DigitMan
Occupant-based DIGITal predictive MANteinance to improve the built environment

Grant number: 2022JMRX2A

WP 3 – Operation predictive module development

T.3.1 – Relevant operation KPIs definition
T.3.2 – Operation data acquisition and post-processing
T.3.3 – Operation predictive module development

DELIVERABLE ID	D3.1
Deliverable Title	KPIs and predictive methods for operation tasks
Date	26/02/2025
Last revision date	23/02/2025
Revision	002
Main partner	POLIMI
Additional partners	UNIBO, UNIVPM
Authors of the contribution	Marco D'Orazio, Graziano Salvalai, Roberto Villa, Riccardo Gulli, Angelo Massafra, Giorgia Predari
Deliverable type	Report and attachments
Number of pages	68

Abstract

This deliverable address operational challenges in building management within public administrations, focusing on university buildings. As local public administrations, universities must strategically manage facilities, balancing effectiveness, efficiency, and stakeholder needs. This includes daily maintenance to prevent degradation and ensure sustained performance. Therefore, a performance-based approach to built asset management is essential. This research introduces a Digital Decision Support System (DDSS) for strategic and operational performance-based management of existing buildings. The DDSS utilizes digital technologies and data techniques to establish a data environment providing insights into energy performance relative to planned occupancy. Supporting both "how-to" (operational) and "what-if" (strategic) analyses, the system aims to increase building managers' understanding of energy behavior. This is achieved by analyzing energy needs and connecting them to occupancy parameters using Key Performance Indicators (KPIs). This knowledge can inform decisions such as prioritizing energy-efficient spaces and identifying areas for energy efficiency improvements. The resulting knowledge informs decisions, prioritizing energy-efficient spaces and identifying improvement areas, with the ultimate goal of enhancing Indoor Environmental Quality (IEQ) while minimizing energy consumption.

Keywords

Building Energy Simulations (BEM), Building Information Models (BIM), Key Performance Indicators (KPIs), Machine Learning (ML), Building Operation

Approvals

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
UNIVERSITÀ POLITECNICA DELLE MARCHE
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
POLITECNICO MILANO 1863

Role	Name	Partner	Date
Coordinator	Marco D'Orazio	UNIVPM	26/02/2025
Task leader	Graziano Salvalai	POLIMI	26/01/2025

Revision versions

Revision	Date	Short summary of modifications	Partner
001	09/01/2025	Document creation and drafting	POLIMI
002	23/02/2025	Text editing and revision	UNIVPM

Summary

1 Introduction

This report focuses on operational issues that characterise building management practices within public administrations, as illustrated in the following pages. As justified in previous reports, the research focuses on university buildings, selected as they represent a key segment of the public administration-managed stock within the extensive, nationally administered built heritage.

University buildings are notable for their diverse functions, broad geographical spread, and the necessity to operate in line with both contextual needs and organisational strategies. As local public administrations, universities must define and implement strategies for managing facilities and infrastructure in line with strategic plans. Such strategies have to support institutional and managerial activities organically, aiming to balance the effectiveness and efficiency of the facilities with the needs and expectations of the personnel, students, and other stakeholders. On the other hand, university administrations must ensure their structures' ordinary and extraordinary maintenance daily to prevent degradation and maintain performance over time.

During all these strategic and operational processes, these institutions have to justify -with appropriate supporting documentation/tools- how they manage these aspects within a systemic perspective that takes into consideration various complex factors, including, among the many, functionality, safety, comfort, energy use, and costs, all areas transversally interconnected with the broader issues related to occupancy and sustainability. They must also continuously evaluate the availability of suitable structural and infrastructural resources for their schools, faculties, and departments (or equivalent units) to support teaching, research, third mission and social initiatives. This systematic assessment should be conducted at varying intervals, depending on the diverse categories of resources and the related critical factors. In this context, projecting built asset management towards a performance-based dimension becomes undoubtedly crucial.

As described in this text, this project's Digital Decision Support System (DDSS) aims to realise strategic and operational performance-based management of existing buildings by leveraging available digital technologies and combining them with efficient data integration, processing, and visualisation techniques. Specifically, this part of the research focuses on creating a data environment -i.e., a DDSS- capable of sharing knowledge about the energy performance of selected case studies and connecting this performance to information about planned occupancy conditions, to be used both in the short term ("How-To" logic) and long term ("What-If" logic). The goal is to assist building managers in becoming more aware of the energy behaviour of the buildings they manage while planning occupancy. Through the DDSS, this can be achieved by analysing the energy needs of different spaces and associating them with occupancy parameters using specific key performance indicators (KPIs). Such knowledge could lead, for instance, to managers prioritising using the most energy-efficient areas inside the building during appropriate hours and seasons and identifying improvement areas for enhancing energy efficiency.

2 State of the art

The Architecture, Engineering, Construction, and Operation (AECO) sector is responsible for a significant portion of the world's energy consumption and environmental impacts [1–3]. Despite extensive research in the field, this sector is characterised by a lack of comprehensive innovation and has limited potential for long-term transformative change due to its inherent nature [4].

A crucial perspective for developing a sustainable and energy-efficient built heritage involves optimising building operations in the short term, with a focus on reducing operational costs, minimising environmental impacts, and lowering energy consumption, all while maintaining comfort for occupants.

Addressing this issue is particularly important, considering that operational expenses and energy consumption can account for up to 75% of the total costs incurred during construction [5].

Advanced digital methods and tools, capable of generating valuable knowledge in the form of information, are therefore needed to provide decision support to building administrators [6]. In particular, filling the 'knowledge gap' in building management is crucially important in relation to energy management, among the numerous themes [7,8]. The complexity of such a framework demands a holistic approach to integrate both economic-financial asset management and technical-functional management within the context of performance-based strategies, aiming to the balance between functional requirements, energy demands, and environmental impacts, combining the 'static knowledge' of the containers -the buildings- with the 'dynamic knowledge' of the contents -the users and the activities inside them.

The Digital Twin (DT) paradigm is evolving to facilitate new methods of sharing existing building information, aiming at optimising cost-benefit during their usage, thereby addressing these challenges [9–12]. However, at present, public administrations and existing buildings are unprepared to integrate management processes with a full DT perspective, DTs being a highly complex technology still in scientific definition and buildings being scarcely equipped with advanced smart infrastructures. In this transitional period, the digitalisation of management processes should leverage the available technologies and competencies of the administrations, taking into account their level of smart readiness and following an incremental approach to digitalisation.

2.1 Energy related issues in university campuses

The pilots on which test the DDSS were chosen within the Italian university heritage. University buildings, in fact, are highly representative of the extensive research sample presented in WP1 given their unique functional complexity.

The management and usage of these facilities are directly linked to public administrations, which are typically the major owners of such assets. In their various and complex forms, these administrations are often characterised by systemic constraints that do not allow for a transformative capacity aimed at optimising costs and benefits from a performance-based perspective. This incapability often results in the inefficient use of public finances. Furthermore, due to their widespread presence across different territories, the usual decentralisation of information in analog formats makes it difficult to have an up-to-date and complete portfolio overview. This condition leads to difficulty in understanding the state of asset utilisation and planning corrective actions in line with strategic visions.

This part of the research focuses on analysing energy-related issues that influence higher education buildings during their operation. This topic is becoming increasingly relevant since, over the past few years, higher education institutions have globally set target goals for energy savings and emission reductions, leading to the implementation of numerous measures to reduce energy usage [13–15], including the installation of renewable energy sources, the renovation of older buildings, and the promotion of policies for increasing awareness of energy conservation practices among buildings' users. In addition to these measures, other interesting energy improvement strategies -often overlooked but significative for university buildings- can be considered, such as flexible working arrangements and demand-driven building system controls [16].

The energy usage and intensity of buildings on a higher education campus are influenced by several factors, including the climate, building systems, construction type, but also occupancy conditions [17]. Occupancy variables, such as the presence of students and staff members and the activities they do, can significantly impact energy consumption. For example, Gui et al [18], in 'Reducing university energy use beyond energy retrofitting: The academic calendar impacts', showed how the academic calendar, which regulates the occupancy condition of campus buildings, determines their energy consumption in Australian universities. In contrast, Mosteiro-Romero et al. [19] demonstrated how demand-driven building system

controls can save energy in university offices and how such strategies are important to adapting building use to climate change in summer conditions in Singapore.

2.2 Current methods for energy assessment

In recent decades, scientists and engineers have dedicated significant efforts to developing approaches for predicting energy consumption. These approaches can be broadly categorised into three types: building physical energy models (referred to as "white box" models), data-driven models (referred to as "black box" models), and hybrid models (referred to as "grey box" models) [20].

The first category of building energy models, known as the "white box" model, relies on detailed building parameters and heat balance equations. This approach involves modeling the physical characteristics of a building, such as its construction materials, insulation, ventilation systems, and thermal properties, and the contextual factors, such as solar radiation, weather conditions, and occupancy patterns. The modeling and calibration process of "white box" software poses significant challenges for building energy stakeholders due to the extensive input parameters required, leading to time-consuming development on a physical software platform and high simulation economic costs. However, when well calibrated, the physical models' prediction accuracy can be higher than the statistical models [21], as well as their interpretability.

Given the limitations of white box models and the rapid advancements in big data technologies like sub-metering and smart buildings, data-driven models have emerged as a viable alternative in the last decade. Black box models offer a simpler approach by capturing the linear and nonlinear relationships between input and output variables. The main research efforts in the last period focused on investigating deep learning techniques and optimising two key aspects: the significance of features to train models and the choice of algorithms. However, training these models and achieving accurate predictions under different conditions typically require vast amounts of historical data and a lengthy training period [20]. Moreover, while black box models have the advantage of needing less building information for their development, their prediction accuracy fluctuates, particularly when applied to different building scenarios. To address these challenges, a solution known as the "grey box" approach has been suggested by the literature. This method incorporates a simplified physical model and readily available data to simulate building energy demand, effectively combining the benefits of white and black box approaches.

White box models were employed in the presented application because of the absence of measurement data and their higher level of standardisation in already developed software ontologies (compared to the black box and grey box models currently available in literature). By adopting this approach, the connection between input and output for analysis becomes more comprehensible, particularly in terms of educating building managers about building behavior. However, some approximations were made due to the complexity of inputting the many input parameters required by Energy Plus for calculation, as discussed in the previous section.

2.2.1 White-box methods

Building performance simulations (BPS) assess building performance using computer-based mathematical models and applying fundamental physical principles and engineering techniques [22]. Simulation models can be distinguished as linear or nonlinear [23] BPS models are also called 'white box' models, since their approach involves calculating output variables by considering rules established b.y domain knowledge theories, inputting factors usually friendly to technicians, such as physical characteristics of buildings (construction materials, insulation, ventilation systems, and thermal properties, etc.) and contextual factors (solar radiation, weather conditions, occupancy patterns, etc.).

Laying its foundation in the 1970s, BPS has undergone significant evolution over several decades, as traced by Augenbroe et al. in 2002 [24] and later by Oh et al. [25]. According to them, two key aspects drove this technology's maturation: progressively increasing quality assurance and growing integration capabilities between simulation tools and expertise in the design process. Starting in the early 2000s, research efforts were directed towards addressing several key challenges in building simulation. These included the need to efficiently capture intricate geometries and complex systems, improving the integration of simulation with data-driven systems, developing new models capable of handling a wide range of time scales and additional physical phenomena, creating innovative methods for optimisation in the presence of uncertainty and risk, and making notable advancements in the verification and validation processes [26].

In recent times, the field of BPS is currently thriving, experiencing significant research and development, with increasing adoption in practical applications that span various aspects, including energy [27], daylight [28], thermal comfort [29], ventilation [30], indoor air quality [31], acoustics [32], structural and fire safety analysis [33]. These applications can range from the building or its components' scale to district and urban levels [34]. Substantial research efforts in BPS have been focused on three pivotal areas:

- addressing the 'performance gap;'
- enhancing interoperability between BPS and BIM;
- improving calculations thanks to AI.

2.2.1.1 Performance gap

BPS simulations are often affected by the so-called 'performance gap' [35,36]. It means there is a difference between the predicted and the actual performance of a building during its operation. This gap can arise due to uncertainties, wrong assumptions made during the modelling process, or discrepancies between the design and the as-built statuses. Its magnitude can be significant, as scientific reports point to the measured energy consumption potentially reaching 2.5 times the anticipated energy usage [37].

Recently, the issue has become increasingly important, especially concerning the energy question. Indeed, the rapid deployment of automated energy meter reading technology, capable of collecting data at hourly or even sub-hourly intervals, has made the performance gap more visible. In this field, the performance gap may erode trust and nurture scepticism among the actors involved in design and management of buildings. It is crucial to bridge this gap for two key reasons. Firstly, the construction industry needs to prove that both new and existing buildings sustain optimal performance over their entire lifespan. Secondly, ensuring optimal performance is a core prerequisite for adopting innovative FM strategies, like performance-based building management and performance contracting [35].

The most used technique for addressing the performance gap and enhancing the credibility of simulations is known as 'model calibration.' The findings from Chong et al. clearly indicate a notable increase in the adoption of calibration approaches, which can be both automated (such as Bayesian calibration and global sensitivity analysis) and manual (involving detailed audits, expert knowledge, and evidence-based procedures) [36]. As per the authors, energy performance simulation models are typically calibrated based on one or two observed outputs, with the primary sources of data for the calibration process being monthly electricity consumption data obtained from utility bills and hourly indoor dry bulb temperature data monitored though sensors and the thermophysical properties of the building envelope, infiltration rates, internal heat gain densities, and indoor temperature setpoints being the input parameters most used for the calibration. However, it is worth noting that the calibration of BPS models remains challenging in practice due to the absence of clear guidelines and best practices, as well as issues related to reproducibility.

2.2.1.2 BIM-BPS interoperability

With the rise of BIM, BPS has started to be integrated into the design and management practice as part of more complex models that are not limited to a single discipline. This integration optimises the information flow and provides a holistic view of projects for both new constructions and existing buildings [38]. However, there is still a lack of bidirectional interoperability between BPS and BIM tools [39,40]. Current practices often involve creating a second unlinked digital model starting from the BIM for conducting performance simulations. Splitting models can lead to loss of information, repetitive and time-consuming tasks, and miscoordination between designers and performance analysts, ultimately hindering the design process, especially in the early stages.

Among the various topics, the interoperability issue between BIM and Building Energy Modelling (BEM) has been the subject of numerous research studies. These studies concur that many unresolved issues in developing BIM-based building energy modelling still lead to an unoptimal data flow integration [41], which is still a challenge to solve.

For instance, digital environmental simulations can present a computational bottleneck concerning the complexity of geometry [42]. BIM-based simulation models often result in high polygon counts, making simulations more time-consuming and less controlled. BPS tools usually require models with regular squared mesh for tasks like daylight analysis, computational fluid dynamics, or safety pathfinding simulations. Furthermore, BIM and BEM systems might employ different geometry kernels and data dictionaries, impacting their performance and compatibility with various software tools [43]. This discrepancy in underlying data structures can hinder smooth data exchange and integration between the two systems, further complicating the interoperability challenge.

Another gap is related to the available data formats for BIM and BEM. Typically, data transmission between BIM and BEM is performed through two file formats: IFC and Green Building Extensible Markup Language (gbXML) [44]. Both have their unique advantages. IFC is considered the standard format for information exchange in BIM. On the other hand, gbXML is a format based on rectangular-shaped surfaces and their attributes rather than objects and can store nearly all the building information needed for energy simulation. IFC-gbXML integration is not straightforward because these modelling systems rely on different approaches, languages, and protocols, sometimes incompatible. Although academic research has made progress in improving BIM-to-BEM interoperability through IFC and gbXML through 'export-import' procedures, achieving bidirectional data exchange between BIM and BEM through available software is still challenging.

2.2.2 Gray-box methods

In recent times, the need to develop applications for asset management or design has sparked a subsequent need for increasingly faster performance calculations, capable of being developed into real and proper applications and promptly responding to the requests of application users. From this, together with the development of AI, new approaches to performance modelling have emerged to enable faster calculations.

In addition to building simulation models (referred to as 'white box' models), data-driven models (referred to as 'black box' models) and hybrid models (referred to as 'grey box' models) have been introduced [45]. These solutions are valuable alternatives to traditional building performance simulation, but they all have limitations.

The choice between them varies according to the needs. Black box models offer a simpler approach by capturing the linear and nonlinear relationships between input and output variables. Nonetheless, training these models and achieving accurate predictions under different conditions typically require vast amounts of historical data and a lengthy training period. Moreover, while black box models have the advantage of

needing less building information for their development, their prediction accuracy fluctuates, particularly when applied to different building scenarios. To overcome this limitation, the 'grey box' approach incorporates a simplified physical model and readily available data to simulate building energy demand, effectively combining the benefits of white and black box approaches. Grey box and black box models can speed up the calculation time and are effectively usable within tools, but they are less controllable and, sometimes, difficult to train. In contrast, when well calibrated, white box models' prediction accuracy and interpretability can be higher than the statistical models [46]. Nevertheless, white box software's modelling and calibration process poses significant challenges for building energy analysts due to the extensive input parameters required, leading to time-consuming modelling processes.

2.2.3 Black-box methods

As black-box models, data-focused structures are progressively seen as crucial elements in BPS to address the dynamic needs concerning asset management, design optimization, and timely energy efficiency [47,48]. In contrast to "white-box" methodologies, which require the formulation of highly intricate physical models pertaining to geometry and boundary conditions, or "grey-box" approaches that endeavor to strike a balance between fundamental physical principles and statistical inference, the black-box paradigm functions without explicit understanding of the system's internal mechanisms, concentrating solely on the discernible linear and non-linear correlations between system inputs and outputs.

Through the analysis and examination of a dataset that includes historical time-series data reflecting various operational conditions and building behaviors, diverse algorithms are capable of extracting knowledge from the data. This assimilated knowledge empowers these algorithms to predict future performance based on the input data provided. Contextual information, such as temperature, occupancy levels, equipment utilisation, and set points, serves as input variables to forecast fluctuations in energy consumption. Moreover, these models can offer insights into the variability of parameters once energy consumption is established.

As previously articulated, "black-box" methodologies encompass automated learning techniques, commonly referred to as "machine learning," alongside statistical analysis. The most rudimentary form of machine learning approaches comprises "Regression" models, ranging from "Linear" regression, which seeks to establish a linear correlation between input and output parameters, to more intricate models such as "Polynomial" regression and "Non-Linear" regression that encapsulate more sophisticated relationships. "Multiple" regression [49] facilitates the assessment of the simultaneous impact of various factors. "Regularised" regression techniques serve to mitigate over-complexity and overfitting, thereby preserving generalizability. Taking cues from the human brain's layout, Artificial Neural Networks [50–52] utilise linked 'neurons' to interpret information through a non-linear process. These networks are particularly adept at modeling intricate systems and discerning non-linear relationships, such as predicting thermal loads in buildings based on meteorological data, occupancy patterns, and equipment schedules. A variety of ANN architectures exist, including Feedforward Networks, where information flows unidirectionally through the layers of neurons. Time-series data is particularly suited for Recurrent Neural Networks (RNNs) because of their interconnected neurons that form cycles, facilitating the storage of historical insights. Long Short-Term Memory (LSTM) networks adopt a similar structure to RNNs but are more proficient at managing long-term recurrences, thereby circumventing the "vanishing gradient" issue inherent in this algorithm. Deep Learning (DL) harnesses networks featuring numerous obscured layers to obtain abstract data representations. Support Vector Machines (SVMs) ascertain an optimal hyperplane for data segregation or, in the case of regression, the approximation of input-output relationships, with the "kernel trick" facilitating the management of non-linear relationships. Decision Trees (DTs) construct a sequential order of decisions predicated on input values, while Random Forests (RFs) amalgamate predictions from multiple trees to enhance robustness. Gradient Boosting sequentially integrates trees

[49,53–55]. Gaussian Processes (GPs) provide a probabilistic framework for predictions, quantifying uncertainty in a manner distinct from other methodologies. Time series analysis employs models such as ARIMA and SARIMA, which are grounded in the statistical examination of the structure of time series data.

The black-box approach offers the advantage of requiring few parameters for model development. The primary limitation of this methodology lies in the quantity and quality of the data. Typically, trained models rely on time-series data that are limited in temporal span, providing information and observations for only a restricted period and thus sampling a very small number of scenarios. Similarly, the predictive model is sensitive to the building's boundary conditions, such as climate. The building's usage type is also a significant factor influencing usage profiles and energy consumption.

2.3 KPIs for operations What-If/How-To scenarios

For the definition of the KPIs, the requirements of the "What-If" and "How-To" logics, as explained in Section 2.1 – Logics and functionalities of deliverable D1.1 - Definition of an occupant-centric conceptual framework and correlation methodology, were first analysed. The "What-If" logic deals with medium- to long-term predictions to support future decisions. The "How-To" logic deals with short-term predictions to support real-time decisions, based on historical data and designed to interact with real-time information. In the development of the KPIs, five main disciplinary areas were identified: Use, Energy, Costs, Environment, and Well-being. Based on these five macro-areas, several specific objectives were identified, relating to one or both of the aforementioned logics. The results of the objective identification process are presented in the Table 1.

Table 1: Tables of Operation KPIs identified by DIGITMAN.

Code	Name	Area	Logic	Description	Goal
KPI O-1.1	Utilisation Rate	Use	What-If / How-To	Percentage of time that a specific space is occupied compared to the total available time.	It helps determine how long the space is being utilized.
KPI O-1.2	Occupancy Rate	Use	What-If / How-To	Average number of occupants in a space compared to its maximum capacity.	It provides insights into the utilization of individual space and identifies spaces that may be over or underutilized.
KPI O-1.3	Real Time Occupancy Rate	Use	How-To	Measurement of the variation of the occupancy in the spaces in timeseries to identify irregular occupancy pattern	Identify the occupation in real time through both measurement and predictions
KPI O-1.4	Standard Deviation Occupancy Variabiliity	Use	How-To	Measurement of the variation of the occupancy in the spaces in timeseries to identify irregular occupancy pattern	Identify irregular patterns in space use
KPI O-2.1	Energy Need	Energy	What-If	Tons of oil equivalent to the ideal energy needed for heating, cooling, lighting, ventilationg and using the appliances of a building or its zones.	It helps understaing how much electricity is needed for operating the zoneor building.
KPI O-2.2	Electricity Power Need at Peak	Energy	What-If	The highest level of electricity power used within a specific time period,	It enables better load balancing within the building and optimisation the

Code	Name	Area	Logic	Description	Goal
				typically measured in kilowatts (kW).	distribution of energy loads to avoid overloading any particular circuit, system, or equipment.
KPI O-2.3	Natural Gas Power Need at Peak	Energy	What-If	The highest level of heating power used within a specific time period, typically measured in kilowatts (kW).	It enables better load balancing within the building and optimisation the distribution of energy loads to avoid overloading any particular circuit, system, or equipment.
KPI O-2.4	Electricity Consumption per Occupant	Energy	How-To	Measurement of the efficiency of the devices per person	Spot peak consumption of energy in different conditions of use per person
KPI O-3	Energy Costs	Costs	What-If	Measures the expenses associated with electricity, natural gas, and district heating consumption within a building or facility.	It allows for identifying areas of higher energy consumption to implement energy-saving measure and reduce overall energy expenses towards energy efficiency.
KPI O-4	CO_2 Emissions due to Energy Use	Environment	What-If	Amount of CO2 emissions produced as a result of the building's energy consumption from the grid.	It allows for tracking of the environmental impact in terms of CO emissions due to the use of energy within a building or its zones.
KPI O-5.1	Thermal-Hygrometric Setpoint	Well-Being	How-To	Percentage of time during working hour within the comfort tresholds	
KPI O-5.2	Thermal-Hygrometric Rate	Well-Being	How-To	Percentage of time during the day which the indoor temperature setpoints do not meet the desired conditions while the space is occupied. It could also be a comfort measure as PPD or PMV.	To ensure occupant comfort, productivity, and well-being. To identify areas where the HVAC or environmental control systems may need adjustments or improvements to maintain desired comfort levels
KPI O-5.3	Visual Setpoint	Well-Being	How-To	Percentage of time during working hour within the comfort tresholds	
KPI O-5.4	Visual Rate	Well-Being	How-To	The percentage of time that sufficient visual comfort is ensured, such as adequate workplane illuminance or appropriate lighting levels, during the considered period while the space is occupied.	Ensure that occupants have access to suitable lighting conditions that support their visual tasks and promote visual comfort
KPI O-5.5	Air Quality Setpoint	Well-Being	How-To	Percentage of time during working hour within the comfort tresholds	
KPI O-5.6	Air Quality Rate	Well-Being	How-To	Percentage of time during which the CO2/TVOC concentration in the indoor air exceeds the desired	To improve indoor air quality and avoid potential health issues. By ensuring proper ventilation, air circulation, and control of pollutant

Code	Name	Area	Logic	Description	Goal
				setpoint while the space is occupied	sources, the aim is to maintain CO2/TVOC concentrations within recommended levels for occupant comfort and well-being.

3 What-If logic

3.1 Methods and Tools

3.1.1 Overview

The proposed methodology for data collection in digital models is organised into several phases and is interrelated with other WPs' activities. In particular, the general development phases are:

1. Topological BIM (addressed in WP1);
2. BIM-to-BEM (addressed in this WP3's report);
3. BIM-to-BSM (addressed in this WP4);
4. Multicriteria analysis (to be addressed in WP5).

The first, conducted in WP1 and briefly summarised below for ease of reading, involves semi-automatically creating the BIM of some case study buildings, enriched with all the information necessary for energy and safety analyses, also called Topological BIM (TBIM). These models are semi-automatically generated with the objective of being compliant with BPS and BSM.

The second step comprised the automated BIM-to-BEM (Building Energy Model) conversion.

Similarly, the third step consists of the transformation of the BIM into a BSM (Building Safety Model), described in D4.1 – KPIs and predictive methods for safety tasks.

In the final phase, simulations will be run, and the results will be interpreted in an integrated manner using project-defined KPIs to identify the optimal management strategies regarding space use, energy needs, and occupant safety.

The first and third phases are presented in this report, while the last, under development, will be the subject of future reports (WP5).

3.1.2 Topological BIM

The BIM process was conducted leveraging the Topological BIM (TBIM) process, documented in WP1's report D1.1 – Definition of an occupant centric framework, which allowed us to semi-automatically generate a BIM rich with all the information needed for safety verifications. The process is briefly summarized below.

3.1.2.1 3D modeling

The initial substep is to model the building's geometry. This is done by creating a closed 3D BRep object for each building space by retracing the administrators' CAD drawings in a 3D modelling environment. This volume, representing the gross shape of the space boundary, is then transformed into a Topologic "cell", serving as the foundational spatial unit in the digital model.

3.1.2.2 Topology modelling

The topology modelling substep establishes the topological relationships between the model's core elements. Thanks to Topologicpy, the cells are combined into a higher-level spatial entity known as the "cell complex", a digital model consisting of topologically interconnected spaces and binding surfaces ("Collector Model") to transform the geometric cells into topological cells. Although the cells do not

contain any data at this point, they are prepared to be populated with information. For this reason, they are called "Informational Collectors," serving as the primary data aggregators in the modelling process.

3.1.2.3 Information enrichment

In this subphase, conditional modelling is used to assign information to the elements within the Collector Model.

First, functional data is added to the Collectors by attributing "Informational Load Dictionaries" (ILDs) to them. ILDs are JSON dictionaries, each representing a specific space occupancy type (e.g., office, classroom, corridor, etc.) and containing relevant operational and safety data (e.g., thermal setpoints, electricity loads, area per occupant, etc.). To enrich a Collector, a space occupancy type is assigned to it, choosing between the occupancy types modelled in the ILDs, and the corresponding ILD is transferred to the corresponding cell, enriching it with the ILD's embedded data. The relevant data added to the spaces concerning operation evaluations are reported in Table 2.

