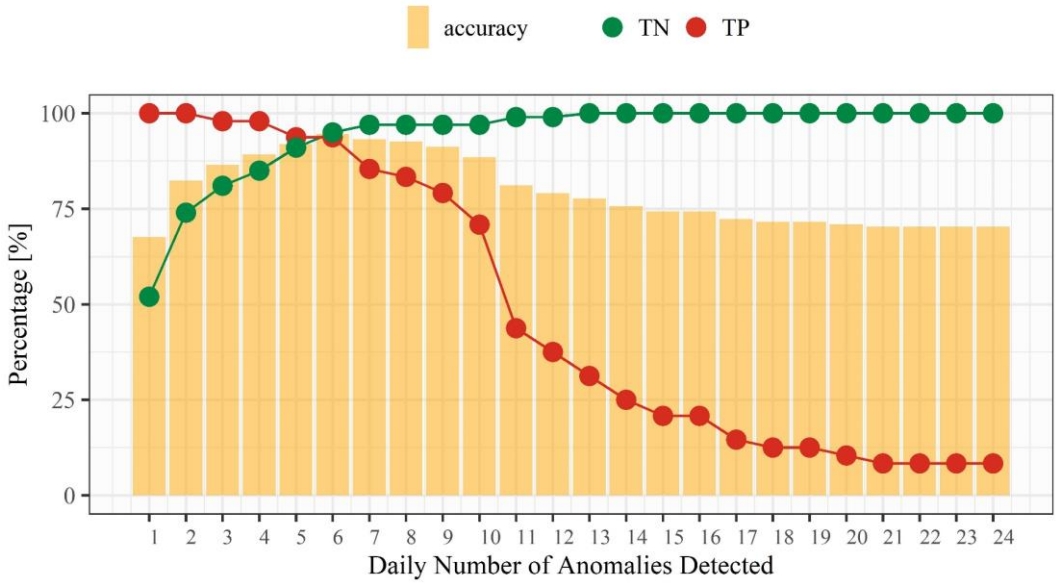

Figure 11. *Overall accuracy, percentage of true positives and percentage of true negatives for each value of* $N_{an.\text{-}per\text{-}day}$

Figure 11 shows the results of the sensitivity analysis. In particular the green and red lines represent the percentage of *true positive* and *true negative* values respectively, while the bars of the histogram correspond to the overall accuracy of the detection tool for each value of $N_{an.\ per\ day}$ calculated as follows (eq. 3):

$$Accuracy = \frac{(TP+TN)}{Number\ of\ days} * 100 \quad (3)$$

The best solution, with an overall accuracy of 94.6 %, was identified when $N_{an.\text{-}per\text{-}day}$ is equal to six. It means that after six out-of-range hourly predictions of the energy consumption, the entire daily load profile was labelled as anomalous.

In Table 3 the confusion matrix of the results of the detection tool related to validation data (setting $N_{an.\text{-}per\text{-}day}$ equal to 6) is reported.

Table 3. *Confusion matrix of the results of the detection tool*

n = 148	Predicted: **anomaly**	Predicted: **normal**	Recall
Actual: **anomaly**	45	3	93.7%
Actual: **normal**	5	95	95%
Precision	90%	96.9%	

The model was able to detect the 93.7% of the anomalous profiles and only the 5% of fault-free days was wrongly predicted as anomalous. The obtained results showed that the anomalous pattern recognition is reliable and robust enough for a real implementation. After the tuning of tool detection parameters, the data included between 01-01-2019 and 25-07-2019 were used for testing the diagnosis capabilities of the process.

Among the 205 days included in the testing dataset, 13 daily load profiles were recognized as anomalous given that for all these days a number of hours higher than $N_{an.per\text{-}day}$ have been identified as punctual anomalies.

