ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Energetics (32nd Cycle)

Enhancing energy management in buildings through data analytics technologies

Marco Savino Piscitelli

* * * * * *

Supervisors
Prof. Alfonso Capozzoli, Politecnico di Torino, Supervisor
Prof. Marco Perino, Politecnico di Torino, Co-Supervisor

Doctoral Examination Committee:

Asst. Prof. Zoltan Nagy, The University of Texas at Austin
Asst. Prof. Clayton Miller, National University of Singapore
Asst. Prof. Cheng Fan, Shenzhen University
Prof. Antonio Rosato, Università degli studi di Napoli Federico II
Prof. Enrico Fabrizio, Politecnico di Torino

Politecnico di Torino
2020

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data. Parts of this Ph.D. dissertation were also previously published in international Journals, also reported in Appendix C of this thesis.

Marco Savino Piscitelli

Marco Savino Piscitelli
Turin, April 30, 2020

Dedicated to the tireless past, present and future members of BAEDA Lab

There can be no value in the whole unless there is value in the parts.

Bertrand Russell

Acknowledgment

Many people supported me during this Ph.D. journey, and deserving to be thanked. Firstly, my supervisor. I truly thank Prof. Alfonso Capozzoli for having spent tons of time on me during these precious years, for his continuous academic and personal support, guidance, suggestions and for having teach me the love for the scientific research without compromises.

Incredible acknowledges to BAEDA fellows for being such incredible colleagues but most of all good friends.

I acknowledge Prof. Marco Perino and the whole TEBE group, for the encouragement in my research and the high scientific support.

My sincere gratitude also goes to Prof. Linda Fu Xiao who provided me with the opportunity to join her research team in Hong Kong involving me in a fantastic work environment.

Lastly, I am truly thankful to my family for always being a reference point during this challenging part of my life path.

Thanks to Angela for her encouragement, understanding and most of all for her endless love.

Abstract

Advanced metering infrastructures are enabling the collection of large amounts of building-related data that are leading to a profound transformation of the energy management paradigm in buildings and energy grids. Building-related data are full of hidden knowledge that can enable significant energy savings when a proper knowledge discovery process is performed. To this purpose advanced Energy Management and Information Systems (EMIS) based on the application of powerful and novel data analytics techniques can be employed. The focus of this dissertation is on the specific segment of EMIS technologies called Decision Support Systems (DSS). DSS include Energy Information Systems (EIS) and Fault Detection and Diagnostic (FDD) systems and can be classified as enabling tools in the building energy management process. Differently from advanced control systems, DSS provide feedbacks to human users (e.g., energy manager, building owner, energy service company) assisting them in improving building performance during operation. The installation of such systems is characterized by a low investment cost and a high energy saving potential making them strategic technologies in the building sector. However, their penetration in the market is still not satisfactory.

In this dissertation four advanced and innovative data analytics based DSS tools (three EIS tools and one FDD tool) at both meter and system level are proposed with the aim of overcoming three main barriers that today thwart the full exploitation of such systems: (i) low level of user engagement, (ii) inadequate detail of the analysis and information provided, (iii) insufficient level of interpretability of the results obtained. For each scale of the analysis considered a novel methodological framework is employed for addressing the main tasks typically required to advanced EIS and FDD systems.

At system level, an EIS tool for the improvement of HVAC scheduling is developed for a town hall building. The tool can effectively reschedule the HVAC system leveraging on the analysis of building occupancy data. The results obtained for the considered case study show that the tool could lead to a potential monthly reduction of the electricity use for HVAC (space heating, space cooling, ventilation and air treatments) that ranges from 12.2% to 15.4% while the average energy saving for the whole analyzed period (4 months) amounts to 14%.

At whole building level, an EIS tool for the automatic detection of anomalous energy trends is developed for a town hall and a university campus. The results obtained for the two case studies demonstrated that the developed tool can predict the typical patterns of building energy consumption during specific periods of the day with an accuracy well over 80%. As a result of the high accuracy in identifying a typical/normal energy behavior, it is possible to achieve a strong anomaly detection capability of the tool when these patterns are violated over time during building operation.

At building portfolio level, an EIS tool for the identification of typical energy use patterns and the classification of energy customers is developed for a stock of 114 industrial buildings. The developed tool is capable to automatically extract from the building portfolio database, 5 groups of typical load profiles and estimate for a new unknown customer its membership to one of them. The tool is based on an evolutionary decision tree and achieves a classification accuracy of about 75% (6% higher than a reference classifier based on recursive partitioning decision tree).

At system component level, an FDD tool for the automatic detection and diagnosis of faults in HVAC systems with a focus on Air Handling Unit (AHU) components is introduced. The tool is developed on the ASHRAE-RP 1312 public dataset and it is capable of detecting up to 11 typical faults (related to valves, fans and dampers) in AHUs during the cooling mode with an overall accuracy of 90%.

All the developed tools leveraged on time series analytics and automated rule extraction techniques with the aim of maximizing the amount of information extracted from building data while maintaining a high level of feedback interpretability. The results obtained demonstrated the added value of data analytics in the process of building energy management and its effectiveness in extracting hidden, useful and actionable knowledge at different scales of analysis.

Findings and outcomes of the present research study are discussed providing a robust reasoning about the optimal design of data analytics processes according to specific mining purposes. Eventually, a wide overview on the lessons learned throughout this research study is proposed for clearly outlining the future application opportunities, and barriers of data analytics technologies in the energy and building sector.

Contents

List of Tables

List of Figures

1 Introduction

The increasing spread of Information and Communication Technologies (ICT) is currently leading to a profound transformation of the energy management paradigm in buildings and energy grids [1]. The building energy management represents a fundamental task for effectively enhancing energy efficiency and reducing the mismatch between the actual and expected energy performance that is often related to incorrect occupant behaviour or equipment and control system malfunctions [2]. Building energy efficiency is a growing policy priority for many countries around the world, as governments seek to reduce wasteful energy consumption for supporting strategic environmental, economic and social goals. The International Energy Agency (IEA) has estimated that in terms of primary energy consumption, buildings represent roughly 40% [3–5].

In this context, Advanced Metering Infrastructures (AMI) are enabling the collection of large amounts of building-related data that can bring significant benefits in characterising actual performance of buildings and spot valuable energy saving opportunities. Pervasive building AMI can generate millions of measurements annually and stakeholders of building portfolios may have to handle overwhelming amounts of data which continue to increase over time [6,7]. However, the current utilization of such large amount of data in buildings is still limited. The reason behind is twofold. Firstly, building data are heterogeneous, often dispersed, with different resolutions, mostly asynchronous and are stored in raw or processed formats [6,8]. Appropriate, data management systems become then necessary to store and prepare large volumes of building related-data. Secondly, the use of conventional techniques based on domain expertise, physical principles and basic statistics are not always effective in extracting knowledge from massive and complex databases [1]. As a consequence, also the building sector is experiencing artificial intelligence momentum, and more and more building management systems are exploiting advanced analytics techniques (i.e., machine learning and data mining techniques) for gaining robust insight into building energy performance patterns and enabling the development of ready-to-implement energy conservation measures [9–11].

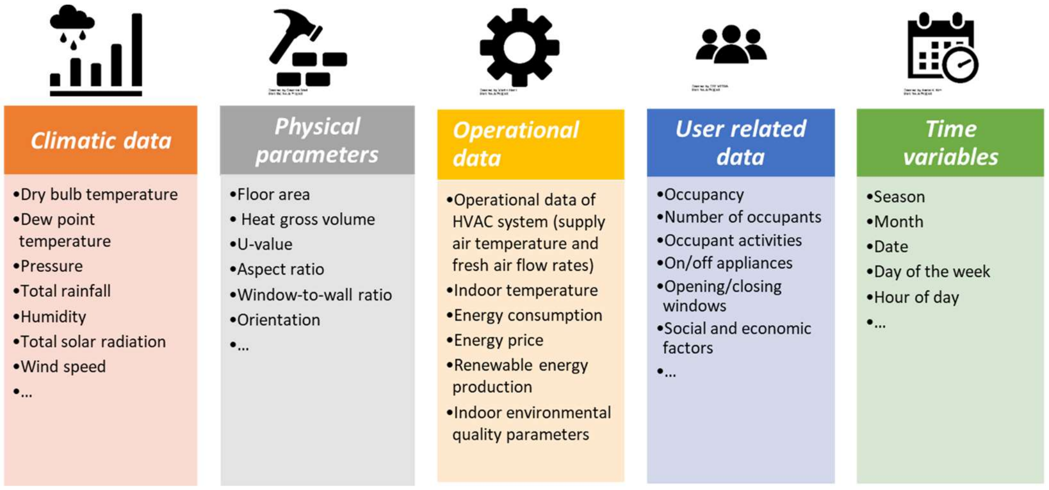
However, while data management issues are related to the solely information technology application field, the fully exploitation of data analytics in building energy management intrinsically involves a knowledge gap between building physics and data science [1].

De facto, gaining insight into building related datasets cannot be achieved by exclusively using advanced techniques but also requires specific domain expertise for extracting, managing and interpreting non-trivial knowledge.

Each building dataset has its own characteristics that significantly contribute to determine the quantity and quality of knowledge that can be potentially extracted through data analytics techniques. In the case of building-related data, some of the most important attributes are listed below according to [1]:

- Spatial scale of interest (e.g., group of buildings, single building, representative building space or energy system);
- Energy services (Heating, cooling, domestic hot water, ventilation, lighting, electrical appliances)
- Data features (e.g., minimum, maximum, mean)
- Measurement accuracy (sensor tolerance)
- Sampling frequency (i.e., annual, monthly, daily, hourly, sub-hourly)
- Monitoring period length (e.g., year, season, month)
- Data source (e.g., sensors, web platform, simulation)

Regardless the type of application, adequate and good quality data are always the cornerstone of effective analysis capable to pinpoint operation issues in buildings energy management and then identify strategic energy saving opportunities [1]. Fig. 1 shows the main groups of variables related to building data that usually are considered during analysis (i.e., climatic data, physical parameters, operational data, user related data and time variables).



Climatic data	Physical parameters	Operational data	User related data	Time variables
•Dry bulb temperature •Dew point temperature •Pressure •Total rainfall •Humidity •Total solar radiation •Wind speed •...	•Floor area • Heat gross volume •U-value •Aspect ratio •Window-to-wall ratio •Orientation •...	•Operational data of HVAC system (supply air temperature and fresh air flow rates) •Indoor temperature •Energy consumption •Energy price •Renewable energy production •Indoor environmental quality parameters	•Occupancy •Number of occupants •Occupant activities •On/off appliances •Opening/closing windows •Social and economic factors •...	•Season •Month •Date •Day of the week •Hour of day •...

Fig. 1 - Building-related data classified according to different categories of influencing factors (adapted from [1])

Clearly, due to its heterogeneity, the knowledge to extract and exploit is complex, legitimating the use of term “big data” in the application field of building automation and energy management [1].

The process of knowledge discovery, on one hand is supported by the growing availability of advanced data analytics techniques and on the other hand is thwarted by the large range of possible applications that involve users, owners and operators at different levels and can be applied at different scales (form single building component to district of buildings). In this context, the fully exploitation of data analytics techniques and their combination still remain challenging to be generalised.

This dissertation aims at proving that the heterogeneity in scale of application, the resolution and data size need the development of proper methodological processes based on advanced data analytics techniques. The main objective is to automatically extract and transfer useful knowledge in the robust way as possible and to translate it into ready-to-implement energy saving strategies in buildings and energy systems.

1.1 Motivations of the research

Building-related data are full of hidden knowledge that can enable significant energy savings when the proper discovery process is performed. In the current paradigm of smart buildings, the building owners, and managers can leverage on more and more sophisticated data analytics-based software, capable to inform and assist them in improving building performance during operation. However, such solutions are valuable if large amount of data is available. This condition is becoming a standard in buildings where modern Building Automation Systems (BAS) monitor hundreds of points with high spatial and temporal detail. BASs are used for controlling building systems and can also provide simple threshold-based alarms when measured data are out of range. However, their analytical capabilities are not enough developed for supporting users in gaining insight into measured data.

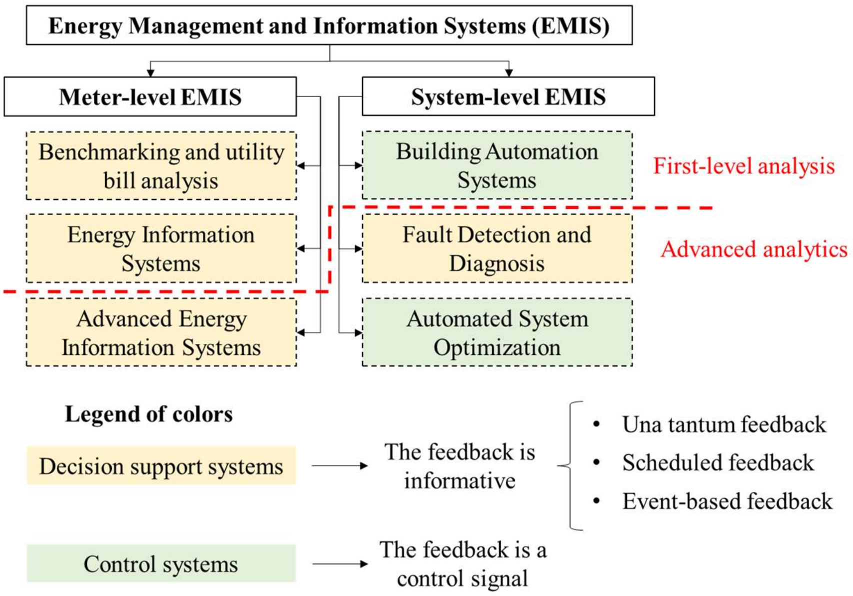

Energy Management and Information Systems (EMIS)
Meter-level EMIS
System-level EMIS
Benchmarking and utility bill analysis
Energy Information Systems
Advanced Energy Information Systems
Building Automation Systems
Fault Detection and Diagnosis
Automated System Optimization
First-level analysis
Advanced analytics
Legend of colors
Decision support systems
The feedback is informative
Una tantum feedback
Scheduled feedback
Event-based feedback
Control systems
The feedback is a control signal

Fig. 2 - EMIS tool classification according to detail of data, detail of analysis and feedback type

To this purpose Energy Management and Information Systems (EMIS) can be employed. EMIS belong to the rapidly evolving family of tools that monitor, analyse, and control building energy use and system performance [12]. Fig. 2 reports a classification of EMIS tools that consider the detail of data, the detail of analysis and the type of feedback provided.

According to [12], a first classification of EMIS tools can be formulated considering if their functionalities are enabled at meter or system level (Fig. 2).

The main difference between meter-level and system-level EMIS consists in the level of data considered in the analysis. The first category of EMIS considers data measurements at a high level (e.g., total energy consumption of a building/system) while system-level EMIS are focused on more detailed data (e.g., component level) related to the operation of specific systems (Fig. 2). For example, meter-level EMIS do not typically provide information as specific as, "cooling coil valve of air conditioning system is stuck".

Another kind of classification can be made considering how much advanced are the analyses performed (Fig. 2). While utility bill analysis and BAS are considered basic tools for controlling building systems and providing information about building performance, this dissertation is focused on EMIS tools that leverage on advanced data analytics-based technologies. A description of advanced EMIS tools is provided in [12] and reported in the following (also see Fig. 2):

Advanced Energy Information Systems (EIS): Advanced EIS are tools focused on meter-level monitored data (e.g., hourly or sub-hourly energy consumption data at whole building level) that are not usually integrated with BAS data. Such tools typically include predictive modeling and pattern recognition analysis for performing tasks such as energy consumption forecasting, anomaly detection, advanced benchmarking, load profiling and schedule optimisation of building energy systems.

Fault Detection and Diagnostic (FDD) systems: FDD tools automatically detect unpermitted deviation of at least one characteristic property of a system from its acceptable, usual, standard condition. Faults are abnormal system states whose identification and diagnosis can lead to significant energy savings. Even though FDD tools exploit BAS data, the feedback that are able to provide is much more detailed and effective than BAS (information about duration, occurrence and impact of faults).

Automated System Optimization (ASO): ASO software analyses BAS data and modifies the control settings for achieving an optimised energy performance of building systems. Differently from EIS and FDD systems, the functionalities of ASO tools are based on a two-way communication paradigm with the BAS making them advanced control solution.

Conversely, EIS and FDD systems can be classified as enabling tools, also called Decision Support Systems (DSS), whose feedback is provided to a human user (e.g., energy manager, building owner, energy service company). In this perspective, while EIS and FDD are powerful tools, they need to be integrated in a robust verification process to achieve the desired impact. In [12] are reported the findings of an implementation campaign of EMIS tools in 96 buildings. It has been demonstrated that DSS users (implementing FDD and/or EIS systems) achieved a median energy saving after two years of implementation of about 9% and 4% respectively. However, users that dedicated adequate staff time and effort in exploiting DSS output (i.e., suggested as best practice) achieved better results even beyond 20% of energy saving.

Determining the cost-effectiveness of DSS is then not straightforward given that their installation does not directly produce savings [12]. Rather, savings or improvement in energy management are achieved by acting on the basis of information these technologies provide. Their effectiveness can be then considered strictly related to three main factors: (i) the level of user engagement, (ii) the detail of the analysis and information provided, (iii) the level of interpretability of the results obtained. The optimal configuration of such aspects can lead to a significant impact on system/building/portfolio energy management and act as enabler for the spread of DSS technologies. In fact, even though the use of DSS represents an asset for the optimal management of energy, the penetration of such technologies in buildings is still not satisfactory. According to [12] the main reasons are related to the following aspects:

- **Software specification and selection**
    - Users are not clear on which analytics tool features they need.
    - Lack of clarity on differences between available analytics tools.
- **Software installation and configuration**
    - Integration problems with existing metering infrastructure and difficulty bringing all the data into a single centralized database.
    - Data quality problems.
    - Inadequate metering infrastructure.
- **Analytics process effectiveness**
    - Users are overwhelmed by data instead of being informed with actionable insights.
    - Difficulty in spotting measures/opportunities in the data.
    - Difficulty in finding root causes of anomalous/faulty operations.
    - Absence of a verification process.
- **Commissioning process**
    - Difficulty in maintaining persistence of savings.
    - Waste of energy due to operation in manual mode of systems.

In this context, this dissertation intends to give a contribution in improving the effectiveness of analytics tools that can be embedded in DSS. To this purpose, relevant applications at different scales are investigated focusing on aspects related to the maximization of knowledge discovered from monitored data, its interpretability, and the way it is transferred to the final user. To this purpose, novel data analytics-based methodologies are developed supporting two different feedback schemes, i.e. una tantum feedback and real/quasi real time feedback:

- *Una tantum feedback*: The results of the analysis performed by the DSS tool are provided to the user as static information such as the identification of energy saving opportunities (e.g. scheduling improvement of building systems) and reference performance patterns in energy consumption (e.g., benchmarking at building portfolio scale).

- *Real time/quasi real time feedback*: The knowledge extraction process of DSS tool is based on continuous analysis of the monitored data and the final user is involved in exploiting the obtained results in real time/quasi real time. In this case the DSS tool could provide several scheduled or event-based feedbacks to the user during the day.

Real/quasi real time feedback can enable a better understanding of the current building/system energy behaviour during operation most of all making it possible to identify poor performance and quickly alarm or suggest solutions. However, feedbacks that are too much frequently sent to the user, transmitted in not engaging way and with a high rate of false alarms could negatively affect the credibility of DSS tools based on real time data analytics.

For this reason, rationalise the number of feedbacks sent to the user, improve the visualization of the results obtained are essential aspects to be taken into account. Therefore, despite DSS tools represent valuable solutions for achieving significant improvement of building performance, their fully exploitation still needs a great research effort especially for what feedback interpretability is concerned.

Nowadays, the rapidly evolving sector of artificial intelligence offers a wealth of new and effective algorithms that in major part are being used also in the building sector. Most of them are open sourced and well documented. In this context, developers can better focus on the application of the algorithms and their combination in robust methodological frameworks of the analysis rather than coding the algorithms themselves [13]. This is extremely desirable for DSS tools, for which the human-in-the-loop paradigm imposes quality constraints (e.g., simplicity in understanding, commissioning and using the tool) not easy to be respected.

Due to these challenging research opportunities, DSS tools based on advanced data analytics processes (i.e., advanced EIS, FDD tools) are investigated throughout this research study.

1.2 Research outline

In order to demonstrate the potential associated to DSS, four main applications related to the implementation of EIS and FDD tools are proposed for different testbeds. Fig. 3 shows the applications investigated, with the reference of the scale of analysis and the feedback scheme assumed. Compared to FDD tools, EISs have a wider range of application in terms of objectives to be pursued and scales of analysis (i.e., system level, whole building level, building portfolio level).

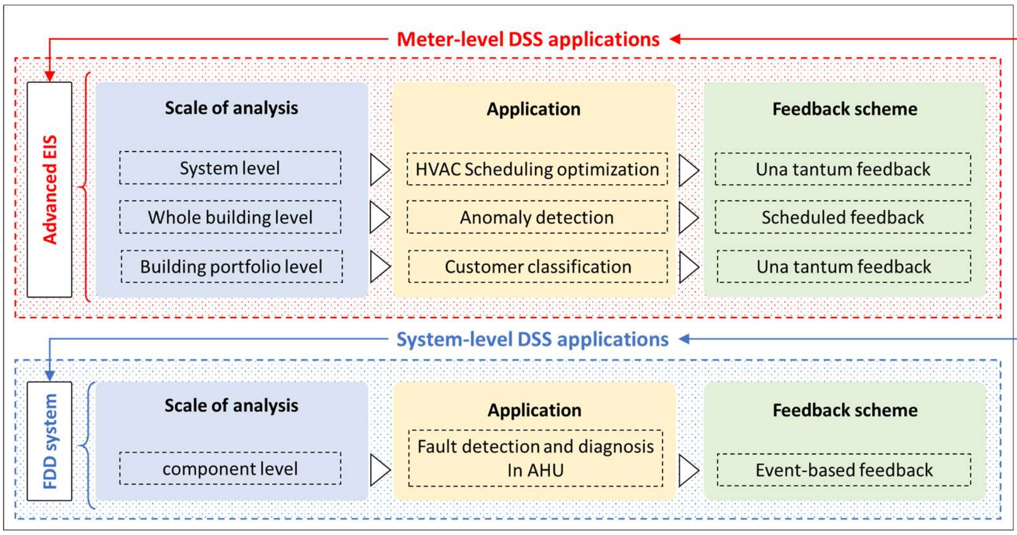


Fig. 3 - Outline of the applications investigated in the thesis with the reference of the scale of analysis and the feedback scheme assumed

For each scale considered an EIS tool was conceived and tested. In particular, novel methodological frameworks of analysis were developed for addressing the following main tasks typically required to advanced EISs (Fig. 2):

- **HVAC scheduling improvements at building system level** [14]. The improvement of HVAC schedules is one of the most effective way for reducing energy waste in building during daily operation. HVAC are responsible of a significant part of the whole building energy consumption and often are operated with fixed schedules that poorly fit the actual occupancy of the building. In that perspective, an EIS tool capable of exploiting measured occupancy data, allows energy managers to properly manage HVAC systems and significantly reduce energy consumption of their buildings. The development of this EIS tool is discussed in section 3.2.
- **Identification of energy consumption reduction opportunities through the detection of anomalous energy trends at whole building level** [2]. Anomaly detection in buildings is often related to FDD analysis conducted at system component level where the scale of analysis is small (e.g., air handling unit components). However, in most of real cases, just few and aggregate variables related to the total energy consumption of the building are monitored and collected. Improving the building energy performance by analysing aggregate data is challenging, especially if several factors such as occupants' behavior, comfort levels, operation schedules of systems generate the existence of different energy consumption patterns not always easily inferable. In this context, an EIS tool capable to automatically detect anomalous energy trends in building energy consumption allows energy managers to be promptly informed when the building is not behaving as expected and to avoid inefficient

energy management procedures. The development of this EIS tool is discussed in section 3.3.

- **Identification of typical energy use patterns and customer classification at portfolio level** [10]. An important functionality that an EIS tool should have is related to its ability of performing analyses at a scale higher than the single building. This opportunity is extremely valuable for users that usually need to manage more than one building simultaneously (e.g., municipalities, demand response aggregators). At this scale of analysis the identification of typical energy use patterns in large building portfolios can reveal knowledge about specific group of buildings that, as a reference, can be useful for designing targeted financial demand response programs, externally benchmarking energy performance of buildings and classifying energy customers. The development of this EIS tool is discussed in section 3.4.

Regarding the DSS applications at system level, an innovative procedure is developed for addressing the following main task:

- **Fault detection and diagnosis in HVAC systems with a focus on Air Handling Unit (AHU) components** (component level) [22]. The optimal management of heating ventilation and air conditioning systems, is a crucial task, considering that such systems account up to 50% of the energy demand in buildings [131]. However, Air Handling Units (AHUs), that are an essential part of HVAC systems, are often inappropriately managed negatively impacting on building energy consumption and on the control of the indoor environment conditions. In this context a robust and novel FDD tool, capable of detecting and diagnosing the main faults of fans, dampers, and valves in AHUs, is proposed in this research study. The development of this FDD tool is discussed in section 4.1.

All the developed tools leveraged on time series analytics frameworks based on the coupling of data mining and machine learning algorithms with the aim of maximizing the amount of information extracted from building monitored data while maintaining a high level of result interpretability.

To this purpose, the developed data-driven methodological frameworks leveraged on the application of automatic rule extraction techniques. Such techniques (e.g., association rules, decision trees) aim at extracting from large amounts of data, inference rules in form of IF-THEN implications that are able to effectively describe all the relations that exist between the variables included in the same dataset. In this way the results of the analysis can be translated in a set of interpretable decision rules that can be easily embedded in DSS, helping managers, owners or service companies in increasing awareness about the measured energy performance of their buildings/systems and achieve demanding energy management targets during daily operation.

1.3 Research questions

As stated in the previous section all the methodological frameworks of analysis, included in the developed DSS tools, are based on time series analytics and automatic rule extraction techniques. Knowledge extracted from building-related time series (e.g., load and occupancy profiles) contains information on how and when building energy use changes during the day for various end uses such as appliances, lighting, ventilation, heating and cooling [15,16] respect to boundary conditions (e.g., weather, time period or user/customer features) influencing their particular variation over time. Advanced DSS tools, at both meter and system level, can provide such level of insight by means of time-series analytics methods.

Algorithms related to (i) sequential and recurrent pattern mining (ii) causality analysis (iii) time series similarity proved to be flexible in their combination and effective in extracting essential knowledge from time series [17–20]. Such advanced techniques play an essential role for addressing emerging issues in building energy management such as identification of energy anomalies, identification and diagnosis of system faults, occupancy and load profiling.

However, the coupling of building physics expertise and diverse analytics techniques still needs significant contributions aimed at developing robust and generalizable analytical frameworks of analysis that provide useful knowledge to be translated in ready-to-implement energy saving and management strategies. For this reason, the primary question addressed through this research is:

- *How it is possible to robustly extract useful knowledge from building time series in order to better understand building behaviour and develop DSS solutions aimed at improving its energy performance?*

This question is articulated into several more specific parts from both analytics and energy point of view:

- *How to combine, in an effective way, time series analytics and automatic rule extraction techniques?*
- *How to prepare time series for mining only useful information from them?*
- *How to deal with highly multivariate time series problems?*
- *How to deal with time series data gathered from different system components, energy systems or even buildings?*
- *How can faulty operation conditions in building systems be detected and promptly diagnosed?*
- *How can anomalous behaviors in building energy management that should be changed be identified?*
- *How to compare the energy performance of a building to its peers?*
- *How to ensure high performance of the analytical process while maintaining high interpretability of the analysis?*

The present dissertation aims at solving and discussing all the aforementioned analytics and energy aspects that typically arise in DSS tool development phase, in a robust way as possible.

1.4 Objectives of the thesis and novelty

The application of data analytics techniques in DSS is a relatively young and fast-growing discipline and clearly its potential has not been fully explored. In the present study much of the effort is devoted in conceiving and testing several novel methodologies for actively contributing to this field of research. Methodological frameworks are developed to prove data analytics effectiveness and scalability specifying how such methodologies need to be employed for specific applications, scales of analysis and feedback schemes and which information should be mined accordingly. In this perspective the main research objectives can be summarized as follow:

- Demonstrate that data analytics-based DSS tools have a high potential in improving energy management during daily operation of buildings. Today, the most spread data analytics based technologies already installed in buildings refer to DSS solutions (i.e., EIS and FDD system [12]). For this reason, advancing research on these energy management solutions represents the most effective way for strongly impact the building automation sector in the short term.
- Address the emerging need of increased automation and robustness in data analytics-based procedures for the advanced characterization of the energy performance in buildings (i.e., from system component up to district level).
- Address the transition from a reactive to predictive approach in building energy management. Advanced DSS tools should leverage on the estimation of building and system behavior over time for helping owners and managers in delivering an optimal indoor environment quality with the possibility of anticipating or early detecting anomalous trends and system failures in their buildings.
- Address the need of high interpretability of the analyses performed by data analytics based DSS tools. DSS solutions include a human-in-the-loop paradigm in the decision-making process and for this reason they require high simplicity in terms of interpretability and opportunity to integrate them into existing systems.
- Rationalize and improve the quality of the feedback schemes especially for real time analytics processes. The process of data analysis often implies a knowledge barrier for users unfamiliar with advanced techniques. For that reason, advanced visualization represents a very important step for improving feedback quality and increasing user engagement resulting in a better exploitation of enabling tools such as DSS.

The main objective of this research study is then demonstrating the added value of data analytics and its contribution in improving DSS tools performance. To this aim robust analytical frameworks of analysis are conceived and tested on real dataset of measured building-related data. The novelty of this research is not related to the set of applications selected (well known in the research field), but it is associated to the approach followed for conducting them. In all the introduced methodologies, supervised and unsupervised rule extraction algorithms are used and combined in an innovative way for achieving the highest performance as possible while maintaining both results and analysis fully interpretable.

1.5 Organization of the thesis

The whole dissertation is divided into 5 chapters organized as showed in Fig. 4. The main content of each chapter is summarized as follow.

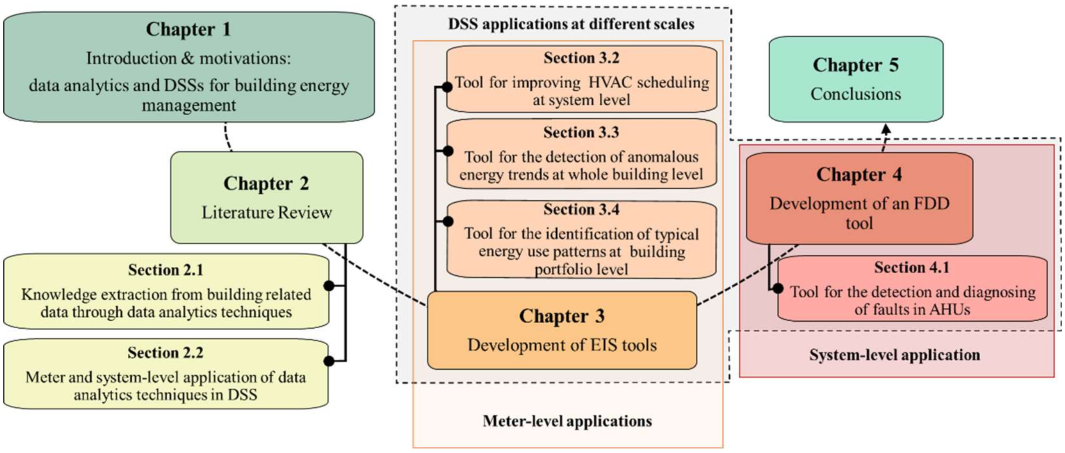

Chapter 1
Introduction & motivations:
data analytics and DSSs for building energy management
Chapter 2
Literature Review
Section 2.1
Knowledge extraction from building related data through data analytics techniques
Section 2.2
Meter and system-level application of data analytics techniques in DSS
DSS applications at different scales
Section 3.2
Tool for improving HVAC scheduling at system level
Section 3.3
Tool for the detection of anomalous energy trends at whole building level
Section 3.4
Tool for the identification of typical energy use patterns at building portfolio level
Chapter 3
Development of EIS tools
Meter-level applications
Chapter 5
Conclusions
Chapter 4
Development of an FDD tool
Section 4.1
Tool for the detection and diagnosing of faults in AHUs
System-level application

Fig. 4 - Conceptual organisation of the thesis

Chapter 1 presents the motivation of this research, the objectives, and the organization of the thesis.

Chapter 2 presents the literature review. The chapter includes two main sections. The section 2.1 introduces a general framework of knowledge extraction from building related data and provides a focus on the data mining and machine learning algorithms used in this research study for conducting time series analysis and automatic rule extraction. On the other hand, section 2.2 reviews the applications of data analytics in DSS for enhancing energy management in buildings. Both meter and system level applications are discussed.

Chapter 3 presents the developed DSS solutions at meter level with reference to advanced EIS tools. In particular section 3.2 presents and discusses the development of an EIS tool for HVAC scheduling improvements at system level (tested on the measured data of a town hall). Section 3.3 presents and discusses the development of an EIS tool for the detection of anomalous energy trends at whole building (tested on the measured data of a university campus and a town hall). Section 3.4 presents and discusses the development of an EIS for the identification of typical energy use patterns at building portfolio level (tested on the measured data of more than 100 commercial and industrial buildings).

Chapter 4 presents the developed DSS solution at system level that consists in an FDD tool conceived for detecting and diagnosing faults of AHU components in HVAC systems.

Eventually chapter 5 summarizes the work presented in this dissertation and gives an outlook about application opportunities, and barriers of data analytics-based technologies in the building sector.

2 Literature review

The scope of the present chapter is to investigate the findings achieved so far in the scientific literature about the use of advanced data analytics techniques and their application in energy and building sector. This chapter provides an extensive overview on the techniques employed to automatically extract information from building related data in order to address emerging tasks in building energy management. The chapter is organised in two main sections. On one hand, section 2.1 presents and discusses the general framework of analysis that is behind the knowledge extraction process when building related data are analysed. On the other hand, section 2.2 reviews all the applications, related to very active fields of research, that benefit from the advancements of data analytics. The applications and range of techniques reviewed in both sections go beyond the scope of the present thesis but fit very well with the current need of analytics capabilities required to advanced DSS tools in buildings that are the focus of this dissertation.

Portions of the present Chapter were already published in the following scientific papers:

- Capozzoli A., Cerquitelli T., Piscitelli M.S. 2016. *Chapter 11 – Enhancing energy efficiency in buildings through innovative data analytics technologies*, in: D. Ciprian, F. Xhafa (Eds.), Pervasive Comput., pp. 353–389. [1] (the portion reused by the author is less than 10% of the material in the book chapter as required by the publisher for a free use of the content)
- Capozzoli A., Piscitelli M.S., Brandi S. 2017. *Mining typical load profiles in buildings to support energy management in the smart city context*. Energy Procedia, 134 pp. 865–874. [11]
- Capozzoli A., Piscitelli M.S., Brandi S., Grassi D., Chicco G. 2018. *Automated load patterns learning and diagnosis for enhancing energy management in smart buildings*. Energy, 157 pp. 336–352. [21]
- 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 pp. 113727. [10]
- Capozzoli A., Piscitelli M.S., Gorrino A., Ballarini I., Corrado V. 2017. *Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings*. Sustainable Cities and Society, 35 pp. 191-208 [14].
- Piscitelli M.S., Mazzarelli D.M., Capozzoli A. Submitted for publication. *Enhancing operational performance of AHUs through an*

advanced fault detection and diagnosis process based on temporal association and decision rules. Energy and Buildings. [22]

- Capozzoli A., Piscitelli M.S., Neri F., Grassi D., Serale G. 2016. *A novel methodology for energy performance benchmarking of buildings by means of Linear Mixed Effect Model: The case of space and DHW heating of out-patient Healthcare Centres*. Applied Energy, 171 pp. 592–607. [52]

2.1 Knowledge extraction process for building related data

The section 2.1 provides an extensive overview on the techniques that are useful for automatically extracting information from data and a specific focus is devoted to the algorithms employed in methodological frameworks discussed in chapter 3 and 4.

Fig. 5 presents a general framework of the knowledge extraction process for building related-data which mainly includes four phases according to [1,23]. Data pre-processing phase consists of two tasks, i.e., data preparation and data characterization (Fig. 5). This phase is fundamental for improving the quality of a given dataset and preparing the data in formats that are suitable for the application of data analytics methods. On the other hand, data characterization is useful for obtaining preliminary knowledge from data in form of visualizations and simple statistics.

Data segmentation is an important phase in the process of analysis, and it is aimed at finding more homogenous sub-datasets in the available database for increasing the effectiveness of knowledge discovery. Data segmentation can be performed through expert-based and statistical-based approaches or by means of pattern recognition techniques.

Knowledge discovery deals with the application of different data analytics techniques, to discover hidden knowledge and patterns in massive data. The knowledge discovery phase could exploit both supervised and unsupervised learning techniques on the basis of the mining target defined by the analyst.

Eventually, knowledge exploitation is aimed at selecting, interpret, and use the knowledge discovered. Therefore, Selected knowledge is used for supporting the final user (e.g., energy manager) in the decision-making process with the final aim of spotting energy saving opportunities and improving energy performance during daily operation of buildings.

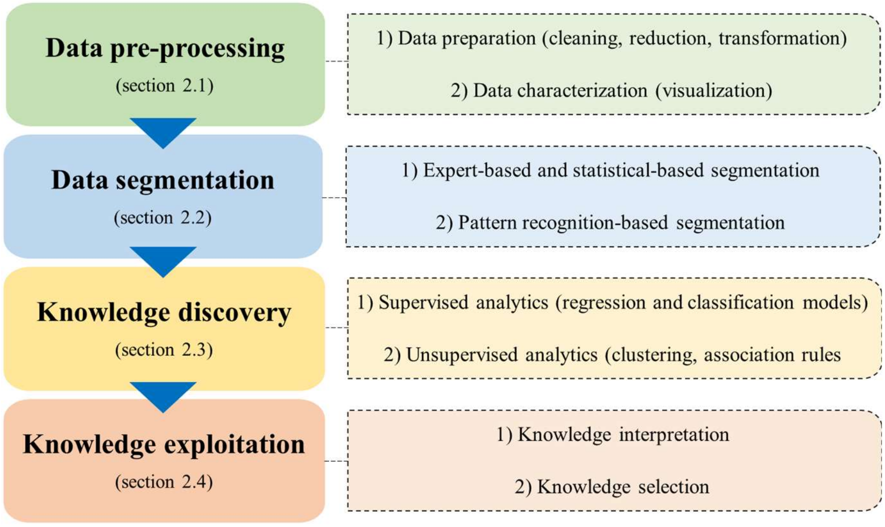

Data pre-processing
(section 2.1)
1) Data preparation (cleaning, reduction, transformation)
2) Data characterization (visualization)
Data segmentation
(section 2.2)
1) Expert-based and statistical-based segmentation
2) Pattern recognition-based segmentation
Knowledge discovery
(section 2.3)
1) Supervised analytics (regression and classification models)
2) Unsupervised analytics (clustering, association rules
Knowledge exploitation
(section 2.4)
1) Knowledge interpretation
2) Knowledge selection

Fig. 5 - Framework of the knowledge discovery process on building energy data and organization of section 2.1 (adapted from [1])

2.1.1 Data pre-processing

The two main tasks that are included in data pre-processing phase are: data preparation and data characterization. While data characterization, aims at providing a first outlook on the analysed dataset (by means of visualizations and simple statistics), data preparation underlies various objectives.

Data preparation consists in three tasks including data cleaning, data transformation and data reduction. In the literature was demonstrated that data preparation is a time-consuming task and it could take up the 80% of the total computational time of the analysis [24]. Moreover, this phase is also crucial for ensuring high performance of data analytics algorithms considering that their effectiveness is largely dependent from the quality of data and from the way in which they are prepared for the analysis.

The purpose of data cleaning task is to solve data quality issues in the dataset, that are mainly related to the presence of missing values and statistical outliers. Such inconsistencies could be generated by noisy and uncertain measurements, sensor faults and insufficient sensor calibration [1]. When dealing with missing values, a number of alternative approaches can be adopted. The simplest approach consists in ignoring records that have missing attributes. In this case the record is removed from the dataset. However, such approach is not recommended in the case of time series data, where each data point has a specific location in the time domain. In that perspective, other approaches such as substitution by mean, by regression/classification model, or by moving average model, make it possible to replace such data with different degrees of approximation. Differently from missing values, statistical outliers included in the analysed dataset should be firstly detected and then replaced. Outliers are records that significantly differ from the other elements in the data sample. The identification of such inconsistencies could be performed by means of simple statistical methods (e.g., box plot analysis (see Fig. 6)) as well as though supervised and unsupervised (e.g., clustering analysis) data analytics techniques.

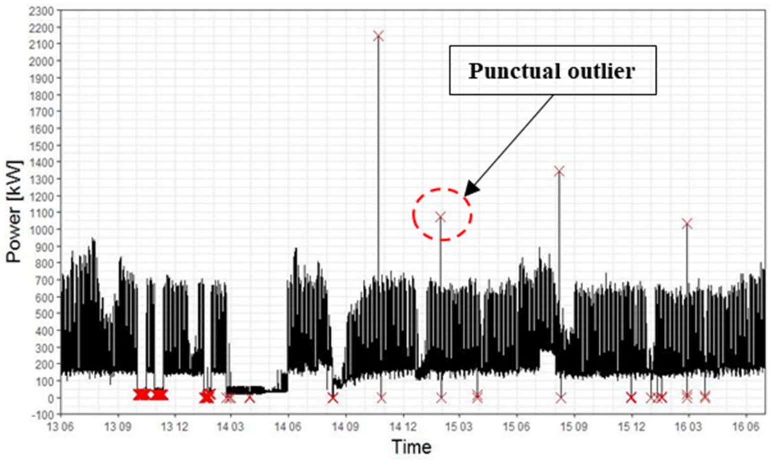
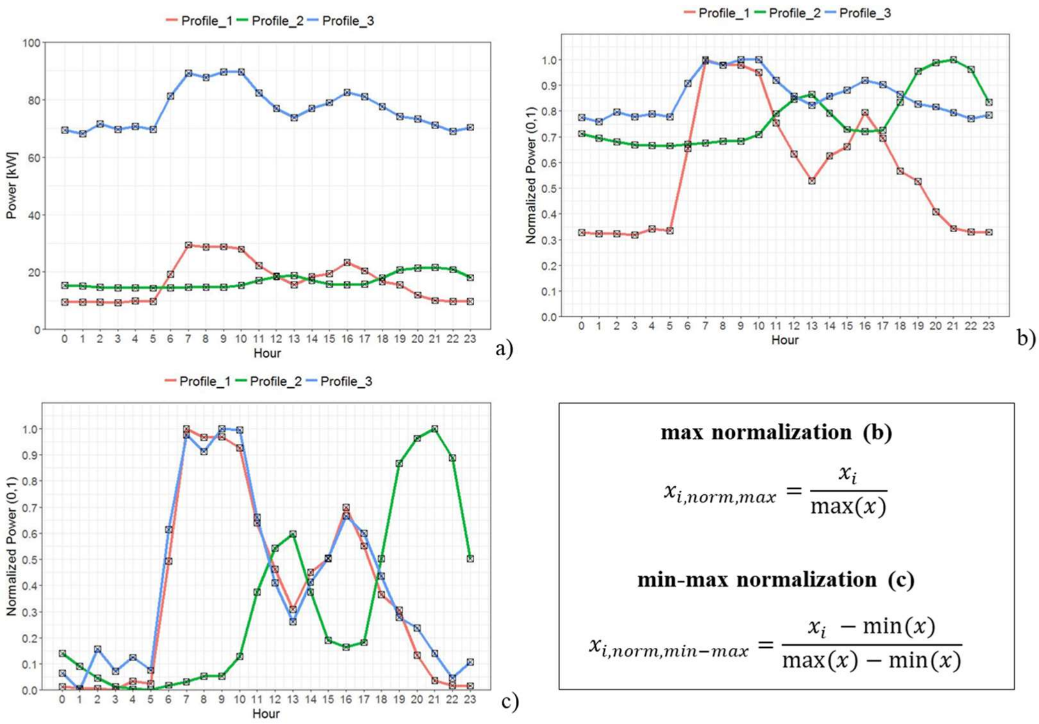


Fig. 6 - Punctual outliers identified through box plot analysis in a time series of 3 years length

When a time series is analysed, the outliers can be classified into two different types: punctual outliers and anomalous data sequences [1]. The first type of outliers is often detected through rolling window-based techniques (e.g., hampel filter), whenever anomalous time series sequences could be detected through further investigation at a later stage. In fact, differently from punctual inconsistencies (that are mainly related to data transmission problems), anomalous sequences could be associated to the monitoring of specific events that are of interest in several pattern recognition application (e.g., fault detection and diagnosis).

The process of data transformation consists in data scaling and data type transformation. The purpose of data scaling is to normalize record attributes so that they become equally important in terms of variability ranges. The methods used for data scaling include min normalization, max normalization, min-max normalization, Z-score normalization, and decimal point normalization [1]. The reasons behind data transformation can be different. For example, data scaling is an essential preliminary step for the development of some supervised analytics methods (e.g., support vector machine, artificial neural network) that can be negatively affected by heterogeneity in the scale of data input. However, in other cases, data normalization is employed for extracting patterns from data that are not sensitive to their magnitude.



max normalization (b)

$$x_{i,norm,max} = \frac{x_i}{\max(x)}$$

min-max normalization (c)

$$x_{i,norm,min-max} = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Fig. 7 - Three different daily load profiles in their original form (a), scaled through max normalization (b), and trough min-max normalization

Fig. 7 shows the effect of different data scaling methods used for the normalization in the (0,1) range of three building daily load profiles. In particular Fig. 7 (a) shows the three load profiles in their original form, Fig. 7 (b) shows the load profiles after max normalization and Fig. 7 (c) shows the same three profiles transformed trough a min-max normalization. In particular, profile 1 and 2 are the closest profiles in Fig. 7 (a) and this is essentially due to their similar magnitude. On the contrary the min-max normalization (Fig. 7 (c)) tends to emphasize only macroscopic shape similarities between the profiles completely neutralizing magnitude effects. In between, max normalization (Fig. 7 (b)) makes the profiles comparable (transforming them in the same variability range) while preserving some features of their original magnitude (e.g., min/max ratio). From the examples provided, it is clear that data scaling represents an important task in preparing data especially for pattern recognition procedures for which the concept of data similarity plays a key role.

The data type transformation is another pre-processing task useful for preparing the data in a suitable format for the application of specific data analytics algorithms. The most common data type transformation consists in transforming numerical data into categorical ones (e.g., High, Medium, Low). As a reference, algorithms such as association rules, can only manage data in categorical format for finding robust relations between itemsets. Available methods for data type transformation include equal-frequency binning, equal-interval binning, and entropy-based discretization [1]. The last pre-processing task is data reduction. In the case of time series, sampling techniques are commonly used for the reduction of the observation sampled at a specific time frequency related to each variable. In some cases, both reduction and transformation of data is needed for specific mining purpose. To this aim, techniques such as Principal Component Analysis, Curvilinear Component Analysis, Sammon Maps, frequency-domain analysis or wavelets [25,26] can be employed. Such methods consist in the elaboration of the initial data into a new vector subspace, for which it is not possible to directly represent the specific properties of the transformed data. However the physical meaning of the original data is lost when they are used [27].

An effective solution is to use more sophisticated data preparation methods that allow to address both tasks while preserving the easily interpretable nature of data. For what concerns time series, one of the most used technique, is the so-called Symbolic Aggregate approXimation (SAX). SAX is an emerging technique in time series analytics that is one of the focus of this research study. For the sake of completeness, Section 1 reports an overview on Symbolic Aggregate approXimation (SAX) with the aim of discussing its advantages and limitations.

2.1.1.1 Data reduction and transformation in time series: Symbolic Aggregate approXimation (SAX)

SAX is one of most promising techniques available to reduce and transform time series, while preserving key information. It is based on the reduction of the time series through a piecewise technique and on its transformation into a symbolic string. The method makes it possible to discretise the time series on the time axis in non-overlapping time windows of equal length by implementing a PAA technique (Fig. 8). PAA performs a constant approximation of the data by replacing the values of the original time series that fall into the same time window with their mean value. In order to transform the PAA results into a symbolic string, the amplitude of the variable (vertical axis) has to be divided into a number of regions that are defined by the analyst. A symbol is associated to each region, and this allows the PAA segments to be encoded. A simple way of addressing this task was proposed in [28], that is, by means of the standardisation of the original time series with a Z-score transformation.

In this way, the desired number of regions is identified, through the definition of breakpoints, on the basis of a hypothesis of a Gaussian distribution. The regions identified, by means of the breakpoints, on the amplitude space of the time series have equal probability of occurring. An example of a breakpoint table is reported in Table 1.

Table 1 - Table of breakpoints for alphabet size A = 3, 4, 5 calculated from a standard Gaussian distribution

		Alphabet size A		
		3	4	5
Breakpoints β	β1	-0.43	-0.67	-0.84
	β2	0.43	0	-0.25
	β3	-	0.67	0.25
	β4	-	-	0.84

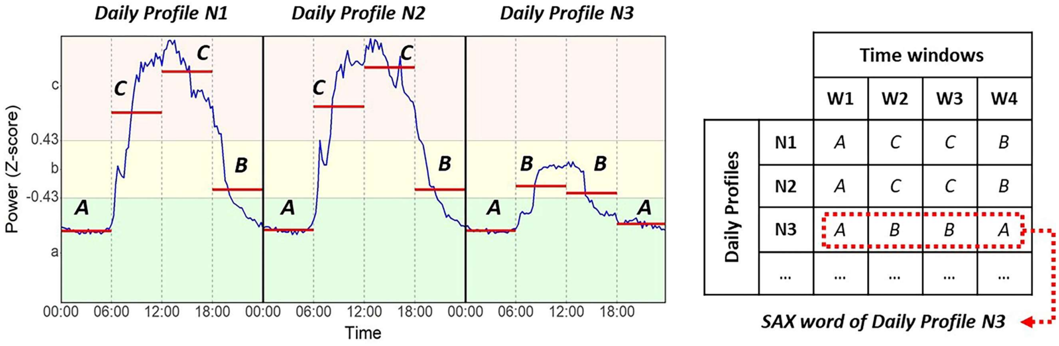
The set of breakpoints β = {β1, …, βA-1} (in Z-score) is calculated according to the chosen alphabet size, in this case A, which corresponds to the desired number of regions and hence to the symbols needed to encode the time series.

However, if the distribution of the time series is not Gaussian, the above explained process, in which a standardised lookup table (Table 1) is used, may generate unequal probability regions on the amplitude axis. If the hypothesis of normality is not satisfied, it is crucial to analyse the actual distribution function of the variable under investigation to find regions with equal probability [25] or regions where the values of the amplitude axis occur frequently [29].

After the identification of the regions, the obtained symbols are concatenated to form a symbolic string. The subsequent step consists of chunking the entire string into a set of N symbolic sub-strings, each with a reference time length T (specified a-priori by the analyst). Each substring contains a certain number of time windows. In this way, the time series data is transformed into a series of continuous symbolic sub-strings that are called SAX words. In short, the SAX technique requires three input parameters:

- the definition of the reference time length *T* of the *N* sub-strings,
- the number *W* of time windows that compose the *N* sub-strings, and
- alphabet size *A* to convert the result of the PAA technique into a symbolic string.

Fig. 8 shows the output of a SAX process applied to the energy consumption time series of a building, with N = 3 days, the daily length T = 24 hours, the number of time windows W = 4 and alphabet size A = 3. The right-hand side of the figure shows a sketch of the NxW matrix in which the evaluated SAX words are stored for the successive analysis. The N1 and N2 daily profiles of two working days are encoded with the same SAX word (i.e., A-C-C-B), while daily profile N3, pertaining to a Saturday, is characterised by a different SAX word (i.e., A-B-B-A), which denotes a lower electrical demand for time windows W2, W3 and W4.



		Time windows			
		W1	W2	W3	W4
Daily Profiles	N1	*A*	*C*	*C*	*B*
	N2	*A*	*C*	*C*	*B*
	N3	*A*	*B*	*B*	*A*
	...	...	...	...	...

Fig. 8 - SAX transformation of a three-day length time series (W=4, A=3) [21]

A further advantage of the SAX transformation is the possibility of computing a distance measure between the SAX words to perform a clustering analysis [25]. To this aim, Lin et al. [28] defined the MINDIST function. Considering two SAX words of the same length, their overall distance is defined by the lower bounding approximation to the Euclidean distance. The ‘lower-bound distance’ corresponds to the distance between the lower limit of the Z-score interval of the symbol located at greater amplitude and the upper limit of the Z-score interval of the other symbol. In other words, the distance between two equal or consecutive symbols (e.g., “a” and “b”) is 0, while MINDIST $\neq$ 0 if the symbols are at least two alphabets apart (e.g., “a” and “c”) [28].

Although SAX introduces certain advantages as far as dimensionality reduction is concerned, the selection of the input parameters W (i.e., time windows) and A (i.e., symbol ranges) is an essential step. Fig. 9 shows two daily load profiles reduced and transformed through Symbolic Aggregate approXimation. For both profiles W and A are set equal to four and three, respectively. In particular, the daily load profiles are encoded with the same SAX word (i.e., a-c-c-a) even if they have significantly different shapes. Indeed, the incorrect number of time windows and symbols, strongly affects the quality of the data reduction and transformation in terms of information loss due to the approximation of the original time series.

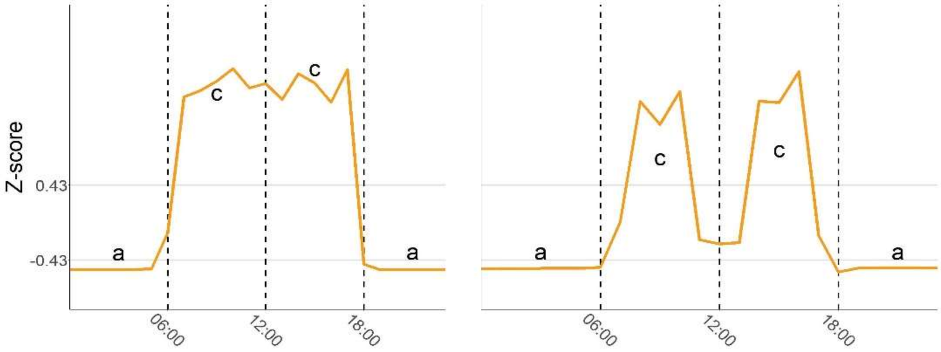


Fig. 9 - Comparison between two load profiles with different shapes encoded with the same symbolic string

In order to face the issue different approaches were proposed in the literature [2,30]. For example, the genetic algorithm NSGA-II was used in [31] to optimise, for a daily length reference period, the number of SAX words generated by setting different alphabet sizes and numbers of time windows. The objective was to maximise data accuracy and compression and to minimise the complexity of a time series transformation, in terms of the number of different generated SAX words. In the literature, different modifications to the original SAX process were proposed for improving its performance and better handling the phase of symbolic encoding of data. In the following, the main characteristics of an enhanced SAX algorithm (Adaptive Symbolic Aggregate approXimation) is described.

2.1.1.2 Adaptive Symbolic Aggregate approXimation (aSAX)

The adaptive symbolic aggregate approximation introduced by Ninh et al. [32] is based on the original SAX, but an adaptive algorithm is used for breakpoint identification. These adaptive breakpoints are evaluated through a pre-processing phase, which is based on the K-means clustering technique. In this case, the hypothesis is not of equal probability, but consists of finding, for a fixed predetermined number of symbols, the partitions that minimise the total representation error after the SAX transformation. The algorithm consists of an iterative process, which starts from the initial conditions, labelled with the superscript (0), and evolves with the generic iteration, labelled with the superscript (*j*). The algorithm inputs are the alphabet size *A* (which corresponds to parameter *k* in the K-means algorithm), and the initial breakpoints $\beta_i^{(0)}$, for *i* = 1, …, *A*, evaluated under the equal probability hypothesis as an effective initialisation of the cluster intervals. The generic step of the algorithm then computes the centroids $c_i^{(j)}$ of all the PAA segments, x_n, that fall between two consecutive breakpoints $[\beta_i^{(j-1)}, \beta_{i+1}^{(j-1)}]$, for *i* = 1, …, *A*−1, as follows (Eq. 1):

$$c_i^{(j)} = \frac{1}{N_i} \sum_{x_n \in \left[\beta_i^{(j-1)}, \beta_{i+1}^{(j-1)}\right]} x_n$$

Eq. 1

where N_i is the total number of segments included in the $[\beta_i^{(j-1)}, \beta_{i+1}^{(j-1)}]$ interval. Subsequently, the new breakpoints $\beta_i^{(j)}$ are moved to the centre of two consecutive centroids. The total representation error, that is, the total residual sum of squares between all the samples and their centroids, is then computed as (Eq. 2):

$$RSS_{tot}^{(j)} = \sum_{i=1}^{A-1} \sum_{x_n \in \left[\beta_i^{(j-1)}, \beta_{i+1}^{(j-1)}\right]} \left(x_n - c_i^{(j)}\right)^2$$

Eq. 2

At the end of each j iteration, it is possible to evaluate the relative representation error reduction,

$\varepsilon^{(j)}$, as follows (Eq. 3):

$$\varepsilon^{(j)} = \frac{RSS_{tot}^{(j-1)} - RSS_{tot}^{(j)}}{RSS_{tot}^{(j-1)}}$$

Eq. 3

The adaptive breakpoint search process is stopped by fixing a minimum threshold, $\bar{\varepsilon}$, when $\varepsilon^{(j)} < \bar{\varepsilon}$. The above described process converges rapidly, due to the initialisation of the effective equally probable breakpoint search process and the reduced dimensionality of the PAA segments. The SAX approximation error of the symbols is then reduced, step by step, until the best set of breakpoints is obtained. The aSAX algorithm proved to be effective in overcoming limitations of the original SAX process significantly reducing the transformation error of encoded time series. Good evidence of this is found in the application of such algorithm in the methodological framework proposed and discussed in section 3.3.3.

2.1.2 Data segmentation

Data segmentation is a fundamental phase for enhancing knowledge extraction from massive databases considering that makes it possible to find more homogenous sub datasets according to the specificity of their own features. In fact, when the heterogeneity among the data points within each sub-dataset is low, a more effective recognition of typical and infrequent patterns can be performed (essential for load profiling process and anomalous energy trend detection). Approaches for the segmentation of data can differ and can be based on domain expertise, statistical methods or data analytics algorithms [11]. A high expertise is required to the analyst in order to determine the segmentation approach to be adopted. For example, with reference to energy consumption data of buildings, a typical expert segmentation consists in separating weekdays data from weekend/holidays data [19] due to the different load conditions that occur during these periods. Depending if the analysis is or not performed on thermal sensitive energy data, the assumptions that need to be taken into account for an expert segmentation could be significantly different. In that case, the use of simple statistical methods (e.g., Pearson correlation, analysis of variance) can support the segmentation process by extracting useful correlations or performing significance tests. For example, the statistical approach can be effective for sub-setting energy consumption data on the basis of a climate-driven segmentation considering a seasonal effect (winter and summer season). However, expert-based and statistical methods are not always able to properly segment data, and in the case of building energy consumption it is due to the existence of various load conditions not always easily inferable.

In this perspective, in order to avoid the identification of noisy or not homogeneous subsets of data, more and more analysts are relying on the application of data analytics techniques, such as cluster analysis, for performing data segmentation [33]. Unlike the expert segmentation, cluster analysis allows homogeneous sets of data to be discovered with an unsupervised approach and without leveraging on a-priori knowledge.

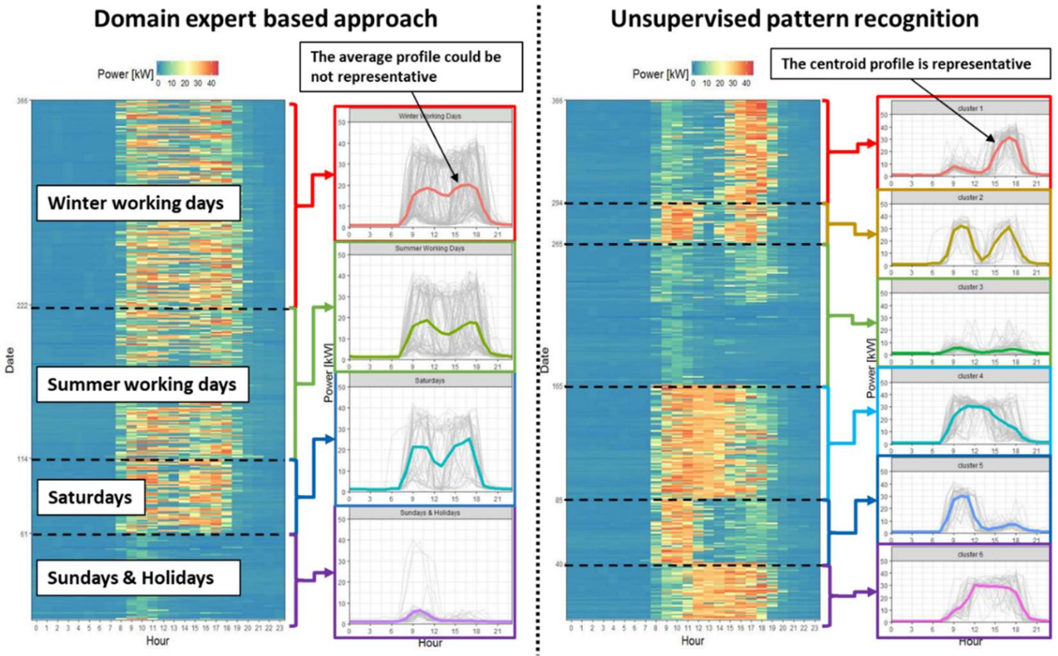


Fig. 10 - Comparison between domain expert based and pattern recognition-based segmentation of daily load profiles of a building

In Fig. 10 is reported an example of very effective data segmentation carried out through the application of a hierarchical clustering algorithm on daily load profiles of a building. The left side of Fig. 10 shows the results obtained by means of a domain expert segmentation (i.e., winter working days, summer working days, Saturdays and holidays) that lead to the identification of groups with low internal similarity. In this case, assuming the average profile as the representative statistical object of each group can produce significant information losses. On the other hand, the segmentation performed through the unsupervised pattern recognition technique (Fig. 10), produces groups of statistical objects (i.e., load profiles) with high internal similarity for which the average profile (i.e., centroid) is representative of the subset of data considered. In each homogeneous group further analysis can be conducted in a later stage for more specific objectives.

As a result, an effective data segmentation of data allows typical and infrequent patterns to be extracted and easily distinguished in massive datasets, significantly supporting the analyst in his/her inference process.

2.1.3 Knowledge discovery process

The knowledge discovery phase covers the actual mining purposes of the whole knowledge extraction process. A wide range of data analytics techniques can be used, and new data mining and machine learning algorithms are emerging in the research field of Artificial Intelligence (AI). The selection of the most appropriate algorithm depends on the problem under investigation, the level and quality of available data and the degree of domain expertise.

According to the literature data analytics techniques can be classified into two main categories, i.e., supervised learning and unsupervised learning techniques [34]. On one hand supervised learning techniques aims at modeling the relationship between output and input variables learnt from a training dataset of historical data. On the other hand, unsupervised learning techniques are not employed for achieving an explicit mining target but aim at automatically extracting underlying and hidden data structures that exist between variables in a dataset [7,35].

A wide range of both supervised and unsupervised data analytics techniques were proposed in the literature to find and model patterns representing interesting knowledge implicitly stored in massive data repositories. As a reference Fig. 11 reports the main data analytics techniques with reference to their supervised (classification/regression) or unsupervised (clustering, association analysis) nature.

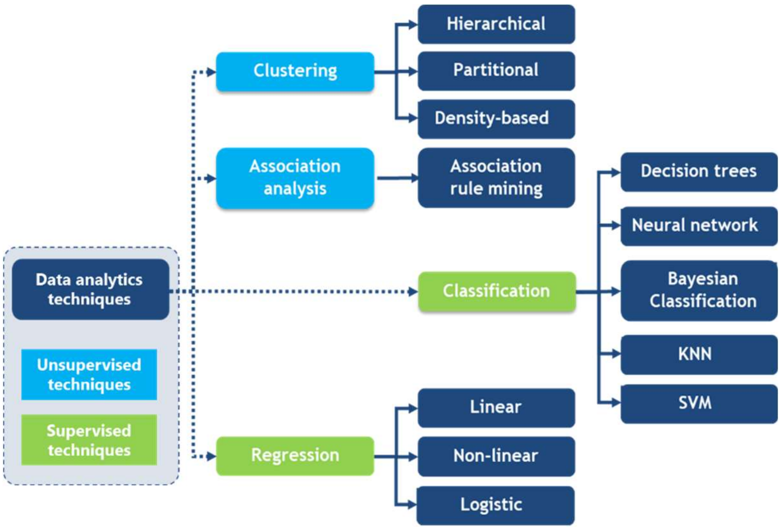


Fig. 11 - Classification of the main supervised and unsupervised data analytics techniques

Although each data analytics technique represents a powerful tool of analysis, its fully exploitation is often related to its combination with other techniques in a multi-step framework of analysis. In energy and building applications unsupervised and supervised techniques are usually employed in sequence (e.g., clustering-then-classification) or parallel (e.g., ensembling of regression models) for achieving demanding knowledge discovery targets. As a result, the development of robust methodological frameworks for building applications requires excellent transversal skills that involve both building physics and data science domain. In this perspective the following sections provide a gentle introduction to the main supervised and unsupervised data analytics techniques that demonstrated, in the literature, to be dominant in the "energy data analytics" field of application (i.e., classification/regression, clustering, association rule mining). Each technique is discussed by highlighting its mining purpose, the main input parameters to be set by the analyst and the way used for setting them. In addition, a specific focus is provided for the algorithms employed for the development of the DSS tools presented in chapters 3 and 4.

2.1.3.1 Supervised data analytics techniques

2.1.3.1.1 Classification and regression

Classification and regression models are the two families of learning algorithms used for developing descriptive or predictive models from a collection of records. Each record can be expressed as a tuple (**x**,y), where **x** represents the explanatory attribute set while y is the target attribute. In particular, the type of target attribute is the key factor that distinguishes classification from regression models. Models designed for categorical target attribute are classification models, on the other hand, models in which y is a numerical continuous attribute are regression models [36]. Classification and regression models include various algorithms as decision tree, neural networks, and support vector machines. Each algorithm employs different learning processes to develop models with high accuracy and generalization capability, i.e., models that are capable to accurately predict previously unknown records. Generally, the development of classification and regression models is performed by splitting the available dataset into *training set*, which is used in the construction phase of the model, and *test set* for validation [1].

Support Vector Machines (SVM) [37] were first proposed in statistical learning theory and can be used for developing both classification and regression model (called Support Vector Regression SVR for regression). SVM is able to identify the optimal hyperplane for separating two or more classes by maximizing the margin between their closest data points. If a linear separator cannot be identified, data are usually transformed into a higher-dimensional space by means of kernel function that makes them linearly separable [38]. The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with the difference that the main objective is to minimize residuals instead of misclassification error. However, the main concept behind is the same: identify the hyperplane which maximizes the margin but considering a tolerance ε of the error.