Table 2. Operation properties added to the spaces of the topological model in the information enrichment step

Property Name	Description	Quantity	Unit
pr_ArtificialLighting	Indication whether this space requires artificial lighting (as natural lighting would be not sufficient). (TRUE) indicates yes (FALSE) otherwise.	Boolean	-
pr_EquipmentPowerDensity	This is typically the maximum electrical power input (in Watts/m2) to using electric appliances in a zone, including PCs, plotters, elevators and other equipment, if present. This value is multiplied by a schedule fraction to get the electric power in a particular timestep.	Power Density	W/m2
pr_Illuminance	Required average illuminance value for this space.	Illuminance	lux
pr_IsCooled	Indication whether this space requires air conditioning provided (TRUE) or not (FALSE).	Boolean	bool
pr_IsHeated	Indication whether this space requires heating provided (TRUE) or not (FALSE).	Boolean	bool
pr_IsMechanicallyVentilated	Indication whether the space is required to have mechanical ventilation (TRUE) or not (FALSE).	Boolean	bool
pr_IsNaturallyVentilated	Indication whether the space is required to have natural ventilation (TRUE) or not (FALSE).	Boolean	bool
pr_IsOccupied	Indication whether the space is permanently occupied (TRUE) or not (FALSE) according to energy modeling purposes. For examples, offices and classrooms are permanently occupied, while circulation spaces or storage spaces not.	Boolean	-
pr_LightingLevel	This is typically the maximum electrical power input (in Watts) to lighting in a zone, including ballasts, if present. This value is multiplied by a schedule fraction to get the lighting power in a particular timestep.	Power	W
pr_LightingPowerDensity	This is typically the maximum electrical power input (in Watts/m2) to lighting in a zone, including ballasts, if present. This value is multiplied by a schedule fraction to get the lighting power in a particular timestep.	Power Density	W/m2

Property Name	Description	Quantity	Unit
pr_MechanicalVentilationRate	*Indication of the requirement of a particular mechanical air ventilation rate, given in air changes per hour. It is calculated during the DigitMan's BIM modelling process.*	*AirChanges*	*[h-1]*
pr_MechanicalVentilationRatePerPerson	*MechanicalVentilationRatePerPerson Standard value for mechanical ventilation of spaces given by UNI EN 10339. For university classrooms, it is equal to 7 L/s per person.*	*AirChanges*	*[L/s pp]*
pr_MechanicalVentilationRatePerVolume	*MechanicalVentilationRate standard value for mechanical ventilation of special spaces (e.g. toilets) given by UNI EN 10339. It should be filled when MechanicalVentilationRatePerPerson is not identifiable according to UNI EN 10339.*	*AirChanges*	*[h-1]*
pr_NaturalVentilationRate	*Indication of the requirement of a particular natural air ventilation rate, given in air changes per hour. It is calculated during the DigitMan's BIM modelling process.*	*AirChanges*	*[h-1]*
pr_OccupancyDensityOperation **	*Design occupancy loading for this type of usage assigned to this space according to UNI EN 10339.*	*Occupancy Density*	*pp/m2*
pr_OccupancyNumber **	*Number of people required for the activity assigned to this space.*	*People Count*	*pp*
pr_OccupancyType *	*Occupancy type for this object. It is defined according to DigitMan's classification system.*	*Text*	-
pr_SpaceHumidityMax	*Max humidity of the space or zone that is required from user/designer view point*	*Relative Humidity*	*%*
pr_SpaceHumidityMin	*Min humidity of the space or zone that is required from user/designer view point*	*Relative Humidity*	*%*
pr_SpaceTemperatureSummerMax	*Maximal temperature of the space or zone for the hot (summer) period, that is required from user/designer view point.*	*Temperature*	*°C*
pr_SpaceTemperatureSummerMin	*Minimal temperature of the space or zone for the hot (summer) period, that is required from user/designer view point.*	*Temperature*	*°C*
pr_SpaceTemperatureWinterMax	*Maximal temperature of the space or zone for the cold (winter) period, that is required from user/designer view point.*	*Temperature*	*°C*
pr_SpaceTemperatureWinterMin	*Minimal temperature of the space or zone for the cold (winter) period, that is required from user/designer view point.*	*Temperature*	*°C*
pr_SpaceWindLoadRating	*Wind load resistance rating for the window and doors located in the space. It is provided according to the national building code (EN15254 for DigitMan).*	*Text*	-

* The data on occupancy types was provided by the building administration and verified through on-site inspections. Subsequently, it was aligned with the OmniClass notation (Table 13); see WP1's report D1.2 – Integrated structuring of a common database.

** For special spaces like classrooms, occupancy properties were set manually space by space and considered equal to the number of seats within the space.

Next, after adding data to the cells, ILD information is transferred to the adjacent faces, binding the cell by executing topological queries. For instance, a partition wall is informed of the occupancy types of the spaces it delimits (e.g., "corridor-office") and respective data (e.g, "unheated-heated").

The faces then undergo additional data enrichment using the "Informational Rulesets" (IRSs), mainly containing construction data about envelope components (e.g., thickness, material, and thermal properties of walls, floors, roofs and openings). These key-value dictionaries contain "conditions" and "styles" applicable to faces. The conditions specify the property values a face should have for applying the IRS to it, while the styles define the new data to be assigned to the face if it meets the conditions. The assignment of IRS data is also carried out through topological queries. Each IRS is applied iteratively to each face within the Collector Model. The condition values are accessed and compared to the face's properties for each face. If a match occurs, the dictionary containing the style data is added to the face; if not, the iteration proceeds to the next face. The outcome is the so-called "Style Model", a Topologic cell complex containing the ILDs' data (operational data) and IRSs' data (construction data).

Relevant operational properties added to the doors thanks to topological and conditional modeling are reported in Table 3.

Table 3: Operation properties automatically added to the doors of the topological model

Property Name	Description	Quantity	Unit
pr_ConstructionName	*Name of the set of materials associated with walls, roofs, floors, windows, and doors modeled in BIM*	*IfcLabel*	-
pr_IsExternal	*Indication whether the element is designed for use in the exterior (TRUE) or not (FALSE). If (TRUE) it is an external element and faces the outside of the building*	*IfcBoolean*	-
pr_MainExposure	*Describes the exposure of the surface by choosing between N, W, E, S, and N-E, N-W, S-E, S-W*	*IfcLabel*	-
pr_SurfaceType	*Type of surface according to EnergyPlus: can be Wall (Exterior, Adiabatic, Underground, Interzone), Roof, Ceiling (Adiabatic, Interzone), Floor (GroundContact, Adiabatic, Interzone), Window, Door*	*IfcLabel*	-
pr_ThermalTrasmittance	*Thermal transmittance coefficient (U-Value) of an element.*	*IfcThermalTransmittanceMeasure*	W/m2K
pr_TopologicalType	*Describes the interface element from a topological perspective (internal vertical, external vertical, internal horizontal, top horizontal, bottom horizontal for surfaces, door, hole, or window for openings)*	*IfcLabel*	-

3.1.2.4 BIM modelling

To finalise the BIM, the apertures are created. They include doors, holes, and windows. Doors represent apertures allowing for horizontal passage between adjacent cells on the same storey, while holes for vertical passage (e.g., between staircases). Windows, instead, links the cells to the external environment. Such apertures are modelled as face elements in Topologicpy, based on the IRS data associated with the faces hosting them.

Subsequently, the Style Model is converted into a Topologic graph to perform graph analysis and detect and correct any errors in the topology modelling process. The outcome is the TBIM, a Topologic cell complex semi-automatically populated with data relevant to BPS analysis, i.e. using BSM. The components in this model (i.e., cells, faces, and apertures) form a network of interconnected objects suitable for direct transformation into BEM and BSM.

As the last substep of the BIM phase, using pyRevit and aligning Topologic's class hierarchy with Revit's and IFC's element classes, the Topologic TBIM is transformed as an Autodesk Revit model. Following this, minor manual adjustments are made to specific instance objects in Revit and direct IFC export is performed. For example, TBIM's apertures, by default placed at the centre of faces, are repositioned as needed, and any errors in construction data assignments to faces and apertures are corrected.

Moreover, the properties that cannot be represented in the ILDs on a functional basis and need to be assigned space by space are added to spaces (such as occupancy properties of special spaces, like classrooms). All these modifications are synchronised with the Topologic model. The outcome is a streamlined BIM model, available in Revit, IFC, and Topologic JSON formats, containing all the essential information for energy and safety analyses.

The properties added to the doors are reported in Table 4, while the output BIM of POLIMI's case study is shown in Figure 1.

Table 4: Safety properties manually added to the doors of the BIM model

Property Name	Description	Quantity	Unit
pr_Area	*Total area of the outer lining of the window.*	*IfcAreaMeasure*	*m2*
pr_FrameMaterial	*Material of the frame of the opening*	*IfcLabel*	*-*
pr_FrameThermalTrasmittance	*Thermal transmittance of the frame of the opening*	*IfcThermalTransmittanceMeasure*	*W/m2K*
pr_FrameThickness	*Thickness of the frame of the opening*	*IfcPositiveLengthMeasure*	*mm*
pr_Glass1Thicnkess	*Thickness of the first (inner) glass layer.*	*IfcPositiveLengthMeasure*	*mm*
pr_Glass2Thicnkess	*Thickness of the second (intermediate or outer) glass layer.*	*IfcPositiveLengthMeasure*	*mm*
pr_Glass3Thicnkess	*Thickness of the third (outer) glass layer.*	*IfcPositiveLengthMeasure*	*mm*
pr_GlassLayers	*Number of glass layers within the frame. E.g. "2" for double glazing.*	*IfcCountMeasure"*	*-*
pr_GlassThermalTrasmittance	*Thermal transmittance coefficient (U-Value) of the gas.*	*IfcThermalTransmittanceMeasure*	*W/m2K*
pr_GlazingAreaFraction	*Fraction of the glazing area relative to the total area of the filling element. It shall be used, if the glazing area is not given separately for all panels within the filling element.*	*IfcNormalisedRatioMeasure"*	*-*
pr_Height	*Total outer heigth of the window lining*	*IfcPositiveLengthMeasure"*	*m*
pr_IsExternal	*Indication whether the element is designed for use in the exterior (TRUE) or not (FALSE). If (TRUE) it is an external element and faces the outside of the building*	*IfcBoolean*	*-*
pr_MainExposure	*Describes the exposure of the surface by choosing between N, W, E, S, and N-E, N-W, S-E, S-W*	*IfcLabel*	*-*
pr_Perimeter	*Total area of the outer lining of the window.*	*IfcPositiveLengthMeasure*	*m*
pr_ShadingCoefficient	*(SC): The measure of the ability of a glazing to transmit solar heat, relative to that ability for 3 mm (1/8-inch) clear, double-strength, single glass. Shading coefficient is being phased out in favor of the solar heat gain coefficient (SHGC), and is approximately equal to the SHGC multiplied by 1.15. The shading coefficient is expressed as a number without units between 0 and 1.*	*IfcNormalisedRatioMeasure*	*-*
pr_SolarHeatGainTransmittance	*(SHGC): The ratio of incident solar radiation that contributes to the heat gain of the interior, it is the solar radiation that directly passes (Tsol or τe) plus the part of the absorbed radiation that is*	*IfcNormalisedRatioMeasure*	*.*

Property Name	Description	Quantity	Unit
	distributed to the interior (qi). The SHGC is referred to also as g-value (g = τe + qi).		
pr_SurfaceType	*Type of surface according to EnergyPlus: can be Wall (Exterior, Adiabatic, Underground, Interzone), Roof, Ceiling (Adiabatic, Interzone), Floor (GroundContact, Adiabatic, Interzone), Window, Door*	*IfcLabel*	-
pr_ThermalTrasmittance	*Thermal transmittance coefficient (U-Value) of the window.*	*IfcThermalTransmittanceMeasure*	*W/m2K*
pr_VisibleLightReflectance	*Fraction of the visible light that is reflected by the glazing at normal incidence. It is a value without unit.*	*IfcNormalisedRatioMeasure"*	-
pr_VisibleLightTrasmittance	*Fraction of the visible light that passes the object at normal incidence. It is a value without unit.*	*IfcNormalisedRatioMeasure"*	-
pr_Width	*Total outer width of the window lining.*	*IfcPositiveLengthMeasure"*	*m*
pr_WindLoadRating	*Wind load resistance rating for this object. It is provided according to the national building code.*	*IfcLabel"*	-

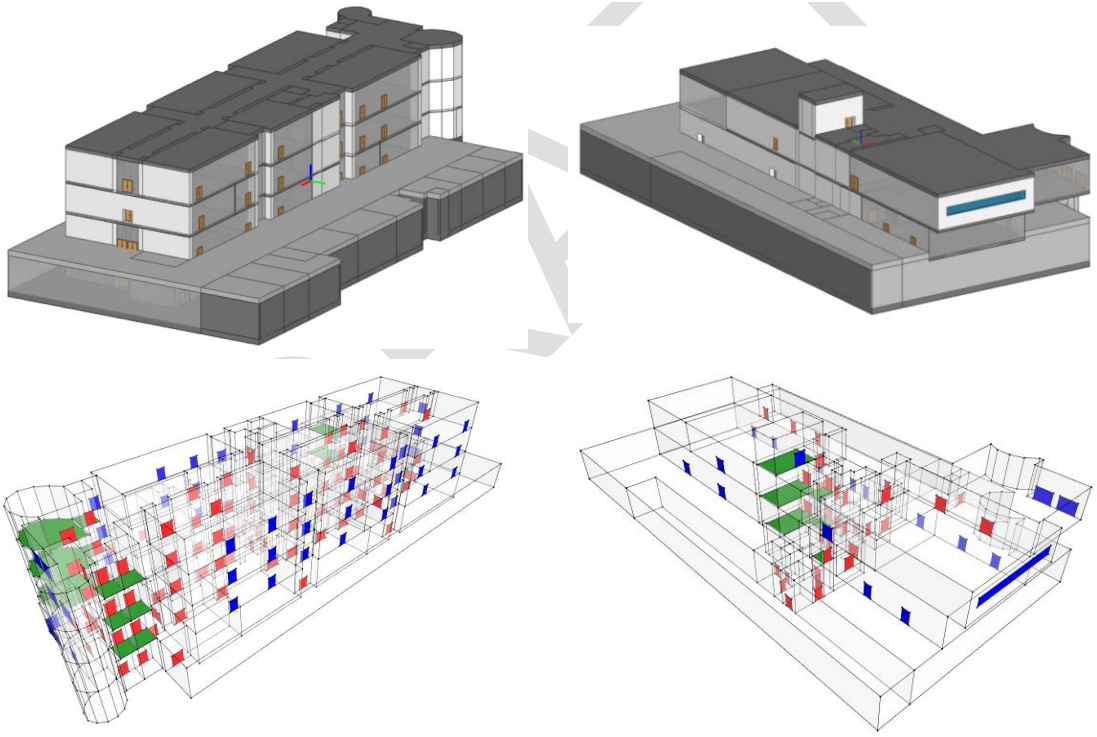

Figure 1: Case study BIM Model (left IFC, right Topologic).

3.1.3 BIM-to-BEM

The BIM-to-BEM conversion uses a Python algorithm that aligns Topologic's and IFC's element classes and properties with those of Ladybug Tools (via the Honeybee Energy APIs) and EnergyPlus (via the Eppy library). Alignment of the ontologies in DIGITMAN is illustrated in Figure 2, while the BIM-to-BEM process, depicted in Figure 3, is explained below.

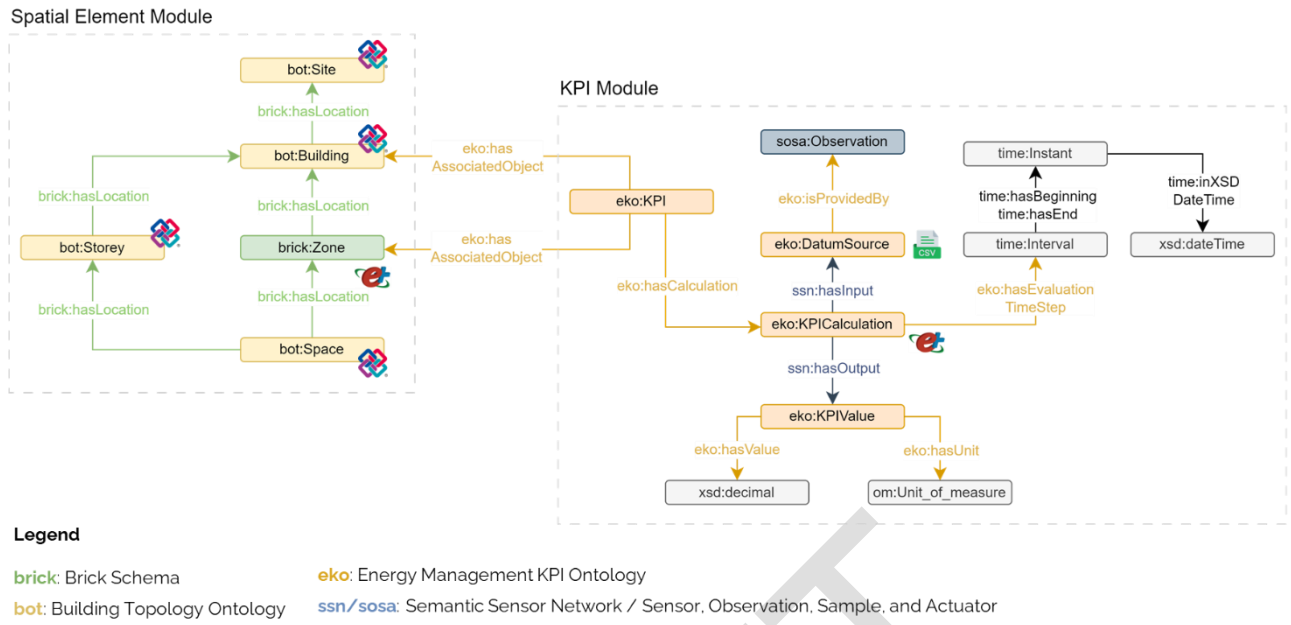

Figure 2. BIM-to-BEM workflow.

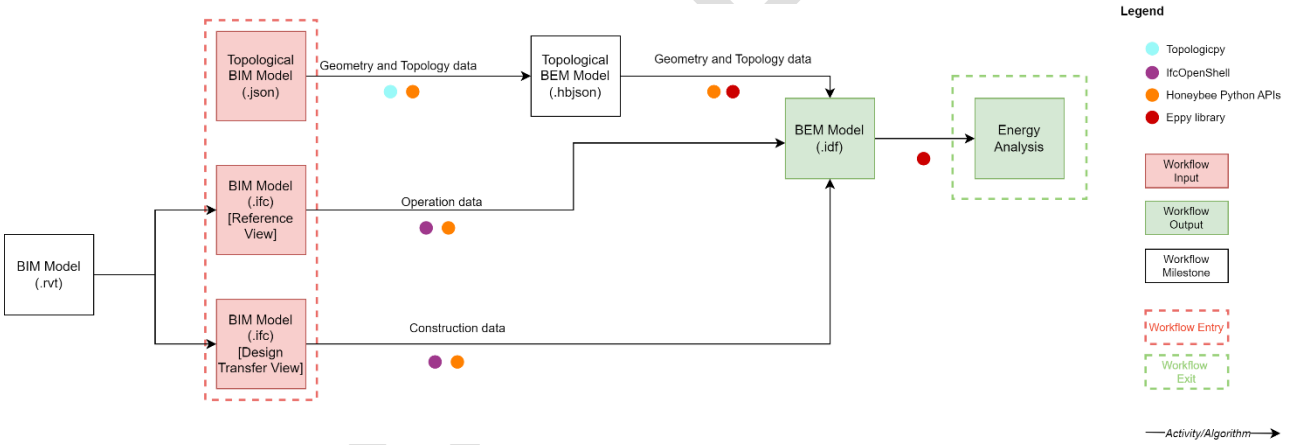

Figure 3. BIM-to-BEM workflow.

3.1.3.1 Zoning and geometry

The Topologic TBIM is used to model the geometry of the BEM.

First, the TBIM cells are aggregated into Topologic cluster objects representing the BEM's thermal zones. The zoning process makes use of the topological methods to cluster neighbour cells and machine learning techniques (i.e., K-Means clustering) to further group them according to their occupancy and operational data, which are retrieved from the IFC model (for this application, the attributes considered for space clustering are: “is occupied”, “is heated”, “is cooled”, and “area per occupant”).

Second, aggregation operations like sum, average, weighted average, minimum, and maximum are executed to transfer the cells' data pertinent to energy analysis to the thermal zones. Following that, by integrating Topologicpy with the Honeybee (HB) APIs, the Topologic faces are converted into HB faces, and the Topologic apertures into HB apertures. Zone by zone, HB faces and apertures are given as input to the HB “Room” method, creating the energy zones forming the basic “HB Model”.

Results of the zoning process are depicted in Figure 4.

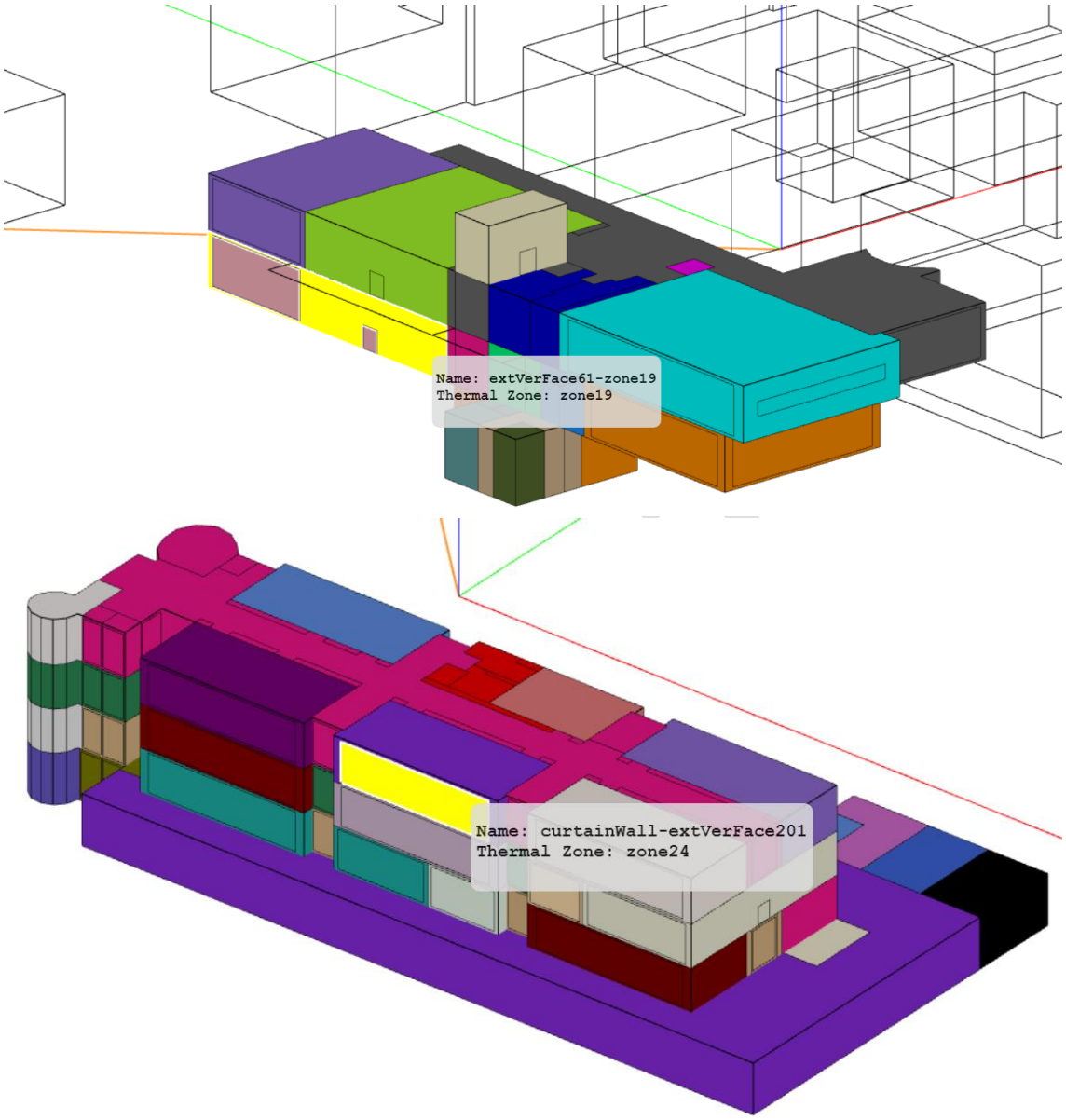

Figure 4. BEM model of Building 10 and Building 09(POLIMI). Faces colored by zone.

3.1.3.2 Boundary conditions

When creating the HB model, the boundary conditions between the HB faces that make up the model (i.e., “outdoor”, “adiabatic”, “ground”) are automatically computed.

Then, the faces of the building adjacent to other buildings are manually set to “adiabatic” (**Errore. L'origine riferimento non è stata trovata.**).

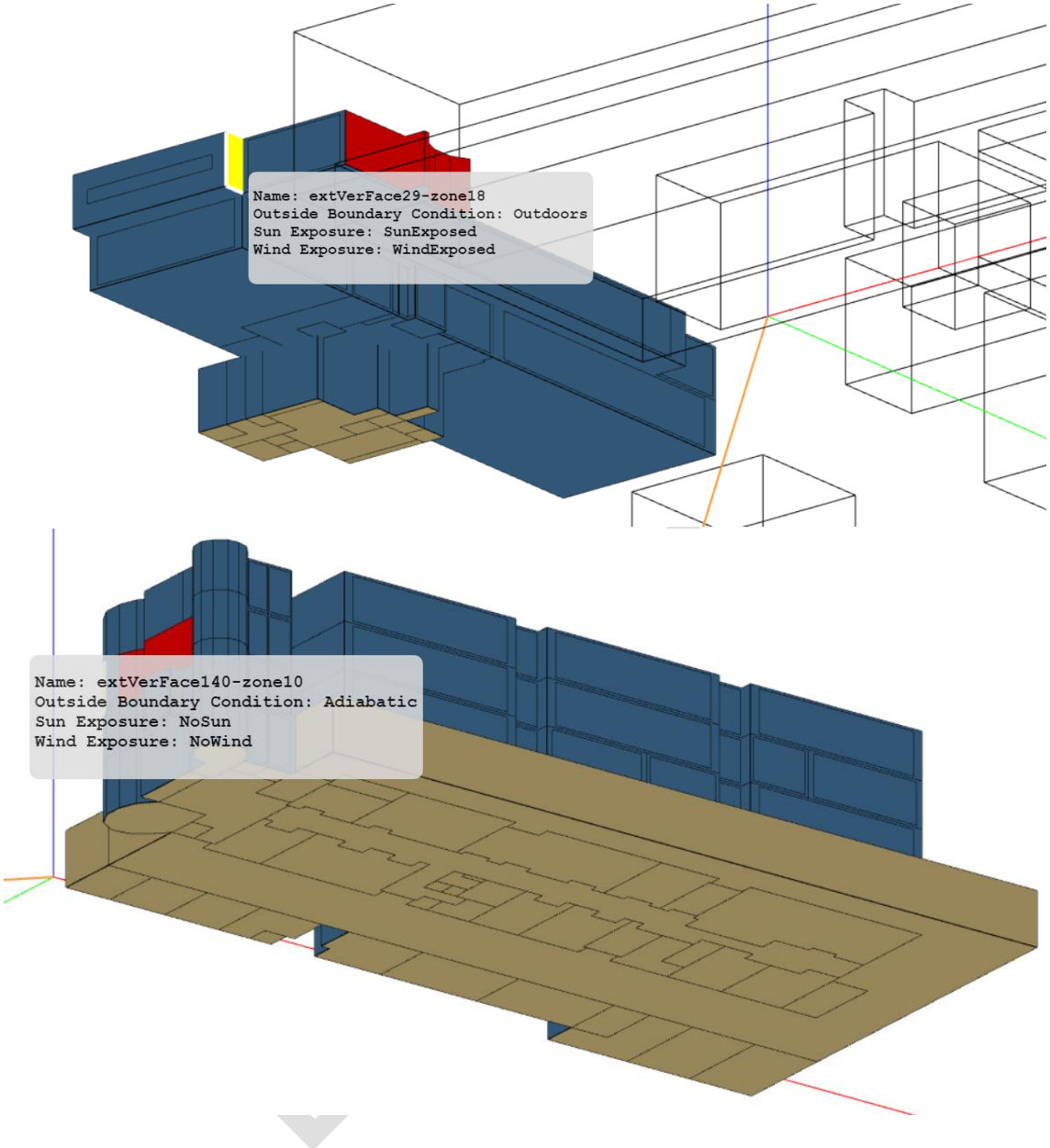

Figure 5. BEM model of Building 10 and Building 09 (POLIMI). Faces colored by boundary condition (red: adiabatic, blue: outdoors, yellow: ground).

3.1.3.3 Construction characteristics and openings

Construction data is the first added to the basic HB Model (Figure 6).

Opaque envelope data is extracted directly from the "IfcMaterialLayerSets" associated with the "IfcWalls" and "IfcSlabs" in the IFC model, getting the thickness, conductivity, density, and specific heat of the "IfcMaterials" composing the layer sets. IfcMaterials are thus converted into HB "EnergyMaterial" objects, which are then used within HB "OpaqueConstruction" objects.

Subsequently, the thermal data of the apertures is transferred to the components of the HB Mode, with glazed component data (such as thermal transmittance, solar heat gain coefficient, and visible

transmittance) extracted from the "IfcWindowTypes" and "IfcDoorType" of the IFC, respectively converted into HB "WindowConstruction" and HB "Door" objects.

