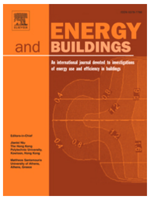
Energy & Buildings 336 (2025) 115633

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



A data-driven building thermal zoning algorithm for digital twin-enabled advanced control

Lina Morkunaite[a,*], Adil Rasheed[b], Darius Pupeikis[a], Vangelis Angelakis[c], Tobias Davidsson[d]

[a] Faculty of Civil Engineering and Architecture, Kaunas University of Technology, Lithuania
[b] Department of Engineering Cybernetics, Norwegian University of Science and Technology, Norway
[c] Department of Science and Technology, Linköping University, Sweden
[d] Technology Development, Akademiska Hus, AB, Sweden

ARTICLE INFO

Keywords:
HVAC zone
Energy flexibility
Dynamic thermal model

ABSTRACT

Effective control of indoor environments is crucial for maintaining occupant comfort and optimizing energy use. However, current building control strategies often fail to achieve these goals, as they rely on static or rule-based approaches that normally do not account for dynamic conditions. While advanced control strategies offer a more adaptive solution, their implementation is challenging due to the need for accurate thermal models, which are resource-intensive to develop. Defining building thermal zones can help to strike a balance between model accuracy and the cost of their development and implementation. However, data-driven approaches for identifying thermal zones remain scarce. This study addresses these gaps by proposing a reusable data-driven thermal zoning algorithm that employs Principal Component Analysis (PCA) and k-means clustering to define building thermal zones. This method allows for the inclusion of numerous parameters, thus increasing the applicability and consistency of the zoning process. Additionally, we propose an algorithm for zones validation, supported by qualitative criteria from literature and standards. The approach is tested in a large educational building, using time-series data from 168 rooms with a total of 262 CO2 and temperature sensors. Results show that the proposed zoning algorithm achieves over 91 % consistency score, depending on the number of selected principal components, clusters, and input parameters available. The derived thermal zones are further validated based on the synthesised qualitative criteria. Finally, the results are visualized in a DT environment, where users can explore color-coded thermal zones alongside real-time sensor data, 3D building geometry, and semantic information.

1. Introduction

Building operations account for 30 % of global energy consumption, contributing to 26 % of global CO2 emissions [1]. Following "The European Green deal" strategy [2] the built environment is under pressure to adapt to the current energy market dynamics and integrate more flexible and responsive energy management systems [3]. Energy prices are inherently fluctuating, and an examination of the electricity price in central Sweden on 21st February 2025 showed a decrease of 86 % between the highest and lowest price during the same day [4]. Furthermore, with the increasing penetration of wind energy and its inherent volatile nature, which can cause fluctuations in the cost of electricity [5], the ability to optimise the timing of electricity usage becomes even more valuable.

Considering the life cycle of a building, the vast majority of energy consumption is attributed to the operational stage [6]. Equipping buildings with sensors and actuators that allow automatic adjustment of heating, cooling, and other systems can bring the needed energy flexibility while maintaining thermal comfort [7]. This approach requires real-time monitoring, analysis, and actuation, leveraging digital tools to measure and calculate when and how to act [8]. The steady decline in sensor and computing power costs over the past few decades, combined with recent advances in machine learning, has accelerated this approach, enabling timely and efficient responses to dynamic conditions [9].

Optimisation of energy usage requires the development of advanced control systems, with accurate and reliable forecasting models providing a basis for control algorithms [10]. In contrast to the currently popular rule-based control (RBC) which relies on predefined rules to maintain

[*] Corresponding author.
E-mail addresses: lina.morkunaite@ktu.lt (L. Morkunaite), adil.rasheed@ntnu.no (A. Rasheed), darius.pupeikis@ktu.lt (D. Pupeikis), vangelis.angelakis@liu.se (V. Angelakis), tobias.davidsson@akademiskahus.se (T. Davidsson).

https://doi.org/10.1016/j.enbuild.2025.115633
Received 5 July 2024; Received in revised form 13 March 2025; Accepted 17 March 2025
Available online 19 March 2025

space comfort set points [11], more advanced control strategies can enhance system efficiency and reduce operational costs [12,13]. Such models can forecast the impact of different control inputs, allowing the control system to make informed decisions [14–16].