Artificial Neural Networks (ANN) [37] simulate biological neural systems. The network consists of several layers: an input layer, *n* hidden layers, and an output layer. Each layer is composed by nodes that are called neurons. Each neuron in a layer takes as input a weighted sum of the outputs of all the neurons in the previous layer, and it applies an activation differentiable function (sigmoid, tangent) to the weighted input. The network is trained with back propagation of the error and iteratively updates the weights of each neuron for minimizing residuals or misclassification error (depending on whether it is a classification or regressive ANN). The updating of neuron weights is performed in the backward direction, that is, from the output layer through each hidden layer down to the first hidden layer [1,38].

Despite those algorithms (i.e., SVM and ANN) can achieve very high accuracy in both classification and regression problems, the models are not easily interpretable for the final user. In order to overcome this limit in the knowledge discovery process, interpretable models such as decision trees are often used. In the following, the main characteristics of two decision tree algorithms (Recursive partitioning tree and evolutionary tree) employed in the methodological frameworks proposed in sections 3.2, 3.3, 3.4 and 4.1 are presented and discussed with the aim of better introduce their advantages.

2.1.3.1.1.1 Classification And Regression Tree (CART) based on recursive partitioning algorithm

Decision trees are machine learning algorithms capable to accomplish both regression and classification tasks through the recursive splitting of the records, included in a dataset, into purer subsets called nodes.

Classification And Regression Tree (CART) is a specific kind of decision tree that is based only on binary splitting. Regardless the learning algorithm considered, decision tree has three types of nodes: the root node that contains the whole learning sample (i) the internal nodes that contain purer subsets of the whole learning sample and are splitted into two child nodes (ii) leaf or terminal nodes that are child nodes pure enough to not be further splitted [39]. In this way, the decision trees output can be translated into a hierarchical tree structure composed by nodes and directed edges (i.e., branches) as showed in Fig. 12. In particular, the leaves represent the predicted class labels/numerical values of the target attribute and the branches represent the conjunctions of the explanatory attributes leading to those class labels/numerical values.

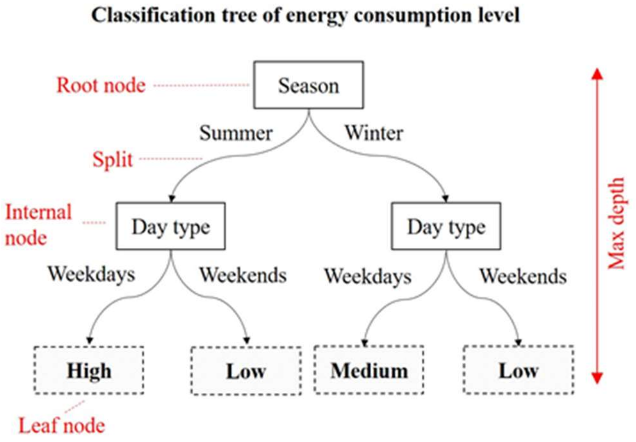


Fig. 12 - Example of decision tree representation [22]

The development of classification/regression tree, as for all predictive models, unfolds through two steps: training and testing of the model. Initially, all the records are grouped in the root node and iteratively the algorithm evaluates the best segmentation of the dataset using a predictor attribute, that minimise the average impurity measure of the child nodes after the split (e.g., Variance, Gini index, Entropy). If there are no stopping rules set by the analyst, the classification/regression tree continuously grows until the impurity in the leaf nodes (in the classification case) or the variance (in the regression case) is zero. In order to avoid this condition of model overfitting, diverse appropriate early stopping criteria can be set in advance (e.g., minimum number of cases into parent and child nodes, maximum tree depth (see Fig. 12), minimum reduction in node impurity/variance after a split). Although the early stop criteria are satisfied, the decision tree could continue to be quite large and/or complex. To this purpose, to set the right tree size, by reducing branches and leaf nodes, it is possible to define a cost-complexity parameter α that can optimize the trade-off between the cost of misclassification/residual sum of squares and the tree complexity [39].

According to [39] starting from T_{max} a sequence of pruned sub-trees T_{max}, T_1,..., T_n exists, where T_{max} corresponds to the fully-grown tree. For any subtree T < T_{max} the number of final nodes |T| correspond to the tree complexity. Then the complexity parameter α (between 0 and ∞) represents the penalty of adding other nodes that do not contribute significantly to the improvement of the overall prediction. Through a linear combination of the misclassification cost of the tree R(T) and its complexity |T|, the cost-complexity $R_\alpha(T)$ can be measured as (Eq. 4):

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right|$$

Eq. 4

Then for each value of α the subtree T(α) < T_{max} which minimises $R_\alpha(T)$ can be found, where (Eq. 5):

$$R_\alpha(T(\alpha)) = \min_{T<T_{max}} R_\alpha(T)$$

Eq. 5

Although α is a continuous variable, there are at most a finite number of subtrees of T_{max}. Thus, the pruning process produces a finite sequence of smaller and smaller subtrees T_l, T_2, T_3, … T_n until the root node is reached. Obviously, a direct search through all possible subtrees to find the minimizer of R(T) is time consuming. For this reason, weakest link cutting method is often used [39]. The weakest link cutting method allowed to find the sequence of α_i which result in the sequence of the smallest minimizing subtrees $T_{\alpha i}$. In the following, the definition of cost-complexity is extended to a single node of the tree and then for a single branch coming out of a node. In particular, for any node $t \in T_{max}$ (Eq. 6):

$$R_\alpha(t) = R(t) + \alpha$$

Eq. 6

Also, for any branch T_t (Eq. 7):

$$R_\alpha(T_t) = R(T_t) + \alpha|T_t|$$

Eq. 7

For α = 0, the inequality $R_0(T_t) < R_0(t)$ is always satisfied. If α gradually increase up to $\alpha = \frac{R(t)-R(T_t)}{|T_t|-1}$, then $R_\alpha(T_t) = R_\alpha(t)$. The node for which this equality is true at the smallest α, is defined as the weakest link and the branch T_t is pruned. This procedure is repeated iteratively until the last subtree collapse in the root node. At the ending of the iterations the final pruned tree can be evaluated by plotting the subtree risks versus their complexity parameters.

Typically, this plot has an initial sharp drop followed by a relatively flat plateau region and then a slow rise. When the decision tree is subjected to a validation procedure (e.g., k-fold cross-validation) it is also possible to compute a standard error for each sub-tree risk. The choice of the best subtree starts from the plateau region of the subtree risks in which is included the minimum cross validated risk achieved. In fact, any sub-tree risk within one time the standard error of the achieved minimum risk can be considered as being equivalent to the minimum [21]. Then the simplest model (with the minimum number of final nodes) among all the identified sub-trees in the plateau region, is chosen.

K-fold cross-validation is usually used for such algorithms. For this kind of method, the original sample of data with M number of objects is divided into k equal sized subsamples. For the k subsamples evaluated, a single subsample is selected as validation dataset for testing the model, and the remaining k-1 subsamples are used for the training. This process is then repeated k times, using a subsample at a time as testing.

As a reference, the explained procedure is employed for developing both classification and regression trees in the data analytics based methodologies presented in sections 3.2, 3.3, 3.4 and 4.1.

2.1.3.1.1.2 Classification And Regression Tree (CART) based on evolutionary partitioning algorithm

As explained in the previous section, the training of a decision tree through recursive partitioning method consists of a forward step-wise approach where at each parent node the best split is evaluated maximizing homogeneity in its child nodes. However, this learning technique leads to solutions that are locally optimal, since the splits are evaluated for minimising a loss function in the next step only [39]. An alternative learning process consists of searching globally optimal trees for example by means of an evolutionary approach. The main steps of the algorithm can be summarised as follow [40]:

- *Setting of the model parameters*: During this step the parameters of the model are set by the analyst. The main parameters are the maximum depth of the trees, the minimum number of observations in a leaf node, the size of tree population (Θ), the variation operator probabilities, the number of iterations, the evaluation function and the complexity parameter.
- *Initialization*: During this step the population of Θ trees is initialized. Each tree is initialized with a root node split that is randomly generated from the input variables.
- *Survivor selection*: In every iteration, each tree (parent solution) is selected once to be modified (generating an offspring solution) by one of the variation operators (i.e., split, prune, major split rule mutation, minor split rule mutation, crossover). The population size Θ remains constant during the evolution and only a fixed subset of the candidate solutions can be stored for the next iteration. The algorithm uses a deterministic crowding approach, where each parent solution competes with its most similar mutation (offspring) for being stored in the population at iteration i_{+1}. In a classification problem the algorithm evaluates among the population of parents and offsprings, the best trees in terms of classification accuracy and complexity.
- *Termination*: The tree with the highest quality according to the evaluation function is returned as the final output of the algorithm at the end of the n iterations. For a large number of iterations (e.g. 10000 iterations) the algorithm terminates when the quality of the best 5% of trees in Θ remains stable for 100 iterations, but not before the ending of 1000 iterations.

The core of the evolutionary learning process consists in the five variation operators implemented by the algorithm at each learning iteration [40]. The main principles of the operators are described below:

- *Split*: the operator randomly selects a leaf node of a tree and assigns a split rule to it. The split rule is randomly generated respect to the input split variable v_r and split point s_r. As a consequence, the leaf node becomes a parent node after the generation of two new child nodes;
- *Prune*: the operator randomly selects an internal node of a tree which has two leaf nodes as successors and prunes it;
- *Major split rule mutation*: the operator randomly chooses an internal node of a tree and modifies the split rule respect to input split variable v_r, and the split point s_r;
- *Minor split rule mutation*: The operator randomly chooses an internal node of a tree and modifies the split rule only respect to the split point s_r of the input variable v_r;
- *Crossover*: The operator randomly selects subtrees from two trees and exchanges them creating two new trees.

It is important to highlight that the globally optimal decision tree algorithm could lead to slightly different solutions depending on the random initialization of the population Θ and the probabilities of variation operators to be applied at each iteration. For this reason, a sensitivity analysis on the tuning of model parameters is highly recommended.

As a reference, the introduced algorithm is employed for developing a classification tree in the data analytics based methodology presented in section 3.4.

2.1.3.2 Unsupervised data analytics techniques

2.1.3.2.1 Clustering

Cluster analysis belongs to the family of unsupervised data mining techniques used for conducting exploratory analysis on massive datasets.

The final aim of clustering is grouping a collection of data objects into subsets (clusters) on the basis of their similarity in a n-dimensional space. A good cluster analysis should lead to the identification of sub datasets that are characterised by high intra-class and low inter-class similarity [1,37,38]. A wide variety of clustering procedures has been introduced in the scientific literature and it is already available on different statistical software. The effectiveness of the different methods was widely discussed in the literature also considering the effect of data normalization (e.g., max normalization) and data reduction (e.g., symbolic aggregate approximation, principal component analysis) techniques on the final results [25].

The most used clustering techniques in the literature are partitive, hierarchical, and density-based algorithms. Partitive clustering (e.g., K-means [41], K-medoids [42]), consists in a division of the data objects into non-overlapping subsets (i.e., clusters) such that each data object can be included only in one subset. K-means is a well-known partitive algorithm that is used for grouping data objects in a pre-determined number of K clusters which are represented by a prototype object called centroid (i.e., mean of the points in the n-dimensional space).

The first step of K-means consists in the setting of the number K of clusters desired to which corresponds a prototype object (centroid) randomly located in the n-dimensional space [43]. Each object in the dataset is then assigned to the closest centroid, and each group of objects assigned to the same centroid is a cluster. The centroid of each cluster is then recalculated as the average of all the object assigned to the cluster. This process is repeated until the objects do not change cluster anymore, and the centroids do not change position.

Given that K-means minimizes the within-cluster sum of squares, this algorithm is particularly effective for the identification of spherical-shaped clusters. However, the randomly initialization of centroid positions may negatively affect the whole iteration process of the algorithm leading to non-optimal solutions. K-means is also sensitive to the presence of outliers, heterogeneous densities of objects, and non-globular shapes of clusters [1]. Despite of all these limitations, K-means is computationally fast and easy to be implemented.

In density-based algorithms (e.g., DBSCAN [44]), a cluster is defined as a dense area of data objects surrounded by an area of low density [43]. DBSCAN is a well-known density-based algorithm capable to evaluate dense groups of objects in databases through the setting of two input parameters [1,38,43]:

- Eps: that is the search space radius of neighbours around a data point p
- minPts: that corresponds to the minimum number of data points that a n-dimensional sphere of radius Eps should contain to define a dense region.

Once these parameters are defined, the algorithm scans the data objects in the dataset and classifies them as (i) core points, (ii) border points or (iii) outliers. In particular, a core point is included in a dense region where at least minPts points are within the distance Eps. Border points are located on the edge of dense regions and are included in n-dimensional spheres of radius Eps that group less than minPts point but at least one core point. As a consequence, all the points that are not reachable from any other point are classified as outliers. Any two core points that are within distance Eps are grouped together in the same cluster. Any border point close enough to a core point is put in the same cluster as the core point. Outliers are instead isolated [1,38,43]. Differently from K-means, DBSCAN can handle clusters of non-globular shapes and outliers, thus increasing cluster homogeneity. The number of clusters is not required as an input parameter, but the user should specify the Eps and MinPts parameters [1,38,43].

Hierarchical algorithms are also widely used for performing data clustering. Differently from partitive algorithms, hierarchical ones allow clusters to have sub-clusters. In this way objects are organized as a set of nested clusters that can be represented with a tree-like structure (i.e., dendrogram). Hierarchical clustering techniques can be classified in agglomerative and divisive algorithms. On one hand, agglomerative algorithms start with the objects as singletons (clusters with one object) and at each step merge the closest pair of clusters according to a proximity measure [43]. On the other hand, divisive algorithms follow the opposite approach recursively splitting objects starting from an all-inclusive cluster.

In agglomerative hierarchical clustering the proximity between two clusters can be computed in different ways according to the linkage method selected [43]. In particular, single linkage method merges clusters assuming cluster proximity as the distance in the n-dimensional space computed between the closest two objects that are in different clusters. Conversely, complete linkage method considers the proximity as the distance between the farthest two objects in different clusters. Average method, instead, defines cluster proximity as the average distance between each object in one cluster and every object in the other cluster. An alternative is the Ward method, that differently from the other linkages, assumes that a cluster is represented by its prototype object (i.e., centroid) and considers the proximity between two clusters in terms of the increase of the sum of square error that results from their merging.

On the basis of the linkage method employed clustering results can significantly vary. Fig. 13 shows the dendrograms of three clustering solutions obtained from the same dataset considering single, complete and average linkage. In particular, on the y-axis there is the height of each node in the plot that corresponds to the fusing distance between its two nested sub-clusters. Cutting all the dendrograms at four clusters, it is possible to understand that the selection of the linkage method highly impacts the cardinality and dispersion of the obtained clusters.

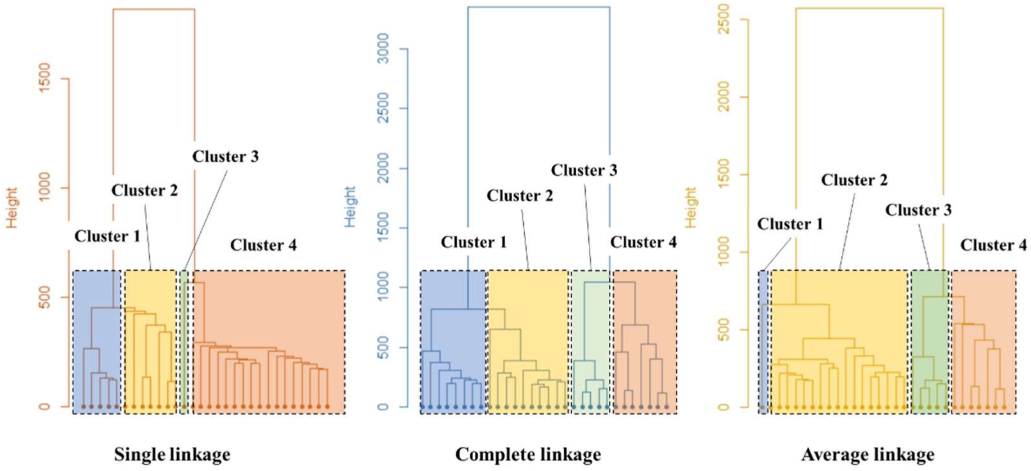


Fig. 13 - Comparison between three dendrograms of hierarchical clustering algorithms assuming different linkage methods (i.e., single, complete and average method)

In the following, the main characteristics of a very effective prototype-based partitive clustering algorithm (Follow The Leader) are discussed. The follow the leader algorithm makes it possible to catch the advantages of k-means without setting a-priori the number of clusters to be found.

2.1.3.2.1.1 Follow the leader clustering algorithm

"Follow The Leader" (FTL) method [45,46] is a partitive clustering technique that differently from K-means does not require the a-priori definition of the number of clusters K, but it is initialized selecting a maximum distance threshold ρ. The dataset is sequentially scanned by the algorithm over a number n of iterations, large enough to ensure the stabilization of the clustering results. In the first iteration, the FTL approach is used to define, as a first attempt, the total number of clusters K and the number of objects assigned to each cluster. During the iterations, if the distance between an object and the cluster centroids computed until that iteration is lower than ρ*, the object will be assigned to the cluster of the closest centroid otherwise a new cluster with one single element is generated. Indeed, the number of clusters and the number of objects belonging to the same cluster may change until the algorithm converges to a stable solution.

Given that the clustering analysis is an unsupervised data mining technique the input parameters are a-priori set by the analyst. For this reason, usually a cluster validity index is needed for supervising their tuning. The selection of an optimal value of input parameter can be then conducted with a "trial and error" procedure. For FTL different values of ρ can be tested and the results in terms of cluster separation and cohesion be compared.

One of the most used cluster validity metric is the Davies-Bouldin Index (DBI) [47]. DBI is based on the concept that for a good partition, inter cluster separation as well as intra cluster cohesion should be as high as possible. As a reference for each clustering result obtained from the setting of different of ρ, the DBI is evaluated according to the following equation (Eq. 8):

$$\mathrm{DBI}(\rho) = \frac{1}{\mathrm{K}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \max_{\mathrm{k} \neq \mathrm{l}} \left(\frac{\delta_{\mathrm{k}} + \delta_{\mathrm{l}}}{\mathrm{d}_{\mathrm{k,l}}} \right)$$

Eq. 8

Where:

- K is the final number of clusters fixing a certain value of the input parameters (i.e., ρ in "follow the Leader" clustering).
- $d_{k,l}$ is the Euclidean distance between centroids of the clusters C_k and C_l.
- δ_k, δ_l are the standard deviations of the distances of objects in clusters C_k and C_l.

The value of the parameter ρ which minimises DBI is considered as the value that leads to the optimal cluster solution. This can be considered as a general validation procedure of clustering results that can be extended also to other kind of algorithms.

As a reference the Follow The Leader method is employed in the data analytics based methodology presented and discussed in section 3.4.3.

2.1.3.2.2 Association rule mining

Association rule mining (ARM) is an unsupervised data mining method to identify all associations and correlations between attribute values in a dataset [48]. The output is a set of association rules that are used to represent patterns of attributes that are frequently associated together (i.e., frequent patterns).

Let $I = \{i_1, i_2, \dots i_d\}$ be the set of all items in a dataset and $D = \{d_1, d_2, \dots d_d\}$ be the set of all transactions. Each transaction d_i contains a subset of items chosen from I. In association analysis, a collection of items is named itemset and the transaction width is defined as the number of items present in a transaction. A transaction d_j contains an itemset X if X is a subset of d_j. An important property of an itemset is its support count, that corresponds to the number of transactions that contain a specific itemset. The support count, σ(X), for an itemset X can be expressed as follows [48] (Eq. 9):

$$\sigma(X) = |\{d\_i \,|X \subseteq d\_i, d\_i \in D\}|$$

Eq. 9

Association rules are usually represented in the form X → Y, where X (also called antecedent) and Y (also called consequent) are disjoint item sets (i.e., $X \cap Y = \emptyset$). Rule quality is usually measured through rule support and confidence. Rule support is the fraction of the total number of transactions in which both the item sets X and Y occur while confidence determines how frequently items in Y appear in transactions that contain X. According to [48] Support and Confidence can be calculated with the following equations (Eq. 10 and Eq. 11):

$$\text{Support}, \mathrm{s}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{\sigma(X \cup Y)}{N}$$

Eq. 10

$$\text{Confidence}, \mathrm{c}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Eq. 11

Where N is the total number of transactions. Therefore, Given a dataset D, whose generic record is a set of items, ARM process discovers all association rules with support and confidence greater than, or equal to, minimum thresholds a-priori defined by the analyst (i.e, MinSup and MinConf).

Furthermore, in order to rank the most interesting rules, the lift index can be used to measure the correlation between antecedent and consequent of the extracted rules. Therefore the lift is intended as the ratio between the observed support to the expected support and is calculated as follow (Eq. 12):

$$\text{Lift } (\mathrm{X} \rightarrow \mathrm{Y}) = \frac{\sigma(X \cup Y)}{\sigma(X) \text{ x } \sigma(Y)}$$

Eq. 12

If lift (X → Y) = 1, itemsets X and Y are statistically independent from each other. Lift values below and above 1 show a negative and positive correlation between itemsets respectively. As a result, rules with a lift value close to 1 are not of interest.

In the context of discrete-value-transactions, association rules can be used as an efficient method for mining also co-occurrences or implications between events in the time domain (Temporal Association Rule Mining (TARM)). The application of TARM algorithms is particularly effective for extracting hidden knowledge from time series that is one of the main objectives investigated in this research study.

For the sake of completeness, the next section, presents and discusses the application association rules in time series for mining relevant pattern through a temporal-based approach. As a reference a TARM algorithm is employed in the data analytics-based methodology presented and discussed in section 4.1.

2.1.3.2.2.1 Temporal association rules

Temporal association rule mining is an extension of sequential pattern mining that is an important data mining method with broad applications that can extract frequent itemset sequences while maintaining their chronological order.

The output of this analysis consists in the automatic identification of IF-THEN rules (IF event A happens THEN event B will also happen) capable to assess how strong is the relation between events frequently associated together. According to the number of time series considered during the analysis, the ARM techniques can be categorized in univariate and multivariate association rules. The aim of extracting association rules from single time series is to identify frequent event sequences that could be useful for example in characterizing chiller operation in complex cooling systems [30]. When multiple time series are considered, ARM techniques can be further divided in intra-transactional and inter-transactional association rules respectively. The first type is aimed at discovering events, in different time series, that frequently happen at the same time. This kind of analysis based on the extraction of co-occurrence, is particularly suitable in finding rules related to the simultaneous operation of different devices or systems in buildings. In that case it would be sufficient to discretize the time series in sequences of 0-1 to preserve only the information related to the system switching ON and OFF. The second type of association rules are the most complex ones since that the co-relations are discovered assuming the existence of a time lag in the event implications. For these rules, the search space in the time domain is represented by a sliding window which length is set in advanced by the analyst. The tuning of this parameter is considered a fundamental step of the analysis due to its impact on the results in terms of total number of rules and consistency with the physical phenomenon they describe.

The problem of finding inter-transactional association rules between events is represented in the following form: $X \xrightarrow{t} Y$.

Therefore, the occurrence of the antecedent X implies the occurrence of the consequent Y within a time t. The temporal relation between events should respect two constraints: the maximum temporal lag between antecedent and consequent and their chronological order.

As regards temporal association rules, an effective approach for their automatic extraction was proposed by Zaki by means of the cSpade algorithm [49]. This algorithm extracts sequential rules considering some constraints defined by the user according to his needs. These constraints may drive the mining of frequent patterns from the database of transactions, for instance by setting the length of the sliding window or a minimum time gap between antecedents and consequents of the rules.

However, since the database of transactions, is generated by using a sample-by-sample sliding window approach, the number of the transactions N results to be very high with items mostly overlapped. For this reason, the calculation of rule support cannot be performed with the canonical formulation. In fact, the value of support calculated according to Eq. 10 can be affected by the high value of the denominator (i.e., the total number of transactions), suggesting the use of a formulation less sensitive to the sample size [50].

In the methodology proposed in section 4.1, the support of an association rule is defined as the ratio between the number of transactions that include both antecedent and the number of transactions that include at least the consequent (Eq. 13).

$$\text{Support}(\mathrm{X} \rightarrow \mathrm{Y}) = P(X, Y) = \frac{N(X, Y)}{N(Y)}$$

Eq. 13

The support calculated with Eq. 13 makes it possible to have high values of support also for large transaction datasets obtained through a sliding window. The support calculated through Eq. 13, assesses the frequency of X U Y on a smaller portion of the total number of transactions (i.e., only the transactions that include the consequent itemset Y). The support is in the range (0-1) and allows an easier extraction of rules to be assumed as reference patterns (i.e., with high support) of the occurrence of a specific condition over time (i.e., consequent itemset Y). However, the confidence can be still calculated according to Eq. 11 only if the consequent itemset Y occurs in a transaction not violating the chronological order respect to the antecedent itemset X.

Also for the inter-transactional approach, the mining of association rules can be summed up as a two-step's procedure. In a first phase, the frequent item sets with a support greater than the MinSup are extracted then the confidence is considered for filtering out rules that consist in weak implications [51].

2.1.4 Knowledge exploitation

The purpose of knowledge exploitation (so-called post-mining phase) is to select, interpret and utilize the knowledge discovered in the previous phase [23]. Different approaches are proposed in the literature to efficiently exploit extracted knowledge and to interpret the patterns discovered (e.g., by means of decision and association rules). The knowledge exploitation usually requires domain expertise to explain the knowledge discovered and to convert it into actionable measures for enhancing building energy performance (e.g., energy system rescheduling, set point optimization, diagnosis of a fault).

Knowledge extracted from energy-related data can be then transferred to different users that often have very different requirements in terms of data insight. For what DSS tools is concerned, extracted knowledge should be informative and, at the same time, easy to be understood and to be exploited in the decision-making process. To this purpose, the developed data-driven methodological frameworks in this research study (presented in chapters 3 and 4), leveraged on the application of automatic rule extraction techniques. Such techniques (e.g., association rules, decision trees) aim at extracting from large amounts of data, inference rules in form of IF-THEN implications that are able to effectively describe all the relations that exist between the variables included in the same dataset. In this way the results of the analysis can be translated in a set of interpretable decision rules that can be easily embedded in DSS, helping managers, owners or service companies in increasing awareness about the measured energy performance of their buildings/systems and achieve demanding energy management targets during daily operation. In addition, also advanced visualizations were used for improving feedback interpretability and increasing user engagement for a better exploitation of enabling tools such as DSS.

2.2 Research context: applications of data analytics technologies for building energy management

The section 2.2 introduces the research context of the dissertation. While in the previous section a general data analytics framework was discussed, in the following the main applications of data analytics-based processes for building energy management are reviewed. Such applications represent the most important functionalities that can be embedded in advanced DSS that are the focus of this study. Fig. 14 shows the conceptual framework at the base of implementing DSS for enhancing energy management in buildings. The flow starts from a set of various data analytics techniques that can be exploited for conducting analysis in the context of different applications (i.e., application layer (Fig. 14)) identified as the main functionalities of an advanced DSS. Each application has its own objective and useful knowledge is extracted accordingly. Once knowledge is extracted it is transferred to the domain expert (e.g., energy manager) for supporting him/her in the definition and implementation of effective energy saving strategies. DSS should also be equipped with verification tools capable to assess the impact of the implemented strategies for supporting the domain expert in verifying the achievement of expected targets (Fig. 14).

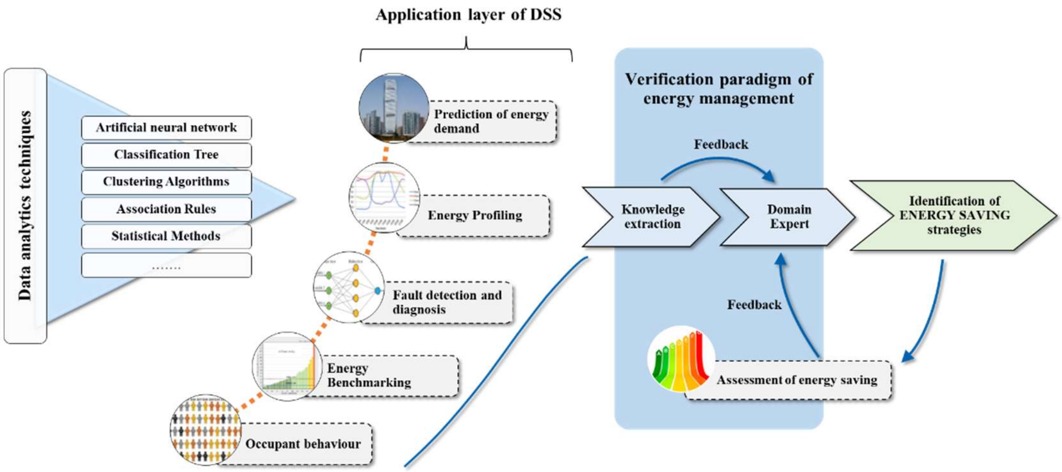

Application layer of DSS
Data analytics techniques
Artificial neural network
Classification Tree
Clustering Algorithms
Association Rules
Statistical Methods
....
Prediction of energy demand
Energy Profiling
Fault detection and diagnosis
Energy Benchmarking
Occupant behaviour
Verification paradigm of energy management
Feedback
Knowledge extraction
Domain Expert
Identification of ENERGY SAVING strategies
Feedback
Assessment of energy saving

Fig. 14 - Conceptual framework behind the use of DSS

In order to ensure the adequate robustness of the whole management process great effort has been devoted in the scientific literature for developing as possible the DSS application layer (i.e. the set of functionalities that an advanced DSS should provide).

Knowledge extracted trough data analytics-based functionalities of DSS contains information for example on how and when building energy use changes during the day with the ability to answer questions such as:

- *how much energy is expected to be consumed at different times of the day?*

- *Which are the typical energy consumption patterns of a building or energy system?*
- *Which are the unexpected/infrequent energy consumption patterns?*
- *Which is the most valuable energy saving opportunity to be investigated?*
- *Are building systems behaving as expected?*
- *How the building is behaving respect to its peers, its past or its intended energy performance?*
- *Which are the main energy consumption trends of a building portfolio?*

In this context, the following sections review analytics-based functionalities that are widely recognised as the most impacting in the building energy management through DSS (Fig. 14) such as (i) energy consumption prediction [53,54], (ii) load profiling [9,10,33], (iii) Fault detection and diagnosis [13,55], (iv) energy benchmarking [52,56,57], (v) characterization of the occupant behaviour [14,58].

In particular, section 2.2.1 presents the main modeling approaches and implications behind the development of prediction models for supporting energy management in buildings. The predictive modeling is at the basis of advanced DSS tools and enables several energy management functionalities that leverage on the estimation of building and system behavior over time such as demand response, fault detection and diagnosis, anomalous energy trend identification, benchmarking of the building energy performance, assessment of the energy saving.

Section 2.2.2 presents and discusses the main opportunities related to the application of load profiling tools for deeply characterizing energy consumption in buildings. Such tools leverage on time series analytics and pattern recognition techniques for extracting both typical and infrequent load profiles in building energy consumption time series also providing information about their shapes and magnitudes. The load profiling applications are discussed at single building level (section 2) highlighting their usefulness in detecting anomalous trends of energy consumption [21]. At building stock level, section 2.2.2.2 provides a wide overview on the main implications the load profiling analysis has on building demand response and demand side management, customer classification and energy benchmarking [10].

Section 2.2.3 discusses the implementation of fault detection and diagnosis (FDD) methodologies for enhancing building energy system performance (especially HVAC systems) during daily operation. FDD tools proved to be essential for achieving demanding energy saving targets in buildings given that up to 20% of energy consumption can be caused by incorrect system configurations and inappropriate operating procedures [59].

Section 2.2.4 discusses the exploitation of benchmarking methodologies that can be embedded in DSS tools for setting credible targets of energy efficiency e.g., through the comparison of the energy performance among similar buildings.

Section 2.2.5 discusses the importance of characterizing occupant behavior in buildings for identifying significant energy saving opportunities. One of the main opportunities is related to the analysis of occupancy patterns in buildings for optimising energy system rescheduling. As a reference, occupancy based rescheduling strategies of HVAC systems proved to be capable of generating a potential savings higher than 10% of energy consumption [14].

The final aim of this chapter is then to discuss a wide research context, for better pointing out broader challenges and opportunities related to the use of data analytics in DSS solutions for enhancing energy efficiency in buildings.

2.2.1 Prediction of building energy consumption

Over the past few decades, researchers have paid much effort to find robust solutions for improving building energy efficiency and usage through various techniques and strategies.

The prediction of building energy consumption proved to be essential for a variety of energy management applications such as demand response, demand side management, fault detection and diagnosis [60], predictive maintenance and optimal control of building systems. According to the prediction horizon of interest, predictive analysis can be categorized as follows: Short Term Forecasting (STF), Medium Term Forecasting (MTF), and Long-Term Forecasting (LTLF). Each of these categories has a different characteristic prediction horizon, which is typically included between few hours (i.e., STF) up to one year (i.e., LTF) and different possible applications can be defined accordingly [61]. The research focus is currently more devoted on short-term prediction given its close linkage to the day-to-day operations.

As a reference, peak demand prediction leverages on STF and makes it possible to inform energy manager about the occurrence of a peak in the building load. It is extremely important, for avoiding penalties in the electricity bill or load shedding and power outages in the case the generation and supply systems are not able to satisfy peak demand [62]. In this perspective, if the final user is timely and adequately informed, he/she can intervene for shifting or shaving the peak and thus avoiding the occurrence of unsatisfactory conditions. However, the use of prediction tools not only enables the implementation of effective energy management strategies during daily operation but also makes it possible to assess their actual impact on building energy consumption. When a Continuous commissioning (CC) of buildings is implemented, prediction tools are essential for benchmarking building energy consumption against its past or intended performance [63]. For example, Fig. 15 illustrates how prediction models can be used in the measurement and verification of energy saving in buildings. Data collected during pre-retrofit period are used to train and test a prediction model capable to provide a robust energy consumption baseline of the building under analysis (Fig. 15). After the implementation of a retrofit action the adjusted baseline is estimated through the model considering the boundary condition of the post retrofit period (e.g., climate condition, number of occupants). Savings or avoided energy consumption are then calculated by comparing the adjusted baseline and the post-retrofit actual data. This kind of application is crucial for partially close the decision-making process in building energy management as it gives to building owners, managers and occupants a feedback on the impact of energy conservation measures.

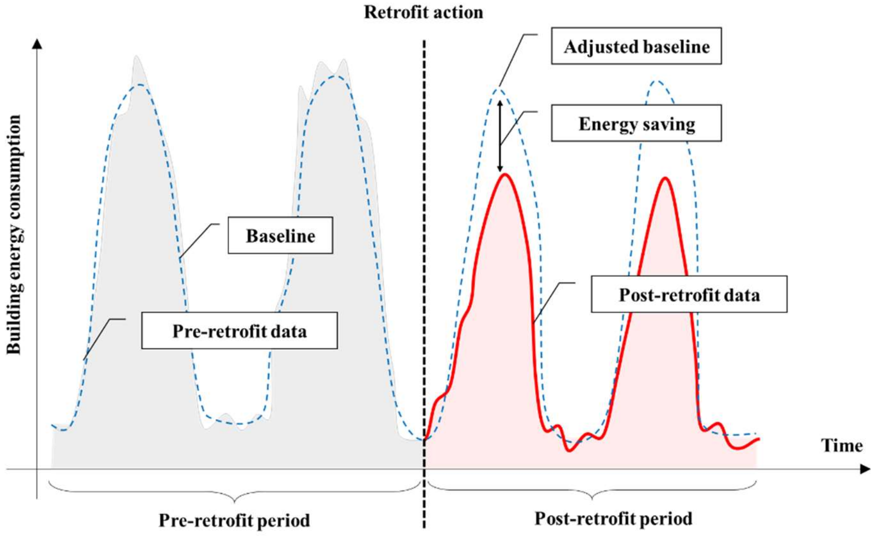


Fig. 15 - Use of prediction models for the measurement and verification of energy saving

Despite the great potential of using predictive-based management solutions, the development of prediction models is often thwarted by the high complexity of the systems inside buildings [1]. This is due to the growing variety of multi-energy plant systems, integration with renewable energy systems, type of loads (e.g., thermal sensitive electrical load) and variable occupant behaviour and presence patterns. These are distinctive characteristics of a building that together with indoor and outdoor environmental conditions (external air temperature, indoor thermal comfort requirements), make the predictive modeling a complex task.

Nowadays, there are many prediction modeling methods available for solving such issues and also achieving high accuracy [64]. A general classification of such methods can be made according to the modeling approach employed in the analysis such as white box models, grey box models and black box models [63]. A white box model is also termed as first principle-based model, which makes use of physical equations for modelling building systems and components. Conversely, a black box model uses data-driven fitting techniques rather than physical knowledge, leveraging on statistical or data analytics-based algorithms. The principle of grey box model lies in the middle between white box model and black box model, it combines both physical knowledge of the system considered and data-driven fitting techniques to derive a robust prediction model. Currently more and more researchers are exploiting advanced data analytics techniques (black box models) for accomplishing predictive tasks in building energy management, and it is mainly due to their high capability in dealing with non-linear problems [64] that often characterise building operation. Various studies have also demonstrated that non-linear techniques could outperform linear ones (e.g., multiple linear regression and autoregressive moving average) in building-related applications [53,65–67].

Artificial Neural Networks (ANN) and Support Vector Machine (SVM) were the two most widely used techniques for this kind of application [61]. Kumar et al. [68] reviewed various ANN methods, including back propagation, recurrent ANN, auto associative ANN and general regression ANN highlighting their high potential in providing robust predictions for various forecasting purposes in buildings. In thermal and electrical energy prediction, ANN was tested with different objectives. In thermal energy prediction, numerous researchers predicted space cooling load [65,69], space heating load [70,71] domestic hot-water heating load [70]. Ben-Nakhi and Mahmoud [72] adopted artificial neural networks to predict next-day building cooling load in order to optimise the HVAC thermal energy storage system operation. It was demonstrated that optimal control strategies based on such predictive modeling approach can increase the operating flexibilities while reducing the operating costs. In different applications, Support vector regression (SVR) proved to be useful for prediction purposes. In [73], good accuracy in cooling load and monthly utility bill prediction was observed. Dong et al. [73] applied SVR to predict monthly utility bills in four commercial buildings located in Singapore. The prediction was based on weather data (ambient temperature, relative humidity and global solar radiation) collected for each of the buildings analysed. The achieved accuracy was close to 96%.

Another group of non-linear models that was widely used for conducting predictive analysis in energy and building applications includes tree-based algorithms such as decision trees and random forests. When such models are based on the development of a single tree they have the advantage of being interpretable as their flowchart-like tree structure can be easily translated in a set of IF-THEN decision rules. Yu et al. [74] used the decision tree to predict and classify the building Energy Use Intensity (EUI) of Japanese residential buildings. In [75] the decision tree method was found to be suitable in improving the criteria of energy efficiency measures in building renovation. Given their nature, decision trees proved to be particularly effective in performing classification tasks. In [76–78] a decision tree was used for predicting different levels of primary energy demand for space heating of about 90,000 flats. The analysis was performed on data gathered from public available energy certificates. Capozzoli et al. [56] developed two different models for estimating the annual heating energy consumption for 80 schools in the province of Turin (Italy). The models developed were a Multiple Linear Regression (MLR) and a regression tree (decision tree with a numeric output). The accuracy of the two models proved to be quite satisfactory and the regression tree performed slightly better than MLR. However, for achieving higher performance, especially for what regression analysis is concerned, ensembling techniques were applied on tree-based models. Algorithms based on decision tree ensembling such as random forests, extreme random forests, bagging trees, gradient boosting trees demonstrated to be a valuable solution for conducting regressive prediction analysis also on data with high granularity (e.g., hourly data). In [54] 12 regression models were used for predicting hourly electrical energy consumption for 482 non-residential buildings. The results showed that decision tree-based models performed better than other models (including ANN and SVM) on two-thirds of the total cohort of buildings generally achieving high accuracy.

For what ensemble learning is concerned, Fan et al. [53] proposed a data mining approach for predicting next-day energy demand and peak power demand of the tallest building in Hong Kong. Eight predictive models including multiple linear regression (MLR), autoregressive integrated moving average (ARIMA), support vector regression (SVR), random forests (RF), multi-layer perceptron (MLP), boosting tree (BT), multivariate adaptive regression splines (MARS), and k-nearest neighbours (kNN) were developed individually and then ensembled by optimising model weights through the application of genetic algorithm (GA). The percentage errors of the ensemble models were 2.32% and 2.85% for the next-day energy consumption and peak power demand respectively, which were higher than those of individual base models.

Future trends in predictive modeling also concern with the exploitation of novel algorithms that belong to the family of deep learning techniques. Deep learning is a powerful solution which is currently used in a wide variety of data analytics tasks, such as image and language recognition. Fan et al. [66] discussed the potentials of deep learning algorithms (fully-connected autoencoders (AEs), convolutional autoencoders (CAEs) and generative adversarial networks (GANs)) demonstrating their potentials in improving the feature engineering for supporting analysts in developing robust, flexible and accurate building energy prediction models.

2.2.2 Load profiling in buildings

The application of data analytics techniques coupled with a robust physics-based expertise can effectively support the implementation of procedures or strategies aimed at enhancing the operational performance of buildings [79]. In particular, the mining of time-series data has recently gained high attention in the scientific literature as a way to describe building load patterns and the boundary conditions (e.g., weather, time period or user/customer features) influencing their particular variation over time. The electrical or thermal load time series are usually characterised by a particular trend with stochastic components and time based cycle at annual, seasonal and daily scales [80].

In the process of building load profiles characterisation (i.e., so-called load profiling), pattern recognition techniques play a key role for the identification of typical operational patterns and trends in a high-dimensional time series [17,81]. Besides this, it can help building managers investigate the discrepancies of energy use characteristics between different seasons, working day and non-working day, day and night, peak and baseload hours, etc. In the analytical process of building load profiling, time series are usually chunked into sub sequences through a fixed length window to obtain time scale-based profiles. In the majority of energy and buildings applications load profiles are usually well described on a daily scale.

The mining of load profiles is an emergent task which enables the implementation of various energy management and diagnostic strategies at both single and multiple buildings level. The process of daily load profiling primarily consists in grouping similar load profiles using domain expert-based procedures, statistical methods or data mining algorithms. For each group of similar loads profiles, a representative load pattern can be extracted. The shape of a load profile is usually representative of an operational pattern of a building and can be used as reference for estimating its expected behaviour over time. Therefore, there are two main expected goals behind load profiling analysis in buildings that can be summarized as follows:

- Identification of typical load patterns e.g., in form of reference load profiles,
- Detection of anomalous load patterns when typical ones were violated.

Depending on whether a single building or a group of buildings are analysed, different implications arise from the process of load profiling. In the first case a detailed diagnostic analysis of energy time series is performed to discover typical energy patterns characterising the operation of building or energy systems and then identify anomalous ones accordingly. In the second case, instead, the objective becomes a load classification to discover typical classes of buildings according to shape similarity [82,83].

However, conducting load profiling analysis in complex systems such as buildings is not an easy task. Indeed, different climatic conditions (external and internal), occupancy patterns, building thermo-physical features and heating/cooling systems operation modes generate the existence of various load patterns (in terms of shape and magnitude), not always easily inferable through domain expert based procedures and statistical methods.

In this perspective, more and more analysts rely on the application of unsupervised pattern recognition techniques such as cluster analysis [84]. Unlike the expert segmentation, cluster analysis allows load patterns to be identified in a not pre-determined time domain. In this way, robust and consistent reference profiles can be discovered.

2.2.2.1 Load profiling at whole building level

The robust characterisation of operational patterns and trends of energy consumption over time (i.e., energy profiling) is a central issue in building energy management, making it possible to better:

- handle the energy demand during the peak times for cost operational saving purposes [85,86],
- define the size of renewable energy system to reduce the operational costs [87],

- define benchmarks that take into account the trend of energy consumption over time [15,88], and
- detect anomalous patterns and profiles [30,89].

Capozzoli et al. [11] introduced a general framework for the mining of typical daily load profiles at both a single and a multiple building level, and discussed various applications exploiting load profiling frameworks as preliminary analysis for supporting the definition of energy management solutions. In fact, load profiling at whole building level is highly desirable on a liberalised energy market for enhancing load forecasting [90–92], implementing targeted demand-side management solutions [93], promoting modifications of the building energy demand and implementing demand response initiatives [11,94]. These implications are potentially crucial for several stakeholders (energy managers, energy service companies, energy network operators and policy makers) for both ordinary energy management and strategic planning activities.

In [95] a motif extraction based methodology was proposed to enhance the operation of a set of chillers serving a data center. Moreover, the sustainability impact was evaluated by means of useful metrics. In [96] a pattern recognition analysis based on a clustering algorithm (k-Shape) was performed in order to discover building energy consumption patterns. These patterns were further utilized to improve the accuracy and robustness of a forecasting model of energy consumption for ten institutional buildings in Singapore based on Support Vector Machines (SVM) algorithm. Also in [97] typical load profiles identification were used as a preliminary step in the development of a forecasting model for the electrical power demand of a supply fan of an Air Handling Unit (AHU). In detail, using a Fuzzy C-Means clustering algorithm, three subsets of homogeneous daily profiles (typical patterns) of the supply fan modulation were discovered while the atypical profiles were removed from the dataset. Then, for each subset a forecasting model combining Autoregressive Neural Network (ANN) and a physical model was built. The development of innovative robust methodologies to automatically detect anomalous energy consumption [60] (profiles with shape/magnitude significantly different from the typical operation patterns) makes it possible to operate a continuous commissioning of the building, also defining rule-based strategies [98] to be implemented in the building energy management system. For building diagnosis purpose also the robust extraction of daily patterns of occupancy data or indoor environmental quality parameters can be extremely useful when it is correlated with energy usage patterns. For example, the occupancy profile can be associated to operation of air conditioning or lighting systems. On the other hand, the building energy usage patterns can be analysed in relation to different components or sub-systems whose mutual interactions and correlations can be discovered by analysing their behaviour over time with a temporal approach in the knowledge discovery. Temporal data mining can support the optimal operation of a building at multiple levels through the extraction of useful cross-sectional relationships between forcing variables and the actual energy consumption by performing a multivariate time series analysis.

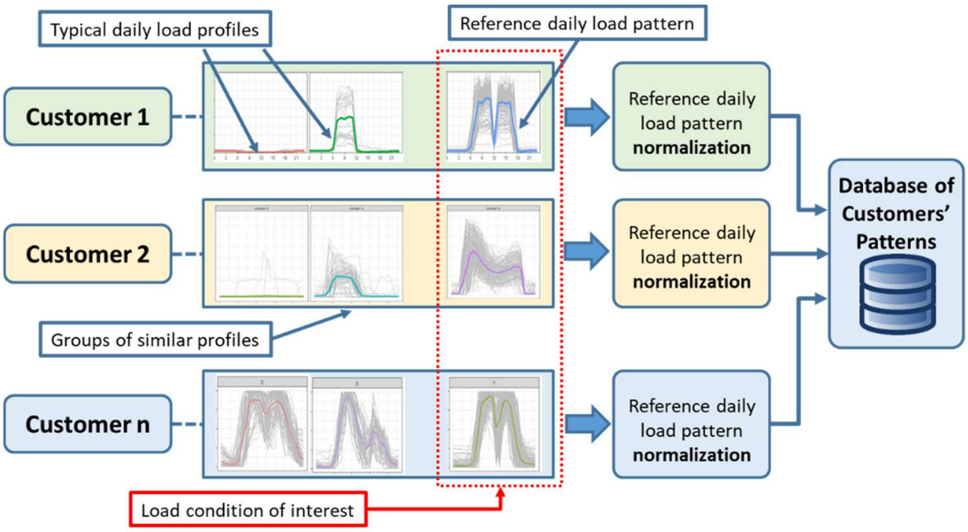
In some cases, can also be beneficial to transform and reduce the daily load profiles to increase the computational cost of analysis [28] and improving the identification of frequent and infrequent patterns [21,89]. To this purpose SAX representation of time series can be employed for transforming the original load profiles information into strings of symbols. In [30] SAX and motif discovery were employed in combination with Temporal Association Rule Mining (TARM) to mine temporal correlations in Building Automation Systems (BAS) data. The extracted knowledge joined with domain expertise was helpful in identifying typical patterns and anomalies, estimating energy performance and detecting opportunities for energy conservation measures. In [99], the time series related to the operational cycles of a solar cooling system was transformed into a symbolic representation and clustered to detect bad, average or good chiller performances. A similar work was conducted in [100], where the SAX process was used to reduce the computational efforts when pattern discovery algorithms were run. SAX was also adopted in [30] to discover frequent patterns in time series related to the energy consumption of the International Commerce Centre in Hong Kong and to efficiently estimate the similarity between each pair of symbolic sub-strings through the Random Projection algorithm. Infrequent pattern recognition was conducted in [89] by means of daily load profiles transformed into SAX words. Infrequent operating patterns of the cooling energy consumption of an international school campus and the overall electricity consumption of an office building were evaluated by setting a threshold of occurrence below which the SAX words representing daily load profiles could be considered as infrequent.

As discussed above the robust recognition of energy patterns from load profiles of building energy consumption is particularly desirable to perform robust energy characterization and diagnosis. In section 3.3 a robust methodological procedure conceived for this purpose is developed on real case study.

2.2.2.2 Load profiling at building portfolio level

A number of load profiling frameworks have been developed in the literature to deal with data coming from multiple buildings usually with the aim to identify, through unsupervised analysis, homogenous groups of typical daily load profiles (i.e., customer classification) characterised by similar shapes and/or magnitude [11,101]. When a group of buildings is analysed a classification process is usually performed. To this purpose a reference daily load pattern needs to be selected for each building (Fig. 16). In fact, the classification process could involve a large number of buildings making it a labour intensive and time-consuming task. For this reason, in most of cases, it is necessary to extract only one representative load pattern from the set of typical daily load profiles of each building. On the basis of the data segmentation, the representative load pattern usually corresponds to the typical profile in a specific time period or to the most populated cluster or to the most occurring motif. After the selection of the reference load pattern for each customer/building, data scaling is necessary in order to compare the different profiles between each other removing the effect of magnitude (Fig. 16). Magnitude differences, resulting from different building design features (e.g.

gross volume, floor area, installed power, etc.) or load conditions, can negatively affect the performance of pattern recognition algorithms in discovering similar shapes among daily load profiles.



Typical daily load profiles
Reference daily load pattern
Customer 1
Customer 2
Customer n
Groups of similar profiles
Load condition of interest
Reference daily load pattern normalization
Reference daily load pattern normalization
Reference daily load pattern normalization
Database of Customers' Patterns

Fig. 16 - Generation of the database of Customers' patterns (i.e., normalized load profiles)

Scaling can be achieved through different approaches. Load profiles in the (0,1) range are obtained normalizing respect to a reference power e.g., the maximum value [102], mean value [87] or between minimum and maximum [84] values of the original daily load profiles. In other cases, a z-score normalisation can be also performed. Consequently, the representative normalised load patterns are stored in a database (Fig. 16) and then grouped through unsupervised pattern recognition algorithms in order to discover typical classes of customers/buildings.

The whole process consists of three different steps: (i) identification of n classes of buildings according to load profile similarity, (ii) definition of the normalised reference load pattern for each customers' class (e.g. centroid) (iii) enrichment of the database with additional attributes (categorical or numerical) for each load profile to perform a supervised classification process.

The first step of the process, in most of the cases, makes use of unsupervised data analytics techniques to identify homogenous groups of customers based on their electrical/thermal daily load profiles [103]. To address that task several algorithms were proposed in the literature and tested on different case studies (e.g., from low voltage to high voltage electric customers).

According to Panapakidis et al. [104] the methods used for the identification of homogenous load profile groups can be categorized as partitional clustering algorithms (e.g., k-means), fuzzy clustering algorithms (e.g., Fuzzy C-means), hierarchical clustering algorithms, neural network based clustering (e.g., self-organizing maps) and algorithms that not belong to the previous categories (e.g., support vector clustering). The k-means algorithm was used with success for the classification of industrial [84] or domestic [105] electricity customers. Fernandes et al. used the Fuzzy C-means for the segmentation of residential gas consumers [106]. In [46] a customer classification process was performed by using a hierarchical clustering process, while Figueiredo et al. characterized the energy consumers by means of a self-organizing maps [107]. Moreover in [108] a support vector clustering process was adopted to segment electrical load patterns.

Despite their proven effectiveness, the robustness of such unsupervised methods is strictly dependent from various factors such as the aggregation algorithm (e.g., complete, single linkage in hierarchical clustering) [103], the dissimilarity distance measure between profiles [88,96], the data normalization technique [11] and number of clusters (i.e., customer groups). Due to such degrees of freedom in the clustering problem formulation, several adequacy indices (based on the measure of inter-cluster similarity and intra-cluster dissimilarity) have been proposed in the literature in order to assess the quality of clustering results [103].

In [103,104] the most popular indices were reviewed such as mean index adequacy (MIA), clustering dispersion indicator (CDI), scatter index (SI), Silhouette index,Variance Ratio Criterion (VRC) and Davies-Bouldin Index (DBI). The use of adequacy indices makes it possible to partially supervise the process suggesting the most suitable number of customer groups to be assumed in the clustering analysis.

The outcome of that step is then the identification of a number of energy customer classes (buildings with similar energy consumption profiles), for which the reference load pattern can be calculated as the centroid or medoid of the profiles grouped together. Subsequently, the customer class label is encoded as a categorical variable to be predicted through a supervised classification model. To this purpose the load profile database is enriched with additional attributes (categorical and/or numerical) to be considered in the classification as predictive variables.

These attributes can be defined a-priori or based on in-field measurement campaign [45]. A-priori indicators are related to the customers' energy contracts and type of commercial activity and then are generally used by energy providers to preliminary characterize their clients. These indicators are static and do not exhibit sensitivity to load profile shape and magnitude [45]. Indeed, if they are used as unique predictors they cannot provide a good characterization of the energy use of customers in the time domain [11]. For this reason, indicators extracted from in-field measurement campaign are employed in order to ensure a higher accuracy of the supervised classification model. These indicators deal with specific features of the load profile shapes and are calculated for each customers' reference load pattern.

These indicators (in the (0,1) range) are capable to capture the normalized variability in daily load profiles, and hourly/sub-hourly load shares with respect to specific reference values (mean, max, min, standard deviation) in different daily periods (e.g. night, lunch time) [45,109].

Once the predictive attributes are selected, the customer classification process goes through the development of a supervised classification model. The classification task aims at assigning unknown customers into pre-identified classes. Decision trees (e.g. C4.5, C5.0, CART) have been often used in the literature to accomplish that task due to their capability in handling both categorical and numerical variables and the high readability of their output in form of decision rules [110,111]. In [112] Ramos et al. used C5.0 algorithm for classifying a portfolio of about 1000 medium voltage customers in groups identified though a clustering analysis. Also in [107] Figueiredo et al. employed the C5.0 algorithm for customer classification purpose. In particular, a different consumer characterization is obtained for each load conditions considered. As a reference for winter working days and weekends the overall classification accuracy is close to 80% leveraging on a set of about 30 decision rules.

Fig. 17 shows the main conceptual steps of the described customer classification process. In particular, once the customer class are identified targeted energy management and demand response strategies can be conceived for each class. At this stage, when new customers are included in the portfolio it is possible to sorting them in the customer classes previously identified.

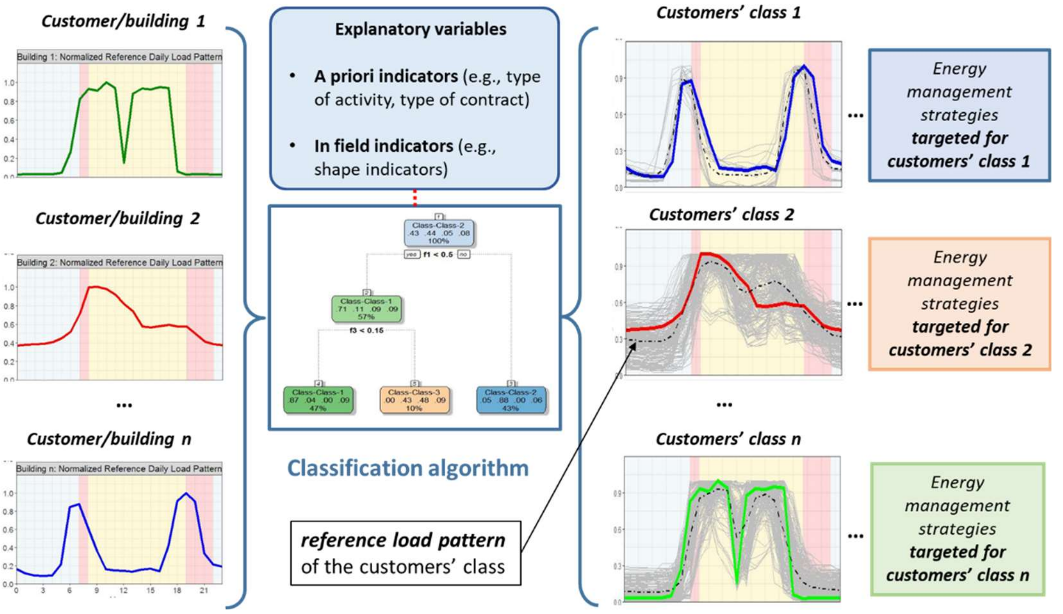

Customer/building 1
Building 1: Normalized Reference Daily Load Pattern
Customer/building 2
Building 2: Normalized Reference Daily Load Pattern
...
Customer/building n
Building n: Normalized Reference Daily Load Pattern
0 3 6 9 12 15 18 21
Explanatory variables
• A priori indicators (e.g., type of activity, type of contract)
• In field indicators (e.g., shape indicators)
Class-Class-2 .43 .44 .05 .08 100%
yes f1 < 0.5 no
Class-Class-1 .71 .11 .09 .09 57%
f3 < 0.15
Class-Class-1 .87 .04 .00 .09 47%
Class-Class-3 .00 .43 .48 .09 10%
Class-Class-2 .05 .88 .00 .06 43%
Classification algorithm
reference load pattern of the customers' class
Customers' class 1
Customers' class 2
...
Customers' class n
Energy management strategies targeted for customers' class 1
Energy management strategies targeted for customers' class 2
Energy management strategies targeted for customers' class n

Fig. 17 - Conceptual framework of the customers' classification process

Rhodes et al. in [83] stated that load profiling of residential customers could serve as a starting point for utilities looking to reduce electricity use during peak times by developing policies that target load shifting. Eventually, in the two-way paradigm of smart grid, load profiling at building portfolio level is particularly beneficial for both energy providers and users that are involved in Demand Response (DR) programs [94,113]. In the current competitive energy retail market, DR programs are designed to be attractive for the consumers and at the same time profitable for the retailers. In incentive-based programs, knowledge of customers' macro-behaviour in energy consumption allows the distribution companies to better manage the grid operation [114] and the interactions between energy consumption and production [94,115] (e.g., indirectly switching certain electric appliances at certain times).

The consequent modification of a load profile allows to flat the demand profile or in some cases to follow the generation pattern for achieving an improved grid stability [116]. For example, virtual thermal storage, through the modification of load profiles of a group of buildings served by a district heating network represents an effective way to increase the share of heat from cogeneration and renewable sources [117].

Load profiling also makes possible to identify energy customers that exhibit more variable load patterns than their peers considering the same load conditions (i.e., season, day type). Classifying these customers is essential as they could be able to change their loads more effectively when involved in demand response programs [118]. In that perspective energy retailers can take advantage from that knowledge in the design stage of dynamic pricing plans. According to the different customer groups in the building portfolio, different energy tariffs can be set for each typical curve in order to maximize the relative profit [119,120]. For instance, in [45] the authors demonstrated how a data-driven customer classification process could be used to modify existing energy tariffs by fixing rate coefficients for each customer class.

Also the customer side is experiencing a revolution in the smart grid environment in terms of demand management opportunities. In fact, thanks to the spread of electrical/thermal energy storages, renewable energy systems and data analytics technologies in buildings [1], user's energy demand is becoming more and more flexible [115,116]. Energy managers can implement, in an easier way, strategies aimed at modifying building energy use to obtain targeted changes in electrical/thermal load profile [116]. In this way, customers can change their load profiles (e.g., consuming less energy during peak hours or shifting the energy use to off-peak hours) in response of variations of energy price over time [121] (i.e., price-based programs) leveraging on energy flexibility and fully exploiting building potential in the energy management [122]. Benchmarking the energy usage in the time domain, through load profiling, is then crucial also for the impact assessment of DMSs and DR initiatives [123,124].

The information about shape and magnitude of electrical power consumption patterns can reveal useful knowledge [125] about building energy flexibility potential and/or in some cases the presence of multiple typical patterns (e.g., seasonality, intra-week variation)[83]. From the design point of view, the in-depth characterization of the energy demand makes it possible to better address the current transition from large centralized generation plants to multi-energy distributed ones that are capable to provide, from different sources, energy at a small scale (e.g., neighbourhood) when it is needed [116]. In fact, the lack of knowledge about building energy use patterns currently represents the main barrier for fully exploiting the benefits of energy management also at micro grid level.

Section 3.4 presents a robust methodological procedure of load profiling for conducting customer classification. The tool is developed at building portfolio level on energy consumption data of real buildings.

2.2.3 Fault detection and diagnosis

Recent years have seen an increasing interest of the scientific community in exploring solutions to improve energy efficiency in buildings by implementing advanced data-analytics based energy management strategies [126]. According to [59], around 20% of energy consumption in buildings is attributable to incorrect system configurations and inappropriate operating procedures that can be effectively detected through automatic analytics processes. For example, in commercial buildings, inefficient system plants waste an estimated 15% to 30% of energy used [127,128]. Due to lack of proper maintenance, failure of components or incorrect installation, building systems are frequently run in faulty conditions where a fault is intended as an unpermitted deviation of at least one characteristic property of the system from the acceptable, usual, standard condition. The objective behind Fault detection and diagnosis (FDD) is twofold. On one hand fault detection consists in the recognition of a fault occurrence, and on the other hand fault diagnosis corresponds to the identification of the causes and the location of the fault [129]. In particular, advanced methods of fault detection are based on mathematical models and on methods of system and process modelling to generate fault symptoms (e.g. residuals). Fault diagnosis methods use causal fault-symptom-relationships by applying methods from statistical decision, artificial intelligence and soft computing.

Although currently underutilized, FDD is a powerful tool for ensuring high efficiency in building operation and FDD products represent a very fast-growing market in the context of building analytics technologies [130]. According to [12] over 30 FDD products are available in U.S. that may be delivered through different implementation models [130]. Despite the existing differences in the way tools are implemented and integrated with the monitoring system, the main tool classification can be performed according to the approach employed for conducting the FDD analysis.

In the study presented in [126] the methods used for performing an FDD analysis can be classified in quantitative model-based, qualitative model-based and process history-based.

The quantitative model-based approach includes all the methods involving engineering models with different levels of detail in the physical description of the system (e.g. white box models). The qualitative model-based methods exploit the system knowledge derived from domain expertise (e.g. rule-based, qualitative models). The last category includes data-driven methodologies exploiting collected operational data of the system under investigation (e.g. Artificial Neural Networks, Association Rules Mining, grey box models). While rule-based methodologies (qualitative approach) are still the norm, vendors are beginning to use data driven methodologies for addressing FDD tasks [130].

In the last few years, the data-driven approach gained more and more interest, thanks to its applicability even in the case engineering models of the building and systems are inadequate or difficult to be developed, or the physics-based knowledge is not wide enough [126]. In this context, particularly promising is the implementation of data analytics techniques which include both supervised and unsupervised algorithms. As reviewed in [13], the main advantages of the data-driven approach based on data analytics, in comparison to traditional approaches rely on the opportunity to

- Learn patterns from system operational data automatically without the use of physical models. The data-driven approach based on data analytics does not require an a-priori understanding of the relationships that exist among faults and their symptoms. Therefore, it would be simple to implement the data driven-based methods.
- Achieve higher fault detection and fault diagnosis accuracy than the knowledge driven-based (qualitative) methods also for faults of low severity levels.
- Perform FDD analysis exploiting a limited number of variables. It means that can enable an optimisation of sensor installation and then significantly reduce the number of required sensors.

Considering the building application field, the most developed data-driven based FDD processes focused on the operation data of Heating Ventilation and Air Conditioning systems, considering that in commercial buildings they could account for 50% of the energy demand [131]. Such systems are essential in buildings for maintaining the desired microclimatic indoor conditions and often a large amount of operation variables is collected through BAS for controlling them. However, the presence of BAS does not ensure that HVAC systems were operated in absence of component or control faults. It has been estimated that the identification and diagnosis of these faults in HVAC can lead to potential savings of about 30% [132].

One of the most sources of component and control faults in HVAC is related to Air Handling Units management. A study conducted on more than 55.000 Air Handling Units of HVAC systems, showed that 90% of them runs with one or multiple faults [133] making them a matter of interest in many FDD applications. In this perspective, the next section provides a wide overview about the use of data analytics for conducting data-driven FDD analysis on real operational data of AHU systems.

2.2.3.1 Fault detection and diagnosis in AHUs

As stated in section 2.1.3, data analytics techniques can be categorized in supervised and unsupervised approach. Both approaches were employed in the literature for conducting FDD analysis in AHUs [134,135].

Even though each component of an AHU can be potentially corrupted by a fault, the most common faults can affect sensors (e.g. offset in the measurement), controlled devices (e.g. blockage or leakage of air damper or coil valves), equipment (e.g. coil fouling or reduced capacity, duct leakage, fan complete failure or deviation in the pressure drop or belt slippage) and controllers (e.g. unstable or frozen control signal for dampers, coils or fan) [136]. Dehestani et al.

proposed a methodology to identify faults related to the fans and the air dampers of an AHU. The methodology used a Multi-Class Support Vector Machine (MC-SVM), for the identification of both pre-labelled faults and new ones [137]. In [138] and [55] a Bayesian Network (BN) was adopted for the diagnosis of faults related to air dampers, cooling coil valve stuck and return fan failure. The BN exploited as input the residuals obtained from a set of limit checking rules and statistical models capable of estimating air temperature, water flow rate, air flow rate and fan power consumption. Mulumba et al. in [139] proposed a methodology to diagnose the presence of several faults affecting air dampers, cooling coil valve and return fan by using a SVM in combination with an autoregressive model with exogenous inputs. Yan et al. in [140] proposed a combination of two supervised techniques to diagnose the blockage of air dampers and coil valve, the duct leakage and the return fan failure. In [140] was developed a classification tree which used as inputs both monitored data (i.e. air temperature and flow rate, fan speed and power, and cooling coil valve position) and residuals obtained from a regression model of the fan speed, while as output the labels of different faults. The methodology developed made it possible to accurately perform fault diagnosis, but without taking into account transient periods of operation. Different classification models for fault detection were also compared in [141] and CART algorithm was identified as the best choice for the detection of steam or chilled water leakage.