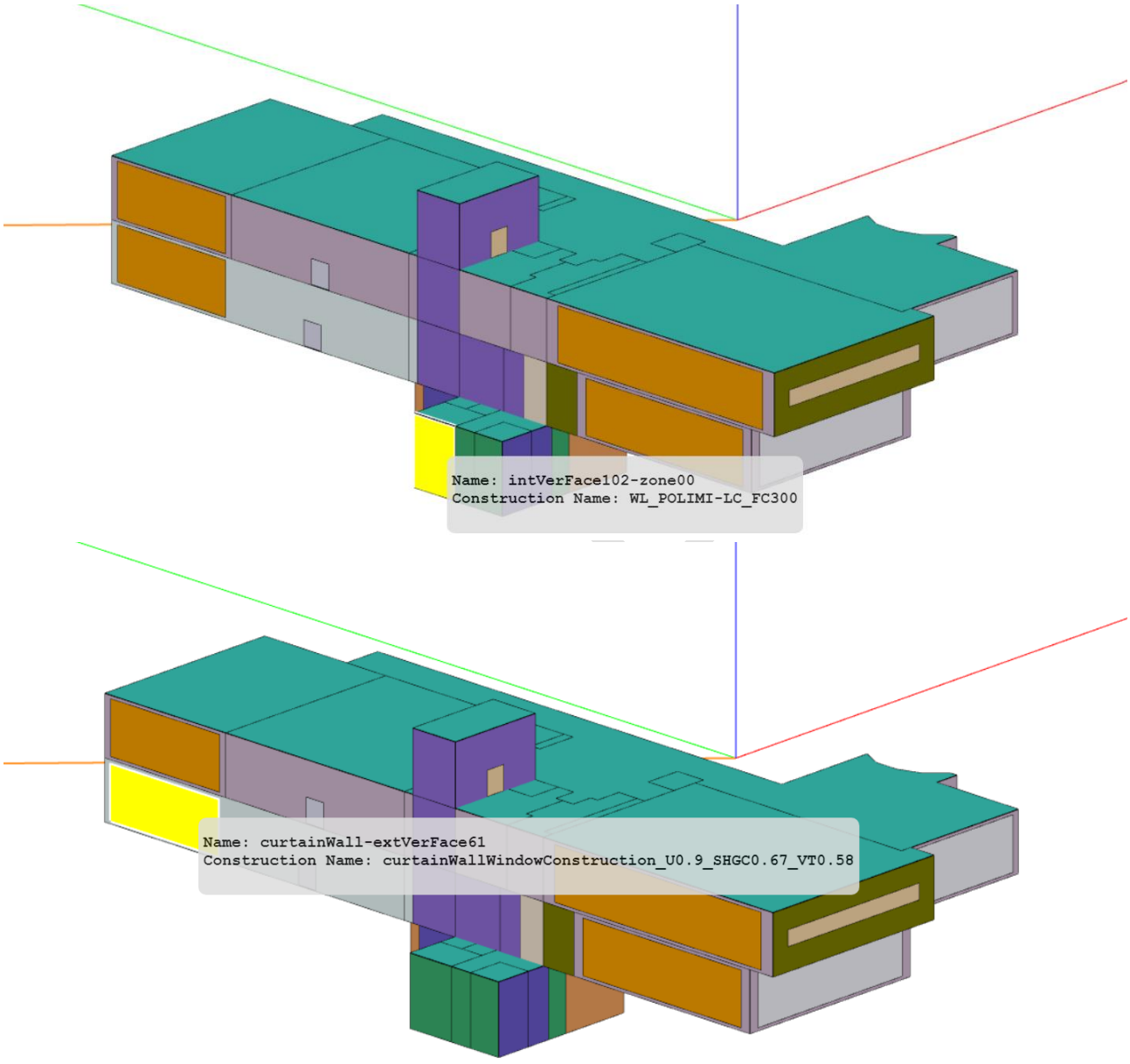

Figure 6. BEM model of Building 10 (POLIMI). Faces colored by construction type.

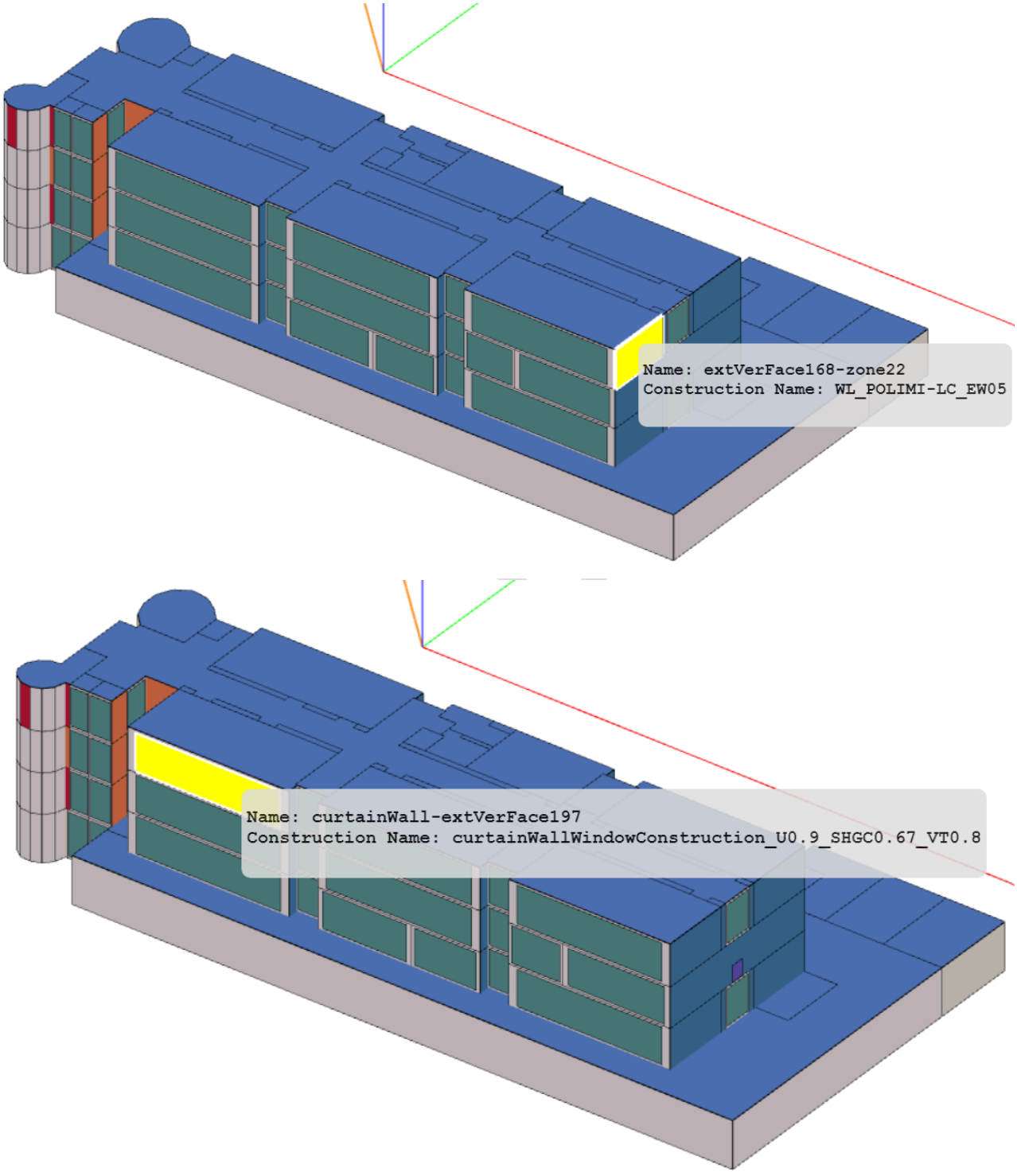

Figure 7. BEM model of Building 09 (POLIMI). Faces colored by construction type.

3.1.3.4 Context shadings

To simulate shade effects on the model, context elements that cast shadows on the building (e.g, vegetation, surrounding buildings, shading systems) are modelled within a 3D modelling environment,

exported as BRep entities, converted into Topologic faces, and then linked to the BEM as HB "Shade" elements (Figure 8).

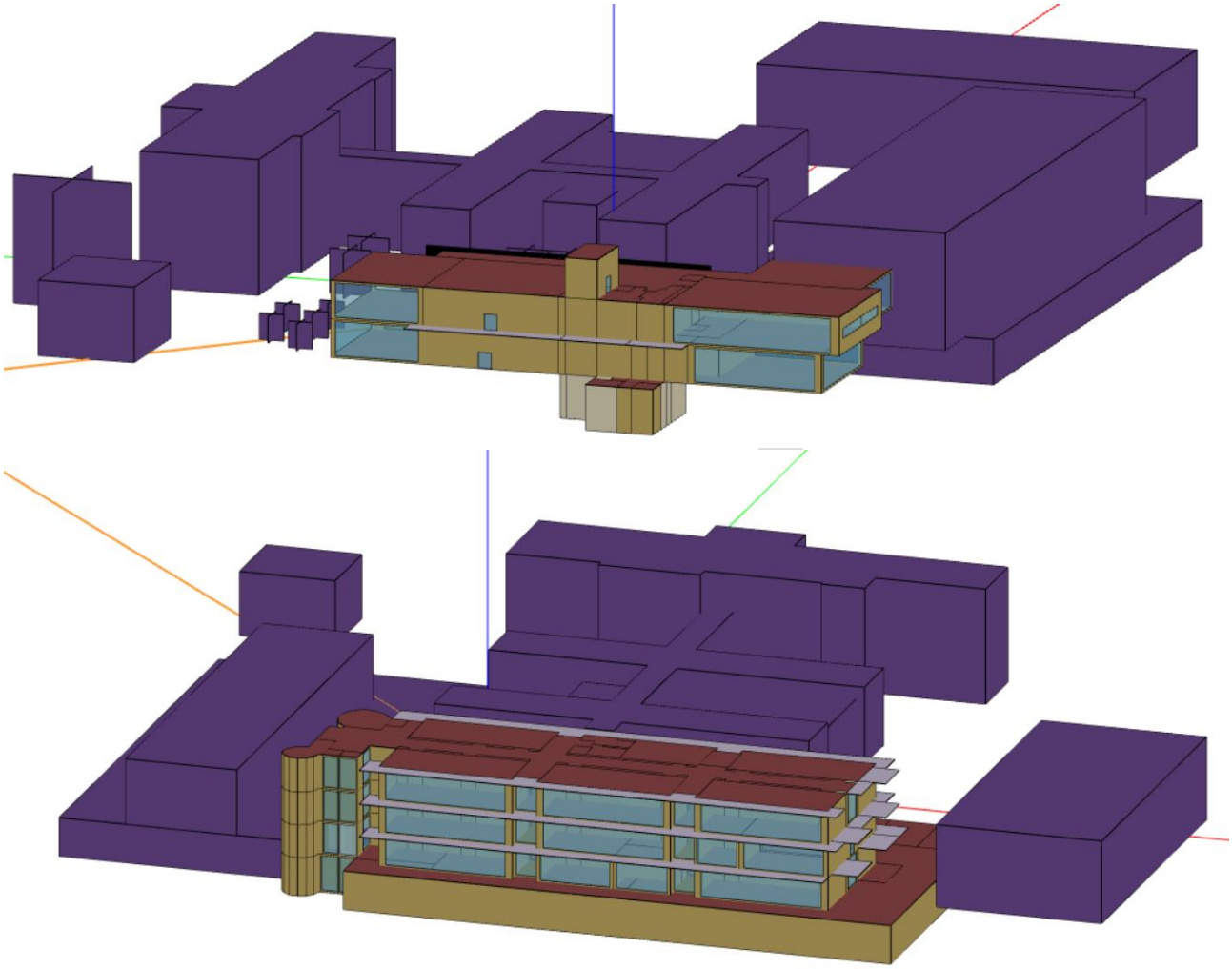

Figure 8: IDF models of the selected pilot site and context shadings. Building 10 on the top, Building 09 on the bottom.

3.1.3.5 Simulation parameters

The simulation parameters are then configured through Eppy. These include the simulation run period, the shadow calculation method, the daylight saving calculation method, and the output variables, which refer to the KPIs to be analysed through the simulation (i.e., zone heating and cooling energy, zone electric energy for lighting and equipment, and zone people occupant count).

The HB model, enriched with this information, is exported as an IDF file. From this point onwards, additional information is assigned to the IDF using the Eppy library.

3.1.3.6 Infiltration

The infiltration rate of the building is set in the IDF by assigning a "ZoneInfiltration:DesignFlowRate" object to all building zones using Eppy, with a flow rate per exterior surface area equal to 0.0001 $m^3/s\text{-}m^2$, referring to a tight building for the case study in question.

3.1.3.7 Schedules

Input schedule data is sourced from JSON files containing schedules formatted as the HB APIs requires for modelling the "ScheduleRulesets". A Python function was then created to transform these data into IDF "Schedule:Compact" objects. The schedules here considered include fraction types, such as occupancy schedules (occupancy density in each zone for every hour of the year) and lighting and equipment schedules (percentage of power usage for lighting and equipment per hour) and boolean types, such as the availability schedules for heating, cooling, and ventilation systems.

For occupancy schedules, which are particularly significant for the project, data directly from the university's asset management service are considered, as explained in the demonstration section.

3.1.3.8 Operational data

Static operational data are instead transferred from the IFC to the IDF. Occupancy data, assigned zone by zone to IDF "People" elements, includes people per floor area in operational conditions and the corresponding schedules. Lighting data, assigned instead to IDF "Lights" objects, consists of the electric power density of lighting devices and relative schedule names. Similarly, equipment data, assigned to "ElectricEquipmen" IDF objects, include the electric power density used by electrical appliances and relative schedules. Instead, thermal and humidity control data are assigned to IDF "HVACTemplate: ZoneIdealLoadsAirSystem" objects. These include temperature and humidity setpoints, related schedules, and availability schedules for HVAC systems. Finally, ventilation data are assigned to "DesignSpecification:OutdoorAir" objects, which specify the method for calculating outdoor air changes, the outdoor airflow per person and zone floor area.

3.1.3.9 Weather data

Climatic data are inputted in the BEM using the EPW format. This data can be downloaded from open-source repositories or custom-created.

In this case, we used air temperature, relative humidity, atmospheric pressure, wind speed, wind direction, and global solar radiation data collected from a weather station close to the analysed building over the last three years. To meet the EPW format's requirements, the dataset was processed by averaging hourly data across years and missing variables, such as dew point and wet bulb temperatures, were derived from the measured data. Meteonorm software [56] was also used to generate missing parameters, such as direct normal radiation, diffuse horizontal radiation, global horizontal radiation, direct normal illuminance, diffuse horizontal illuminance, total sky cover, opaque sky cover and visibility. Figure 9 to Figure 13 summarises the main climatic data from the EPW file.

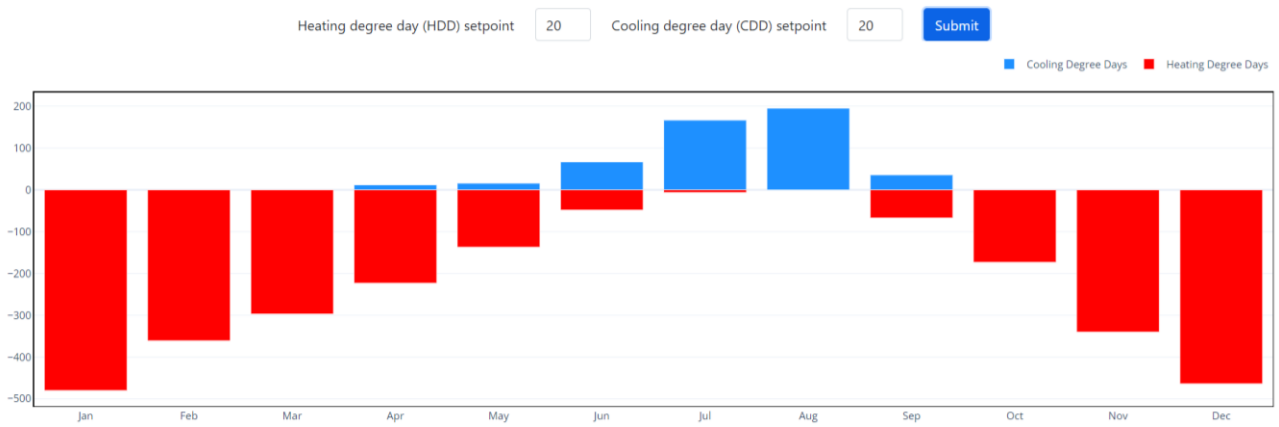

Figure 9: Heating and cooling degree days (EPW analysis in CBE Clima Tool).

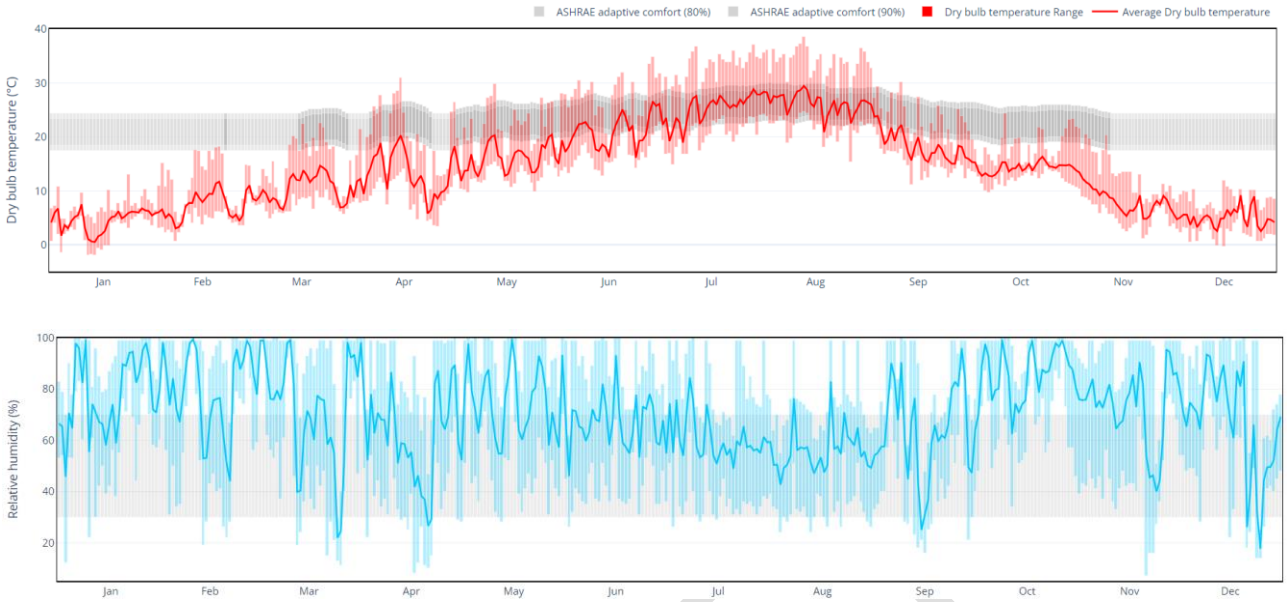

Figure 10: Yearly chart with temperature and humidity values (EPW analysis in CBE Clima Tool).

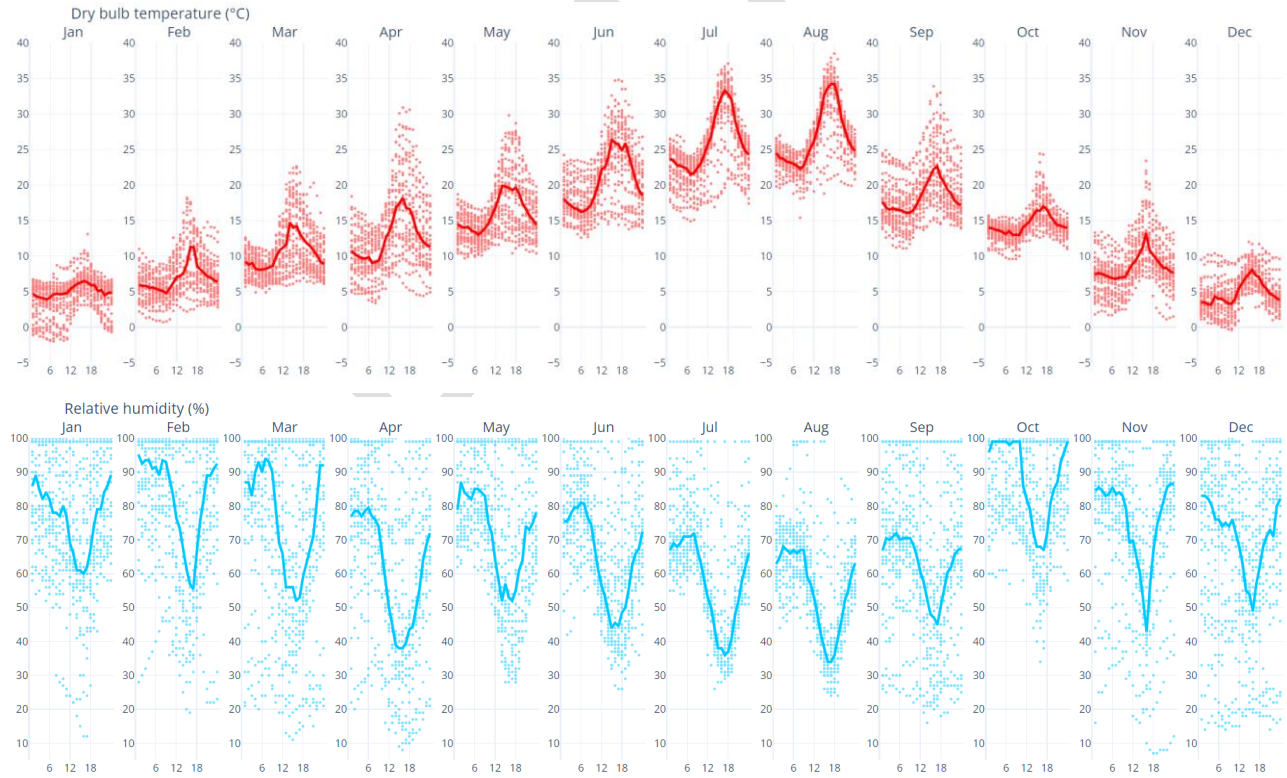

Figure 11: Daily chart with temperature and humidity values (EPW analysis in CBE Clima Tool).

Figure 12: Global and Diffuse Horizontal Solar Radiation (Wh/m²) (EPW analysis in CBE Clima Tool).

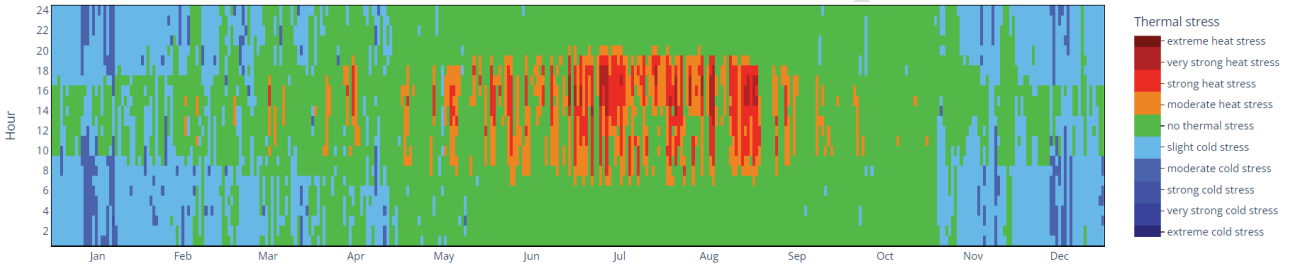

Figure 13: UTCI thermal stress chart (EPW analysis in CBE Clima Tool).

3.2 Results and Demonstration

3.2.1 BEM calibration

After generating the BEM model using the algorithm described above, the model calibration is conducted. In this context, energy calibration refers to the process of adjusting the model parameters so that the computed energy values align with real reference data acquired from sensors. This step is crucial for ensuring the model accurately reflects the physical system by minimising discrepancies between theoretical predictions from simluations and actual measurements.

3.2.1.1 Assumptions and rationale for energy model calibration

Since the goal of the energy model is to estimate the building's energy needs in terms of thermal energy required for heating and cooling, we have decided to calibrate the model by tuning its input data. The aim is to adjust the simulated indoor temperatures so they closely match those measured by sensors.

Our underlying assumption is that the building's energy demand is directly correlated to its thermal losses. Since these losses are proportional to the temperature difference between the indoor and outdoor environments, if the simulated and measured temperatures are equivalent, the model can be considered valid for estimating the building's energy needs.

Furthermore, this calibration approach was chosen because temperature data were the only thermal measurements available for the building. Other data typically used for calibration – such as gas and electricity consumption from utility bills – were available only in aggregated form at the campus level. This aggregation made it impossible to assign specific energy consumption values to the individual buildings analysed. Therefore, with data available from only a single classroom per building and considering that the

building's construction is quite homogeneous throughout, the temperature calibration was performed on that single classroom. This was achieved by adjusting the temperature setpoints during both the summer and winter seasons, as well as modifying the related operational schedules.

3.2.1.2 Calibration data

For the calibration process, a dataset was constructed containing 8,760 rows—each corresponding to one hour of the year—and three columns: a timestamp, measured temperature values (labeled "Real Temperature"), and simulated temperature values (labeled "Simulated Temperature"). The real temperature represents the hourly average temperature recorded by sensors installed in the classrooms. In contrast, the simulated temperature is derived from the "Zone Mean Radiant Temperature" of the thermal zone corresponding to the classroom within the energy model, as calculated by EnergyPlus. The mean radiant temperature (MRT) is defined as the weighted average of the temperatures of the surfaces surrounding an environment, such as walls, floors, ceilings, and windows.

Given that the sensors at POLIMI are wall-mounted, their readings were chosen for the calibration. The sensor data cover the period from February 2024 to January 2025, and the same timeframe was used for the climatic data to ensure consistency. Initially, the data were plotted to assess the discrepancies between the real measurements and the simulated model output, with a summary presented in the figures below.

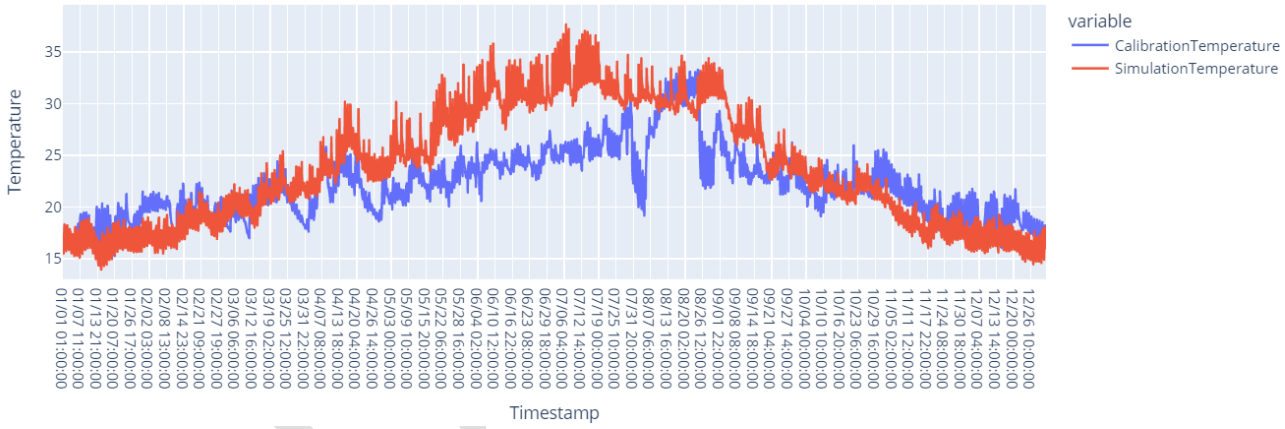

Figure 14: Hourly Real vs Simulated data before the model calibration (year), Building 10.

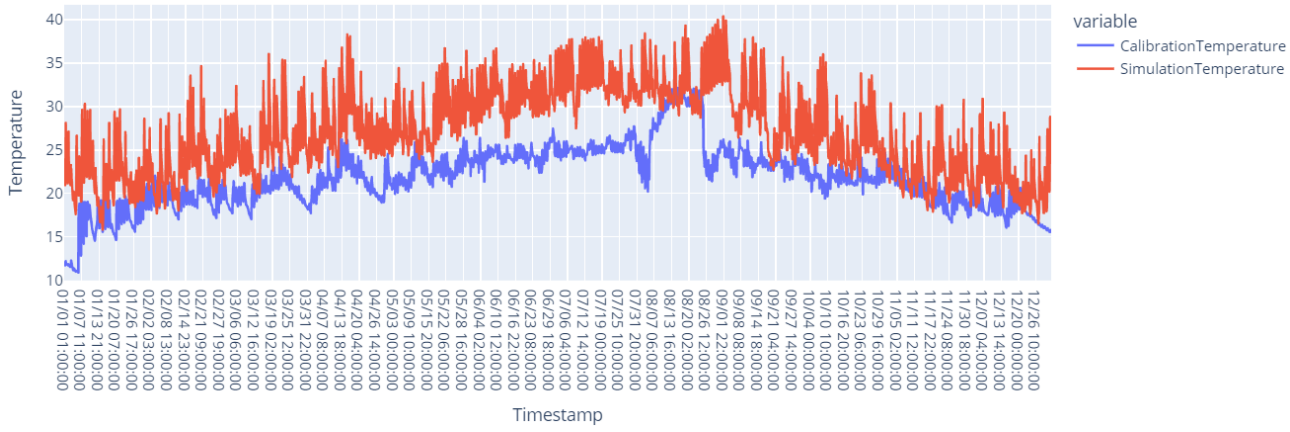

Figure 15: Hourly Real vs Simulated data before the model calibration (year), Building 09.

3.2.1.3 Evaluation of typical weeks

To assess the discrepancies between the simulated and real data, various metrics were employed at different temporal scales. Initially, characteristic weeks were selected for detailed evaluation. These weeks were chosen based on the EPW file corresponding to the nearest location to Lecco—downloaded from https://climate.onebuilding.org/ - and are as follows:

- Winter: February 1st – February 6th
- Summer: July 21st – July 27th
- Spring: April 14th – April 20th
- Autumn: October 13th – October 19th

Special emphasis was placed on the winter and summer weeks, as these periods typically represent the extreme conditions—cold in winter and hot in summer—that most significantly affect the building's energy performance.