Digital twin (DT) is emerging as a promising technology to achieve all the above-mentioned features in building performance monitoring, optimisation and control [17]. A DT is defined as a virtual replica of a physical asset, enabled through data and simulations, that can be used for real-time monitoring, optimisation, and decision making [18]. At a basic level, a DT can function as a data aggregation tool, enabling users to access and interact with data in a holistic and centralized manner. Among other benefits, a building's DT can facilitate data-driven analysis and predictive modeling, enabling more informed decision-making for energy management, occupant comfort, and operational efficiency [19].

Currently, there are three main techniques applied for building dynamic thermal modelling, namely, physics-based (white-box), data-driven (black-box) and hybrid (grey-box) models [20]. Each of these techniques starts with defining the boundary conditions, where determining the thermal zones of the building is essential. Atam and Helsen [21] performed an in-depth analysis of the 3 modelling techniques and related approaches for thermal zoning. However, to this day, there is still no clear agreement on how to define such zones [22].

There are several different strategies for building thermal zoning. For physics-based models, it is common practice to set each room as a separate zone while performing simulations with well-known open-source or commercial software such as Energy+, TRNSYS, IES or IDA-ICE [23]. However, defining each room as a separate thermal zone is impractical for most HVAC systems and can be costly to implement [22]. It also adds additional complexity while defining dynamic thermal models. For example, in grey-box modeling approaches, using multiple zones introduces parameter estimation complexity, making calibration and implementation challenging [24]. Another approach for thermal zoning is based on standards such as ASHRAE (American Society of Heating, Refrigerating and Air Conditioning Engineers), that recommends to separate interior and parameter spaces; separate orientations with significant amount of glazing; and separate top bottom and middle floors [25]. The authors in [26] developed a tool that automatically defines zones based on these criteria. While this approach can be useful in the early design stage, it often lacks precision in the operational stage, as it does not incorporate real data, potentially leading to suboptimal zoning configurations [27].

A new automatic thermal zoning method for commercial buildings energy simulation was recently proposed by M. Shin and J. S. Harberl [28]. The method includes identification of similarly performing grid units in terms of building's heating/cooling requirements using the linear correlation coefficient. The results showed that the proposed thermal zoning method can reduce the heating/cooling loads of the case study building by 11 %–27 % compared to a single zone model. However, this method is mostly applicable in building's early design phase where the final layout of the spaces and HVAC equipment is still to be determined. Furthermore, it requires additional data of the building geometry and includes intermediate simulation steps via energy modelling software.

Defining thermal zones while building dynamic thermal models is crucial for balancing development costs with their benefits [29]. Currently, a number of scientific works are present exploring different modelling approaches for advanced control [30]. However, to our knowledge, none propose a reusable performance-data-based algorithm for building thermal zoning. This is also identified as a gap in [22], revealing the lack of quantitative methods for selecting thermal zones. Therefore, in this study, we propose a data-driven algorithm based on principal component analysis (PCA) and k-means clustering to determine similarly performing rooms in the building and cluster them into distinct zones.

PCA is a widely known dimensionality reduction technique that has been used for many applications in various fields, including building energy analysis, with its use in predicting building energy consumption dating back to 1993 [31]. The most common application closely related to building thermal zoning is parameter dimensionality reduction for building energy modelling [32–34]. PCA applies very well to building thermal zoning as well, allowing one to include as many zone thermal performance defining parameters as needed. In addition, not only the mean but also the standard deviation (STD) values are used to define each of the input parameters to account for the possible high variation in the data. The extracted dominant principal components (PCs) are then used to determine the thermal zones by another widely adopted k-means algorithm [35]. K-means algorithm was also successfully implemented for the thermal zoning of buildings at the city level [36]. The analysis performed on data collected in 274 cities allowed the development of a new thermal design zoning scheme for buildings in China.

The thermal zoning algorithm developed in this study has been implemented in a complex educational building on the Linköping University campus in Sweden. This case study utilized CO2 and temperature time-series data collected from sensors in 168 rooms. Although this study focusses on these specific data sets, the algorithm is versatile and can be applied to any time-series data relevant to defining building thermal zones. This method is particularly advantageous for complex buildings where multiple parameters can affect the thermal performance of a zone or when qualitative data, such as room function, occupancy schedules, or room locations, are not available.