The unsupervised methods proved to be particularly flexible for their nature in exploring data set without any a priori constraint, contrary to the supervised models [7].

In [142] the authors proposed an unsupervised methodology to identify energy wastes and faults of a fan in an AHU by exploiting Association Rules Mining (ARM). This type of algorithm requires a strong expertise by the analyst for the interpretation of the results considering that the rule set extracted could include also not interesting information for the identification of anomalous operation of the air conditioning system [142]. Many studies made use of ARM for the identification of faults in different types of HVAC sub-systems including district heating substation, AHU and chillers [13]. In order to help the domain expert in the interpretation of ARM results, in [143] was proposed a methodology to reduce the number of rules to be analysed and to effectively group them for distinguishing the faulty from the normal operation. Furthermore, the temporal relation among the energy consumption of different HVAC components was studied in [30,144] to determine the presence of faults and prevent a reduction of energy performance over time.

A combination of a supervised with unsupervised methods (e.g., decision tree and clustering analysis) for the detection of anomalous energy consumption in a group of smart office buildings was proposed in [60,145]. As a reference, Dey et al. achieved high values of accuracy in the automatic FDD on fan coil units operation by combining MC-SVM and cluster analysis [146].

In [147] Du et al. proposed a methodology to identify faults of temperature, flow rate and pressure sensors in a VAV system by implementing Artificial

Neural Networks (ANNs) in combination with a signal decomposition technique (i.e. Wavelet analysis). In [148], ANN was combined with clustering analysis to diagnose faults related to cooling coil valve, air damper and temperature sensors in an AHU. In the first step, ANN was used for the estimation of supply air and water temperature to perform a residual analysis, then the methodology leveraged on clustering analysis for the fault diagnosis stage. Guo et al. used a Hidden Markov Model (HMM) for the fault detection phase and a cluster analysis for the identification of various types of faults such as the blockage of dampers, frozen fan or unstable cooling coil valve control signal [149]. In [150,151] an unsupervised data-driven approach was used to identify the presence of cooling coil valve blockage, heating coil valve leakage and air damper blockage, by analysing the error generated from the reduction of variables by means of Wavelet Transform and Principal Component Analysis. Successively, fault diagnosis was performed by analysing the trend of each variable during faulty conditions in order to identify the variable mostly influenced by the fault source.

In [152] the authors proposed a methodology to diagnose the stuck of the recirculation damper and of the cooling coil valve and the decreasing of the supply fan speed in an AHU. In particular an SVM was used in combination with a white box model, exploiting the residuals obtained comparing actual and simulated fault-free values of supply and mixed air temperature, and cooling coil outlet water temperature. Wu et al. [59] combined a quantitative model-based method with an unsupervised data-driven method to diagnose sensor faults, air damper blockage or frozen fan. Specifically, the variables considered were firstly reduced (i.e., by means of Principal Component Analysis); successively the presence of faults was investigated comparing actual monitored data with the estimation of airflow rate and energy calculated by using energy and pressure-flow simplified balance equations. In other works, qualitative-based approach was used to perform automatic FDD in combination with the data-driven approach. In [153], the detection of faults occurring in an AHU was performed by exploiting "IF-THEN" expert rules related to the residuals of mixed air temperature, return air flow rate, supply air static pressure and cooling coil valve control signal, generated with different General Regression Neural Networks. In [154] was proposed the integration of expert rules with Bayesian Networks in order to better isolate faults in AHU. Such approach made it possible to exploit the violation of expert rules to better detect the co-occurrence of multiple faults at the same time.

The proposed literature review demonstrated how much active is the FDD research field and the high contribution that data analytics methodologies bring.

In this perspective, Chapter 4 presents and discusses a novel FDD tool, based on data analytics techniques, developed on measured AHU operation data.

2.2.4 Benchmarking analysis

The main goal of a benchmarking system is to evaluate the divergence between the energy performance of a building/system and a reference baseline.

Benchmarking methods can be categorized according to the considered type of baseline. Four types of baselines can be considered in existing benchmarking methods: previous performance of similar buildings (i.e., external benchmarking), current performance of similar buildings (i.e., external benchmarking), previous performance of the same building (i.e., internal benchmarking), and intended performance of the same building (i.e., internal benchmarking) [63]. The first two types of baselines are used by regulators, public authorities, or private building portfolio managers to encourage owners to improve energy efficiencies of their buildings [155]. On the other hand, internal benchmarking techniques are exploited et single building level for energy tracking and continuous commissioning purpose.

According to the modelling approach considered benchmarking analysis can be further classified in calculation-based and data-driven benchmarking system [156]. The calculation-based benchmarking system compares the observed energy consumption with a simulated benchmark, representing an archetype or a theoretical energy consumption [157]. Calibrated simulation tools, belonging to the so-called white box methods, are by now the main instrument to assess the energy performance of buildings and to evaluate the possible scenarios for energy retrofit [158–162]; they also provide the most reliable results in the design stage of a building [163]. This approach is however of limited use for large building stocks because it is time-consuming, labour intensive [164], and it requires detailed building information which is not always easily available within a large dataset [165]. The data-drive benchmarking process compares the observed energy consumption with a benchmark value obtained from actual energy consumption data. The most common data-driven benchmarking processes proposed in the literature are performed through statistical models [166], data analytics techniques [9,10,52,167,168] and simple normalization of the energy consumption with respect to floor area and/or volume as a way to computing the mean or median value [156].

With the rapid growth of stored data in the building sector and the necessity to extract knowledge from these large datasets to improve the building performance, the data-driven benchmarking analysis is becoming the most promising approach. The choice of the most suitable strategy (simple normalization, statistical models and data analytics techniques) to develop a benchmarking process depends mainly on the quantity and the quality of the available information and on properties of the available dataset.

When the pieces of information available are exclusively related to building energy consumption (e.g. total energy consumption, space heating, space cooling, lighting, etc.), a simple normalization is the most common way to obtain the benchmark value. To this purpose, buildings are firstly segmented and classified according to their building end-use category as residential, industrial, commercial and then Key Performance Indicators are calculated.

This approach, relying only on one the calculation of simple KPIs, was used for example in [5,169]. In [5] the average energy consumption for different building types in the US was computed, in order to make available comprehensive building energy information useful to plan efficient energy policies for the future. In [169], the annual total electricity use intensity is used to identify the benchmark value of the California State Teachers' Retirement System Headquarters (CalSTRS). To quantify the potential energy savings, this value was thus compared with a median value of annual total electricity use intensity obtained from 31 buildings selected from the database of California's Commercial End Use Survey.

Other simple normalization methods evaluate different performance indicators simultaneously. For instance, in [170] *Technique for Order Preference by Similarity to Ideal Solution* (TOPSIS) was used to introduce a multi-criteria benchmarking approach. This technique made it possible to compare building energy performances considering multiple indicators for a more comprehensive evaluation. Lee at al. [171] demonstrated that this multi-criteria approach is more consistent than using a single performance indicator.

However, in many cases, buildings belonging to the same end-use category can exhibit significantly different patterns in their energy consumption [9,121]. In such cases, benchmarking methods related to the energy use intensity (e.g., kWh/m^{2}y) of the building are not able to fully characterise the energy behaviour of a customer over time. On the contrary, knowledge extracted from energy consumption time series (i.e., load profiling analysis) contains information on how and when building energy use changes during the day for various end uses such as appliances, lighting, ventilation, heating and cooling [15,16].

For this reason, data analytics techniques, such as cluster analysis, were proposed in the literature to find homogeneous groups of buildings having the same energy consumption pattern (i.e., energy profiling) [9–11,167], or similar energy features [168]. As a reference, when a clustering analysis is conducted for segmenting buildings in heterogeneous building portfolios, the benchmarking process analysis can be conducted using the centroids of the clusters as reference patterns/values. Also supervised data analytics techniques were often used for benchmarking energy use in buildings taking into account various significant influencing factors(e.g. weather conditions, building envelope, building operational modes, occupant behaviour etc.) [172]. These benchmarking processes, based on the development of prediction models, represent fast and accurate management tools. The main used models are Artificial Neural Networks (ANNs) [65,70,71,173], Support Vector Machine (SVM) [73,174], Gaussian Process Regression (GPR) [175,176], Multiple Linear Regressions (MLR) [56,177–180].

Sharp [181] developed a stepwise linear regression model in order to evaluate the main factors influencing EUI (Energy Use Intensity) for office buildings. The main variables considered were building size, number of workers, number of computers, occupancy, operating schedule, presence of an economizer and presence of a chiller. Regression residuals (i.e. the difference between the monitored and the estimated energy consumption) were used as measures of energy inefficiency. Also in [179,180] the residuals of the regression model were used as measure of the building energy efficiency. Similarly, the Energy Star

[182] uses the Energy Efficiency Ratio (defined as the ratio between the actual and the estimated energy consumption) to perform the building energy benchmarking.

Wong et al. [183] used an ANN model in the evaluation of energy performance of office buildings located in Hong Kong. The results showed that the ANN model achieved more accuracy in the prediction of electricity use for periods in which the energy end-use was clear (i.e. summer cooling and winter heating).

In [52], a novel methodology was proposed to perform a benchmarking analysis particularly suitable for heterogeneous samples of buildings. The methodology exploited Linear Mixed Effects Model to take into account both fixed effects shared by all individuals within a dataset and the random effects related to specific groups/classes of individuals in the population. The groups of individuals within the population were classified through a decision tree. The benchmarking analysis was tested for a case study of 100 out-patient Healthcare Centres in Northern Italy, finally resulting in 12 different frequency distributions for space and Domestic Hot Water heating energy consumption, one for each class of homogeneous class of buildings. From the median value of each frequency distribution, reference values were extracted to be used in a benchmarking analysis.

Benchmarking analysis is one of the main functionalities in DSS and in this dissertation is explored at building portfolio level on real energy consumption data of more than 100 buildings. Specifically, section 4.3 presents a robust methodological procedure with twofold objective. On one hand the aim is the identification of typical load profiles in a building portfolio (i.e., energy use benchmarks). On the other hand, a second objective is to develop a tool for the classification of new buildings included in the portfolio (i.e., customer classification). These two objectives are achieved by developing a unique multifunctional DSS tool.

2.2.5 Characterisation of occupant behaviour

Occupant behaviour is one of the major factors influencing building energy consumption and introducing sources of uncertainty in building energy use prediction and simulation [184–186]. Currently the exploitation and characterization of occupant-related data in buildings is insufficient thus limiting opportunities of building design optimizations and energy management improvements [184]. Occupant behaviour is associated with various actions that have a direct or indirect impact upon building energy consumption such as adjustment of thermostat settings, opening and closing of windows, dimming and switching of lights, use of blinds, turning on/off of HVAC systems, presence and movement in building spaces[184].

Occupant actions in building can be categorized in (i) adaptive actions, and (ii) non-adaptive actions [186–188].

On one hand, adaptive actions are intended as reactions of occupants (a) to adapt the indoor environment to their needs or preferences or (ii) adapt themselves to the environment e.g. by clothing adjustments and movement in building spaces [186]. On the other hand, non-adaptive actions are related to occupant presence and operation of plug-ins and electrical appliances [186]. Both adaptive and non-adaptive actions are influenced by typological factors called "drivers" that are related to external and individual parameters. According to [189] drivers of occupant actions in buildings can be classified as follow:

- Physical drivers such as internal and external environment, such as temperature, wind speed, humidity;
- Contextual drivers, related to the building features, such as floor area, insulation, type of heating system;
- Psychological drivers related to the ways people reacts to satisfy their needs;
- Physiological drivers such as age, activity level and health of occupants;
- Social driver related to interactions between occupants.

Quantifying the effect of occupant behaviour on building energy consumption and the potential energy saving achievable through its modification remain primary challenges. According to the literature a potential reduction of energy consumption in the range of 10-20% and 5-30% can be achieved for residential and commercial buildings respectively [184–186].

Ouyang and Hokao in [190] investigated the energy saving potential in 124 households in China by improving behaviour of occupants. Specifically, the stock of households was segmented into two groups. The occupants included in the first group, received suggestions on how improve their behaviour for reducing energy consumption, while maintaining unchanged occupant habits in the second group. By comparing monthly household electricity consumption of both groups, it was found that energy-conscious behaviour lead to an energy saving higher than 10%.

In [191] the authors simulated the effect of occupant behaviour by means of the energy simulation software ENERWIN. The first step of the analysis was aimed at collecting (by means of surveys) data and information on typical occupancy patterns and operation schedules of electrical devices in 30 residences in Kuwait. In the second step of analysis, the patterns obtained in the previous step were used as input data of the simulation software ENERWIN replacing its default data. Results demonstrated that the annual electricity consumption would rise by 21% when the realistic occupant behaviour patterns were used instead of the default settings provided by the software.

The research studies presented, demonstrated that occupant behaviour significantly influences energy management in buildings making its fully characterization highly desirable for improving energy performance. However, occupant behaviour is a complex phenomenon to be effectively characterized due to its stochastic nature and dependency from social factors. To this purpose, with a certain level of approximation, it can be simplified and represented quantitatively by understanding the relation that exist between drivers and behaviour [184]. In this perspective the growing availability of data in buildings, also related to occupant-related information, makes the use of data analytics techniques particularly suitable to automatically extract robust behavioural patterns from massive datasets [1,58,189,192].

In [193] a methodology for characterizing and improving occupant behaviour in residential buildings was proposed. End-use loads of various electrical devices were categorized into two levels (i.e., main and sub-category) and were analysed to indirectly infer the corresponding two-level occupant activities (i.e., general and specific behaviours). In particular, data clustering and classification were performed for analysing the main-use loads with the aim to identify general energy-inefficient behaviours of occupants. In a second step of analysis, association rules were extracted to also characterize energy-inefficient specific occupant behaviour. The methodology was implemented in a group of residential buildings and demonstrated the effectiveness of data analytics techniques for the identification of behaviours to be modified for achieving energy saving.

In [189] was proposed a data analytics based methodology conceived for the extraction and modeling of window opening and closing patterns in an office building naturally ventilated. A multi-step framework of analysis was developed. Firstly, a logistic regression model was developed to identify the most relevant factors influencing opening and closing of windows. Successively, a clustering analysis was performed to extract typological behavioural patterns considering motivational aspects, opening duration, opening frequency and window positions. Eventually, association rules were mined among the cluster patterns for the identification of two reference office user profiles. for which different natural ventilation strategies as well as robust building design recommendations that may be appropriate. Such advanced characterization of occupant behaviours provided a set of behavioural rules that can be used to support specific operation and maintenance and develop ventilation-based energy saving strategies.

In this context, especially for commercial and office buildings, great attention has been paid to the collection and processing of occupancy data in buildings, since different system operations can be optimized by characterizing and managing occupants' presence [194–196]. In fact also for occupant presence patterns a great difference from the default profiles suggested in building energy modelling software could exist [58].

Currently, most heating, cooling, ventilation and lighting systems are operated considering buildings as occupied with a fixed schedule that is assumed to not change over time. In the majority of cases, this assumption differs significantly from the actual occupants' presence. As a consequence, knowledge extractable from occupancy data can lead to considerable energy savings achievable by operating energy system (especially for centralized HVAC systems) with optimized occupancy-based schedules [197].

In the literature several researchers focused on the analysis of actual occupancy data for characterising and managing occupant presence improving the performance of HVAC systems in buildings. In [197], through the analysis of actual occupancy data, was introduced an optimised HVAC schedule by reducing occupancy diversity through the aggregation of occupants with similar patterns in the same part of the building. Fig. 18 shows the main conceptual steps behind occupancy diversity reduction in buildings for optimising HVAC system rescheduling. In particular, thanks to the adoption of unsupervised pattern recognition techniques it is possible to construct the database of typological occupancy profiles. In this way it is possible to reduce the occupancy diversity in each thermal zone by aggregating in the same portion of the building occupants that exhibit similar behaviour (i.e., occupancy profile). As a consequence, the schedule of each HVAC system can be optimised and diversified according to the actual occupied period of the served thermal zone generating a potential reduction of the daily operation hours of the system.

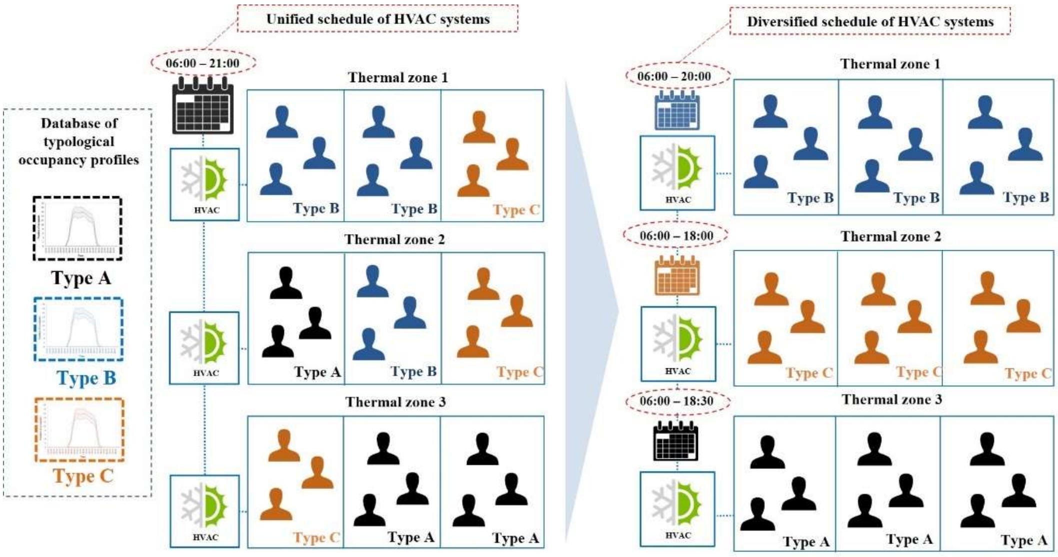


Fig. 18 - Main conceptual steps behind occupancy diversity reduction in buildings

Yang, Ghahramani and Becerik-Gerber in [198] investigated the concept of occupancy diversity reduction in buildings and provided a methodology to quantify its impact on HVAC energy consumption. Five reference buildings were analysed, considering different set point controls before and after eliminating occupancy diversity. The energy efficiency at both zone and building level were found to be significantly affected by occupancy diversity. A reduction of the energy consumption of the HVAC systems of about 18% was found when occupancy diversity was minimized at building level.

In [199] was presented an experimental study on two occupancy-based HVAC system control strategies implemented in a Building Automation System (BAS). The experimental analysis was performed in a single room in a university campus in Florida equipped with a VAV (Variable Air Volume) system. The occupancy data required for the control strategies were collected through an occupant's presence sensor. Both occupancy-based strategies were found to guarantee 40% of energy savings during the experiments, compared with a baseline control that did not use occupancy measurements.

Another study, aimed at calculating energy savings due to the adoption of an occupancy-based control strategy, was carried out in [200]. Operational zoning and intermittent HVAC system operation strategies were evaluated for common built mosques in Saudi Arabia, considering that the mosques were occupied for five periods of one hour per day (for prayers). Since it was difficult to accurately define the occupancy profiles, the percentage of occupants was derived as a function of the pray duration. Compared to a continuous operation mode, the intermittent mode led to 30% of savings as a result of an appropriate operation zoning.

In section 3.2 a robust methodological procedure is developed, exploiting actual occupancy data for obtaining scheduling improvements of HVAC systems in buildings.

2.3 Discussion of the literature review

The application of data analytics techniques represents a powerful opportunity to extract useful information from the building-related data to enhance energy efficiency of buildings during their operation. The applications discussed in the previous sections demonstrated the usefulness of data analytics technologies in several fields of building energy management especially for what concerns DSS solutions (i.e., not including a control signal as output).

The information obtained from DSS tools enables building owners to operate their buildings more efficiently avoiding energy waste over time. Differently from other kinds of EMIS systems, DSS read data from the monitoring system, analyse them but do not have a direct communication with the Building Automation System (BAS) for adjusting the control parameters of building energy systems during operation. In this perspective, while DSS are powerful tools, they need to be integrated in a robust verification process to achieve the desired impact.

As demonstrated in the literature review, advanced data analytics techniques are today capable to address emerging energy management tasks that represent essential functionalities of advanced DSS solutions; i.e., energy consumption prediction, identification of energy anomalies, identification and diagnosis of faults in energy systems, benchmarking and characterization of the occupant behavior. However the effective coupling of building physics and data science, at the basis of advanced DSS, still needs significant contributions aimed at developing robust and generalizable frameworks of analysis that on one hand extract useful knowledge from measured data and on the other hand support the final user in defining ready-to-implement energy saving and management strategies.

A plethora of both general purpose and tailored algorithms are available for each data analytics technique, and in most cases no algorithm is universally superior [1]. Several aspects determine which algorithm performs best, including data volume, data quality and the target of analysis. The selection of an optimal algorithm, as well as the tuning of its parameters, needs to be supervised by an experienced computer scientist, seeking a good trade-off between generalizability, robustness, interpretability, and accuracy. The whole process requires a considerable amount of expertise and effort. Thus, new scalable approaches that are highly interpretable for the user and capable to automatically extract actionable knowledge from massive energy-related data repositories will fuel the next generation of energy management and information systems [1].

This dissertation seeks to address each of these challenges through the development of four DSS tools useful at different levels of building energy management (from system component level up to building stock level) leveraging on both building physics and data science expertise.

3 DSS applications at meter-level: development of advanced Energy Information Systems (EIS) tools

This chapter discusses in detail the development of data analytics-based methodologies that can be integrated in DSS. The focus is on meter-level analyses, typically performed by means of advanced Energy Information System (EIS) tools.

Portions of the present chapter were already published in the following scientific papers:

- Capozzoli A., Piscitelli M.S., Brandi S. 2017. *Mining typical load profiles in buildings to support energy management in the smart city context*. Energy Procedia, 134 pp. 865–874. [11]
- Capozzoli A., Piscitelli M.S., Brandi S., Grassi D., Chicco G. 2018. *Automated load patterns learning and diagnosis for enhancing energy management in smart buildings*. Energy, 157 pp. 336–352. [21]
- 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 pp. 113727. [10]
- Capozzoli A., Piscitelli M.S., Gorrino A., Ballarini I., Corrado V. 2017. *Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings*. Sustainable Cities and Society, 35 pp. 191-208 [14].

3.1 Advanced Energy Information Systems (EIS)

Advanced EIS are enabling tools (i.e., DSS) that provide the needed analytical capability to building owners and energy managers as they are conceived for automatically extracting knowledge from building related data. The information gathered through EIS tools provides insight into building energy use and system performance enabling building owners to operate their buildings more efficiently [12].

Differently from other kinds of energy management DSS tools, EISs read data at meter-level, analyse them and provide informative outputs to a human user (e.g., energy manager, building owner, energy service company)[12,201].

Advanced EISs not only allow new forms of building energy management to be pursued but at the same time significantly reduce the complexity of performance commissioning in existing buildings. According to the Building Commissioning Association (BCA) Existing Building Commissioning (EBCx) is defined as a systematic process aimed at improving the performance of buildings and energy systems by means of low/no cost and capital-intensive measures and ensuring their effect persists over time [202].

Advanced EISs are today capable to enhance such process (i.e., building commissioning) and it is mainly due to the exploitation of data analytics methods. In particular, as emerged from section 1.3, time series analytics and automated rule extraction techniques play an essential role in the knowledge discovery phase for maximizing both the amount of extracted information and its interpretability.

In this context, the effective coupling of building physics and data science needs significant contributions aimed at developing robust and generalizable methodological frameworks for bridging the gap between the growing availability of measured data and the need of actionable knowledge. As stated in section 1.1, the effectiveness of DSS solutions (and then also of EIS) can be considered strictly related to three main factors: (i) the level of user engagement, (ii) the detail of the analysis and information provided, (iii) the level of interpretability of the results obtained. Such aspects should be then properly considered for achieving significant improvements in building energy management. In order to demonstrate such potential, three main opportunities related to the implementation of EISs are investigated in this dissertation for different testbeds at different scales (Fig. 19).

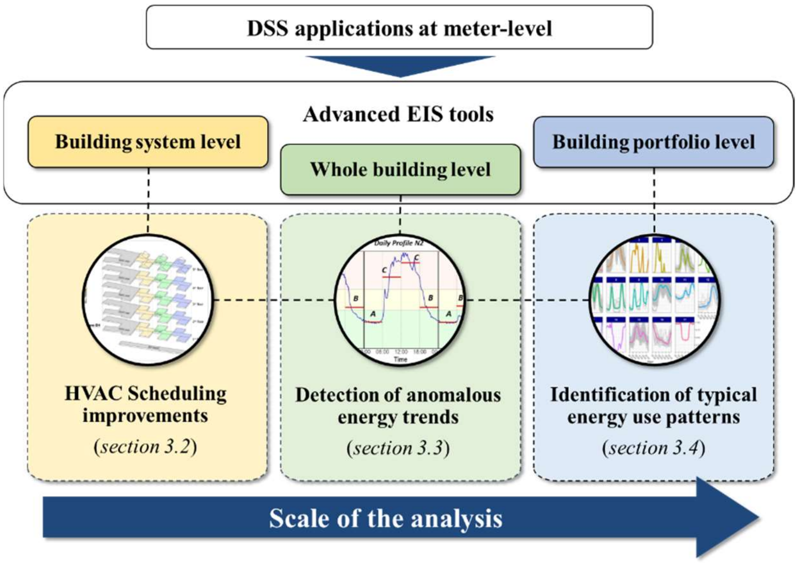


Fig. 19 - Advanced EIS tools developed at different scales and organization of Chapter 3

Novel frameworks of analysis are developed at meter-level for addressing the following main tasks typically required to advanced EISs (Fig. 19):

- HVAC scheduling improvements at building system level (discussed in in section 3.2).
- Identification of energy consumption reduction opportunities through the detection of anomalous energy trends at whole building level (discussed in section 3.3).
- Identification of typical energy use patterns and customer classification at portfolio level (discussed in section 3.4).

All the developed tools leverage on methodological procedures that exploit time series analytics and rule extraction techniques for ensuring high interpretability of the results.

The conceived EIS tools can be then easily translated in a set of decision rules and embedded in DSS helping managers, owners or service companies in increasing awareness about the energy performance of their buildings and achieve demanding energy management targets during daily operation.

3.2 Development of an EIS tool for scheduling improvements at building system level

Advanced EISs provide today valuable opportunities for extracting in a robust way useful hidden knowledge from monitored building-related data and developing effective and ready to implement energy saving strategies in buildings. In this perspective one of the most frequently implemented operational improvement through EISs at system level, deals with the optimization of energy system schedules (e.g., lighting system, HVAC system).

Given the new paradigm of pervasive monitoring in buildings more and more data (not only related to energy consumption and system operative variables) are becoming available, completely changing the approach also to traditional energy management strategies such as HVAC system scheduling optimization. In particular, in the last few years, great attention has been paid to the collection and processing of occupancy data in buildings, since different system operations can be considered directly or indirectly based on the occupants' presence [194–196].

The next section presents the main research challenges related to HVAC scheduling improvements in buildings and introduces the motivations and novelty of proposed methodological approach.

3.2.1 Motivations and novelty of the proposed approach

Currently, most heating, cooling, ventilation and lighting systems are operated considering buildings as occupied with a fixed schedule that is assumed the same over time. In the majority of cases, this assumption differs significantly from the actual occupants' presence. As a consequence, EISs capable to exploit knowledge gathered from occupancy data can lead to considerable energy savings achievable by operating energy system with optimized occupancy-based schedules [197]. Starting from the literature reviewed in section 2.2.5, an advanced EIS tool for HVAC schedule improvement is developed and presented for demonstrating the impact related to the analysis of actual occupancy data in buildings.

The proposed process is effective as it can be generalized and is capable of driving energy managers in the definition of the most advantageous HVAC schedule when occupancy data are available. The aim is to evaluate optimised HVAC system operation schedules by displacing occupants with similar occupancy patterns and, similar arrival and exit times, in the same building thermal zone (i.e., by reducing occupancy diversity in the building). The novelty of the proposed methodology concerns the fact that the displacement of occupants from one thermal zone to another is proposed considering more than one typical occupancy profile. This is useful for fully exploiting the knowledge of occupant behaviours that can vary over different days of a week. In this way, the occupancy diversity in buildings can be furtherly reduced leading to higher energy savings achievable through system rescheduling. Despite, the usefulness of occupant related data, privacy issues could exist when they are not properly analysed.

For this reason, the proposed pattern recognition analysis is performed on aggregated occupancy information (i.e., related to groups of occupants) in order to preserve the privacy of each individual occupant [203].

The aim is to develop an EIS tool capable to define an optimal and fixed distribution of groups of occupants in the sub-zones of the building under analysis (considering fixed constraints, such as room capacity and occupancy pattern similarity) in order to reduce the operating hours of the HVAC system in the thermal zones of the building. The results obtained for the considered case study show that the HVAC scheduling improvement could determine a potential monthly reduction of the electricity use for HVAC (space heating, space cooling, ventilation and air treatments) that ranges from 12.2% to 15.4% while the average energy saving for the whole analysed period (4 months) amounts to 14%.

The developed methodology for schedule optimization, the description of the case study and obtained results are presented and discussed in detail in the following sections. In particular, section 3.2.2 provides a presentation of the case study considered for conducting the analysis; section 3.2.3 presents the developed methodological framework; section 3.2.4 and section 3.2.4.2 present the results obtained in terms of recognized occupancy patterns and displacement of the occupants. Eventually, section 3.2.6 discusses the results and contains the concluding remarks related to this specific EIS tool. As a remark part of the content of the section 3.2 was published as scientific article in the Elsevier journal "Sustainable city and societies" [14].

3.2.2 Case study used for developing the EIS tool at building system level

The EIS tool for HVAC scheduling optimization is developed through the analysis of anonymised occupancy data collected in the office building of the town hall of Zaanstad (Netherlands).

The Zaanstad Town Hall (Fig. 20) is a five-storey building located in the North Holland province, to the northwest of Amsterdam, in the Netherlands. It is characterised by a conditioned net floor area of about 23,000 m^2 and by a conditioned net volume of about 60,000 m^3. The Town Hall was built above a bus station, that is currently located on ground floor. The rest of the building includes a public area and the employees' offices.

Since the Town Hall was built in 2011, the building envelope is characterised by quite a high thermal performance. The opaque envelope is made up of a lightweight metal frame structure with a continuous external thermal insulation layer ($U = 0.27$ W·m^{-2}K^{-1}), while the windows are low-e double glazed and are filled with air ($U = 1.10$ W·m^{-2}K^{-1}). Each window is equipped with internal solar shading devices that are manually controlled.

The building is divided into five thermal zones: the bus station, the public area and three office thermal zones (named CD, EF and GH, respectively), as shown in Fig. 20. Each thermal zone has two kind of sub-zones: offices and corridors on each floor. In detail 30 office sub-zones are included in the three thermal zones considered (10 offices for each thermal zone).

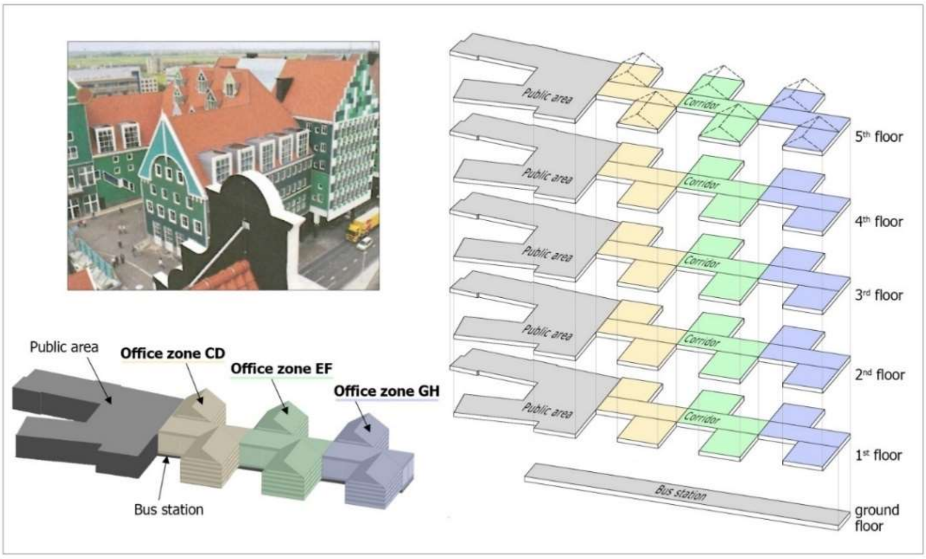


Fig. 20 - Zaanstad Town Hall: picture and geometric model [14]

Each thermal zone has an individual HVAC system with a centrally controlled temperature set point. In addition, each office is equipped with a thermostat that allows the occupants to increase or decrease the temperature set point by ± 2 °C. The HVAC system is usually turned on at 6:00 am and turned off at 9:00 pm during working days, while it is switched off during holidays and weekends. Each thermal zone is connected to a heat/cold storage, coupled to a geothermal heat pump, which provides space heating and space cooling, and to an air handling unit. Condensing gas boilers integrate the space heating and domestic hot water energy needs in each thermal zone.

The lighting system is controlled by means of presence sensors that have been installed in each office. The entire building is equipped with a BAS that includes several sensors, actuators and user interfaces and allows the energy manager and the energy providers to control and manage the HVAC system and the lighting system.

The following data are monitored on an hourly basis in each office: indoor air temperature, relative humidity of the indoor air, temperature set point and occupancy presence. Furthermore, the delivered electricity and natural gas are monitored for each building zone. The BAS is also connected to a meteorological station that is located on the roof of the building, which monitors the outdoor air temperature, the relative humidity, the wind speed and the wind direction.

Another monitoring system, called *FlexWhere,* tracks the presence of the employees in each office. This system monitors user's workstation login and visualises this information through different monitors located at the entrance of each room/working unit. All the workstations are connected on cloud, thus flexibility in moving employees through the building offices is guaranteed.

The *FlexWhere* system stores data that gives information on the current state of the workstations every 15 minutes. A 0 value indicates that a workstation is empty while 1 indicates it is occupied.

In this way, if an employee is temporary out of office when the workstation is still logged in, the workplace is classified as occupied. This situation points out some limitations of this approach for the assessment of occupant presence. However, the information related to arrival and exit times can instead be considered robust and reliable.

In order to develop the EIS tool for schedule optimization, the anonymised and aggregated occupancy data provided by the *FlexWhere* system are used to extract typical occupancy patterns for each office sub-zone.

Fig. 21 shows the hourly box plots of the measured number of occupants for the whole building and for each thermal zone considered. The boxplots are built considering both occupancy data related to weekdays and weekends (when the offices are unoccupied).

The descriptive statistics allows the occupancy patterns to be detected easily at a high level of aggregation and common assumptions to be made that were useful for the subsequent occupancy learning process.

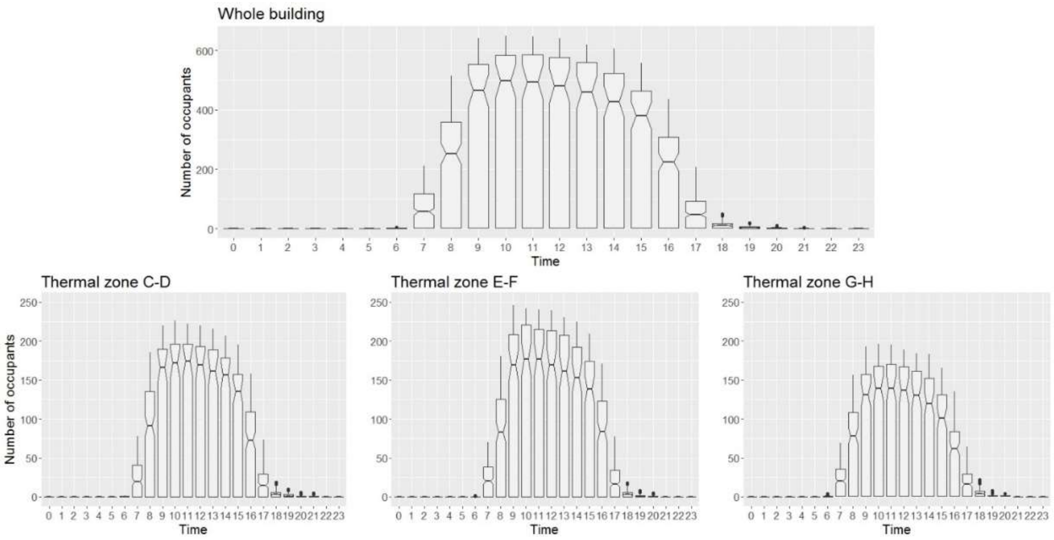


Fig. 21 - Hourly measured number of occupants in the whole building and in single thermal zones [14]

Fig. 21 shows that the arrival time at the building and thermal zone level varies over a very narrow range. For this reason, it can be assumed as a recurrent pattern that occurs at around 07:00 a.m. As a consequence, performing a schedule optimisation based on the arrival time would not produce any advantage, since the office sub-zones show very similar entry times of the occupants.

The exit time, instead, is affected by a greater variability, ranging from 17:00 to 20:00, depending on the thermal zone considered. This is further highlighted in Fig. 22 (a) and (b), which show the box plots of the number of occupants of two representative office sub-zones (office C1 and office D0), that belong to the same thermal zone (C-D).

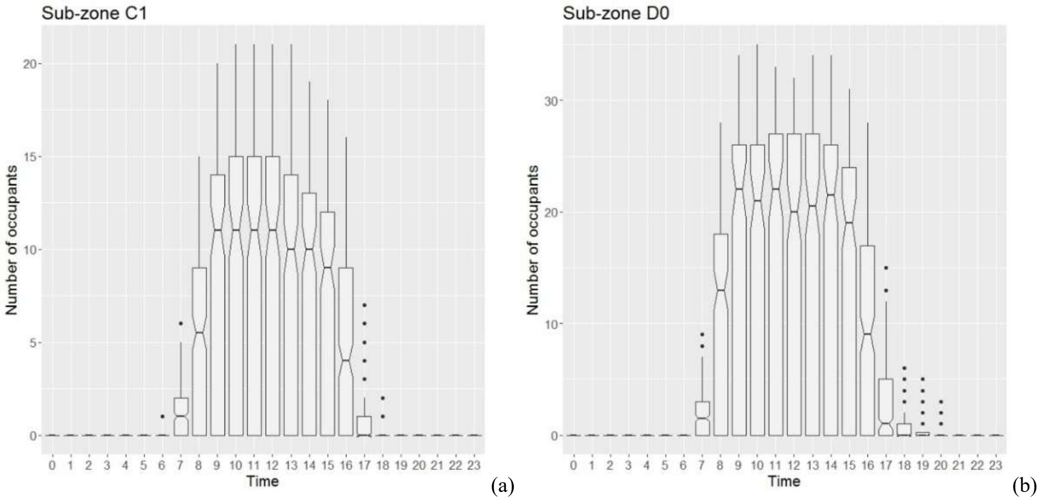


Fig. 22 - Hourly measured number of occupants in sub-zone C1(a) and sub-zone D0 (b) [14]

Therefore, the occupant displacement process may only affect the HVAC stop time of each thermal zone to any great extent. The start of the HVAC system is currently scheduled at 06:00 a.m. for all the thermal zones to guarantee thermal comfort on the arrival of the first employee that is assumed to occur at 07:00 a.m. Moreover, through dynamic simulation was estimated that one hour is the shortest boost period possible to ensure a comfortable temperature at 07:00 a.m. in the building. For this reason, all the successive procedures are aimed at optimising the HVAC schedule considering only the convenience of re-scheduling the shutdown time of the HVAC system for each thermal zone.

Fig. 23 shows a scatterplot of the maximum number of occupants recorded in a single timestep (i.e., 15 min) during the monitored period versus the design capacity for each office sub-zone.

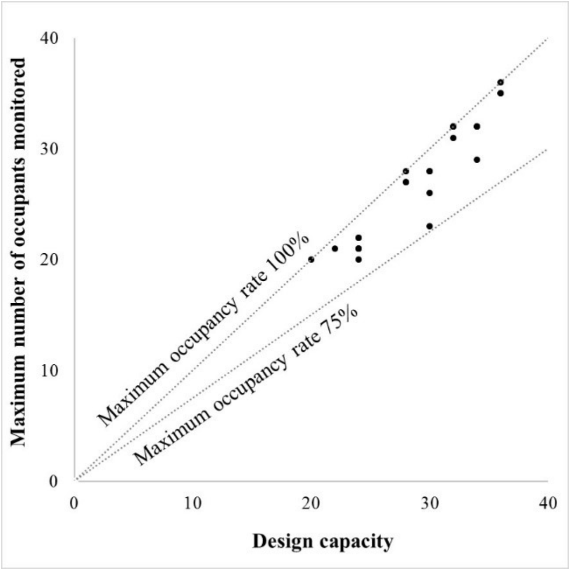


Fig. 23 - Maximum number of occupants vs. design capacity [14]

The office sub-zones are characterised by a design capacity that ranges from 20 to 36 occupants, with a value of the maximum occupancy rate for the monitored period that is never lower than 75% (Fig. 23). Considering that office sub-zones are characterised by a relatively high variability in terms of design capacity, and the actual maximum occupancy rate is not always equal to 100%, a preliminary labelling analysis is required. Indeed, the office sub-zones are labelled in order to understand which groups of occupants can be moved from an office to another one avoiding room capacity issues. The occupants that work in office sub-zones with the same capacity label can be then interchanged in the displacement process for reducing occupancy diversity in each thermal zone.

3.2.3 Implemented methodology for HVAC schedule optimisation exploiting occupancy data

As described in detail in Section 3.2.2, the HVAC schedule optimization EIS tool is developed for a building composed of three thermal zones for which the HVAC system can control the individual loads and ventilation rates. Each thermal zone is composed of 10 office sub-zones, and each office sub-zone is considered as the minimum level of aggregation of occupancy data.

The general framework of the whole methodological process of analysis unfolds over two different stages, which are shown in Fig. 24 and Fig. 25 respectively. The first stage aims at optimising the HVAC schedule according to the actual arrival and exit times of the occupants, by displacing the occupants with the most similar occupancy patterns to the same thermal zone. To this aim, a preliminary characterisation of the typical occupancy profiles of each sub-zone is performed by means of a data analytics-based process.

The second stage of the analysis aims at assessing the energy performance impact of the optimised occupancy-based HVAC schedules that are obtained in the first stage. To this purpose, a calibrated simulation model of the building is used.

3.2.3.1 Occupancy pattern analysis and reconfiguration process

This first stage of the analysis aims at finding an optimised HVAC schedule according to actual patterns of occupants' presence. The first stage of the methodology is shown in Fig. 24, and it is structured in three phases as follows:

1. Data preparation;
2. Recognition and classification of the occupancy patterns;
3. Reconfiguration process of the thermal zones through the occupants' displacement.

The aim of the first phase is to prepare the occupancy data, organising them in daily occupancy profiles aggregated at office sub-zone level. Successively, in the second phase, the typical daily occupancy profiles of each sub-zone are identified and classified considering robust explanatory variables (e.g. season, month, day of the week). Eventually, groups of occupants with similar occupancy patterns are displaced in the same thermal zone in order to reduce occupancy diversity as possible. The new occupants' configuration makes it possible to rationalise the number of operating hours of the HVAC system in each thermal zone.

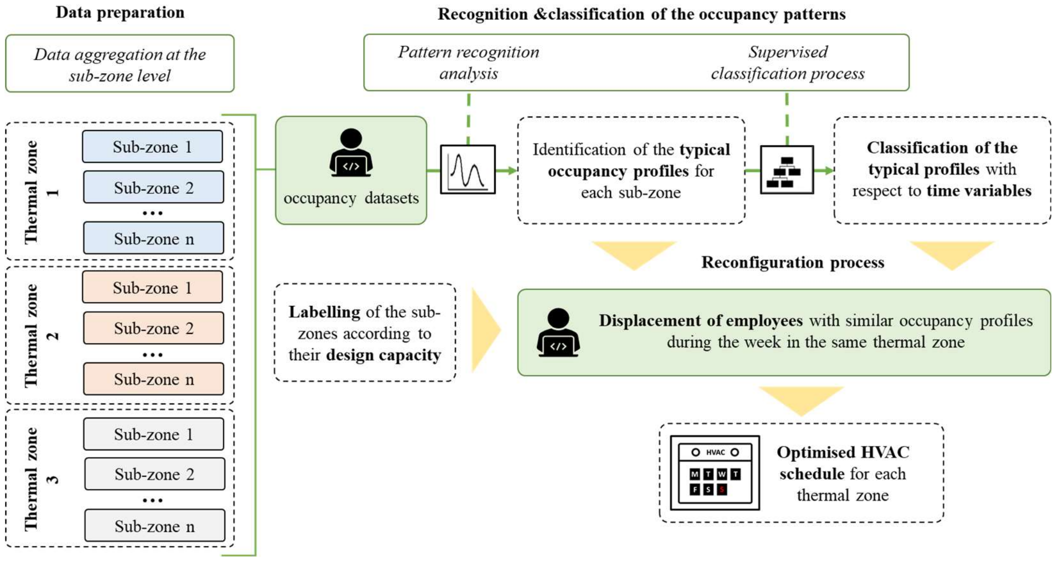

Data preparation
Data aggregation at the sub-zone level
Recognition &classification of the occupancy patterns
Pattern recognition analysis
Supervised classification process
Thermal zone 1
Sub-zone 1
Sub-zone 2
Sub-zone n
Thermal zone 2
Sub-zone 1
Sub-zone 2
Sub-zone n
Thermal zone 3
Sub-zone 1
Sub-zone 2
Sub-zone n
occupancy datasets
Identification of the typical occupancy profiles for each sub-zone
Classification of the typical profiles with respect to time variables
Reconfiguration process
Labelling of the sub-zones according to their design capacity
Displacement of employees with similar occupancy profiles during the week in the same thermal zone
Optimised HVAC schedule for each thermal zone

Fig. 24 - Occupancy pattern analysis and reconfiguration framework (adapted from [14])

3.2.3.1.1 Data preparation

A data aggregation processes is performed for preparing the time series of occupancy data in the proposed methodology. Data aggregation is a prerequisite for the analysis, considering that the displacement process should be investigated by aggregating and anonymizing occupancy data, in order to avoid privacy issues. The sub-zone aggregation level allows information to be extracted from occupancy time series of groups of about 20-30 occupants. These groups are large enough to preserve the privacy of the individual occupants, and at the same time sufficiently homogenous to define robust and representative occupancy patterns.

In a second step, the sub-zone occupancy time series are chunked in fixed daily sequences. The daily sequences, which represent the occupancy profiles, are then organized in an MxN matrix, where M is the number of daily occupancy profiles and N is the number of measurements per day (i.e., for a timestep of 15 min., N is equal to 96).

3.2.3.1.2 Recognition and classification of the occupancy patterns

Once the occupancy data are aggregated at a sub-zone level, a pattern recognition analysis is performed to discover the typical occupancy patterns of each sub-zone. The typical occupancy patterns of each sub-zone are extracted by clustering the daily occupancy profiles with a time interval of 15 minutes. The statistical objects to be clustered are represented by vectors of 96 components (daily occupancy profiles), where each component corresponds to the number of occupants in a sub-zone during a specific timestep. The outcome of this process is that *n* groups of daily occupancy profiles are defined for each sub-zone, and the typical occupancy patterns are evaluated by calculating the centroid of the profiles clustered together. Once the typical occupancy patterns are evaluated for each sub-zone, a supervised classification process is performed in order to characterise the patterns, considering time variables (e.g. month, day of the week, season) as input attributes. The pattern recognition is performed using a partitive clustering algorithm (i.e., k-means), while the classification process is performed adopting a binary recursive decision tree based on CART algorithm [204]. These two methods are well-known algorithms in the field of occupant behaviour and occupancy characterisation [58,205]. They proved to be effective in different occupancy pattern recognition applications and energy performance analysis in buildings due to their flexibility and their easy implementation.

3.2.3.1.3 Reconfiguration of the thermal zones through occupants' displacement

A labelling process of the sub-zones, considering their maximum capacity and the maximum occupancy rate recorded in the monitored period, is firstly carried out in the reconfiguration phase. The labelling procedure makes it possible to evaluate the physical constraints that need to be considered for the occupants' displacement. The reconfiguration of the thermal zones is based on the moving of groups of occupants that work together in the same office sub-zones to other sub-zones with similar capacities.

The results of the labelling procedure, together with the outcomes of the first two phases of the analysis, are used to aggregate groups of employees with similar occupancy profiles in the same thermal zone, according to the occupancy pattern and the constraints related to each sub-zone capacity. This third phase allows an optimised operation schedule of the HVAC system to be set up in each thermal zone.

3.2.3.2 Energy performance assessment

The second stage of the methodological process is aimed at assessing the impact of the HVAC schedule optimization on the energy performance of the building through a forward simulation approach.

A forward simulation model is built, according to the framework set out in Fig. 25, and data related to the climate, users, equipment, lighting (input data) and to the building features (fixed parameters) are introduced. The historical energy consumption data are then used to calibrate the model. The output of this model is used to build a baseline model for the impact evaluation of the strategy. A second forward model is built by implementing the optimised HVAC operation schedules, obtained from the occupancy patterns analysis and reconfiguration process (first stage), in the calibrated tailored baseline model. Eventually, the assessment of the impact is performed through the comparison between the outcomes of these two forward models.

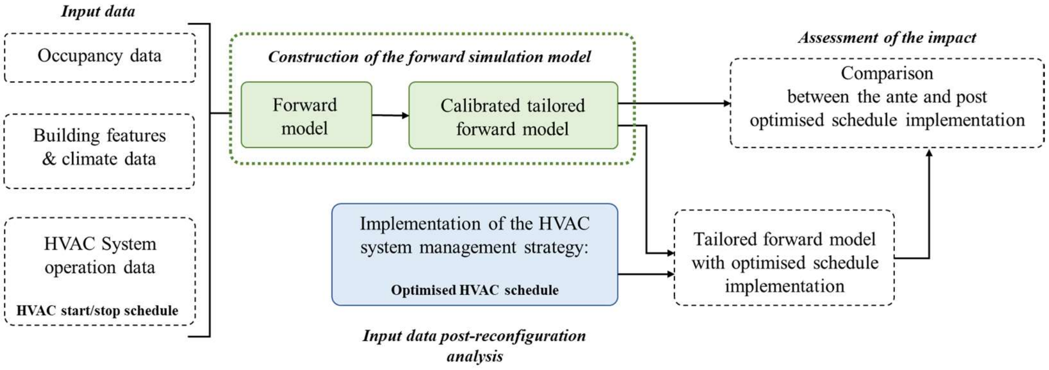

Input data
Occupancy data
Building features & climate data
HVAC System operation data
HVAC start/stop schedule
Construction of the forward simulation model
Forward model
Calibrated tailored forward model
Assessment of the impact
Comparison between the ante and post optimised schedule implementation
Implementation of the HVAC system management strategy:
Optimised HVAC schedule
Tailored forward model with optimised schedule implementation
Input data post-reconfiguration analysis

Fig. 25 - Energy performance assessment (adapted from [14])

This procedure is useful to assess the impact that the EIS tool can have on the final energy consumption of a building by optimizing HVAC schedules on the basis of actual occupancy data. The frameworks proposed hereafter can be easily and smoothly generalized to different kinds of building types and air conditioning systems.

3.2.4 Results obtained from HVAC schedule optimisation analysis

3.2.4.1 Recognition and classification of the occupancy patterns

In order to perform the pattern recognition analysis, the building occupancy data are firstly aggregated at office sub-zone level in order to perform analysis on occupancy profiles representative of about 20-30 occupants. In this way, thirty occupancy datasets (one for each office sub-zone) are identified without affecting occupant privacy.

In the methodological process proposed, privacy plays a key role, considering that the information related to the displacement of single occupants is often recognized as sensitive data [203]. The main potential drawback of this kind of data aggregation is that the average occupancy profile, related to each group of employees, may not be representative of the whole sample of employees if the occupants have very different habits in the same sub-zone. However, the assumption of considering an average profile for each sub-zone as a common occupancy pattern for a group of occupants that work in the same sub-zone is verified and demonstrated hereafter.

On the other hand, better results, in terms of occupancy diversity reduction, can be achieved by displacing one occupant at a time, although this could generate exclusion and marginalization processes. For this reason, the analyses are focused on aggregated occupancy profiles in order to keep the work context of each occupant unchanged after the displacement process.

The typical occupancy profiles for each office sub-zone are identified through a k-means clustering algorithm (discussed in section 2.1.3.2.1). In detail, two parameters are used as partitioning performance criteria for selecting the optimal number of clusters. The first parameter is the average within cluster distance, which is calculated by averaging the distance between the centroid and all the examples in a cluster. This parameter is a good indicator of inter-cluster similarity. The Davies-Bouldin index [206] is selected as a second indicator. The Davies-Bouldin index is based on an inter-cluster to intra-cluster distance ratio. Clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) have a low Davies–Bouldin index. The value of k that produced the set of clusters with the smallest Davies–Bouldin index is considered as the best number of partitions based on this parameter.

Once the occupancy profiles are clustered, the typical occupancy profiles of each office sub-zone are evaluated averaging all the profiles in the same cluster and then expressed in the form of daily profiles with a timstep of 15 minutes. Three or four typical occupancy patterns are discovered, depending on the sub-zone considered. Fig. 26 shows the centroids of the four clusters evaluated for one of the thirty sub-zones (sub-zone D0).

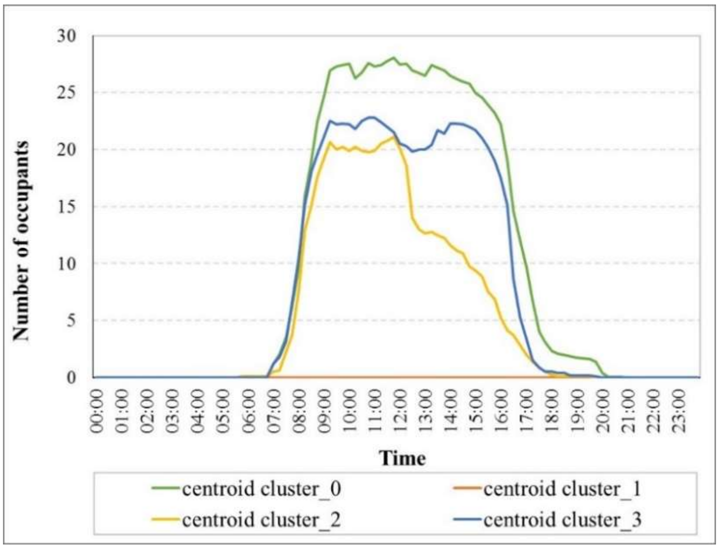


Fig. 26 - Typical occupancy patterns of sub-zone D0 [14]

Similar results are obtained for the remaining sub-zones. It can be observed that all the days during which the office sub-zone D0 is completely unoccupied (holidays and weekends) are grouped in cluster 1, while the remaining clusters describe three different occupancy patterns over the monitored period that need to be further characterised.

Moreover, it is verified that the typical occupancy profiles of each sub-zone are characterised by a low deviation, thus validating the assumption of considering an average profile as being representative of all the occupants that work in the same sub-zone. As a reference, Fig. 27 shows the average occupancy profile and the standard deviation of a cluster related to the sub-zone C1. It can be observed that the arrival and exit times are affected by lower deviation than the middle hours of the day. This outcome is also valid for all the clusters associated to the remaining sub-zones.

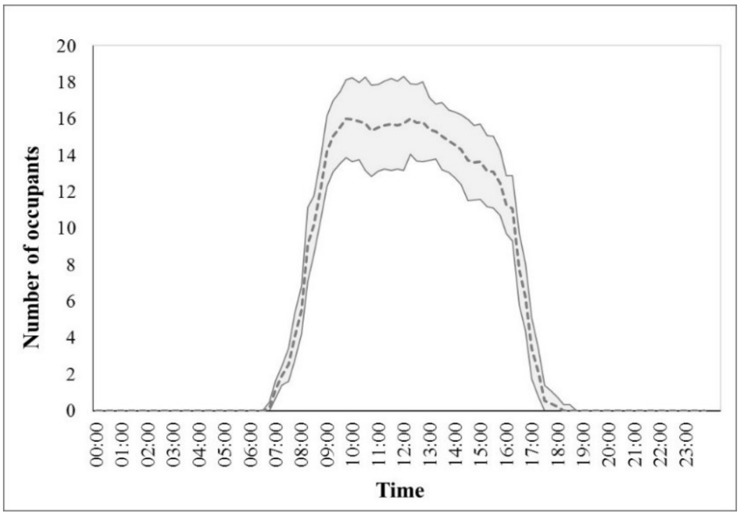


Fig. 27 - Occupancy profile sub-zone C1 (cluster 3) [14]

This fact demonstrates that even though the number of occupants may change slightly from day to day, the uncertainty related to their arrival and exit times can be considered negligible.

In the proposed data analytics framework, the occupancy patterns recognized by the cluster analysis are subsequently classified in order to learn the occupancy schedule rules. To this aim, a supervised classification process is developed in order to associate each typical occupancy pattern evaluated through k-means to a specific time reference period.

A classification tree (discussed in section 2.1.3.1.1.1) is built for each sub-zone considering the variables "*month*" and "*day of the week*" as input attributes. These variables proved to be able to explain the properties of different typical occupancy patterns discovered [205].

Fig. 28 shows the output of the classification process for the occupancy patterns of sub-zone D0 in the form of a decision tree. The classifier does not include "*month*" as a splitting variable. This means that, during the monitored time period (i.e., four months), the occupancy patterns are closely related to the day of the week, but independent from the month of the year. It is found, for sub-zone D0, that:

- The objects classified as belonging to cluster 0 are daily occupancy profiles of Monday, Tuesday and Thursday in 85.2% of the cases;
- The objects classified as belonging to cluster 1 are daily occupancy profiles of Saturday and Sunday in 100% of the cases;
- The objects classified as belonging to cluster 2 are daily occupancy profiles of Friday in 88.9% of the cases;
- The objects classified as belonging to cluster 3 are daily occupancy profiles of Wednesday in 66.7% of the cases.

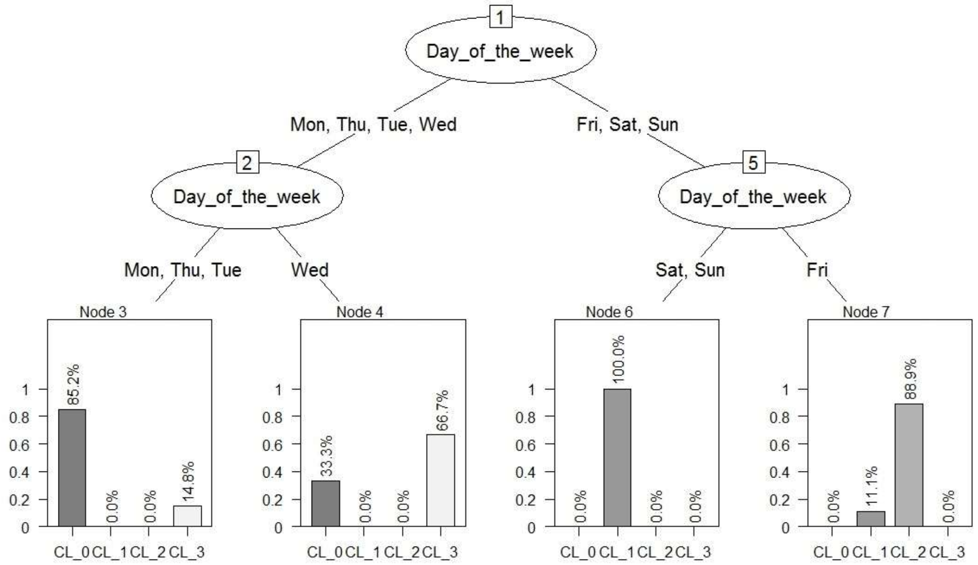


Fig. 28 - Classification of the occupancy patterns of sub-zone D0 [14]

The Gini index is used in the implemented classification and regression tree to establish the degree of impurity of each node. The k-fold cross-validation method is used to evaluate the accuracy of the classification tree. The classification tree is initially developed by setting the minimum number of cases in the parent and child nodes (10 and 8 cases, respectively), and the maximum decrease in the impurities of each split is set equal to 0.01. The tree-growing process is stopped before decision tree is generated in its maximum size. In this way, by means of an early stopping rule it is possible to overcome the problem of model overfitting.

The classification process is useful for characterizing the patterns of each sub-zone that need to be considered in the thermal zone reconfiguration process. Thanks to the coupling of the cluster analysis and the classification tree, it is possible to identify, with a high level of accuracy, the group of weekdays that have similar occupancy profiles for which the process of occupants' displacement can be extended. In fact, the reconfiguration phase is aimed at aggregating group of employees with similar occupancy profiles in the same thermal zone as long as the considered occupancy profiles referred to the same time period. As an alternative, seven average daily occupancy profiles – one for each day of the week – could be computed, but this would increase the complexity and the computational cost of the reconfiguration process. Moreover, the developed classification trees could also be used for occupancy schedule prediction purposes, if adequately trained, tested and validated for a monitoring period of at least an entire year [205].

As a final result, each sub-zone is characterised by at least 3 occupancy patterns composed of:

- One representative occupancy pattern for Monday, Tuesday and Thursday, that are generally the days with the highest occupancy rate for all the sub-zones;
- One representative occupancy pattern for weekends and holidays, when all the sub-zones are unoccupied;
- One representative occupancy pattern for Friday, that is generally the day with the lowest occupancy rate for all the sub-zones.

For what Wednesday is concerned, it results to be the day characterized by the greatest diversity between the occupancy patterns among the sub-zones.

On one hand, for sub-zones characterized by 3 typical patterns, Wednesday has an occupancy profile similar to Friday or to the group of days that includes Monday, Tuesday and Thursday. On the other hand, for some sub-zones Wednesday exhibits a completely different occupancy profile from all the other days of the week justifying the existence of an additional pattern.

In this way, excluding the weekend pattern (representative of the unoccupied period), it is not possible to identify less than three reference periods for the working days of a week:

- *First reference period*: Monday, Tuesday and Thursday;
- *Second reference period*: Friday;
- *Third reference period*: Wednesday.

In the following section, the displacement process is performed considering the convenience of displacing occupants in the same thermal zone according to the similarity of occupancy profiles that refer the same reference period. In fact, all the potential occupants' displacements that could generate improvements in re-scheduling the operation of the HVAC system are considered in a three-step reconfiguration process (one step for each reference period).

3.2.4.2 Reconfiguration of thermal zones through the occupants' displacement

3.2.4.2.1 Capacity labelling of the sub-zones

As stated in section 3.2.3.1.3, a preliminary capacity labelling of each sub-zone is performed to evaluate the physical constraints of thermal zone reconfiguration. In this phase, the sub-zones are labelled according to their maximum capacity and the maximum number of occupants recorded during the monitored period. This first procedure makes it possible to identify the physical constraints that need to be taken into account during the occupants' displacement.

The reconfiguration process is based on the moving of groups of occupants that work together in the same office sub-zones to other sub-zones with similar capacities in order to group together occupants with similar presence patterns in the same thermal zone. In order to maximize the potential number of interchanges both design capacity (DC) and maximum occupancy rate (MNO) of each sub-zone are considered. The sub-zones C0 - E0 - G0 - H0 (4 offices located at ground floor) are excluded from the labelling process, because they are not connected to FlexWhere system. The capacity labelling procedure unfolds over the following steps:

1. The sub-zones are ordered according to their design capacity (DC_1, DC_2, DC_3…DC_n with $DC_j > 0$).
2. The maximum number of occupants recorded during the monitored period is evaluated for each sub-zone (MNO_1, MNO_2, MNO_3…MNO_n with $MNO_j \geq 0$).
3. The minimum design capacity (DC_1) is selected as a reference value.
4. According to the design capacity order, the first capacity label was assigned to those sub-zones that verify the following condition (Eq. 14):

   $$MNO_j \leq DC_1$$

   Eq. 14

   for $2 \leq j \leq n$

5. According the design capacity order, the first sub-zone for which the condition (Eq. 14) is not verified becomes the new reference sub-zone, and its design capacity is set as the reference value for the labelling process of the remaining sub-zones.

As a result of this process, five labels are assigned to the office sub-zones (Fig. 29) as follow:

- Label A: office sub-zone with DC equal to 20 occupants;
- Label B: office sub-zones with DC ranging from 22 to 24 occupants;
- Label C: office sub-zones with DC ranging from 28 to 30 occupants;
- Label D: office sub-zones with DC ranging from 32 to 34 occupants;
- Label E: office sub-zones with DC equal to 36 occupants.

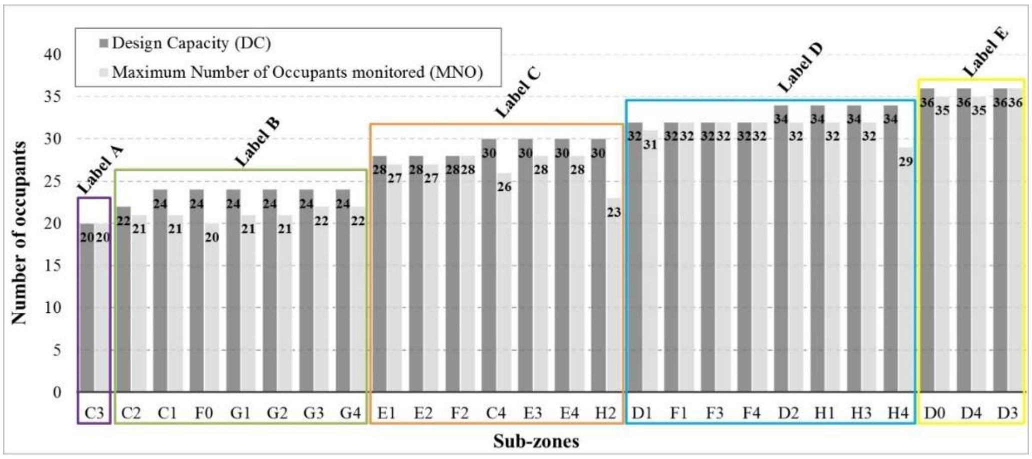


Fig. 29 - Capacity labels of the sub-zones [14]

Although some sub-zones are involved in the labelling process, they are successively excluded from the occupants' displacement analysis for the following reasons:

- Sub-zones D0-D3-D4 (**Label E**): These sub-zones are characterised by the highest design capacity and are located in the same thermal zone (thermal zone C-D). For this reason, it was not possible to move their occupants to other thermal zones.
- Sub-zone C3 (**Label A**): This is the smallest sub-zone, in terms of design capacity. The group of occupants in this office sub-zone could be displaced to any other sub-zone, but the opposite is never possible.