3.2.1.4 Calibration metrics for the selected weeks

The calibration was performed on these weeks using the following indicators:

- **MSE (Mean Squared Error):** This represents the average of the squared differences between the simulated and measured values. A high MSE indicates significant errors between the two datasets. Because the errors are squared, larger discrepancies are emphasised.
- **RMSE (Root Mean Squared Error):** This is the square root of the MSE, which brings the error back to the same unit as the original data (°C in this case). It indicates, on average, by how much the simulated values deviate from the measured ones. A high RMSE suggests that the simulation model might not be accurately reproducing the real temperature values.
- **Mean Temperature Difference in the Week (MTDW):** For each hour of the week, the difference between the real and simulated data is calculated, and then the average of these differences is taken. This value ensures that, on average, the temperature difference remains low, so that the overall daily balance is maintained. Unlike RMSE, this metric takes into account the positive and negative signs of the differences.
- **Error in the Week:** This is defined as the ratio between the temperature difference and the mean temperature of the week, expressed as a percentage. It provides a measure of the magnitude of the error relative to the average measured temperature.

Table 5 and Table 6 provide an overview of the principal values evaluated before the model calibration, while Figure 16-Figure 23 depict the comparison between real and simulated data before calibration.

Table 5: Analysis of Real vs Simulated data before the model calibration (winter week), Building 10.

Period	MTDW	RMSE	MSE	Error
Winter week	*+3.2 °C*	*+ 3.2 °C*	*+10.3 °C²*	*+ 15.8 %*
Summer week	*- 4.8 °C*	*+ 4.7°C*	*+24.7 °C²*	*- 18.2 %*
Spring week	*- 2.3°C*	*+ 2.8 °C*	*+7.83 °C²*	*- 9.7 %*
Autumn week	*+ 0.3 °C*	*+0.8 °C*	*+0.66 °C²*	*+ 1.4 %*

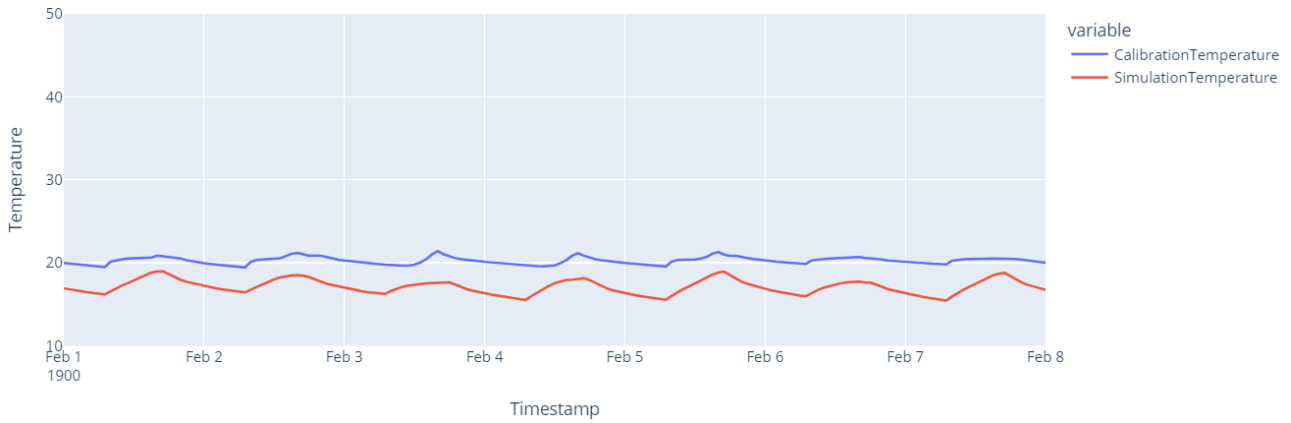

Figure 16: Hourly Real vs Simulated data before the model calibration (winter week), Building 10.

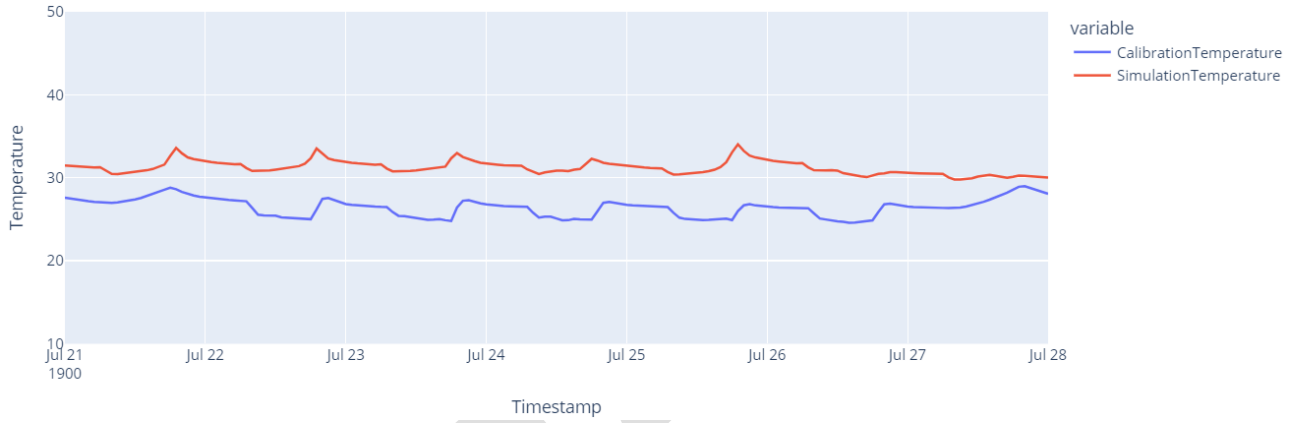

Figure 17: Hourly Real vs Simulated data before the model calibration (summer week), Building 10.

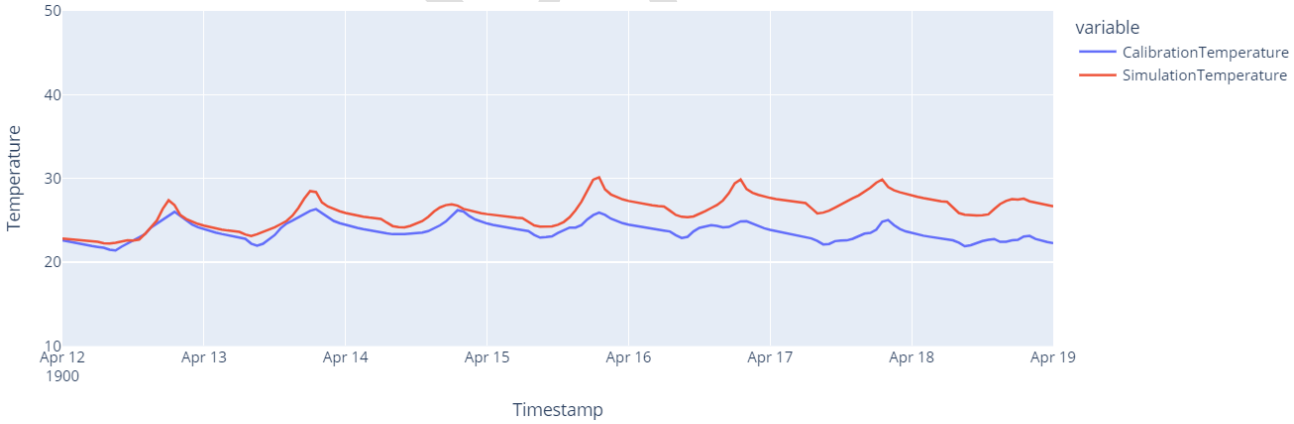

Figure 18: Hourly Real vs Simulated data before the model calibration (spring week), Building 10.

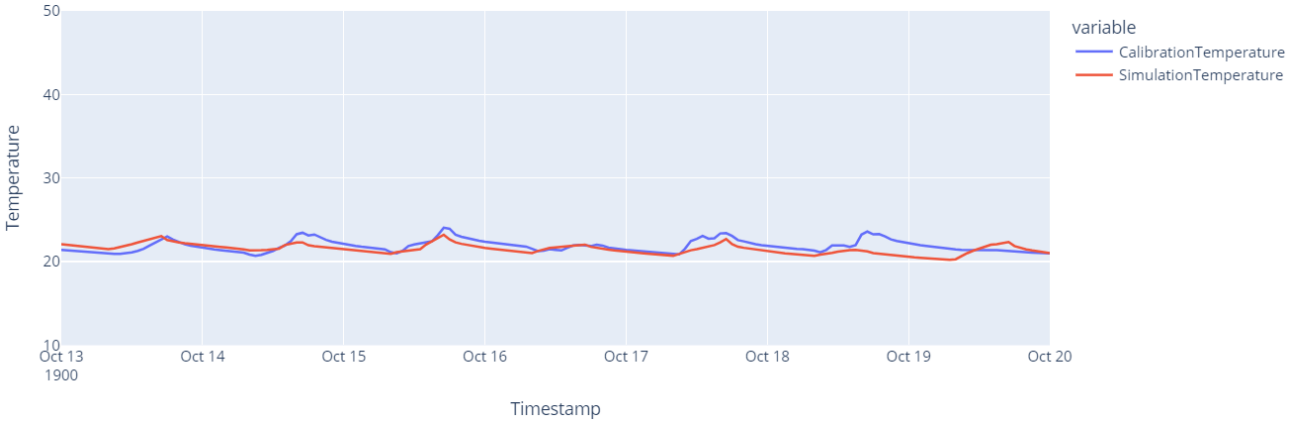

Figure 19: Hourly Real vs Simulated data before the model calibration (autumn week), Building 10.

Table 6: Analysis of Real vs Simulated data before the model calibration (winter week), Building 09.

Period	MTDW	RMSE	MSE	Error
Winter week	*- 1.8 °C*	*+ 2.8 °C*	*+ 7.86 °C²*	*- 9.1 %*
Summer week	*- 6.8 °C*	*+ 7.0 °C*	*+ 48.46 °C²*	*- 26.7 %*
Spring week	*- 7.2 °C*	*+ 7.7 °C*	*+ 59.85 °C²*	*- 30.5 %*
Autumn week	*- 3.0 °C*	*+ 3.5 °C*	*+ 12.26 °C²*	*- 13.7 %*

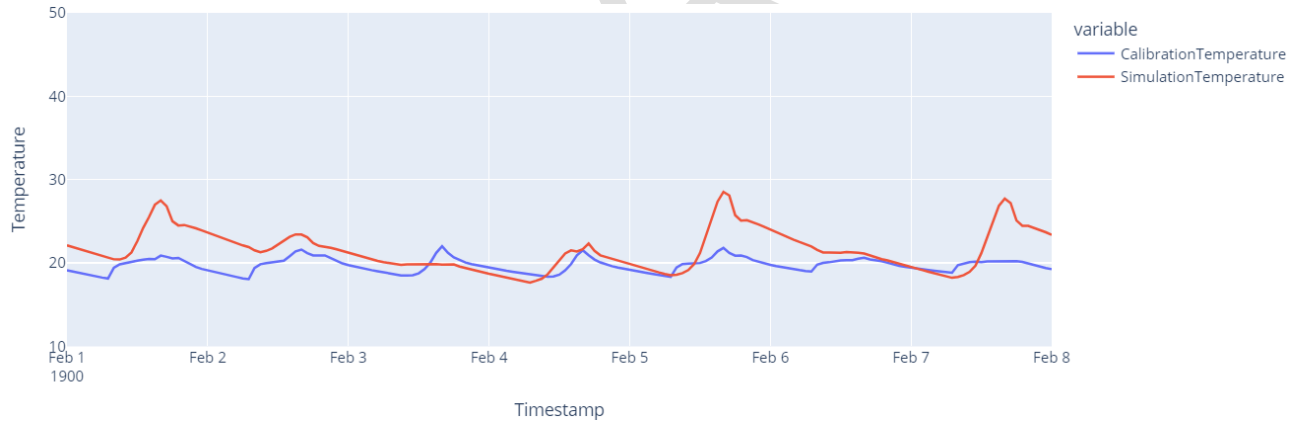

Figure 20: Hourly Real vs Simulated data before the model calibration (winter week), Building 09.

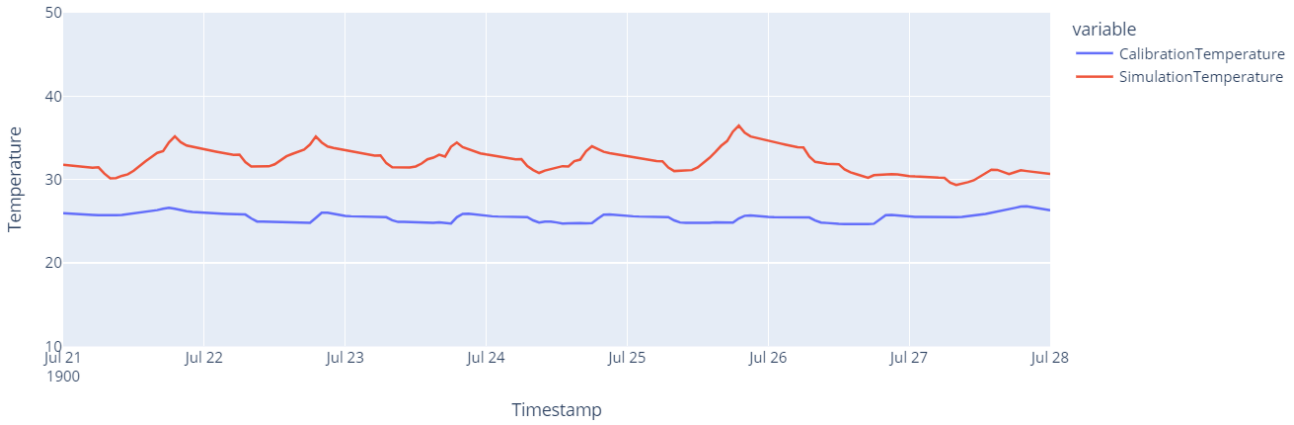

Figure 21: Hourly Real vs Simulated data before the model calibration (summer week), Building 09.

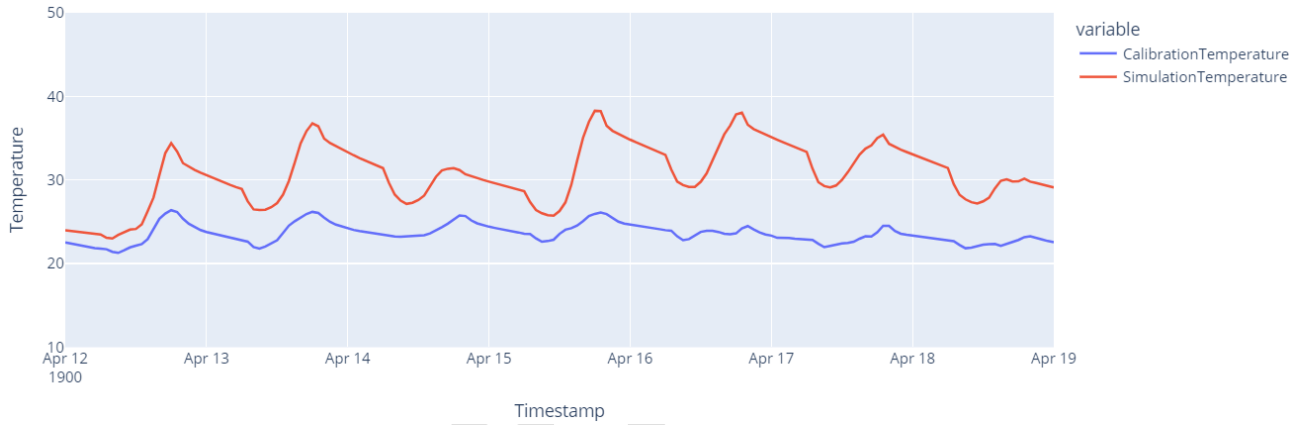

Figure 22: Hourly Real vs Simulated data before the model calibration (spring week), Building 09.

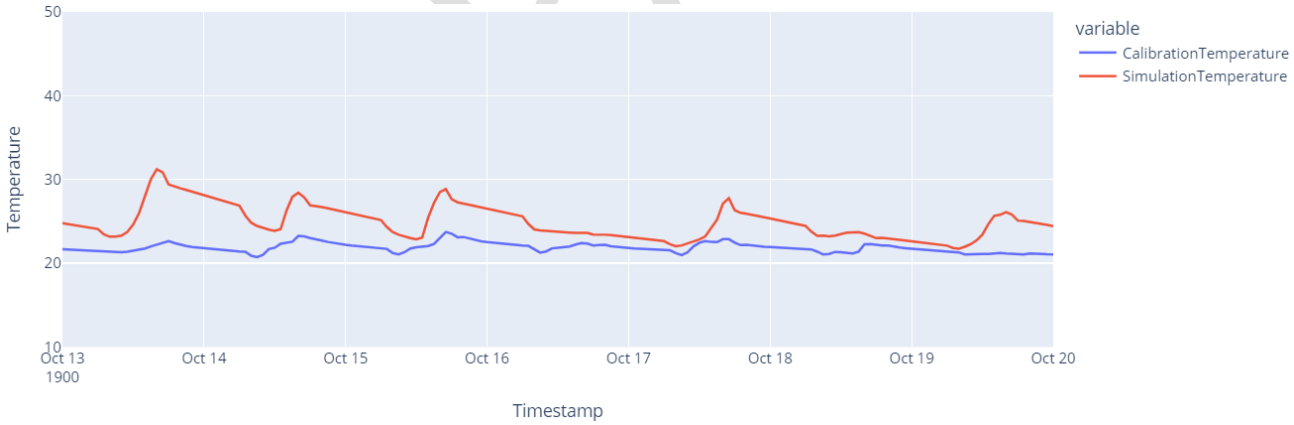

Figure 23: Hourly Real vs Simulated data before the model calibration (autumn week), Building 09.

The model tuning was carried out through the following modifications:

- Debugging the Model:
  - Ventilation Schedule: The ventilation system availability schedule, which was originally set to OFF during the night, was changed to ALWAYS ON.
  - Window Shading Devices: The shading devices for the windows (such as sunshades), initially modeled as Context Shades, were removed and reconfigured as

WindowMaterial:Blind along with WindowShadingControl objects in the IDF, after detecting some anomalies.
  - Solar Distribution Calculation: The solar distribution method was changed from FullInteriorAndExterior to FullExterior to enable a faster calculation when blinds are in place.
  - Simulation Timestep: The simulation timestep was set to 4, which sped up the simulation process.
  - Temperature Setpoint Schedules: The temperature setpoint schedules, initially exported as Schedule:Constant, were modified to Schedule:Compact. This object allows for a more flexible and modulated definition of setpoints throughout the day (e.g., differentiating between day and night cycles).
- Tuning Adjustments:
  - Heating Setpoint: The heating setpoint was iteratively adjusted until it was ultimately set at 22°C.
  - Heating Setback: The heating setback was iteratively adjusted to 18°C.
  - Heating Operation Schedule: The heating setpoint is applied between 07:00 and 18:00, with the setback applied during the remaining hours.
  - Cooling Setpoint: The cooling setpoint was iteratively adjusted until it reached 22°C.
  - Cooling Setback: The cooling setback was iteratively adjusted to 24°C.
  - Cooling Operation Schedule: The cooling setpoint operates between 07:00 and 18:00, with the setback applied outside this period.

The results of these adjustments are detailed in Table 7 and illustrated in Figure 24-Figure 28.

Table 7: Analysis of Real vs Simulated data after the model calibration (winter week), Building 10.

Period	MTDW	RMSE	MSE	Error
Winter week	*+ 1.6 °C*	*+ 1.7 °C*	*+ 2.8 °C²*	*+ 7.9 %*
Summer week	*+ 1.1 °C*	*+ 1.9 °C*	*+ 3.5 °C²*	*+ 4.2 %*
Spring week	*+ 0.7 °C*	*+ 1.4 °C*	*+ 2.1 °C²*	*+ 3.0 %*
Autumn week	*+ 1.2 °C*	*+ 1.5 °C*	*+ 2.3°C²*	*+ 5.5 %*

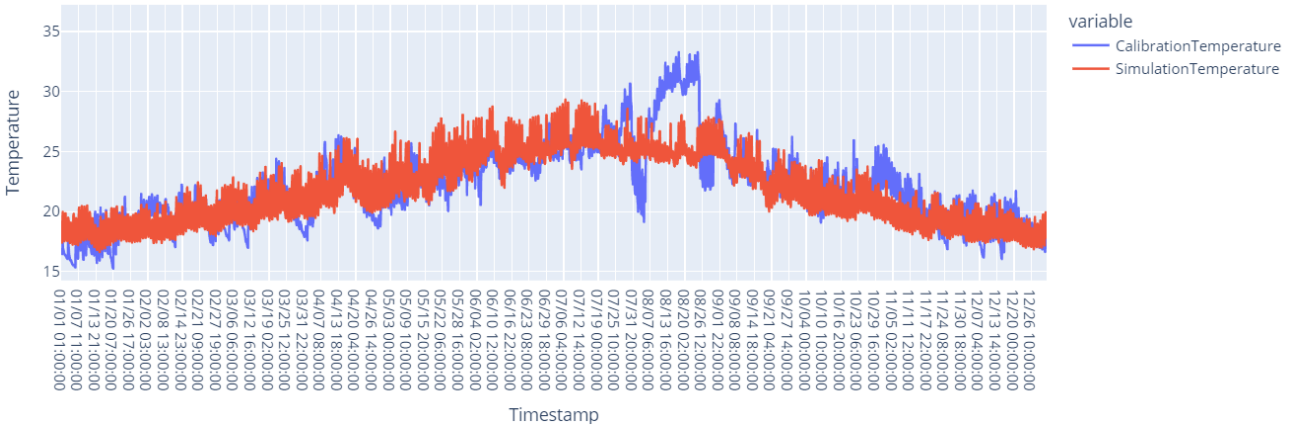

Figure 24: Hourly Real vs Simulated data before the model calibration (year), Building 10.

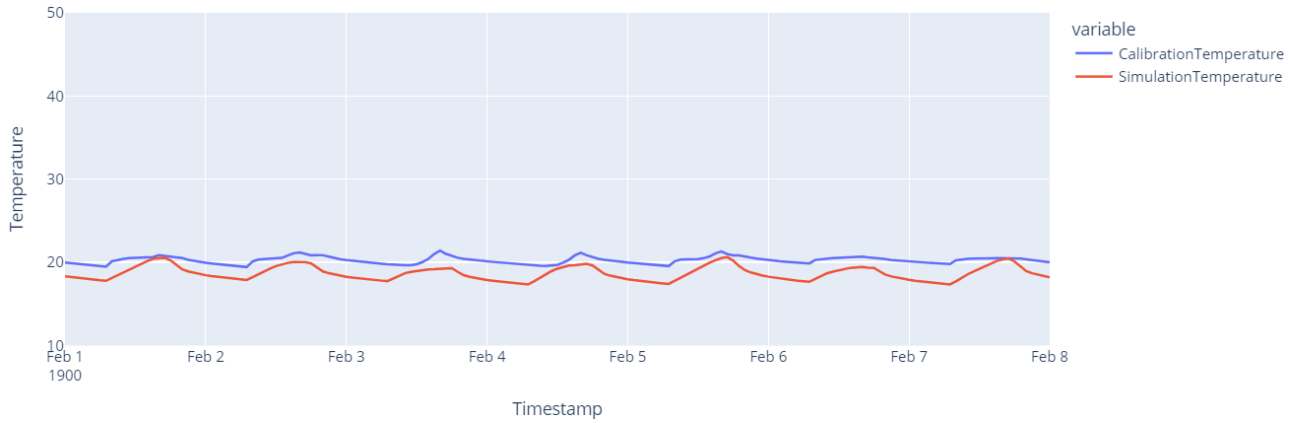

Figure 25: Hourly Real vs Simulated data after the model calibration (winter week), Building 10.

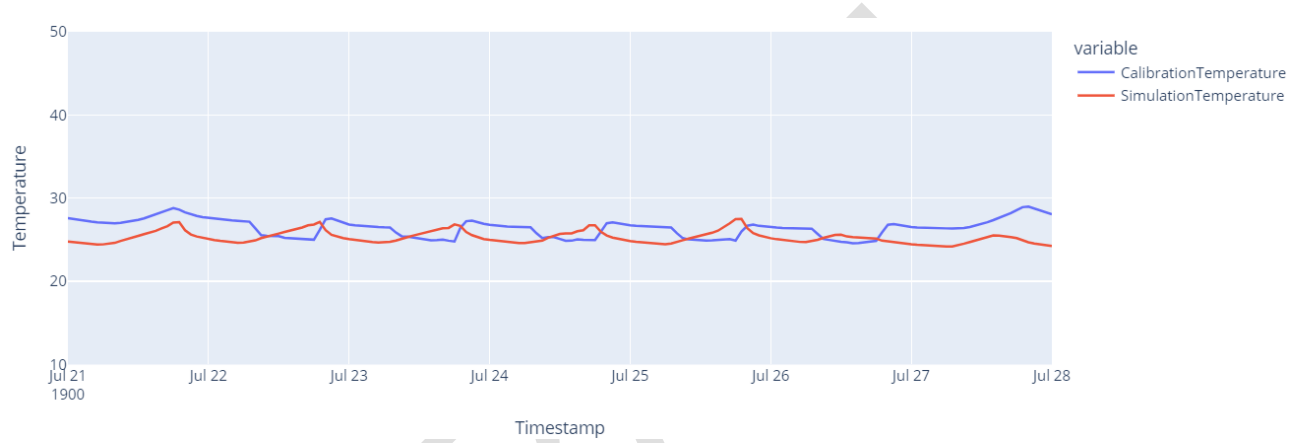

Figure 26: Hourly Real vs Simulated data after the model calibration (summer week), Building 10.

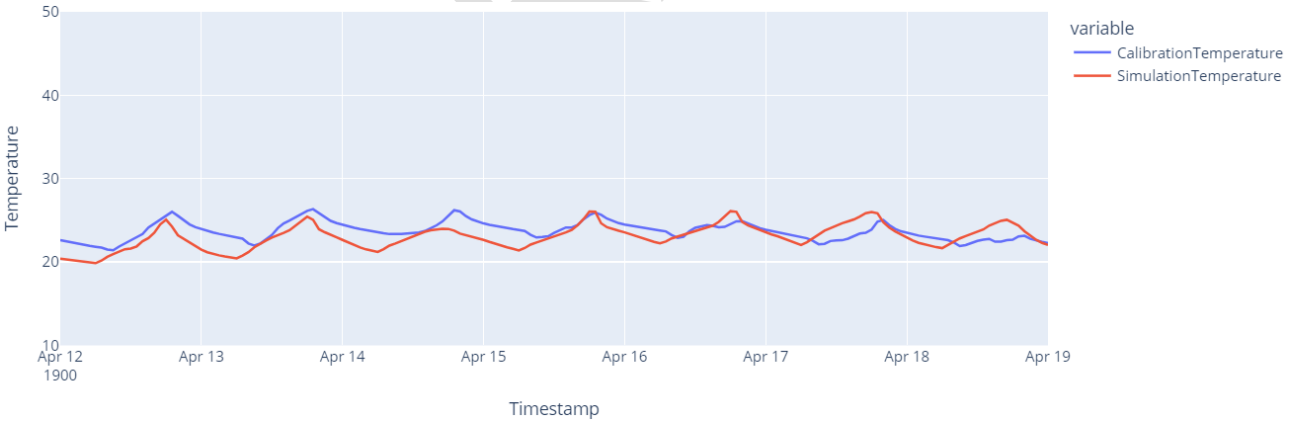

Figure 27: Hourly Real vs Simulated data after the model calibration (spring week), Building 10.

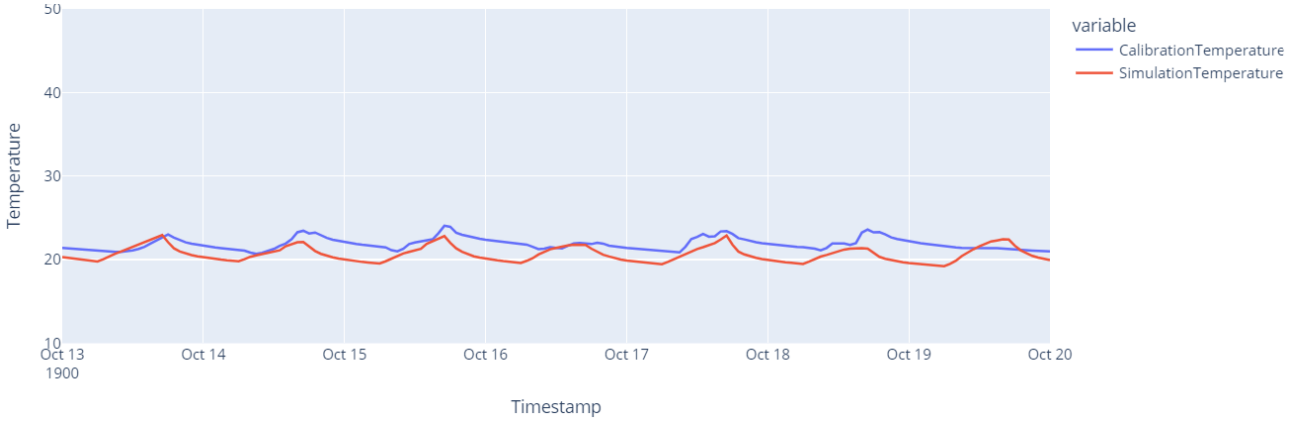

Figure 28: Hourly Real vs Simulated data after the model calibration (autumn week), Building 10.

Table 8: Analysis of Real vs Simulated data after the model calibration (winter week), Building 10.