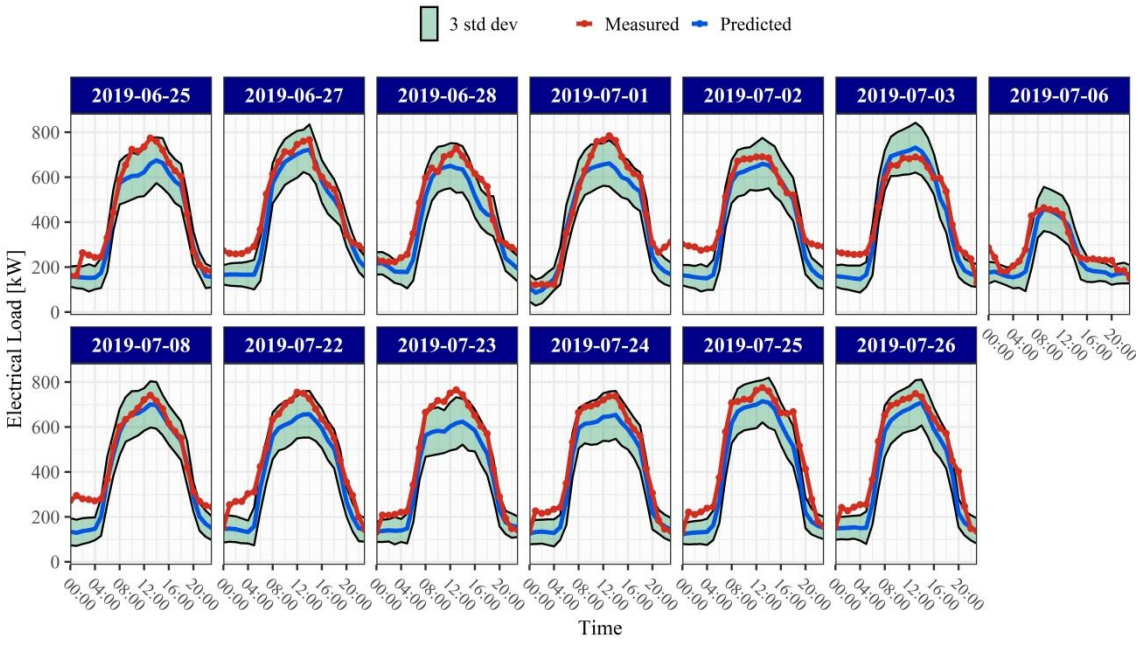

Figure 12. *Daily load profiles recognized as anomalous in the testing dataset of 2019: in red actual load profile, in blue estimated load profiles, in green the 3 standard deviation range of residuals*

During the deployment of the FDD tool, when a load profile is detected as anomalous (i.e., more than 6 hours are characterised by an anomalous power demand) three different conditions could occur:

- The punctual anomalies detected are noisy values related to problems in the transmission or storing of data.
- The punctual anomalies detected are related to the absence of the monitored data (missing values) and the detector was not able to compute the residuals
- The punctual anomalies refer to energy consumption data that are consistent regarding data quality but actually describe an anomalous pattern.

In order to handle all the possible cases, when a daily pattern is detected as anomalous, it is subjected to an analysis of inconsistencies. First missing values are detected and replaced by means of a linear interpolation; when 3 consecutive hours include missing values the daily profile is filtered out. Moreover, the punctual outliers, are detected by means of the Hampel filter method (Pearson 1999).

When the load profile considered has been checked, the residuals (related to the reference estimation model) are recalculated and if the number of punctual anomalies is still higher than $N_{an.\ per\ day}$, the profile goes through the diagnosis module.

In the test performed, all the 13 profiles detected still overcame the threshold of $N_{an.\text{-}per\text{-}day}$ after the detection of inconsistencies (Figure 12). According to the methodology, they were compared with the centroids of the anomalous profiles previously identified in the pattern recognition stage. Specifically, the detected anomalous profiles were assigned to the clusters with the closest centroids only if the distance between those profiles is lower than a fixed threshold otherwise a new cluster is generated. In this case study a threshold

of 240 kWh (assuming 10 kWh per 24 hours) was assumed as a distance threshold for the generation of a new cluster.

Figure 13 shows in which cluster the detected profiles (red lines) were grouped due to high similarity with the centroids (yellow line). Moreover, also the profiles of the energy consumption of the mechanical room (blue lines) are shown in the figure. According to the results obtained it is possible to see that two profiles were labelled as "*wd_start*", three profiles were labelled as "*wd_start_stop_2*" and one profile was labelled as "*wd_stop*". Those profiles are very closed to the centroids of the clusters they were assigned to, highlighting how the diagnosis process is accurate in distinguishing the different types of anomalies. None of the detected profiles during the testing process were assigned to clusters "*sat_start_stop*", "*sun_start_stop*" and "*wd_start_stop_1*".

Moreover 7 profiles out of 13 overcame the threshold distance for being assigned to the pre-identified groups of anomalies, so that two new groups of infrequent patterns were automatically generated. The profiles, labelled as "*new_anomaly_1*" are similar to the profiles labelled as "*wd_start*", however, during these days the cooling system was not continuatively operated during night-time, but the chillers were turned ON about 5 hours before the schedule time. Moreover, a single profile was labelled as "*new_anomaly_2*". Such profile refers to a Saturday during which the cooling system was turned OFF in the very late night (about 2:00 a.m.) and then turned ON again after 2 hours. In addition, instead of to be turned OFF at 4:00 p.m. the cooling system was operated about 8 hours longer than the normal schedule time.