Table 2 - sub-zone capacity label [14]

Thermal zone	ID Sub-zone	Design capacity	Maximum number of occupants	Capacity label
C-D	C1	24	21	B
	C2	22	21	B
	C3	20	20	-
	C4	30	26	C
	D0	36	35	-
	D1	32	31	D
	D2	34	32	D
	D3	36	36	-
	D4	36	35	-
E-F	E1	28	27	C
	E2	28	27	C
	E3	30	28	C
	E4	30	28	C
	F0	24	20	B
	F1	32	32	D
	F2	28	28	C
	F3	32	32	D
	F4	32	32	D
G-H	G1	24	21	B
	G2	24	21	B
	G3	24	22	B
	G4	24	22	B
	H1	34	32	D
	H2	30	22	C
	H3	34	32	D
	H4	34	29	D

The occupants that work in these office sub-zones cannot be moved elsewhere, and for this reason their typical occupancy profiles are used as constraints in the reconfiguration analysis. These profiles are considered as target patterns, with respect to which it is necessary to ensure similarity during the reconfiguration process. Eventually, only the groups of occupants that work in office sub-zones with the same capacity labels can be interchanged. Table 2 reports the complete list of thermal zones, sub-zones, design capacity, maximum number of occupants and capacity labels (sub-zones D0, D3, D4 and C3, which are excluded from the occupants' displacement analysis, are in bold).

3.2.4.2.2 Reconfiguration of the thermal zones

After the pattern recognition, the classification analysis and the capacity labelling, it is possible to initialize the thermal zone reconfiguration process. In this phase the convenience of aggregating occupants with similar occupancy profiles in the same thermal zone is explored. The analysis is conducted evaluating the similarity between occupancy patterns that refer to the same reference time period (same group of weekdays). The concept of similarity involves only the occupants' exit time, given that the reconfiguration process is aimed at optimising the stop schedule of the HVAC system for a typical week. As already discussed in section 3.2.2, the start schedule of the HVAC system is considered to be already optimal.

For this reason, a reference occupants' exit time is extracted from each occupancy pattern evaluated through the clustering analysis. This time corresponds to the last time the presence of at least 1 occupant is observed. Considering all the occupancy profiles, the extracted exit times range between 16:15 and 20:15 (i.e., exit time interval). In order to reduce the computational cost of the reconfiguration process the exit time interval is divided into time windows of fixed length of 30 minutes. The splitting of the exit times into time windows is particularly useful to reduce the order of the optimisation problem. Then each time window is labelled with a symbol as follow:

- Symbol **a**: reference exit time equal to 16:45;
- Symbol **b**: reference exit time equal to 17:15;
- Symbol **c**: reference exit time equal to 17:45;
- Symbol **d**: reference exit time equal to 18:15;
- Symbol **e**: reference exit time equal to 18:45;
- Symbol **f**: reference exit time equal to 19:15;
- Symbol **g**: reference exit time equal to 19:45;
- Symbol **h**: reference exit time equal to 20:15.

In this way, an exit time vector, composed of three letters (one for each reference period during the typical week), can be associated to each group of occupants. For example if a group of occupants has a reference exit time for Monday, Tuesday and Thursday equal to 17:45, a reference exit time for Wednesday equal to 17:15 and a reference exit time for Friday equal to 16:15, its exit time vector can be expressed through the symbol sequence *c-b-a*.

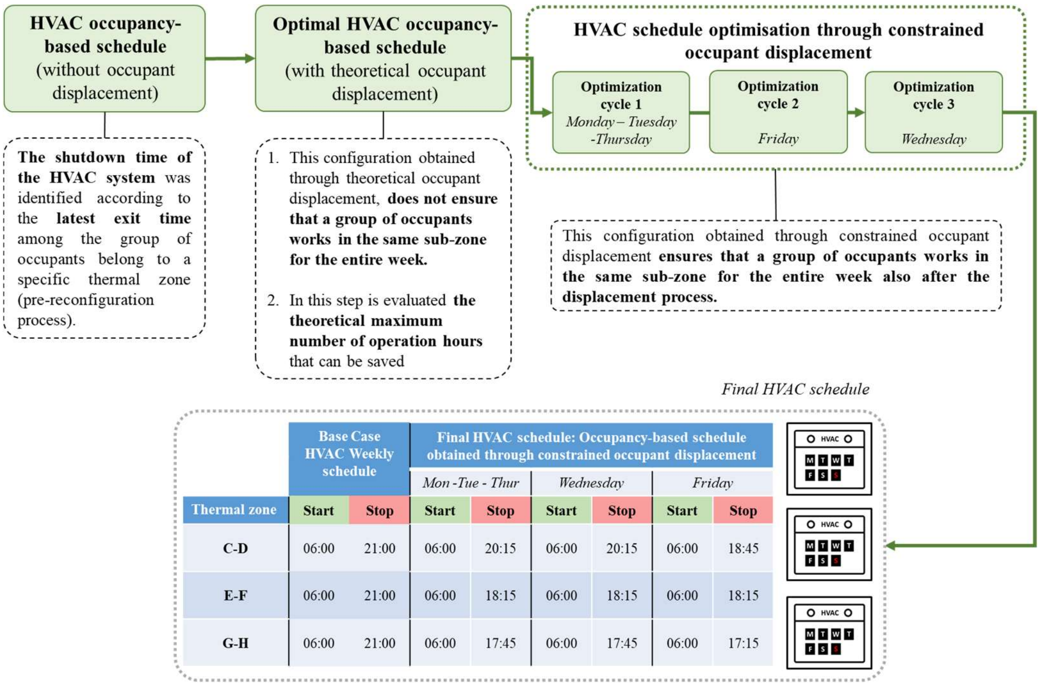
Starting from the encoding of exit times during the three reference periods, two main hypotheses can be assumed for carrying out the thermal zone reconfiguration process through occupant displacement:

- *Constrained occupant displacement*: A group of occupants can be moved only one time and then the new work location does not change for the entire week;

- *Theoretical occupant displacement*: A group of occupants can be moved one time for each of the reference time periods. In this way, for a group of occupants, the work location can change up to three times during the week.

Despite the second hypothesis could lead to a higher reduction of HVAC operation hours, it is characterized by poor feasibility. For this reason, in the reconfiguration process each displacement of occupants involve the entire week, thus ensuring that the thermal zone configuration does not change from one reference period to another.

However, the second hypothesis is also tested, but only with the aim of evaluating a theoretical reduction of HVAC operation hours. Fig. 30 shows the reconfiguration procedure in form of flowchart.



HVAC occupancy-based schedule (without occupant displacement)

The shutdown time of the HVAC system was identified according to the **latest exit time** among the group of occupants belong to a specific thermal zone (pre-reconfiguration process).

Optimal HVAC occupancy-based schedule (with theoretical occupant displacement)

1. This configuration obtained through theoretical occupant displacement, **does not ensure that a group of occupants works in the same sub-zone for the entire week.**
2. In this step is evaluated **the theoretical maximum number of operation hours** that can be saved

HVAC schedule optimisation through constrained occupant displacement

Optimization cycle 1 *Monday – Tuesday -Thursday*

Optimization cycle 2 *Friday*

Optimization cycle 3 *Wednesday*

This configuration obtained through constrained occupant displacement **ensures that a group of occupants works in the same sub-zone for the entire week also after the displacement process.**

Final HVAC schedule

	Base Case HVAC Weekly schedule		Final HVAC schedule: Occupancy-based schedule obtained through constrained occupant displacement					
			Mon -Tue - Thur		*Wednesday*		*Friday*	
Thermal zone	**Start**	**Stop**	**Start**	**Stop**	**Start**	**Stop**	**Start**	**Stop**
C-D	06:00	21:00	06:00	20:15	06:00	20:15	06:00	18:45
E-F	06:00	21:00	06:00	18:15	06:00	18:15	06:00	18:15
G-H	06:00	21:00	06:00	17:45	06:00	17:45	06:00	17:15

Fig. 30 - Process of reconfiguration of thermal zones through occupant displacement (adapted from [14])

The steps of the process are described hereafter:

1 *Identification of the HVAC occupancy-based schedule (without occupant displacement):*

The first step of the procedure is aimed at evaluating the HVAC stop time for each thermal zone, according to the actual occupant presence patterns without considering any occupant displacement.

In this case, the objective is to evaluate the reduction of HVAC operation hours (N*) for the three thermal zones during a week, in comparison to the base case scenario, where the system is operated with a fixed schedule from 6:00 to 21:00. The shutdown time of the HVAC system is identified for each reference time period according to the greatest exit time symbol associated to the last occupant group that leave the thermal zone. The reduction of HVAC operation hours (N*) is found to be 27.15 hours per week, and this figure is taken as a reference value to calculate the additional improvement that could be achieved through the occupant displacement process.

2 *Identification of the optimal HVAC occupancy-based schedule (with theoretical occupant displacement):*

All the possible displacements of occupant groups are computed independently for each reference period, in order to aggregate occupants that have the encoded exit time symbols as similar as possible in the same thermal zone. As a reference, occupants that work in the same sub-zone can be moved to another sub-zone with the same capacity label only if it is located in a different thermal zone.

The aim of this step is to find a further theoretical reduction of HVAC operation hours (N_{th}) for the three thermal zones and for all the reference periods respect to the solution obtained in the step 1. This can be considered a theoretical limit, because the optimization is performed taking into account each reference time period independently from the others. In other words, a group of occupants can be moved one time for each of the reference time periods. In this way, in order to reduce the occupancy diversity in the thermal zone, for a group of occupants the work location can change up to three times during the week. The chosen shutdown time of the HVAC system, after the reconfiguration, corresponds to the exit time of the last occupant group that leave the thermal zone. The further theoretical cumulated reduction of the HVAC operation hours for all thermal zones, is found to be:

- 6 hours for the Monday-Tuesday-Thursday reference period;
- 1 hour for the Wednesday reference period;
- 2 hours for the Friday reference period.

The optimal solution for a typical week converges to a theoretical further reduction of N_{th}= 9 hours with respect to the N*=27.15 hours obtained in the previous step. Starting from this preliminary result, the following reconfiguration process is performed through a constrained occupant displacement process moving the groups of occupants only one time in order to maintain unchanged the work location for the entire week. The main objective is to obtain a reduction of HVAC operation hours as close as possible to the theoretical solution N_{th}. To this purpose one optimization cycle for each reference period is run for moving occupants as shown in Fig. 31.

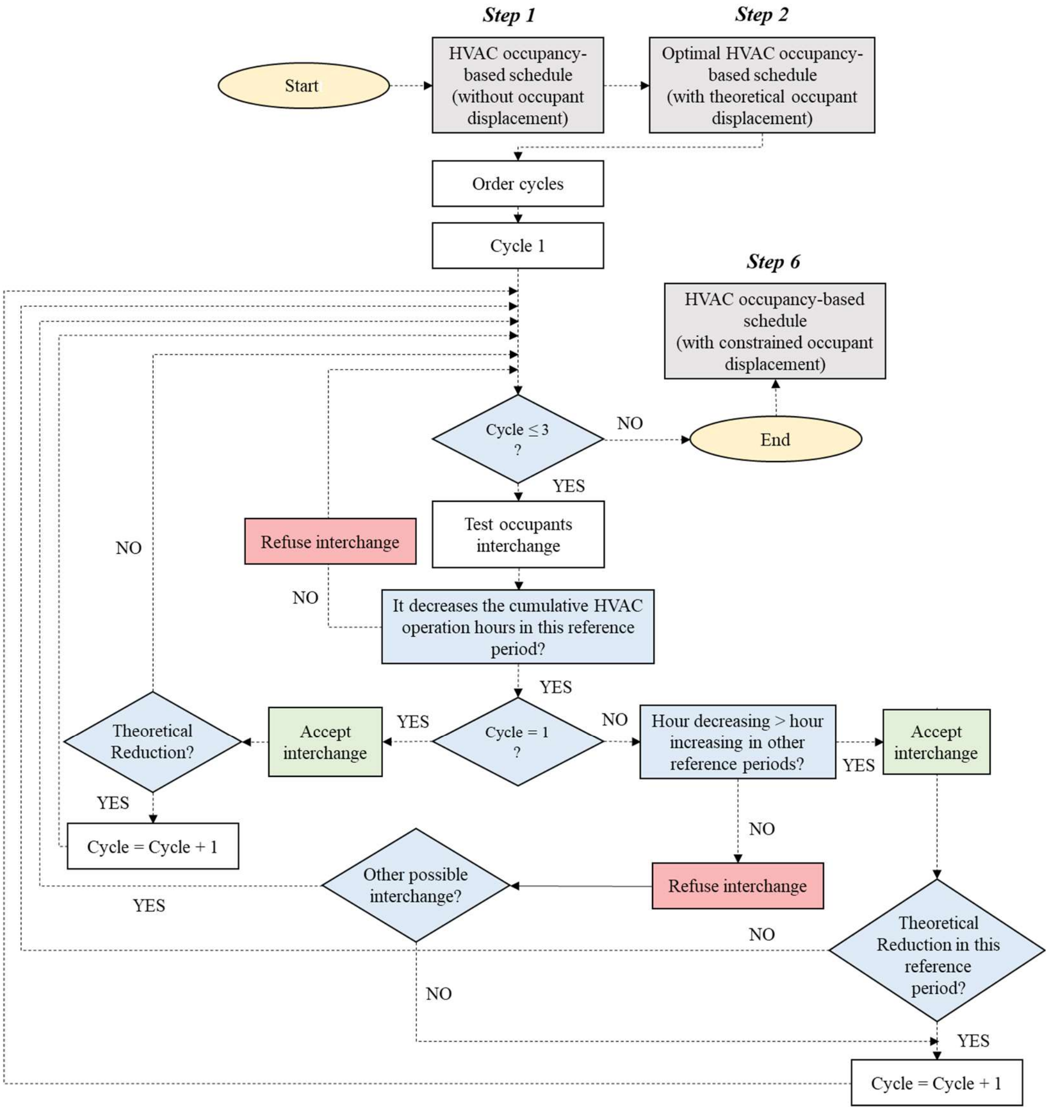

Step 1
Step 2
Start
HVAC occupancy-based schedule (without occupant displacement)
Optimal HVAC occupancy-based schedule (with theoretical occupant displacement)
Order cycles
Cycle 1
Step 6
HVAC occupancy-based schedule (with constrained occupant displacement)
Cycle ≤ 3 ?
NO
End
YES
Test occupants interchange
Refuse interchange
NO
NO
It decreases the cumulative HVAC operation hours in this reference period?
YES
Theoretical Reduction?
Accept interchange
YES
Cycle = 1 ?
NO
Hour decreasing > hour increasing in other reference periods?
YES
Accept interchange
YES
Cycle = Cycle + 1
NO
Other possible interchange?
Refuse interchange
YES
NO
Theoretical Reduction in this reference period?
NO
YES
Cycle = Cycle + 1

Fig. 31 - HVAC schedule optimisation through constrained occupant displacement

3 *First optimisation cycle of the HVAC occupancy-based schedule (with constrained occupant displacement):*

The reconfiguration process is conceived for optimising the displacement of each group of occupants that worked in the same sub-zones to another sub-zone with the same capacity label in order to reduce occupancy diversity in the thermal zone.

The first optimisation cycle (Fig. 31) only takes into account the Monday-Tuesday-Thursday reference period. The cycle ended when no other change generated an improvement, that is, a reduction in the operation time of the HVAC system in all the thermal zones, with respect to the HVAC stop schedule evaluated in step 1 for the same reference period.

4 *Second optimisation cycle of the HVAC occupancy-based schedule (with constrained occupant displacement):*

The second optimisation cycle (Fig. 31) only takes into account the Friday reference period. The optimisation cycle ends when no other change generates an improvement, that is, a reduction in the sum of the operation hours of the system in all the thermal zones, with respect to the HVAC stop schedule evaluated in step 1 for Friday. Moreover, a displacement is considered admissible, if it generates a greater reduction in the total operating hours on Friday than the potential increase of operating hours for the previously optimised reference time period.

5 *Third optimisation cycle of the HVAC occupancy-based schedule (with constrained occupant displacement):*

The third optimisation cycle (Fig. 31) only takes into account the Wednesday time reference period. The third optimization cycle ends when no other change generated an improvement, that is, a reduction in the sum of the operation hours of the system in all the thermal zones, with respect to the HVAC stop schedule evaluated in step 1 for Wednesday. A displacement is considered admissible if it generates a greater reduction in the total operating hours on Wednesday than the increase in the sum of operating hours for the other previously optimised reference time periods.

6 *Identification of the HVAC occupancy-based schedule obtained through the constrained occupant displacement:*

At the end of the reconfiguration process (Fig. 31) the final shutdown schedule of the HVAC system for each reference time period corresponds to the greatest exit time symbol associated to the last occupant group that leave the thermal zone.

The reconfiguration process does not converge to a single solution, but instead converges to *n* equivalent solutions. Table 3 shows the optimal solution for the HVAC occupancy-based schedule obtained through the constrained occupant displacement process. Each group of occupants is labelled by specifying the sub-zone of origin before the reconfiguration process. For example, the label O_{C2} refers to the group of occupants that before the reconfiguration process used to work in the sub-zone C2. However, after the reconfiguration process the sub-zone C2 is occupied by the group of occupants labelled as O_{G3}. This displacement of occupants is admissible, because the capacity constraint is respected. In fact, the sub-zone C2 and the sub-zone G3 have the same capacity label (i.e., label B). Due to the constraints considered in the analysis, the best solution generated a reduction of 8.5 operation hours respect to the solution evaluated in step 1. The deviation from the theoretical optimal solution (step 2) can be considered negligible and it is equal to 30 minutes.

Table 3 - Final optimised HVAC occupancy-based schedule obtained through the constrained occupant displacement process [14]

				Exit time symbol*					
		Group of occupants		Mon–Tue–Thu		Wednesday		Friday	
Thermal zone	Sub-zone**	Ante reconfig.	Post reconfig.	Ante reconfig.	Post reconfig.	Ante reconfig.	Post reconfig.	Ante reconfig.	Post reconfig.
C-D	C1	O_{C1}	O_{G1}	*b*	*d*	*b*	*b*	*a*	*b*
	C2	O_{C2}	O_{G3}	*c*	*d*	*c*	*d*	*c*	*c*
	C3	O_{C3}	O_{C3}	*d*	*d*	*d*	*d*	*b*	*b*
	C4	O_{C4}	O_{H2}	*a*	*h*	*a*	*h*	*a*	*e*
	D0	O_{D0}	O_{D0}	*g*	*g*	*b*	*b*	*b*	*b*
	D1	O_{D1}	O_{F3}	*c*	*d*	*c*	*d*	*b*	*c*
	D2	O_{D2}	O_{F1}	*c*	*c*	*a*	*d*	*a*	*e*
	D3	O_{D3}	O_{D3}	*d*	*d*	*d*	*d*	*c*	*c*
	D4	O_{D4}	O_{D4}	*e*	*e*	*e*	*e*	*d*	*d*
	HVAC system stop time			**19:45**	**20:15**	**18:45**	**20:15**	**18:15**	**18:45**
E-F	E1	O_{E1}	O_{E1}	*c*	*c*	*c*	*c*	*c*	*c*
	E2	O_{E2}	O_{E2}	*d*	*d*	*d*	*d*	*b*	*b*
	E3	O_{E3}	O_{E3}	*d*	*d*	*c*	*c*	*c*	*c*
	E4	O_{E4}	O_{E4}	*d*	*d*	*d*	*d*	*d*	*d*
	F0	O_{F0}	O_{C2}	*b*	*c*	*b*	*c*	*a*	*c*
	F1	O_{F1}	O_{H3}	*c*	*d*	*d*	*c*	*e*	*c*
	F2	O_{F2}	O_{F2}	*c*	*c*	*c*	*c*	*c*	*c*
	F3	O_{F3}	O_{H1}	*d*	*c*	*d*	*c*	*c*	*c*
	F4	O_{F4}	O_{F4}	*c*	*c*	*b*	*b*	*b*	*b*
	HVAC system stop time			**18:15**	**18:15**	**18:15**	**18:15**	**18:45**	**18:15**
G-H	G1	O_{G1}	O_{F0}	*d*	*b*	*B*	*b*	*b*	*a*
	G2	O_{G2}	O_{G2}	*c*	*c*	*c*	*c*	*b*	*b*
	G3	O_{G3}	O_{C1}	*d*	*b*	*d*	*b*	*c*	*a*
	G4	O_{G4}	O_{G4}	*c*	*c*	*c*	*c*	*b*	*b*
	H1	O_{H1}	O_{D2}	*c*	*c*	*c*	*a*	*c*	*a*
	H2	O_{H2}	O_{C4}	*h*	*a*	*h*	*a*	*e*	*a*
	H3	O_{H3}	O_{D1}	*d*	*c*	*c*	*c*	*c*	*b*
	H4	O_{H4}	O_{H4}	*b*	*b*	*b*	*b*	*b*	*b*
	HVAC system stop time			**20:15**	**17:45**	**20:15**	**17:45**	**18:45**	**17:15**

*Exit time symbols: ***a*** = 16:45, ***b*** = 17:15, ***c*** = 17:45, ***d*** = 18:15, ***e*** = 18:45, ***f*** = 19:15, ***g*** = 19:45, ***h*** = 20:15.
**The sub-zones excluded from the occupant displacement analysis are in bold.

3.2.5 Impact assessment of HVAC scheduling optimisation

3.2.5.1 Boundary conditions, assumptions and calibration of the model

A forward simulation model is developed to assess the energy performance of the case study before and after the implementation of the HVAC occupancy-based schedule obtained through the occupant displacement process. The analysis is aimed at estimating the potential impact of an advance EIS capable to exploit actual occupancy data for supporting energy managers in achieve scheduling improvements of the HVAC system in their buildings. The calculation is carried out by means of a detailed simulation tool, *EnergyPlus 8.5* [207], and using the building geometry interface of *DesignBuilder 5.0*.

The public area and the bus station are excluded, as they are not affected by the implementation of the energy management strategy proposed by the EIS. The three thermal zones (C-D, E-F, G-H) are modelled separately, considering a multi-zone calculation, without thermal coupling between zones.

In the cases where the use of the building is known, the actual data are considered, otherwise standard values according to ISO 18523-1 [208] are taken into account.

The maximum heat load related to occupancy is derived from ISO 18523-1, and is scaled with respect to the normalized actual occupancy profile for each sub-zone evaluated in Section 3.2.4.1. The schedules of the heat loads (related to lighting and appliances) and of the ventilation rate are built according to the normalised actual occupancy profiles that are gathered from measurements. The temperature set-point for heating is derived from hourly time-step measurements.

The energy baseline model is calibrated considering the actual weather data and the building energy consumption in the period from January to April 2016. The monitored variables are the outdoor air temperature, relative humidity, wind speed and wind direction; the remaining weather data are derived from the IWEC (*International Weather for Energy Calculations*) data set. The available energy consumption data are the metered hourly values of the overall delivered electricity for each office zone, including all energy uses, i.e. heating, cooling, mechanical ventilation, lighting and appliances. No breakdown of the HVAC consumption into final energy uses is available.

In order to compare the results of the energy performance simulation with the actual energy consumption, some elaborations on the hourly global amount of electricity consumption data are carried out considering the following assumptions:

- The electricity consumption during the unoccupied hours (i.e. night hours, weekends and holidays) is attributed to the energy uses that are considered independent from the building occupancy (e.g. server room, stand-by parasitic power).
- The electricity consumption during the occupied hours is considered to depend on the building occupancy and is therefore mainly due to space heating, space cooling, ventilation and air treatments, lighting and plug loads.

Consequently, the global energy consumption is split between the unoccupied periods and the occupied periods, and the model calibration is performed considering only the energy consumed in the occupied periods.

Fig. 32 shows an example of metered hourly electricity consumption data, referring to the first week of April 2016, for office zone G-H. The energy consumption in the occupied and unoccupied periods is marked with different shades of grey.

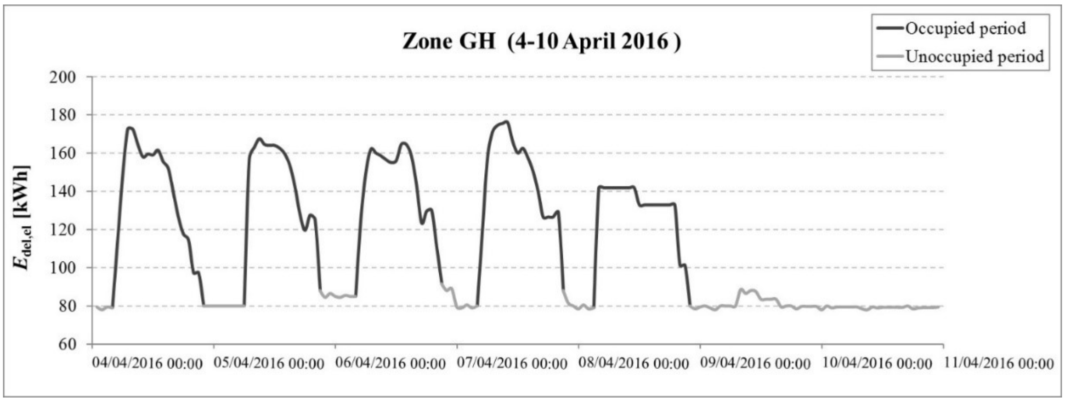


Fig. 32 - Monitored hourly electricity consumption of zone G-H during the first week of April 2016 [14]

Fig. 33 (a) and Fig. 33 (b) show the comparison between the monitored and simulated monthly electricity consumptions for the whole office building and for each thermal zone for the period January-April 2016.

The monthly deviations between the actual and the estimated consumptions are very low; the deviations for the whole office building range from 0.3% in April to 2.2% in January. For what concerns the single zones, the highest deviation can be observed for the thermal zone EF (3.11%). On average, the forward energy simulation model overestimates the actual consumption values by 0.45%.

The accuracy of the model calibration is verified according to ASHRAE Guideline 14 [209], which provides acceptable tolerances of the calibration through the use of two indexes, as follows: (i) MBE_{month} = ±5%, and (ii) $C_{\mathrm{V}}(RMSE_{\mathrm{month}})$ = 15%. For the conducted simulations, MBE_{month} and $C_{\mathrm{V}}(RMSE_{\mathrm{month}})$ are 0.45% and 1.51% respectively. These results denote an accurate calibration of the baseline model. Since a breakdown of the HVAC consumption is not available, a monthly basis calibration process was preferred to an hourly calibration. This choice is considered more appropriate and compatible with the details of the available monitored data.

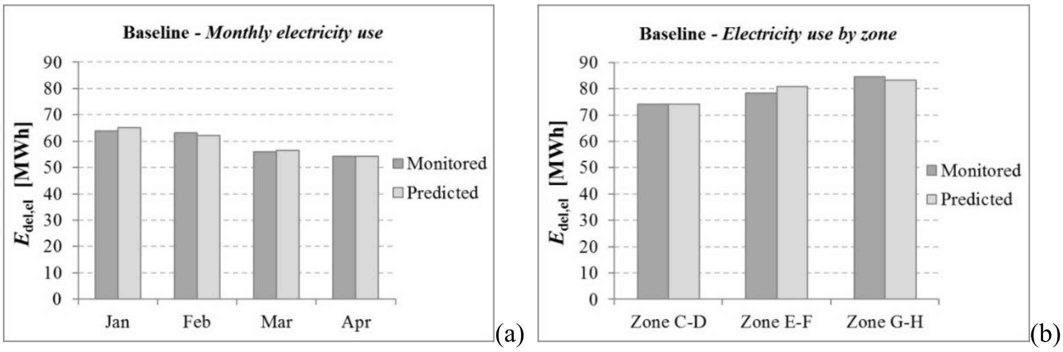


Fig. 33 - Comparison between the monitored and predicted electricity consumption values on a monthly basis for the whole office part (a), and by considering the office zone for the whole analysed period (b) [14]

3.2.5.2 Assessment of the energy savings

The forward simulation models of the case study before and after the implementation of the HVAC occupancy-based schedule are run considering the same boundary conditions (e.g. weather data) in order to make them comparable. The estimated energy saving obtained refers to the electricity use for HVAC (space heating, space cooling, ventilation and air treatments) during an occupied period of four months. Fig. 34(a) and Fig. 34(b) show the estimated energy saving for the whole office part of the building, and for each thermal zones. The results show that the HVAC scheduling improvement leads to a monthly energy saving that range from 12.2% to 15.4% while the average energy saving for the whole analysed period (4 months) amounts to 14%.

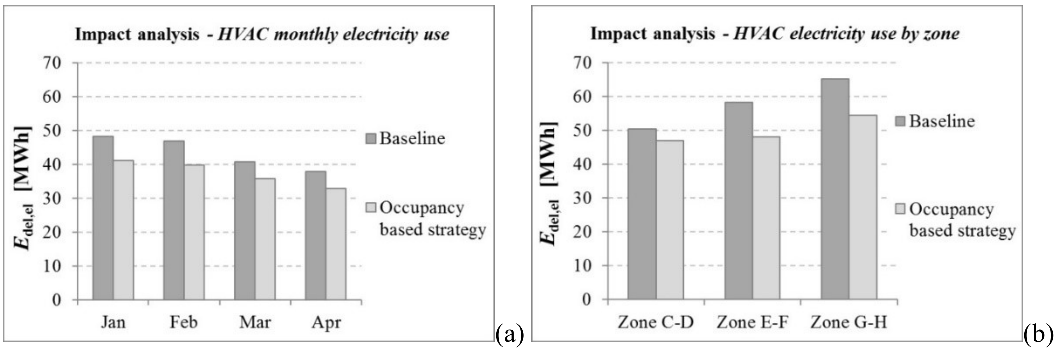


Fig. 34 - Comparison between the electricity use for HVAC before (baseline) and after the implementation of the strategy, on a monthly basis for the whole office part (a), and considering the office zone for the whole analysed period (b) [14]

3.2.6 Discussion

The developed EIS tool aims at performing an automatic recognition and classification of building occupancy patterns for the improvement of HVAC scheduling. In the developed data analytics process, a cluster analysis and a decision tree are coupled in a complementary way, without overlapping, in terms of extracted knowledge. In fact, while the k-means allows the analyst to discover hidden occupancy profiles from actual data, the decision tree provides a unique set of IF-THEN rules used to classify them. The high interpretability of the results obtained allows the final user to easily identify groups of weekdays with common occupancy profiles, for which the rescheduling of the HVAC system should be performed. The novelty of the proposed EIS tool consists in its ability to consider multiple occupancy patterns at the same time in the thermal zone reconfiguration process HVAC scheduling improvement. The advantage is the possibility to exploit in the analysis a variable presence behaviour in occupying the building over time [58]. This characterisation increases the number of constraints that need to be taken into account, but also offers the opportunity of achieving greater energy savings. Another advantage is that the developed EIS tool is capable of effectively handling occupancy aggregated data. In this way, it is possible to overcome privacy issues for individual occupants, whose specific habits are not deducible from the analysis.

As far as the analysed case study is concerned, the starting time of the HVAC system is considered already optimised, with respect to the expected arrival time of the occupants. For this reason, the rescheduling is performed exclusively on the shutdown time of the system in each thermal zone for each reference period. The final schedule computed by the EIS tool could determine a potential reduction of the electricity use for HVAC (space heating, space cooling, ventilation and air treatments) during the monitored periods of about 14%. Moreover, in order to assess the impact of the reconfiguration process, according to [197,198], two different strategies are compared:

- *strategy 1* – HVAC occupancy-based schedule obtained without performing any occupant displacement (*step 1* of the reconfiguration process)
- *strategy 2* – HVAC occupancy-based schedule obtained through the constrained occupant displacement (*step 6* of the reconfiguration process).

The results obtained through a forward simulation model show that the implementation of the *strategy 1* can reduce the electricity use for HVAC of about 10% (about 17 MWh), while the implementation of *strategy 2* can generate a further reduction of 4.2% (7.3 MWh) at the whole building level for the monitored period during the heating season (from January 2016 to April 2016).

Such figures demonstrate the powerfulness of data analytics based EIS in improving daily energy management of buildings. The developed EIS tool proved to be effective in drawing low-cost real-life management solutions, being capable to handling both multiple occupancy patterns and physical constraints in the occupant displacement process (e.g. sub-zone design capacity). In this perspective, the EIS tool shows enough flexibility for including in the future further constraints to the occupant displacement process such as the similarity of the employees' tasks, the presence of specific working groups or of special workplaces.

Prima facie, EIS tools capable to exploit occupancy data offers very interesting opportunities to understand and manage the presence of occupants in buildings leading to low/no cost and capital-intensive energy saving measures. On the other hand, obtain occupancy data with high quality and resolution, could be complex and expensive. Low-quality data could be responsible of non-robust and wrong reasoning from the final user, de facto, erasing the potential positive effect of EIS implementation. For this reason, domain expertise in building/energy applications always represents a cornerstone for supervising the knowledge extraction from data and ensuring that the information acquired is credible enough for being considered as actionable.

3.3 Development of an EIS tool for the automatic detection of anomalous energy trends at whole building level

The following sections describe and discuss the development of a real-time analytics based EIS tool capable to perform an automatic anomalous trend detection in building energy consumption time series. To this purpose the EIS tool should have two main functionalities that can be summarized as follows:

- Robust Identification of building typical energy consumption patterns over time;
- Exploitation of the typical pattern knowledge for the detection of anomalous trends.

The main objective is then to conceive a methodological framework of analysis that allows the final user to gain insights into energy consumption time series at whole building level and enables the identification of incorrect energy management procedures that are responsible of energy wasting during operation.

The methodology exploits time series analytics techniques and automatic rule extraction methods (decision trees) for developing a load profiling framework that can be embedded in a EIS.

As discussed in section 2 load profiling is an application field of data analytics in building energy management that is aimed at providing information on the actual energy use pattern at system, whole building and portfolio level. Besides this, it can help building managers to effectively investigate and characterise the building energy behaviour among different load conditions such as winter and summer season, working and not-working day, peak and off-peak hours.

Differently from the HVAC schedule optimiser tool, previously presented in section 3.2, the developed anomaly detector provides to the user a number of feedbacks during the day, leveraging on advanced data visualizations and highly interpretable results (in forms of IF-THEN inference rules) for systematically support the exploitation of the knowledge extracted. In this way it is possible to identify poor performance and quickly alarm or suggest solutions.

The next section presents the main research challenges related to the detection of anomalous trend in building energy consumption time series and introduces the motivations and novelty of the proposed methodological approach.

3.3.1 Motivation and novelty of the proposed approach

In the analysis of building related data, the characterisation of the building load profiles plays a key role to fully understand the building energy usage patterns (both normal and anomalous patterns).

In this field of application, different questions concerning temporal energy pattern characterisation and anomaly detection need to be answered, such as:

- *How can crucial information from various time series be extracted to characterise energy consumption and to identify saving opportunities in buildings?*
- *Can a tailored methodological process for energy pattern discovery also be flexible as far as the building typology, the detail of data, the set of explanatory variables and the sampling data frequency are concerned?*
- *How can an anomaly detection EIS tool, based on the predictive modeling of building energy consumption, which is easy to use, and has few explanatory variables, be developed?*
- *How does the presence of thermal sensitive loads influence the structure of the procedure proposed to characterise electrical energy use patterns?*

In order to contribute to answering these questions, in the following is proposed and discussed the development of a novel EIS tool. Such tool uses meter-level data to characterise energy consumption at whole building level and to detect anomalous energy patterns in quasi real time in order to reduce energy waste and operating costs. The EIS tool is conceived to be general for different types of buildings and is tested for two different case studies which differ in volume, building end use, data sampling frequency, set of explanatory variables and heating/cooling system configuration.

The analysed datasets refer to the overall building electrical demand of two public buildings (i.e., a university campus and a town hall) and are gathered from actual energy management systems. In order to limit the amount of data to be handled, attention is focused on the selection of a suitable data size reduction technique. In detail an enhanced SAX representation (discussed in section 2.1.1.2) is employed in the proposed methodology to address the aforementioned issue while increasing the computational efficiency [28].

SAX (discussed in section 2.1.1.1) is one of most promising techniques suitable to reduce the size of a time series, preserving key information. It is based on the reduction of the time series through a piecewise technique and on its transformation into a symbolic string. Given its remarkable flexibility and faster computation, the SAX method is widely used in the energy and building sector as a pre-processing step (data reduction and transformation), or to rapidly characterise building operation energy patterns. In the developed methodology, and adaptive SAX transformation is coupled with a regression tree model (discussed in section 2.1.1.2 and 2.1.3.1.1.1) in an innovative way in order to optimise the reduction of the time series assuming aggregation intervals of unequal length and minimise as possible the transformation error of the original time series. Moreover, differently from the existing literature, after the encoding of the time series in symbols, motif and discord recognition is performed at the aggregation interval level by developing predictive classification models (discussed in section 2.1.3.1.1.1) for each time window.

The results obtained for the two case studies demonstrated that the developed classifiers can predict the typical patterns of building energy consumption during each time window of the day with an accuracy well over 80%. As a result of the high the accuracy of the classifiers (final nodes with very high occurrence probability of a certain energy consumption pattern), it is possible to achive a strong anomaly detection capability of the EIS tool when the classification rules are violated during building operation.

Furthermore, a preliminary high-level energy diagnosis step is also included in the framework of analysis using additional electrical energy consumption datasets related to the building heating and cooling needs. The performed diagnosis is considered as preliminary, because it considers as potential cause of the anomalies detected at whole building the incorrect operation of the HVAC system that in both buildings is responsible of the most impacting energy demand. After a validation phase, the process has been also implemented on a virtual server of one of the two case study considered (i.e., Politecnico di Torino campus) for working on-line. This EIS tool can be easily translated in a set of interpretable rules helping campus managers in the early detection of anomalous energy patterns and preliminary diagnosis of their most probable associated causes.

The subsequent sections are organised as follows. Section 3.3.2 describes the case studies used for developing the EIS tool for detecting anomalous energy trends in building energy consumption. Section 3.3.3 describes the methodological framework adopted for conducting the analysis. Section 3.3.4 presents the results obtained for the selected case studies. Eventually section 3.3.5 discusses the results and contains the concluding remarks related to this specific EIS application.

As a remark part of the content of the section 3.3 was published as scientific article in the Elsevier journal “Energy” [2].

3.3.2 Case studies used for developing the EIS tool at whole building level

The methodology is tested on two case studies in order to demonstrate the flexibility and adaptability of the conceived EIS tool in real building applications. The selected case studies are substantially different from each other in terms of location of the building, building typology, type of equipment, operating schedules, monitored variables and data sampling frequency.

The first application (Case study 1) refers to the overall electrical energy consumption of a town hall located in Spain (with hourly sampling frequency of data), whereas the second application (Case study 2) is related to the total electrical energy consumption of a part of the university campus of the Politecnico di Torino, Italy (with 15-min sampling frequency of data). Fig. 35 reports the carpet plots of the one-year electrical average power demand for hourly and 15-min time intervals for case study 1 and case study 2, respectively. The carpet plot is a visualisation technique that depicts numerical values (i.e., hourly or sub-hourly electrical demands) using a colour palette and assuming a dimensional filling grid (time of the day vs. day of the year).

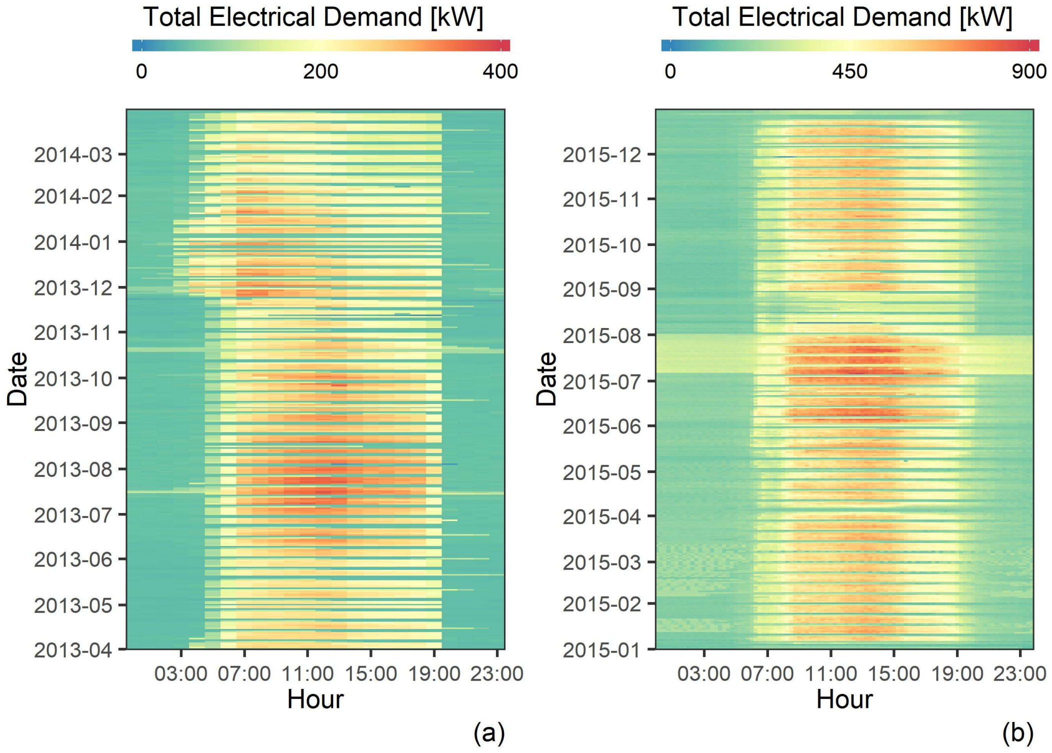


Fig. 35 - Carpet plot visualisation of the total electrical demand for Case study 1 (a) and Case study 2 (b) [2]

3.3.2.1 Case study 1 – Sant Cougat town hall

The first case study refers to the town hall of Sant Cugat del Vallés. Sant Cugat del Vallés is located in north-east Spain and is characterised by Mediterranean weather conditions. The considered building, which was built in 2007, is a six-storey glazed building with an overall floor area of about 8,600 m^2. Two of the six-storeys are underground and are used for parking, as well as for housing technical equipment and archives, whereas the other four storeys are used for public activities, that is, as offices, meeting rooms and changing rooms. The building envelope is characterised by a flat roof with low-performance skylights (thermal transmittance U = 5.70 W m^{-2} K^{-1}) and by a vertical transparent envelope composed of steel frame and double-glazed windows filled with air (thermal transmittance U = 2.70 – 3.00 W m^{-2} K^{-1}). External solar shading devices are installed, and the natural lighting is controlled manually through internal curtains. The building is divided into twenty-eight thermal zones, and the indoor thermal comfort and air quality are met with a multi-zone all-air system with partial recirculation. Each thermal zone is equipped with individual Air Handling Units (AHU), which are connected to the centralised system. The centralised system is composed of two heat/cold storage systems, coupled to an air-to-water heat pump and electrical chiller. The storage capacity for hot water is 2500 litres, whereas it is 2000 litres for chilled water.

The heat pump is characterised by a rated heating and cooling capacity of 503 kW and 570 kW, respectively. Since the heat pump alone is unable to meet the cooling needs, an electrical chiller, with a rated cooling capacity of 150 kW, has also been installed. The building is equipped with a BAS that allows the energy manager to control and manage operation of the HVAC system. The data used in this analysis pertain to the April 2013 to March 2014 period, have an hourly timestamp and include: the total building electrical load, the electrical load of the heat pump, chiller and circulation auxiliaries, the external air temperature and the internal air temperature in one of the most representative zones. Table 4 summarises the variables collected for this case study.

Table 4 - Case study 1 - Summary of the variables [2]

Variable	Description	Type	Unit of measure
Day	Day of the week	Categorical	[-]
E_total	Total electrical demand of the building	Numerical	[kW]
E_H/C system	Electrical demand of the Heating/Cooling system	Numerical	[kW]
T_external	External air temperature	Numerical	[°C]
T_internal	Internal air temperature	Numerical	[°C]

Fig. 35 (a) shows that the most significant variations of the electrical load mainly occur in the time interval between 05:00 a.m. – 08:00 p.m., albeit with different trends over the year. By analysing the total electrical demand, one can infer that the HVAC system is usually turned on at 5:00 a.m., except for a limited period of the year during which the system is turned on at 3 a.m., and it is switched off at 8:00 p.m. The average incidence of the HVAC system on the total electrical demand is 35% for the analysed period. When the building is unoccupied, and the plug-loads are switched off, the base load of the building is about 50 kW. Moreover, it is possible to note that the electrical demand peak during the winter season occurs in the early morning, whereas it occurs in the middle of the day during the summer and autumn.

3.3.2.2 Case study 2 – the Politecnico di Torino campus

The second case study refers to a part of the Politecnico di Torino campus, that is served by the same medium-voltage transformer room. The overall floor area, which is over 20.000 m^2, includes several facilities. The area is divided into central administration offices, which host more than 300 employees, and academic spaces, which include more than 20 lecture halls and 4 information technology labs. Moreover, a bar and a large canteen are located in the public spaces, and their yearly electricity consumption accounts for 17% of the total consumption. One of the building’s data-centres, whose consumption represents 14% of the analysed part of the campus, is located underground. The heating and cooling system is composed of two different circuits, which are used to produce hot and chilled water. The heating circuit is served by a heat exchanger that is connected to the district heating system, while the cooling circuit is instead served by a closed-loop geothermal plant composed of two chillers and one water-to-water heat pump. The chillers and the heat pump are connected in parallel and have a total rated cooling capacity of 1120 kW and 590 kW, respectively.

The overall yearly consumption of these systems, including circulation auxiliaries for both the heating and cooling circuits, accounts for 15% of the total consumption. Data pertaining to the year 2015, with a time stamp of 15 minutes, were analysed for this case.

Table 5 - Case study 2 - Summary of the variables [2]

Variable	Description	Type	Unit of measure
Day	Day of the week	Categorical	[-]
E_total	Total electrical demand of the building	Numerical	[kW]
E_H/C system	Electrical demand of the Heating/Cooling system	Numerical	[kW]
T_external	External air temperature	Numerical	[°C]
T_internal	Internal air temperature	Numerical	[°C]
Occupancy	Number of occupants in the central administration offices of the Politecnico di Torino	Numerical	[-]

Table 5 summarises the variables considered for this case study. from Fig. 35(b) it can be inferred, that the building energy systems are usually turned on at 6.00 a.m., a period in which the building begins to be occupied, and are switched off at 19.00 p.m. The total electrical demand increases over the time interval between 9 a.m. (the time at which teaching and office activities begin) until 4 p.m., with the greatest electrical demand in the middle of the day, since all the activities, including the canteen activities, take place at this time. During the summer period, there is a higher electrical demand in the afternoon hours than in the other periods of the year. This is due to a higher operation of the cooling system to meet the cooling needs.

3.3.3 Implemented methodology for the automatic detection of anomalous trends in building energy consumption time series

The methodology is based on the application of an enhanced SAX transformation, coupled with classification and regression trees, in order to perform an advanced energy consumption characterisation and an anomalous trend detection analysis. The methodology process is performed through a multistep data analytics procedure. The whole process is tested on the total electrical energy consumption data of two buildings which have different end uses. One-year, hourly/sub-hourly electrical energy consumption data are available for each building, together with other influencing variables (e.g., climatic, occupancy data). The general framework unfolds over several different stages, as shown in Fig. 36.

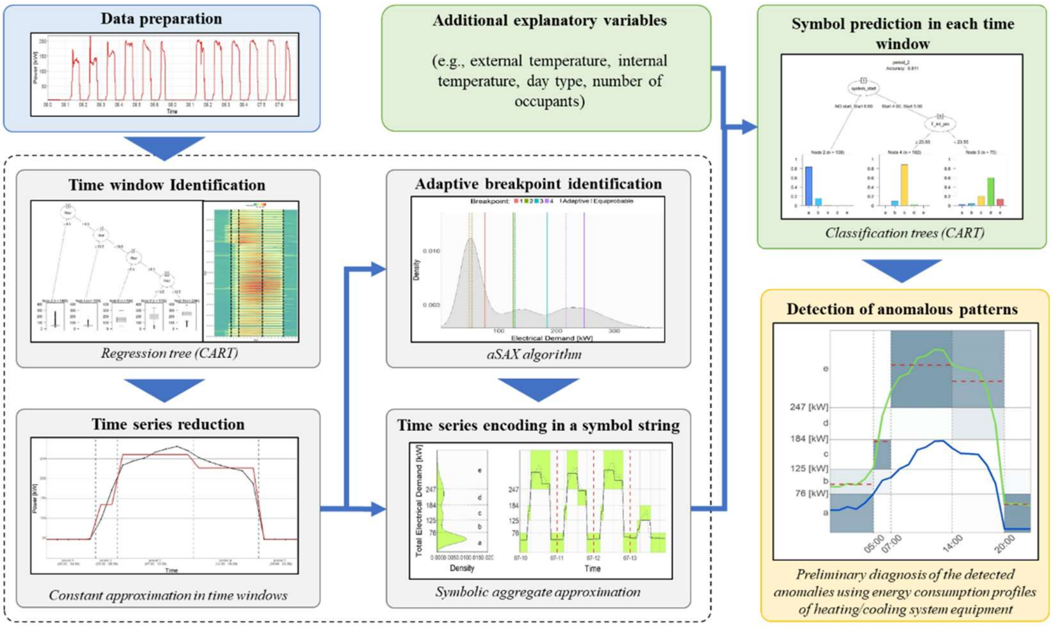

Data preparation
Additional explanatory variables
(e.g., external temperature, internal temperature, day type, number of occupants)
Symbol prediction in each time window
Classification trees (CART)
Time window Identification
Regression tree (CART)
Adaptive breakpoint identification
aSAX algorithm
Time series reduction
Constant approximation in time windows
Time series encoding in a symbol string
Symbolic aggregate approximation
Detection of anomalous patterns
247 [kW]
184 [kW]
125 [kW]
76 [kW]
05:00
07:00
14:00
20:00
Preliminary diagnosis of the detected anomalies using energy consumption profiles of heating/cooling system equipment

Fig. 36 - Framework for advanced characterisation of building energy consumption time series and anomalous trend detection (adapted from [2])

The first stage (Fig. 36) is aimed at data preparation. Data pre-processing is a crucial task to prepare the time series for the analysis. At this stage, the energy consumption time series are analysed in order to identify any missing value and/or punctual outlier to be removed and replaced. The second stage of the analysis is aimed at transforming the energy consumption time series by implementing an enhanced SAX process. In detail, two preliminary hypotheses are formulated in different way from the classic SAX implementation presented in Section 1. The first hypothesis is related to the length of the non-overlapping W windows on the time axis. In the literature, the time windows are generally assumed to have the same length [30,31,89]. However, this hypothesis could cause a significant loss of information in many applications, in terms of approximation error of the original time series. For example, when analysing building energy consumption data, the symbolic sub-strings have a daily length (i.e., $T = 24$ hours), which constrains the length that each time window can assume if the hypothesis of equal-length is satisfied. In the case of an hourly-based time series, if T has a length of 24 hours, the time windows can have sizes of 2, 3, 4, 6, 8 or 12 hours, which results in equally sized time windows, each with duration T/W.

To overcome this limitation, unequal time window lengths can be identified in order to approximate the original time series in a much better way than when the equal ones are used [210]. This time series approximation, called Adaptive Piecewise Constant Approximation (APCA), is based on the same principle as PAA, but it offers the advantage of being able to conduct a finer aggregation of areas of the time series where the amplitude variation of the variable is higher than in the areas with low amplitude variation over time [210].

In the proposed approach, the evaluation of time windows is conducted using a Regression Tree algorithm (described in section 2.1.3.1.1.1) [204]. The regression tree is used to optimise the size and the number of time windows through a cost-complexity process, searching for a trade-off between the approximation error and the number of time windows. The regression tree is developed using the hourly/sub-hourly electric power demand as the numerical target attribute, and the time of the day as the ordinal explanatory attribute. The time attribute is set as an ordinal variable to identify non-overlapping time windows that are then used to segment the daily reference period. This choice is important, because if the time variable is set as a categorical one (i.e., without preserving the order of the possible values in the splits of the regression tree), the identification of non-overlapping time windows would not be ensured. In such a case, the regression tree could indiscriminately group together some electric power demand values pertaining to the early morning with those referring to the night.

In addition, in the developed methodology, the second hypothesis, which is formulated in a different way from that of the classic implementation of SAX, rejects the equal probability of the symbols in order to encode each approximated constant segment on the vertical axis. This difference is specifically introduced by considering the nature of the energy consumption data. In fact, in some applications, where a data transformation and reduction of the time series is needed, the evaluation of equal probability regions may not be the best choice, in terms of approximation quality. For example, the operation of a chiller in a building depends exclusively on the occupancy rate during the day. Therefore, assuming a 10-hour operation per day (08:00 – 18:00) during working days, the chiller is turned off most of the time (14 hours per day). In that case, the breakpoints evaluated by considering equal probability symbol regions could produce very narrow intervals in correspondence to low electrical power values and wide intervals for high values (considering that the most frequent electrical power values during the day are close to zero). This could result in losing key information when the constant approximated segments are encoded in symbols.

In the proposed methodology, without any standardisation of the original data, and rejecting the equal probability hypothesis, the aSAX (Adaptive Symbolic Aggregate ApproXimation [32]) algorithm is adopted for the evaluation of breakpoints. The aforementioned methods (i.e., Regression Tree and aSAX) used in the methodological process to reduce and transform the data are briefly introduced and discussed in Section 2.1.1.2 and 2.1.3.1.1.1.

After the data transformation, the entire time series encoded in a unique string of symbols is chunked into *N* sub-strings of a daily length (i.e., $T = 24$ hours) in order to obtain constant time-scale based sequences. The *N* symbolic sub-strings are made up of a certain number, *W,* of time windows encoded in alphabetic symbols and organised in an NxW matrix. In this way, each daily load profile is represented by a SAX word that is then used as the input for the successive anomaly detection analysis. At this stage, the probability of each symbol occurring in each time window, under specific boundary conditions, is evaluated by means of a classification tree, which is based on additional explanatory variables (e.g., external temperature, internal temperature, day type, month). In this way, if the occurrence probability estimated with the classification tree and associated with a symbol is very low, it is likely that the energy consumption in the corresponding sub-daily time window is abnormal.

Furthermore, the post-mining stage of the analysis is performed using additional datasets for the two case studies in order to further support the preliminary diagnosis of detected anomalous patterns at whole building level.

In this perspective, the developed EIS tool based on the proposed methodology can be effectively used to support the implementation of advanced targeted anomaly diagnosis in a specific time window of the entire time domain. The EIS tool can easily and smoothly be generalised for different kinds of building types and climates.

3.3.4 Results obtained from load profile characterization analysis

3.3.4.1 Application of the customised SAX process

In order to perform an advanced characterisation of the total energy consumption for the two case studies, a reduction and transformation of the energy consumption time series are carried out using the proposed transformation process.

The time windows of the daily load profiles are evaluated using a regression tree. As described in Section 3.3.3, a regression tree allows unequal time window lengths to be identified. The regression tree is developed using the total electrical load as the numeric target variable and the time of the day as an ordinal predictive attribute. In order to identify the optimal number and size of the time windows, the regression tree is subjected to a cost-complexity pruning process. Assuming the time as an ordinal variable, any non-overlapping time windows with homogeneous electricity consumption values are identified. In order to preserve the accuracy of the model in the leaf nodes during the operation hours of the building systems (e.g., from 07:00 to 19:00), only working days are taken into account, and days with a low standard deviation of the electricity demand (e.g., Sundays, holidays) are excluded. In fact, the tree splitting process 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 (e.g., the minimum number of cases in the parent and child nodes, maximum tree depth, minimum reduction in node variance after splitting). For the analysed case studies, the selected stopping criterion is based on the minimum number of objects in a child node in order to identify time windows with a length of at least two hours (i.e., $W_{min.\ length}$120 min.), as follows (Eq. 15):

$$Obj_{min} = \left(M_{days,tot} - M_{days,excluded}\right) * \frac{W_{min.\ length}}{timestep}$$

Eq. 15

where Obj_{min} the minimum number of objects in a child node, $M_{days,tot}$ is the total number of daily load profiles, $M_{days,excluded}$ is the number of daily load profiles excluded from the dataset (e.g., Sundays, holidays), $W_{min.\ lengt}$ the minimum length of the time window (expressed in minutes) and *timestep* (expressed in minutes) is the measurement sampling frequency. For example, if a measurement campaign of one year (with 250 working days and the other days being excluded), a time window length of 120 minutes and a sampling frequency of 15 minutes are assumed, Obj_{min} equal to 2000 objects.

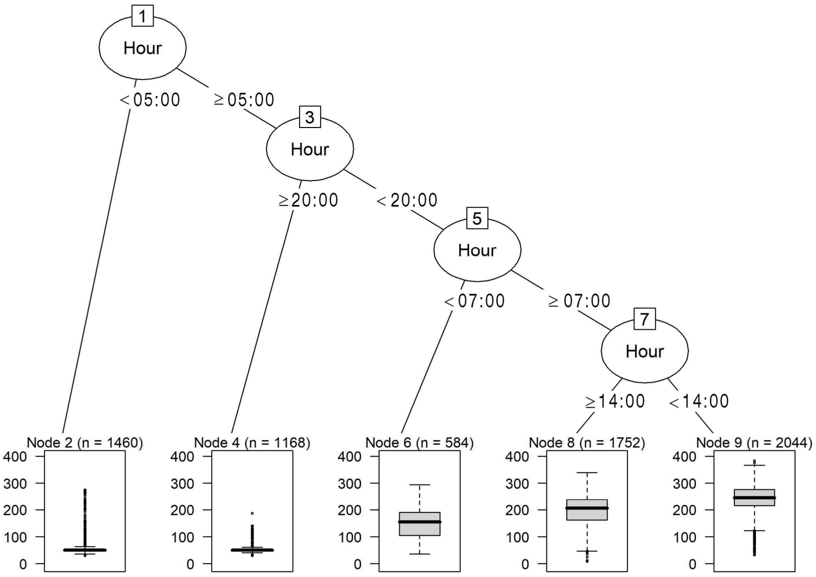


Fig. 37 - Identification of sub-daily time windows by means of the CART algorithm for Case study 1 [2]

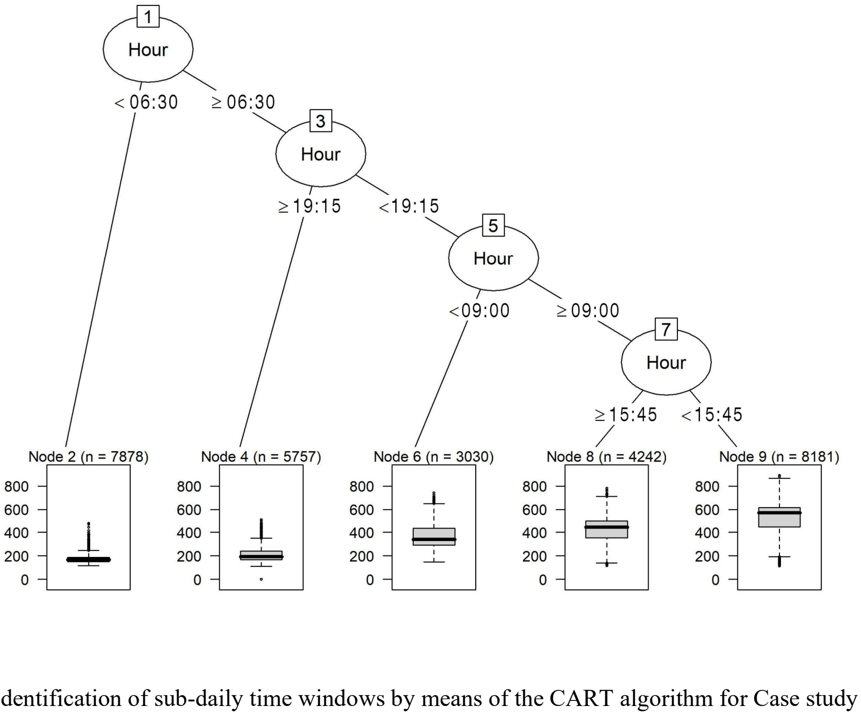


Fig. 38 - Identification of sub-daily time windows by means of the CART algorithm for Case study 2 [2]

Fig. 37 and Fig. 38 report the outputs of the regression model for both case studies in the form of decision trees. The cost-complexity pruning process was performed for both case studies.

The pruning procedure of the regression tree is repeated iteratively, and smaller and smaller subtrees are found until the root node is reached. At the end of the iterations, the final pruned tree can be evaluated by plotting the relative errors of the subtrees versus their complexity parameters (cp). This kind of plot usually shows an initial sharp drop, followed by a relatively flat region (Fig. 39).

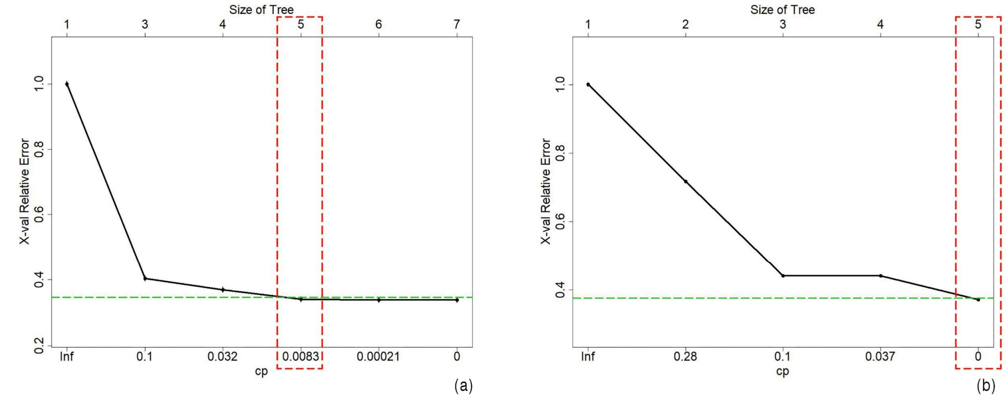


Fig. 39 - Optimal size of the regression tree for Case study 1 (a) and Case study 2 (b) [2]

When the decision tree is subject to a validation procedure (e.g., k-fold cross-validation), it is also possible to compute a standard error for each relative error of the sub-tree. The choice of the best subtree starts from the flat region of the subtree errors that includes the minimum cross validated error that is achieved. In fact, the values falling within one standard error of the achieved minimum risk (i.e., 1-SE rule) identify statistically equivalent sub-trees [204]. The simplest model (with the minimum number of final nodes) of all the identified sub-trees in the flat region is then chosen.

As shown in Fig. 39, both of the decision trees have five final nodes, which are determined when the relative error goes below the user-defined threshold (green dashed line), and which correspond to five time windows of different lengths.

Table 6 reports the obtained time windows with reference to their durations. For both buildings, period 1 and period 5 are related to the night hours during which the buildings are unoccupied, while periods 2, 3 and 4 are representative of the operation hours of the building systems. It is possible to note that the lengths of the evaluated time windows are significantly different from each other, thus highlighting the importance of assuming time windows of unequal lengths to achieve an optimal time series reduction.

Table 6 - Sub-daily time windows for case study 1 and case study 2 [2]

Case study	**Time windows**				
1) Sant Cougat town hall	**Period 1**	**Period 2**	**Period 3**	**Period 4**	**Period 5**
	00:00 – 04:59 5 hours	05:00 – 06:59 2 hours	07:00 – 13:59 7 hours	14:00 – 19:59 6 hours	20:00 – 23:59 4 hours
2) Politecnico di Torino	**Period 1**	**Period 2**	**Period 3**	**Period 4**	**Period 5**
	00:00 – 06:29 6 hours and 30 minutes	06:30 – 08:59 2 hours and 30 minutes	09:00 – 15:44 6 hours and 45 minutes	15:45 – 19:14 3 hours and 30 minutes	19:15 – 23:59 4 hours and 45 minutes

Period 2 has the shortest duration (about 2 hours) for both case studies. In fact, period 2 represents the ramp-up of the daily load profile during which the heating/cooling systems are usually turned on and the employees start to occupy the building. The regression tree isolates this period in a specific time window, which is characterised by a high load variation over a very short time, thus reducing the global constant approximation error to a great extent. After the identification of the time windows, the entire time series is reduced through a constant approximation process by replacing the electrical demand values that fall into the same time window with the relative mean value. The time windows identified through the regression trees are also assumed for the previously excluded daily load profiles. It is in fact verified that the days that showed a limited variation of the electrical demand over time (e.g., flat profiles of Sundays and holidays) are not influenced to any great extent by the windowing process, in terms of the constant approximation error of the time series.

In the successive step, the constant approximated segments are encoded in symbols. For this purpose, the aSAX algorithm is implemented, as discussed in section 3.3.3. Alphabet size *A*, which corresponds to the number of symbols, is assumed equal to the number of the previously identified sub-daily time windows. Fig. 40 shows the frequency histogram of the electrical demand time series reduced through the constant approximation process.

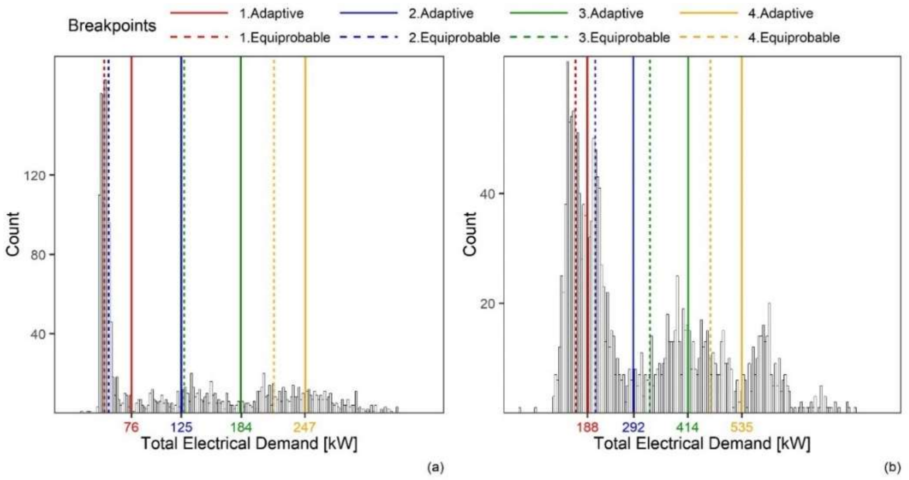


Fig. 40 - Identification of Adaptive breakpoints through the aSAX algorithm for case study 1 (a) and case study 2 (b) [2]

The hypothesis of Gaussian distribution of the constant approximated time series is rejected for both case studies. As explained in section 3.3.3, the equal probability hypothesis for the breakpoint evaluation is used for the initialisation of the aSAX algorithm. The initial position of the breakpoints (dotted lines in Fig. 40) produces very narrow intervals in correspondence to low values of electrical demand and wide intervals for high values. The final adaptive breakpoint positions evaluated after about 20 iteration steps of the aSAX algorithm (continuous lines in Fig. 40) make it possible to identify balanced regions that minimise the representation error resulting from the encoding of the symbols. Table 7 reports the breakpoints of each symbol for both case studies.

Table 7 – Breakpoints of each symbol for case study 1 and case study 2 [2]

Case study	**Symbol**				
1) Sant Cougat town hall	**a** 0 kW - 76 kW	**b** 76 kW - 125 kW	**c** 125 kW - 184 kW	**d** 184 kW - 247 kW	**e** > 247 kW
2) Politecnico di Torino	**a** 0 kW - 188 kW	**b** 188 kW - 292 kW	**c** 292 kW - 414 kW	**d** 414 kW - 535 kW	**e** > 535 kW

Fig. 41 reports the output of the customised SAX process for a sequence of twenty consecutive time windows for case study 1. The left-hand side of the figure also shows the frequency histogram of the constant approximated segments of the reduced time series and the breakpoints evaluated through the aSAX algorithm.