Period	MTDW	RMSE	MSE	Error
Winter week	*+1.1 °C*	*+ 2.4 °C*	*+ 5.79 °C²*	*+ 5.6 %*
Summer week	*+ 1.0 °C*	*+ 2.0 °C*	*+ 4.19 °C²*	*+ 3.9 %*
Spring week	*+ 0.1 °C*	*+ 2.3 °C*	*+ 5.40 °C²*	*+ 0,4 %*
Autumn week	*+ 1.1 °C*	*+ 2.3 °C*	*+ 5.27 °C²*	*+ 5.0 %*

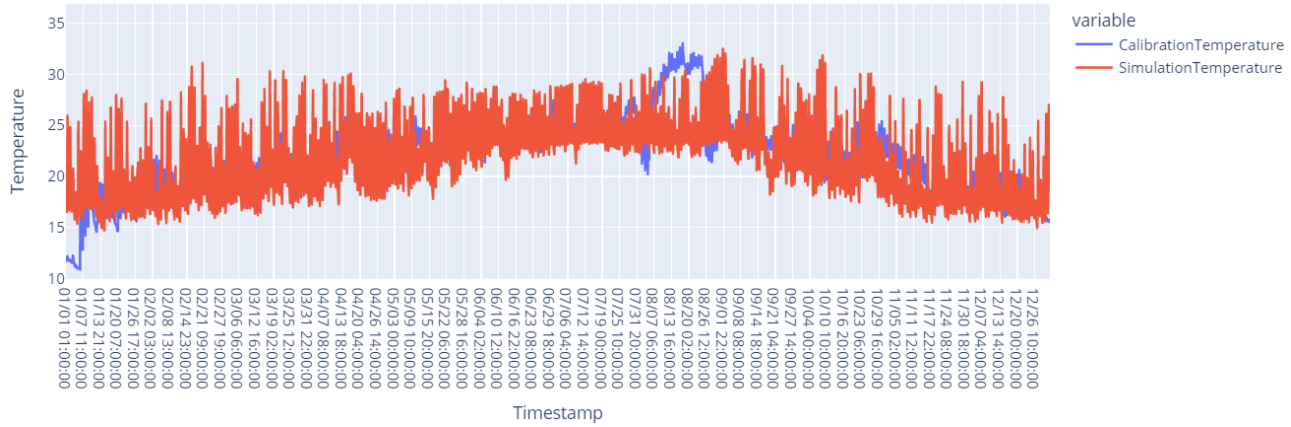

Figure 29: Hourly Real vs Simulated data before the model calibration (year), Building 09.

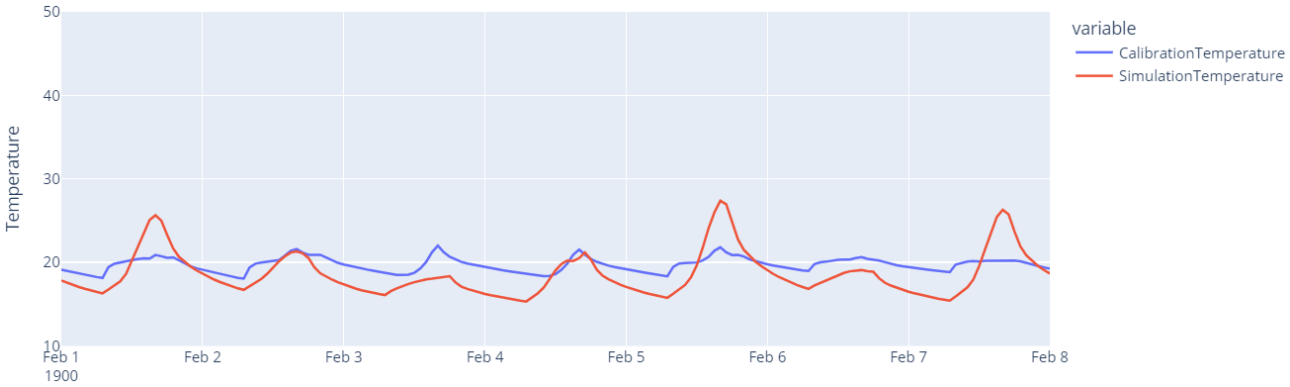

Figure 30: Hourly Real vs Simulated data after the model calibration (winter week), Building 09.

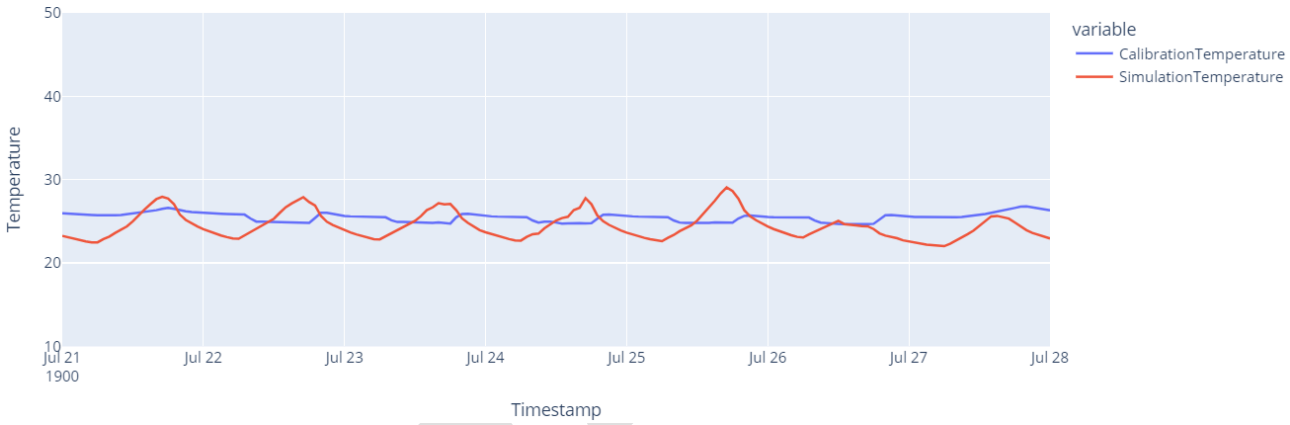

Figure 31: Hourly Real vs Simulated data after the model calibration (summer week), Building 09.

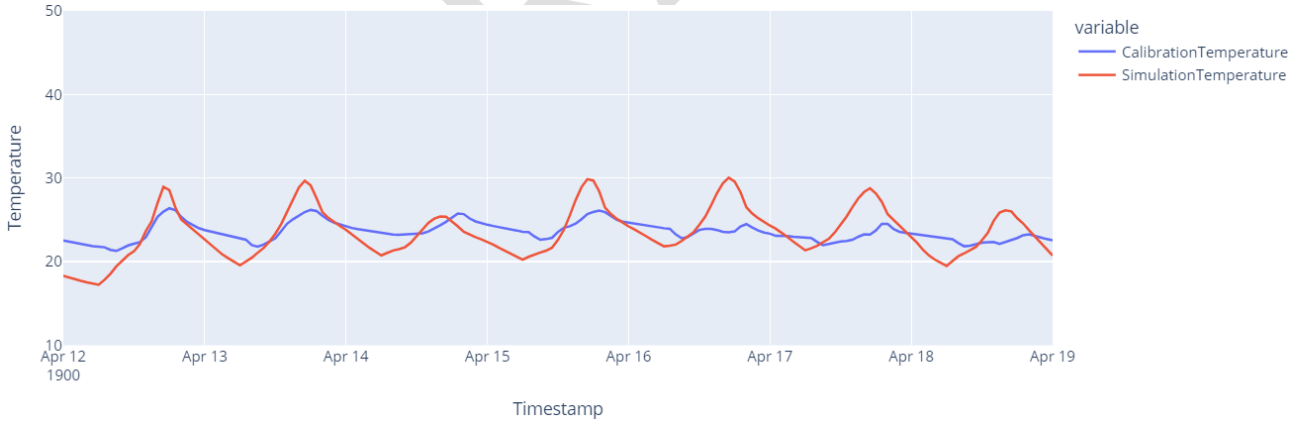

Figure 32: Hourly Real vs Simulated data after the model calibration (spring week), Building 09.

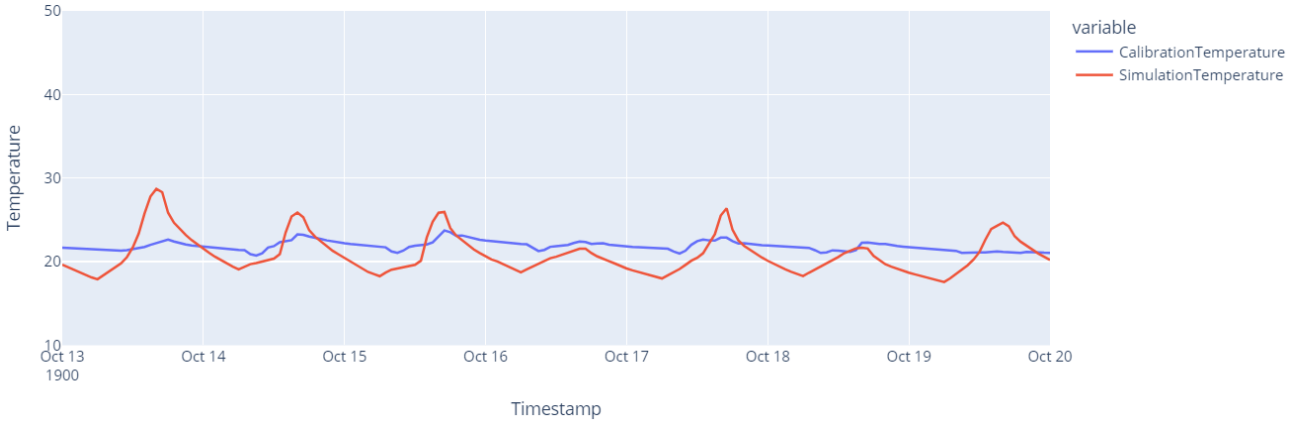

Figure 33: Hourly Real vs Simulated data after the model calibration (autumn week), Building 09.

3.2.1.5 Calibration on aggregated data

During the iterative tuning of the temperature setpoints, additional aggregated graphs were analysed on both a daily and monthly basis, as shown in the figures below. In particular, the final graph was used to inform the ultimate decision.

Building 10

Figure 34 presents a comparison of the monthly average temperatures, while Figure 35 displays the difference between the simulated and real temperature values. In summary, with the exception of August—which shows seemingly anomalous data (see Figure 36)—the difference between the real and simulated monthly average temperatures is within ±1 °C.

This outcome is considered satisfactory given the study's objectives (i.e., an error of less than 5%, corresponding to a mean error of under 1 °C on a baseline temperature of 20 °C).

Regarding the issues observed in August, several factors may explain the anomalies:

- Connectivity problems: Between July and September, both classrooms experienced connectivity issues that resulted in significant data loss. Measurements lasting fewer than four hours were interpolated, whereas those exceeding four hours were left blank.
- Maintenance limitations: Due to the aforementioned connectivity problems, maintenance activities— which require access to the internet network—could not be performed during the campus summer break.
- Access restrictions: During the first three weeks of August, access to the first and second floors was restricted to students. This may have led to variations in the ventilation system setpoints or even forced the shutdown of the cooling battery systems.

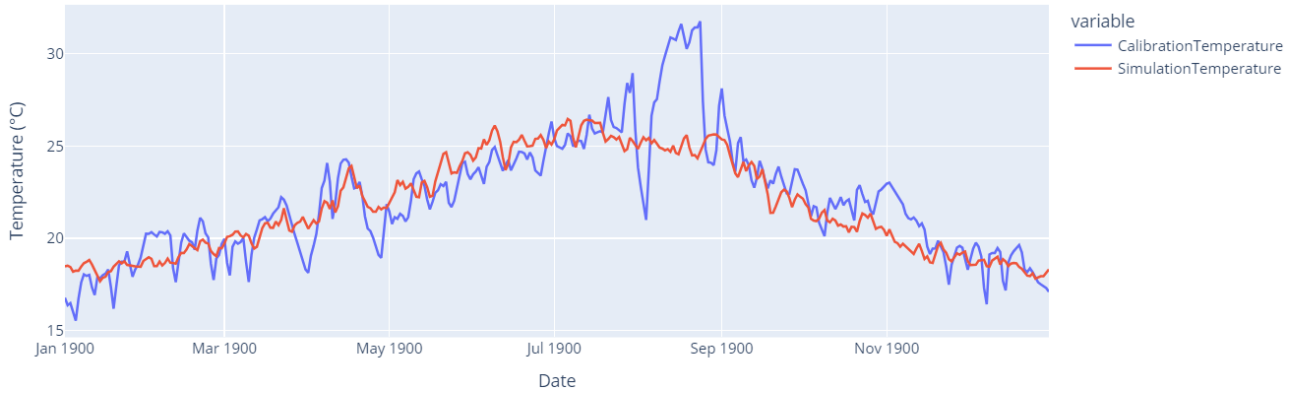

Figure 34: Daily Real vs Simulated data after the model calibration, Building 10.

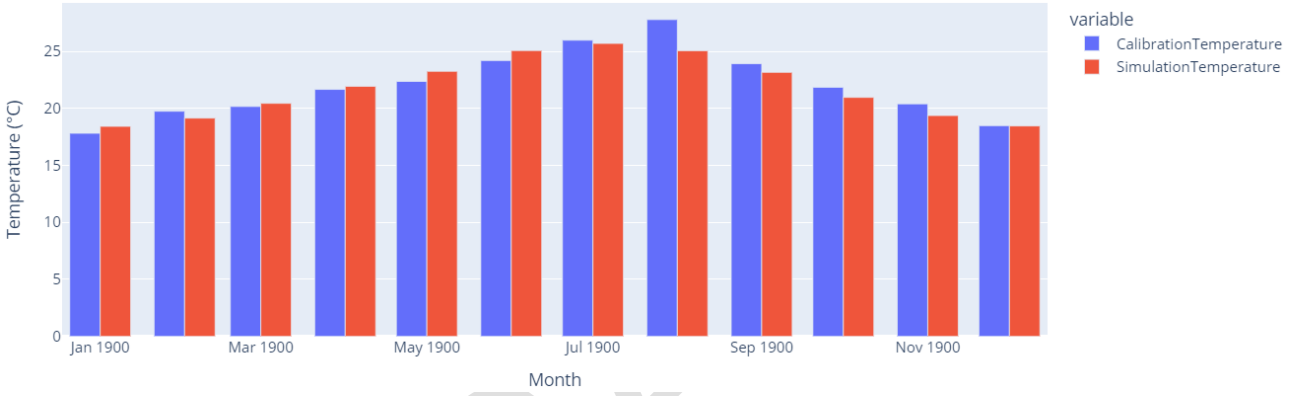

Figure 35: Monthly Real vs Simulated data after the model calibration, Building 10.

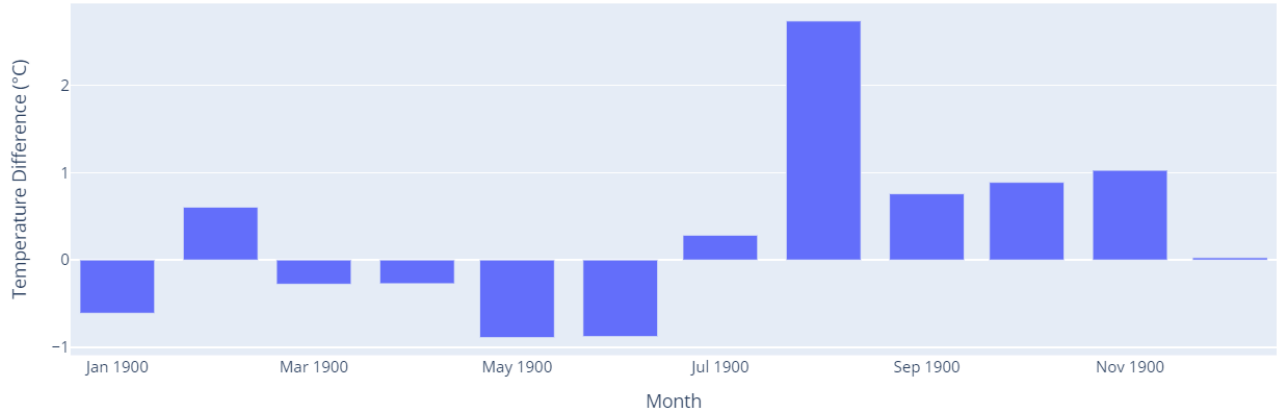

Figure 36: Monthly difference between mean real temperature and mean simulated temperature, after the model calibration, Building 10.

Building 09

Figure 37 presents a comparison of the monthly average temperatures, while Figure 37 displays the difference between the simulated and real temperature values. In summary, with the exception of August and January —which show seemingly anomalous data (see Figure 39)—the difference between the real and simulated monthly average temperatures is within ±1 °C.

This outcome is considered satisfactory given the study's objectives (i.e., an error of less than 5%, corresponding to a mean error of under 1 °C on a baseline temperature of 20 °C).

Regarding the issues observed in August, several factors may explain the anomalies:

- Connectivity problems: Between July and September, both classrooms experienced connectivity issues that resulted in significant data loss. Measurements lasting fewer than four hours were interpolated, whereas those exceeding four hours were left blank.
- Maintenance limitations: Due to the aforementioned connectivity problems, maintenance activities— which require access to the internet network—could not be performed during the campus summer break.
- Access restrictions: During the first three weeks of August, access to the first and second floors was restricted to students. This may have led to variations in the ventilation system setpoints or even forced the shutdown of the cooling battery systems.

Regarding the problems observed in the months of January and August, several factors may explain the anomalies. During the periods where data inconsistencies were encountered, the sensors experienced connectivity issues, and, due to the Campus closure periods for holidays, it was not possible to intervene to restore them. Also, given the closure period, changes were made to the schedules that could not be tracked, and this likely affected the simulation results.

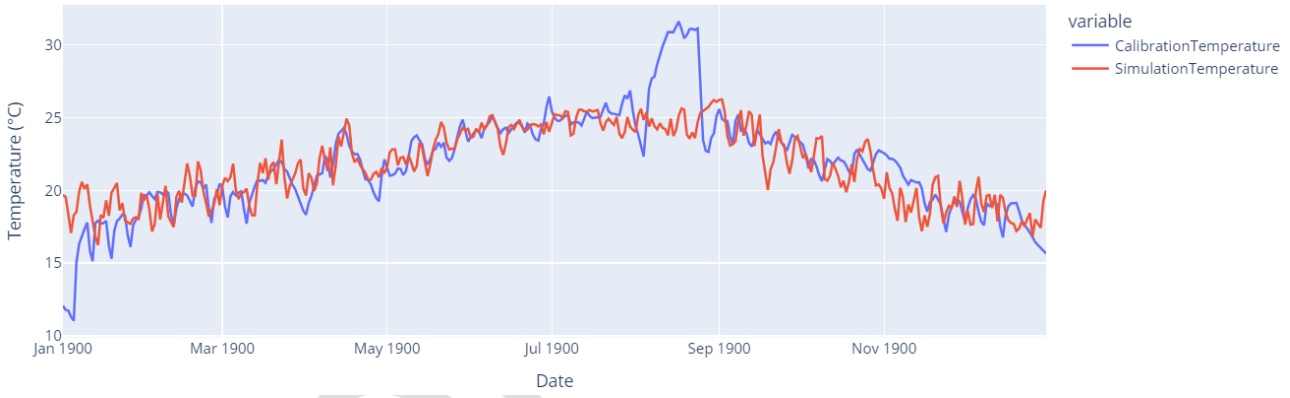

Figure 37: Daily Real vs Simulated data after the model calibration, Building 09.

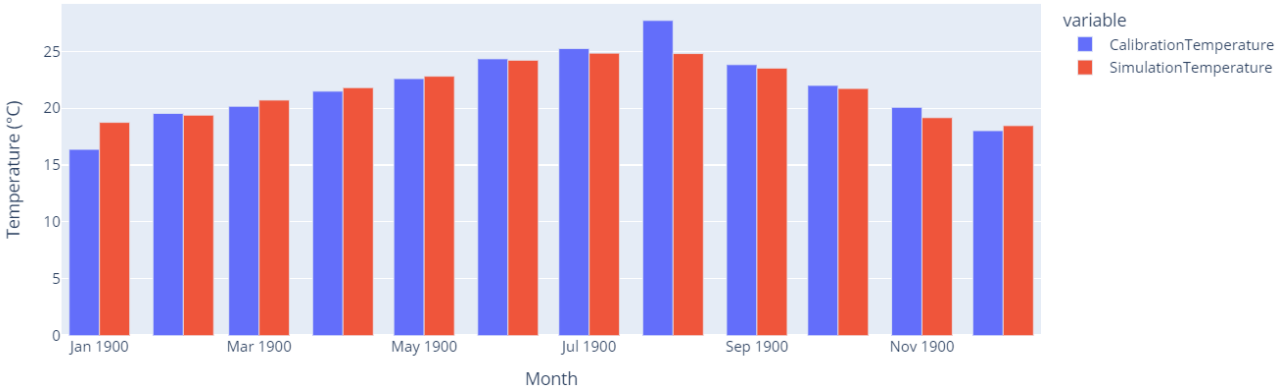

Figure 38: Monthly Real vs Simulated data after the model calibration, Building 09.

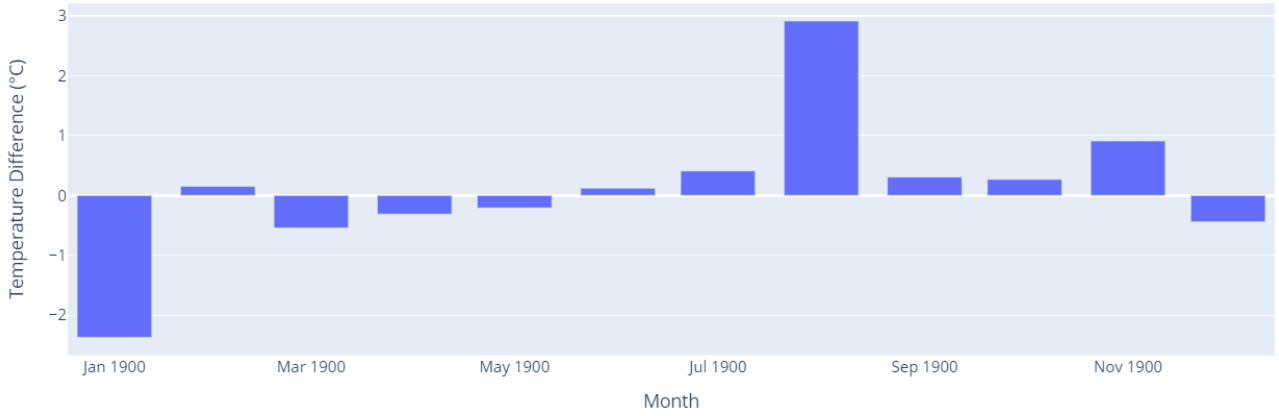

Figure 39: Monthly difference between mean real temperature and mean simulated temperature, after the model calibration, Building 09.

3.2.2 BEM simulation

Following the successful calibration of the energy model, a simulation was conducted using EnergyPlus version 24.1.0 to calculate KPIs that reflect the building's current energy status for the year 2024.

3.2.2.1 Simulation settings

The simulation was set up with specific controls: zone sizing calculations were enabled, while system and plant sizing calculations were disabled. The time resolution was configured to 4 timesteps per hour, utilising a full exterior solar distribution approach. The PolygonClipping method was employed for the shadow calculation, with a periodic update every 30 timesteps and a maximum allowance of 15,000 figures for shadow overlap, ensuring detailed modeling of solar shading effects while keeping good runtime (40 seconds for Building 10 on an ordinary laptop). The overall simulation run period was defined as the whole year, covering January 1 to December 31, 2024.

3.2.2.2 Output variables

The simulation outputs considered in this study encompass a range of variables that capture both energy needs (used for calculating the KPIs) and indoor thermal conditions within the building zones (used for calibrating the model). All the variables were reported hourly for each zone for the entire year.

The output variables of energy demand are:

- **Zone Ideal Loads Supply Air Total Cooling Energy (TCE)**: provides an aggregate measure of the energy required from the ideal supply air system to cool the zones and maintain comfort during warmer periods. It was assumed to correspond to the energy demand for cooling the zones.
- **Zone Ideal Loads Supply Air Total Heating Energy (THE)**: provides an aggregate measure of the energy required from the ideal supply air system to heat the zones and maintain comfort during colder periods. It was assumed to correspond to the energy demand for heating the zones.
- **Zone Electric Equipment Electricity Energy (TEE)**: records the energy consumption of all electric equipment in the zone. Differently from the previous, this variable was not calibrated due to the lack of real data for the calibration.
- **Zone Lights Electricity Energy (TLE):** measures the energy used for lighting. This variable was also not calibrated due to the lack of real data for the calibration.

Additionally, there is the output variable "Zone Ideal Loads Supply Air Standard Density Volume Flow Rate" which, on a zone-by-zone basis, allows calculating the airflow rate (in m^3/s) delivered by the ideal loads supply air system serving the zone. This variable is used to calculate:

- **Zone Ventilation Electricity Energy (TVE)**: the electrical energy required by the air handling units (AHUs) to mechanically ventilate the zone. This variable is computed in Python by processing the EP results file, assuming that the electrical power of the AHU serving the zone remains constant. More precisely, for each zone and for each hour, if the Zone Ideal Loads Supply Air Standard Density Volume Flow Rate is greater than 0, then the power of the AHU serving the zone is taken into account, multiplied by the time unit to obtain kWh. Finally, the data are aggregated and summed at both the zone and building levels.

To these variables, we added a metric for counting the occupancy count in the zones according to the occupancy schedules in the IDF:

- **Zone People Occupant Count (POC)**: calculated hourly, offers insights into the number of individuals in each zone.

3.2.2.3 Simulation results

Results from EP simulation are provided at the building level in Table 9 and from Figure 40 to Figure 44. This data can be aggregated zone by zone or daily, monthly or yearly, as shown in Figure 45 and Figure 47.

Table 9: Simulation results at the building level from Energy Plus, Building 10 (calibrated).

Output Variable	ID	Value	Normalised Value
Zone Ideal Loads Supply Air Total Heating Energy	*TCE*	*1,058,655 kWh*	*176.11 kWh/m²* *35.42 kWh/m³*
Zone Ideal Loads Supply Air Total Cooling Energy	*THE*	*360,503 kWh*	*59.97 kWh/m²* *12.06 kWh/m³*
Zone Lights Electricity Energy	*TEE*	*27,358 kWh*	*4.55 kWh/m²*
Zone Lights Electricity Energy	*TLE*	*14,068 kWh*	*2.34 kWh/m²*
Zone Ventilation Electric Energy	*TVE*	*185,748 kWh*	*30,9 kWh/m²* *6,22 kWh/m³*

Table 10: Simulation results at the building level from Energy Plus, Building 09 (calibrated).

Output Variable	ID	Value	Normalised Value
Zone Ideal Loads Supply Air Total Heating Energy	*TCE*	*1,383,730 kWh*	*154,76 kWh/m²* *31.18 kWh/m³*
Zone Ideal Loads Supply Air Total Cooling Energy	*THE*	*1,097,850kWh*	*122.79 kWh/m²* *24.74 kWh/m³*
Zone Lights Electricity Energy	*TEE*	*209,174 kWh*	*23,39 kWh/m²*
Zone Lights Electricity Energy	*TLE*	*69,567 kWh*	*7.78 kWh/m²*
Zone Ventilation Electric Energy	*TVE*	*459,929 kWh*	*51,44 kWh/m²* *10,36kWh/m³*

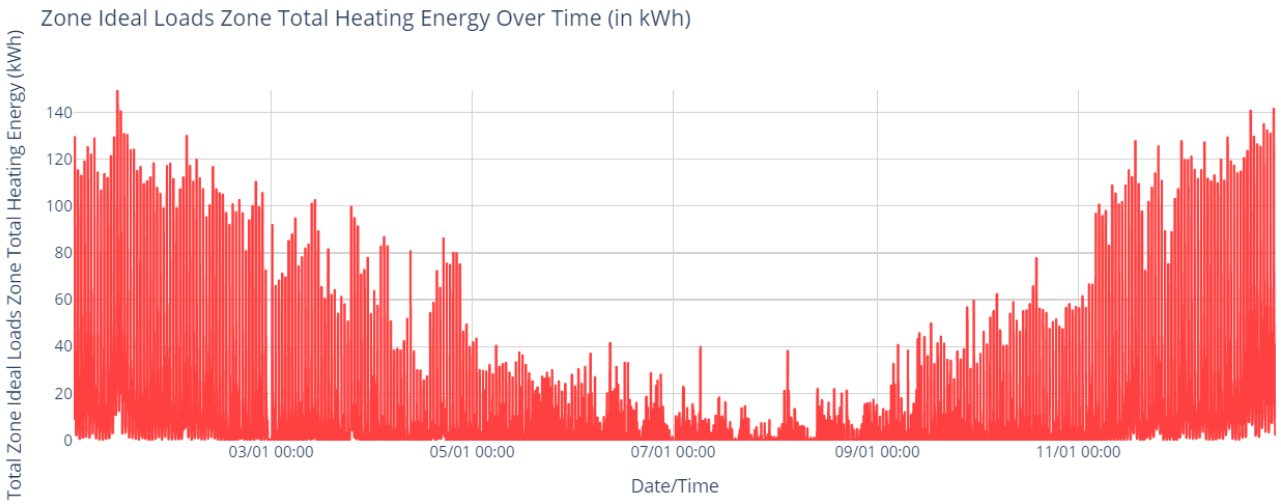

Figure 40: Hourly TCE resulting from the EP simulation, Building 10.