In addition, a two-fold validation is applied. Firstly, a statistical validation algorithm is employed that splits the data set into two parts and reapplies PCA and k-means clustering to compare the resulting clusters with the original ones. Secondly, the clustered zones are validated using a qualitative method comprised of previous scientific works and available standards [22,25,37]. Finally, the defined thermal zones are visualized in a DT environment, where the user can see the color-coded thermal zones and access real-time sensor data, 3D building geometry, and semantic data transferred from the IFC file.

This article is structured as follows. Section 2 presents the essential theory for the concepts and algorithms used in this work. Section 3 provides information on the case study, collected data and its preprocessing. It also provides all the information required to reproduce the results presented in this article, including the thermal zoning algorithm, the quantitative validation algorithm, the synthesised qualitative thermal zoning criteria and integrated DT architecture. Section 4 explores the results obtained from the analysis and presents the developed DT. Finally, Section 5 concludes the work and offers recommendations for areas of future research.

2. Essential concepts and definitions

2.1. Digital twins

Multiple authors have proven the benefits of DTs to optimise and control HVAC systems [3,39,40]. A DT is a virtual representation of a physical asset that integrates real-time sensor data with simulation models to enhance monitoring, analysis, and decision-making. The authors in [38] present a DT capability level scale adapted from a DNV GL report [41] that divides a DT into six distinct levels. These levels are 0-Standalone, 1-Descriptive, 2-Diagnostic, 3-Predictive, 4-Prescriptive and 5-Autonomous (Fig. 1). The standalone DT can exist even before the asset is built and can consist of only 3D models with accompanying semantic data. When the asset is in place and is equipped with sensors, data can be streamed in real-time to create a descriptive DT, giving more insight into the state of the asset. When analytic tools are applied to the incoming data stream to diagnose anomalies, the DT advances to a diagnostic level. At the first three levels, the DT can provide information/insight only about the past and present. However, a predictive DT can describe the future state of the asset. Using a predictive DT, one can perform scenario analysis to provide recommendations to push the asset to the desired state. This is then referred to as the prescriptive level. Lastly, the asset updates the DT at the autonomous level, and the DT

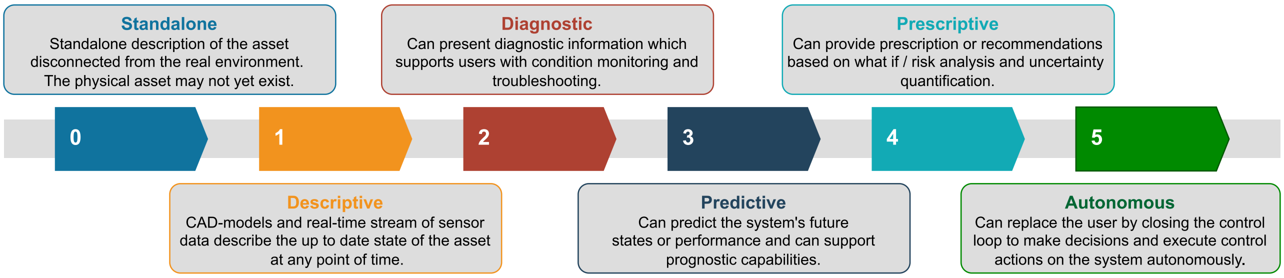

Standalone
Standalone description of the asset disconnected from the real environment. The physical asset may not yet exist.
Diagnostic
Can present diagnostic information which supports users with condition monitoring and troubleshooting.
Prescriptive
Can provide prescription or recommendations based on what if / risk analysis and uncertainty quantification.
0
1
2
3
4
5
Descriptive
CAD-models and real-time stream of sensor data describe the up to date state of the asset at any point of time.
Predictive
Can predict the system's future states or performance and can support prognostic capabilities.
Autonomous
Can replace the user by closing the control loop to make decisions and execute control actions on the system autonomously.

Fig. 1. Description of capability levels of a digital twin. Adapted from [38].

controls the asset autonomously. Achieving Capability Level 5 in a Digital Twin (DT) requires fully autonomous control, making it the most advanced stage of DT development. Thermal zoning can aid in reaching this level by streamlining the control process and optimizing resource allocation in terms of modelling and implementation [42].