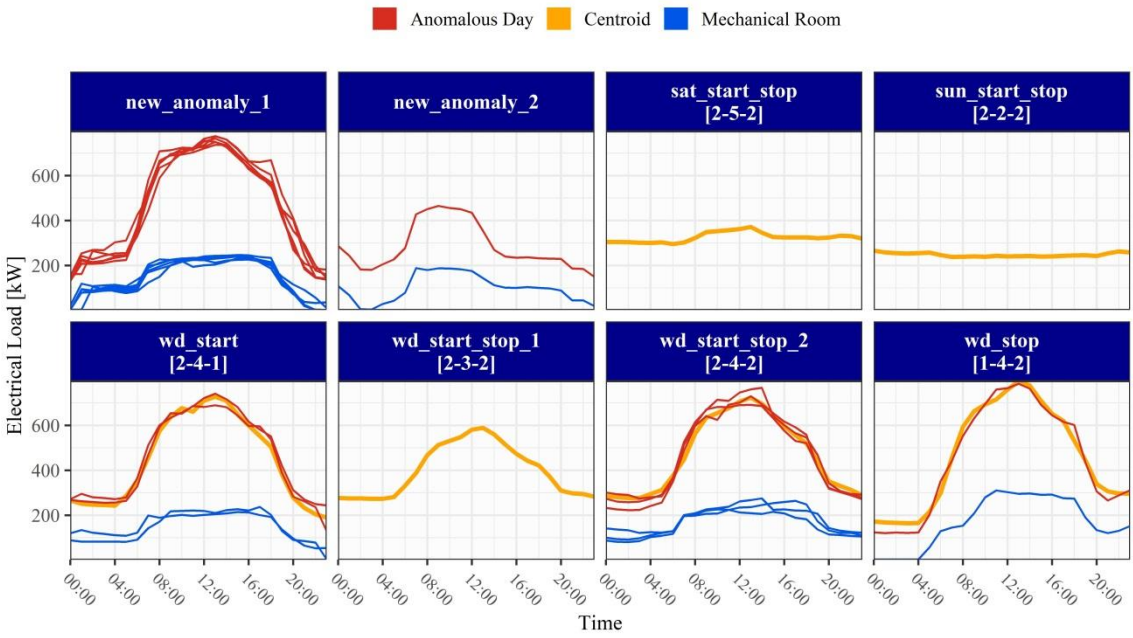

Figure 13. *Classification of the detected profiles in the pre-identified groups of anomalies: in red the detected anomalous profiles of the substation C, in yellow the centroids of the pre-identified groups of anomalies, in blue the profiles of the energy consumption of the mechanical room.*

The results obtained proved that the anomalous profiles extracted in the pattern recognition phase represent a consistent fault library capable to effectively summarise the most important anomalous patterns of the substation C. In addition, the entire procedure proved to be flexible and upgradable over time considering the opportunity to automatically generate new classes of anomalies in a robust way.

6. Discussion and concluding remarks

The need of reducing and optimising the use of energy during building operation is a key issue to face for achieving current decarbonisation targets. Nowadays the implementation of Building Energy Management Systems (BEMS) makes the collection of a large amount of building related data possible. The analysis of these data represents a valuable opportunity for better managing the energy demand of buildings during their operation (Yu et al., 2013) and for making occupants aware of their energy intensive behaviours

(Yan et al. 2017). Data analytics-based energy management makes it possible on one side to gain deep knowledge of building energy behaviour and on the other side to predict its evolution over time.

This study proposed a novel methodology for the development of a robust reference model for detecting abnormal daily electrical energy consumption patterns of a transformer substation located in the Politecnico di Torino. The model estimates the hourly energy consumption of the substation that can approach or deviate from the measured one. In the first case the substation is behaving as expected by the model, which has been trained on fault-free data. In the second case the deviation can be generated by malfunctions of the metering system or by an abnormal consumption/management of facilities/systems served by the substation. The last hypothesis was tested, by verifying the capability of the model in detecting all the anomalous patterns that were filtered out from the dataset during the pattern recognition phase. By means of the fine tuning of the neural network parameters and the identification of the best value of $N_{an.\text{-}per\text{-}day}$, the detection rate of anomalous daily load profiles was high, and the false alarm rate was low enough to consider the procedure robust and reliable if implemented on real BEMS.