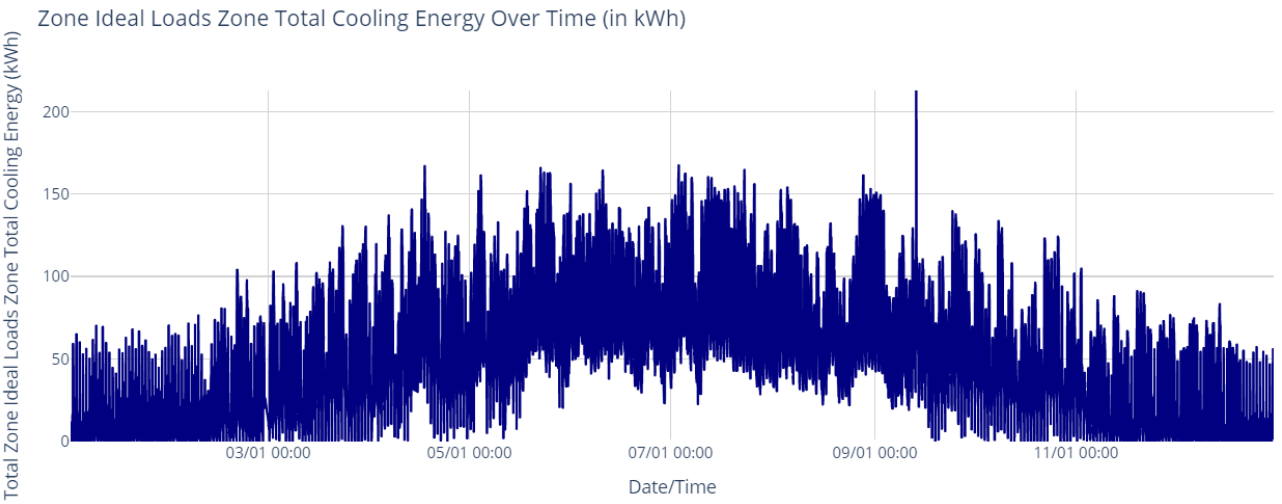

Figure 41: Hourly THE resulting from the EP simulation, Building 10.

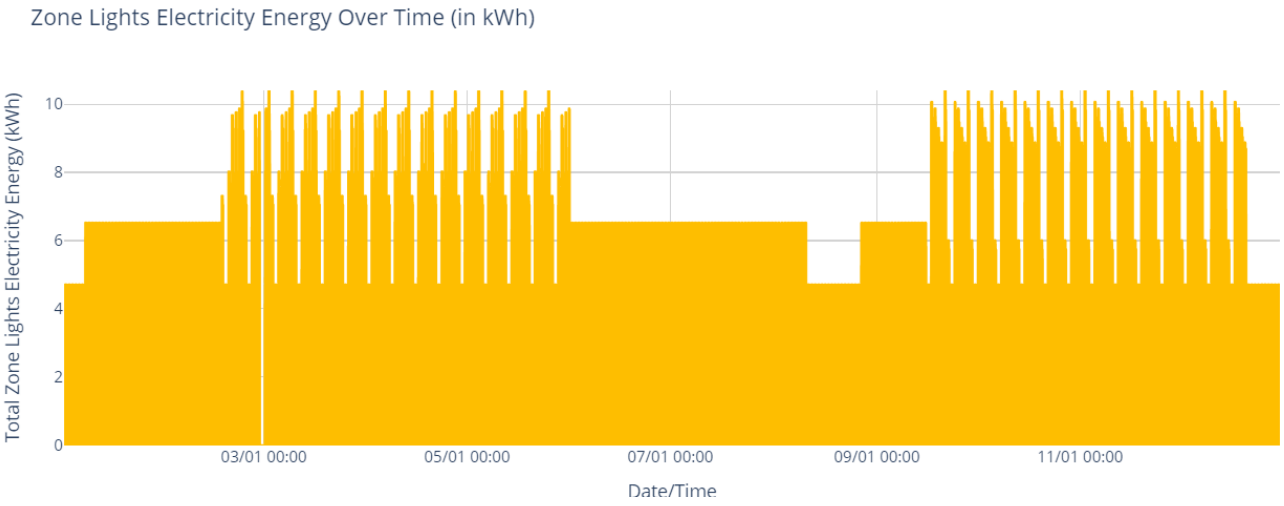

Figure 42: Hourly TEE resulting from the EP simulation, Building 10.

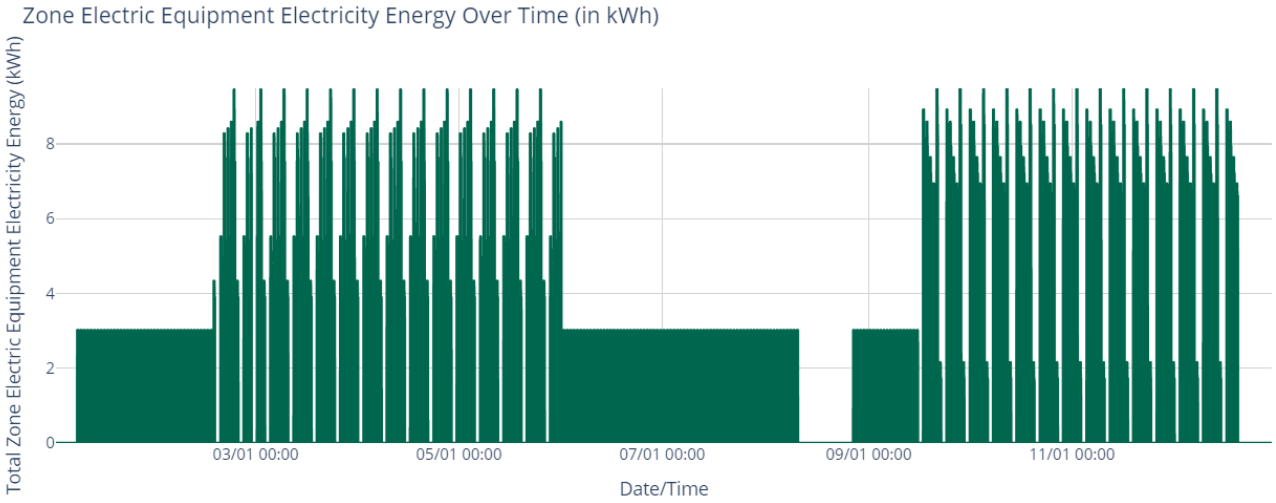

Figure 43: Hourly TLE resulting from the EP simulation, Building 10.

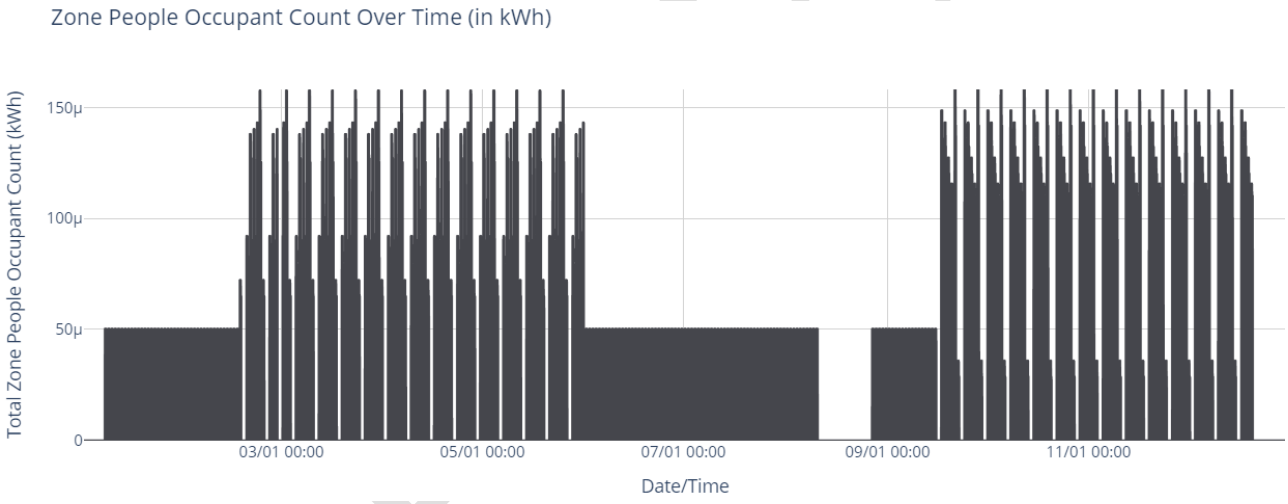

Figure 44: Hourly POC resulting from the EP simulation, Building 10.

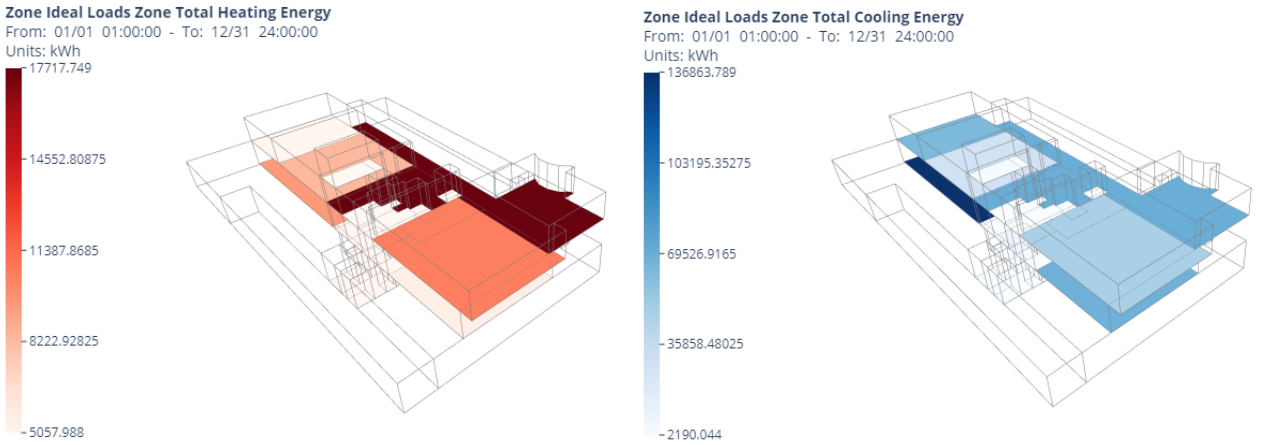

Figure 45: THE and TCE by zone, Building 10.

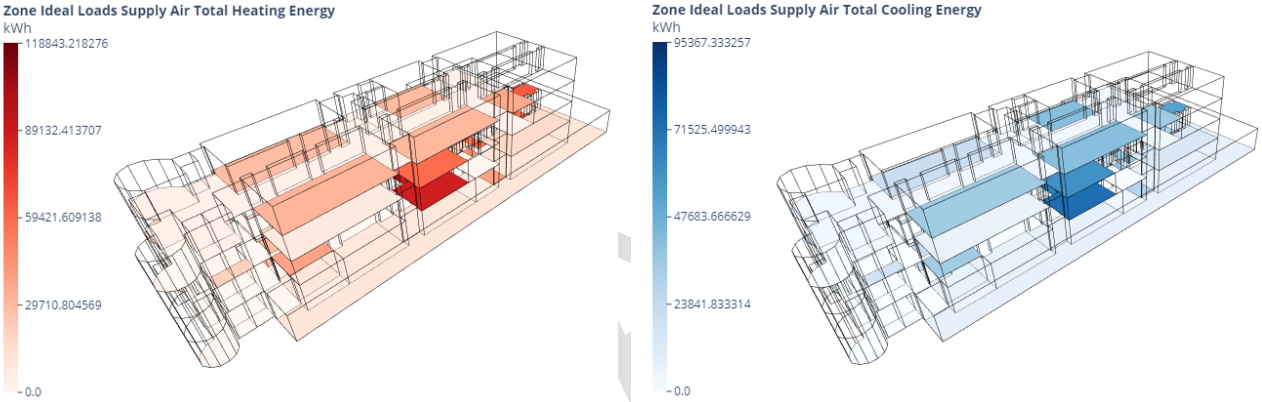

Figure 46: THE and TCE by zone, Building 09.

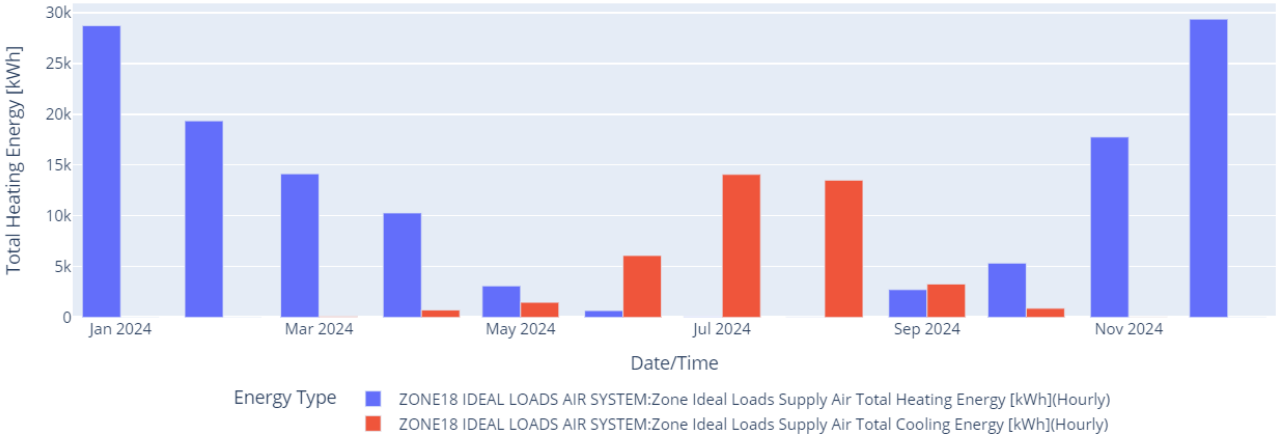

Figure 47: Monthly THE and TCE for the classroom used for calibration, Building 10.

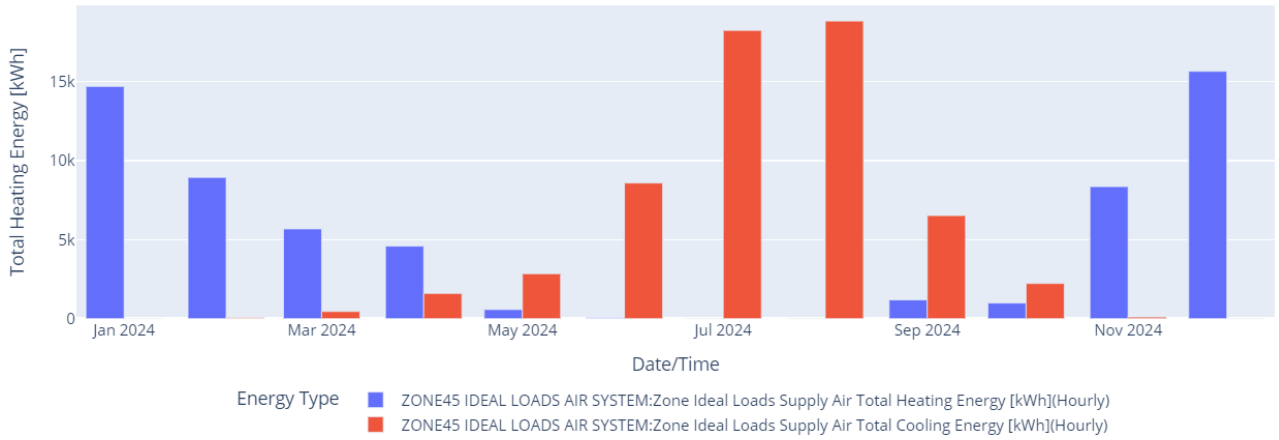

Figure 48: Monthly THE and TCE for the classroom used for calibration, Building 09.

3.2.2.4 Calculation of DIGITMAN's KPIs

The following KPIs are computed starting from the EP results file.

KPI O-1.1 - Utilization Rate

This is the percentage of time that a specific space is occupied compared to the total available time. It helps determine how long the space is being utilised.

The calculation procedure is as follows:

1. For each occupied zone, count the number of hours during which the "People Occupant Count" is greater than one (i.e., when at least one person is present in the zone, indicating that the zone is occupied).
2. For each zone, compute the zone utilisation rate by dividing the number of occupied hours by 8,760 (the total number of hours in a year).
3. For the entire building, calculate the weighted average of the zone utilisation rates, using the maximum number of occupants in each zone as the weighting factor.

KPI O-1.2 - Occupancy Rate

This is the average number of occupants in a space compared to its maximum capacity. It provides insights into the utilisation of individual space and identifies spaces that may be over or underutilised.

The calculation procedure is as follows:

1. For each occupied zone, for every hour of the year, divide the People Occupant Count by the maximum number of occupants for that zone to obtain the hourly occupancy rate.
2. For each zone, calculate the average of the hourly occupancy rates to obtain the zone's yearly occupancy rate.
3. For the entire building, calculate the weighted average of the zone occupancy rates, using the maximum number of occupants in each zone as the weighting factor.

KPI O-2.1 - Energy Need

This is equivalent to the tons of oil equivalent to the ideal energy needed for heating, cooling, lighting, ventilation and using the appliances of a building or its zones. It helps understaing how much electricity is needed for operating the zone or building.

The calculation procedure is as follows:

1. For each zone, obtain the annual values for THE, TCE, TLE, TEE, and TVE.
2. If natural gas is used for heating:
    a. Multiply THE (in kWh, as provided by EnergyPlus) by the conversion factor from kWh to Smc.
    b. Multiply the resulting THE (in Smc) by the conversion factor from Smc to TEP.
    c. Sum TCE, TLE, TEE, and TVE to obtain the total electricity consumption, and multiply this sum by the conversion factor from kWh to TEP.
    d. Add the converted THE (in TEP) to the total electricity (in TEP).
3. If electricity is used for heating:
    a. Sum THE, TCE, TLE, TEE, and TVE to obtain the total electricity consumption, and multiply this sum by the conversion factor from kWh to TEP.

KPI O-3 - Energy Costs

Measures the expenses associated with electricity, natural gas, and district heating consumption within a building or facility. It allows for identifying areas of higher energy consumption to implement energy-saving measures and reduce overall energy expenses towards energy efficiency.

The calculation procedure is as follows:

1. For each zone, obtain the annual values for THE, TCE, TLE, TEE, and TVE.
2. If natural gas is used for heating:
    a. Multiply THE (in kWh, as provided by EnergyPlus) by the conversion factor from kWh to Smc.
    b. Multiply the resulting THE (in Smc) by the conversion factor from Smc to EUR, getting heating energy cost (HEC).
    c. Sum TCE, TLE, TEE, and TVE to obtain the total electricity consumption, and multiply this sum by the conversion factor from kWh to EUR, obtaining electricity energy cost (EEC).
    d. Sum HEC and EEC to get the total costs (in EUR).
3. If electricity is used for heating:
    a. Sum THE, TCE, TLE, TEE, and TVE to obtain the total electricity consumption, and multiply this sum by the conversion factor from kWh to EUR.

KPI O-4 - CO2 Emissions due to Energy Use

Amount of CO2 emissions produced as a result of the building's energy consumption from the grid. It allows for tracking of the environmental impact in terms of CO emissions due to the use of energy within a building or its zones.

The calculation procedure is as follows:

1. For each zone, obtain the annual values for THE, TCE, TLE, TEE, and TVE.
2. If natural gas is used for heating:
    a. Multiply THE (in kWh, as provided by EnergyPlus) by the conversion factor from kWh to Smc.
    b. Multiply the resulting THE (in Smc) by the conversion factor from Smc to kgCO2/eq, getting hypothetical emission due to heating use (HEE).
    c. Sum TCE, TLE, TEE, and TVE to obtain the total electricity consumption, and multiply this sum by the conversion factor from kWh to kgCO2/eq, obtaining hypothetical emission due to electricity use (EEE).
    d. Sum HEE and EEE to get the total costs (in kgCO2/eq).
3. If electricity is used for heating:

a. Sum THE, TCE, TLE, TEE, and TVE to obtain the total electricity consumption, and multiply this sum by the conversion factor from kWh to kgCO2/eq.

The conversion factors used are:

- kwh_to_smc [57] = 10.944
- smc_to_tep [58] = 0.000082
- kWh_to_tep [58] = 0.0008598
- smc_to_EUR [59] = 0.4987
- kWh_to_EUR [60] = 0.164
- smc_to_CO2 [61] = 2.02
- kWh_to_CO2 [61] = 0.467

4 How-To logic

Regarding the "How-to" methodology, this section will explain the operations performed for Task “T3.2 – Operation data acquisition and post-processing”, and Task “T3.3 – Operation predictive methodology development”. The first task, concerning the acquisition of operational data on indoor environments, was described in WP1 and is briefly summarized below. It involves the process of downloading the data and concatenating them to obtain a single file for each specific classroom. The concatenation phase also includes data post-processing operations and the extraction of ancillary information.

The development of T3.3 involves "black-box" methodologies for energy assessment. In the methodology section, the algorithms that were considered and selected for the creation of the predictive modules will be detailed.

4.1 Methods and Tools

4.1.1 Overview and database structure

The Data Acquisition techniques, previously detailed in deliverables related to WP1 and based on the Brick schema for measurement classification, were subsequently implemented in the following phases, which are briefly summarized below:

1. Definition of the overall database of measurements, related to the monitored classrooms in the POLIMI building, as shown in Section **Errore. L'origine riferimento non è stata trovata.**.
2. Selection of pre-processing methodologies to obtain the parameters considered in the KPIs.
3. Definition of the selected methodologies for data set analysis and the creation of a decision support system using predictive methods for the recognition and prediction of temperature and consumption profiles of the case studies.
4. Multi-criteria analysis (to be developed in WP5).

The final application and validation are performed on the monitored classrooms within the POLIMI building stock, described in D1.1 of WP1.

This section presents the structure of the database obtained from the data acquired during the measurement period by the sensors. The first phase consists of aggregating data from all sources and acquisition systems; in this phase, all data are organized to standardize them. Subsequently, through scripts written in Python, all data are aligned and concatenated. It is in this phase that derived measurements are calculated, static information is inserted, and the parameters necessary for the calculation of KPIs are extracted, with the aim of having a single dataframe containing all the necessary measurements.

The first process executed is the calculation of the daily mean temperature based on the data from the weather station. The Running Mean Temperature (RMT) is then calculated using a moving average with a three-day window, which is crucial for labeling and classifying days according to their respective seasons. The possible classifications are Winter and Summer, with the condition that if the RMT is below 15°C, the season is classified as winter; otherwise, it is classified as summer. Based on this, indoor temperatures are classified, as presented in D1.1, at the various measurement points. Following this initial classification, the comfort level within the indoor spaces is then classified according to the various measurement points. The calculation of PMV (Predicted Mean Vote) and PPD (Predicted Percentage of Dissatisfied) is performed using the pyThermalComfort library [62]. Since the measurement period constituting the dataset extends for almost a year, it is necessary to parameterize some variables, both on a daily basis and based on the

operation of ventilation systems. The metabolic rate (MET) of the occupants is set to 1.2, which is the value assigned to students in a classroom. The estimation of clothing (clo) is performed based on the temperature measured at 6 a.m. If the temperature is below -5°C, the clo value is 1. If the temperature is between -5°C and 5°C, the calculation is:

$$clo = 0.818 - 0.0364 * t_{out}$$

Between 5°C and 26°C, the calculation is:

$$clo = 10^{-0.1635 - 0.0066 * t_{out}}$$

Any temperature above that sets the clothing value to 0.46. Air velocity was determined based on the periods when the machinery was operating. Knowing the flow rates of the machinery and the size of the supply vents, it is possible to estimate the air velocity, an estimate validated through measurements in the classrooms. When the AHU is off, an air velocity of zero is assumed, but a perceived velocity is still considered, expressed by the following formula:

$$v = v_{air} + 0.3 \times (met - 1)$$

Where v_{air} is the initial air velocity, set to 0, the met value represents the metabolic rate of the occupant, and the term *0.3 × (met - 1)* adjusts the air velocity according to the activity level. Hence all the parameters have been calculated, the final calculation for PMV and PPD can be finished and appended to the dataframe.

Occupancy estimation was therefore performed based on the CO_2 levels measured inside the classroom. This estimation also depends directly on the operation of the ventilation systems and requires two separate formulas. The first formula is used when the ventilation systems are in operation, taking into account that the building does not have operable windows and therefore lacks natural ventilation, estimating the number of people based on the increase or decrease in carbon dioxide. The formula, derived from the literature[63], is presented below:

$$n = \frac{Q}{k * e^{-\frac{Q}{V}(i-s)}} * (C_i - C_0 - (C_s - C_0) * e^{-\frac{Q}{V}(i-s)}$$

Where *n* represents the number of occupants, *Q* is the air flow rate, *V* is the volume of the room, C_s is the CO_2 concentration at the initial instant, C_i is the concentration at the i-th instant, and C_0 is the CO_2 concentration outside the building, which is assumed to be constant throughout the day and is measured inside the classroom after one hour of machinery operation, knowing the air volume has been changed six times, it is assumed that the CO_2 level is equal to the outdoor level, with the classroom still empty. Finally, *k* is the per capita CO_2 production rate for people who are seated and studying; it is assumed that the value is constant and equal to 0.0056 m^3/s[64]. The second formula is used, precisely, when the systems are off and measures the variations of CO_2 over time, relating them to the volume. The formula is as follows:

$$n = \frac{(C_i - C_s) * V}{k * (i - s)}$$

The final steps consist of normalizing the occupancy values for the calculation of KPIs, where the occupancy, derived from the schedules provided by the administrative office, is scaled as a percentage value, and the same process is performed for the estimated occupancy values. The delta between the estimated actual occupancy and the scheduled occupancy is then calculated to determine the Standard Deviation of Occupancy Variability. These estimations are made for the calculation of the KPIs related to Use, namely the O-1 KPIs.

4.1.2 Data preprocessing

Data preparation is the initial step performed to enable analysis and the application of machine learning methodologies to the data. The operations involve loading and preliminary cleaning of the data. Since the information is distributed across multiple CSV files, a function is developed to import each dataset,

appropriately rename columns, and remove unnecessary ones. To improve dataset quality and eliminate redundant or insignificant variables, a filter based on a list of patterns is applied, removing columns that do not make a relevant contribution to the analysis. The output of the function consists of two distinct DataFrames, corresponding to the two source files, which are subsequently concatenated into a single data structure to allow a joint analysis of information from different sources.

After this initial standardization phase, the dataset is further refined by filtering based on two main criteria: seasonality and occupancy pattern. Only data corresponding to specific seasonal intervals (summer, winter) are selected in order to analyze specific climatic conditions. Concurrently, the dataset is filtered according to the occupancy level of the spaces, excluding situations not representative of the normal use of the spaces. These two criteria allow for the isolation of data portions more consistent with the analysis objectives, reducing the presence of irrelevant or potentially distorting elements.

To ensure the reliability of the analysis, an outlier detection and removal process based on the Interquartile Range (IQR) was implemented. This method allows for the identification and removal of anomalous values that significantly deviate from the central distribution of the data. For each numerical variable, the first quartile (Q1) and the third quartile (Q3) are calculated, from which the Interquartile Range (IQR = Q3 - Q1) is obtained. Values that fall outside the range defined by [Q1 - 1.5 × IQR, Q3 + 1.5 × IQR] are considered outliers and replaced with null values (NaN). However, this methodology is not applied to variables related to the operation of the Air Handling Units (AHUs) because these systems operate in discrete mode with distinct ON/OFF states. When the AHU is deactivated, some parameters may assume very low or null values, which would be incorrectly classified as outliers by the IQR method. To avoid the removal of valid data, these variables were excluded from the filtering process, ensuring a correct representation of their operating conditions.

A critical aspect of the dataset is the presence of discontinuities due to interruptions in data transmission, which introduce missing values (NaN). To avoid altering originally absent data, a mask identifying the NaNs already present in the dataset is generated before applying the IQR method. Once the outliers are removed, the new missing values are interpolated only if they result from the removal phase and were not already absent in the original dataset. Interpolation is performed using a linear method, which ensures a gradual transition between adjacent points, preserving the temporal consistency of the series.

After data cleaning, the dataset is segmented into daily intervals, as shown in Figure 49 to Figure 53, dividing the time series into subsets corresponding to individual days. For each day, a complete time interval with a 10-minute resolution, corresponding to the original sampling frequency, is created.