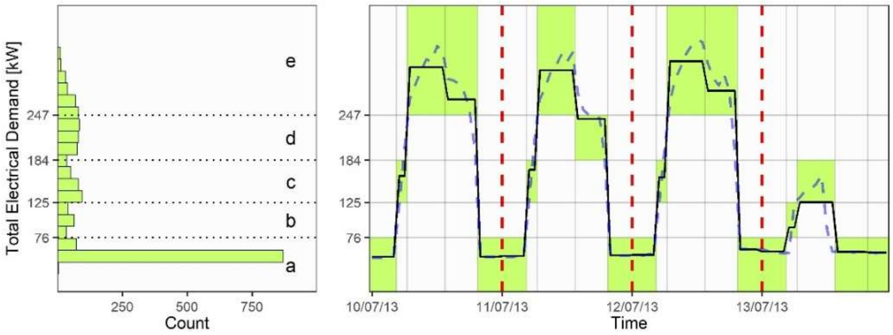


Fig. 41 - Symbolic transformation for a sequence of twenty time windows (i.e., four days) for case study 1 [2]

The successive step consists in chunking the entire transformed time series into a set of sub-strings. As previously discussed, the reference sub-string has a daily length, and it is found to be composed of five consecutive non-overlapping time windows for the analysed case studies.

After the chunking of the time series, the transformed daily load profiles are organised in an NxW matrix, where N are the daily sub-strings (i.e., N = 365 days), and W is the number of identified time windows (i.e., W = 5 time windows for both case studies).

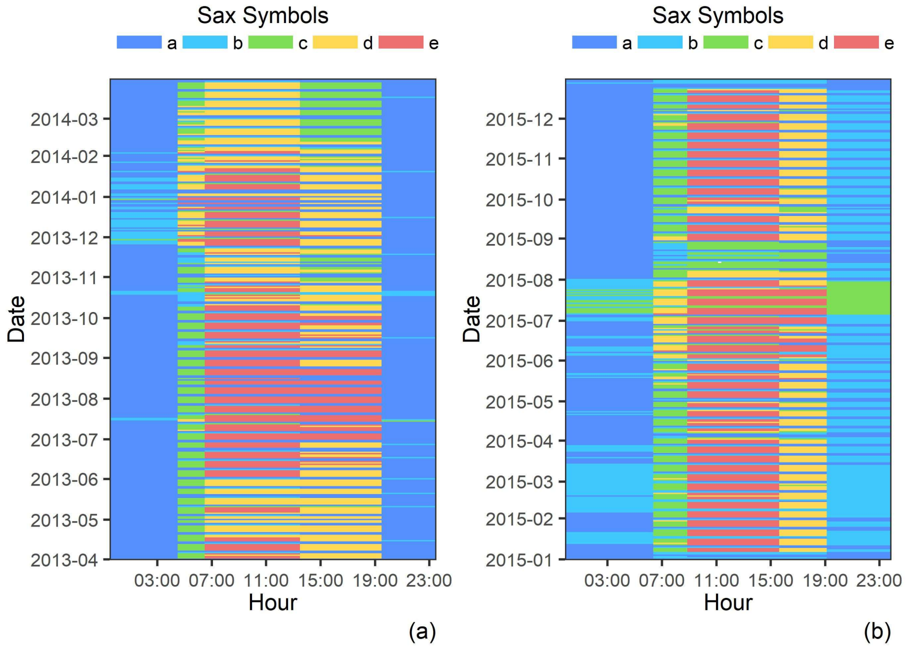


Fig. 42 - Carpet plot visualization of the SAX symbols for Case study 1 (a) and Case study 2 (b) [2]

Fig. 42 (a) and (b) report the carpet plots of the NxW matrix for case study 1 and case study 2, respectively. Through this effective visualisation, it is possible to quickly understand how the symbols are distributed among the time windows in the daily load sub-strings.

The occurrence frequency of each symbol is then calculated for each time window (Fig. 43 and Fig. 44) for the whole monitored period in order to perform a preliminary characterisation of the data after their encoding in symbols. It can be seen, for case study 1, that the symbol "*a"* has an occurrence of about 90% for periods 1 and 5, thus highlighting a close correlation between the electricity consumption in these time windows and the OFF state of the HVAC system, which mostly influences the reduction in the total electrical demand during the night hours.

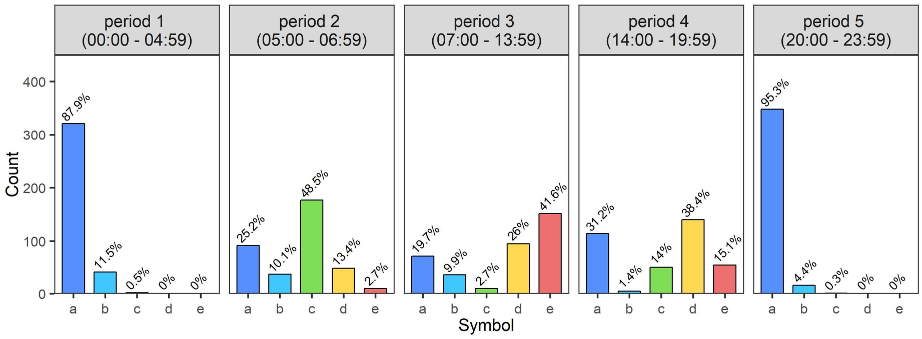


Fig. 43 - Occurrence frequency of each symbol in each time window for case study 1 [2]

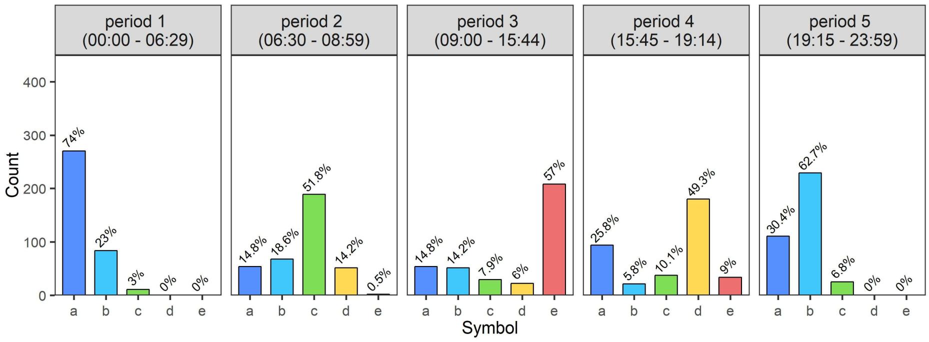


Fig. 44 - Occurrence frequency of each symbol in each time window for case study 2 [2]

The symbols do not exhibit any specific trends for other periods, thus suggesting the need to further investigate the dependencies of each symbol on the boundary conditions of influence. To this purpose, a classification tree is developed for each time window, using the SAX symbol as a categorical target variable and the additional available variables as predictive attributes.

Table 8 - Summary of the variables used for the classification process for each time window [2]

Variable	Description	Case study 1	Case study 2
Day	Day of the week	✓	✓
T_ext	Average external temperature	✓	✓
T_int	Average internal temperature	✓	
T_int_pre	Average internal temperature in the previous period	✓	
Occ	Number of occupants		✓
Sym_pre	Symbol in the previous period	✓	✓

Table 8 summarises the variables used for the characterisation of each time window.

As a reference, Fig. 45 reports the output of the classification tree developed in period 3 (07:00 – 13:59) for case study 1. For this period, it is possible to identify the boundary conditions that could explain the occurrence of each symbol with a high probability by means of a classification tree.

The building of case study 1 is equipped with a heat pump that supplies both hot and cold water according to its operating mode (heating or cooling mode). This implies that the electric consumption of the heat pump is thermal sensitive and during winter operation is inversely proportional to the outside temperature, while the opposite occurs in the summer period.

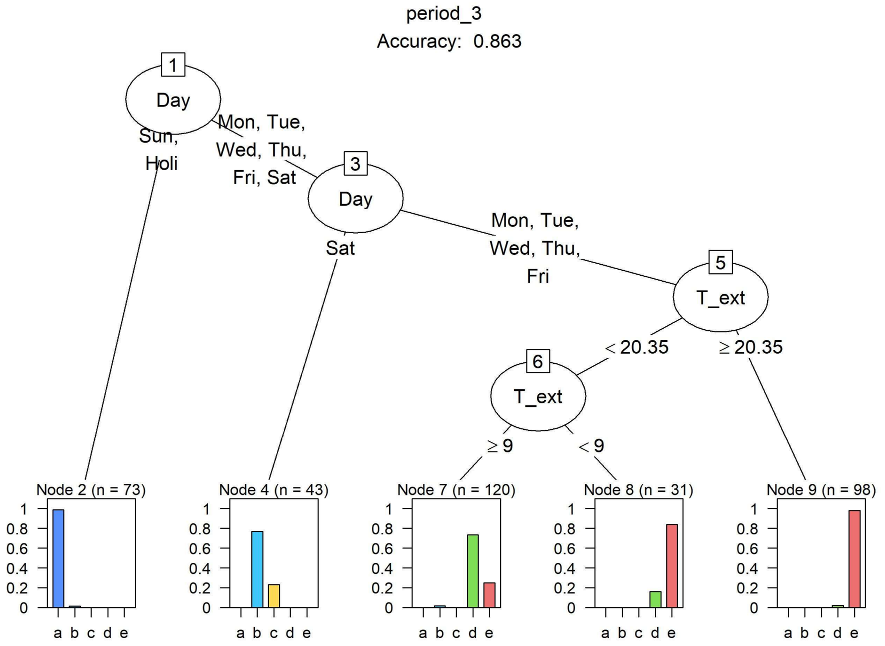


Fig. 45 - Classification tree developed for period 3 (07:00 – 13:59) for Case study 1 [2]

The classifcation tree shown in Fig. 45 is able to distinguish two characteristic power demand symbols for the working days: "*d*" and "*e*". A higher consumption (i.e., symbol “*e*”) is associated with the hot season (T_ext > 20.35°C) and the cold season (T_ext < 9°C), while a lower consumption (i.e., symbol “*d*”) is associated with the mild season (9°C < T_ext < 20.35°C). Symbol “*a*” only occurrs in period 3 during Sundays and Holidays (tree node 2), while symbol “*b*” is typical of Saturdays and had an occurrence probability of about 77% (tree node 4). At the same time, these decision rules extracted from the classification tree developed in period 3 also make it possible to identify a very low occurrence probability associated with a symbol, given certain boundary conditions. For example, the symbol “*b*” has an occurrence probability lower than 1% during Sundays and Holidays (tree node 2). Therefore, the symbol “*b*”, due to its low occurrence during these days, can be considered as a discord candidate that needs to be further investigated in the diagnostic phase. In fact, the higher the accuracy of the decision rules is (final nodes with very high occurrence probability of a certain symbol), the higher the consequent anomaly detection capability when the rules are violated during building operation.

Table 9 - Decision rules for case study 1 [2]

Time window	Decision rules	Symbol	Accuracy
Period 1 (00:00 – 04:59)	IF ***system_start*** = is turned OFF	→ ***a***	*98%*
	IF ***system_start*** = is turned ON at 04:00 a.m. AND ***T_int*** ≥ 23,43 °C	→ ***a***	*80%*
	IF ***system_start*** = is turned ON at 04:00 a.m. AND ***T_int*** < 23,43 °C	→ ***b***	*79%*
Period 2 (05:00 – 06:59)	IF ***Day*** = Holiday OR Sunday OR Saturday	→ ***a***	*83%*
	IF ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday AND ***T_int_pre*** *(period 1)* ≥ 23,55 °C	→ ***c***	*88%*
	IF ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday AND ***T_int_pre*** *(period 1)* < 23,55 °C	→ ***d***	*60%*
Period 3 (07:00 – 13:59)	IF *Day* = Holiday OR Sunday	→ ***a***	*99%*
	IF *Day* = Saturday	→ ***b***	*77%*
	IF ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday AND 9 °C ≤ ***T_ext*** < 20,35 °C	→ ***d***	*73%*
	IF ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday AND ***T_ext*** ≥ 20,35 °C	→ ***e***	*98%*
	IF ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday AND ***T_ext*** < 9 °C	→ ***e***	*84%*
Period 4 (14:00 – 19:59)	IF ***Sym_pre*** = a OR b OR c	→ ***a***	*96%*
	IF ***T_ext*** < 24,1 °C AND ***Sym_pre*** *(period 3)* = *"d"* AND ***T_int*** < 25,55 °C	→ ***c***	*69%*
	IF ***T_ext*** < 24,1 °C AND ***Sym_pre*** *(period 3)* = *"d"* AND ***T_int*** ≥ 25,55 °C	→ ***d***	*75%*
	IF ***T_ext*** < 24,1 °C AND ***Sym_pre*** *(period 3)* = *"e"*	→ ***d***	*94%*
	IF ***Sym_pre*** = *"d"* OR *"e"* AND ***T_ext*** *(period 3)* ≥ 24,1 °C	→ ***e***	*79%*
Period 5 (20:00 – 23:59)	-	→ ***a***	*95%*

Table 10 - Decision rules for case study 2 [2]

Time window	Decision rules	Symbol	Accuracy
Period **1** (00:00 – 06:29)	-	→ **a**	*74%*
Period **2** (06:30 – 08:59)	IF ***0cc*** < 47 AND ***Day*** = Holiday OR Sunday	→ **a**	*81%*
	IF ***0cc*** < 47 AND ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday OR Saturday AND ***T_ext*** < 18,8 °C	→ **b**	*86%*
	IF ***0cc*** < 47 AND ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday OR Saturday AND ***T_ext*** ≥ 18,8 °C	→ **c**	*78%*
	IF ***0cc*** ≥ 47 AND ***T_ext*** < 18 °C	→ **c**	*93%*
	IF ***0cc*** ≥ 47 AND ***T_ext*** ≥ 18 °C	→ **d**	*80%*
Period **3** (09:00 – 15:44)	IF ***0cc*** < 189,5 AND ***Day*** = Holiday OR Sunday	→ **a**	*84%*
	IF ***0cc*** < 189,5 AND ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday OR Saturday AND ***T_ext*** < 22,35 °C	→ **b**	*81%*
	IF ***0cc*** < 189,5 AND ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday OR Saturday AND ***T_ext*** ≥ 22,35 °C	→ **c**	*77%*
	IF ***0cc*** ≥ 189,5	→ **e**	*93%*
Period **4** (15:45 – 19:14)	IF ***0cc*** < 94 AND ***Day*** = Holiday OR Sunday OR Saturday	→ **a**	*85%*
	IF ***0cc*** < 94 AND ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday AND ***T_ext*** < 17,3 °C	→ **b**	*90%*
	IF ***0cc*** < 94 AND ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday AND ***T_ext*** ≥ 17,3 °C	→ **c**	*100%*
	IF ***0cc*** ≥ 94 AND ***T_ext*** < 28,95 °C	→ **c**	*88%*
	IF ***0cc*** ≥ 94 AND ***T_ext*** ≥ 28,95 °C	→ **d**	*74%*
Period **5** (19:15 – 23:59)	IF ***Day*** = Holiday OR Sunday OR Saturday	→ **a**	*75%*
	IF ***Day*** = Monday OR Tuesday OR Wednesday OR Thursday OR Friday	→ **b**	*87%*

Table 9 and Table 10 report all the decision rules extracted from the classification trees developed for case study 1 and case study 2 developed for each period. It can be observed that the set of additional variables used by the classification trees differ among the periods. Furthermore, it can be noticed that period 5, pertaining to case study 1, and period 1, referring to case study 2, are not associated with any decision rule. These periods are characterised by a very high occurrence (over 70% over the whole year) of only one symbol. In these cases, the additional available variables are not able to further characterise the occurrence of other symbols.

The classification process leads to very robust results, with a global accuracy that ranges between 80% and 90%. The methodological process proves to be flexible, both with respect to the timestamp of the data (hourly or sub-hourly) and to the set of input variables employed in the classification phase. As previously discussed, a discord can be detected when a different symbol from the one estimated through the decision rules is observed in a specific time window, given certain boundary conditions.

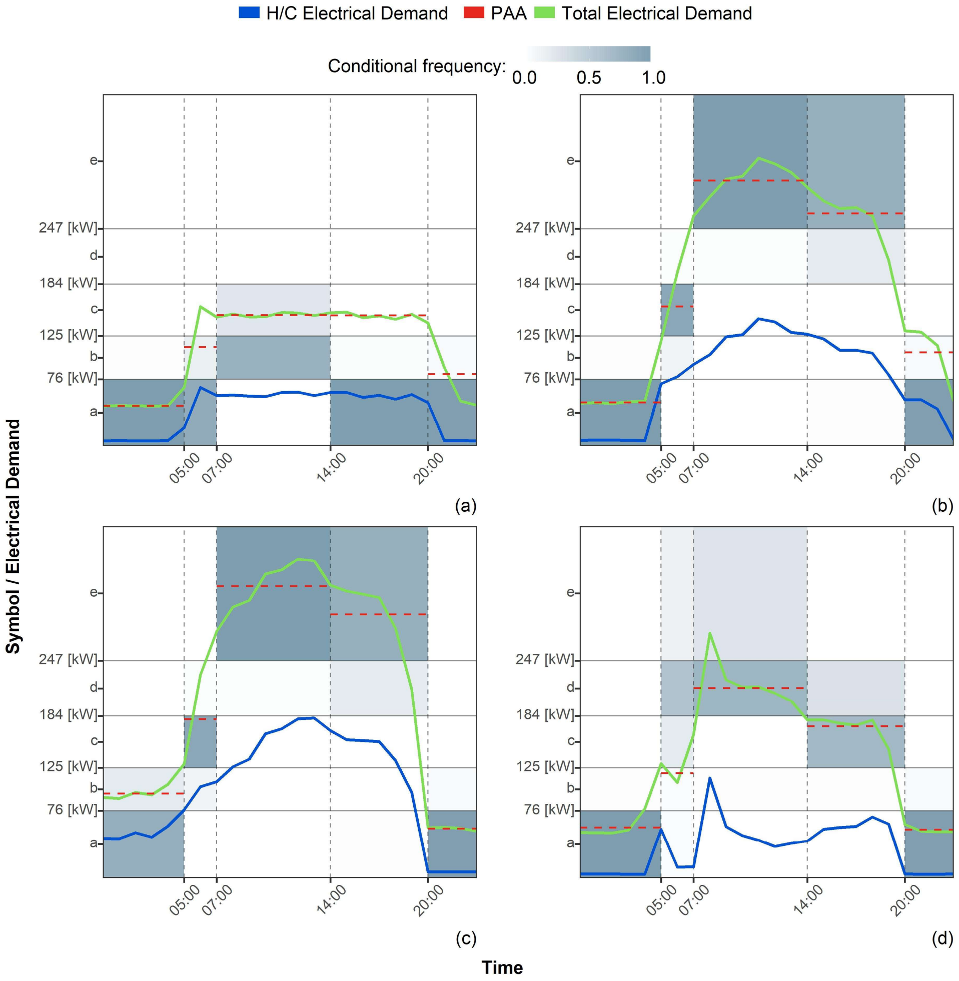


Fig. 46 - Anomalous patterns related to the building heating/cooling system operations for Case study 1 [2]

Fig. 46 and Fig. 47 show various daily sub-strings containing anomalous candidates for both case studies. In order to perform a preliminary diagnosis in which the causes of the occurrence of an infrequent symbol are searched for in a time window, the corresponding electrical demand of the heating/cooling system which influences the total building electrical demand the most, is also analysed.

Fig. 46 and Fig. 47 report the original (not reduced) daily profiles of the total building electrical demand (green line), the constant approximation of the total building electrical demand (red dashed line) and of the cooling/heating system electrical demand (blue line). The figures also report the occurrence probabilities of symbols extracted from the leaf nodes of the classification trees developed for each time window. The symbol occurrence probabilities are visualised by filling the regions of the amplitude space with different shades of grey. The higher the symbol occurrence probability is, the darker the colour of the symbol region. Thanks to this effective visualisation, the final user of the EIS tool can quickly become aware of a potential anomalous energy consumption when the mean value of the total electrical demand (PAA segment) falls into a time window filled with a lighter colour. Fig. 46 reports four representative daily sub-strings for case study 1, where infrequent values of the total electrical demand are detected for at least one time window during the day. Fig. 46 (a) shows the daily load profiles (the electrical demand for the whole building and for the heating/cooling system) for a Saturday. By means of the decision rules reported in Table 9, it is possible to assess *a-a-b-a-a* as the estimated word (made up of five consecutive symbols) related to the whole building electrical demand. The real sequence of symbols that instead emerged is *a-b-c-c-b,* which is significantly different from the expected word. Excluding the first time window of the day, where a perfect match between the real and the expected symbols is observed, the total electricity consumption of the building can be considered infrequent in time windows 2, 3, 4 and 5. In order to establish whether these detected infrequent patterns correspond to an anomalous energy management operation, the daily electrical load profile of the heating/cooling system is analysed. It can be observed that the system is not turned off at 12:30 a.m., as expected for Saturday, thus generating an over-consumption during periods 4 and 5. Moreover, the total electrical demand variation of the building during operation perfectly matches the electrical load variation of the heat pump, thus suggesting that the system is operated when no other electrical load is present in the building. For this reason, it can be inferred that the system is turned-on during the considered day although the building is unoccupied.

Fig. 46 (b) and (c) report two daily sub-strings that refer to working days in the summer season. The total electrical demand for these two days is found to be higher than that expected for periods 5 and 1, respectively. By analysing the trend of the electrical demand of the cooling system, it can be inferred that these over-consumptions are due to an incorrect operation during the night hours. The last discord reported in Fig. 46 (d) corresponds to a delayed start-up of the system after 06:00 in the morning, which results in an under-consumption during period 2.

By comparing Fig. 46 (c) and (d), it is possible to observe the flexibility of the procedure in predicting, for the ramp-up time window (period 2), different symbols between the winter and summer seasons (i.e., according to different boundary conditions).

For the sake of completeness, Fig. 47 (a) and (b) show anomalous profiles generated by an incorrect operation of the heating/cooling system during unoccupied periods for case study 2. An over consumption can be observed in Fig. 47 (a), due to the incorrect operation of the heating system during a holiday; Fig. 47 (b), instead, shows the case of an infrequent electrical demand during the night (periods 1 and 5), due to continuous operation of the chiller system throughout the day.

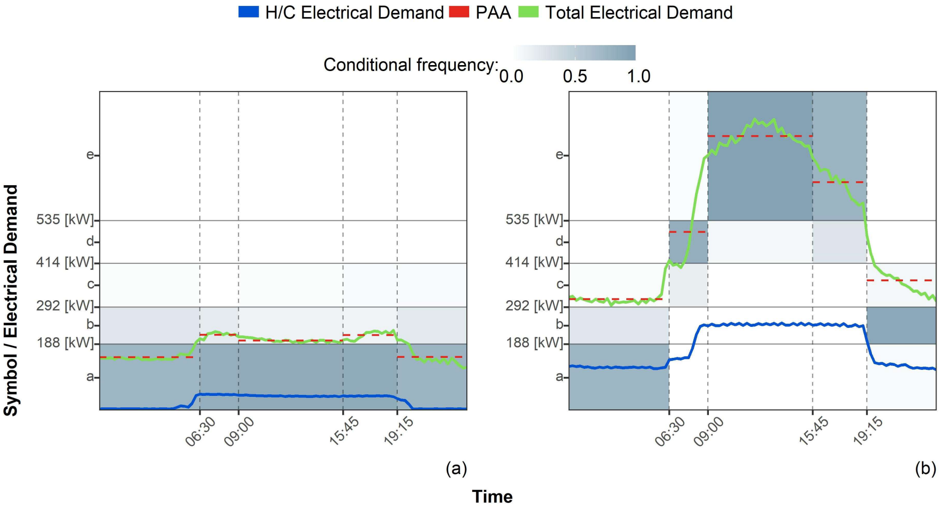


Fig. 47 - Anomalous patterns related to the building heating/cooling system operations for Case study 2 [2]

3.3.5 Discussion

The performed analysis focused on whole building level in order to demonstrate the potential of an EIS tool for detecting anomalous trends in building energy consumption time series. For developing the EIS tool, data reduction, transformation and data analytics methods are coupled in a complementary way in order to detect infrequent/unexpected patterns of whole building energy consumption at a sub-daily time scale.

Early detection of anomalous trends in energy consumption trough EIS tools can prevent the occurrence of abnormal events and reduce energy waste over time. In this context, the proposed anomaly detection EIS tool (based on the extraction of decision rules) can identify abnormal energy consumption in representative and specific time windows of the day. In order to demonstrate the implications also for a preliminary fault diagnosis, some anomalous trends of the total electrical load are examined in a post-mining phase, using additional datasets of the electrical energy consumption related to heating and cooling needs.

Considering that SAX is based on the reduction and approximation of a time series, the information loss, due to SAX implementation, needs to be considered during the analysis. The shapes and magnitudes of the energy profiles of buildings are influenced a great deal by occupancy and system operation schedules (e.g., start-up and shut-down of the heating/cooling system, occupants' arrival and departure times), which often result in high energy consumption variations over very short time periods. Strong load variations over time represent very important features of the daily energy profile, which should be automatically detected and isolated in specific time windows in order to reduce the constant approximation error. Managing the windowing process is a complex issue; an erroneous setting of the number of time windows and of the number of symbols could negatively affect the capability of the EIS tool to identify energy patterns that can be considered frequent and typical for irregularly occupied buildings or buildings characterised by a high number of anomalous consumption patterns and inefficient operating settings over time.

The impact of input parameters on the generation of symbolic sub-strings was evaluated in [89] by performing a sensitivity analysis on both the number of equal-length time windows and the number of symbols. It was found that increasing the size of the input parameters makes the interpretation of the results difficult in a manual way. On the other hand, a smaller number of parameters can generate a decreasing level of detail and the loss of key information.

In this context, the main objective of the proposed methodology, at the basis of the EIS tool, is to exploit the potential of SAX, while customising the process according to the specific data features, and at the same time to develop an automatic, but supervised, procedure for the tuning of input parameters.

The regression trees make it possible to perform numerical estimations, by segmenting the dataset through splitting criteria that are evaluated on explanatory variables. Setting the time of the day as an ordinal input variable makes it possible to identify subsequent time windows of different lengths on daily load profiles, where the constant approximation produces a low approximation error. The number of the time windows is defined by means of the complexity parameter α (varying between 0 and ∞), which represents the penalty of adding other time windows that do not contribute significantly to the improvement of the overall approximation error of the regression tree.

The great advantage of this tool is the self-tuning capability of the process to find the most appropriate lengths and number of time windows in order to effectively reduce the time series. Moreover, the number of symbols (alphabet size) necessary to encode the reduced time series is set equal to the number of sub-daily time windows. When the alphabet size is defined, the adaptive breakpoint identification process, performed with the aSAX algorithm [32], is completely automated.

After the encoding of the time series in symbols, infrequent pattern recognition is performed. Differently from other applications (e.g., as the work published in [89]) for the developed EIS tool the discords in the time series are detected at a single time window level by developing predictive classification models for each time window. A set of decision rules is then extracted from the classification trees to estimate the mostly frequently occurring SAX symbol for a time window, given certain boundary conditions. As a reference, in the case of the characterisation of thermal sensitive electrical loads (i.e., the electrical demand of chillers or heat pumps), the climatic conditions (external and internal), occupants' presence, the thermo-physical features of a building and the operation modes of the heating/cooling systems could help explain the existence of infrequent but not anomalous patterns, which are not always easy to be inferred. For this reason, and also thanks to the progressive electrification of heating/cooling systems, the pervasive monitoring of indoors/outdoors variables of influence is an important aspect to be taken into account in EIS design.

In this context, despite EIS tools are becoming more and more accurate and comprehensive in characterizing the actual energy behaviour of the building during its operation, the effectiveness of the information transfer to the user still remain a matter of concern.

In the developed EIS tool the results obtained through a real-time analysis of building-related data are provided to the user by means of scheduled feedbacks during the day. Each feedback is expected to be sent at the end of each time window and consists in a symbol estimation (explained trough IF-THEN rules), and powerful graphical visualisation that allow stakeholders to have an immediate picture of anomalous trends that deviate from the frequent/expected energy consumption patterns in quasi real time [35,211] and then avoiding further energy waste during the subsequent hours.

In this way, it is possible to compare the expected behaviour and the actual energy consumption at the end of a preidentified characteristic time period thus reducing the number of interactions with the user during a day while ensuring high consistency and interpretability of the knowledge transferred.

In the next section the last EIS tool developed is presented and discussed. In that case the analysis is focused on a larger scale than the single building with the aims of identifying typical energy use patterns and classifying energy customers at building portfolio level.

3.4 Development of EIS tool for the identification of typical energy use patterns and the classification of energy customers at building portfolio level

The progressive introduction of Advanced Metering Infrastructure (AMI) in the last years is enabling the collection of a huge amount of building energy consumption data [212,213]. In this context, data analytics based EIS can be exploited by energy suppliers or portfolio managers to gain insight into energy consumption patterns for a vast number of buildings [11]. A significant amount of research has been conducted in the field of building characterization using measured meter data [57,214].

This field of research often deals with the exploitation of various extracted temporal features from smart meter data [214] (e.g., load shape features, weather-dependency features, load pattern specificity, load diversity, long and medium-term volatility) for the segmentation and classification of large stock of buildings according to their energy behaviour.

Traditionally, when a portfolio of buildings is analysed, energy customers are segmented and classified according to their building end-use category as residential, industrial, commercial and so on. However, in many cases, buildings belonging to the same category can exhibit significantly different patterns in their energy consumption [9,121]. In such cases, benchmarking methods related to the calculation of energy use intensity metrics (e.g., kWh/m^{2}y) of the building are not able to fully characterise the energy behaviour of an energy customer over time. Conversely, knowledge extracted from energy consumption time series (i.e., load profiling analysis) contains information on how and when building energy use changes during the day for various end uses such as appliances, lighting, ventilation, heating and cooling [15,16].

A number of load profiling frameworks were developed in the literature to deal with data coming from multiple buildings usually with the aim to identify, through unsupervised analysis, homogenous groups of typical daily load curves (i.e., customer classification) characterised by similar shapes and/or magnitude [11,101]. In this context, advanced EIS tools capable to effectively mine the energy consumption patterns of buildings in large portfolios not only provide more robust energy benchmarks [9,15] but can also support the development of energy management initiatives and demand response programs [94] targeted to specific segments of users [215].

In order to provide a contribution in the aforementioned research fields, in the following is proposed and discussed the development of an EIS tool used to characterise energy consumption at building portfolio level with the aim of extracting typical patterns in the time domain.

The developed EIS tool also can address a robust customer classification process and enabling the classification of new unknown customers through a non-intrusive approach which does not make use of in-field load monitoring data.

Specifically, the identification of typical patterns is performed by analysing hourly daily load profiles grouped together using the “Follow the Leader” clustering algorithm (details in section 2.1.3.2.1). Successively, the classification of customers is performed developing a decision tree as a supervised classifier (details in section 2.1.3.1.1.2). The predictive attributes are gathered from monthly energy bills of each customer and from additional information on customers’ habits collected by means of phone survey.

The next section presents the main research challenges related to the identification of typical energy use patterns and customer classification in building portfolios and introduces the motivations and novelty of proposed methodological approach.

3.4.1 Motivations and novelty of the proposed approach

In the literature, the customer classification problem has been widely discussed by several researchers. Overviews on data mining based methodologies for customer classification are provided in [11,103,216]. As stated in section 2.2.2.2, this task unfolds over four main methodological stages: i) identification of *n* classes of customers with similar energy consumption profiles; ii) definition of the reference load pattern for each class; iii) enrichment of the database with predictive attributes; iv) development of a supervised classification model.

Although the clustering phase for the identification of typical energy use patterns is well investigated in the literature, little focus has been devoted to classification phase and in particular to the nature of the predictive attributes employed for developing the classifier. As previously explained in section 2.2.2.2, in most of cases the classification attributes are directly extracted from the load curves as done in [109,217]. These variables show an excellent explanatory potential; however, they can be computed only through an intrusive approach. This is usually unfeasible since energy providers or building managers not always have at their disposal such information when dealing with a new building to be included in their portfolio.

Indeed, the main issues to be addressed for developing a customer classification EIS tool are the following:

- Most of the analytical effort presented in literature is devoted to the pattern recognition phase (i.e., clustering phase) often neglecting the development of classifiers capable to estimate, for an unknown customer, its most probable cluster label and representative profiles;
- When a classification model is developed, in most of the cases the input attributes are gathered from in-field energy monitoring campaigns. It means that such a classifier can be used by an energy

provider/retailer/manager only for classifying buildings, whose energy consumption profiles are already available;

- The output of the customer classification process mainly consists in estimating normalised reference shapes of load profiles (e.g, (0,1) range) without providing any information about their magnitude;
- In most of the applications only one reference load pattern per customer is considered for the subsequent clustering analysis. This assumption while allowing the dataset to be reduced, in some cases can constraint the exploration of different load conditions (e.g., seasonal patterns).

In that perspective, the main objective is to develop and test an EIS tool that contributes to facing the aforementioned issues in a robust way as possible. The methodological framework of the analysis focuses on electrical load patterns of a stock of industrial and commercial buildings and relies on the application of data analytics techniques. The analysed dataset refers to the overall electrical demand of more than 100 energy customers of an Italian energy provider (eVISO s.r.l.). The EIS tool is conceived to be general for different types of buildings and is tested for a portfolio of buildings which significantly differ in volume and building end use.

In particular, the representative load profiles are grouped with a "Follow the Leader" clustering algorithm (discussed in section 2.1.3.2.1) [45,46]. In the post-clustering phase, a globally-optimal decision tree [218] is employed to build a supervised classification model and compared against a traditional recursive partitioning decision tree (the classification algorithms are explained in section 2.1.3.1.1.1 and section 2.1.3.1.1.2 respectively). The predictive attributes are extracted from monthly energy bills of each customer and from additional information about customers' habits collected by means of phone survey. The proposed tool can be then employed by energy retailers, energy managers and demand response operators to identify representative groups of customers in large heterogeneous building portfolios and to estimate for an unknown customer its most probable reference load profile by exploiting easy-to-collect and non-intrusive data and information (e.g., billing data, working time schedules).

The rest of the chapter 3 is organised as follows. Section 3.4.2 provides a description of the analysed dataset used for developing an EIS tool for the identification of typical energy patterns in building portfolio (benchmarking) and for the classification of new energy customers. Section 3.4.3 presents the methodology adopted for developing the EIS tool. Section 3.4.4 presents the results obtained for the analysed case study. Eventually, section 3.4.5 discusses the results and contains the concluding remarks related to this specific EIS application.

As a remark, part of the content of the section 3.4 was published as scientific article in the Elsevier journal "Applied Energy" [10].

3.4.2 Case study used for developing the EIS tool at building portfolio level

The customer classification EIS tool is developed starting from the monitored data of 114 electrical customers of the Italian Energy Provider (eVISO s.r.l.). The buildings are located in Piedmont (North-Western region of Italy) and are characterized by similar climate conditions. 17 customer typologies (i.e., building end uses) are considered in the analysis.

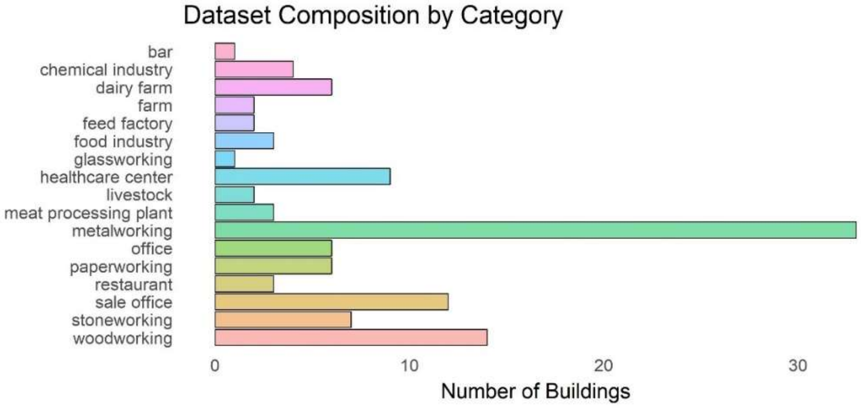
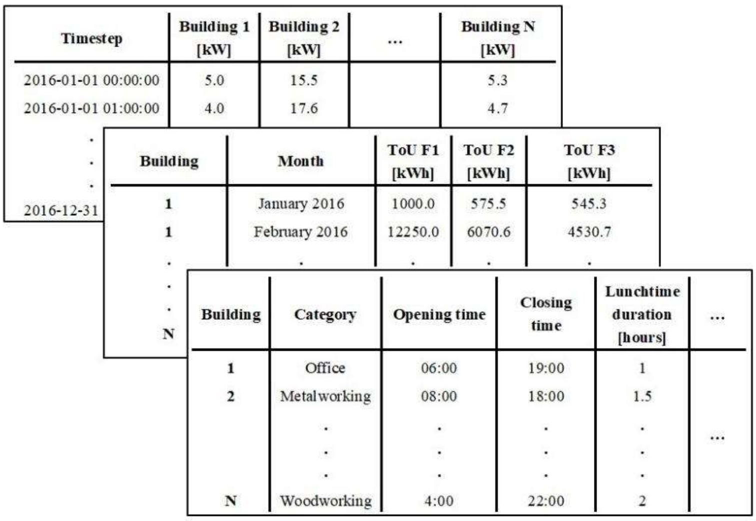


Fig. 48 - Number of customers for each category (adapted from [10])

In particular, from Fig. 48 it can be inferred that the majority of the analysed buildings are manufacturing industries (i.e., metal-working, wood-working, stone-working).
The analysed data consists in three different datasets:

- Electrical power dataset: it includes at least 4 months of measured hourly power demand of the 114 customers from “00:00:00 2014-01-01” to “23:00:00 2017-01-31”;
- Energy bills dataset: it includes the monthly billing information for the 114 customers;
- Additional info dataset: it includes features of the 114 customers such as building typology and working time.

Electrical power data were collected by means of smart meters installed by the energy provider while monthly billing data and the additional information were retrieved through short phone surveys and energy bills. For conducting the analysis, data are analysed and presented in anonymous form due to privacy issues related to the customer’s portfolio of the energy provider.



Timestep	Building 1 [kW]	Building 2 [kW]	...	Building N [kW]
2016-01-01 00:00:00	5.0	15.5		5.3
2016-01-01 01:00:00	4.0	17.6		4.7
.				
.				
.				
2016-12-31				

Building	Month	ToU F1 [kWh]	ToU F2 [kWh]	ToU F3 [kWh]
1	January 2016	1000.0	575.5	545.3
1	February 2016	12250.0	6070.6	4530.7
.	.	.	.	.
.				
.				
N				

Building	Category	Opening time	Closing time	Lunchtime duration [hours]	...
1	Office	06:00	19:00	1	
2	Metalworking	08:00	18:00	1.5	
	.	.	.	.	...
	.	.	.	.	
	.	.	.	.	
N	Woodworking	4:00	22:00	2	

Fig. 49 - Example of raw data structure [10]

Fig. 49 shows an extraction of records from the available dataset in order to provide an understanding of raw data structure. The year 2016 is selected as reference period for conducting the analysis. During 2016 the *electrical power* and *energy bills* datasets present the minimum amount of missing values and all the additional info are available.

Fig. 50 also shows the box plots of the average electrical power demand in the three italian ToU time slots for the buildings in the same category. The high diversity of the sample, in terms of building typology and energy consumption, represents an asset for developing this kind of EIS tool in the perspective of extracting knowledge as generalizable as possible.

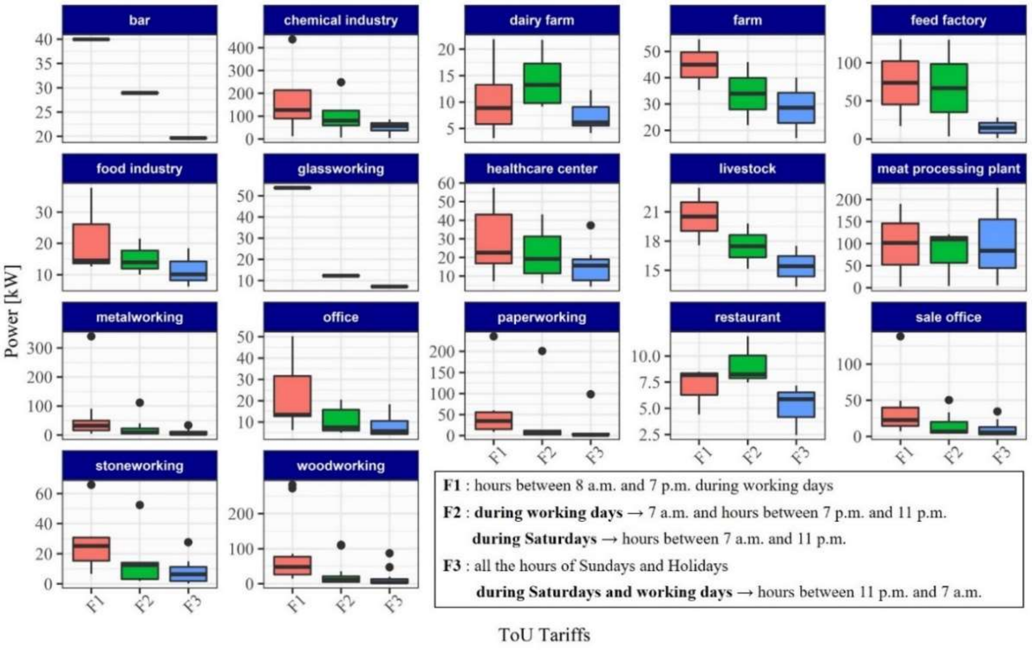


Fig. 50 - Box plots of the average electrical power demand in the three time slots related to different Italian electrical energy tariffs (F1, F2, F3) for the buildings in the same category [10]

3.4.3 Implemented methodology for the identification of typical energy use patterns and for the classification of energy customers

The methodology relies on the application of a clustering algorithm coupled with a decision tree, to perform a robust classification of a number of electrical industrial and commercial customers belonging to the same building portfolio. The whole process is developed and tested on the dataset previously described in section 3.4.2. The general framework unfolds over four different stages (Fig. 51) below introduced.

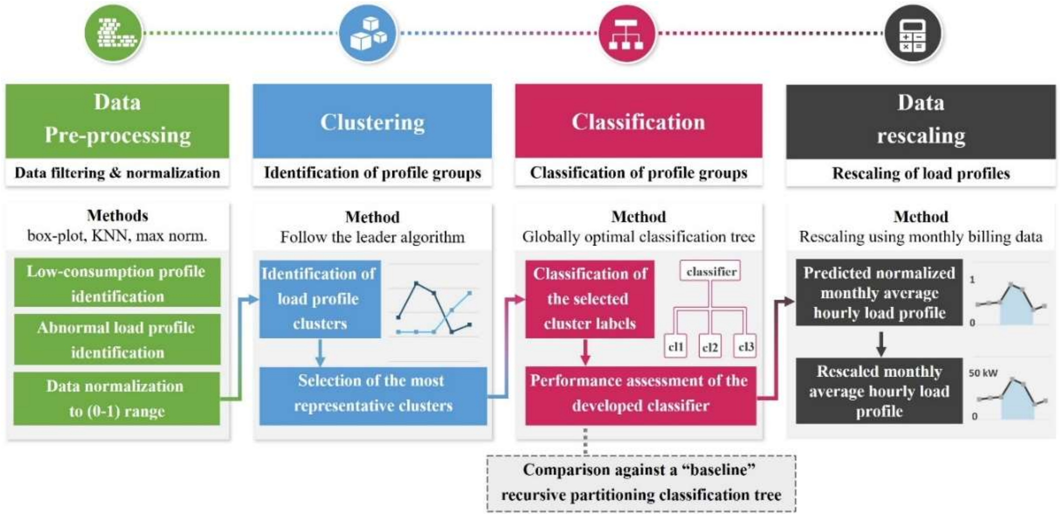

Data Pre-processing
Data filtering & normalization
Methods
box-plot, KNN, max norm.
Low-consumption profile identification
Abnormal load profile identification
Data normalization to (0-1) range
Clustering
Identification of profile groups
Method
Follow the leader algorithm
Identification of load profile clusters
Selection of the most representative clusters
Classification
Classification of profile groups
Method
Globally optimal classification tree
Classification of the selected cluster labels
classifier
cl1
cl2
cl3
Performance assessment of the developed classifier
Comparison against a "baseline" recursive partitioning classification tree
Data rescaling
Rescaling of load profiles
Method
Rescaling using monthly billing data
Predicted normalized monthly average hourly load profile
1
0
Rescaled monthly average hourly load profile
50 kW
0

Fig. 51 - General methodological framework of the analysis [10]

3.4.3.1 Data pre-processing for cleaning, filtering and normalizing building load profiles

The first stage is aimed at filtering and preparing the data. Data pre-processing is a mandatory task for any analytical process applied to data collected by means of smart meters. The time series of energy consumption for each building is chunked into daily sub-sequences. After the segmentation, only load profiles of working days are considered.

Punctual outliers are removed from daily load profiles and replaced through linear interpolation. Furthermore, also outliers at daily energy trend level are detected and removed. These profiles are characterised by very low or infrequent variation in energy demand over time.

The first type of abnormal patterns is represented by days during which the electrical load is significantly lower than the other working days. These days may include holidays or days that are not correctly identified and labelled as no-working days.

The identification process of these profiles is conducted separately for each building for each month. For each daily load profile, the daily power demand standard deviation is calculated. In this way, through a box plot analysis for each building for each month, the low variation profiles are identified according to the following equations (Eq. 16 and Eq. 17):

$$OUT_{SD} = Q1_{SD} - 1.5 \cdot IQR_{SD}$$

Eq. 16

$$IQR_{SD} = Q3_{SD} - Q1_{SD}$$

Eq. 17

Where Q1 and Q3 are, the first and third quartile of the frequency distribution of the standard deviation of daily power demand respectively and IQR represents the interquartile range. All the profiles labelled as OUT_{SD} are the lower outliers of the distribution and are removed from the set of data.

To the second type of abnormal patterns belong the days which electrical power demand is significantly different, in terms of magnitude and shape, from the other working days. The identification of such profiles is carried out separately for each month for each building in the dataset. To this purpose, the k-Nearest-Neighbours (KNN) algorithm is employed.

The algorithm computes the distance matrix between all the elements (i.e., daily load profiles) in a specific month and identifies for each profile the set of its K nearest neighbours. The number of K neighbours and the distance metric are set by the analyst.

In this case study K is assumed equal to 4 and the distance metric adopted is the Euclidean distance computed as follow (Eq. 18):

$$\boldsymbol{d} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{q_i} - \boldsymbol{p_i})^2}$$

Eq. 18

The algorithm returns for each element the distance values of its 4 nearest neighbours. These 4 values are then averaged into one single value and its frequency distribution among the months is computed. The outliers of these distributions represent the daily load profiles that significantly differ from their nearest neighbours and are identified according to the following equations (Eq. 19 and Eq. 20):

$$OUT_{KNN} = Q3_{KNN} + 1.5 \cdot IQR_{KNN}$$

Eq. 19

$$IQR_{KNN} = Q3_{KNN} - Q1_{KNN}$$

Eq. 20

Where Q1 and Q3 are, the first and third quartile of the frequency distribution of the average distance of each profile from its neighbours respectively and IQR represents the interquartile range. All the profiles labelled as OUT_{KNN} are the upper outliers of the distribution and are removed from the set of data.

Once the abnormal load profiles are identified and filtered out, the monthly reference load profiles for each building are calculated by averaging the remaining working daily load profiles in each month.

At this stage, in order to facilitate the subsequent grouping of similar profiles, also a normalization of data is carried out. The data normalization, especially for multidimensional problems, is necessary to compare load profiles of different buildings to each other, removing the effect of the amplitude variability of data attributes. For energy profiling tasks, amplitude differences related to different load conditions can negatively affect the performance of pattern recognition algorithms in discovering similar shapes among profiles. To this purpose the Normalized Monthly Reference Load Profiles (NMRLPs) in the (0,1) range are obtained normalizing each monthly reference load profiles respect to its maximum average power according to the following equation (Eq. 21):

$$\widehat{\boldsymbol{x}}_{i,m} = \frac{\boldsymbol{x}_{i,m}}{\max(\boldsymbol{x}_{i,m})}$$

Eq. 21

Where $\boldsymbol{x}_{i,m}$ is the vector representing the monthly reference load profile of the i-*th* customer for the m-*th* month and $\max(\boldsymbol{x}_{i,m})$ corresponds to its maximum value.

3.4.3.2 Clustering of the Normalized Monthly Reference Load Profiles (NMRLPs)

The second stage of the analysis is aimed at grouping similar NMRLPs in clusters which are representative of specific energy consumption patterns. The unsupervised segmentation is performed by means of “Follow the Leader” clustering algorithm [45,46] using the Euclidian distance as dissimilarity measure. Details on the clustering method are provided in section 2.1.3.2.1.

3.4.3.3 Classification of the identified clusters

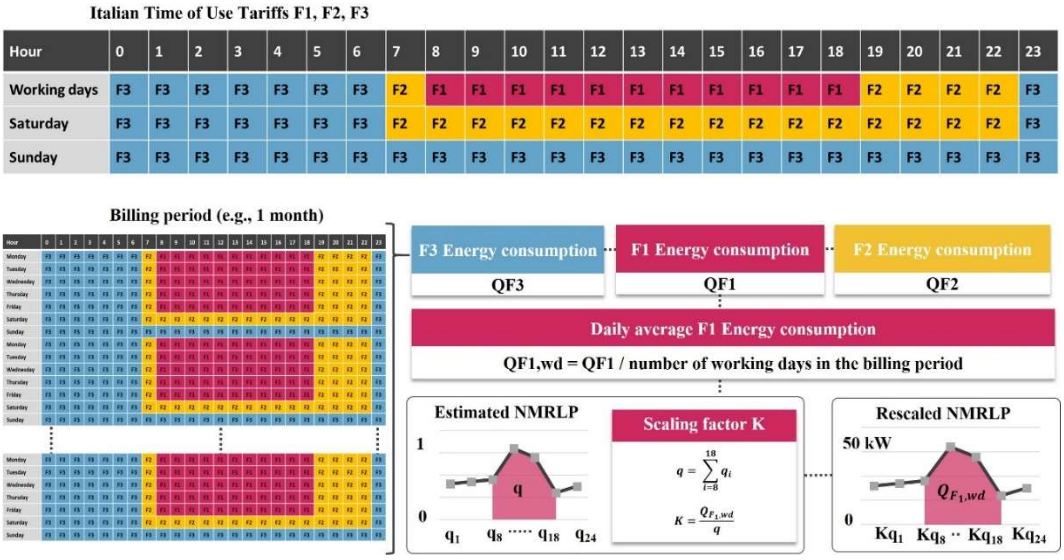
The clusters evaluated in the previous stage are analysed and described, and the labels of the most representative ones are used as target variables in a classification process. More in detail, a proposed model consisting in a globally optimal decision tree [110,218] is compared with a baseline model consisting of a recursive partitioning classification tree algorithm [204]. The proposed model makes use of stochastic optimisation methods (i.e., evolutionary algorithms) that can lead to much more accurate classification than locally optimal decision trees [218]. Both the classification models (i.e., baseline and proposed) are developed using the cluster labels as target variable, and additional building features as input variables. The classifiers are able to predict, for a new building included in the portfolio, the most probable NMRLP on monthly basis only using a-priori knowledge (e.g., occupant arrival and exit time) and billing data. As a consequence, an energy provider or energy manager may be able to easily estimate, for a new building, the monthly average hourly load profile based on the membership to a customer class previously identified. Details on the classification algorithms are provided in sections 2.1.3.1.1.1 and 2.1.3.1.1.2.

3.4.3.4 *Data rescaling* of the estimated Normalized Monthly Reference Load Profiles (NMRLPs)

The final stage of the process consists in the rescaling of NMRLPs. In fact, after the estimation of the NMRLP for a new building, it becomes essential to evaluate the magnitude associated to these normalized profiles. To address this task only historical billing data are used as shown in Fig. 52.

In Italy, from electrical energy bills, it is possible to associate energy consumption data to hours with specific Time of Use (ToU) tariffs. The Italian ToU tariffs consist of three different daily time slots (Fig. 52):

- F1 time slot (peak hours): it includes hours between 8 a.m. and 7 p.m. during working days;
- F2 time slot (off-peak hours): during working days this slot includes one hour in the morning (7 a.m.) and hours between 7 p.m. and 11 p.m. During Saturdays it includes hours between 7 a.m. and 11 p.m.;
- F3 time slot (off-peak hours) which comprises the remaining hours not included in the F1 and F2 time slots (i.e., Sundays, Holidays and night hours between 11 p.m. and 7a.m.).



Italian Time of Use Tariffs F1, F2, F3

Hour	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Working days	F3	F3	F3	F3	F3	F3	F3	F2	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1	F2	F2	F2	F2	F3
Saturday	F3	F3	F3	F3	F3	F3	F3	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F2	F3
Sunday	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3	F3

Billing period (e.g., 1 month)
F3 Energy consumption
QF3
F1 Energy consumption
QF1
F2 Energy consumption
QF2
Daily average F1 Energy consumption
QF1,wd = QF1 / number of working days in the billing period
Estimated NMRLP
1
0
q
q_1 q_8 ····· q_{18} q_{24}
Scaling factor K
$q = \sum_{i=8}^{18} q_i$
$K = \frac{Q_{F_1,wd}}{q}$
Rescaled NMRLP
50 kW
0
$Q_{F_1,wd}$
Kq_1 Kq_8 ·· Kq_{18} Kq_{24}

Fig. 52 - Methodological process for the rescaling of the Normalized Monthly Reference Load Profiles (NMRLPs) [10]

For developing the EIS tool only working days were analysed for computing NMRLPs for each building. For this reason, in order to rescale these normalized load profiles, only the energy consumption referred to the F1 slot during the billing period was considered, since the other slots are also included in weekends and holidays.

Assuming a monthly billing period, the total energy consumption in the time slot F1 for that period ($\boldsymbol{Q_{F_1}}$) is divided for the number of working days to calculate the daily average energy consumption $\boldsymbol{Q_{F_1,wd}}$ expressed in kWh. The scaling factor K is then calculated as follow (Eq. 22):

$$K = \frac{Q_{F_1,wd}}{q}$$

Eq. 22

Where q is the normalized daily average energy consumption of the estimated NMRLP during the F1 time slot (i.e., 8 a.m. – 18 p.m.) calculated as follow (Eq. 23):

$$q = \sum_{i=8}^{18} q_i * T$$

Eq. 23

Where $\boldsymbol{q_i}$ is the i-*th* normalized average power of the NMRLP and T is the timestep of the load profile expressed in hours. After the evaluation of the scaling factor, each $\boldsymbol{q_i}$ of the NMRLP is multiplied by K. Assuming that K is calculated starting from an average energy balance on about the 50% of the hours of a day (F1 time slot), it can be considered a reliable and representative scale factor for an entire working day. For this reason, the factor K is then used also to rescale $\boldsymbol{q_i}$ not included in the F1 time slot.

Following this framework, the rescaling process proves to be straightforward and robust.

The entire methodological process, behind the functionalities of the EIS tool, is tested using a sampling composed by 13 buildings, for which one-year of hourly data are available. The approximation error referred to classification and rescaling of NMRLPs is then evaluated in a testing phase.

3.4.4 Results obtained from typical energy use pattern recognition and customer classification analysis

3.4.4.1 Data pre-processing results

To develop the customer classification EIS tool, raw data are prepared and processed. The main objective of pre-processing phase is to evaluate the NMRLPs in a robust way. Data pre-processing unfolds over different stages that makes it possible to automatically filter out from the dataset weekends, daily load profiles with low standard deviation and abnormal daily load profiles. For the year 2016 the "*electrical power dataset*" is composed by 41.724 daily load profiles. After data pre-processing the dataset is reduced of about the 42% of the total number of daily load profiles (Fig. 53(a)). In particular are filtered out:

- The 31% of the total amount of load profiles referred to weekends or holidays;
- The 8% of the total amount of load profiles labelled as working days that have low standard deviation of power during the day;
- The 3% of the total amount of load profiles labelled as working days that are characterized by abnormal/infrequent patterns.

The final dataset is then composed by 24.310 daily load profiles.

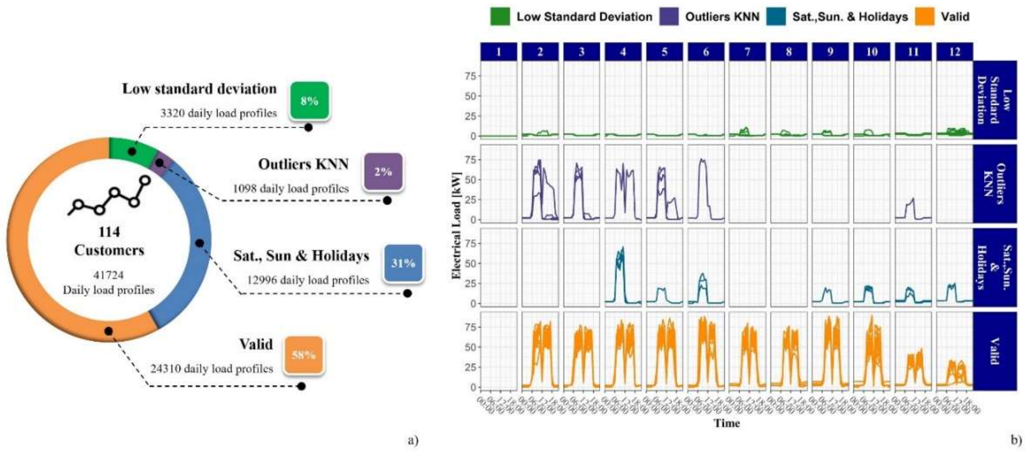


Fig. 53 - Percentage of valid and excluded load profiles after pre-processing analysis (a) valid and excluded daily load profiles grouped by month for a randomly selected customer (b) [10]

Fig. 53(b) shows the impact of data pre-processing for a randomly selected building in terms of valid and excluded daily load profiles. It is possible to notice that, after the data filtering, the remaining daily load profiles (in orange) exhibit high homogeneity in each month. This ensure that averaging those profiles per month, leads to a robust evaluation of reference patterns. At the end of the entire process the available set of monthly reference load profiles is made up of 1.249 NMRLPs normalised in the range (0,1). It is important to highlight that, although a reference period of 1 year is considered for the analysis, the number of NMRLPs per customer may be different from twelve (i.e., one per month) due to the presence of missing data or the filtering of entire months during the pre-processing phase (e.g. August). On average per each customer about 10 NMRLPs are available.

3.4.4.2 Clustering Results

In order to find similar groups of NMRLPs a clustering analysis is performed. The “Follow the Leader” algorithm is employed to this purpose as previously explained in section 3.4.3. The initialization of the algorithm consists in choosing an optimal value of the parameter ρ. To do this a sensitivity analysis is conducted, using the Davies Bouldin index (DBI) as reference metric for cluster validation (according to the process described in section 2.1.3.2.1). Considering that monthly reference load profiles are normalized, ρ represents an a-dimensional threshold distance between load profiles in the range (0,1). The clustering results are evaluated for different values of ρ in a range between 0,8 and 2,0 with an incremental step of 0,05. For each setting of ρ the corresponding number of clusters and DBI is calculated. Fig. 54 shows the results of the sensitivity analysis.

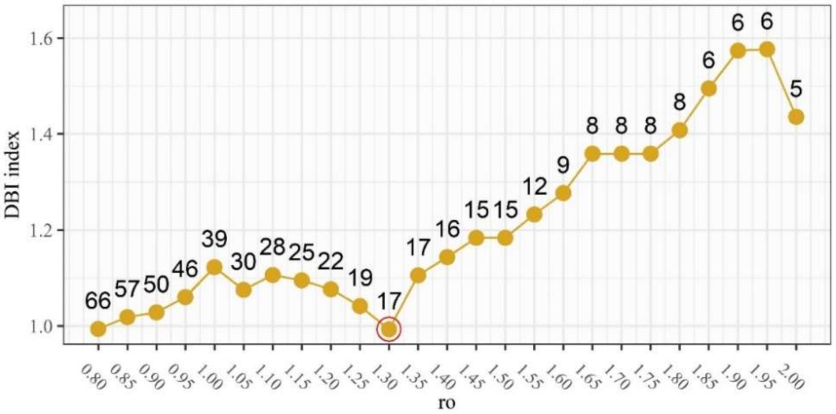


Fig. 54 - Identification of optimal value ρ* with the corresponding number of clusters for the initialization of "Follow the Leader" algorithm [10]

The Fig. 54 shows that the optimal value ρ* of the parameter ρ (that minimize the DBI) is equal to 1,30 and corresponds to the identification of 17 clusters. It means that for ρ = ρ* the resulting clusters exhibit optimal inter cluster separation and intra cluster cohesion. The 17 clusters obtained have different cardinalities and are shown in Fig. 55 with the evidence of their centroids.

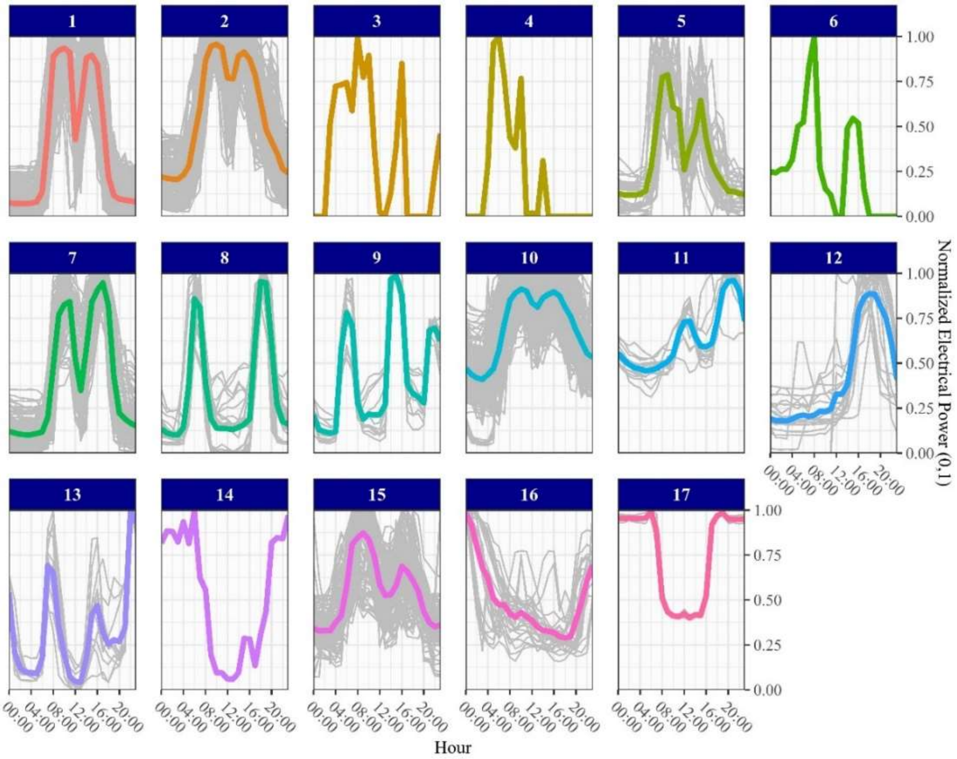


Fig. 55 - Clusters of load profiles identified through the "Follow The Leader" algorithm [10]

For classification purpose, only the labels of the most representative and populated clusters are considered. The selection process unfolds over a descriptive analysis of the clusters obtained.

The Fig. 56 shows the scatter plot of the number of buildings (x-axis) versus the number of NMRLPs (y-axis) grouped in each cluster. The horizontal and vertical dashed red lines represent the average number of NMRLPs and of buildings per cluster respectively. In this way the 17 clusters are segmented according to two main plane regions.

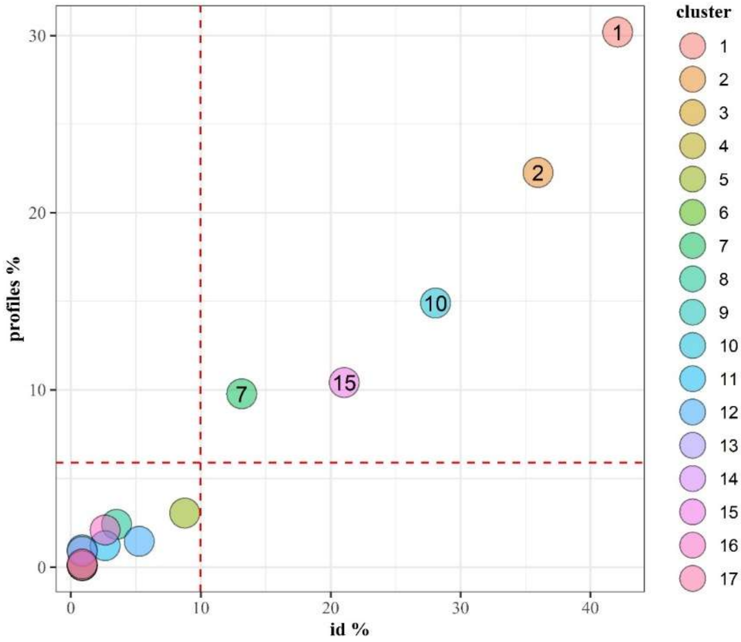


Fig. 56 - Scatter plot of the number of customers (x-axis) versus the number of NMRLPs (y-axis) grouped in each cluster [10]

The first region includes clusters in the left-bottom corner of the plot. These clusters group together few NMRLPs and buildings that are characterized by patterns that significantly differ from the rest of the dataset. In particular those clusters can be described as follow:

- Clusters 3, 4, 6 and 14 include one single NMRLP. These profiles correspond to specific months during which the energy consumption patterns of some buildings are infrequent. Although those profiles are not filtered out during the pre-processing phase, the “Follow-the-Leader” algorithm is able to isolate them.
- Clusters 9, 13 and 17 include all the NMRLPs of one single building. These buildings show infrequent energy consumption patterns compared to rest of the dataset and high intra cluster cohesion.
- Within Cluster 8 are grouped together buildings with the same end-use which is related to milk production activities (i.e., dairy farms).
- Within clusters 11, 12, 16 are grouped together buildings with end-use related to food-service activities (i.e., food industry). These are the only

buildings characterised by a power demand during night hours higher than in the morning ones.

- Cluster 5 includes several buildings with different end-uses. However, the total number of NMRLPs that are grouped in this cluster corresponds to around the 3% of the total.

The second region includes clusters in the right-top corner of the plot. In these clusters is grouped about the 90% of the total number of NMRLPs available in the dataset corresponding to 103 out of 114 initial customers. Centroids of clusters 1, 2, 7, 10 and 15, are then generated from the most typical recognised patterns in the dataset. All these patterns are characterized by higher power demand during morning and afternoon hours than the night ones. Moreover, a reduction of power demand during the middle hours of the day occur due to the effect of a lunch-break. Although these clusters show similar trends some differences can be pointed out (see Fig. 55):

- *Cluster 1* groups profiles for which power demand is high between “07:00” and “18:00” (i.e., around the 90% of the maximum power) with a strong decrease between “12:00” and “14:00” due to the lunch-break (i.e., the power demand is around the 50% of the maximum power);
- *Cluster 2* groups profiles for which the night power demand is higher than the profiles included in cluster 1 and the effect of lunch-break is less intense. Moreover, the power demand is still high also after “18:00”;
- *Cluster 7* groups profiles similar to cluster 1 but for which the power demand peak occurs in the afternoon hours after the lunch-break hours;
- *Cluster 10* groups profiles with the highest power demand during night hours (i.e., the power demand is around the 30-40% of the maximum power) compared to the other clusters (i.e., cluster 1,2,7,15);
- *Cluster 15* groups profiles for which the power demand is higher in the morning hours than in the afternoon hours after the lunch-break.