The pattern recognition analysis showed that anomalous events usually occurred during consecutive days. Therefore, such detection tool can support building managers to early recognize abnormal events avoiding their protraction over a long time. Detecting in advance anomalous patterns leads to a significant reduction of energy waste during building operation. For instance, comparing the expected energy consumption with the actual one during the 48 days grouped in the anomalous clusters, a potential saving of around 40 MWh could be estimated.

In the methodological framework, also a diagnosis phase was conceived. Potential abnormal events were compared with the centroids of the anomalous clusters identified

in the pre-processing stage. In that case the abnormal daily load profile was assigned to its closest a-priori labelled cluster making the diagnosis of the most probable cause possible. This diagnosis procedure is open and upgradable over time. In fact, in case of low match with the a-priori labelled anomalies, new patterns are automatically added to the anomaly library. This work led to the development of a flexible intelligent tool useful for operating a continuous commissioning of the campus facilities by integrating several tailored analytics layers.

Future works will be aimed at extending the tool capabilities and testing the process on other data sets for benchmarking purposes (Miller, 2019). The methodology consists of a multi-step process that can be reproduced and implemented also substituting algorithms employed in this study (e.g. clustering algorithm, prediction model) according to user necessities. The proposed algorithms cannot be considered always as the best general solution for the tasks addressed in this study (i.e., data clustering, forecasting). However, once a machine learning algorithm is selected its carefully development (e.g., selection of input variables, training and testing) and tuning (e.g., setting of hyperparameters) are essential steps for obtaining the best performance achievable. As a reference, in this study the sensitivity analysis performed on ANN model hyperparameters made it possible to identify a performance gap of 5.4% (in terms of MAPE) between the best and worst configuration used for training the model. Future works will also be aimed at enhancing the capability of this tool in recognising new kinds of anomalies or rare events and at optimally scheduling the re-training of the prediction models. Moreover, an advanced visualisation will be conceived and developed for supporting different actors (e.g., energy manager, users) in enhancing their awareness of building operation saving energy and costs.

Acknowledgements

The authors express their gratitude to Living Lab of PoliTo for providing data and to Eng. Giovanni Carioni for the support in data preparation and collection. The authors also gratefully acknowledge the support of this research by the Research Grant Council of the Hong Kong SAR (152133/19E).

References

Ahmad MW, Mourshed M, Yuce B, Rezgui Y. 2016. Computational intelligence techniques for HVAC systems: A review. *Building Simulation*. 9: 359–398.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. Classification and regression trees. New York: Routledge.

Capozzoli A, Lauro F, Khan I. 2015. Fault detection analysis using data mining techniques for a cluster of smart office buildings. *Expert Systems with Applications*; 42: 4324–4338.

Capozzoli A, Piscitelli MS, Brandi S. 2017. Mining typical load profiles in buildings to support energy management in the smart city context. *Energy Procedia*. 134: 865-874

Capozzoli A, Piscitelli MS, Brandi S, Grassi D, Chicco G. 2018. Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings, *Energy*. 157: 336-352

Chicco G, Napoli R, Postolache P, Scutariu M, Toader C. 2005. Emergent electricity customer classification. *IEEE Proceedings-Generation, Transmission and Distribution*. 152: 164–172.

Chicco G, Napoli R, Piglione F. 2006. Comparisons among clustering techniques for electricity customer classification, *IEEE Transactions on Power Systems.* 21(2): 933-940

Chou J, Telaga AS. 2014. Real-time detection of anomalous power consumption. *Renewable and Sustainable Energy Reviews*. 33:400–11.

Corder G W, Foreman D I. 2014. Nonparametric statistics: A step-by-step approach. John Wiley & Sons.

Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2019 on the energy performance of buildings.

Davies DL, Bouldin DW. 1979. A cluster separation measure. *IEEE Transactions on pattern analysis and machine intelligence*. 1: 224–227.

Dehestani D, Eftekhari F, Guo Y, Ling S, Su S, Nguyen H. 2013. Online Support Vector Machine Applicationfor Model Based Fault Detection and Isolationof HVAC System. *International Journal of Machine Learning and Computing*;1:66–72.

Du Z, Fan B, Jin X, Chi J. 2014. Fault detection and diagnosis for buildings and HVAC systems using combined networks and subtractive clustering analysis, *Building and Environment*. 73: 1-11.