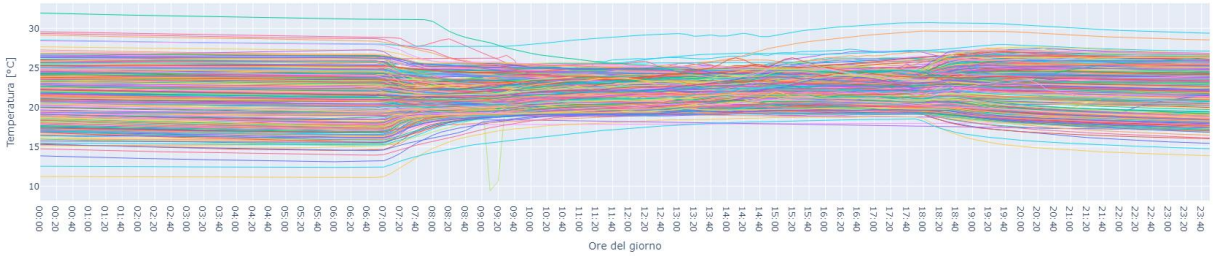

Figure 49: Daily sequences of the Indoor Temperature

Figure 50: Daily sequences of the CO_2

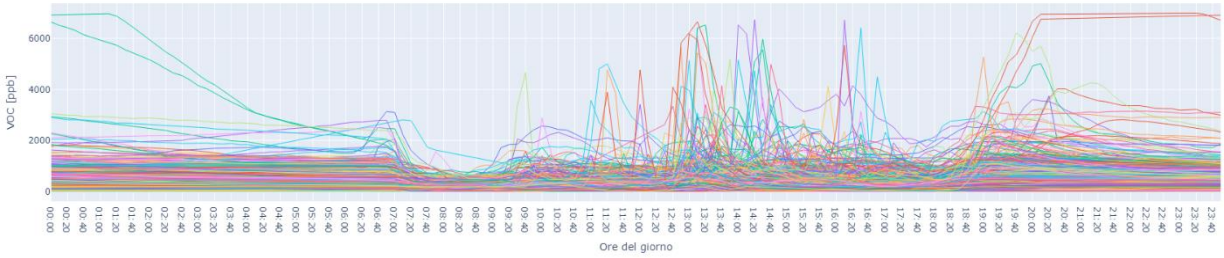

Figure 51: Daily sequences of the VOCs

Figure 52: Daily sequences of the PM2.5

Figure 53: Daily sequences of Thermal Power

However, not all daily sequences contained a complete set of measurements due to possible interruptions in data collection or the presence of missing values. To ensure the reliability of the analysis, only days in which all required measurements are available for the variables of interest are selected. This process ensures that each daily segment is representative of the complete system dynamics, avoiding distortions due to partial or incomplete data. Once segmented, the data are organized into separate DataFrames for each monitored parameter, allowing a detailed analysis of daily trends and facilitating the subsequent clustering and modeling phase. The entire process described constitutes the initial phase of data pre-processing, which guarantees the quality and consistency of the dataset before the application of advanced analysis techniques.

4.1.3 Data postprocessing

Following the pre-processing phase, our focus shifted to the identification of latent temporal patterns within the extensive dataset. To address this challenge, we turned to clustering, an unsupervised learning technique has been employed to allure the identification of similar observations, thus revealing intrinsic structures in the data that would otherwise be difficult to discern. The resulting reduction in dimensionality and complexity is a crucial step in making the data more manageable and interpretable, paving the way for subsequent modeling analyses. Given the multifaceted nature of our data, a multi-algorithm approach was adopted, comparing the performance of four distinct clustering techniques. This strategy allowed us to assess which algorithm was best suited to capture the specificities of the different variables under consideration.

The first algorithm considered was K-Means, a classic and widely used partitional method. K-Means operates iteratively to minimize the sum of squared distances between data points and the centroid (the mean) of the cluster to which they are assigned. Its simplicity and speed make it a natural starting point; however, this method is sensitive to outliers and assumes that clusters are spherical and of similar size.

In parallel, we explored K-Medoids, a more robust variant of K-Means. Instead of utilizing the mean as the cluster center, K-Medoids selects an actual data point (the medoid) that minimizes the sum of dissimilarities with the other points in the cluster. This characteristic renders it less influenced by outliers and more suitable for data with non-spherical distributions.

In order to overcome the limitations of partitional methods, two density-based algorithms were introduced: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and its hierarchical derivative, HDBSCAN (Hierarchical DBSCAN). The distinguishing feature of DBSCAN is its capacity to identify clusters of arbitrary shape and to automatically detect noise points (outliers). The operation of HDBSCAN is predicated on two key parameters: eps, which defines the radius of a neighborhood, and “min_samples”, which establishes the minimum number of points within that neighborhood for a point to be considered a "core point" and give rise to a cluster. The HDBSCAN algorithm extends DBSCAN by constructing a hierarchy of clusters at different density scales. This approach facilitates the identification of clusters of varying sizes and densities, thereby offering enhanced flexibility and adaptability to complex data structures. HDBSCAN necessitates the definition of “min_cluster_size”, which represents the minimum size of a cluster, and “min_samples”, which plays a role analogous to that in DBSCAN, albeit with some interpretative nuances.

4.1.3.1 Clustering phase and hypertuning

Prior to the implementation of the clustering algorithms, the data underwent a standardization transformation using PowerTransformer. This operation is imperative when variables manifest skewed or non-normal distributions, a prevalent occurrence in environmental data. PowerTransformer implements a non-linear transformation (specifically, a Yeo-Johnson transformation) that aligns the data to a distribution more closely resembling the Gaussian. This step has been shown to enhance the efficacy of algorithms such as K-Means and K-Medoids, which are particularly sensitive to deviations from normality and implicitly assume a certain symmetry in the data. To assess the quality of the obtained clusters and guide the selection of the optimal model, two complementary metrics were adopted: the Silhouette Score and the Davies-Bouldin Index (DBI).

- The Silhouette Score is a metric that evaluates the cohesion and separation of clusters by calculating the difference between the average distance from other points in the same cluster (cohesion) and the average distance from the nearest cluster (separation). A Silhouette value close to +1 indicates a satisfactory grouping, with the point well-placed in its own cluster and far from others. Conversely, values approaching 0 signify the presence of overlap between

clusters, while negative values suggest that the point in question may have been erroneously assigned to a different cluster.

- The DBI, conversely, emphasizes the internal dispersion of clusters and their mutual distance. A low DBI signifies compact and well-separated clusters, whereas high values indicate dispersed or overlapping clusters.

Each clustering algorithm is characterized by one or more hyperparameters, which influence the algorithm's behavior and performance. Achieving optimal results necessitates identifying the combination of hyperparameters that is most appropriate for the particular characteristics of the data under consideration. This process, termed hypertuning or hyperparameter optimization, was systematically executed for each algorithm. In the case of K-Means and K-Medoids, the number of clusters (n_clusters) varied within a range from 2 to 20. For each setting of "n_clusters", the Silhouette Score, the DBI, and their delta were calculated, and the configurations that yielded the optimal results were recorded. For DBSCAN, two parameters were optimized: eps, varied between 0.1 and 3.0 with increments of 0.1, and "min_samples", explored between 2 and 50 with unit increments. Only configurations that generated at least 3 clusters, excluding those that produced an excessive number of small clusters or a single dominant cluster, have been considered. In the case of HDBSCAN, have been varied "min_cluster_size" between 5 and 50 with increments of 5, and "min_samples" between 2 and 50 with increments of 2. Again, configurations with at least 3 clusters have been favored, discarding those that did not meet this criterion. For DBSCAN and HDBSCAN, the Silhouette Score was the sole metric of evaluation, excluding the DBI. This exploration of the hyperparameter space enabled the identification of the configuration that optimally captured the underlying structure of the data for each algorithm and variable, the results of the best clusters are presented in Table *11*. It is noteworthy that the DBSCAN and HDBSCAN algorithms did not consistently demonstrate optimal performance. In the majority of cases, such as those involving CO2, VOC, and PM, the algorithms identified only a single cluster.

Table 11: Parameters identified for each clustering method after hyperparameters tuning

Cluster	*Method*	*Number of cluster identified*	*Silhouette score*	*DBI score*
Temperature	*K-Means*	*2*	*0.5618*	*0.5810*
	K-Medoids	*2*	*0.5637*	*0.5740*
	DBSCAN	*3*	*0.3078*	-
	HDBSCAN	*3*	*0.1983*	-
CO_2	*K-Means*	*2*	*0.2798*	*1.3210*
	K-Medoids	*2*	*0.2688*	*1.3738*
	DBSCAN	*1*	*0.0841*	-
	HDBSCAN	*3*	*-0.0421*	-
VOC	*K-Means*	*2*	*0.3020*	*1.2232*
	K-Medoids	*2*	*0.3001*	*1.2181*
	DBSCAN	*1*	*-0.0912*	-
	HDBSCAN	*3*	*-0.2660*	-
PPM2.5	*K-Means*	*2*	*0.3894*	*0.9713*
	K-Medoids	*2*	*0.3921*	*0.9458*
	DBSCAN	*165*	*-0.0668*	-
	HDBSCAN	*3*	*-0.1345*	-
AHU_Total Thermal Power	*K-Means*	*8*	*0.6451*	*0.5248*
	K-Medoids	*4*	*0.6449*	*0.6288*
	DBSCAN	*1*	*0.5679*	-
	HDBSCAN	*2*	*0.6021*	-

Following the clustering process, an analysis was conducted to verify the balance of the clusters and their distribution. The purpose of the analysis is to verify the quality of the clusters identified after the hypertuning process. As anticipated, the process did not yield robust clusters for the pollutants; this can be observed from the results of the balance analysis. For that reason k-Means clusters has been selected as clustering algorithm due to the general performance and overall balance in cluster distribution. Table 12 shows the results of the analysis, highlighting the strong imbalance for the DBSCAN and HDBSCAN methods, where the label -1, representing an invalid cluster, was assigned in the majority of the clusters. For the sake of brevity, the summary values have been included as a link to the table, as there are 165 clusters.

Table 12: Cluster Balance Analysis

K-Means			K-Medoids			DBSCAN			HDBSCAN		
Cluster N	Items	%	Cluster N	Items	%	Cluster N	Items	%	Cluster N	Items	%
Temperature											
1	243	60,599%	0	247	61,596%	-1	105	26,185%	-1	155	38,653%
0	158	39,401 %	1	154	38,404%	1	101	25,187%	0	98	24,439%
-	-	-	-	-	-	0	98	24,439%	1	76	18,953%
-	-	-	-	-	-	2	97	24,190%	2	72	17,955%
CO_2											
1	422	51,463%	2	281	34,268%	-1	742	90,488%	-1	510	62,195%
0	398	48,537 %	0	204	24,878%	0	78	9,512%	2	175	21,341%
-	-	-	3	179	0,21829268	-	-	-	1	77	9,390%
-	-	-	1	156	0,1902439	-	-	-	0	58	7,073%
VOC											
0	452	55,122 %	0	474	57,805%	-1	765	93,293%	-1	699	85,244%
1	368	44,878 %	1	346	42,195%	0	55	6,707%	1	105	12,805%
-	-	-	-	-	-	-	-	-	2	9	1,098%
-	-	-	-	-	-	-	-	-	0	7	0,854%
PM2.5											
DigitMan_Table_ClusterResults.xlsx											
Thermal Power											
0	61	40,397 %	1	64	42,384%	-1	91	60,265%	0	83	54,967%
2	33	21,854 %	0	51	33,775%	0	60	39,735%	1	61	40,397%
4	29	19,205 %	3	33	21,854%	-	-	-	-1	7	4,636%
1	25	16,556%	2	3	1,987%	-	-	-	-	-	-
5	1	6,623%	-	-	-	-	-	-	-	-	-
3	1	6,623%	-	-	-	-	-	-	-	-	-

The clustering results based on temperature data (Figure 54: Temperature cluster trend lines) have produced the following daily profiles, which clearly encompass indoor temperatures during both summer (cluster 0, red line) and winter (cluster 1, blue line) periods. In Figure 55 are presented the daily profiles associated with the daily average profile of the clusters.

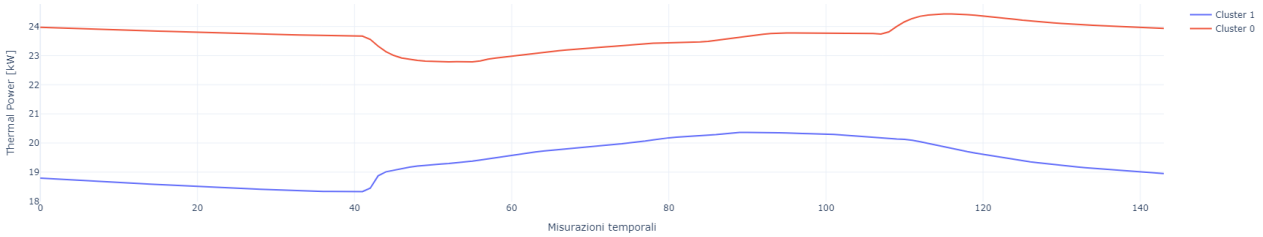

Figure 54: Temperature cluster trend lines

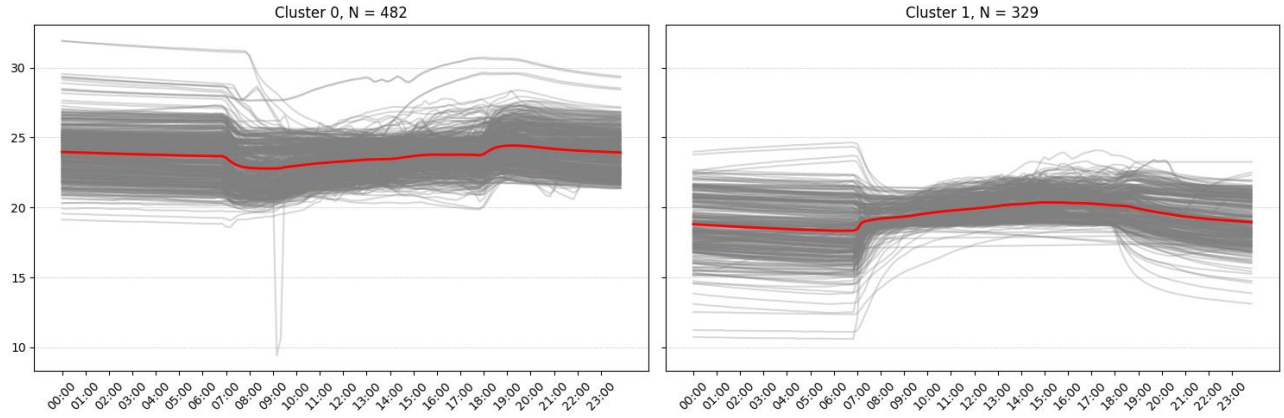

Figure 55: Cluster visualization of daily profiles associated with the daily average profile of the Temperature cluster

The clustering results based on thermal power data (Figure 56) have produced the following daily profiles, associated with corresponding daily profiles (Figure 57).

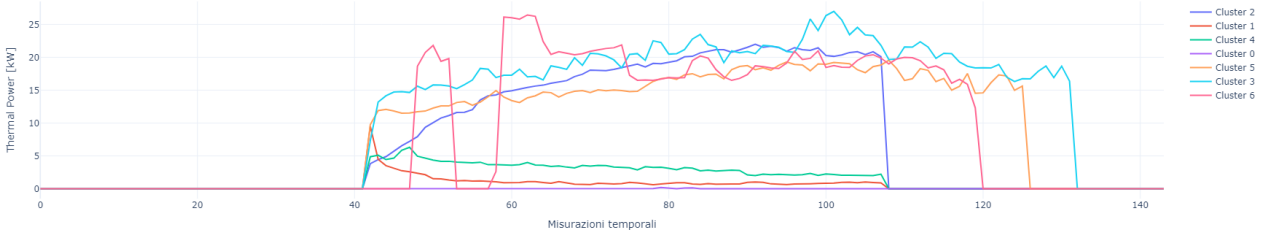

Figure 56: Thermal Power cluster trend lines

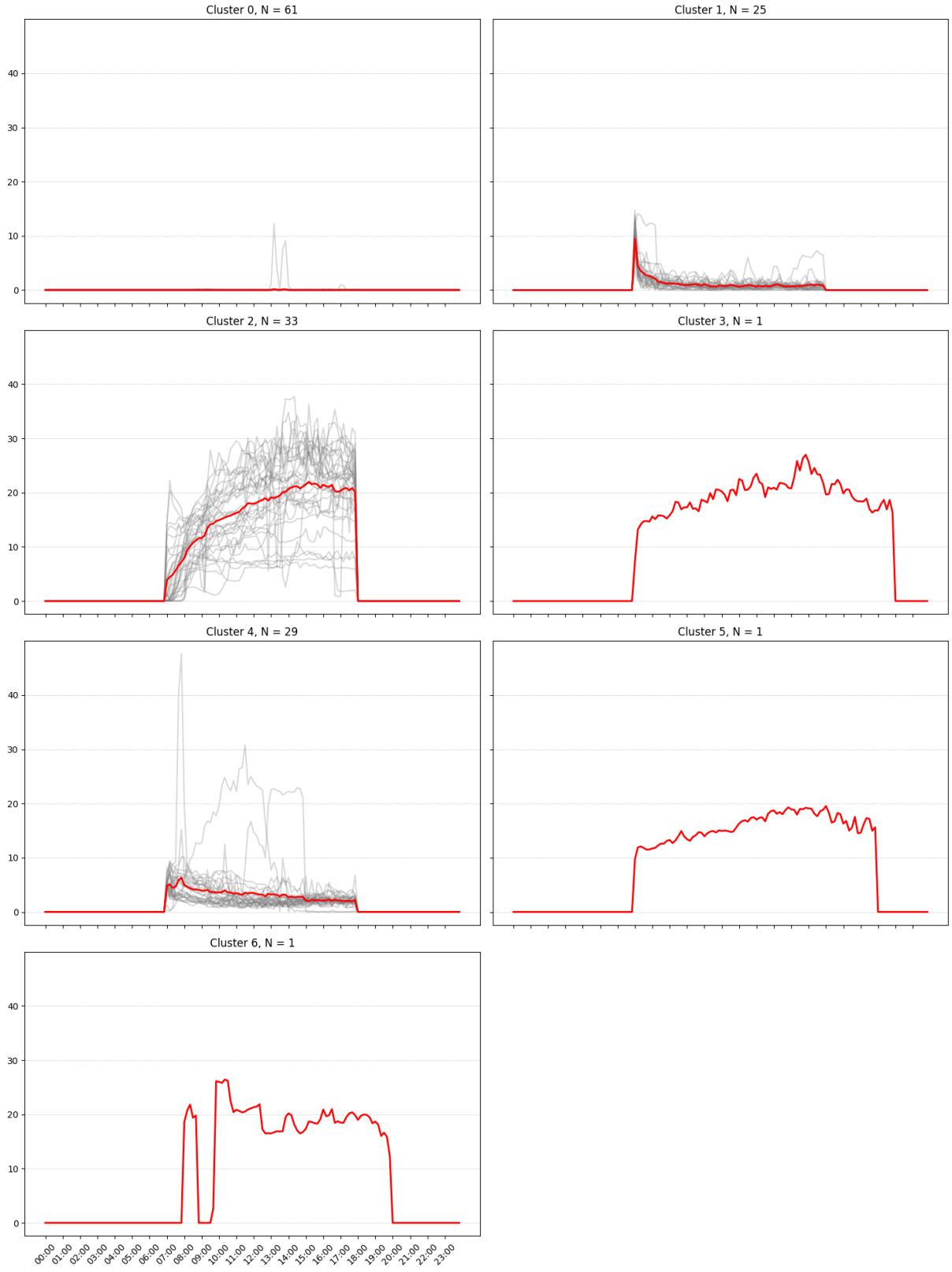

Figure 57: Cluster visualization of daily profiles associated with the daily average profile of the Thermal Power cluster

Following the completion of the clustering phase and the selection of the optimal models, a significant data aggregation phase was initiated. This phase had three objectives: (1) the consolidation of the results obtained, (2) the integration of information from disparate sources, and (3) the preparation of the data for the subsequent construction of predictive models. The initial step involved the addition of the cluster labels (assigned by the optimal algorithms) to the original transposed data, which represent the time series of the different variables. This enabled the association of each temporal observation with its cluster membership. Subsequently, a transformation was implemented, transitioning from a "wide" format (comprising one column for each hourly measurement) to a "long" format (comprising a single column for the hour and a column for the measured value). This transformation, designated as unpivoting or melting, was executed using the Pandas melt function. The "long" format is more conducive to time series analysis and modeling, as it enables the application of functions and transformations to the data.

The subsequent critical step entailed the incorporation of supplementary information from the original pre-processed DataFrame. This DataFrame encompassed variables measured at varying scales (building, classroom, single station) and with disparate temporal frequencies. To reintegrate this information into the aggregated data, we constructed a lookup dictionary. This dictionary maps the combination of date and time to specific values of variables present in the original DataFrame.

The culmination of this process was a refined, structured, and enriched dataset, prepared for the subsequent construction of classification models, such as decision trees, which will be the focus of future phases of our study. This dataset signifies an effective synthesis of the information contained in the original data, integrating the temporal patterns identified through clustering with the contextual variables that are pertinent for predictive modeling.

4.2 Predictive methods

4.2.1 Decision Tree – Random Forest

The Random Forest algorithm is an ensemble learning method that utilises a collection of decision trees for classification or regression tasks. To minimise correlation among the individual trees, each tree is constructed using a randomly resampled subset of the training data, a technique known as bootstrap aggregating (bagging). During prediction, each tree independently classifies the input, effectively casting a "vote" for its predicted class. The final prediction of the Random Forest is determined through a majority vote, aggregating the predictions of all individual trees.

4.2.1.1 Features selection – Correlation Matrix

The aggregated dataset (df_Temperature_DT.csv or df_TP_DT.csv), comprising predictor variables (features) and the target variable (cluster label), served as the starting point. Preliminary operations were executed, including the removal of the 'Date' column, deemed unnecessary for modeling as temporal information was already encoded within other variables. Missing values were handled conservatively, replaced with -1 to prevent data loss and treat them as a separate informative category. The dataset was then partitioned into two sets: X, containing the predictor variables, and y, representing the target variable.

Prior to model construction, a correlation analysis was undertaken to elucidate the relationships between the predictor variables and the target variable. This analysis served two primary purposes: firstly, to identify the most promising features, recognizing that variables exhibiting a high absolute correlation with the target are potentially more informative for classification; and secondly, to identify potential redundancies, where highly correlated predictor variables might provide similar information, suggesting the possibility of removing some for model simplification.

Two tools were used for the correlation analysis:

1. The first tool was the general correlation matrix, which is a method of displaying the Pearson correlations between all pairs of predictor variables.
2. The second tool was the correlation with the target, which calculates the Pearson correlation between each predictor variable and the target variable. This provides a direct measure of the potential importance of each feature for classification.

The analysis yielded features with a correlation (in absolute value) greater than a predefined threshold (0.4), enabling the selection of the most relevant variables (Figure 58).

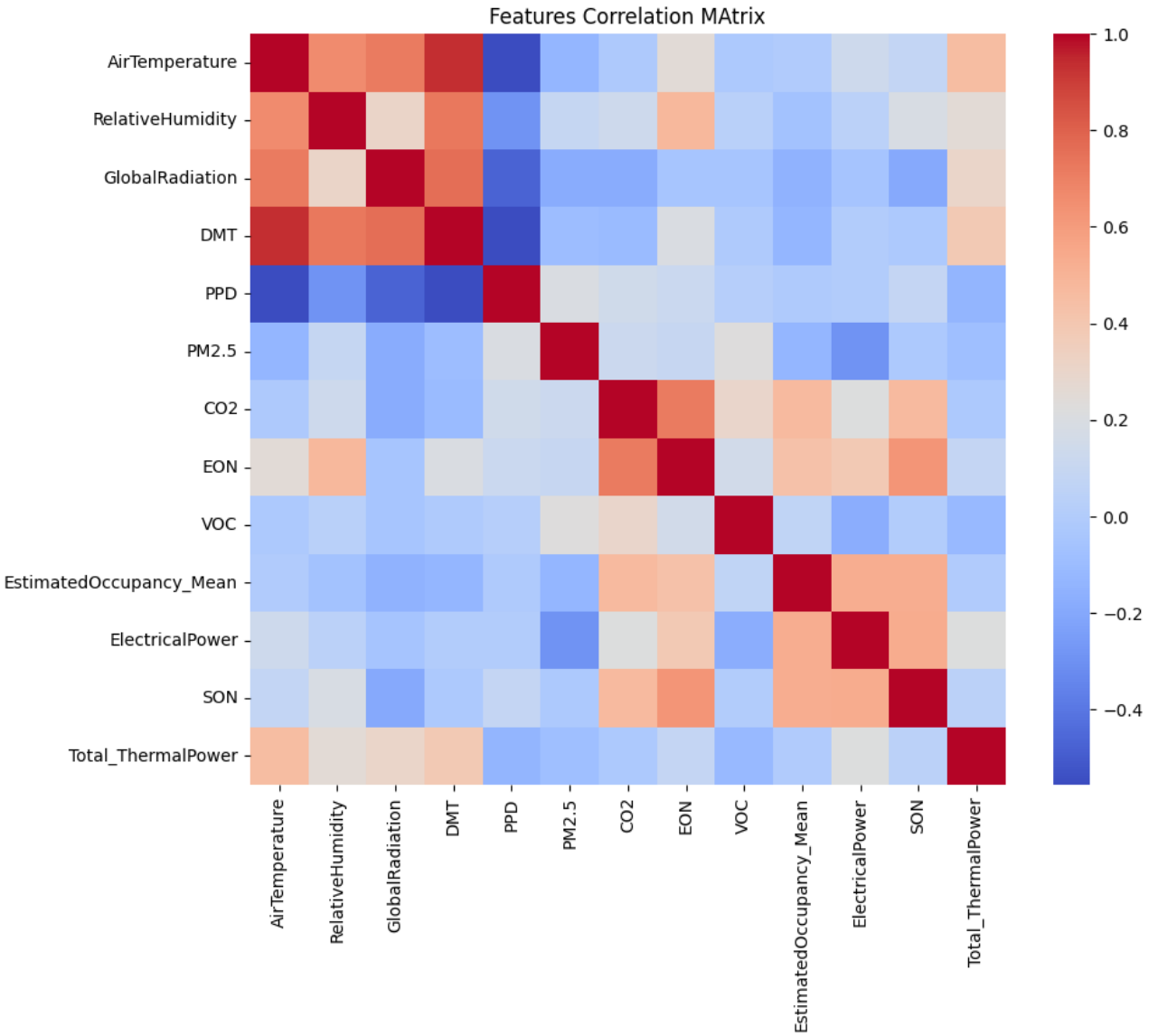

Figure 58: Correlation Matrix based on K-Means cluster of Temperature

The selected features are Air Temperature, Global Radiation, DMT (Daily Mean Temperature), Relative Humidity, PPD (People Percentage Dissatisfied), PM2.5, CO2, Total Thermal Power, Estimated Occupancy Mean, SON (Scheduled Occupancy Normalized).

4.2.1.2 Model Hypertuning

Decision trees, and especially Random Forests, have several hyperparameters that affect their performance. Finding the optimal combination of hyperparameters is crucial to obtain an accurate and generalizable model. Two optimization techniques have been studied:

- The Bayesian optimization constructs a probabilistic model of the objective function (model accuracy) and uses it to intelligently select the next hyperparameters to evaluate, balancing exploration (trying new combinations) and exploitation (leveraging combinations that have yielded good results in the past). It is particularly efficient when evaluating the objective function is expensive (in terms of time or resources).
- Grid Search method, which is more exhaustive but potentially more resources expensive, evaluates all hyperparameter combinations within a predefined grid. It is appropriate when the hyperparameter space is relatively small and the goal is to explore every possible combination.

Bayesian optimization

Bayesian optimization was used to fine-tune the hyperparameters of the Random Forest model. This process relied on the definition of a specific objective function. This function, when given a set of hyperparameters, would train a Random Forest model and subsequently return to its average accuracy. Accuracy was determined using a cross-validation technique, which is described in more detail in the following sections. The search space for the Bayesian optimization included a range of plausible values for the key Random Forest hyperparameters. These hyperparameters included the number of trees in the forest, the maximum depth of individual trees (max_depth), the minimum number of samples required to split an internal node, and the minimum number of samples required to form a leaf node.

The Bayesian optimizer was configured to perform a series of iterations. It started the process with 5 randomly selected starting points within the defined search space. This was followed by 15 iterations, guided by the probabilistic model built internally by the optimizer. During each iteration, different hyperparameter combinations were evaluated and the probabilistic model was updated based on the observed performance. At the end of all iterations, the optimizer returned the hyperparameter combination that gave the highest average accuracy across the cross-validation folds.

Grid Search (GridSearchCV)

For the grid search, the following hyperparameters were considered: The number of trees in the forest, the maximum depth of the trees, the minimum number of samples required to split an internal node, the minimum number of samples required in a leaf node, and finally the function measuring the quality of a split (either "gini" or "entropy"). GridSearchCV evaluated all combinations of values within a predefined range for each of these hyperparameters. It then returned the hyperparameter combination that produced the best average accuracy through cross-validation. In some cases, consideration was given to selecting a model with parameters within one standard deviation of the best performing model.

4.2.1.3 Crossfold validation

To reliably evaluate model performance and to guide both Bayesian optimization and grid search, cross-validation was used. This technique involves partitioning the training dataset into multiple subsets (folds) and iteratively using one-fold as the validation set and the remaining folds as the training set. A 10-fold cross-validation was used, meaning that the dataset was divided into 10 parts. For each hyperparameter combination, 10 models were trained, each using 9-fold for training and 1-fold for validation. The average accuracy across these 10 models was used as an estimate of the model's performance for that specific hyperparameter combination. Cross-validation provides a more robust estimate of model performance than a single train/test split because it reduces the variability due to random selection of training and test data. Once the optimal hyperparameters were identified (via Bayesian optimization or grid search), the

final model was trained using the entire training data set. The performance of the final model was then evaluated on the test dataset (which had not been used for either hyperparameter optimization or cross-validation), providing an unbiased estimate of its ability to generalize to new data.