In each of these clusters at least the 10% of the buildings are grouped as well as about the 10% of the NMRLPs. This ensures the representativeness of such groups for customer classification purpose. For this reason, only the labels of clusters 1,2,7,10,15 are used in the subsequent phase and encoded as the categorical target variables of the decision tree. As demonstrated in other studies [45,46], the FTL algorithm is capable to identify the most relevant clusters within the given dataset. The algorithm proves to be able to in handle outliers isolating anomalous/infrequent patterns in separate clusters that are easily identified and filtered out.

3.4.4.3 Classification Results

In the classification phase of the methodological framework two classification models, which are based on different learning process, are compared in terms of accuracy for predicting the cluster labels assigned to each group of NMRLPs evaluated in the clustering stage. In detail, a traditional recursive partitioning decision tree is selected as baseline, while a globally optimal decision tree is proposed as improved alternative.

Decision trees are robust and highly readable algorithm and at this stage are used to predict, for new customers, their monthly average hourly load profiles based on the membership to one of the clusters previously identified by means of FTL algorithm. It is important to notice that the prediction is monthly-based, and then a customer could have NMRLPs belonging to different clusters for each month. Therefore, in this case, the decision trees allow to finely characterise also customers with multiple typical NMRLPs among the year (e.g., presence of seasonal-based patterns).

To develop the models, the input attributes are selected from the available datasets. The variables included in the model can be easily acquired through short phone survey and from customer energy bills. In this way the input data collection can be considered as a non-intrusive process, since in-field energy monitoring is not needed. The input variables considered for both the "baseline" and "proposed" classifier are summarised in Table 11. All the input variables are treated as numeric or ordinal attributes, while the target variable (i.e., cluster labels) as a categorical attribute.

Table 11 - Input variables for both "baseline" and "proposed" classifiers [10]

	Description	Unit	Name
Monthly-scale Variables	Energy Consumption in time slot F1 / Total Energy consumption	-	F1
	Energy Consumption in time slot F2 / Total Energy consumption	-	F2
	Energy Consumption in time slot F3 / Total Energy consumption	-	F3
	Energy Consumption in time slot F1 / Energy Consumption in time slot F2	-	F1_2
	Energy Consumption in time slot F1 / Energy Consumption in time slot F3	-	F1_3
	Energy Consumption in time slot F2 / Energy Consumption in time slot F3	-	F2_3
Customer-features	Working start time	[h]	opening
	Working end time	[h]	closing
	Lunch break duration	[h]	d_lt

Before developing the classification models, from each customer cluster at least one customer is sampled (with all its NMRLPs) to be used as testing. The testing dataset consists of 13 customers and 142 NMRLPs. Training and testing datasets are identified in order to obtain nearly the 85% of the initial population size in the training set, avoiding the presence of the NMRLPs of the same customer in both of them. Moreover, in order to roughly maintain the same share of cluster objects in the two sets, from each cluster a number of customers is sampled proportional to the cluster cardinality.

In order to perform a robust and reliable comparison, for both the "baseline" and "proposed" classifier the minimum number of elements in each leaf node (*minbucket*) and the maximum depth reachable by the tree (*maxdepth*) are set equal to 20 and 3, respectively. The *minbucket* is set equal to two times the average number of MRLPs for each customer, ensuring the presence of at least two customers classified in each leaf node of the tree. On the other hand, the maximum tree depth is set large enough to develop an accurate tree but not too much complex for avoiding overfitting issues. Considering that a *maxdepth* equal to 3 levels already limits the complexity of the possible solutions to a maximum number of 8 leaf nodes (as a consequence of three levels of binary splits), the complexity parameter α was set for both the classifiers equal to 0 in order avoid an additive penalty index in the evaluation function of the model.

Table 12 - Configurations of variation operator probabilities (globally optimal decision tree) [10]

Setting of the variation operators	**Probabilities**				
	Crossover	**Major mutation**	**Minor mutation**	**Split**	**Prune**
c20m40sp40	20 %	20 %	20 %	20 %	20 %
c10m30sp60	10 %	15 %	15 %	30 %	30 %
c00m50sp50	-	25 %	25 %	25 %	25 %
c40m20sp40	40 %	10 %	10 %	20 %	20 %
c10m10sp80	10 %	5 %	5 %	40 %	40 %
c50m00sp50	50 %	-	-	25 %	25 %

For the "proposed" decision tree, based on the evolutionary learning algorithm, further hyper parameters need to be set (as explained in section 2.1.3.1.1.2). The parameters to be tuned are the population size Θ, the maximum number of iterations and the variation operator probabilities. Six different configurations of variation operator probabilities, three different number of maximum iterations and four population sizes are tested. This analysis unfolds over two steps, as presented in [218].

In the first step, the 18 configurations generated by combining six different settings of variation operator probabilities (Table 12) and three maximum number of iterations (i.e., 500, 1000, 10000) are analysed. Each combination is tested for 100 different random initialisations of the population Θ, which is fixed at 100 trees (default value). Each solution is evaluated computing its misclassification error. Fig. 57 shows the box plots of the 100 misclassification errors for each of the 18 combinations of the trees developed on the entire dataset.

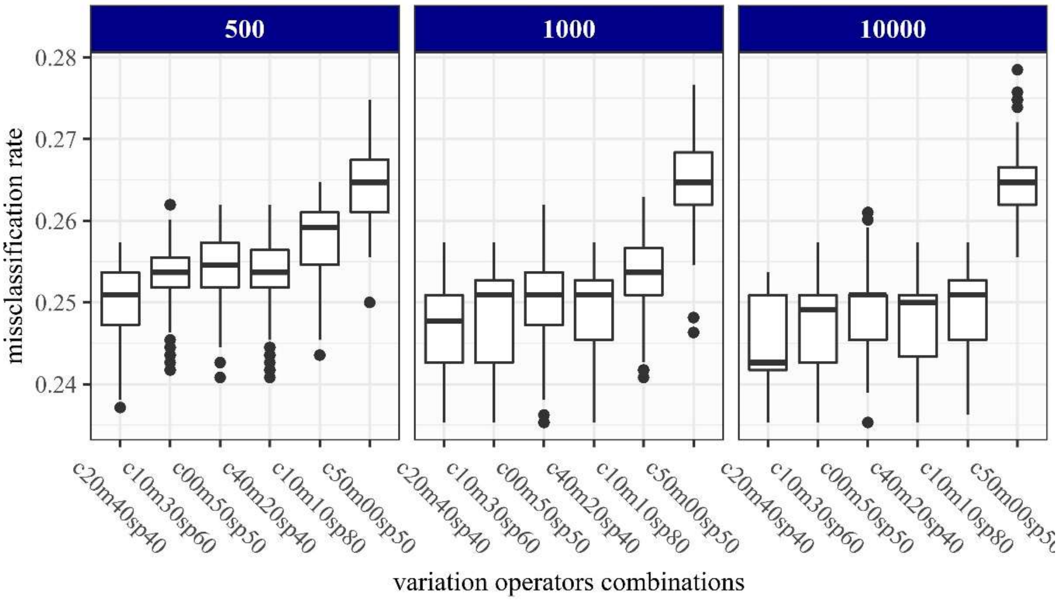


Fig. 57 - Misclassification rates for 18 configurations of iteration number and variation operator probabilities (globally optimal decision tree) [10]

From this first step it is possible to infer that the misclassification rate of the decision trees decreases with the increasing of the maximum number of iterations reaching its best median value for 10000 iterations and variation operator probabilities set at *c20m40sp40.* In the second step of the analysis the impact of the population size Θ on the misclassification rate is evaluated considering 100 different random initialisations of populations with size of 25, 50, 100, 250 and 500 trees respectively. In this step the number of iterations and variation operator probabilities are set at 10000 and *c20m40sp40* respectively, that correspond to the optimal values previously identified.

Fig. 58 shows that the cardinality of population size positively affects the overall performance of the decision tree, reaching the minimum median value of the misclassification rate for a population of 500 trees.

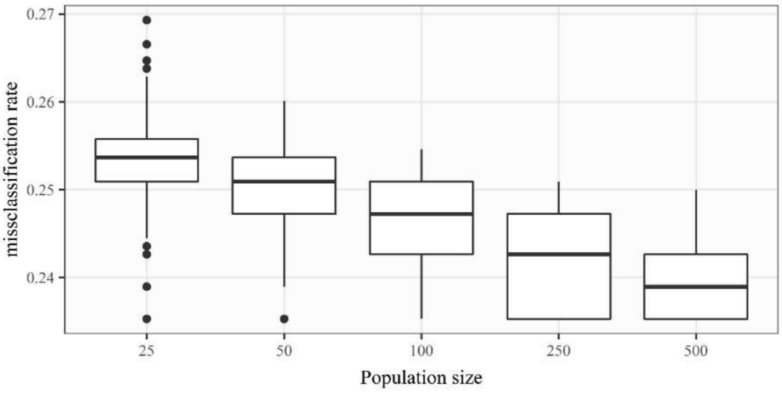


Fig. 58 - Misclassification rates for populations with size of 25, 50, 100, 250 and 500 trees respectively (globally optimal decision tree) [10]

According to the performed sensitivity analysis, the globally optimal tree was then developed on the training dataset with the following parameter setting: Θ equal to 500, number of iterations set to 10000 and the variation operator probabilities set to *c20m40sp40*.

Fig. 59 and Fig. 60 show the final decision trees (i.e., “baseline” vs “proposed”), developed on the training dataset. The two trees differ in terms of number of leaf nodes and input variables used for the split generation. The globally optimal decision tree is capable to converge into a more detailed and accurate solution following decision rule paths different from the other model.

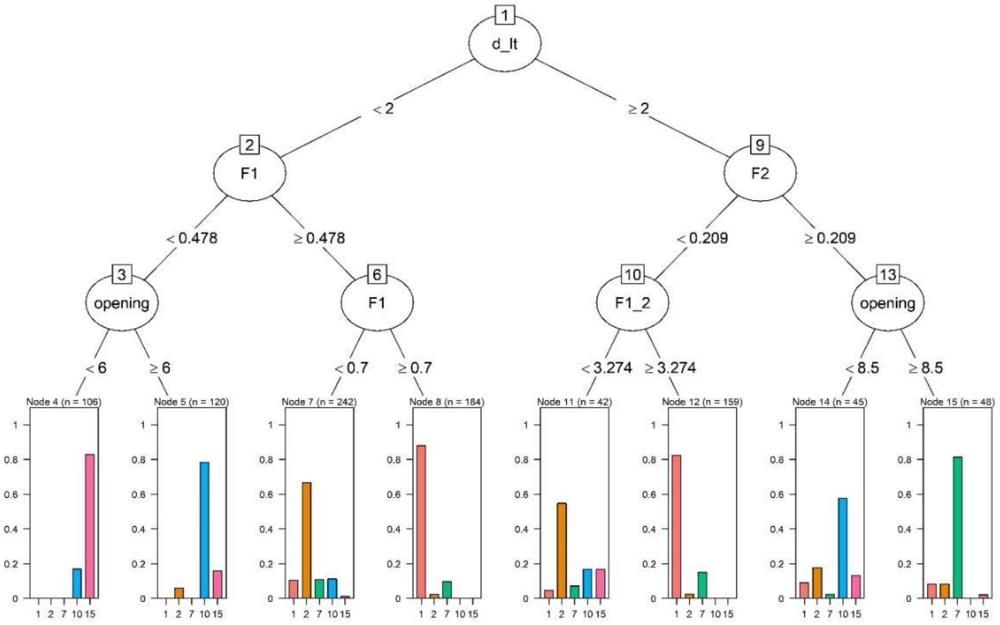


Fig. 59 - Globally optimal decision tree [10]

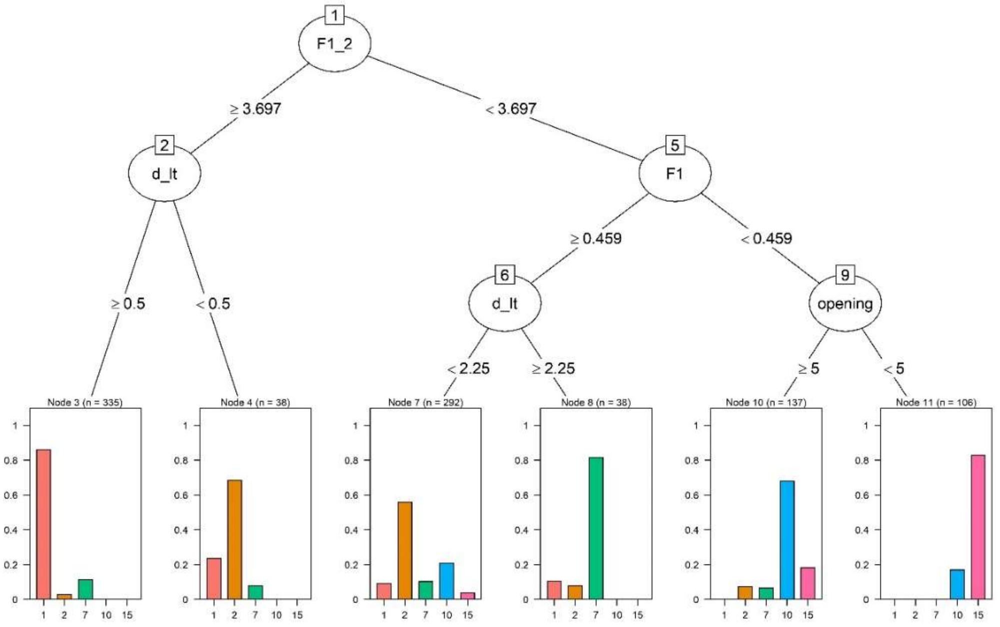


Fig. 60 - Recursive partitioning decision tree [10]

In fact, the locally optimal decision tree at each parent node evaluates the best split, maximizing homogeneity in the next step only. On the contrary, the globally optimal decision tree is capable to leverage on less accurate internal splits in order to reach a higher final performance of the classifier. Table 13 and

Table 14 report the decision rules extracted from the two classifiers.

Table 13 - Decision rules extracted from globally optimal classifier [10]

Cluster	Node	Decision Rules	Profiles	Accuracy
1	8	IF **d_lt** < 2 AND **F1** ≥ 0.504 AND **F1** ≥ 0.701	184	88 %
	12	IF **d_lt** ≥ 2 AND **F2** < 0.208 AND **F1_2** ≥ 3.27	159	82.4%
2	7	IF **d_lt** < 2 AND **F1** ≥ 0.504 AND **F1** < 0.701	230	67.8 %
	11	IF **d_lt** ≥ 2 AND **F2** < 0.208 AND **F1_2** < 3.27	42	54.8 %
7	15	IF **d_lt** ≥ 2 AND **F2** ≥ 0.208 AND **opening** ≥ 08:30	48	81.7 %
10	5	IF **d_lt** < 2 AND **F1** < 0.504 AND **opening** ≥ 06:00	131	75.6 %
	14	IF **d_lt** ≥ 2 AND **F2** ≥ 0.208 AND **opening** < 08:30	45	57.8 %
15	4	IF **d_lt** < 2 AND **F1** < 0.504 AND **opening** < 06:00	107	82.2 %

Table 14 - Decision rules extracted from recursive partitioning classifier [10]

Cluster	Node	Decision Rules	Profiles	Accuracy
1	3	IF **F1_2** ≥ 3.697 AND **d_lt** ≥ 0.5	335	86 %
2	4	IF **F1_2** ≥ 3.697 AND **d_lt** < 0.5	38	68.4 %
	7	IF **F1_2** < 3.697 AND **F1** ≥ 0.459 AND **d_lt** < 2.25	292	55.8 %
7	8	IF **F1_2** < 3.697 AND **F1** ≥ 0.459 AND **d_lt** ≥ 2.25	38	81.6 %
10	10	IF **F1_2** < 3.697 AND **F1** < 0.459 AND **opening** ≥ 05:00	137	67,9 %
15	11	IF **F1_2** < 3.697 AND **F1** < 0.459 AND **opening** < 05:00	106	83 %

Both models suggest that NMRLPs grouped within cluster 1 and cluster 2 are characterised by a higher monthly energy consumption during time slot F1 respect to other clusters. However, the energy consumption during F2 hours are more significant for cluster 2 compared to cluster 1. According to the globally optimal decision tree solution, customers whose MRLPs are grouped within cluster 7 are characterized by working activities starting later than 08:30 a.m., while such a feature is not extractable from baseline solution (recursive partitioning decision tree). Within cluster 10 and cluster 15, are grouped MRLPs for which the energy consumption during time slot F2 are higher compared to other clusters. The difference between the cluster 10 and 15 consists in an earlier starting of working activities for customers in cluster 15 than of others in cluster 10. Those cluster features are exploited by both models, however, the globally optimal decision tree shows a more detailed description by using one more decision rule. The rules are furtherly applied on the testing set to evaluate "out-of-sample" performances of the two models.

Table 15 reports the overall misclassification errors of the two models for the training and testing datasets. It is possible to see that the proposed globally optimal decision tree performs better than the locally optimal one both in training and testing. The accuracy in testing session increases of about 6%.

Although for the "proposed" model the setting of parameters is not straightforward and the computational cost is quite high, the algorithm is capable to reach results significantly better than the "baseline" approach in terms of generalizability and accuracy of the model.

Table 15 - Overall misclassification errors of recursive partitioning and globally optimal decision tree for the training and testing dataset [10]

	Misclassification error	
	Globally optimal decision tree	***Recursive partitioning decision tree***
training	23.5 %	27.1 %
testing	24.6 %	30.9 %

3.4.4.4 Rescaling results

The last phase of the methodological process consists in the rescaling of the estimated NMRLPs. In fact, after the classification of the 13 customers of the testing dataset, their estimated NMRLPs are rescaled in order to obtain a reference hourly power demand profile expressed in kW. The NMRLPs are rescaled by multiplying their 24 values (one for each hour of the day) by the scaling factor K (as explained in section 3) obtained by using the actual monthly energy consumption in the F1 time slot. The rescaled NMRLPs are compared to the actual ones in order to evaluate the overall performance of the methodological framework.

In particular the Pearson correlation computed between real and rescaled profiles is select as validity index. For instance, in the case of a customer with 10 monthly reference load profiles, the correlation coefficient is calculated among 24 x 10 data points expressed in kW. On average for the entire testing set, consisting in 142 profiles, a strong linear correlation coefficient equal to *0.895* is obtained (Fig. 61).

It means that the EIS tool is capable to return, for an unknown customer, a set of estimated monthly reference load profiles that are accurate in terms of both magnitude and shape.

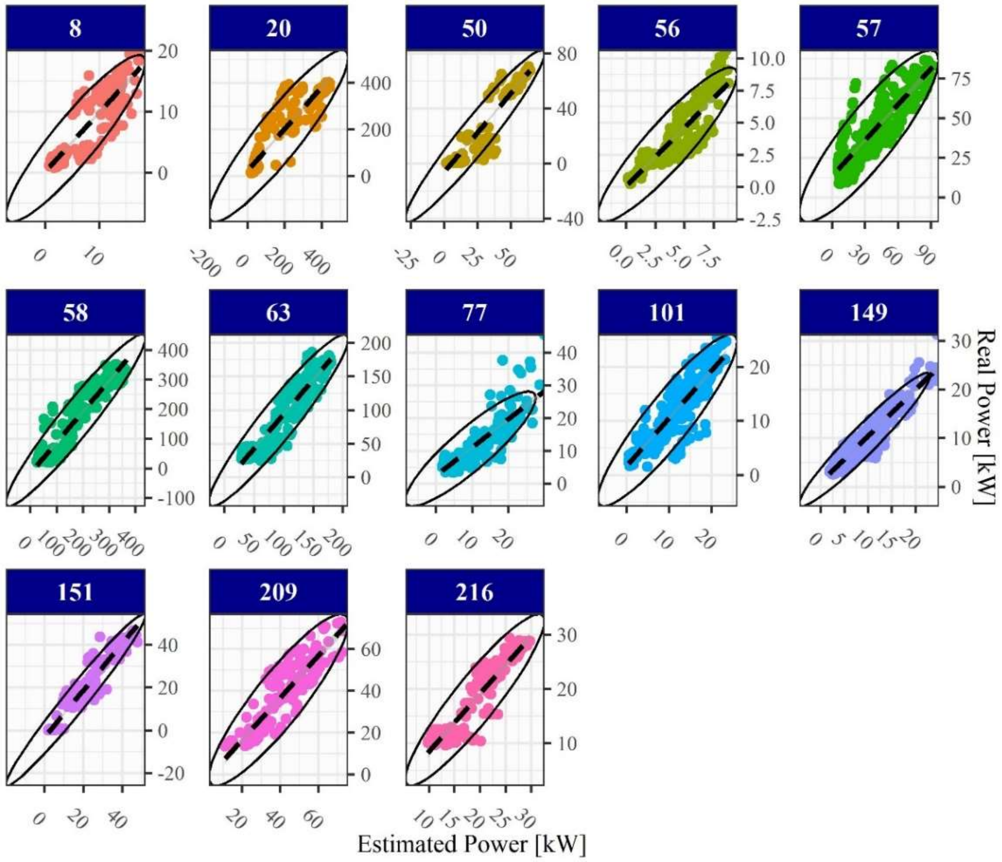


Fig. 61 - Linear correlations between actual and rescaled estimated energy profiles for each customer of the testing set [10]

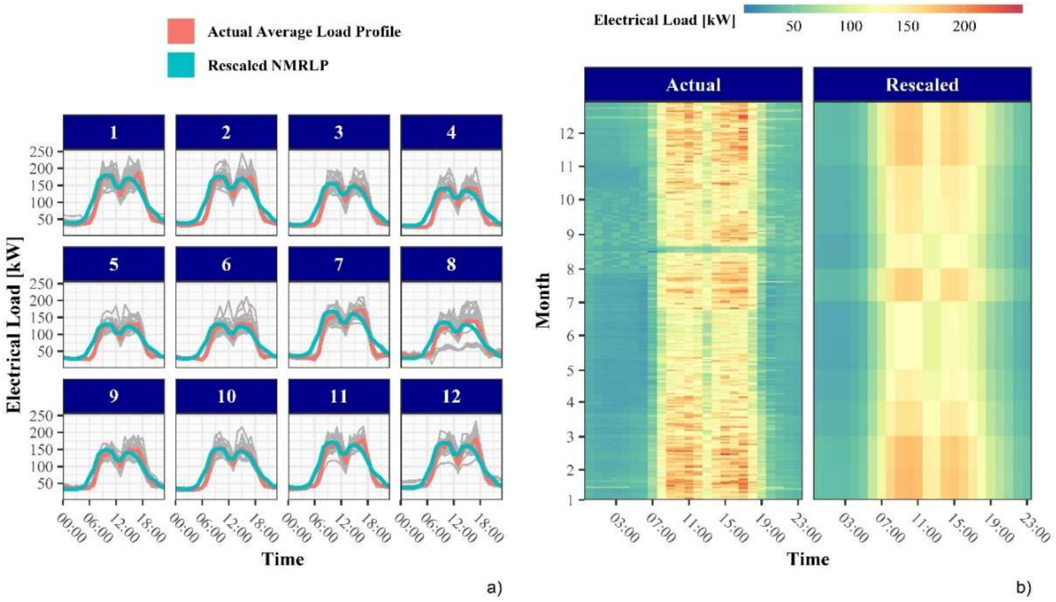


Fig. 62 - Actual load profiles of the working days (grey lines), actual average load profiles (red lines) and rescaled load profiles (blue lines) of a randomly selected customer from the testing set (a) carpet plot of actual load profiles together with the carpet plot reconstructed on monthly basis through the rescaled estimated load profiles (b) [10]

As a reference, Fig. 62 (a) shows the results of the rescaling process for each month of a randomly selected customer from the testing set. For each month, grey lines show the actual load profiles of the working days, the red line is the actual average profile and blue line is the rescaled NMRLP. In addition, Fig. 62 (b) shows the carpet plot of actual load profiles together with carpet plot reconstructed on monthly basis through the rescaled estimated load profiles. Both figures show how the process performs proving its robustness and effectiveness.

3.4.5 Discussion

The developed EIS tool provides a robust solution for the automatic identification of typical energy use patterns and classification of energy customers in building portfolios.

To this purpose, supervised and unsupervised data analytics techniques are combined in the methodological framework of the analysis. In the pattern recognition phase the "follow the leader" method is used for identifying typical energy use patterns of the most significant customer groups in a building portfolio. At the same time the fine tuning of the pattern recognition process allows the EIS tool to isolate infrequent or anomalous patterns in separate groups. The algorithm belongs to the family of partitive clustering techniques, but differently from K-means it requires a distance threshold instead of the number of desired clusters K as input parameter. It brings advantages in terms of algorithm flexibility. In fact, the use of a distance threshold makes it possible to better manage infrequent/anomalous patterns without performing a preliminary outlier detection for preserving clustering performances. The setting of the threshold ρ is supervised by using the Davies Bouldin Index as a cluster validity metric allowing the optimal value ρ* to be automatically identified. The cluster analysis results in the identification of 17 customer groups among the portfolio, that are characterized by different cardinality and reference shapes of the load profiles. Even if the high diversity of patterns represents an asset in a customer classification process, a large customer database is required to adequately represent each of them.

For the case study considered the developed EIS tool considers 5 cluster labels in the classification phase given that the remaining 12 groups include few customer profiles or discord ones. Excluding Clusters 3, 4, 6 and 14 that group one single discord NMRLP, the others are candidates for being considered in the classification process when further NMRLPs will be stored in the customer database. In that perspective, the EIS tool can be considered open and furtherly upgradable considering that more cluster labels could be taken into account in the future for developing an extended classification. One of the most recently developed algorithms for decision tree based on globally optimal learning process is tested and compared with the well-known one-step-forward approach. This proposed classification model leads to an improved accuracy of 6% for the testing data set in comparison to the baseline classification model. Differently from more straightforward decision tree models, the globally optimal algorithm requires a high computational cost and the tuning of model parameters represents a non-trivial and time-consuming task. For the case study analysed, the higher accuracy achieved, and the limited database volume make its implementation reasonable. The algorithm is capable to accomplish the classification task by fully exploiting few input variables collected through a non-intrusive approach.

This aspect represents one of the strengths of the methodological process at the basis of the EIS tool, given that it allows final users to preliminary characterize electric or thermal energy customers in a very detailed way without using in-field monitoring data [219]. The opportunity to estimate, for an unknown customer, its most probable NMRLPs is highly desirable for several stakeholders (e.g., energy suppliers, local and national authority, energy manager of a large building portfolio) in the smart city environment.

As a consequence, the knowledge that can be extracted through this kind of EIS tool can enable the definition of both robust energy performance benchmarks and effective energy management strategies conceived for specific customer groups. As a reference, on the basis of their representative load profiles, specific groups of customers can be involved in targeted financial demand response programs (e.g., Time Of Use tariff, Critical Peak Pricing, Real-Time Pricing).

These programmes are getting more and more attention as retailers keep looking for a better way to balance loads and at the same time increase their profitability. On the other hand, such programs are designed to be attractive also for the consumers (building portfolio managers) as they can exploit a deeper knowledge of their energy patterns to reduce the total energy bill cost of their building portfolios. In this context the modification of a load profile plays a critical role not only from an economical point of view but also in terms of grid stability.

The developed EIS tools can also be employed for tracking the changes of power consumption patterns over time. By benchmarking customer flexibility (in terms of demand modification) it is possible to assess which could be the influence and the impact of specific Demand Side Management (DSM) and Demand Response (DR) initiatives for a group of customers or even at larger scale (e.g. district). Furthermore, the rescaling process, adopted in the methodological framework of analysis, makes it possible to associate a magnitude to the normalised load profiles once they are classified. The results prove that introduced the methodological process allows to robustly estimate for an unknown customer, a set of monthly reference load profiles that are accurate in terms of both magnitude and shape.

An EIS tool capable to estimate the shape of a load profile together with its magnitude enables the fully characterization of building energy use patterns from the very early customer engagement stage.

4 DSS application at system-level: development of a fault detection and diagnostic (FDD) tool

This chapter discusses in detail the development of a data analytics-based methodology that can be integrated in a DSS. The focus is on system-level analysis, performed by means of Fault Detection and Diagnostic (FDD) tool. The developed tool is conceived for conducting automated FDD analysis on HVAC system data that are specifically referred to the operation of Air Handling Unit (AHU) components.

This content is developed as a publication submitted to the Elsevier journal "Energy and Buildings":

- Piscitelli M.S., Mazzarelli D.M., Capozzoli A. Submitted for publication. *Enhancing operational performance of AHUs through an advanced fault detection and diagnosis process based on temporal association and decision rules*. Energy and Buildings. [22]

4.1 Development of FDD tool for the detection and diagnosis of faults at system level

FDD tools belong to the family of software solutions that automate the process of detecting faults and improper operation of building systems and help user to promptly diagnose their potential causes [126]. FDD tools represent an essential part of DSS that is focused on system-level applications often exploiting data collected through Building Automation Systems (BAS). Such tools are typically integrated to BAS as separate software and are able to provide detailed information about the duration and frequency of faults, with reference of their relative cost, energy impact and priority level [12,130].

The literature review reported in sections 2.2.3 and 2.2.3 demonstrated how much active the scientific research is in the field of data analytics application for FDD in HVAC systems with a specific reference to AHUs. The opportunity to approach the well-known task of FDD from this innovative point of view is mainly due to the growing availability of experimental data related to both normal and faulty operation of systems. In this context some projects, supported by the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE) made very comprehensive field surveys, laboratory tests and performance evaluations about the performance of HVAC systems also in faulty conditions. The outcomes of such projects (e.g., ASHRAE Project 1312-RP and 1043-RP) enabled a great spread of FDD methodologies exploiting experimental measured data. Among the papers reviewed in section 2.2.3, several published studies focused on the ASHRAE RP-1312 data set for developing and testing FDD methodologies for AHUs [55,138,220–225].

Despite those papers discuss the results of FDD methodologies on the same data set, not always the assumptions behind the analysis are the same. The main differences are related to the operation mode considered (cooling, heating, spring), the number and the type of faults analysed, the regime of operation considered (transient, non-transient). However, from the analysis of these works, some general considerations can be made:

- In most of the cases the analysis is performed for the summer period achieving high values of accuracy in diagnosing faults (over 90% of accuracy),
- The analysis is performed for data collected with sampling frequency of 1-min (original granularity of the dataset),
- Data-driven models used for characterizing the normal behaviour of the AHU lack of interpretability (SVM, ANN)
- In most of the cases, the fault diagnosis is performed through interpretable classifiers (decision trees, Bayesian networks).

The present chapter presents and discussed the development of an FDD tools starting from the ASHRAE RP-1312 data set. The main objective is to introduce an FDD methodology for AHU systems that is based on data analytics techniques, fully interpretable and rule-based. Indeed, the rule-based approach can satisfy the user need of simplicity and interpretability while the data-driven nature of the tool can enable the learning of system operational patterns automatically [13]. Another objective is also to reduce the granularity of the dataset while maintaining good performance in fault diagnosis. In fact, analyse data with a high sampling frequency could expose the FDD tool to instabilities when deployed for operating in real time (presence of punctual anomalies, missing values, sensor network latency).

In the proposed methodological framework of analysis, at the basis of the developed tool, two rule extraction methods (association rule mining, decision tree) are employed for conducting FDD analysis in AHU system by exploiting the reduction and transformation of multiple time series related to the operation variables of the system.

The next section presents the main motivations and research challenges related to the automatic extraction of interpretable rules in multivariate FDD problems and introduces the motivations and novelty of the proposed approach.

4.1.1 Motivations and novelty of the proposed approach

The complexity of an AHU system with multiple operational parameters and temporal interactions among them makes challenging the effective characterisation of its behaviour.

The operation of an AHU system is characterized by two major time-regimes, transient and non-transient respectively. The transient operation typically occurs when the AHU is started-up and is approaching the steady state conditions, or when it is shutdown or disturbed from its non-transient regime. These disturbances could be caused by either variation of thermal loads or by feedback controls and during transient periods some variables can exhibit strong variation in short time and a significant temporally lagged response respect to the control signals. In addition, the behaviour of an AHU system varies as its mode of operation changes during the day and the year i.e., off mode, heating mode, free cooling mode, and cooling mode. As a consequence, a robust data analytics-based FDD tool should be able to automatically determine the mode of operation of the system to prevent false alarms from being generated. For example, normal behaviour during summer season may be faulty if the system is operating in heating mode (winter season). In order to avoid this condition, FDD tools in AHU systems are characterized by a hierarchical architecture that makes it possible to exploit only the portion of knowledge that is consistent with the specific operation mode considered. In this perspective, when using data analytics-based FDD tools it is necessary for the training data to be exhaustive as possible for each operation mode.

However, given their complexity data analytics-based FDD tools often lack in interpretability. In this context, the use of rule-based methods for FDD can satisfy the user need of simplicity in terms of understanding the FDD tool, using it, commissioning it, integrating it with existing BAS and updating it. For this reason, great attention is paid on the application of advanced supervised and unsupervised rule extraction methods (i.e., decision trees, association rule mining) with reference to multivariate problems.

The operation of an AHU is a perfect case that can be effectively described through the analysis of multiple time series (defined as series data points indexed in time order) associated to each operational variable of the system. Considering that the majority of AHU operational data are gathered from continuous-time continuous-variable signals by high frequency sampling a suitable data reduction (aggregation in the time domain) and discretization (quantization of the signal value) become often necessary for knowledge extraction applications.

This is a challenging task considering that each variable has its own behaviour and distribution and, as a consequence, the optimal time aggregation and value discretization of the signal need to be identified with the aim of minimizing the information loss and of maximizing the mining performance. Such preparation of time series is an essential step in FDD methodologies based on rule extraction techniques (e.g., based on association rule mining algorithms or decision trees) that, in the literature, have been used for effectively mining co-occurrences or implications between discrete values and events in the time domain during HVAC operation [30,140,142,226].

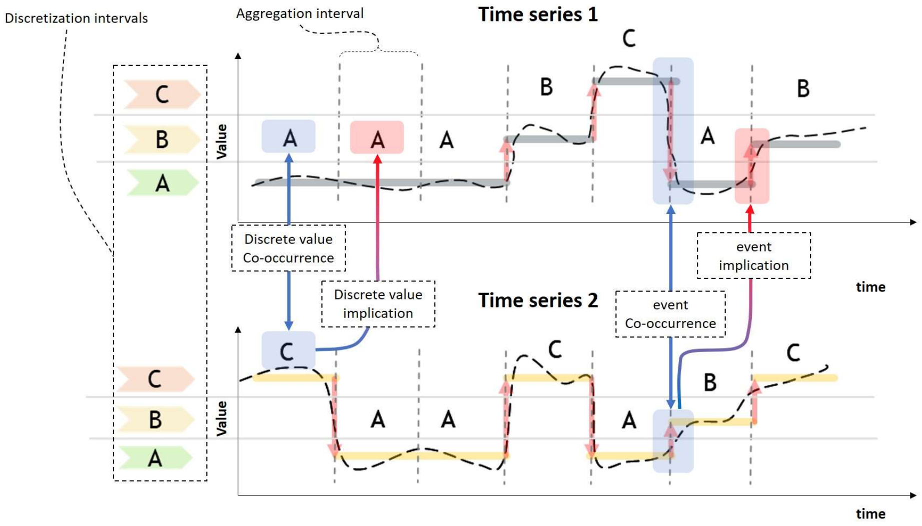


Fig. 63 - Graphical representation of co-occurrence and implication between discrete values and events among multiple time series

Fig. 63 depicts in graphical form the concepts, previously introduced, of discrete value, event (change of discrete value between two contiguous aggregation intervals), co-occurrence and implication of discrete values and events with reference to two time series encoded in symbols.

When multiple time series are considered, rule extraction techniques can be categorized in intra-transactional and inter-transactional respectively. The first type of extraction is aimed at discovering co-occurrences between discrete values and events that frequently happen at the same time among different time series

(Fig. 63). The second type of rule extraction is more complex, considering that in this case the occurrences of discrete values and events among different time series, are searched taking into account the existence of a time lag (Fig. 63). During transient period of AHU operation, the latter approach is particularly favourable in describing phenomena that are characterized by temporal dependences between discrete values and events representative of the system operation (e.g., change of status in fan speed and the corresponding effect on supply air temperature).

In order to develop an FDD tool capable to be flexible in relation to different conditions of operation in AHUs, two rule extraction methodologies tailored for both transient and non-transient periods are introduced. The developed framework is aimed at preventing anomalous running modes in AHUs which can lead to significant energy waste over time and/or discomfort conditions in the built environment.

The analysis relies on temporal abstraction as a pre-processing stage. Temporal abstraction is aimed at reducing and transforming time series in discrete-time and discrete-value signals through aggregation on the time axis and discretization of the value in order to perform the extraction of interesting co-occurrences and implications. To this purpose the adaptive Symbolic Aggregate approximation (described in section 2.1.1.2) algorithm is used.

Furthermore, by means of temporal association rules (described in section 2.1.3.2.2.1) in form of IF-THEN implications strong relations between events (i.e., change of discrete value between contiguous aggregation intervals) are automatically mined in the transient period of AHU operation (i.e., start-up phase) considering an intra-transactional approach for characterizing the fault-free behaviour of the system. Similarly, during the non-transient period of operation a set of classification trees (described in section 2.1.3.1.1.1) are developed for extracting reference patterns in form of decision rules. Potential faulty conditions are then detected when the discovered association and decision rules are violated over time. Successively, the identified anomalous patterns (during the non-transient period) are exploited for performing a diagnosis of the most probable faults associated to a specific kind of rule violation by means of a classification algorithm (described in section 2.1.3.1.1.1). In particular the developed FDD tool is capable of detecting up to 11 typical faults (of valves, fans and dampers) in AHUs with an overall accuracy of 90% leveraging on a set of intuitive and interpretable decision rules.

By combining and integrating data analytics techniques, the conceived FDD methodology introduces the following innovative aspects:

- An adaptive process of data reduction and transformation is employed to enhance the knowledge extraction from multiple time series. In complex systems as AHUs, the number of monitored variables and their sampling frequencies could be very high. Extracting only key information from large data set is essential for reducing redundancy, complexity and computational cost. The developed methodology makes it possible to achieve good performance in FDD (comparable to other studies focused on the same dataset [55,138,220–225]) leveraging only on the analysis of significant discrete intervals of the operational variables over time.

- The start-up period of AHU operation is isolated and treated separately by developing a tailored analytics module (instead of being filtered out as happened in other studies focused on the same dataset [55,138,220–225]). During transient period of operation time lags occur for example between a change of status in fan speed and the corresponding effect on supply air temperature. Such a condition could compromise the assumption of discrete value and event co-occurrence when the reference behaviour of the system is characterised. For this reason, temporal association rules are extracted, following an intra-transactional approach, for discovering associations between events during transient periods, across multiple time series, that frequently happens within a time lag.
- The characterization of normal behaviour during the non-transient period is completely automated and performed by using a set of estimation models based on decision trees. In comparison to other studies focused on the same dataset [55,138,220–225], the reference behaviour of the AHU is evaluated estimating the most probable discrete value of each influencing operational variable in relation with all the other ones monitored. In that way, all the existing relations between variables are exploited through several estimation models for detecting potential faulty conditions. Such approach introduces high flexibility and generalizability in the formulation of the FDD problem.
- A fault diagnosis during non-transient period of AHU operation is performed by employing a decision tree capable to extract rules for the classification of typical faults. The diagnosis process exploits the residuals evaluated by means of a set of estimation models (i.e., difference between real and expected discrete value of influencing variables) as input attributes for the classification of the most probable faults.

As previously stated, several studies considered the RP-1312 data set in the analysis, achieving an accuracy in fault diagnosing over 90%. As a consequence, the main objective of this analysis is not to improve the (already high) FDD performance achieved on the RP-1312 dataset, but rather to demonstrate the opportunity to achieve high performance as well through a fully interpretable and simplified data-driven approach, based on rule extraction techniques.

The remining sections of chapter 4 are organised as follow. Section 4.1.2 presents the case study analysed for developing the FDD tool; then section 4.1.3 provides a description of the proposed methodology. Successively, section 4.1.4 presents the results obtained from the application of the methodology. Eventually, Section 4.1.5 discusses the results and contains the concluding remarks related to this specific DSS application.

4.1.2 Case study used for developing the FDD tool at system level

In order to test the validity and the effectiveness of the developed FDD tool, operational data related to two AHUs collected in the framework of the Research Project ASHRAE RP-1312 [136] are analysed. The system investigated is a Variable Air Volume (VAV) AHU. A VAV system is able to modulate the air flow rate according to the variation of the building load and it is typically made up of 4 subsystem controllers acting on supply air temperature, dampers and valves, supply air static pressure and return air flow rate. Specifically, the control logic maintains the supply air temperature set-point acting on damper and valve positions, according to the mode of operation (i.e. heating, cooling with partial mixing of outdoor air, cooling with 100% of outdoor air, cooling with minimum outdoor air).

Furthermore, even the static pressure of the supplied air and the difference between the supply and return air flow rate is controlled. The return air flow rate is modulated acting on the mixing dampers and the return fan speed, while the system maintains the supply air static pressure set point. As a result, the difference between the supply and return air flow rate is kept constant [227].

The analysed dataset is particularly interesting and includes several running conditions for two AHUs in faulty and fault-free operation. The faulty operation is obtained by artificially implementing a number of different faults. The site, where monitoring data were collected, is a test facility simulating a typical commercial building occupancy schedule.

The monitoring data were gathered from two AHUs of the facility (AHU-A and B), which are perfectly identical form technical and operational points of view and serve specular zones. The zones served by AHU-A and B face east and west orientation respectively in order to be comparable also under thermal load aspects. The AHUs are characterised by a mixing chamber to mix return air with outdoor air by means of dampers. Each AHU is equipped with heating and cooling coils and VAV devices to locally adjust the supply air temperature. However, the control volume considered excludes the local VAV devices.

Fig. 64 shows a schematic configuration of the system with the indication of monitored variables (description is provided in Table 17).

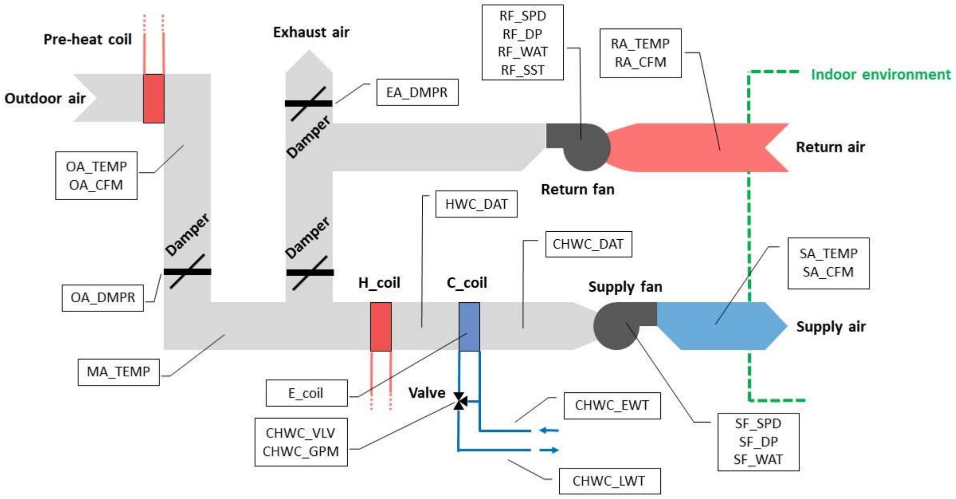


Fig. 64 - Scheme of the AHU analysed (refer to Table 17 for variable encoding)

In the context of the ASHARE project different faults were implemented one per time, each for a whole day, only in the AHU-A, in order to analyse the effects of each fault independently. The AHU-B was always run in fault-free conditions to have a reference of the normal operation. The data collection was conducted over three seasons and only the monitoring data of the summer season are in the following considered.

The dataset consists of multiple time series (one for each variable monitored) with a length of 33 days and a sampling time of 1 minute. In particular, 22 out of 33 days are tagged as fault-free days while the remaining 11 days correspond to different faulty conditions. Table 16 reports the number of fault-free and faulty days, the description of each fault and the tags used for labelling each day included in the monitoring campaign.

Table 16 - Tags and descriptions of faults

Fault Tag	Description	Number of days
CCVS15	Cooling coil valve stuck at 15%	1
CCVS65	Cooling coil valve stuck at 65%	1
CCVSFC	Cooling coil valve stuck fully closed	1
CCVSFO	Cooling coil valve stuck open	1
EASFC	Exhaust air damper stuck fully closed	1
EASFO	Exhaust air damper stuck fully open	1
Normal	Normal operation	22
OAS45	Outdoor air damper stuck 45	1
OAS55	Outdoor air damper stuck 55	1
OASFC	Outdoor air damper stuck fully closed	1
RFCF	Return fan complete failure	1
RFF30	Return fan at fixed speed (30%)	1

A feature selection is preliminarily performed on the basis of expert system considerations to focus the analysis only on relevant variables.

As a result, the variables considered for the implementation of the FDD tool are: electrical load, pressure drop and speed of fans, flow rate and temperature of the air measured in different parts of the system, damper position, valve position, water flow rate and energy exchanged in the cooling coil.

Table 17 reports the list of the 23 variables considered for the analysis with the specification of variable labels, descriptions, ID n° and unit of measure.

Table 17 - List of variables considered in the analysis

Variable	Description	ID n°	Unit
SF_WAT	Supply fan power	1	W
RF_WAT	Return fan power	2	W
SA_CFM	Supply air flow rate	3	m^3/h
RA_CFM	Return air flow rate	4	m^3/h
OA_CFM	Outdoor air flow rate	5	m^3/h
SA_TEMP	Supply air temperature	6	°C
MA_TEMP	Mixed air temperature	7	°C
RA_TEMP	Return air temperature	8	°C
HWC_DAT	Heating coil air temperature	9	°C
CHWC_DAT	Cooling coil air temperature	10	°C
SF_DP	Supply fan pressure drop	11	Pa
RF_DP	Return fan pressure drop	12	Pa
SF_SPD	Supply fan speed	13	%
RF_SPD	Return fan speed	14	%
OA_TEMP	Outdoor air temperature	15	°C
CHWC_EWT	Cooling coil input water temperature	16	°C
CHWC_LWT	Cooling coil output water temperature	17	°C
CHWC_GPM	Cooling coil water flow rate	18	m^3/h
E_ccoil	Cooling coil power	19	kW
CHWC_VLV	Cooling coil valve position	20	%
EA_DMPR	Exhaust air damper position	21	%
OA_DMPR	Outdoor air damper position	22	%
RF_SST	Return fan start/stop signal	23	-

For the application of the proposed FDD methodology, the available data sample is split into two datasets. The first one is used for the characterization of the normal operative condition of the system, while the latter is used for the fault detection and diagnosis. The first dataset is composed of 20 days tagged as “Normal” (training dataset), while the second by the rest of the days including 2 “Normal” days and 11 “Faulty” days (testing dataset).

4.1.3 Implemented methodology for the detection and diagnosis of faults in AHU

The methodology relies on the application of both supervised and unsupervised algorithms, to perform robust fault detection and diagnosis in AHUs.

The framework unfolds over different stages as shown in Fig. 65. Two different analytics modules are proposed for developing an FDD tool tailored for both transient and non-transient conditions of the AHU operation. For that purpose, in the methodological framework, a data segmentation phase is preliminarily carried out in order to split the data according to the regime of operation they belong to (i.e. transient or non-transient). In the following sections are then described the pre-processing analysis applied to the entire dataset and the two analytics modules tailored for transient and non-transient periods.

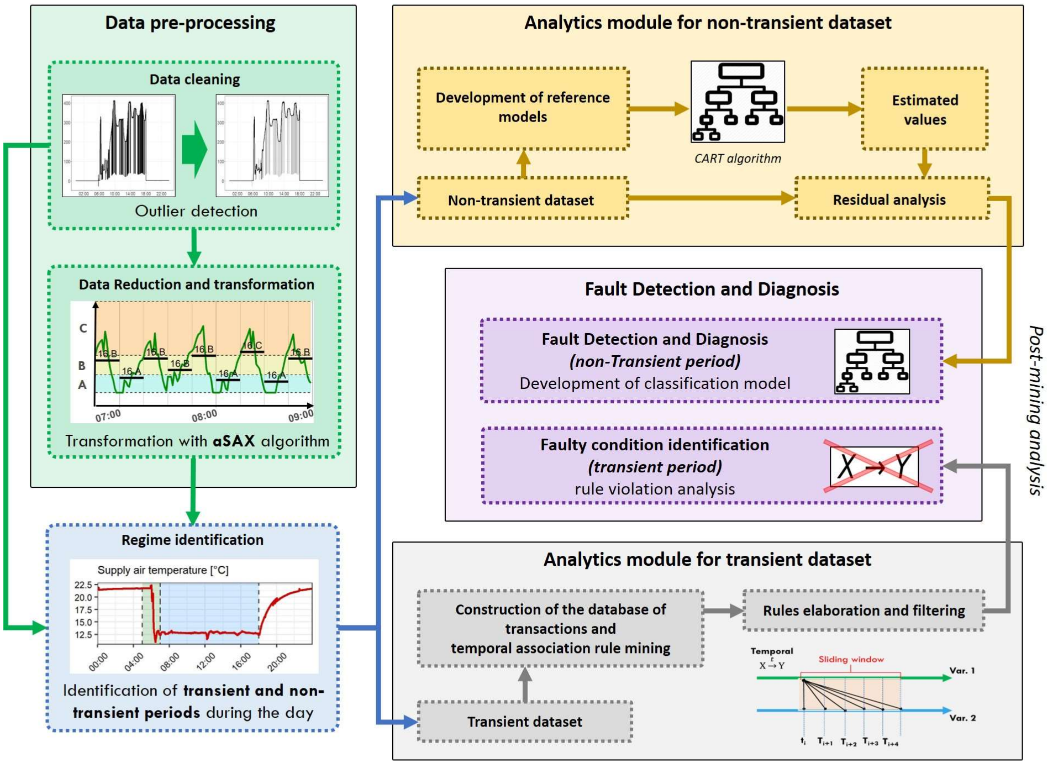

Data pre-processing
Data cleaning
Outlier detection
Data Reduction and transformation
C
B
A
16 B
16 A
16 C
07:00
08:00
09:00
Transformation with aSAX algorithm
Regime identification
Supply air temperature [°C]
22.5
20.0
17.5
15.0
12.5
00:00
04:00
08:00
12:00
16:00
20:00
Identification of transient and non-transient periods during the day
Analytics module for non-transient dataset
Development of reference models
CART algorithm
Estimated values
Non-transient dataset
Residual analysis
Fault Detection and Diagnosis
Fault Detection and Diagnosis
(non-Transient period)
Development of classification model
Faulty condition identification
(transient period)
rule violation analysis
X → Y
Post-mining analysis
Analytics module for transient dataset
Construction of the database of transactions and temporal association rule mining
Rules elaboration and filtering
Transient dataset
Temporal
X → Y
Sliding window
Var. 1
Var. 2

Fig. 65 - General framework of the analysis

4.1.3.1 Data pre-processing stage

The pre-processing stage consists of three main tasks i.e., cleaning, reduction, and transformation, typically accomplished for preparing the data sets. In detail, outlier detection and replacement are firstly performed (for each time series) by using the Hampel filter method [228]. 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 replaced with the median.

The monitoring data are available in time-series with a sampling time of 1-minute, which would make the analysis onerous to be performed. For this reason, in a successive step a data reduction and transformation process are performed by means of the adaptive Symbolic Aggregate Approximation (aSAX) method [229] (described in section 2.1.1.2). This algorithm is employed for reducing the time series through a piecewise technique aggregating data with a fixed length window from 1 minute to 15 minutes and then for transforming it into a symbolic string. The objective is to maximise data compression and minimise the complexity of the time series while preserving important information.

Fig. 66 reports an example of data reduction and transformation through aSAX algorithm for a portion of the time series related to the variable encoded with the ID n° 16 according to Table 17 (i.e., cooling coil input water temperature). The figure shows the time series after the application of the Hampel filter (green curve) and the time series in form of constant approximated piecewise (black lines). Furthermore, Fig. 66 also shows the result of the aSAX transformation of the time series into a symbolic string. The variable can assume three discrete values encoded with the symbols 16.A, 16.B or 16.C according to the region the piecewise segments of 15-min fall in.

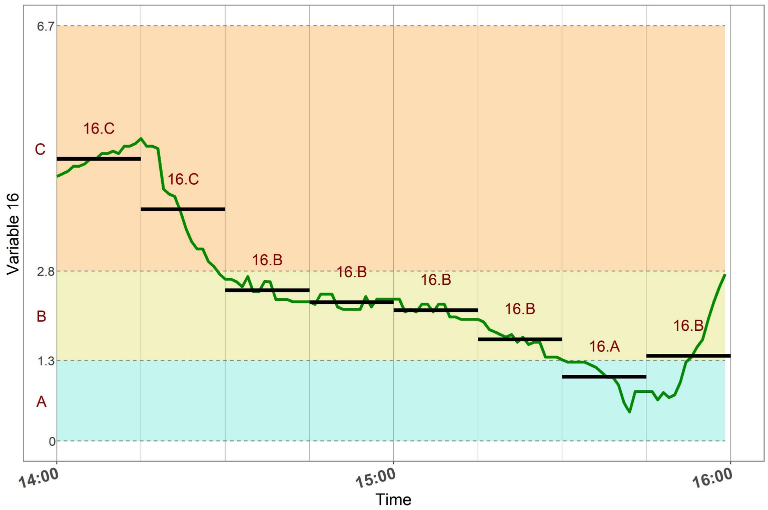


Fig. 66 - Example of aSAX transformation for a numerical variable

The obtained symbol sequence is 16.C-16.C-16.B-16.B-16.B-16.B-16.A-16.B from which it is possible to infer that the original time series is characterized by changes in the pattern at times 14:30, 15:30, 15:45 that in the present methodology are intended as events.

As a result, time series are transformed in discrete-time discrete-value sequences of equidistant symbols making it possible to extract events from them.

4.1.3.2 Regime identification

At this stage, a regime identification is performed, at daily scale, to detect when transient and non-transient conditions typically occur during the AHU operation.

To this purpose an automatic regime detector is used to identify the transient period and separate it from the non-transient one. The details of the detector used are the same as that reported in [138,230]. The transient identification is performed on data with sampling time of 1 minute, specifically analysing the cooling coil valve position, the supply air temperature, supply fan speed and the supply air static pressure. Then the frequency of transient data points during the day is evaluated for each 15-min aggregation interval derived from the data reduction phase (Fig. 67).

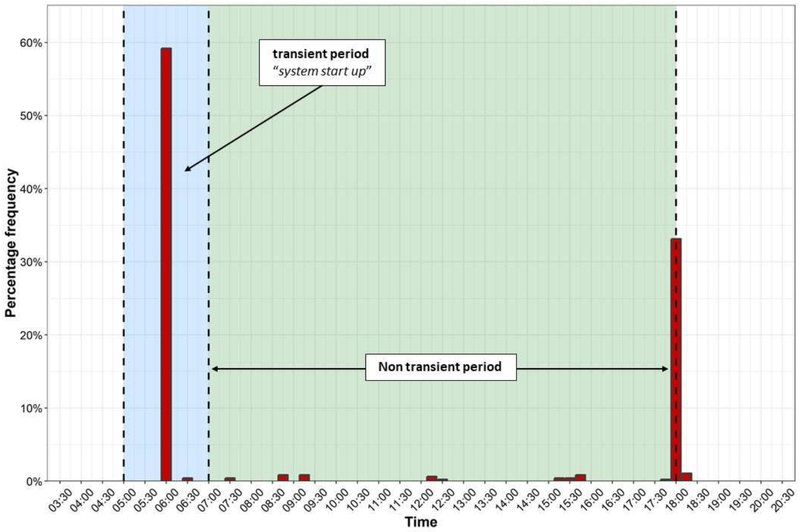


Fig. 67 - Identification of the transient period

Thanks to this analysis it is possible to establish during which aggregation interval of 15-min (of the reduced daily time series) a transient condition has the highest frequency of occurrence.

Starting from such aggregation interval of 15 min, the transient period is evaluated considering a time window of two hours (i.e., blue area of the plot) that includes one hour before and one hour later the aggregation interval considered (Fig. 67).

As can be noticed from Fig. 67, transients occur at the start-up and the shut-down. Among the two transient periods, only the start-up transient is investigated because during this period the system dynamics affects the successive operation, while in the other case the system is thereafter turned off.

As a result, the non-transient period is supposed to start at the end of the start-up time interval and ends when the system is turned off.

Therefore, excluding the night hours, during which the AHU is certainly not operated, the dataset is segmented as follows:

- From 05:00 to 07:00: transient period labelled as "*system start-up*";
- From 07:00 to 18:00: non-transient period.

4.1.3.3 Implemented FDD methodology for the transient period

As explained in section 4.1.3.2 the segmentation phase makes it possible to distinguish transient from non-transient periods during the day.

The main flow of FDD research has been carried out in a steady-state approach [55,136,140,231,232], because operating characteristics in a steady state is relatively more credible and reproducible than in a transient state [231]. Transient data are characterised by great variation in the time domain and require specific machine learning and data mining algorithms to be employed for properly reflect the system dynamics. The herein proposed FDD tool provides a tailored methodology for such condition of operation.

An overall procedure is developed to obtain temporal association rules that are representative of frequent relationships between events in multiple time series, using a time window and a time lag.

Temporal association rules are an interesting extension of association rules that include a temporal constraint, which leads to different forms of IF-THEN implication over time. When an event leads to the occurrence of another event, there may be causal relationship or certain correlation between them. The corresponding mining purpose is to find out the reference fault-free association rules between events and time in a temporal transaction dataset, whose violation can suggest the presence of faulty conditions during the start-up period of the AHU system. The extraction makes it possible to find those sequences of events that appear many times among monitored fault-free days and have a high rate of occurrence (i.e., reference rules).

The reference association rules are searched in the 20 days tagged as fault-free (training dataset) while the remaining 2 fault-free days and 11 faulty days (testing dataset) are used in the successive fault detection phase.

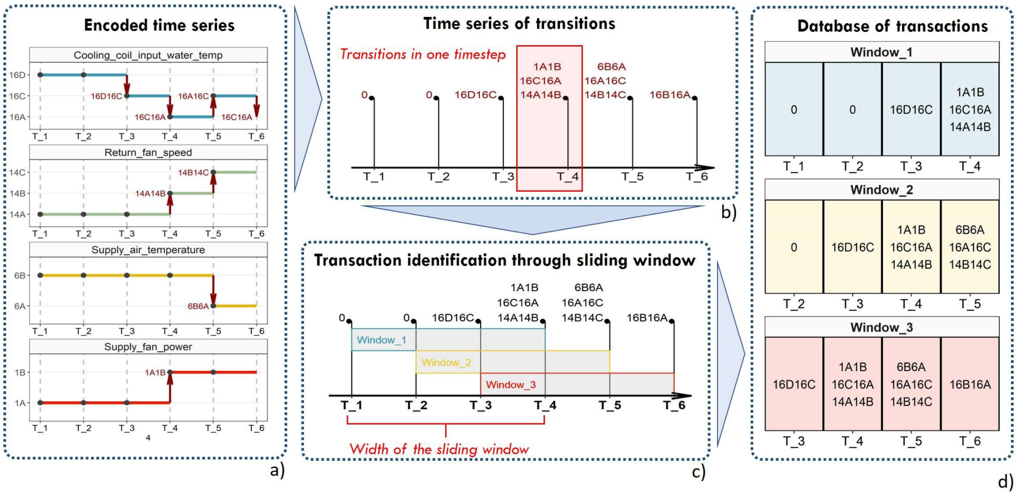


Fig. 68 - Procedure for the construction of the database of transactions

Before extracting reference temporal association rules from data, it is necessary to create the database of transactions T following the framework reported in Fig. 68.

The first step consists of putting together all the transitions that occur in each time series into a unique multivariate time series of transitions.

In particular, according to the symbolic transformation performed during the pre-processing stage a transition in a time series is a kind of event that corresponds to the change of symbol (i.e., encoded discrete values of the variable) in a specific timestep across two consecutive aggregation intervals.

As an example, Fig. 68 (a) shows six timesteps of four time series (i.e., cooling coil input water temperature, return fan speed, supply air temperature, supply fan power). The time series "supply fan power" corresponds to the operation variable of the AHU encoded with the ID n° 1 and assumes only two discrete values (encoded with the symbols 1A and 1B) along the six timesteps considered. In the same way the time series "*return fan speed*" that corresponds to the operation variable of the AHU encoded with the ID n ° 14, assumes three discrete values (encoded with the symbols 14A, 14B and 14C) among the six timesteps. If two consecutives aggregation intervals are encoded with the same symbol no transition (i.e., event) is detected. Otherwise, during a specific timestep, a transition (i.e., event) is encoded reporting the ID n° of the variable and the two symbols included in the change of discrete value. For example, according to Fig. 68 (a), at the first timestep T_1 for any time series, a transition does not occur and then 0 is stored in the time series of transitions (Fig. 68 (b)). On the contrary, at the fourth timestep T_4, a transition occurs for the time series 1, 14 and 16. In particular, for time series 1 and 14, occurs a change from symbol "A" to symbol "B" (events encoded as "1A1B" and "14A14B" respectively) while for time series 16 the variable changes symbol from "C" to "A" (event encoded as "16C16A"). Once the encoded events are stored in the multivariate time series of transitions (Fig. 68 (b)), the database of transactions is constructed by chunking this time series considering a fixed-length sliding time window (Fig. 68 (c)). Fig. 68 (d) shows how the encoded transitions for each timestep are stored in the database of transactions. For example, assuming a sliding window that includes four timesteps, the database T can be represented by a $4\,x\,n$ transition matrix where n corresponds to the maximum number of sliding windows which can be contained in the time series of transitions.

Considering that the time windows are sliding a timestep by time, two consecutive rows in the database T (Fig. 68 (d)) differ only for a single item. As a reference considering a time series of transitions with 6 timesteps and a sliding window that includes 4 timesteps, the database of transactions is a $4\,x\,3$ transition matrix given that the maximum number of complete time windows is equal to 3 (Fig. 68 (d)). After the construction of the database T, the temporal association rules are searched among transactions.

The cSpade algorithm [49] is selected for the extraction of the rules from the inter-transactional database setting in advance three fundamental parameters: minimum confidence, minimum support, and maximum time lag between antecedent and consequent item sets (equal to the sliding window length).

According to the proposed methodology, the first two parameters (i.e., confidence and support) should be as high as possible, to ensure that the extracted rules are much frequent as possible and then representative of the normal behaviour of the system.

Once the reference rule set is identified, it is used for detecting the presence of potential faults in a testing dataset.

In particular, a temporal association rule is expressed as a logical IF-THEN implication where the presence of an event (i.e., antecedent) implies the occurrence of another event (i.e., consequent) within a certain time lag. According to this formulation, three potential violations can occur when such rules are applied on a testing set of data:

i) absence of the antecedent itemset,
ii) absence of consequent itemset,
iii) absence of antecedent and consequent item sets.

In that perspective, the violation analysis helps physical interpretation of rules making it possible to assess their sensitivity to the presence of specific faults or group of them.

4.1.3.4 Implemented FDD methodology for the non-transient period

The methodology employed for performing an FDD analysis during non-transient period relies on three fundamental phases that can be generalized as follows:

- Development of reference models through classification trees, representative of the normal behaviour (fault-free condition) of the system under analysis;
- Comparison between the estimated behaviour of the system and the actual one (i.e., evaluation of model residuals) for detecting potential faulty conditions;
- Analysis of the model residuals for diagnosing the most probable cause associated to a specific fault (fault diagnosis).

The first step of the process consists of a robust characterization of the fault-free operation of the AHU during the non-transient period (i.e., from 07:00 to 18:00). To this purpose, several estimation models (i.e., classification trees) are developed on a portion of the available non-transient dataset. In detail 20 days tagged as fault-free are considered at this stage (training dataset) while the remaining 2 fault-free days and 11 faulty days (testing dataset) are used in the successive diagnostic phase.

For the development of the estimation models (i.e., classification trees), all the variables concerning the operation of the AHU (e.g., supply fan power, return fan power, supply air flow rate) are selected once at a time as target attribute while the remaining ones are used as input attributes. However, features related to external forcing variables to the AHU system (i.e., cooling coil input water temperature, outdoor air temperature) are used only as input attributes.

In that way, 21 classification trees are developed for providing a robust benchmark of the fault-free operation. To that purpose, a classification tree based on recursive partitioning algorithm (described in section 2.1.3.1.1.1) is employed as a supervised classifier. The developed classification trees estimate for each target variable and for each 15-min aggregation interval included in the non-transient period, the most probable discrete value (encoded as symbol) according to the relationship that exists between all the input variables and the dependent attribute. Successively all the classification trees developed are put together in the same estimation layer as shown in Fig. 69.

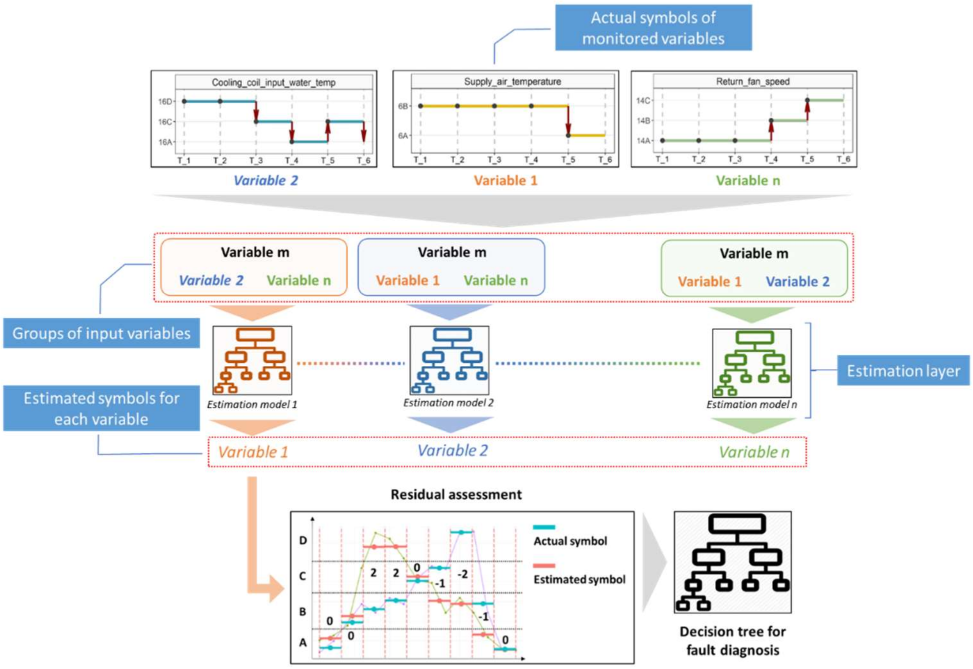


Fig. 69 - Analytics module for the non-transient period

At this stage, the estimation process can be summarized as follow:

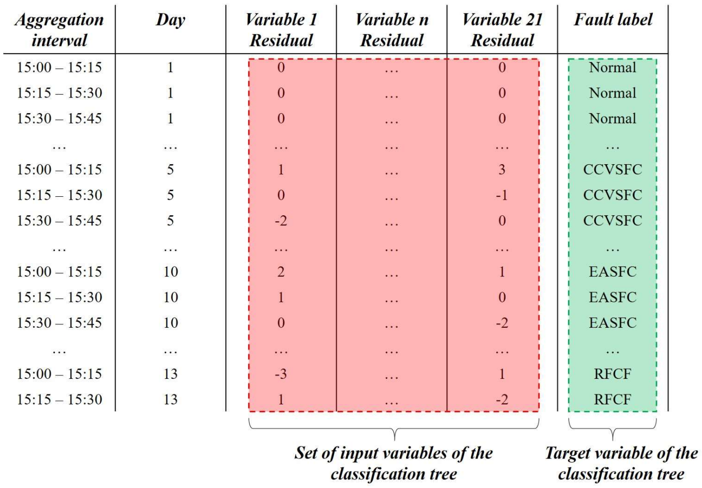
- At each aggregation interval (i.e., 15 min.) the monitored variables are encoded into symbols through the aSAX method (i.e., pre-processing stage);
- The set of encoded variables goes through the estimation layer (that consists of 21 classification trees) providing an estimation of each target variable for the considered aggregation interval;
- The actual symbols are compared with the estimated ones.