DOE Office of Energy Efficiency and Renewable Energy. 2012. *Buildings Energy Databook.*

Fan C, Xiao F, Zhao Y, Wang J. 2018. Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data. *Applied Energy*. 211:1123–35

Fan C, Xiao F, Madsen H, Wang D. 2015. Temporal knowledge discovery in big BAS data for building energy management. *Energy and Buildings*. 109: 75–89.

González PA, Zamarreño JM. 2005. Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. *Energy and Buildings*. 37(6):595–601.

Hinton G, Srivastava N, Swersky K. 2012. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Han, H., B. Gu, Y. Hong, and J. Kang. 2011a. Automated FDD of multiple-simultaneous faults (MSF) and the application to building chillers. *Energy and Buildings* 43(9):2524–32.

Han, H., B. Gu, T. Wang, and Z.R. Li. 2011b. Important sensors for chiller fault detection and diagnosis (FDD) from the perspective of feature selection and machine learning. *International Journal of Refrigeration.* 34(2):586–99.

Katipamula S, Kim W. 2018. A review of fault detection and diagnostics methods for building systems. *Science and Technology for the Built Environment.* 24: 3–21.

Ku K, Jeong S. 2018 Building electric energy prediction modeling for BEMS using easily obtainable weather factors with Kriging model and data mining. *Building Simulation*. 11: 739–751.

Li D, Zhou Y, Hu G, Spanos CJ. 2016. Fault detection and diagnosis for building cooling system with a tree-structured learning method. *Energy and Buildings*, 127:540–51.

Liang J, Du R. 2007. Model-based fault detection and diagnosis of HVAC systems using Support Vector Machine method. *International journal of refrigeration*. 30(6): 1104-1114

Miller C. 2019. More Buildings Make More Generalizable Models - Benchmarking Prediction Methods on Open Electrical Meter Data. *Machine Learning and Knowledge Extraction*. 1(3): 974-993

Miller C, Nagy Z, Schlueter A. 2018. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews*. 81: 1365–1377.

Miller C, Nagy Z, Schlueter A. 2015. Automated daily pattern filtering of measured building performance data. *Automation in Construction*. 49: 1–17.

Nair V, Hinton G E. 2010. Rectified linear units improve restricted boltzmann machines. *In Proceedings of the 27th international conference on machine learning (ICML-10)*. 807-814

Panapakidis I, Christoforidis G. 2018. Optimal Selection of Clustering Algorithm via Multi-Criteria Decision Analysis (MCDA) for Load Profiling Applications. *Applied Science*. 8: 237–279.

Pearson RK. 1999. Data cleaning for dynamic modeling and control. *1999 Eur Control Conf*. 2584–9.

Piscitelli M S, Brandi S, Capozzoli A. 2019. Recognition and classification of typical load profiles in buildings with non-intrusive learning approach. *Applied Energy*. 255, 113727.

Piscitelli M S, Chiabrera E, Brandi S, Capozzoli A. 2018. A tool for anomaly detection of energy consumption in buildings: the case of Politecnico di Torino campus. (*4th Asia Conference of International Building Performance Simulation Association - ASim2018*), 3-5 December 2018 Hong Kong (China)

Qiu S, Feng F, Li Z, Yang G, Xu P, Li Z. 2018. Data mining based framework to identify rule based operation strategies for buildings with power metering system. *Building Simulation*.12: 195–205.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. http://www.r-project.org/.

Tan P-N, Steinbach M, Kumar V. 2006. Classification: Basic Concepts, Decision Trees, and Model Evaluation. Introduction to Data Mining. 67: 145–205.

Yan D, Pan S, Wang X, Wei Y, Zhang X, Gal C, Ren G, Shi Y, Wu J, Xia L, Xie J, Liu J. 2017. Cluster analysis for occupant-behavior based electricity load patterns in buildings: A case study in Shanghai residences. *Building Simulation*. 10: 889–898.

Yan R, Ma Z, Kokogiannakis G, Zhao Y. 2016. A sensor fault detection strategy for air handling units using cluster analysis, *Automation in Construction*. 70:77–88.

Yu Z, Fung BCM, Haghighat F. 2013. Extracting knowledge from building-related data - A data mining framework. *Building Simulation*. 6: 207–222.

Zhao Y, Wen J, Xiao F, Yang X, Wang S. 2017. Diagnostic Bayesian networks for diagnosing air handling units faults – part I: Faults in dampers, fans, filters and sensors. *Applied Thermal Enginereeng*. 111:1272–86.

Zhao Y, Wen J, Wang S. 2015. Diagnostic Bayesian networks for diagnosing air handling units faults - Part II: Faults in coils and sensors. *Applied Thermal Enginereeng*. 90:145–57.