4.2.2 Long Short-Term Memory (LSTM)

LSTM networks are a specialized type of Recurrent Neural Network (RNN) designed to address the vanishing gradient problem that can hinder the training of standard RNNs on long sequences. Unlike standard RNNs, which have a simple repeating module, LSTMs incorporate a complex memory cell structure with "gates" (input, forget, and output gates) that regulate the flow of information. These gates allow the LSTM to selectively remember or forget information over long periods, enabling it to capture long-range dependencies in sequential data. This makes LSTMs particularly well-suited for tasks involving time series analysis, natural language processing, and other applications where the context of past information is crucial for accurate predictions. LSTMs are trained using backpropagation through time (BPTT), a modified version of the backpropagation algorithm adapted for sequential data.

4.2.2.1 Preparation layer

Unlike decision trees, LSTMs require a specific three-dimensional time series input format:

- Hourly observations were grouped by day, resulting in a sequence of 144 measurements for each day.
- Creating a three-dimensional array where the first dimension is represented by the number of days, the second dimension represents the length of the time sequence (144 daily measurements), and finally the third dimension represents the number of predictor variables (features) used for each measurement.
- A normalization of the data has been performed by using a min-max scaler of the feature values within the range [0, 1]. This step is important to improve the stability and convergence speed of the neural network training.
- Since the target variable (cluster) is categorical, it was transformed into a one-hot representation. In this encoding, each class is represented by a binary vector, where a single element is 1 (indicating class membership) and all others are 0.

4.2.2.2 LSTM model architecture

An LSTM model has been designed with the following architecture:

- LSTM layer: A single LSTM layer of 50 units. The LSTM unit is the core of the network and contains memory cells and gates that control the flow of information. The “input_shape” argument specifies the shape of the input sequences (144, number of features). The hyperbolic tangent (tanh) activation function was used, a common choice for LSTMs.
- Dropout layer: A dropout layer with a rate of 30% was added. Dropout randomly deactivates a fraction of connections during training, helping to prevent overfitting and improve model generalization.
- Dense (Intermediate) Layer: A dense (fully connected) layer with 25 units and a Rectified Linear Unit (ReLU) activation function. This layer adds further learning capacity to the model.
- Dense Layer (Output): A final Dense layer with a number of units equal to the number of classes (clusters) and a Softmax activation function. Softmax produces a probability distribution over the classes such that the sum of the probabilities is 1.

4.2.2.3 Model training

Prior to training, the LSTM model was compiled, a step that defines the optimization algorithm, the loss function and the evaluation metrics. The optimization algorithm chosen was the Adam optimizer, a widely used and efficient stochastic gradient descent algorithm. The loss function chosen was “categorical_crossentropy”, which is well suited for multi-class classification tasks where the target variable is represented in a one-hot coded format. Finally, the performance evaluation of the model was primarily based on accuracy, which was tracked as the key metric during training.

The data set was divided into training and test sets, with 80% used for training and 20% for testing. The LSTM model was then trained using the training set. The training process was configured with a total of 100 epochs, where each epoch represents a complete iteration over the entire training dataset. A batch size of 128 was used, meaning that 128 samples were processed in each mini-batch during gradient descent optimization. Crucially, the test set was also provided as validation data during training. This allowed continuous monitoring of the model's performance on unseen data throughout the training process. By observing the model's performance on the validation set, it was possible to identify potential signs of overfitting, where the model begins to memorize the training data rather than generalizing to unseen examples.

4.3 Results

This study explored the application of machine learning techniques, specifically Random Forest (an advanced implementation of decision trees) and Long Short-Term Memory (LSTM) recurrent neural networks, to the classification of temporal patterns in multivariate environmental data. Several considerations and evaluations were made in the study. Different data preprocessing methods were investigated, and it was found that raw data without any filtering should be used, while for all other parameters, pre-clustering operations were necessary. However, it was found that the outlier removal process did not significantly affect the clustering methods and, consequently, the subsequent steps. Nevertheless, it was decided to perform it in order to obtain a slight gain in cluster regularity. The above mentioned applies to the temperature related clusters.

4.3.1 Decision Tree

Two approaches to hyperparameter optimization were evaluated for the Random Forest model: Bayesian Optimization and Grid Search (GridSearchCV). Both methods yielded comparable results in terms of accuracy, although Bayesian optimization proved to be more efficient in terms of computational time, requiring the evaluation of fewer hyperparameter combinations. In addition to the hypertuning methods, an option that considered the model with a parameter that had an accuracy value not less than the best model by a value of one standard deviation ("one standard error rule") was evaluated.

The optimal configuration determined for the temperature dataset includes the following hyperparameters

- n_estimators: Number of trees in the forest, ranging from 10 to 200.
- max_depth: Maximum depth of the trees, ranging from 1 to 8.
- min_samples_split: Minimum number of samples needed to split a node, ranging from 1000 to 3000.
- min_samples_leaf: Minimum number of samples required in a leaf node, varying between 100 and 1600.
- criterion: Function to measure the quality of a split, the choice is between 'gini' (default) and 'entropy'.

The feature selection based on the Pearson correlation with the target variable proved to be effective in reducing the dimensionality of the problem, focusing attention on the most informative variables (with a minimum correlation threshold of 0.4) and improving the interpretability of the model.

The final results (Table 13), evaluated on an independent test set, indicate an accuracy of approximately 0.967 for Grid Search and 0.954 for Bayesian Optimization for the temperature model.

Table 13: Performance Log of the Decision Tree Hyperparameters search

Deliverable	Content	Type	Link
3.1.4.1.1	Log of the hyperparameters tuning phase	Log	DigitMan_Log_DecisionTree_Hyperparameters.txt

The results of the Random Forest algorithm, generated with GridSearchCV hypertuning are shown in Figure 59. The rules are reported in Table 14.

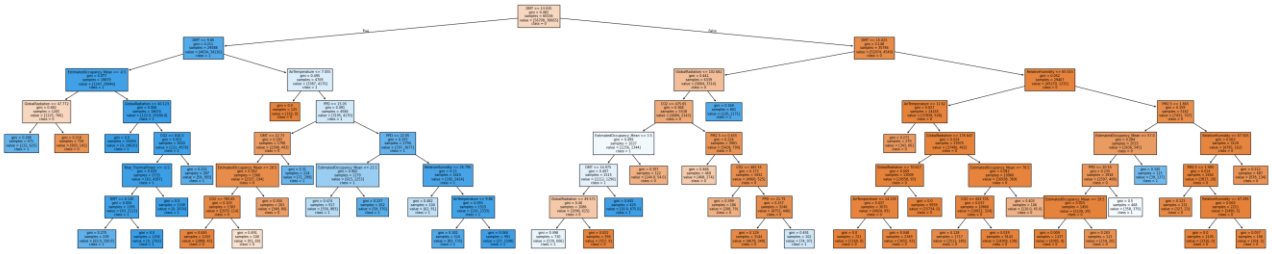

Figure 59: Graph representation of Decision Tree with GridSearchCV optimization

The results of the Random Forest algorithm, generated with Bayesian hypertuning are shown in Figure 60. The rules are reported in Table 14.

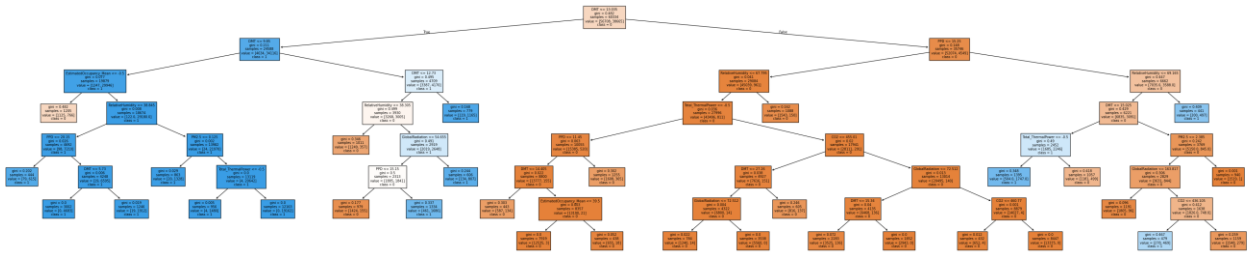

Figure 60: Graph representation of Decision Tree with Bayesian optimization

4.3.2 LSTM

The LSTM model, designed to capture long-term dependencies in time series, required specific data preparation, with hourly observations transformed into three-dimensional daily sequences. The optimal model architecture includes an LSTM layer of 50 units, followed by a dropout layer (rate 0.3), an intermediate dense layer (25 units, ReLU activation), and an output dense layer (softmax activation).

Model training, performed for 100 epochs with a batch size of 128, showed stable convergence, with a final accuracy on the test set of 0.584 on the train set and a value accuracy of 0.59 on the validation data set. The quality of the trained model can be estimated by analyzing the trend loss function over epochs, for both the training set (train loss) and the validation set (validation loss). As shown in Figure 61, the progressive and simultaneous decrease in both curves indicate learning and generalization ability. Another metric to analyze is the evolution of accuracy during training, shown in Figure 62. As can be seen, the gap between the training and validation accuracy curves is extremely narrow. While this indicates that the model is not overfitted, the limited growth in accuracy across epochs suggests underfitting. This implies that the model possesses insufficient capacity to capture the complexity of the data.

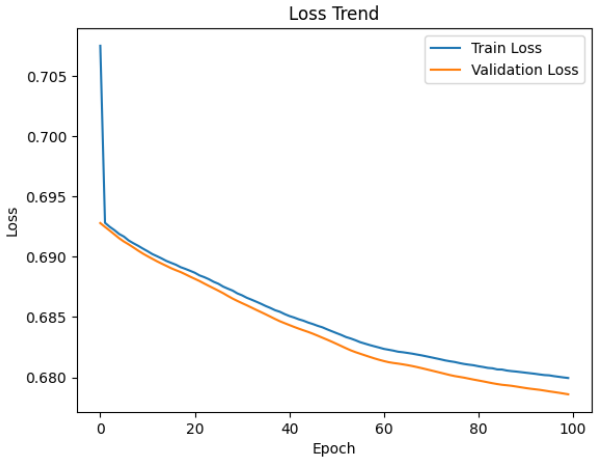

Figure 61: Loss trend graph

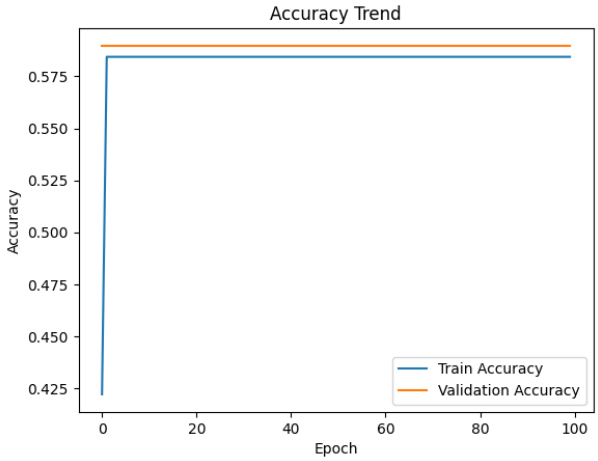

Figure 62: Accuracy trend graph

4.4 Conclusions and future perspectives

Although both models demonstrated good classification capabilities, the Random Forest model achieved higher accuracy compared to the LSTM model, especially with the Grid Search optimization. This suggests that for this particular dataset and classification task, the non-linear relationships between the features and the target are better captured by an ensemble of decision trees. The effectiveness and speed of the grid search technique for the Random Forest model is also highlighted. In contrast, the Bayesian optimization, although more sophisticated, did not lead to significant improvements in accuracy despite the reduction in computational time.

The LSTM model, although less accurate, provides a complementary approach that is able to explicitly model temporal dependencies. Future research directions might include:

- Integration of additional variables: The inclusion of exogenous variables (such as weather forecasts) could enrich the models and improve their performance.
- Comparison with other algorithms: Exploring other machine learning algorithms (e.g., support vector machines, gradient boosting) could provide additional benchmarks.
- Sensitivity analysis: A more in-depth analysis of the influence of individual features on classification could provide useful insights for interpreting the results.

In conclusion, this study demonstrates the potential of machine learning techniques, in particular Random Forest and LSTM, for the analysis and classification of complex temporal patterns in environmental data. The results obtained provide a basis for the development of decision support systems in the study focus areas. The methodology used and explained above has produced different models with different accuracies. Table 14 below summarizes the contents of the models obtained from the data collected from the case study. These models include different formats such as Joblib, H5 and Keras.

Table 14: ML models produced based on DIGITMAN methodology

Deliverable	Content	Type	Link
3.1.4.2.1 Random Forest Model - GridSearch	Random Forest rules calculated with GridSearchCV	Model	Temperature_RF_GS_Model.joblib
3.1.4.2.2	Random Forest rules calculated with Bayesian Optimization	Model	Temperature_RF_BS_Model.joblib

Deliverable	Content	Type	Link
Random Forest Model - Bayesian			
3.1.4.2.3 LSTM Model – H5 export	Long Short-Term Memory trained model. H5 legacy format	Model	Temperature_LSTM_model.h5
3.1.4.2.3 LSTM Model – Keras export	Long Short-Term Memory trained model. Keras compatibility format	Model	Temperature_LSTM_model.keras

5 References

1. AbdelAzim, A.I.; Ibrahim, A.M.; Aboul-Zahab, E.M. Development of an Energy Efficiency Rating System for Existing Buildings Using Analytic Hierarchy Process – The Case of Egypt. *Renewable and Sustainable Energy Reviews* **2017**, *71*, 414–425, doi:10.1016/J.RSER.2016.12.071.
2. Mattoni, B.; Guattari, C.; Evangelisti, L.; Bisegna, F.; Gori, P.; Asdrubali, F. Critical Review and Methodological Approach to Evaluate the Differences among International Green Building Rating Tools. *Renewable and Sustainable Energy Reviews* **2018**, *82*, 950–960, doi:10.1016/J.RSER.2017.09.105.
3. Zuo, J.; Zhao, Z.Y. Green Building Research–Current Status and Future Agenda: A Review. *Renewable and Sustainable Energy Reviews* **2014**, *30*, 271–281, doi:10.1016/J.RSER.2013.10.021.
4. Buil4People (B4P) Co-Programmed European Partnership Biennial Full Report.
5. Bortolini, R.; Forcada, N. Analysis of Building Maintenance Requests Using a Text Mining Approach: Building Services Evaluation. *Building Research & Information* **2020**, *48*, 207–217, doi:10.1080/09613218.2019.1609291.
6. Wang, K.; Guo, F.; Zhang, C.; Hao, J.; Schaefer, D. Digital Technology in Architecture, Engineering, and Construction (AEC) Industry: Research Trends and Practical Status toward Construction 4.0. *Construction Research Congress 2022: Project Management and Delivery, Controls, and Design and Materials - Selected Papers from Construction Research Congress 2022* **2022**, *3–C*, 983–992, doi:10.1061/9780784483978.100.
7. Succar, B.; Poirier, E. Lifecycle Information Transformation and Exchange for Delivering and Managing Digital and Physical Assets. *Autom Constr* **2020**, *112*, 103090, doi:10.1016/J.AUTCON.2020.103090.
8. Gómez-Gil, M.; Sesana, M.M.; Salvalai, G.; Espinosa-Fernández, A.; López-Mesa, B. The Digital Building Logbook as a Gateway Linked to Existing National Data Sources: The Cases of Spain and Italy. *Journal of Building Engineering* **2023**, *63*, 105461, doi:10.1016/J.JOBE.2022.105461.
9. Sheikh Khan, D.; Kolarik, J. Can Occupant Voting Systems Provide Energy Savings and Improved Occupant Satisfaction in Buildings?—A Review. *Sci Technol Built Environ* **2022**, *28*, 221–239, doi:10.1080/23744731.2021.1976017.
10. Jin, Y.; Yan, D.; Chong, A.; Dong, B.; An, J. Building Occupancy Forecasting: A Systematical and Critical Review. *Energy Build* **2021**, *251*, 111345, doi:10.1016/J.ENBUILD.2021.111345.
11. Lee, D.; Lee, S.H.; Masoud, N.; Krishnan, M.S.; Li, V.C. Integrated Digital Twin and Blockchain Framework to Support Accountable Information Sharing in Construction Projects. *Autom Constr* **2021**, *127*, 103688, doi:10.1016/J.AUTCON.2021.103688.
12. Ding, Y.; Han, S.; Tian, Z.; Yao, J.; Chen, W.; Zhang, Q. Review on Occupancy Detection and Prediction in Building Simulation. *Build Simul* **2022**, *15*, 333–356, doi:10.1007/S12273-021-0813-8.
13. *SCOPE OF EMISSIONS*;
14. Marshall, W. *UWE Carbon Management Plan 2013-2020*; 2017;
15. MIT Campus Greenhouse Gas Emissions Reduction Strategy | MIT Sustainability Available online: https://sustainability.mit.edu/resource/mit-campus-greenhouse-gas-emissions-reduction-strategy (accessed on 28 February 2025).

16. Mosteiro-Romero, M.; Miller, C.; Chong, A.; Stouffs, R. Elastic Buildings: Calibrated District-Scale Simulation of Occupant-Flexible Campus Operation for Hybrid Work Optimization. *Build Environ* **2023**, *237*, 110318, doi:10.1016/J.BUILDENV.2023.110318.
17. Wadud, Z.; Royston, S.; Selby, J. Modelling Energy Demand from Higher Education Institutions: A Case Study of the UK. *Appl Energy* **2019**, *233–234*, 816–826, doi:10.1016/J.APENERGY.2018.09.203.
18. Gui, X.; Gou, Z.; Lu, Y. Reducing University Energy Use beyond Energy Retrofitting: The Academic Calendar Impacts. *Energy Build* **2021**, *231*, 110647, doi:10.1016/J.ENBUILD.2020.110647.
19. Mosteiro-Romero, M.; Miller, C.; Chong, A.; Stouffs, R. Elastic Buildings: Calibrated District-Scale Simulation of Occupant-Flexible Campus Operation for Hybrid Work Optimization. *Build Environ* **2023**, *237*, 110318, doi:10.1016/J.BUILDENV.2023.110318.
20. Mêda, P.; Calvetti, D.; Hjelseth, E.; Sousa, H.; Aigbavboa, C.; Ejohwomu, O.; Roberts, C. Incremental Digital Twin Conceptualisations Targeting Data-Driven Circular Construction. *Buildings 2021, Vol. 11, Page 554* **2021**, *11*, 554, doi:10.3390/BUILDINGS11110554.
21. Resources: The Roadmap | Centre for Digital Built Britain Completed Its Five-Year Mission and Closed Its Doors at the End of September 2022 Available online: https://www.cdbb.cam.ac.uk/DFTG/DFTGRoadmap (accessed on 28 February 2025).
22. Wilde, P. de Building Performance Analysis Pieter de Wilde. **2018**.
23. Szukits, Á.; Móricz, P.; Móricz petermoricz, P.; Szukits, Á.; Móricz, P. Towards Data-Driven Decision Making: The Role of Analytical Culture and Centralization Efforts. *Review of Managerial Science 2023 18:10* **2023**, *18*, 2849–2887, doi:10.1007/S11846-023-00694-1.
24. Augenbroe, G.; Verheij, H.; SCHWARZMÜLLER, G. Project Web Sites with Design Management Extensions. *Engineering, Construction and Architectural Management* **2002**, *9*, 259–271, doi:10.1108/EB021221/FULL/PDF.
25. Oh, S.; Haberl, J.S. Origins of Analysis Methods Used to Design High-Performance Commercial Buildings: Whole-Building Energy Simulation. *Sci Technol Built Environ* **2016**, *22*, 118–137, doi:10.1080/23744731.2015.1063958.
26. de Wilde, P. Building Performance Simulation in the Brave New World of Artificial Intelligence and Digital Twins: A Systematic Review. *Energy Build* **2023**, *292*, 113171, doi:10.1016/J.ENBUILD.2023.113171.
27. Yu, J.; Chang, W.S.; Dong, Y. Building Energy Prediction Models and Related Uncertainties: A Review. *Buildings 2022, Vol. 12, Page 1284* **2022**, *12*, 1284, doi:10.3390/BUILDINGS12081284.
28. Wong, I.L. A Review of Daylighting Design and Implementation in Buildings. *Renewable and Sustainable Energy Reviews* **2017**, *74*, 959–968, doi:10.1016/J.RSER.2017.03.061.
29. Zhao, Q.; Lian, Z.; Lai, D. Thermal Comfort Models and Their Developments: A Review. *Energy and Built Environment* **2021**, *2*, 21–33, doi:10.1016/J.ENBENV.2020.05.007.
30. Aflaki, A.; Esfandiari, M.; Mohammadi, S. A Review of Numerical Simulation as a Precedence Method for Prediction and Evaluation of Building Ventilation Performance. *Sustainability 2021, Vol. 13, Page 12721* **2021**, *13*, 12721, doi:10.3390/SU132212721.
31. Abdalla, T.; Peng, C. Evaluation of Housing Stock Indoor Air Quality Models: A Review of Data Requirements and Model Performance. *Journal of Building Engineering* **2021**, *43*, 102846, doi:10.1016/J.JOBE.2021.102846.
32. Prinn, A.G. A Review of Finite Element Methods for Room Acoustics. *Acoustics 2023, Vol. 5, Pages 367-395* **2023**, *5*, 367–395, doi:10.3390/ACOUSTICS5020022.

33. Kasereka, S.; Kasoro, N.; Kyamakya, K.; Doungmo Goufo, E.F.; Chokki, A.P.; Yengo, M. V. Agent-Based Modelling and Simulation for Evacuation of People from a Building in Case of Fire. *Procedia Comput Sci* **2018**, *130*, 10–17, doi:10.1016/J.PROCS.2018.04.006.
34. Dabirian, S.; Panchabikesan, K.; Eicker, U. Occupant-Centric Urban Building Energy Modeling: Approaches, Inputs, and Data Sources - A Review. *Energy Build* **2022**, *257*, 111809, doi:10.1016/J.ENBUILD.2021.111809.
35. De Wilde, P. The Gap between Predicted and Measured Energy Performance of Buildings: A Framework for Investigation. *Autom Constr* **2014**, *41*, 40–49, doi:10.1016/J.AUTCON.2014.02.009.
36. Chong, A.; Gu, Y.; Jia, H. Calibrating Building Energy Simulation Models: A Review of the Basics to Guide Future Work. *Energy Build* **2021**, *253*, 111533, doi:10.1016/J.ENBUILD.2021.111533.
37. Menezes, A.C.; Cripps, A.; Bouchlaghem, D.; Buswell, R. Predicted vs. Actual Energy Performance of Non-Domestic Buildings: Using Post-Occupancy Evaluation Data to Reduce the Performance Gap. *Appl Energy* **2012**, *97*, 355–364, doi:10.1016/J.APENERGY.2011.11.075.
38. Energy Performance – BuildingSMART Italia Available online: https://www.buildingsmartitalia.org/utenti/pubblicazioni/energy-performance/ (accessed on 28 February 2025).
39. Chen, W.; Chen, K.; Cheng, J.C.P. Towards an Ontology-Based Approach for Information Interoperability between BIM and Facility Management. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2018**, *10864 LNCS*, 447–469, doi:10.1007/978-3-319-91638-5_25/FIGURES/11.
40. Chen, S.; Jin, R.; Alam, M. Investigation of Interoperability between Building Information Modelling (BIM) and Building Energy Simulation (BES). *International Review of Applied Sciences and Engineering* **2018**, *9*, 137–144, doi:10.1556/1848.2018.9.2.9.
41. Bonomolo, M.; Di Lisi, S.; Leone, G. Building Information Modelling and Energy Simulation for Architecture Design. *Applied Sciences 2021, Vol. 11, Page 2252* **2021**, *11*, 2252, doi:10.3390/APP11052252.
42. Panagiotidou, V.; Körner, A. From Intricate to Coarse and Back A Voxel-Based Workflow to Approximate High-Res Geometries for Digital Environmental Simulations. *Proceedings of the International Conference on Education and Research in Computer Aided Architectural Design in Europe* **2022**, *1*, 491–500, doi:10.52842/CONF.ECAADE.2022.1.491.
43. Costa, G.; Sicilia, Á. Web Technologies for Sensor and Energy Data Models. *Buildings and Semantics* **2022**, 51–68, doi:10.1201/9781003204381-4.
44. Porsani, G.B.; de Lersundi, K.D.V.; Gutiérrez, A.S.O.; Bandera, C.F. Interoperability between Building Information Modelling (BIM) and Building Energy Model (BEM). *Applied Sciences 2021, Vol. 11, Page 2167* **2021**, *11*, 2167, doi:10.3390/APP11052167.
45. Collao, J.; Lozano-Galant, F.; Lozano-Galant, J.A.; Turmo, J. BIM Visual Programming Tools Applications in Infrastructure Projects: A State-of-the-Art Review. *Applied Sciences 2021, Vol. 11, Page 8343* **2021**, *11*, 8343, doi:10.3390/APP11188343.
46. Gartners Top 10 Technology Trends 2017 Available online: https://www.gartner.com/smarterwithgartner/gartners-top-10-technology-trends-2017 (accessed on 28 February 2025).
47. Yang, S.; Wan, M.P.; Chen, W.; Ng, B.F.; Dubey, S. Model Predictive Control with Adaptive Machine-Learning-Based Model for Building Energy Efficiency and Comfort Optimization. *Appl Energy* **2020**, *271*, 115147, doi:10.1016/J.APENERGY.2020.115147.

48. Villano, F.; Mauro, G.M.; Pedace, A. A Review on Machine/Deep Learning Techniques Applied to Building Energy Simulation, Optimization and Management. *Thermo 2024*, *Vol. 4*, *Pages 100-139* **2024**, *4*, 100–139, doi:10.3390/THERMO4010008.
49. Capozzoli, A.; Grassi, D.; Causone, F. Estimation Models of Heating Energy Consumption in Schools for Local Authorities Planning. *Energy Build* **2015**, *105*, 302–313, doi:10.1016/j.enbuild.2015.07.024.
50. Beccali, M.; Ciulla, G.; Lo Brano, V.; Galatioto, A.; Bonomolo, M. Artificial Neural Network Decision Support Tool for Assessment of the Energy Performance and the Refurbishment Actions for the Non-Residential Building Stock in Southern Italy. *Energy* **2017**, *137*, 1201–1218, doi:10.1016/j.energy.2017.05.200.
51. Nutkiewicz, A.; Yang, Z.; Jain, R.K. Data-Driven Urban Energy Simulation (DUE-S): A Framework for Integrating Engineering Simulation and Machine Learning Methods in a Multi-Scale Urban Energy Modeling Workflow. *Appl Energy* **2018**, *225*, 1176–1189, doi:10.1016/j.apenergy.2018.05.023.
52. Jovanović, R.; Sretenović, A.A.; Živković, B.D. Ensemble of Various Neural Networks for Prediction of Heating Energy Consumption. *Energy Build* **2015**, *94*, 189–199, doi:10.1016/j.enbuild.2015.02.052.
53. 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, doi:10.1016/j.energy.2017.05.191.
54. Yang, J.; Ning, C.; Deb, C.; Zhang, F.; Cheong, D.; Lee, S.E.; Sekhar, C.; Tham, K.W. K-Shape Clustering Algorithm for Building Energy Usage Patterns Analysis and Forecasting Model Accuracy Improvement. *Energy Build* **2017**, *146*, 27–37, doi:10.1016/j.enbuild.2017.03.071.
55. Robinson, C.; Dilkina, B.; Hubbs, J.; Zhang, W.; Guhathakurta, S.; Brown, M.A.; Pendyala, R.M. Machine Learning Approaches for Estimating Commercial Building Energy Consumption. *Appl Energy* **2017**, *208*, 889–904, doi:10.1016/j.apenergy.2017.09.060.
56. Intro - Meteonorm (En) Available online: https://meteonorm.com/en/ (accessed on 25 February 2025).
57. Measures Converter Available online: https://www.snam.it/en/storage/tools/converter.html (accessed on 28 February 2025).
58. Conversione, Fattori Di - ENEA - Dipartimento Unità per l'efficienza Energetica Available online: https://www.efficienzaenergetica.enea.it/glossario-efficienza-energetica/lettera-c/conversione-fattori-di.html (accessed on 28 February 2025).
59. Arera: Valore CMEMm - Servizio Di Tutela Della Vulnerabilità Available online: https://www.arera.it/area-operatori/prezzi-e-tariffe/valore-cmemm-vulnerabili (accessed on 28 February 2025).
60. Arera: Valori Della Materia Energia per Il Servizio a Tutele Graduali Available online: https://www.arera.it/consumatori/valori-della-materia-energia-per-il-servizio-a-tutele-graduali (accessed on 28 February 2025).
61. The Covenant of Mayors for Climate and Energy Reporting Guidelines - Publications Office of the EU Available online: https://op.europa.eu/en/publication-detail/-/publication/ac865f28-dedb-11e6-ad7c-01aa75ed71a1 (accessed on 28 February 2025).
62. Tartarini, F.; Schiavon, S. Pythermalcomfort: A Python Package for Thermal Comfort Research. *SoftwareX* **2020**, *12*, doi:10.1016/J.SOFTX.2020.100578.
63. Franco, A.; Leccese, F. Measurement of CO2 Concentration for Occupancy Estimation in Educational Buildings with Energy Efficiency Purposes. *Journal of Building Engineering* **2020**, *32*, 101714, doi:10.1016/J.JOBE.2020.101714.

64. Standard 55 – Thermal Environmental Conditions for Human Occupancy Available online: https://www.ashrae.org/technical-resources/bookstore/standard-55-thermal-environmental-conditions-for-human-occupancy (accessed on 25 February 2025).