The latter step consists in the evaluation of the model residuals (i.e., the difference between actual and predicted values).

In particular, the difference between two equal symbols is assumed to be zero, while the residual differs from zero if the symbols are at least one alphabet apart. For example, if the estimated and actual symbol for a variable is equal to “A” and “B” respectively, the residual between those symbolic discrete-values is equal to 1 (Fig. 69).

Considering that the estimation models are trained on fault-free data, at the end of the estimation process it is possible to assess how much the input data differ from the reference fault-free behaviour of the AHU through the analysis of residuals. Understanding which variables are out of range and assessing the severity of those deviations enables the detection of possible faulty conditions. In order to test the developed FDD tool, all the days excluded from the training set of the reference models (i.e., 2 fault-free days and 11 faulty days) are considered. In particular, each day included in the testing dataset is labelled as "Normal" or with the tag of one of the faults reported in Table 16.

The time series of the 13 days are pre-processed (aggregated in intervals of 15-min and encoded in symbols) and put through the estimation layer (i.e., 21 classification trees) generating a dataset of residuals as shown in Fig. 70.



Aggregation interval	*Day*	*Variable 1 Residual*	*Variable n Residual*	*Variable 21 Residual*	*Fault label*
15:00 – 15:15	1	0	…	0	Normal
15:15 – 15:30	1	0	…	0	Normal
15:30 – 15:45	1	0	…	0	Normal
…	…	…	…	…	…
15:00 – 15:15	5	1	…	3	CCVSFC
15:15 – 15:30	5	0	…	-1	CCVSFC
15:30 – 15:45	5	-2	…	0	CCVSFC
…	…	…	…	…	…
15:00 – 15:15	10	2	…	1	EASFC
15:15 – 15:30	10	1	…	0	EASFC
15:30 – 15:45	10	0	…	-2	EASFC
…	…	…	…	…	…
15:00 – 15:15	13	-3	…	1	RFCF
15:15 – 15:30	13	1	…	-2	RFCF

Set of input variables of the classification tree
Target variable of the classification tree

Fig. 70 - Structure of the database used for developing the classification tree of fault diagnosis

At this stage, a further classification tree is developed to predict the label of each faulty or Normal condition (Fig. 70) for performing the fault diagnosis. This classification tree estimates the most probable label (e.g., CCVSFC, EASFC, RFCF or Normal) according to the residuals evaluated for each variable as an outcome of the estimation layer.

In the dataset reported in Fig. 70 the target variable is the fault tag, and the same tag is assigned to all of the 44 aggregation intervals of 15-min that belong to the same day (included in the 11 hours of "non transient" operation of the AHU from 7:00 to 18:00), generating a total amount of 572 instances on which develop the classifier.

The developed FDD tool is then trained and tested on real data of an AHU operated in cooling mode for 33 non consecutive days during the summer season (22 "normal" days and 11 "faulty" days). Therefore, the decision and association rules extracted through the proposed methodology can be considered valid only for the operation mode under consideration. In this perspective, rule-based tools can be easily integrated in FDD systems with hierarchical architecture capable to exploit only the useful knowledge during specific conditions. For instance, the use of automatic detector makes it possible to call specific sets of rules depending on the operating mode of the AHU: off mode, heating mode, free cooling mode, and mechanical cooling mode [233].

4.1.4 Results obtained from fault detection and diagnosis analysis

4.1.4.1 Data pre-processing results

According to the methodological framework introduced in Section 4.1.3, a data pre-processing stage is preliminarily implemented. Firstly, outliers are filtered out by implementing the Hampel filter on the 1-minute time series. For each sample of the time series, the filter computes the standard deviation and the median of a window composed of the current sample and $\frac{Len-1}{2}$ adjacent samples on each side of the current sample. Len is the window length and is set equal to 31 minutes.

After data cleaning a data reduction is performed by means of a piecewise aggregate approximation (PAA) process with the aim of approximating the time series of each considered variable to the mean value calculated in non-overlapped time intervals with a fixed length of 15 min. Successively, the trasformation of the reduced variables in symbols is carried out by implementing the aSAX algorithm [229].

The algorithm is initialised for each variable by identifying the number of symbols (i.e., discretization intervals) and the initial positions of the breakpoints (i.e., borders of the discretization intervals) with a hierarchical cluster analysis using the Ward linkage method [48] (described in section 2.1.3.2.1). Through the clustering algorithm, it is possible to obtain the optimal number of discretization intervals (i.e., number of symbols) by computing several cluster validation metrics. This process is completely automated and performed through Nbclust package [234] available in the statistical software R. The number of discretization intervals is constrained from 2 to 4 considering only data referred to the period of operation of the system (i.e., ON-hours of the system).

When the optimal positions of the adaptive breakpoints are found and each variable is encoded in symbols, the operation conditions of the AHU are considered fully characterised. Successively, data related to OFF-hours of the system operation are analysed to find possible additional intervals. In particular, if during OFF-hours a variable typically assumes values that are out of the identified ranges of discretization, a new lower or upper half-open interval is appended to the previous ones.

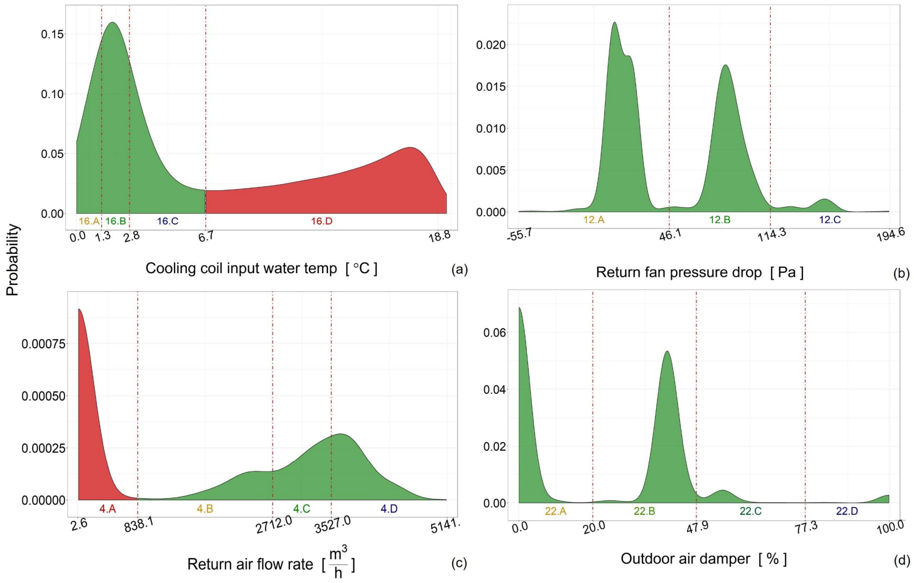


Fig. 71 - Distributions and breakpoint identification for some variables

Fig. 71 shows the encoding process performed for 4 variables (i.e., cooling coil input water temperature, return fan pressure drop, return air flow rate, outdoor air damper) randomly selected from the set of inputs. It can be observed that for two variables an additional OFF-hours discretization interval (i.e., red area of the distributions in Fig. 71 (a) and (c)) is added to the other ranges of values for the symbol encoding (i.e., ID n° = 16, symbol = D and ID n° = 4, symbol = A).

As a reference, Table (a) in Appendix A summarizes the transformation results obtained, with the specification of the numerical range corresponding to each symbol for all the analysed operational variables.

At this stage, transient and non-transient periods are identified, and the data set is consequently segmented. In particular, it is labelled as transient start-up period the time interval between 5:00 and 7:00, while the period from 7:00 to 18:00 is considered as non-transient period.

4.1.4.2 Fault detection analysis for the transient period (system start-up)

The encoded time series are analysed for extracting temporal association rules in the start-up period of system operation. In detail, the transitions of the variables (i.e., change from a symbolic discrete-value to another one) are preliminarily encoded and the inter-transactional database is created considering a sliding window of 60 minutes. The width of the sliding window is chosen to be large enough to include any effect of the system dynamics, but tight enough to ensure that the occurrence of a consequent itemset is related to a physics-based implication with its antecedent itemset.

Considering that the fault detection methodology is conceived for extracting reference association rules of normal operation, the inter-transactional database is created from the fault-free dataset, by selecting rules with high values of support and confidence. The ARM process is carried out by implementing the cSpade algorithm [49]. The analysis implemented in R [235], including the rule extraction phase which is performed by using the “cSpade” function of the “arules” package [236].

Considering that the application of the cSpade algorithm is conducted on the inter-transactional database, the values of support and confidence associated with the rules extracted are evaluated according to formulations reported in section 2.1.3.2.2 (Eq. 11 and Eq. 13) and introduced in [50,237].

Typically, the main issue related to association rules mining consists in handling and filtering a large number of rules extracted and eventually identify those that are of interest [20]. To tackle this problem and facilitate the mining of useful knowledge from extracted rules, a post-mining phase is performed.

The post-mining phase is aimed at solving various practical issues, such as interestingness, redundancy, generalization, visualization and interpretability of association rules.

To this purpose additional quality metrics are introduced: the daily support of the rule (i.e. SUPP.DAY) calculated for both fault-free (SUPP.DAY$_{\text{NORMAL}}$) and the faulty days (SUPP.DAY$_{\text{FAULTY}}$) and the actual time lag between the antecedent and consequent of a rule (ACTUAL TIME LAG). In more detail, the SUPP.DAY$_{\text{NORMAL}}$ is defined as the percentage of fault-free days during which a single association rule (R$_i$) occurred, while SUPP.DAY$_{\text{FAULTY}}$ is calculated for the faulty days (Eq. 24 and Eq. 25).

$$SUPP.DAY_{NORMAL}, R_i = \frac{\text{N° of Free−fault days with of the occurrence of the rule R}_i}{\text{Tot. N° of Free−fault days}}$$

Eq. 24

$$SUPP.DAY_{FAULTY}, R_i = \frac{\text{N° of Faulty days with of the occurrence of the rule R}_i}{\text{Tot. N° of Faulty days}}$$

Eq. 25

The ACTUAL TIME LAG is introduced to evaluate the most frequent temporal distance between the first occurrence of an antecedent and the last occurrence of the corresponding consequent of a specific rule. As a consequence, even though the rules are searched with a sliding window of 60 minutes, the user can have a feedback about the most frequent time interval within a consequent itemset occurs given the presence of its antecedent itemset.

The ACTUAL TIME LAG is calculated for each rule by computing the cumulative frequency of occurrences of the temporal distance between antecedent and consequent. For each rule a cumulated frequency threshold of 80% is considered in order to evaluate this metric.

Fig. 72 shows the frequency distribution of the ACTUAL TIME LAG for two rules. The rule on the left (i.e., rule 1077) occurs for more than the 80% of the time with an actual time lag between the antecedent itemset and consequent itemset of 15 minutes, while for the rule on the right (i.e., rule 15268) the 80% of occurrences has a characteristic time lag lower or equal to 30 minutes.

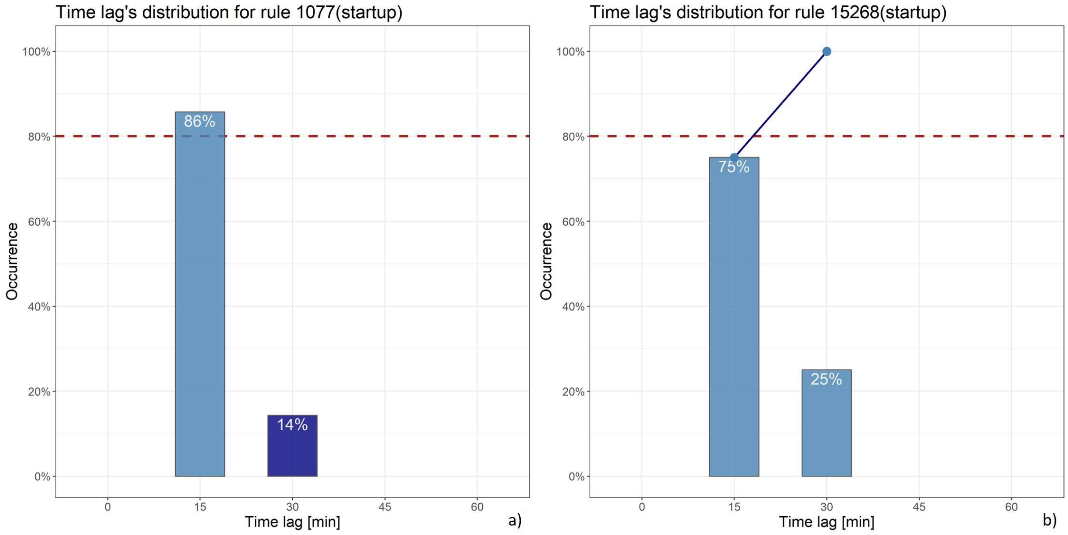


Fig. 72 - Distribution of the time lags for rule 1077 (a) and rule 15268 (b) – (refer to Table (b) in Appendix B for the description of the rules)

At this stage, more than 15,000 rules are extracted from the start-up dataset of fault-free days, assuming minimum support and minimum confidence equal to 0.7 and not including drivers of system's operation as potential consequent events (i.e. outdoor air temperature and cooling coil input water temperature).

After the rule extraction, the values of support and confidence are recalculated considering only the occurrences of each rule within the evaluated ACTUAL TIME LAG (instead of the window of 60-min), reducing the set of rules to 7,419 rules.

Since the rules extracted should be representative of the fault-free operation of the system, only the rules, which in the testing dataset frequently occur in normal days and rarely in the faulty ones, are of interest for the problem under investigation. To this purpose, after the application of the 7,419 temporal association rules to the testing dataset, only the rules with a $\text{SUPP.DAY}_{\text{NORMAL}}$ equal to 1 (i.e., the rule occurring for each day labelled as "normal" included in the testing dataset) and a maximum value of $\text{SUPP.DAY}_{\text{FAULTY}}$ equal to 0.3 are considered with the final result of obtaining 465 reference rules (SUPP.DAY values are set by the user).

As a general approach, the parameters are set in order to obtain a limited number of interesting rules, which respect the following conditions i) each rule occurs during fault-free condition with high support and confidence, ii) each rule has high probability to be violated during faulty conditions regardless from the fault type. In this perspective, general rules that are sensitive to more fault types at the same time are preferred to those violated only for specific faults.

However, according to ASHRAE project RP-1312, during the faulty day tagged as CCVSFO (i.e., cooling coil valve stuck open), the blockage of the cooling valve in fully open position was implemented from 8:00 to 18:00 and hence out of the start-up period of the system. For this reason, the day tagged as CCVSFO is not considered in the calculation of $\text{SUPP.DAY}_{\text{FAULTY}}$.

The introduced metrics allow an enhanced comprehension of the rule set, making it possible to discriminate rules with high support and confidence occurring during both fault-free and faulty days, from the rules, robust as well, occurring only during the normal operation of the system.

Fig. 73 shows for each day in the testing dataset (composed by 11 different faulty days and 2 Normal days) the percentage of rules (out of the 465 considered) which occur and/or are violated, with specification of the kind of violation detected. In particular the label "presence" indicates that the rule occurred with its antecedent and consequent while the labels "*antecedent*", "*consequent*" and "*absence*" indicate three different types of violation. In detail, the label "*antecedent*" denotes that a rule is violated because of the only presence of the antecedent; the label "*consequent*" indicates that a rule is violated because of the only presence of the consequent; the label "*absence*" indicates the complete violation of a rule because of the absence of both antecedent and consequent. The characterisation of the rules in terms of type of violation helps the interpretation of the path which determines a specific fault. In fact, the presence of the only antecedent, the only consequent, rather than the absence of both item sets, correspond to different behaviours of the system in relation to the presence of the considered faults.

The results obtained can be described according to the severity of rule violation for each day representative of a specific fault implementation or normal operation. To this purpose four different groups of days can be identified and in the following described.

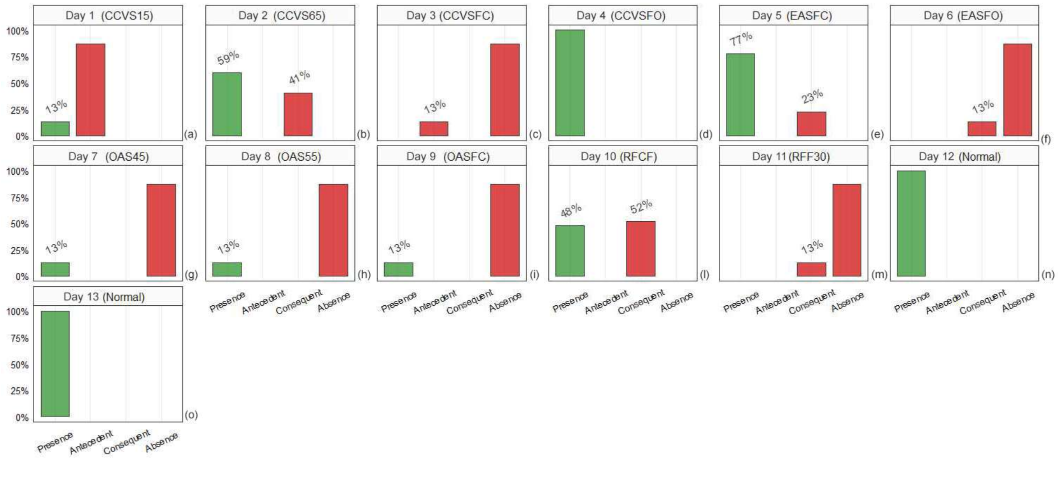


Fig. 73 - Characterization of the presence or the violation of the extracted rules for the testing days (refer to Table 16 for the encoding of faults)

The first group includes days characterized by the presence of the 100% of the 465 rules tested. This is the case of days in Fig. 73 (d), (n) and (o) tagged as Normal and the faulty day tagged as CCVSFO. Such condition suggests, as expected, that during the faulty day CCVSFO the start-up of the system can be considered normal.

The second group instead, includes the days in Fig. 73 (a), (c), (f), (g), (h), (i) and (m) that are characterized by a net prevalence of rule violations (more than

70%). Moreover, for those days, the presence of a fault is also associated to a specific kind of violation of the rules. As a reference, in case of CCVSFC, EASFO, OAS45, OAS55, OASFC and RFF30 the rules are violated mainly due to the absence of both antecedents and consequents, while only in the in case of CCVS15 the rule is violated for the absence of consequent.

The third group includes the day in Fig. 73 (e) for which, during the start-up period, the percentage of violations is lower than the percentage of valid occurrences of the rules. Such condition suggests that during this day the behaviour of the system is similar to the normal one limiting the number of violations. The main reason is that such fault does not strongly affect the system operation making the detection process less sensible to its presence. This result agreed with the findings of the ASHRAE-RP 1312 project, during which the analysed dataset was generated.

The last group includes days in Fig. 73 (b) and (l) that are characterized by a similar amount of violated and not violated rules (violation rate between 40% and 60%). These two faults seem to affect the performance of the system differently from other faults respect to which hypothetically they should exhibit high similarity (i.e., CCVS15 and RFF30). Regarding the fault CCVS65 (Fig. 73 (b)), the cooling coil valve is stuck at 65% and therefore the supply air flow is overcooled. In this case, the system reacts opening the heating coil valve and operating in fully recirculation mode for increasing the supply air temperature. Consequently, the failure of the cooling coil valve does not affect the capability of the system in reaching the supply set-point temperature, but the operation of the other components is different from the normal condition.

On the opposite, during the day (Fig. 73 (a)) tagged as CCVS15 (included in group 2) the cooling coil valve is almost closed limiting the heat exchange with the supply air flow that does not reach the set point temperature. Such case is representative of the complete failure of the system in maintaining the desired conditions of the indoor environment, as a matter of fact, justifying a higher rule violation rate for CCVS15 respect to CCVS65.

Regarding the fault RFCF (Fig. 73 (l)), the system is operated implementing the complete failure of the return fan despite its speed control signal is correctly elaborated. Instead, during the day in Fig. 73 (m) tagged as RFF30 (included in group 2), the return fan is not corrupted, but it is subjected to a faulty control signal. In this case the high number of rules violated for RFF30 suggests a higher sensitivity of the extracted rules to frequent transitions of the fan speed discrete values rather than fan power ones.

Some key figures related to the 465 extracted rules are in the following described. The rules are characterized by an ACTUAL TIME LAG that lies between 15 and 30 minutes. The evaluation of the ACTUAL TIME LAG can be then an essential step for reducing the intrinsic latency of the FDD tool during real implementation. Indeed, the ACTUAL TIME LAG gives the opportunity to check the occurrence of a rule within a time interval smaller than the width of the sliding window used for the rule extraction (in this case study is equal to 60 min.).

Table 18 reports the transitions in the antecedent and consequent item sets with the corresponding occurrence frequency.

Table 18 - Occurrence frequency of each event included in the antecedent and consequent item sets

Itemset	**Variable**	**Event**	**Frequency**
Antecedent	Return Fan Speed	RF_SPD [A-B]	87%
	Cooling coil input water temperature	CHWC_EWT [D-C]	31%
	Return fan power	RF_WAT [A-B]	24%
	Exhaust air damper position	EA_DMPR [A-B]	23%
	Cooling coil output water temperature	CHWC_LWT [C-B]	22%
	Supply fan speed	SF_SPD [A-B]	22%
	Cooling coil input water temperature	CHWC_EWT [C-A]	21%
	Supply fan power	SF_WAT [A-B]	21%
	Return fan start/stop signal	RF_SST [A-B]	18%
	Return air flow rate	RA_CFM [A-B]	16%
	Return fan pressure drop	RF_DP [A-B]	12%
	Return fan pressure drop	SF_DP [A-B]	9%
Consequent	Return Fan Speed	RF_SPD [B-C]	87%
	Supply Air Temperature	SA_TEMP [B-A]	49%
	Cooling coil output water temperature	CHWC_LWT [B-A]	46%
	Cooling coil air temperature	CHWC_DAT [B-A]	36%

In particular, the number of different consequent item sets is 13, resulting from a combination of 4 different events, while the antecedent item sets are 99, resulting from the combination of 12 different events.

Table (b) in Appendix B reports the obtained rule set including the most representative rules.

The rules extracted are meaningful since they can be interpreted as chains of events that characterise the normal operation of the AHU in reaching the set-point conditions during the start-up period. Indeed, extracted rules can be expressed as IF-THEN implications to be verified within a specific time interval. As a reference, the rule n° 8661 (included in Table (b) in Appendix B) can be written and interpreted as follow: IF (RF_SPD [A-B] and CHWC_LWT [C-B] and EA_DMPR [A-B]) occur THEN (CHWC_DAT [B-A] and RF_SPD [B-C]) will occur within 30 minutes with the 100% of confidence during a normal day.

In detail, the antecedent itemset includes transitions related to the return fan speed, the cooling coil output water temperature and the exhaust air damper position that imply the occurrence of consequent transitions related to the temperature of the air after the cooling coil and return fan speed.

In order to further improve the interpretability of the rules a novel visualization is also proposed. Fig. 74 shows an example of this visualization, referred to the profiles of the variables involved in rule n°1253 (included in Table (b) in Appendix B).

Fig. 74 shows the trend of the variables in terms of real profile (i.e. green curve) and PAA (i.e. black segments). Regardless of the approximation introduced by the PAA, the behaviour of the variables during the transient period is preserved, as can be seen by looking at the supply air temperature trend (SA_TEMP). In fact, during the start-up period the supply fan speed (SF_SPD) initially ramps up and then it is reduced to a constant level. The transitions of the antecedent itemset are reported in red, while the consequent itemset in blue. The PAA is represented in a window of 60 minutes, while with a darker shade of grey the length of the ACTUAL TIME LAG (i.e., 15 minutes) is reported. On the y-axis are shown the values used for the discretization of each variable.

The rule in Fig. 74 shows a typical behaviour of the system at the start-up period, in terms of the variation of “Supply fan speed” (SA_SPD), “Exhaust air damper” (EA_DMPR),“Return fan power” (SA_CFM), and “supply air temperature” (SA_TEMP).

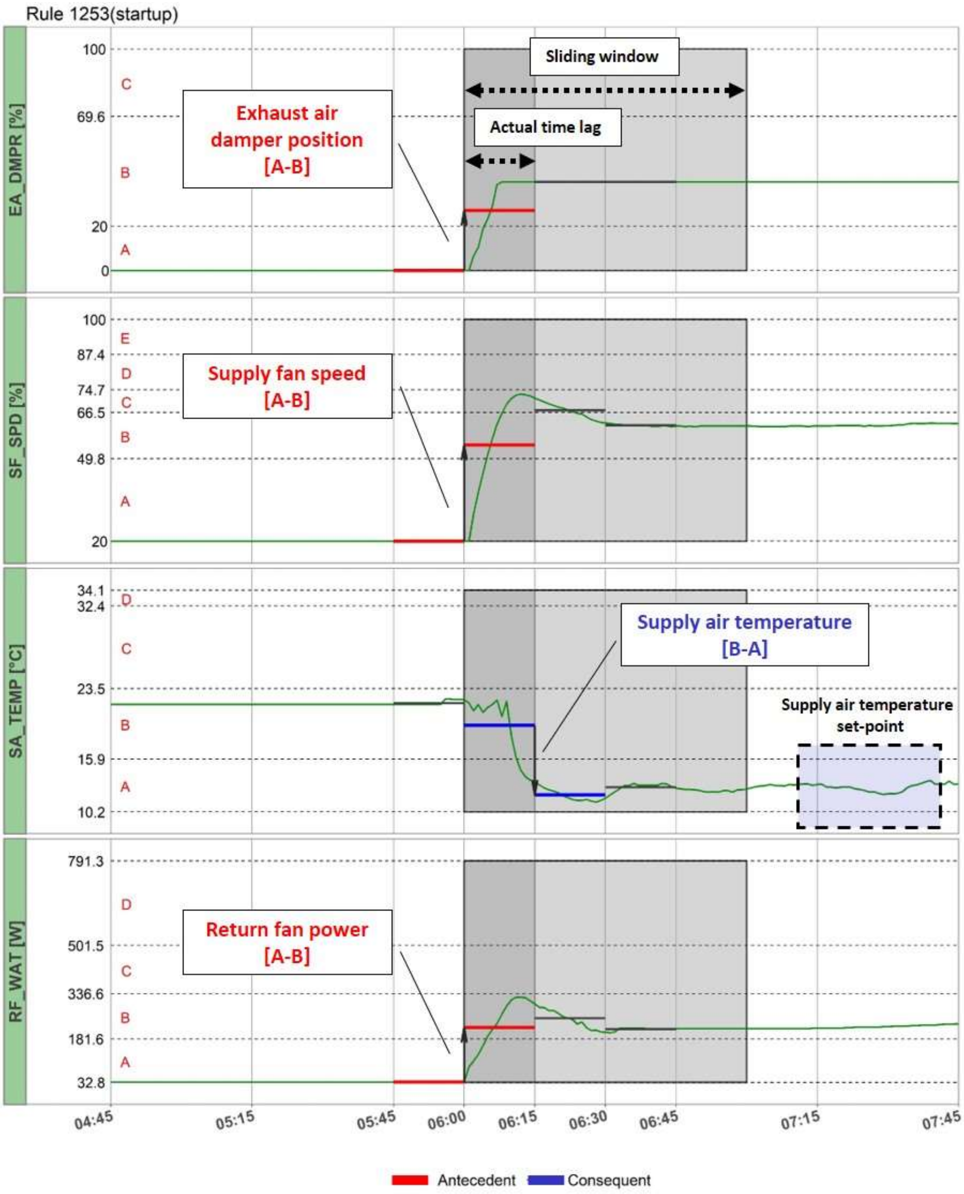


Fig. 74 - Visualization of an extracted temporal association rule (refer to Table 17 for variable encoding)

According to this rule, usually at the time scheduled for the start-up (i.e. 6:00 a.m.), the supply fan receives the start signal contemporary to the opening of the exhaust air damper while the return fan power increases (change from A to B).

After 15 minutes from the occurrence of the first antecedent transition in the event chain, according to the rule supply air temperature decreases from symbolic discrete-value B to A until the reaching of the desired set-point.

This proved that the chain of events related to each association rule provides information about the expected behaviour in terms of discrete-value changes among influencing variables of the AHU during normal operation.

4.1.4.3 Fault detection and diagnosis during non-transient period

The first step of the methodology for non-transient period, is aimed at developing a classification tree model for each variable to predict the normal operation of the system. For the development of the reference estimation models, all the variables related to the operation of the AHU are selected once at a time as target attribute while the remaining ones are used as input attributes. However, features related to external forcing variables to the AHU system (i.e., cooling coil input water temperature, outdoor air temperature) are used only as input attributes.

As a result, 21 reference models are built for providing a robust benchmark of fault-free operation. Moreover, the variables used as input are also considered with a maximum backward lag of four time steps (i.e. 60 minutes). Indeed, the decision trees are able to predict the discrete values (i.e., symbol) of a target variable considering the discrete values of the input variables both in the same and previous aggregation intervals.

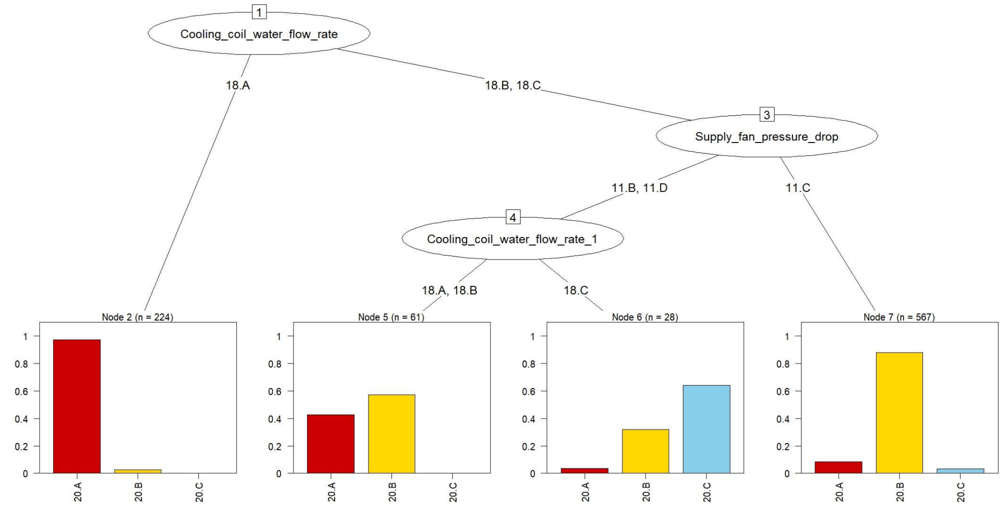


Fig. 75 - Decision tree for the estimation of the symbolic discrete-values of the cooling coil valve position

Fig. 75, reports as an example the classification tree developed for predicting the discrete values (i.e., symbol) of the variable "*cooling coil water valve position*" (i.e., variable tagged as CHWC_VLV with ID n° = 20), with an overall accuracy of 88%. The algorithm selects as input variables the "*cooling coil water flow rate*" (i.e., variable tagged as CHWC_GPM with ID n° = 18) and the "*pressure drop of the supply fan*" (i.e., variable tagged as SF_DP with ID n° = 11). From this classification tree, it is possible to extract useful decision rules for straightforwardly characterizing all the implications between discrete values (i.e., symbols) that typically occur during the fault-free operation of the AHU. Table 19 reports all the IF-THEN decision rules extracted from the CT shown in Fig. 75 with the reference of the accuracy achieved in each leaf node. The accuracy is referred to each single leaf node assuming that the predicted label of the node corresponds to the label of the majority of the objects.

Table 19 - Decision rules for the estimation of the symbolic discrete-value of cooling coil valve position

Rule number	Decision rules	CHWC_VLV discrete-value	N° of objects	Leaf node accuracy
1)	IF CHWC_GPM = 18_A	20_A	224	95%
2)	IF CHWC_GPM = 18_B or 18_C AND SF_DP = 11_B or 11_D AND CHWC_GPM (lag -1) = 18_A or 18_B	20_B	61	55%
3)	IF CHWC_GPM = 18_B or 18_C AND SF_DP = 11_B or 11_D AND CHWC_GPM (lag -1) = 18_C	20_C	28	65%
4)	IF CHWC_GPM = 18_B or 18_C AND SF_DP = 11_C	20_B	567	85%

For example, according to rule 4, the value of the response variable "*cooling coil valve position*" is equal to 20_B (i.e. CHWC_VLV lies in the interval 41 – 75 [%]) if the "*cooling coil water flow rate*" is equal to 18_B or 18_C (i.e. CHWC_GPM lies in the interval 0,89 – 2.7 [m^3/h]) and the "*supply fan pressure drop*" is equal to 11_C (i.e. SF_DP lies in the interval 562 – 770 [Pa]). Once all the estimation models are trained and validated, the residual analysis is performed by using a testing dataset including both faulty and fault-free data (i.e., 2 fault-free and 11 faulty days). Therefore, the difference between the actual discrete state of a variable and that estimated by the decision tree during an aggregation interval determines the detection or not of a potential faulty condition, since the predicted discrete state should be considered as the reference condition (fault-free). The values of the residuals can be equal to zero in case of absence of deviation from the normal conditions, positive if the actual value is higher than expected, or negative if the actual value is lower than expected.

Eventually, in order to perform fault diagnosis, it is developed an additional classification tree which uses as input the residuals obtained from the estimation performed through the 21 classification trees in the estimation layer and as output the tags related to the various faults considered.

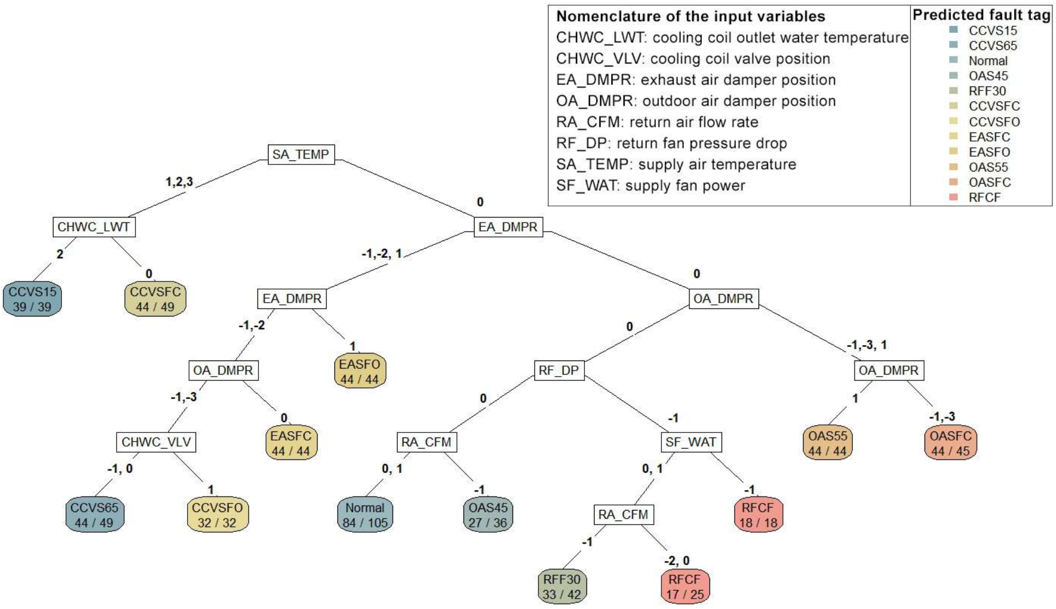


Fig. 76 - Classification tree for the fault diagnosis during the non-transient period

Fig. 76 shows the classification model obtained, which can classify with a set of intuitive rules the faults considered with an overall accuracy of the 90%.

The variables involved as input for the classification are the supply air temperature (i.e. SA_TEMP), the position of the outdoor and exhaust air dampers (i.e. OA_DMPR, EA_DMPR), the cooling coil outlet water temperature (i.e. CHWC_LWT), the cooling coil valve position (i.e. CHWC_VLV), the power demand of the supply fan (i.e. SF_WAT), the pressure drop of the return fan (i.e. RF_DP) and the return air flow rate (i.e. RA_CFM). The classification tree developed can diagnose 11 different faults and the normal condition as well.

The latter is predicted by following the decision tree path (Fig. 76) including all zeros (i.e. residual equal to zero) in the splits for the variables SA_TEMP, EA_DMPR, OA_DMPR, RF_DP and RA_CFM. By referring to Fig. 76, some other rules are described in the following.

The first split made by the classification tree concerns with the supply air temperature, which identifies the faults related to a blockage of the cooling coil valve at 0% (CCSFC) or at 15% (CCVS15) if the air temperature presents values higher than normal (i.e., SA_TEMP residuals = 1, 2, 3),.

In some cases, the faults can be diagnosed by analysing the variables directly related to the corrupted component, such as the blockage of the exhaust and outdoor air dampers at 0%, 55% or 100% (i.e. OASFC, OAS55, EASFC, and EASFO). In other cases, a series of deviation from the normal condition for different variables can be considered as symptoms for a specific fault. That is the case, for example, of anomalous energy exchange in the cooling coil due to blockage of the cooling coil valve at 65% (CCVS65) or 100% (CCVSFO). These faults are diagnosed in the case both the air dampers are completely closed (i.e. negative values of residuals), but the supply air temperature does not present a deviation from the normal condition. In this case, the system tries to counterbalance the excessive decrease of the temperature of the air by operating in fully recirculation mode.

The effect of a fault related to the return fan can be easily identified, since the pressure drop at the return fan is lower than expected while supply air temperature and air damper positions are normal. The discrimination between the complete failure of the return fan (RFCF) and the case in which the speed is fixed at 30% (RFF30) can be performed by evaluating the severity of the reduction of the return air flow rate rather than the reduction of the power demand of the supply fan.

The FDD tool is based on a multiclass classifier for fault diagnosis and when in operation sorts data into either fault-free (i.e., normal) or faulty classes. All the evaluation metrics for a multiclass classification model can be understood in the context of a binary classification model (where the classes are "positive" and "negative"). These metrics are derived from the following categories:

- True Positives (TP): Objects labelled as positive and predicted to be positive.
- False Positives (FP): Objects labelled as negative and predicted to be positive.
- True Negatives (TN): Objects labelled as negative and predicted to be negative.
- False Negatives (FN): Objects labelled as positive and predicted to be negative.

The multiclass classification problem can be seen as a set of many binary classification problems and its performance can be assessed labelling as "positive" each class once at time. In the context of the presented multiclass diagnostic classifiers some metrics can be calculated:

- Accuracy (A): Objects of items correctly identified as either truly positive or truly negative out of the total number of items i.e., $\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$.
- Recall (R): Number of objects correctly identified as positive out of the total actual positives i.e., $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$. The recall is calculated for each class and then averaged among classes for a global performance assessment of the CT.
- Precision (P): Number of objects correctly identified as positive out of the total items predicted as positive i.e., $\frac{\mathrm{TP}}{\mathrm{TP+}}$. The precision is calculated for each class and then averaged among classes for a global performance assessment of the CT.
- False Positive Rate (FPR), Type I error: Number of objects wrongly identified as faulty out of the total actual fault-free data i.e., $\frac{\mathrm{FP}}{\mathrm{FP+T}}$. In FDD processes, this error means that data belonging to fault-free class (negative) are incorrectly labelled as faulty (positives) generating false alarms.

- False Negative Rate (FNR), Type II error : Number of objects wrongly predicted as fault-free out of the total actual faulty data i.e., $\frac{\mathrm{FN}}{\mathrm{FN+TP}}$. In FDD processes, this error means that data belonging to one of the fault classes (positives) are incorrectly labelled as fault-free (negative) generating missing detection opportunities.

In particular, the developed classification tree exhibits the following performances A = 90%, R = 89%, P = 91%, FNR = 4%, FPR = 4%. The performance of the classification tree can be also assessed with the detail of each class considered. To this purpose, Table 20 reports the Confusion Matrix (CM) of the classification tree. The CM, in form of table (actual class vs predicted class), allows an effective analysis of the performance of the classification tree making it possible to identify confusion between all the considered classes (i.e., mislabelling of objects belonging to a class and classified into another one).

In particular, rows of the table correspond to the actual classes while columns to the predicted ones. At this stage it is possible to evaluate in each class the proportion of prediction actually correct (i.e., Precision) and the proportion of actual values predicted correctly (i.e., Recall).

Table 20 - Precision and recall for classification tree of fault diagnosis during non-transient period

	CCVS15	**CCVS65**	**Normal**	**OAS45**	**RFF30**	**CCVSFC**	**CCVSFO**	**EASFC**	**EASFO**	**OAS55**	**OASFC**	**RFCF**	**Total**	**Recall**
CCVS15	**39**	0	0	0	0	5	0	0	0	0	0	0	44	89%
CCVS65	0	**44**	0	0	0	0	0	0	0	0	0	0	44	100%
Normal	0	0	**84**	3	0	0	0	0	0	0	1	0	88	96%
OAS45	0	0	17	**27**	0	0	0	0	0	0	0	0	44	61%
RFF30	0	0	2	1	**33**	0	0	0	0	0	0	8	44	75%
CCVSFC	0	0	0	0	0	**44**	0	0	0	0	0	0	44	100%
CCVSFO	0	5	2	5	0	0	**32**	0	0	0	0	0	44	73%
EASFC	0	0	0	0	0	0	0	**44**	0	0	0	0	44	100%
EASFO	0	0	0	0	0	0	0	0	**44**	0	0	0	44	100%
OAS55	0	0	0	0	0	0	0	0	0	**44**	0	0	44	100%
OASFC	0	0	0	0	0	0	0	0	0	0	**44**	0	44	100%
RFCF	0	0	0	0	9	0	0	0	0	0	0	**35**	44	80%
Total	39	49	105	36	42	49	32	44	44	44	45	43	572	**Average 89%**
Precision	100%	90%	80%	75%	79%	90%	100%	100%	100%	100%	98%	81%	**Average 91%**	

Thanks to this methodology the faults in the dataset were diagnosed with both high precision and recall, as can be seen in Table 20. The lowest values of precision and recall are related the fault "*outdoor air damper stuck at 45%*" (OAS45), for which part of the records are mislabelled as "Normal" (i.e., 17 out of 44 objects, that correspond to the 39% of data labelled as OAS45 and to the 89% of the total amount of False Negatives). This condition is due to the fact that the stuck of the outdoor air damper at 45% does not invalidate the operation of the system which is similar to the fault-free one during the non-transient period. It is worth nothing that all the assumptions taken, and results obtained are related to a specific operative condition of the system (i.e., cooling mode). The set of rules extracted can be then considered a valid FDD solution if only applied on data consistent with the initial hypotheses. Despite this, even though the analysis is related to a portion of the possible operative conditions of an AHU, the performance achieved suggests good perspectives in applicability and generalizability of the proposed tool.

4.1.5 Discussion

The developed FDD tool is based on two different analytics modules proposed for the analysis of both transient and non-transient conditions of AHUs. The fault detection during the start-up period is performed with an innovative approach by searching frequent and non-anomalous relationships between events in a temporal transaction set using temporal association rules. A temporal association rule is expressed as a logical IF-THEN implication where the presence of an event (i.e., antecedent) implies the occurrence of another event (i.e., consequent) within a certain time lag. According to this approach, the violation of a rule or group of rules may suggest the occurrence of abnormal conditions during system operation. Three potential rule violations are considered for detecting faults during the start-up period: i) absence of the antecedent, ii) absence of consequent, iii) absence of antecedent and consequent.

The used rules are extracted by expert knowledge from a large set of possible rules, are representative of the normal operation of the AHU and are characterised by high physical interpretability. The introduction of innovative parameters (e.g. SUPP.DAY in faulty and normal conditions, support and confidence in the ACTUAL TIME LAG) allow a robust selection of the most interesting association rules, minimising the effort required in the post-processing stage. Furthermore, an effective visualization of the temporal association rules is introduced with the aim of supporting energy managers in the interpretation of the temporal associations between operational variables in real-time.

The FDD tool, for the non-transient period, provides a robust benchmark of the fault-free operation of the AHU by training and testing 21 classification trees. The classification trees are able to predict the discrete values (i.e., symbol) of a target operational variable considering the values of the input variables both in the same and previous aggregation interval. The classification trees show high performance (i.e., high accuracy, precision and recall) in modeling all the variable relations that are characteristic of the operative condition of interest (i.e., 20 days of AHU operated in cooling mode).

Eventually, an additional classifier is embedded in the tool in order to perform fault diagnosis. The diagnosis shows an overall accuracy of 90% and is performed by means of a set of intuitive rules easy to be implemented for detecting up to 11 typical faults in AHUs. However, the set of rules extracted can be considered valid only if applied on data consistent with the initial hypotheses (i.e., AHU operated in cooling mode).

Overall, the results obtained are characterised by robustness and high interpretability proving the effectiveness of the proposed FDD tool for ensuring a correct energy and operational management of the ventilation and air-conditioning process. Even though the tool is tailored for the case study analysed, the outcomes of the analysis can be considered flexible and generalizable. The methodologies are conceived for being automatic and for effectively managing the redundancy, interpretability and physical meaningfulness of the association and classification rules. Moreover, the developed FDD tool is conceived for quasi real-time implementation, paying attention to the optimisation of its computational cost. To this purpose, the preliminary discretisation of the variables, performed trough the aSAX algorithm, proved to be particularly effective in extracting the crucial operational conditions of the AHU reaching the optimal trade-off between data reduction and information loss. Moreover, the association rules are extracted from an event-based dataset (i.e., database of transactions) where only information about the discrete-value changes of the operational variables is stored. As a consequence, the computational cost related to the mining of rules is strongly reduced, increasing the feasibility of such approach in real case studies. In more detail the rule extraction phase takes more or less 10 minutes. It means that the most onerous parts of the analysis are represented by the pre-processing and post-mining phases. In the pre-processing phase the assessment of the optimal quantization of the time series through aSAX is validated by using more than 20 metrics (cluster validity indices included in the R Nbclust package [234]). Such calculation takes more than 10 minutes. In the post mining phase, the recalculation of support and confidence of each rule within the evaluated ACTUAL TIME LAG (instead of the window of 60-min), and the violation analysis performed on the testing dataset take about 20 minutes. For what concern the analysis of non-transient data, the development of each classification tree takes few seconds of computation and can be considered a task easily parallelizable. As a result, the impact of the analysis of non-transient data can be considered negligible in terms of computational cost compared to the pre-processing, rule extraction and rule post-mining. Indeed, in the perspective of a real-time implementation of the FDD tool, the update of the discretization intervals, set of association rules and estimation models can be accomplished during night-time while the fault detection and diagnosis tool can be run online during operation. For what concern the pre-processing stage, during the real-time operation the Hampel filter can still be used but considering that its intrinsic latency equal to *(Len-1)/2* should be added to the latency of the FDD process in detecting faults (in this case the latency of the FDD tool is equal to the length of the aggregation interval i.e., 15 min). For avoiding high latency in the analysis, *Len* can be reduced. As an alternative, other pre-processing algorithms, particularly suitable for the analysis of data streams, can be employed for detecting statistical outliers in real-time (i.e., before time t_{+1} and without any look ahead) [238].

A major future effort to build upon this work is the expansion of the tool to other operation modes and systems and to integrate it with knowledge driven-based analysis for better addressing the implementation issues that are characteristic of data analytics-based FDD tools. Indeed, such tools need a proper amount of data for the development of diagnosis models and cannot extrapolate beyond the range of training data [13]. It means that their capability in automatically extracting pattern from actual performance data is strictly related to the availability of pre-labelled monitored data (typically derived from AHU recommissioning or simulated data). On the contrary, knowledge driven-based approach (i.e., quantitative approach) can introduce domain knowledge and user experience into the FDD process [13], especially in the case initial information is not enough for deploying an FDD tool. In this perspective, a perfect integration of both approaches represents the main opportunity for significantly improve robustness, accuracy, and generalizability of FDD tools conceived for applications in building energy systems.

5 Conclusions

The present dissertation was aimed at demonstrating the effectiveness of data analytics-based DSS tools for improving energy management and enhancing energy efficiency in buildings. The growing availability of building related data is currently changing the decision-making process for optimising building daily operation making it possible to exploit the great potential of data analytics-based technologies also in the building sector. However, the knowledge extraction from massive building datasets is not an easy task and it requires skills in both data science and building physics.

The research activity outlined in the present dissertation was undertaken in this framework with the aim to actively contributing to bridging the gap between these two research areas.

To this purpose both meter-level and system-level DSS applications were explored and advanced data analytics-based EIS and FDD tools were proposed. Each tool was conceived for being implemented on a specific scale and for addressing tasks that are relevant in the building research field (from system component up to building portfolio level) bringing the following innovative aspects:

- **HVAC scheduling improvements at building system level**
  The improvement of HVAC schedules is one of the most effective way for reducing energy waste in buildings during daily operation. HVAC are often responsible of the largest amount of the building energy consumption and typically are operated with fixed schedules that poorly fit with the actual occupancy of the building. In that perspective, the developed EIS tool is capable to effectively analyze measured occupancy data and extracting from them typical patterns in form of daily profiles. The scope behind the identification of such patterns is twofold. On one hand, it allows the reduction of occupancy diversity in buildings by means of the proper displacement of occupants among building thermal zones. On the other hand, it makes possible to reduce the operation hours of the HVAC system by modifying its schedule according to the actual presence of occupants. The developed EIS tool fully exploits both opportunities through an innovative process of analysis that is also capable to deal with variable behaviors of occupants during the week (multiple typical occupancy profiles) always preserving their privacy. The results obtained for the considered case study show that the HVAC scheduling improvement could determine a potential monthly reduction of the electricity use for HVAC (space heating, space cooling, ventilation and air treatments) that ranges from 12.2% to 15.4% while the average energy saving for the whole analysed period (4 months) amounts to 14%.

- **Identification of energy consumption reduction opportunities through the detection of anomalous energy trends at whole building level**
  Although data availability is increasing in buildings, in most of real cases, just few and aggregate variables related to the total energy consumption of the building are measured and stored. Improving building energy performance by analysing data at a such high level is challenging, especially for buildings characterised by the existence of various load conditions. In that perspective the developed EIS tool is capable to automatically detect anomalous energy trends in building energy consumption time series exploiting a small set of input variables. The tool is based on an innovative methodology that performs a transformation of the whole building energy consumption time series by coupling Symbolic Aggregate approXimation and decision trees. The main advantage introduced is the possibility of reducing data volume and at the same time performing advanced pattern recognition analysis on data referred to characteristic periods of the day. The results obtained for the two case studies demonstrated that the developed classifiers can predict the typical patterns of building energy consumption during each considered periods of the day with an accuracy well over 80%. As a result of the high the accuracy of the classifiers (final nodes with very high occurrence probability of a certain energy consumption pattern), it is possible to achive a strong anomaly detection capability of the EIS tool when the classification rules are violated during building operation.The tool is able to distinguish infrequent from anomalous sub-daily patterns based on specific boundary conditions in a fully interpretable way, helping users in early detecting anomalous energy patterns and diagnosing their most probable associated cause during building operation.

- **Identification of typical energy use patterns at building portfolio level** i.e., customer classification. The knowledge of typical energy use patterns in large building portfolios is extremely valuable for designing targeted financial demand response programs, benchmarking the energy performance of buildings amongst their peers, and identifying strategic modifications of the building energy consumption curve. The developed EIS tool was capable to automatically extract from a building portfolio database, 5 groups of typical load profiles (i.e., benchmarking) and estimate for a new unknown customer its membership to one of them (i.e., customer classification). The tool is based on an evolutionary decision tree and achieves a classification accuracy of about 75% (6% higher than a reference classifier based on recursive partitioning decision tree). The main innovative aspects introduced are twofold. The first one is the development of a non-intrusive classification model that does not take into account attributes based on in-field load monitoring as input variables but only exploits historical billing data and a-priori knowledge (e.g. type of activity, voltage level, type of contract, occupant arrival and exit time). The second aspect is related to the capability of the tool in providing for a new customer not only a normalized reference load profile but also an estimation of its magnitude. Classifying for a customer the expected shape of typical load profile together with its magnitude enables the exploration of demand response opportunities at the very early stage of the customer engagement phase.

- **Fault detection and diagnosis at AHU component level**
  The optimal management of heating ventilation and air conditioning systems, is a crucial task, considering that such systems account for 50% of the energy demand in commercial buildings [131]. However, Air Handling Units (AHUs), that are an essential part of HVAC systems, are often inappropriately managed negatively impacting on building energy consumption and on the control of the indoor environment conditions.
  The developed FDD tool is based on a novel application of temporal association rules and decision trees for the extraction and identification of dominant pattern in AHU time series. Such patterns, in form of IF-THEN rules, are capable to robustly characterize the normal fault-free operation of the system during both transient and non-transient condition. The rules are completely interpretable and their violation in the time domain allows system inefficiencies and failures to be detected and be associated to faults of fans, dampers, and valves. The FDD tool is capable of detecting up to 11 typical faults (of valves, fans and dampers) in AHUs with an overall accuracy of 90%. The novelty of the proposed tool consists in the evolution of the FDD task from an expert-threshold-based analysis to an unsupervised-event-based one. In this way it is possible to learn robust FDD policies without a-priori knowledge of the system configuration by exploiting the knowledge of only relevant events extracted from multiple time series.

Regardless to the final objective of each tool two main investigations are at the basis of methodology development:

- Exploitation of time series analytics (e.g., sequential pattern mining, causality analysis, time series similarity) for extracting hidden patterns from building related data in the time domain,
- Implementation of supervised and unsupervised algorithms that provide results in terms of interpretable IF-THEN rules (decision trees, association rule mining algorithms).

These two aspects were considered as key points to overcome the main barriers that today thwart the fully exploitation of data analytics-based technologies for building energy management. Currently, DSS software represents an effective solution for gaining insight into building data and converting it in actionable knowledge. However, the knowledge gap that exists between building professionals and data scientists significantly affect the impact that such tools could have in improving building operation. Indeed, non-expert users if not adequately supported in the knowledge exploitation phase, tend to not trust results and suggestions provided by data analytics-based systems. For this reason, in some cases, trivial but highly understandable analyses were preferred to very complex but detailed ones. The great challenge to face is then maximising the extraction of hidden patterns from data provided that their interpretability is guaranteed.

In that perspective, the developed tools significantly contributed to achieve this demanding target in the robust way as possible. Most of the findings and outcomes of the present research work were already discussed in detail in the previous chapters. Therefore, the goal of this final chapter is to provide mainly a wide overview on the lessons learned in the framework of this research study.

Data pre-processing: the "sword of Damocles" hanging over data analysts

One of the main barriers that can be encountered while developing data analytics-based processes is the low quality of data. Data volume is worth nothing if it is not supported by high quality data. In this perspective data pre-processing represents a mandatory step in the process of analysis. Data pre-processing requires very good skills in data analysis, and it could take up to 80% of the whole computational time. However, in the discussed applications the aim of data pre-processing was twofold. On one hand data cleaning techniques (e.g., application of hampel filter) were used for detecting and replacing statistical outliers and increasing data quality. On the other hand, reduction and transformation techniques were employed for improving knowledge extraction according to specific mining targets.

The definition of a good pre-processing procedure cannot be seen as an isolated task to be performed at the beginning of the analysis. In fact, the pre-processing and preparation of data impact the entire flow of analysis and can significantly enhance the performance of data analytics algorithms. Good evidence of this aspect was found in the development of the EIS tool for anomaly detection and the FDD tool (sections 3.3 and 4.1). For these applications, an enhanced SAX algorithm was used for reducing and transforming time series of building/energy system data. SAX made it possible to significantly reduce the dataset size and at same time to better identify relevant patterns in the time domain. Indeed, the encoding of time series in sequences of symbols perfectly match with the use of automatic rule extraction techniques (e.g., decision tree, association rules) and then increasing the whole process interpretability.

Another crucial aspect of data pre-processing is related to its configuration in the case of on-line deployment of data analytics processes. The pre-processing of data streams needs to be designed differently from off-line static conditions, considering that the managing of missing values and the identification of statistical outliers should be performed in real-time. In that case the pre-processing of data could introduce latency issues in the process of analysis to be properly examined.

"In the wrong hands, all tools are weapons": data analytics and privacy issues

Pervasive monitoring and control systems enable the opportunity to collect a large amount of data in buildings and to provide fine-grained and optimised controls for heating, cooling, ventilation, lighting, and other building energy systems. However, the information that is collected, especially if referred to occupancy data, could potentially be used for undesirable purposes, generating then privacy issues [203]. The characterization of occupant behaviour and presence in buildings represents a key strategy to improve system management and efficiency through low/no cost and capital-intensive measures.

Particularly, the EIS tool developed for HVAC schedule improvements and occupancy diversity reduction, demonstrated its potential in reducing building energy demand by exploiting occupancy data, provided that information about occupant was aggregated and anonymised. However, regarding the opportunity to reduce occupancy diversity in specific parts of the building, the aggregation of data represents a great constraint. In fact, the analysis of individual occupant locations may yield much better results in terms of occupancy diversity reduction in building. However it can reveal contextual information about the individuals' habits, interests, activities, and relationships exposing occupants to mobbing practice, denigration and social reputation or economic damage [203]. In that perspective, the design of the data analytics processes should always take into account potential privacy issues finding a trade-off between the amount of knowledge extracted and the protection of sensible information.

"In theory, there is no difference between theory and practice.…but, in practice, there is" (Jan L. A. van de Snepscheut): **Scientific research vs real-life implementation**

The development of data analytics methodologies of analysis has been widely explored in the scientific literature but in most of the cases only through off-line tests. Despite off-line tests are essential for assessing the reliability of data analytics processes, crucial aspects related to data volume management, computational cost, updating of models, decline in accuracy are often neglected. In this research study rule extraction techniques were employed for two main reasons. On one hand such techniques are characterised by a high degree of interpretability, that represents a huge benefit for the final user. On the other hand, outputs in form of small sets of IF-THEN rules can be easily embedded in energy management systems and updated with a low computational cost (as a reference the computation of a recursive partitioning decision tree takes few seconds).

Another aspect to be considered is related to the actual availability of a proper amount of data for the development of data analytics-based tool. Indeed, despite their capability in automatically extracting pattern from actual performance data they cannot extrapolate beyond the range of training data [13]. Especially in the case of new installation of monitoring systems, data analytics-based tools can be enabled after a certain amount of time. On the contrary, an approach based on domain expertise can introduce a-priori knowledge into the data analytics process [13], particularly useful in the first period of data collection. In this perspective, a perfect integration of both approaches will represent the main opportunity for significantly improving robustness, accuracy, and generalizability of advanced data analytics tools and significantly reducing the time lag between their installation and utilization.

"The whole is greater than the sum of its parts": is the integration of EMIS tools always possible?

The main problem faced throughout this research study has been the availability of heterogeneous sets of building monitored variables among the case studies analysed. As a consequence, each scale of application was investigated with reference to different buildings. That is in fact no problem, because each tool was properly developed and tested demonstrating its robustness and effectiveness. However, it was not possible to integrate tools in a unique multilevel EMIS solution that allows information to be exchanged between tools and to achieve further improvements in building energy management. The most valuable solution is to re-think the paradigm behind data collection and data analysis. Monitoring systems are often designed with the only aim of measuring operation variables useful for the control of building systems neglecting other kind of measurements that could instead enable the implementation of advanced processes of analysis. Next generation of building monitoring systems should be then conceived, from their early design stage, in the perspective of maximising the set of functionalities that EMIS systems could have when being installed in buildings.

Having skills in data analytics is not enough

The diversity of data analytics techniques and their combination require good skills in data science. However, most of the effort is needed for understanding which analytical process can support the analyst in achieving a specific energy management target. It means that a strong background in building physics is always required for the fully exploitation of building related data.

In the context of the modern building industry, engineers rarely have both scientific backgrounds, and this could introduce several limitations. In fact, while data scientists tend to approach the energy modeling problems mainly looking at the achievable algorithm performance, energy and building engineers give priority to understandability and respect of physics laws of the analysis. As a result, on one hand data scientists are exposed to the risk of extracting counterintuitive knowledge from building related data. On the other hand, building end energy engineers typically are able to infer only trivial and obvious knowledge from large energy datasets. It is clear that the building sector is experiencing a relevant transition phase making essential the introduction of a new hybrid professional figure that is transversal to both energy and analytics application fields. This is a long process during which the scientific research will play a fundamental role in driving the required technological and knowledge transfers.

“By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it” (Eliezer Youdkowsky)**: Toward Explainable Artificial Intelligence (XAI)**

The EU General Data Protection Regulation (GDPR) went into effect on May 25, 2018. The regulation had a great impact on Artificial Intelligence (AI) companies and professionals especially due to the introduced “right to explanation” mandate. The basic concept is that when a decision is automatically generated, the final user has the right to receive an explanation about the generation process of that decision. Despite the regulation is essentially focused on personal data protection and safeguarding of personal rights, the debate on AI explainability also took place in cross-cutting sectors, with no exclusion of the energy and building one. As a reference, the advent of novel techniques (e.g., deep learning methods) and learning approaches (e.g., reinforcement learning) makes it possible to integrate more and more sophisticated AI-based systems in the building energy management process. However, for building professionals is a big challenge to fully understand the inference mechanism learnt by such AI systems and then they could express mistrust towards their outputs. To enhance confidence in AI, data scientists are today focusing on the development of new Explainable AI (so-called XAI) systems that will have the ability to explain their rationale, their weaknesses, their strengths and how they will perform in the future.

A major effort to build upon this research work will be then focused on fully addressing all the mentioned challenges that are behind the next generation of “intelligent” buildings.

Acknowledgment

Particular thanks to ASHRAE for the permission given for the use of the data and documents from ASHRAE RP-1312.

© ASHRAE www.ashrae.org (“ASHRAE RP-1312 Tools for Evaluating Fault Detection and Diagnostic Methods for Air-Handling Units.”), (2011).

References

[1] Capozzoli A, Cerquitelli T, Piscitelli MS. Chapter 11 – Enhancing energy efficiency in buildings through innovative data analytics technologies, in: D. Ciprian, F. Xhafa (Eds.), Pervasive Comput., 2016: pp. 353–389.

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

[3] Cao X, Dai X, Liu J. Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. Energy Build 2016; 128: 198–213.

[4] Allouhi A, El Fouih Y, Kousksou T, Jamil A, Zeraouli Y, Mourad Y. Energy consumption and efficiency in buildings: Current status and future trends. J Clean Prod 2015; 109: 118–130.

[5] Pérez-Lombard L, Ortiz J, Pout C. A review on buildings energy consumption information. Energy Build 2008; 40: 394–398.

[6] Marinakis V. Big Data for Energy Management and Energy-Efficient Buildings. Energies 2020; 13: 1555.

[7] Fan C, Xiao F, Li Z, Wang J. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy Build 2018; 159: 296–308.

[8] Bhattarai BP, Paudyal S, Luo Y, Mohanpurkar M, Cheung K, Tonkoski R, Hovsapian R, Myers KS, Zhang R, Zhao P, Manic M, Zhang S, Zhang X. Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions. IET Smart Grid 2019; 2: 141–154.

[9] Park JY, Yang X, Miller C, Arjunan P, Nagy Z. Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. Appl Energy 2019; 236: 1280–1295.

[10] Piscitelli MS, Brandi S, Capozzoli A. Recognition and classification of typical load profiles in buildings with non-intrusive learning approach. Appl Energy 2019; 255: 113727.

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

[12] Kramer H, Lin G, Granderson J, Curtin C, Crowe E. Synthesis of Year One Outcomes in the Smart Energy Analytics Campaign Building Technology and Urban Systems Division. 2017.

[13] Zhao Y, Li T, Zhang X, Zhang C. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. Renew Sustain Energy Rev 2019; 109: 85–101.

[14] Capozzoli A, Piscitelli MS, Gorrino A, Ballarini I, Corrado V. Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. Sustain Cities Soc 2017; 35: 191–208.

[15] Luo X, Hong T, Chen Y, Piette MA. Electric load shape benchmarking for small- and medium-sized commercial buildings. Appl Energy 2017; 204: 715–725.

[16] Zakovorotnyi A, Seerig A. Building energy data analysis by clustering measured daily profiles. Energy Procedia 2017; 122: 583–588.

[17] Fu TC. A review on time series data mining. Eng Appl Artif Intell 2011; 24: 164–181.

[18] Shumway RH, Stoffer DS. Time Series Analysis and Its Applications, 2017.

[19] Mitsa T. Temporal Data Mining, 2010. http://dl.acm.org/citation.cfm?id=1809755.

[20] Pei J, Han J, Wang W. Mining sequential patterns with constraints in large databases. Int Conf Inf Knowl Manag Proc 2002; 18–25.

[21] Capozzoli A, Savino M, Brandi S, Grassi D, Chicco G. Automated load patterns learning and diagnosis for enhancing energy management in smart buildings. Energy 2018; 157: 336–352.

[22] Piscitelli MS, Mazzarelli DM, Capozzoli A. Submitted for publication. Fault detection and diagnosis for air handling units in buildings through temporal association and decision rules. Energy Build.

[23] Fan C, Xiao F, Yan C. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Autom Constr 2015; 50: 81–90.

[24] Xiao F, Fan C. Data mining in building automation system for improving building operational performance. Energy Build 2014; 75: 109–118.

[25] Notaristefano A, Chicco G, Piglione F. Data size reduction with symbolic aggregate approximation for electrical load pattern grouping. IET Gener Transm Distrib 2013; 7: 108–117.

[26] Wrinch M, EL-Fouly THMHMM, Wong S. Anomaly detection of building systems using energy demand frequency domain analysis. 2012 IEEE Power Energy Soc Gen Meet 2012; 2012: 1–6.

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

[28] Lin J, Keogh E, Lonardi S, Chiu B. A Symbolic Representation of Time Series , with Implications for Streaming Algorithms. Proc 8th ACM SIGMOD Work Res Issues Data Min Knowl Discov 2003; 2–11.

[29] Reinhardt A, Koessler S. PowerSAX: Fast motif matching in distributed power meter data using symbolic representations, in: Proc. 9th IEEE Int. Work. Pract. Issues Build. Sens. NetworkApplications (SenseApp 2014), IEEE, Edmonton, Canada, 2014: pp. 531–538.

[30] Fan C, Xiao F, Madsen H, Wang D. Temporal knowledge discovery in big BAS data for building energy management. Energy Build 2015; 109: 75–89.

[31] Fonseca JA, Miller C, Schlueter A. Unsupervised load shape clustering for urban building performance assessment. Energy Procedia 2017; 122: 229–234.

[32] Pham ND, Le QL, Dang TK. HOT a SAX : A Novel Adaptive Symbolic Representation for Time Series Discords Discovery, in: N.T. Nguyen, M.T. Le, J. Świątek (Eds.), Intell. Inf. Database Syst. ACIIDS 2010. Lect. Notes Comput. Sci., Springer, Berlin, 2010: pp. 113–121.

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

[34] Fan C, Xiao F, Yan C. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Autom Constr 2015; 50: 81–90.

[35] Miller C, Nagy Z, Schlueter A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. Renew Sustain Energy Rev 2018; 81: 1365–1377.

[36] Tan P-N, Steinbach M, Kumar V. Classification: Basic Concepts, Decision Trees, and Model Evaluation. Introd to Data Min 2006; 67: 145–205.

[37] Bhatia P. Introduction to Data Mining, in: Data Min. Data Warehous., 2019: pp. 17–27.

[38] Ramasubramanian K, Singh A. Machine Learning Using R. apress 2017

[39] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees, 1984.

[40] Grubinger T, Zeileis A, Pfeiffer K-P. evtree : Evolutionary Learning of Globally Optimal Classification and Regression Trees in R . J Stat Softw 2015; 61:1-29.

[41] Juang B H RL. he segmental k-means algorithm for estimating parameters of hidden markov models. IEEE Trans Acoust Speech Signal Process 1990; 38: 1639–1641.

[42] Kaufman L RPJ. Finding groups in data: An introduction to cluster analysis. Wiley 1990.

[43] Tan P-N, Steinbach M, Kumar V. Chap 8 : Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining. Pearson, Boston, 2005.

[44] Ester M, Kriegel H P, Sander J XX. A density-based algorithm for discovering clusters in large spatial databases with noise. Knowl Discov Data Min 1996; 226–231.

[45] Chicco G, Napoli R, Postolache P, Scutariu M, Toader C. Customer Characterization Options for Improving the Tariff Offer. IEEE Trans Power Syst 2003; 18: 381–387.

[46] Chicco G, Napoli R, Postolache P, Scutariu M, Toader C. Emergent electricity customer classificatio. IEE Proceedings-Generation, Transm Distrib 2005; 152: 164–172.

[47] Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell 1979; PAMI-1: 224–227.

[48] Tan P-N, Steinbach M, Kumar V. Introduction to Data Mining, Pearson, Boston, 2005.

[49] Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Mach Learn 2001; 42: 31–60.

[50] Martínez-De-Pisón FJ, Sanz A, Martínez-De-Pisón E, Jiménez E, Conti D. Mining association rules from time series to explain failures in a hot-dip galvanizing steel line. Comput Ind Eng 2012; 63: 22–36.

[51] Kaur G. Association Rule Mining: A survey. Int J Comput Sci Inf Technol 2014; 5: 2320–2324.

[52] Capozzoli A, Piscitelli MS, Neri F, Grassi D, Serale G. A novel methodology for energy performance benchmarking of buildings by means of Linear Mixed Effect Model: The case of space and DHW heating of out-patient Healthcare Centres. Appl Energy 2016; 171: 592–607.

[53] Fan C, Xiao F, Wang S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energy 2014; 127: 1–10.

[54] Miller C. More Buildings Make More Generalizable Models—Benchmarking Prediction Methods on Open Electrical Meter Data. Mach Learn Knowl Extr 2019; 1: 974–993.

[55] Zhao Y, Wen J, Wang S. Diagnostic Bayesian networks for diagnosing air handling units faults - Part II: Faults in coils and sensors. Appl Therm Eng 2015; 90: 145–157.

[56] Capozzoli A, Grassi D, Causone F. Estimation models of heating energy consumption in schools for local authorities planning. Energy Build 2015; 105: 302–313.

[57] Liu X. Smart Meter Data Analytics: Systems, Algorithms, and Benchmarking. ACM Trans Database Syst 2016; 42: 1–39.

[58] D'Oca S, Hong T. Occupancy schedules learning process through a data mining framework. Energy Build 2015; 88: 395–408.

[59] Wu S, Sun JQ. Cross-level fault detection and diagnosis of building HVAC systems. Build Environ 2011; 46: 1558–1566.

[60] Capozzoli A, Lauro F, Khan I. Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Syst Appl 2015; 42: 4324–4338.

[61] Ahmad a. S, Hassan MY, Abdullah MP, Rahman H a., Hussin F, Abdullah H, Saidur R. A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renew Sustain Energy Rev 2014; 33: 102–109.

[62] Molina-Solana M, Ros M, Ruiz MD, Gómez-Romero J, Martin-Bautista MJ, Martín-Bautista MJ, Martin-Bautista MJ. Data Science for Building

Energy Management: a Review. Renew Sustain Energy Rev 2017; 70: 598–609.

[63] Li Z, Han Y, Xu P. Methods for benchmarking building energy consumption against its past or intended performance : An overview. Appl Energy 2014; 124: 325–334.

[64] Zhao HX, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012; 16: 3586–3592.

[65] Li Q, Meng Q, Cai J, Yoshino H, Mochida A. Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. Energy Convers Manag 2009; 50: 90–96.

[66] Fan C, Sun Y, Zhao Y, Song M, Wang J. Deep learning-based feature engineering methods for improved building energy prediction. Appl Energy 2019; 240: 35–45.

[67] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms. Appl Energy 2017; 195: 222–233.

[68] Kumar R, Aggarwal RK, Sharma JD. Energy analysis of a building using artificial neural network: A review. Energy Build 2013; 65: 352–358.

[69] Li Z, Huang G. Re-evaluation of building cooling load prediction models for use in humid subtropical area. Energy Build 2013; 62: 442–449.

[70] Aydinalp M, Ugursal VI, Fung AS. Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks. Appl Energy 2004; 79: 159–178.

[71] Mihalakakou G, Santamouris M, Tsangrassoulis a. On the energy consumption in residential buildings. Energy Build 2002; 34: 727–736.

[72] Ben-Nakhi AE, Mahmoud MA. Cooling load prediction for buildings using general regression neural networks. Energy Convers Manag 2004; 45: 2127–2141.

[73] Dong B, Cao C, Lee SE. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005; 37: 545–553.

[74] Yu Z, Haghighat F, Fung BCM, Yoshino H. A decision tree method for building energy demand modeling. Energy Build 2010; 42: 1637–1646.

[75] Mikučionienė R, Martinaitis V, Keras E. Evaluation of energy efficiency measures sustainability by decision tree method. Energy Build 2014; 76: 64–71.

[76] Capozzoli A, Grassi D, Piscitelli MS, Serale G. Discovering knowledge from a residential building stock through data mining analysis for engineering sustainability. Energy Procedia 2015; 83: 370–379.

[77] Attanasio A, Piscitelli MS, Chiusano S, Capozzoli A, Cerquitelli T. Towards an automated, fast and interpretable estimation model of heating energy demand: A data-driven approach exploiting building energy certificates. Energies 2019; 12(7): 1-25.

[78] Capozzoli A, Serale G, Piscitelli MS, Grassi D. Data mining for energy analysis of a large data set of flats. Proc Inst Civ Eng Eng Sustain 2017; 170: 3–18.

[79] Capozzoli A, Serale G, Marco Savino P, Grassi D. Data mining for energy analysis of a large data set of fl ats. Proc Inst Civ Eng Eng Sustain 2017; 170: 3–18.

[80] Dudek G. Neural networks for pattern-based short-term load forecasting: A comparative study. Neurocomputing 2016; 205: 64–74.

[81] Iglesias F, Kastner W. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. Energies 2013; 6: 579–597.

[82] Panapakidis IP, Papadopoulos TA, Christoforidis GC, Papagiannis GK. Pattern recognition algorithms for electricity load curve analysis of buildings. Energy Build 2014; 73: 137–145.

[83] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy 2014; 135: 461–471.

[84] Tsekouras GJ, Hatziargyriou ND, Dialynas EN. Two-stage pattern recognition of load curves for classification of electricity customers. IEEE Trans Power Syst 2007; 22: 1120–1128.

[85] Fernandes MP, Viegas JL, Vieira SM, Sousa JM. Analysis of residential natural gas consumers using fuzzy c-means clustering. 2016 IEEE Int Conf Fuzzy Syst FUZZ-IEEE 2016 2016; 1484–1491.

[86] Marszal-Pomianowska A, Heiselberg P, Kalyanova Larsen O. Household electricity demand profiles - A high-resolution load model to facilitate modelling of energy flexible buildings. Energy 2016; 103: 487–501.

[87] Spertino F, Chicco G, Ciocia A, Corgnati S, Di Leo P, Raimondo D. Electricity consumption assessment and PV system integration in grid-connected office buildings. 2015 IEEE 15th Int Conf Environ Electr Eng EEEIC 2015 - Conf Proc 2015; 255–260.

[88] Ma Z, Yan R, Nord N. A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. Energy 2017; 134: 90–102.

[89] Miller C, Nagy Z, Schlueter A. Automated daily pattern filtering of measured building performance data. Autom Constr 2015; 49: 1–17.

[90] Zhou K Le, Yang SL, Shen C. A review of electric load classification in smart grid environment. Renew Sustain Energy Rev 2013; 24: 103–110.

[91] Jota PRSS, Silva VRBB, Jota FG. Building load management using cluster and statistical analyses. Int J Electr Power Energy Syst 2011; 33: 1498–1505.

[92] Beccali M, Cellura M, Lo Brano V, Marvuglia A. Short-term prediction of household electricity consumption: Assessing weather sensitivity in a Mediterranean area. Renew Sustain Energy Rev 2008; 12: 2040–2065.

[93] Do Carmo CMR, Christensen TH. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. Energy Build 2016; 125: 171–180.

[94] Wang Y, Chen Q, Kang C, Zhang M, Wang K, Zhao Y. Load profiling and its application to demand response: A review. Tsinghua Sci Technol 2015; 20: 117–129.

[95] Patnaik D, Marwah M, Sharma R, Ramakrishnan N. Sustainable Operation and Management of Data Center Chillers Using Temporal Data Mining. Proc 15th ACM SIGKDD Int Conf Knowl Discov Data Min 2009; 1305–1314.

[96] Yang J, Ning C, Deb C, Zhang F, Cheong D, Eang Lee S, Sekhar C, Wai Tham K, Lee SE, Sekhar C, Tham KW. k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy Build 2017; 146: 27–37.

[97] Le Cam M, Daoud A, Zmeureanu R. Forecasting electric demand of supply fan using data mining techniques. Energy 2016; 101: 541–557.

[98] Peña M, Biscarri F, Guerrero JI, Monedero I, León C. Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach. Expert Syst Appl 2016; 56: 242–255.

[99] Habib U, Zucker G. Finding the Different Patterns in Buildings Data Using Bag of Words Representation with Clustering. Proc - 2015 13th Int Conf Front Inf Technol FIT 2015 2016; 303–308.

[100] Alam MJE, Muttaqi KM, Sutanto D. A SAX-based advanced computational tool for assessment of clustered rooftop solar PV impacts on LV and MV networks in smart grid. IEEE Trans Smart Grid 2013; 4: 577–585.

[101] Tureczek AM, Nielsen PS. Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data. Energies 2017; 10: 1–19.

[102] Chicco G, Napoli R, Piglione F. Comparisons among clustering techniques for electricity customer classification. IEEE Trans Power Syst 2006; 21: 933–940.

[103] Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 2012; 42: 68–80.

[104] Panapakidis I, Alexiadis M, Papagiannis G. Evaluation of the performance of clustering algorithms for a high voltage industrial consumer. Eng Appl Artif Intell 2015; 38: 1–13.

[105] Mcloughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015; 141: 190–199.

[106] Fernandes MP, Viegas JL, Vieira SM, Sousa JMC. Segmentation of residential gas consumers using clustering analysis. Energies 2017; 10: 2047–2073.

[107] Figueiredo V, Rodrigues F, Vale Z, Gouveia JB. An electric energy consumer characterization framework based on data mining techniques. IEEE Trans Power Syst 2005; 20: 596–602.

[108] Chicco G, Ilie IS. Support vector clustering of electrical load pattern data. IEEE Trans Power Syst 2009; 24: 1619–1628.

[109] Piao M, Ryu KH. Subspace Frequency Analysis-Based Field Indices Extraction for Electricity Customer Classification. ACM Trans Inf Syst 2016; 34: 1–18.

[110] Biscarri F, Monedero I, García A, Guerrero JI, León C. Electricity clustering framework for automatic classification of customer loads. Expert Syst Appl 2017; 86: 54–63.

[111] Zhong S, Tam KS. Hierarchical Classification of Load Profiles Based on Their Characteristic Attributes in Frequency Domain. IEEE Trans Power Syst 2015; 30: 2434–2441.

[112] Ramos S, Duarte JM, Duarte FJ, Vale Z. A data-mining-based methodology to support MV electricity customers' characterization. Energy Build 2015; 91: 16–25.

[113] Wang F, Zhen Z, Wang B, Mi Z, Wang Z, Li K. A Baseline Load Estimation Approach for Residential Customer based on Load Pattern Clustering. Energy Procedia 2018; 142: 2042–2049.

[114] Grigoraş G, Bobric E-C. Clustering Based Approach for Customers' Classification From Electrical Distribution Systems. UPB Sci Bull, Ser C 2015; 77: 219–226.

[115] Siano P. Demand response and smart grids - A survey. Renew Sustain Energy Rev 2014; 30: 461–478. http://dx.doi.org/10.1016/j.rser.2013.10.022.

[116] Gelazanskas L, Gamage KAA. Demand side management in smart grid: A review and proposals for future direction. Sustain Cities Soc 2014; 11: 22–30.

[117] Verda V, Guelpa E, Sciacovelli A, Acquaviva A, Patti E. Thermal peak load shaving through users request variations. Int J Thermodyn 2016; 19: 168–176.

[118] Jang D, Eom J, Jae Park M, Jeung Rho J. Variability of electricity load patterns and its effect on demand response: A critical peak pricing experiment on Korean commercial and industrial customers. Energy Policy 2016; 88: 11–26.

[119] Chen CS, Hwang JC, Huang CW. Application of load survey systems to proper tariff design. IEEE Trans Power Syst 1997; 12: 1746–1751.

[120] Wang K, Zhang M, Wang Z, Li R, Li F, Wu H. Time of use tariff design for domestic customers from flat rate by model-based clustering. Energy Procedia 2014; 61: 652–655.

[121] Panapakidis IP, Christoforidis GC. Implementation of modified versions of the K-means algorithm in power load curves profiling. Sustain Cities Soc 2017; 35: 83–93.

[122] Jalali MM, Kazemi A. Demand side management in a smart grid with multiple electricity suppliers. Energy 2015; 81: 766–776.

[123] Azaza M, Wallin F. Smart meter data clustering using consumption indicators: Responsibility factor and consumption variability. Energy Procedia 2017; 142: 2236–2242.

[124] Benítez I, Quijano A, Díez JL, Delgado I. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. Int J Electr Power Energy Syst 2014; 55: 437–448.

[125] Khan I, Huang JZ, Luo Z, Masud MA. CPLP: An algorithm for tracking the changes of power consumption patterns in load profile data over time. Inf Sci (Ny) 2018; 429: 332–348.

[126] Kim W, Katipamula S. A review of fault detection and diagnostics methods for building systems. Sci Technol Built Environ 2018; 24: 3–21.

[127] Schein J, Bushby ST, Castro NS, House JM. A rule-based fault detection method for air handling units. Energy Build 2006; 38: 1485–1492.

[128] Katipamula S, Brambley MR. Review article: Methods for fault detection, diagnostics, and prognostics for building systems—a review, part II. HVAC R Res 2005; 11: 169–187.

[129] Isermann R. Fault-Diagnosis Systems: an introduction from fault detection to fault tolerance, Springer Science & Business Media, 2006.

[130] Granderson J, Lin G, Singla R, Mayhorn E, Ehrlich P, Vrabie D. Commercial Fault Detection and Diagnostics Tools: What They Offer, How They Differ, and What's Still Needed, 2018.

[131] Office of Energy Efficiency & Renewable Energy (EERE) U.S. Department of Energy, DOE Office of Energy Efficiency and Renewable Energy. Buildings energy databook, 2012.

[132] Yan K, Zhong C, Ji Z, Huang J. Semi-supervised learning for early detection and diagnosis of various air handling unit faults. Energy Build 2018; 181: 75–83.

[133] Proctor J. Residential and Small Commercial Central air Conditioning; Rated Efficiency isn't Automatic. Present Public Sess ASHRAE Winter Meet 2004.

[134] Beghi A, Brignoli R, Cecchinato L, Menegazzo G, Rampazzo M, Simmini F. Data-driven Fault Detection and Diagnosis for HVAC water chillers. Control Eng Pract 2016; 53: 79–91.

[135] Xue P, Zhou Z, Fang X, Chen X, Liu L, Liu Y, Liu J. Fault detection and operation optimization in district heating substations based on data mining techniques. Appl Energy 2017; 205: 926–940.

[136] Wen, Jin; Li S. ASHRAE 1312-RP Tools for Evaluating Fault Detection and Diagnostic Methods for Air-Handling Unit. 2011.

[137] Dehestani D, Eftekhari F, Guo Y, Ling S, Su S, Nguyen H. Online Support Vector Machine Applicationfor Model Based Fault Detection and Isolationof HVAC System. Int J Mach Learn Comput 2011; 1: 66–72.

[138] Zhao Y, Wen J, Xiao F, Yang X, Wang S. Diagnostic Bayesian networks for diagnosing air handling units faults – part I: Faults in dampers, fans, filters and sensors. Appl Therm Eng 2017; 111: 1272–1286.

[139] Mulumba T, Afshari A, Yan K, Shen W, Norford LK. Robust model-based fault diagnosis for air handling units. Energy Build 2015; 86: 698–707.

[140] Yan R, Ma Z, Zhao Y, Kokogiannakis G. A decision tree based data-driven diagnostic strategy for air handling units. Energy Build 2016; 133: 37–45.

[141] Mchugh MK. Data-Driven Leakage Detection in Air-Handling Units on a University Campus. ASHRAE Annu Conf 2019;

[142] Yu Z, Haghighat F, Fung BCM, Zhou L. A novel methodology for knowledge discovery through mining associations between building operational data. Energy Build 2012; 47: 430–440.

[143] Zhang C, Xue X, Zhao Y, Zhang X, Li T. An improved association rule mining-based method for revealing operational problems of building heating, ventilation and air conditioning (HVAC) systems. Appl Energy 2019; 253: 113492.

[144] Fan C, Sun Y, Shan K, Xiao F, Wang J. Discovering gradual patterns in building operations for improving building energy efficiency. Appl Energy 2018; 224: 116–123.

[145] Khan I, Capozzoli A, Corgnati SP, Cerquitelli T. Fault detection analysis of building energy consumption using data mining techniques. Energy Procedia 2013; 42: 557–566.

[146] Dey M, Rana SP, Dudley S. Smart building creation in large scale HVAC environments through automated fault detection and diagnosis. Futur Gener Comput Syst 2018;

[147] Du Z, Jin X, Yang Y. Fault diagnosis for temperature, flow rate and pressure sensors in VAV systems using wavelet neural network. Appl Energy 2009; 86: 1624–1631.

[148] Du Z, Fan B, Jin X, Chi J. Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. Build Environ 2014; 73: 1–11.

[149] Guo Y, Wall J, Li J, West S. Intelligent Model Based Fault Detection and Diagnosis for HVAC System Using Statistical Machine Learning Methods, in: ASHRAE 2013 Winter Conf., 2013: pp. 1–8.

[150] Li S, Wen J. A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform. Energy Build 2014; 68: 63–71.

[151] Jin X, Du Z. Fault tolerant control of outdoor air and AHU supply air temperature in VAV air conditioning systems using PCA method. Appl Therm Eng 2006; 26: 1226–1237.

[152] Liang J, Du R. Model-based Fault Detection and Diagnosis of HVAC systems using Support Vector Machine method. Int J Refrig 2007; 30: 1104–1114.

[153] Lee WY, House JM, Kyong NH. Subsystem level fault diagnosis of a building's air-handling unit using general regression neural networks. Appl Energy 2004; 77: 153–170.

[154] Dey D, Dong B. A probabilistic approach to diagnose faults of air handling units in buildings. Energy Build 2016; 130: 177–187. http://dx.doi.org/10.1016/j.enbuild.2016.08.017.

[155] Chung W. Review of building energy-use performance benchmarking methodologies. Appl Energy 2011; 88: 1470–1479.

[156] Wang S, Yan C, Xiao F. Quantitative energy performance assessment methods for existing buildings. Energy Build 2012; 55: 873–888.

[157] Lee W L, Yik F W H, Burnett J. A method to assess the energy performance of existing commercial complexes. Ind Built Env 2003, 12: 311-327.

[158] Fabrizio E, Corrado V, Filippi M. A model to design and optimize multi-energy systems in buildings at the design concept stage. Renew Energy 2010; 35: 644–655.

[159] Hong T, Piette MA, Chen Y, Lee SH, Taylor-lange SC, Zhang R, Sun K, Price P. Commercial Building Energy Saver: An energy retrofit analysis toolkit. Appl Energy 2015; 159: 298–309.

[160] Maria G, Hamdy M, Peter G, Bianco N, Hensen JLM. A new methodology for investigating the cost-optimality of energy retrofitting a building category. Energy Build 2015; 107: 456–478.

[161] Lee SH, Hong T, Piette MA, Taylor-Lange SC. Energy retrofit analysis toolkits for commercial buildings: A review. Energy 2015; 89:1087-1100

[162] Tahsildoost M, Sadat Z. Energy retrofit techniques : An experimental study of two typical school buildings in Tehran. Energy Build 2015; 104: 65–72.

[163] Al-homoud MS. Computer-aided building energy analysis techniques. Build Environ 2001; 36: 421–433.

[164] Filogamo L, Peri G, Rizzo G, Giaccone A. On the classification of large residential buildings stocks by sample typologies for energy planning purposes. Appl Energy 2014; 135: 825–835.

[165] Zhang Y, Neill ZO, Dong B, Augenbroe G. Comparisons of inverse modeling approaches for predicting building energy performance. Build Environ 2015; 86: 177–190.

[166] Lee W-S, Lee K-P. Benchmarking the performance of building energy management using data envelopment analysis. Appl Therm Eng 2009; 29: 3269–3273.

[167] Petcharat S, Chungpaibulpatana S, Rakkwamsuk P. Assessment of potential energy saving using cluster analysis : A case study of lighting systems in buildings. Energy Build 2012; 52: 145–152.

[168] Gao X. a New Methodology for Building Energy Benchmarking: an Approach Based on Clustering Concept and Statistical Models. Energy Build 2013; 84: 607–616.

[169] Hong T, Yang L, Hill D, Feng W. Data and analytics to inform energy retrofit of high performance buildings. Appl Energy 2014; 126: 90–106.

[170] Wang E. Benchmarking whole-building energy performance with multi-criteria technique for order preference by similarity to ideal solution using a selective objective-weighting approach. Appl Energy 2015; 146: 92–103.

[171] Lee W, Lin L. Evaluating and ranking the energy performance of office building using technique for order preference by similarity to ideal solution. Appl Therm Eng 2011; 31: 3521–3525.

[172] Yu Z (Jerry), Haghighat F, Fung BCM. Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustain Cities Soc 2016; 25: 33–38.

[173] Karatasou S, Santamouris M, Geros V. Modeling and predicting building's energy use with artificial neural networks: Methods and results. Energy Build 2006; 38: 949–958.

[174] Li Q, Meng Q, Cai J, Yoshino H, Mochida A. Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. Appl Energy 2009; 86: 2249–2256.

[175] Heo Y, Zavala VM. Gaussian process modeling for measurement and verification of building energy savings. Energy Build 2012; 53: 7–18.

[176] Manfren M, Aste N, Moshksar R. Calibration and uncertainty analysis for computer models - A meta-model based approach for integrated building energy simulation. Appl Energy 2013; 103: 627–641.

[177] Vu DH, Muttaqi KM, Agalgaonkar AP. A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. 2015; 140: 385–394.

[178] Braun MR, Altan H, Beck SBM. Using regression analysis to predict the future energy consumption of a supermarket in the UK. Appl Energy 2014; 130: 305–313.

[179] Chung W, Hui Y V., Lam YM. Benchmarking the energy efficiency of commercial buildings. Appl Energy 2006; 83: 1–14.

[180] Chung W. Using the fuzzy linear regression method to benchmark the energy efficiency of commercial buildings. Appl Energy 2012; 95: 45–49.

[181] Sharp T. Energy Benchmarking In Commercial Office Buildings. ACEEE Summer Study Energy Effic Build 1995; 4: 321–329.

[182] Energy Star. Energy Star Performance Ratings — Technical Methodology, Environmental Protection Agency, Washington, DC, 2011.

[183] Wong SL, Wan KKW, Lam TNT. Artificial neural networks for energy analysis of office buildings with daylighting. Appl Energy 2010; 87: 551–557.

[184] Hong T, Taylor-Lange SC, D'Oca S, Yan D, Corgnati SP. Advances in research and applications of energy-related occupant behavior in buildings. Energy Build 2016; 116: 694–702.

[185] D'Oca S, Hong T, Langevin J. The human dimensions of energy use in buildings: A review. Renew Sustain Energy Rev 2018; 81: 731–742.

[186] Hong T, Yan D, D'Oca S, Chen C fei. Ten questions concerning occupant behavior in buildings: The big picture. Build Environ 2017; 114: 518–530.

[187] O'Brien W, Gunay HB. The contextual factors contributing to occupants' adaptive comfort behaviors in offices - A review and proposed modeling framework. Build Environ 2014; 77: 77–87.

[188] Richard J. de Dear, Gail Schiller Brager. Developing an adaptive model of thermal comfort and preference. ASHRAE Trans 1998; 104: 1–18.

[189] D'Oca S, Hong T. A data-mining approach to discover patterns of window opening and closing behavior in offices. Build Environ 2014; 82: 726–739.

[190] Ouyang J, Hokao K. Energy-saving potential by improving occupants' behavior in urban residential sector in Hangzhou City, China. Energy Build 2009; 41: 711–720.

[191] Al-Mumin A, Khattab O, Sridhar G. Occupants' behavior and activity patterns influencing the energy consumption in the Kuwaiti residences. Energy Build 2003; 35: 549–559.

[192] Yu Z, Fung BCM, Haghighat F, Yoshino H, Morofsky E. A systematic procedure to study the influence of occupant behavior on building energy consumption. Energy Build 2011; 43: 1409–1417.

[193] Yu Zhun Jerry ZJ, Haghighat F, Fung BCM, Morofsky E, Yoshino H. A methodology for identifying and improving occupant behavior in residential buildings. Energy 2011; 36: 6596–6608.

[194] Ekwevugbe T, Brown N, Pakka V, Fan D. Improved Occupancy Monitoring in Non-Domestic Buildings. Sustain Cities Soc 2017; 30:311-327

[195] Ioannidis D, Tropios P, Krinidis S, Stavropoulos G, Tzovaras D, Likothanasis S. Occupancy driven building performance assessment. J Innov Digit Ecosyst 2016; 3: 57–69.

[196] Yang J, Santamouris M, Lee SE. Review of occupancy sensing systems and occupancy modeling methodologies for the application in institutional buildings. Energy Build 2016; 121: 344–349.

[197] Yang Z, Becerik-Gerber B. The coupled effects of personalized occupancy profile based HVAC schedules and room reassignment on building energy use. Energy Build 2014; 78: 113–122.

[198] Yang Z, Ghahramani A, Becerik-Gerber B. Building occupancy diversity and HVAC (heating, ventilation, and air conditioning) system energy efficiency. Energy 2016; 109: 641–649.

[199] Goyal S, Barooah P, Middelkoop T. Experimental study of occupancy-based control of HVAC zones. Appl Energy 2015; 140: 75–84.

[200] Budaiwi I, Abdou A. HVAC system operational strategies for reduced energy consumption in buildings with intermittent occupancy: The case of mosques. Energy Convers Manag 2013; 73: 37–50.

[201] Granderson J, Berkeley L, Piette MA, Ghatikar R. Building Energy Information Systems : State of the Technology and User Case Studies. Lawrence Berkeley National Laboratory Lawrence Berkeley National Laboratory 2011.

[202] Building Commissioning Association. New construction building commissioning best practice, 2018.

[203] Jia R, Dong R, Sastry SS, Spanos CJ. Privacy-enhanced architecture for occupancy-based HVAC control. Proc - 2017 ACM/IEEE 8th Int Conf Cyber-Physical Syst ICCPS 2017 (Part CPS Week) 2017; 177–186.

[204] Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the rpart routines, 1997.

[205] Liang X, Hong T, Shen GQ. Occupancy data analytics and prediction: A case study. Build Environ 2016; 102: 179–192.

[206] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979; 1: 224–227.

[207] Documentation, EnergyPlus. "Engineering reference-EnergyPlus 8.5".The Reference to EnergyPlus Calculation. 2016;

[208] International Organisation for Standardisation (ISO). ISO 18523-1.Energy performance of buildings - Schedule and condition of building, zone and space usage for energy calculation - Part 1: Non-residential buildings, 2016.

[209] American Society of Heating Refrigetation and Air-Conditioning Engineers, (ASHRAE). Measurement of Energy, Demand, and Water Savings, 2014.

[210] Chakrabarti K, Keogh E, Mehrotra S, Pazzani M. Locally adaptive dimensionality reduction for indexing large time series databases. ACM Trans Database Syst 2002; 27: 188–228.

[211] Ma Z, Yan R, Li K, Nord N. Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering. Energy Build 2018; 166: 284–295.

[212] Tureczek A, Nielsen PS, Madsen H. Electricity consumption clustering using smart meter data. Energies 2018; 11: 1–18.

[213] Pérez-Chacón R, Luna-Romera JM, Troncoso A, Martínez-Alvarez F, Riquelme JC. Big data analytics for discovering electricity consumption patterns in smart cities. Energies 2018; 11: 1–19.

[214] Miller C, Meggers F. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. Energy Build 2017; 156: 360–373.

[215] Arco L, Casas G, Nowé A. Clustering methodology for smart metering data based on local and global features, in: IML '17 Proc. 1st Int. Conf. Internet Things Mach. Learn., 2017: pp. 1–13.

[216] Panapakidis I, Christoforidis G. Optimal Selection of Clustering Algorithm via Multi-Criteria Decision Analysis (MCDA) for Load Profiling Applications. Appl Sci 2018; 8: 237–279.

[217] Bicego M, Farinelli A, Grosso E, Paolini D, Ramchurn SD. On the distinctiveness of the electricity load profile. Pattern Recognit 2018; 74: 317–325.

[218] Grubinger T, Zeileis A, Pfeiffer K-P. evtree : Evolutionary Learning of Globally Optimal Classification and Regression Trees in R. J Stat Softw 2015; 61: 1–29.

[219] Vercamer D, Steurtewagen B, Van Den Poel D, Vermeulen F. Predicting Consumer Load Profiles Using Commercial and Open Data. IEEE Trans Power Syst 2016; 31: 3693–3701.

[220] Li D, Zhou Y, Hu G, Spanos CJ. Optimal Sensor Configuration and Feature Selection for AHU Fault Detection and Diagnosis. IEEE Trans Ind Informatics 2017; 13: 1369–1380.

[221] Li S, Wen J. Application of pattern matching method for detecting faults in air handling unit system. Autom Constr 2014; 43: 49–58.

[222] Yan K, Huang J, Shen W, Ji Z. Unsupervised learning for fault detection and diagnosis of air handling units. Energy Build 2020; 210: 109689.

[223] Yan Y, Luh PB, Pattipati KR. Fault diagnosis of HVAC: Air delivery and terminal systems. IEEE Int Conf Autom Sci Eng 2017; 2017-Augus: 882–887.

[224] Zhong C, Yan K, Dai Y, Jin N, Lou B. Energy efficiency solutions for buildings: Automated fault diagnosis of air handling units using generative adversarial networks. Energies 2019; 12: 1–11.

[225] Gao T, Boguslawski B, Marié S, Béguery P, Thebault S, Lecoeuche S. Data mining and data-driven modelling for air handling unit fault detection. E3S Web Conf 2019; 111

[226] Li G, Hu Y, Chen H, Li H, Hu M, Guo Y, Liu J, Sun S, Sun M. Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions. Appl Energy 2017; 185: 846–861.

[227] Yu Y, Woradechjumroen D, Yu D. A review of fault detection and diagnosis methodologies on air-handling units. Energy Build 2014; 82: 550–562.

[228] Pearson RK. Data cleaning for dynamic modeling and control. Eur Control Conf ECC 1999 - Conf Proc 2015; 2584–2589.

[229] Pham ND, Le QL, Dang TK. HOT aSAX: A novel adaptive symbolic representation for time series discords discovery. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2010; 5990 LNAI: 113–121.

[230] Li S. A Model-Based Fault Detection and Diagnostic Methodology for Secondary HVAC Systems. Drexel Univ 2009.

[231] Kim M, Yoon SH, Domanski PA, Vance Payne W. Design of a steady-state detector for fault detection and diagnosis of a residential air conditioner. Int J Refrig 2008; 31: 790–799.

[232] Roh CW, Kim M, Kim HS, Kim MS. Design Method Of Steady State Detector For Multi-Evaporator Heat Pump System With Decomposition Analysis Technique. 2010.

[233] Dexter A, Pakanen J. Demonstrating Automated Fault Detection and Diagnosis Methods in Real Buildings, in: Proc. VTT Symp. 217, 2001: p. 381.

[234] Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set Malika. J Stat Softw 2014; 61.

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

[236] Hahsler M, Grun B, Hornik K. arules – A Computational Environment for Mining Association Rules and Frequent Item Sets. J Stat Softw 2005; 14: 1–6.

[237] Martínez-de-Pisón Ascacíbar FJ, Pernía Espinoza A, Martínez F, Roberto, Escribano García R, Guillén Rondón P, Conti Guillén D. System for uncovering hidden knowledge in real time for the analysis of environmental and agricultural processes. XIII Int Conf Proj Eng 2009.

[238] Ahmad S, Purdy S. Real-Time Anomaly Detection for Streaming Analytics. arXiv preprint arXiv:1607.02480, 2016.

Appendix A

Table (a) - Discretization intervals for all the analysed variables.

Variable	ID	Unit	Sym. A	Sym. B	Sym. C	Sym. D	Sym. E
SF_WAT	1	[W]	< 522 **OFF**	522 – 1265 **ON**	1265 – 2440 **ON**	> 2440 **ON**	-
RF_WAT	2	[W]	< 181 **ON**	181 - 337 **ON**	336 – 502 **ON**	> 502 **ON**	-
SA_CFM	3	[m^3/h]	< 591 **OFF**	591 – 2276 **ON**	2276 – 3414 **ON**	3414 – 4706 **ON**	> 4706 **ON**
RA_CFM	4	[m^3/h]	< 838 **OFF**	838 – 2712 **ON**	2712 – 3527 **ON**	> 3527 **ON**	-
OA_CFM	5	[m^3/h]	< 477 **ON**	477 – 1146 **ON**	> 1146 **ON**	-	-
SA_TEMP	6	[°C]	< 15,9 **ON**	15,9 – 23,5 **ON**	23,5 – 32,4 **ON**	> 32,4 **ON**	-
MA_TEMP	7	**[°C]**	**< 20,3** **ON**	20,3 – 30,8 **ON**	> 30,8 **ON**	-	-
RA_TEMP	8	[°C]	< 25,7 **ON**	25, 7 – 31,5 **ON**	>31,5 **ON**	-	-
HWC_DAT	9	[°C]	< 20,2 **ON**	20,2 – 26 **ON**	26 – 35,7 **ON**	> 35,7 **ON**	-
CHWC_DAT	10	[°C]	< 14,4 **ON**	14,4 – 22 **ON**	22 – 30,7 **ON**	> 30,7 **ON**	-
SF_DP	11	[Pa]	< 324 **OFF**	324 – 562 **ON**	562 – 770 **ON**	> 770 **ON**	-
RF_DP	12	[Pa]	< 46 **ON**	46 – 114 **ON**	> 114 **ON**	-	-
SF_SPD	13	[%]	< 50 **OFF**	50 – 67 **ON**	67 – 75 **ON**	75 – 87 **ON**	> 87 **ON**
RF_SPD	14	[%]	< 30 **OFF**	30 – 43 **ON**	43 – 57 **ON**	57- 69 **ON**	> 69 **ON**
OA_TEMP	15	[°C]	< 18,3 **ON**	> 18,3 **ON**	-	-	-
CHWC_EWT	16	[°C]	< 1,3 **ON**	1,3 – 2,8 **ON**	2,8 – 6,7 **ON**	> 6,7 **OFF**	-
CHWC_LWT	17	[°C]	< 13,3 **ON**	13,3 – 19,9 **ON**	19,9 – 21,2 **ON**	> 21,2 **OFF**	-
CHWC_GPM	18	[m^3/h]	< 0,9 **ON**	0,9 – 1,7 **ON**	> 1,7 **ON**	-	-
E_ccoil	19	[kW]	< 11,7 **ON**	> 11,7 **ON**	-	-	-
CHWC_VLV	20	[%]	< 41 **ON**	41 – 75 **ON**	> 75 **ON**	-	-
EA_DMPR	21	[%]	< 20 **ON**	20 – 70 ON	> 70 **ON**	-	-
OA_DMPR	22	[%]	< 20 **ON**	20 – 47 **ON**	47 – 77 **ON**	> 77 **ON**	-
RF_SST	23	[-]	< 0,5 **OFF**	> 0.5 **ON**	-	-	-

Appendix B

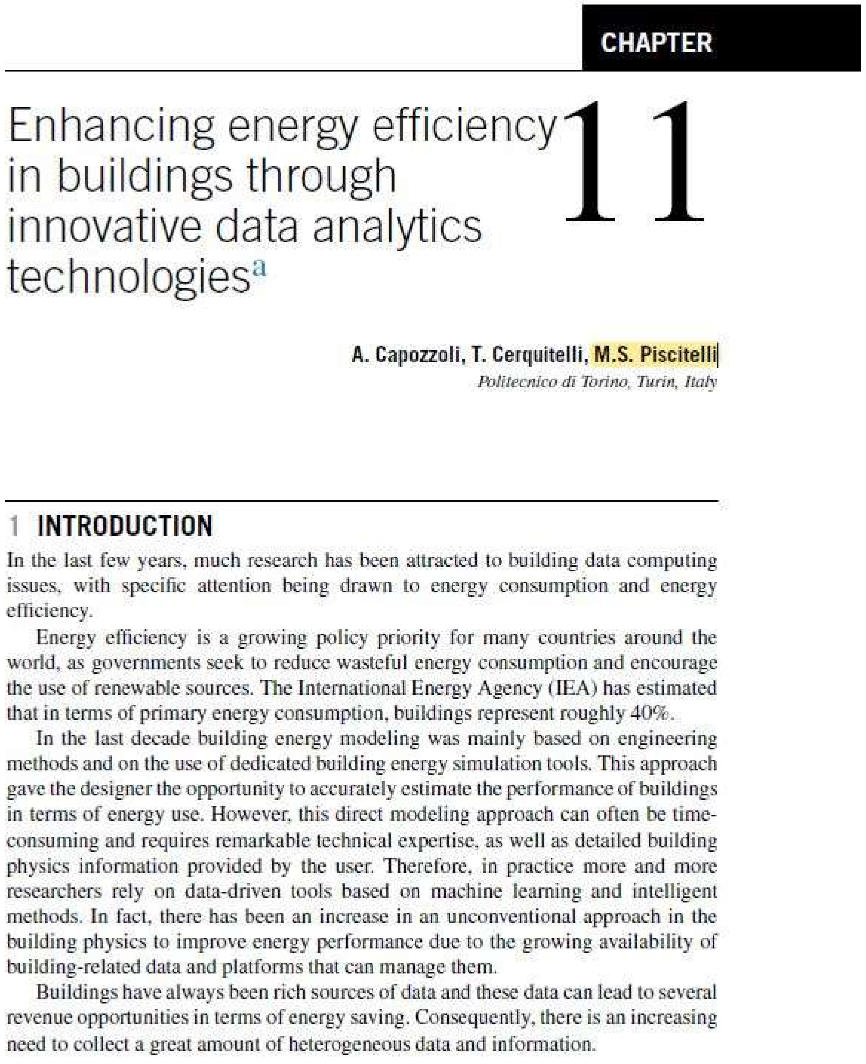
Table (b) - Most representative extracted temporal association rules

ID N°	Antecedent	Consequent	Supp.	Conf.	ACTUAL TIME LAG	SUPP. DAY FAULTY
1077	SF_SPD [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_DAT [B-A]	0.70	0.8	15	0.27
1078	SF_WAT [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_DAT [B-A]	0.70	0.8	15	0.27
1526	SF_SPD [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_DAT [B-A], CHWC_LWT [B-A]	0.70	0.8	15	0.27
1527	SF_WAT [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_DAT [B-A], CHWC_LWT [B-A]	0.70	0.8	15	0.27
1864	SF_SPD [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_DAT [B-A], CHWC_LWT [B-A], SA_TEMP [B-A]	0.75	0.8	15	0.27
1865	SF_WAT [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_DAT [B-A], CHWC_LWT [B-A], SA_TEMP [B-A]	0.75	0.8	15	0.27
8661	**RF_SPD [A-B], CHWC_LWT [C-B], EA_DMPR [A-B]**	**CHWC_DAT [B-A], RF_SPD [B-C]**	**0.9**	**1**	**30**	**0.09**
8750	RF_SPD [A-B], EA_DMPR [A-B], RF_SST [A-B]	CHWC_DAT [B-A], RF_SPD [B-C]	0.8	0.89	30	0
6255	RF_SPD [A-B], CHWC_LWT [C-B], RA_CFM [A-B]	CHWC_DAT [B-A], RF_SPD [B-C], CHWC_LWT [B-A]	0.89	0.8	15	0.18
6256	RF_SPD [A-B], CHWC_LWT [C-B], RF_SST [A-B]	CHWC_DAT [B-A], RF_SPD [B-C], CHWC_LWT [B-A]	0.89	0.8	15	0.09
6226	RF_SPD [A-B], CHWC_LWT [C-B], RF_SST [A-B]	CHWC_DAT [B-A], RF_SPD [B-C], SA_TEMP [B-A]	0.89	0.8	15	0.09
6936	RF_SPD [A-B], CHWC_LWT [C-B]	CHWC_DA T [B-A], RF_SPD [B-C], SA_TEMP [B-A]	0.889	0.8	15	0.18
1933	SF_SPD [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_LWT [B-A], SA_TEMP [B-A]	0.75	0.8	15	0.27
1934	SF_WAT [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_LWT [B-A], SA_TEMP [B-A]	0.75	0.8	15	0.27
5257	RF_SPD [A-B], EA_DMPR [A-B], RA_CFM [A-B]	RF_SPD [B-C]	0.82	1	30	0.18
5259	RF_SPD [A-B], EA_DMPR [A-B], RF_SST [A-B]	RF_SPD [B-C]	0.82	1	30	0.09
6415	RF_SPD [A-B], CHWC_LWT [C-B], RF_WAT [A-B]	RF_SPD [B-C], CHWC_LWT [B-A]	0.8	0.8	15	0.09
6416	RF_SPD [A-B], CHWC_LWT [C-B], RF_DP [A-B]	RF_SPD [B-C], CHWC_LWT [B-A]	0.8	0.8	15	0.09
6126	RF_SPD [A-B], CHWC_LWT [C-B], RF_WAT [A-B]	RF_SPD [B-C], CHWC_LWT [B-A], SA_TEMP [B-A]	0.89	0.8	15	0.09
8309	RF_SPD [A-B], CHWC_LWT [C-B], EA_DMPR [A-B]	RF_SPD [B-C], CHWC_LWT [B-A], SA_TEMP [B-A]	0.78	0.78	15	0.09
15268	RF_SPD [A-B], EA_DMPR [A-B], RF_WAT [A-B]	RF_SPD [B-C], SA_TEMP [B-A]	0.8	0.89	30	0
15269	RF_SPD [A-B], EA_DMPR [A-B], RF_DP [A-B]	RF_SPD [B-C], SA_TEMP [B-A]	0.8	0.89	30	0
1253	**SF_SPD [A-B], EA_DMPR [A-B], RF_WAT [A-B]**	**SA_TEMP [B-A]**	**0.70**	**0.8**	**15**	**0.27**
1254	SF_WAT [A-B], EA_DMPR [A-B], RF_WAT [A-B]	SA_TEMP [B-A]	0.70	0.8	15	0.27
1406	SF_WAT [A-B], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_DAT [B-A], SA_TEMP [B-A]	0.70	0.8	15	0.27
5240	CHWC_EWT [D-C], EA_DMPR [A-B], RF_WAT [A-B]	CHWC_DAT [B-A], SA_TEMP [B-A]	0.70	0.8	30	0.27

Table (b) reports 26 rules (two for each unique consequent transaction) extracted from the transient dataset with the specification of the event chains of antecedent and consequent, the value of support and confidence within the ACTUAL TIME LAG and its duration (evaluated on the training dataset), and the SUPP.DAY$_{\text{FAULTY}}$ (evaluated on the testing dataset).

Appendix C

This Appendix lists the papers published by the author that have been included/partially included in this dissertation.



CHAPTER

Enhancing energy efficiency in buildings through innovative data analytics technologies[a] 11

A. Capozzoli, T. Cerquitelli, M.S. Piscitelli

Politecnico di Torino, Turin, Italy

1 INTRODUCTION

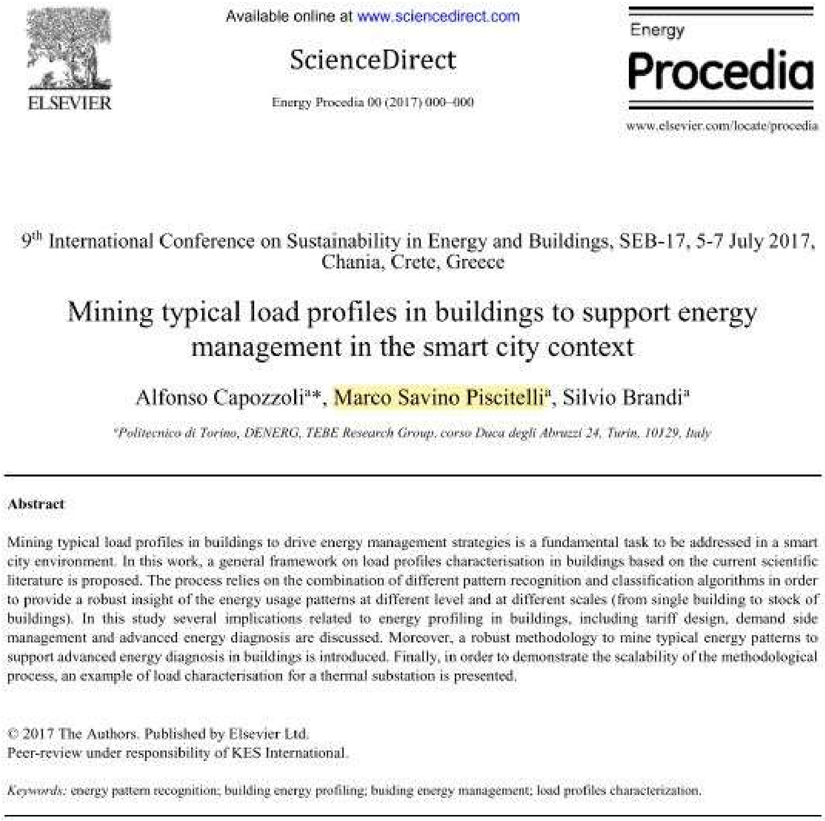
In the last few years, much research has been attracted to building data computing issues, with specific attention being drawn to energy consumption and energy efficiency.

Energy efficiency is a growing policy priority for many countries around the world, as governments seek to reduce wasteful energy consumption and encourage the use of renewable sources. The International Energy Agency (IEA) has estimated that in terms of primary energy consumption, buildings represent roughly 40%.

In the last decade building energy modeling was mainly based on engineering methods and on the use of dedicated building energy simulation tools. This approach gave the designer the opportunity to accurately estimate the performance of buildings in terms of energy use. However, this direct modeling approach can often be time-consuming and requires remarkable technical expertise, as well as detailed building physics information provided by the user. Therefore, in practice more and more researchers rely on data-driven tools based on machine learning and intelligent methods. In fact, there has been an increase in an unconventional approach in the building physics to improve energy performance due to the growing availability of building-related data and platforms that can manage them.

Buildings have always been rich sources of data and these data can lead to several revenue opportunities in terms of energy saving. Consequently, there is an increasing need to collect a great amount of heterogeneous data and information.

Capozzoli A., Cerquitelli T., Piscitelli M.S. 2016. *Chapter 11 – Enhancing energy efficiency in buildings through innovative data analytics technologies*, in: D. Ciprian, F. Xhafa (Eds.), Pervasive Comput., pp. 353–389. [1] (the portion reused by the author is less than 10% of the material in the book chapter as required by the publisher for a free use of the content)



Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 00 (2017) 000–000

Energy Procedia

www.elsevier.com/locate/procedia

9th International Conference on Sustainability in Energy and Buildings, SEB-17, 5-7 July 2017, Chania, Crete, Greece

Mining typical load profiles in buildings to support energy management in the smart city context

Alfonso Capozzoli[a]*, Marco Savino Piscitelli[a], Silvio Brandi[a]

[a]Politecnico di Torino, DENERG, TEBE Research Group, corso Duca degli Abruzzi 24, Turin, 10129, Italy

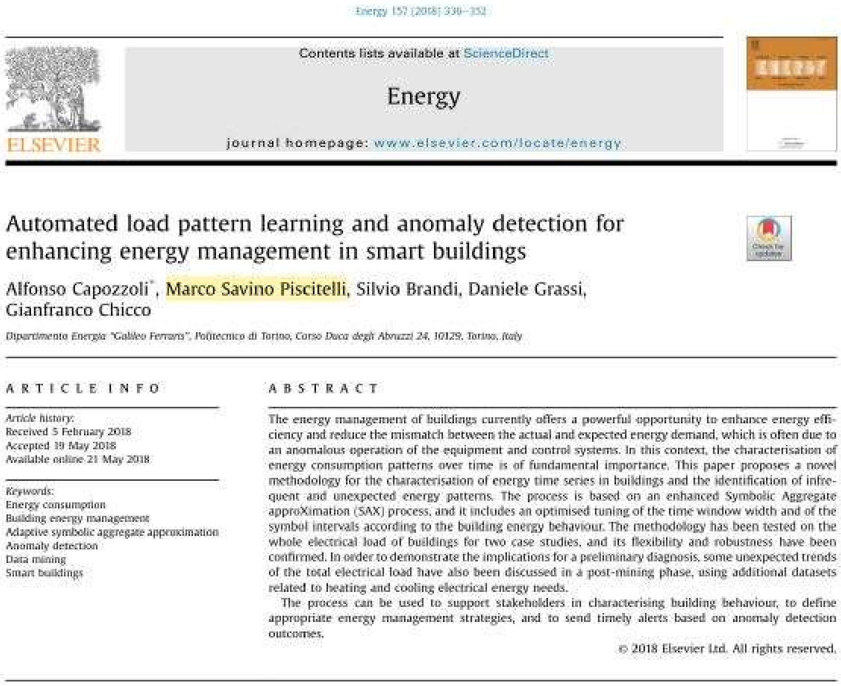
Abstract

Mining typical load profiles in buildings to drive energy management strategies is a fundamental task to be addressed in a smart city environment. In this work, a general framework on load profiles characterisation in buildings based on the current scientific literature is proposed. The process relies on the combination of different pattern recognition and classification algorithms in order to provide a robust insight of the energy usage patterns at different level and at different scales (from single building to stock of buildings). In this study several implications related to energy profiling in buildings, including tariff design, demand side management and advanced energy diagnosis are discussed. Moreover, a robust methodology to mine typical energy patterns to support advanced energy diagnosis in buildings is introduced. Finally, in order to demonstrate the scalability of the methodological process, an example of load characterisation for a thermal substation is presented.

© 2017 The Authors. Published by Elsevier Ltd.
Peer-review under responsibility of KES International.

Keywords: energy pattern recognition; building energy profiling; buiding energy management; load profiles characterization.

Capozzoli A., Piscitelli M.S., Brandi S. 2017. *Mining typical load profiles in buildings to support energy management in the smart city context*. Energy Procedia, 134 pp. 865–874. [11]



Energy 157 (2018) 336–352

Contents lists available at ScienceDirect

Energy

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

Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings

Alfonso Capozzoli*, Marco Savino Piscitelli, Silvio Brandi, Daniele Grassi, Gianfranco Chicco

Dipartimento Energia "Galileo Ferraris", Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

ARTICLE INFO

Article history:
Received 5 February 2018
Accepted 19 May 2018
Available online 21 May 2018

Keywords:
Energy consumption
Building energy management
Adaptive symbolic aggregate approximation
Anomaly detection
Data mining
Smart buildings

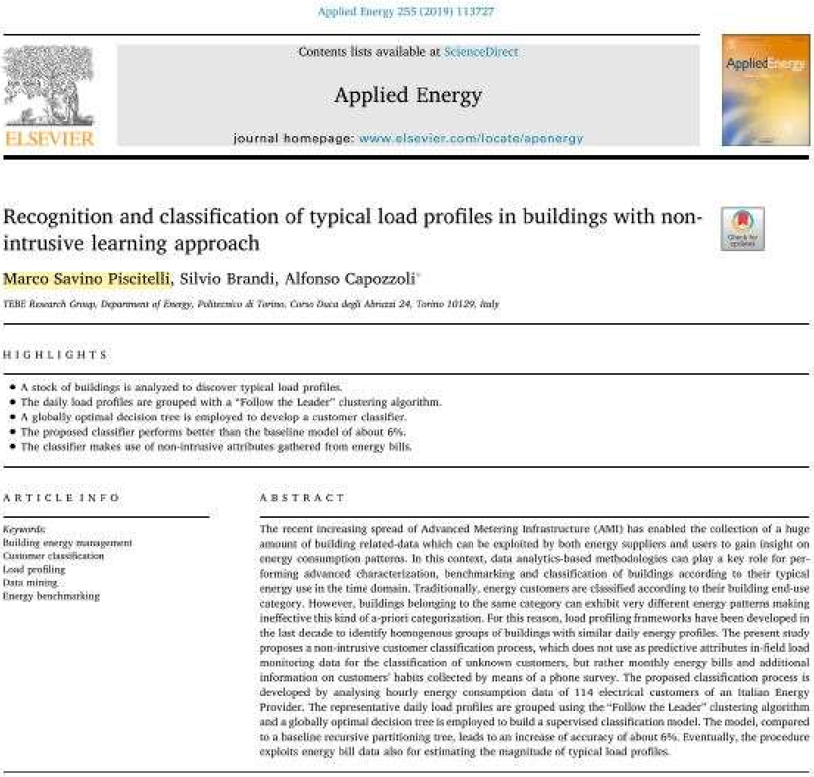
ABSTRACT

The energy management of buildings currently offers a powerful opportunity to enhance energy efficiency and reduce the mismatch between the actual and expected energy demand, which is often due to an anomalous operation of the equipment and control systems. In this context, the characterisation of energy consumption patterns over time is of fundamental importance. This paper proposes a novel methodology for the characterisation of energy time series in buildings and the identification of infrequent and unexpected energy patterns. The process is based on an enhanced Symbolic Aggregate approXimation (SAX) process, and it includes an optimised tuning of the time window width and of the symbol intervals according to the building energy behaviour. The methodology has been tested on the whole electrical load of buildings for two case studies, and its flexibility and robustness have been confirmed. In order to demonstrate the implications for a preliminary diagnosis, some unexpected trends of the total electrical load have also been discussed in a post-mining phase, using additional datasets related to heating and cooling electrical energy needs.

The process can be used to support stakeholders in characterising building behaviour, to define appropriate energy management strategies, and to send timely alerts based on anomaly detection outcomes.

© 2018 Elsevier Ltd. All rights reserved.

Capozzoli A., Piscitelli M.S., Brandi S., Grassi D., Chicco G. 2018. *Automated load patterns learning and diagnosis for enhancing energy management in smart buildings*. Energy, 157 pp. 336–352. [21]



Applied Energy 255 (2019) 113727

Contents lists available at ScienceDirect

Applied Energy

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

Recognition and classification of typical load profiles in buildings with non-intrusive learning approach

Marco Savino Piscitelli, Silvio Brandi, Alfonso Capozzoli*

TEBE Research Group, Department of Energy, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino 10129, Italy

HIGHLIGHTS

- A stock of buildings is analyzed to discover typical load profiles.
- The daily load profiles are grouped with a "Follow the Leader" clustering algorithm.
- A globally optimal decision tree is employed to develop a customer classifier.
- The proposed classifier performs better than the baseline model of about 6%.
- The classifier makes use of non-intrusive attributes gathered from energy bills.

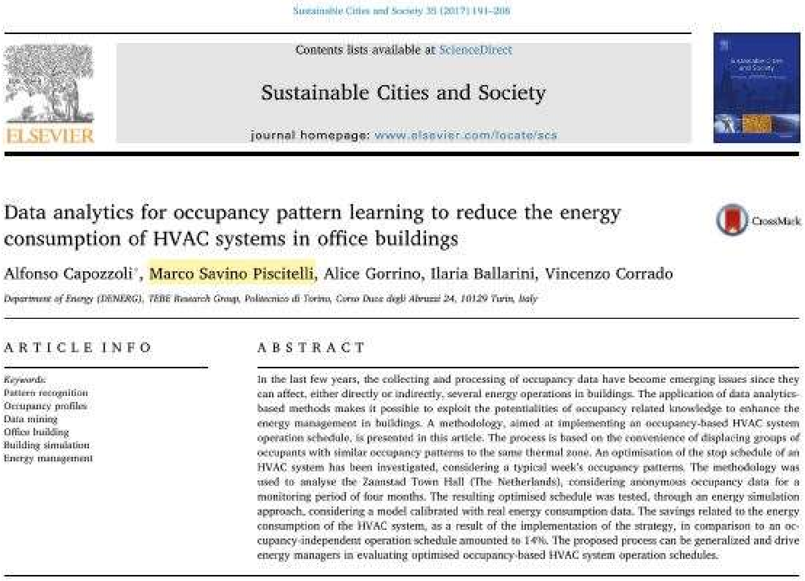
ARTICLE INFO

Keywords:
Building energy management
Customer classification
Load profiling
Data mining
Energy benchmarking

ABSTRACT

The recent increasing spread of Advanced Metering Infrastructure (AMI) has enabled the collection of a huge amount of building related-data which can be exploited by both energy suppliers and users to gain insight on energy consumption patterns. In this context, data analytics-based methodologies can play a key role for performing advanced characterization, benchmarking and classification of buildings according to their typical energy use in the time domain. Traditionally, energy customers are classified according to their building end-use category. However, buildings belonging to the same category can exhibit very different energy patterns making ineffective this kind of a-priori categorization. For this reason, load profiling frameworks have been developed in the last decade to identify homogenous groups of buildings with similar daily energy profiles. The present study proposes a non-intrusive customer classification process, which does not use as predictive attributes in-field load monitoring data for the classification of unknown customers, but rather monthly energy bills and additional information on customers' habits collected by means of a phone survey. The proposed classification process is developed by analysing hourly energy consumption data of 114 electrical customers of an Italian Energy Provider. The representative daily load profiles are grouped using the "Follow the Leader" clustering algorithm and a globally optimal decision tree is employed to build a supervised classification model. The model, compared to a baseline recursive partitioning tree, leads to an increase of accuracy of about 6%. Eventually, the procedure exploits energy bill data also for estimating the magnitude of typical load profiles.

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 pp. 113727. [10]



Sustainable Cities and Society 35 (2017) 191–208

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings

Alfonso Capozzoli*, Marco Savino Piscitelli, Alice Gorrino, Ilaria Ballarini, Vincenzo Corrado

Department of Energy (DENERG), TEBE Research Group, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy

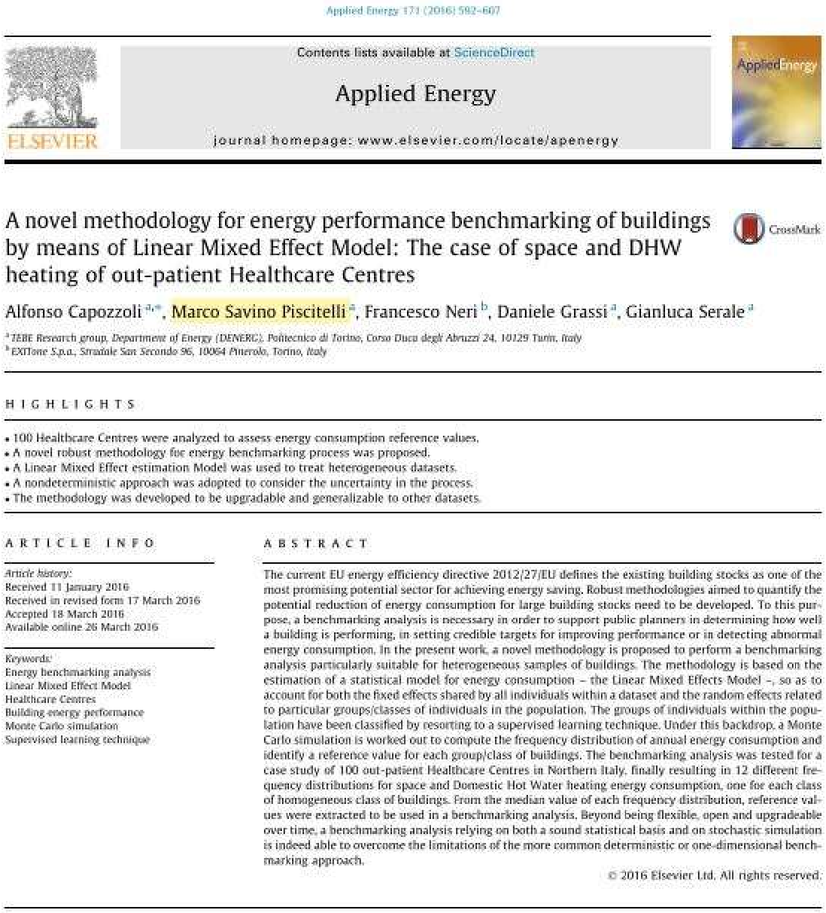
ARTICLE INFO

Keywords:
Pattern recognition
Occupancy profiles
Data mining
Office building
Building simulation
Energy management

ABSTRACT

In the last few years, the collecting and processing of occupancy data have become emerging issues since they can affect, either directly or indirectly, several energy operations in buildings. The application of data analytics-based methods makes it possible to exploit the potentialities of occupancy related knowledge to enhance the energy management in buildings. A methodology, aimed at implementing an occupancy-based HVAC system operation schedule, is presented in this article. The process is based on the convenience of displacing groups of occupants with similar occupancy patterns to the same thermal zone. An optimisation of the stop schedule of an HVAC system has been investigated, considering a typical week's occupancy patterns. The methodology was used to analyse the Zaanstad Town Hall (The Netherlands), considering anonymous occupancy data for a monitoring period of four months. The resulting optimised schedule was tested, through an energy simulation approach, considering a model calibrated with real energy consumption data. The savings related to the energy consumption of the HVAC system, as a result of the implementation of the strategy, in comparison to an occupancy-independent operation schedule amounted to 14%. The proposed process can be generalized and drive energy managers in evaluating optimised occupancy-based HVAC system operation schedules.

Capozzoli A., Piscitelli M.S., Gorrino A., Ballarini I., Corrado V. 2017. *Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings*. Sustainable Cities and Society, 35 pp. 191-208 [14].



Applied Energy 171 (2016) 592–607

Contents lists available at ScienceDirect

Applied Energy

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

A novel methodology for energy performance benchmarking of buildings by means of Linear Mixed Effect Model: The case of space and DHW heating of out-patient Healthcare Centres

Alfonso Capozzoli [a,*], Marco Savino Piscitelli [a], Francesco Neri [b], Daniele Grassi [a], Gianluca Serale [a]

[a] TEBE Research group, Department of Energy (DENERG), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy
[b] EXITone S.p.a., Stradale San Secondo 96, 10064 Pinerolo, Torino, Italy

HIGHLIGHTS

- 100 Healthcare Centres were analyzed to assess energy consumption reference values.
- A novel robust methodology for energy benchmarking process was proposed.
- A Linear Mixed Effect estimation Model was used to treat heterogeneous datasets.
- A nondeterministic approach was adopted to consider the uncertainty in the process.
- The methodology was developed to be upgradable and generalizable to other datasets.

ARTICLE INFO

Article history:
Received 11 January 2016
Received in revised form 17 March 2016
Accepted 18 March 2016
Available online 26 March 2016

Keywords:
Energy benchmarking analysis
Linear Mixed Effect Model
Healthcare Centres
Building energy performance
Monte Carlo simulation
Supervised learning technique

ABSTRACT

The current EU energy efficiency directive 2012/27/EU defines the existing building stocks as one of the most promising potential sector for achieving energy saving. Robust methodologies aimed to quantify the potential reduction of energy consumption for large building stocks need to be developed. To this purpose, a benchmarking analysis is necessary in order to support public planners in determining how well a building is performing, in setting credible targets for improving performance or in detecting abnormal energy consumption. In the present work, a novel methodology is proposed to perform a benchmarking analysis particularly suitable for heterogeneous samples of buildings. The methodology is based on the estimation of a statistical model for energy consumption – the Linear Mixed Effects Model –, so as to account for both the fixed effects shared by all individuals within a dataset and the random effects related to particular groups/classes of individuals in the population. The groups of individuals within the population have been classified by resorting to a supervised learning technique. Under this backdrop, a Monte Carlo simulation is worked out to compute the frequency distribution of annual energy consumption and identify a reference value for each group/class of buildings. The benchmarking analysis was tested for a case study of 100 out-patient Healthcare Centres in Northern Italy, finally resulting in 12 different frequency distributions for space and Domestic Hot Water heating energy consumption, one for each class of homogeneous class of buildings. From the median value of each frequency distribution, reference values were extracted to be used in a benchmarking analysis. Beyond being flexible, open and upgradeable over time, a benchmarking analysis relying on both a sound statistical basis and on stochastic simulation is indeed able to overcome the limitations of the more common deterministic or one-dimensional benchmarking approach.

Capozzoli A., Piscitelli M.S., Neri F., Grassi D., Serale G. 2016. *A novel methodology for energy performance benchmarking of buildings by means of Linear Mixed Effect Model: The case of space and DHW heating of out-patient Healthcare Centres*. Applied Energy, 171 pp. 592–607. [52]