2.2. Building's thermal zone

Buildings thermal zones has several interchangeably used terms such as thermal zone, HVAC zone, thermal block, etc. One of the commonly used definitions is proposed by ASHRAE Standard 90.1-2022 [25] that defines an HVAC zone as *a space or group of spaces within a building with heating and cooling requirements that are suciently similar so that desired conditions (e.g. temperature) can be maintained throughout using a single sensor (e.g. thermostat or temperature sensor)*. As this is the most comprehensive definition, it is as well adapted for this research. In fact, one of the most important criteria in defining thermal zones is the possibility of applying tailored control mechanisms to each of them, which in turn requires the installation of dedicated software and hardware. As HVAC systems in buildings are commonly complex, this poses an important boundary condition while developing dynamic thermal models [43].

Several authors have investigated the impact of thermal zoning on different control applications. Authors in [44] explored the impact of thermal zoning on heat energy consumption in typical Canadian housing. The results showed up to 36 % in energy savings while using a zonal control system. The authors also discussed the importance of the ability of control systems to learn from a building's thermal behaviour and adjust the HVAC operations accordingly. Furthermore, the study in [45] demonstrated that dynamically adjusting thermal zones based on occupancy can lead to additional energy savings.

Reynolds et al. [46] developed zone-level artificial neural networks to optimize heating set-point schedules for each zone in a small office building. Using a genetic algorithm, the approach achieved a 25 % reduction in energy consumption over a test week and further reduced energy costs by 27 % by dynamically adjusting heating schedules based on price fluctuations.

Nevertheless, there is still a great deal of uncertainty involved in defining the thermal zones of the building, mainly due to the constantly changing dynamic parameters related to environmental conditions and the influence of the user [47]. Outdoor weather parameters that are highly dependent on seasonality can cause a variation of thermal zones in a building [48]. Therefore, in some cases, it is worth considering defining seasonal thermal zones. Currently, in the vast majority of applications, thermal zones of buildings are defined based on qualitative methods. ASHRAE 90.1-2022 [25], IBPSA (International Building Performance Simulation Association) [49], CIBSE (Chartered Institution of Building Services Engineers) [50], and the well-known scientific literature [51–53] identify similar criteria such as solar irradiation, orientation, envelope exposure, occupancy, schedules, HVAC distribution type, function, etc.

Further in this study (Section 3.5), we synthesize these sources to identify the six most commonly used qualitative criteria for validating data-driven zones.

2.3. Principal component analyis

PCA is a dimensionality reduction technique used to transform n-dimensional data of X containing features $x_1, x_2, \dots, x_n$ into r-dimensional data of Z containing features $z_1, z_2, \dots, z_r$ with a new coordinate system in which the variance of the data along the axes (principal components) is maximised. This is achieved through several steps. First, the data are standardised by subtracting the mean and dividing by the standard deviation for each feature.

$$X_c = \frac{X - \mu}{\sigma} \tag{1}$$

where μ and σ are the mean and standard deviation of each of the individual features. Next, this matrix X_c is decomposed using the singular value decomposition [54] as follows:

$$X_c = USV^T \tag{2}$$

where U and V^T are called the left and right singular vectors, respectively. S is a diagonal matrix with the diagonal elements equal to the root of positive eigenvalues of $X_c X_c^T$ and $X_c^T X_c$. After this process only the first r vectors are retained and the reconstructed matrix is computed using the expression

$$\hat{X} = U_r S_r V_r^T \tag{3}$$

2.4. K-mean clustering

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a data set into K distinct, non-overlapping subsets (clusters). The goal of the algorithm is to minimise the variance within each cluster. Here are the key steps of the K-means algorithm:

- Initialisation: Randomly select K data points as initial centroids.
- Assignment: Assign each data point to the nearest centroid, forming K clusters.
- Update Centroids: Recalculate the centroids of each cluster taking the mean of all data points assigned to that cluster.
- Repeat: Repeat steps 2 and 3 until convergence (when the centroids do not change significantly or a predefined number of iterations is reached).

The objective function to minimise in K-means is the sum of squared distances between the data points and their assigned centroids. Let C_i be the set of data points assigned to the centroid i, and μ_i be the centroid of the cluster i. Then the objective function J is given by:

$$J = \sum_{i=1}^{K} \sum_{j=1}^{|C_i|} \|x_j - \mu_i\|^2 \tag{4}$$

Now, to determine the optimal value of K, the elbow method is commonly used. The idea is to run the K-means algorithm for a range of values of K and plot the within-cluster sum of squares (WCSS) for each K. WCSS is the sum of squared distances between each point and its assigned centroid (a.k.a. inertia). The elbow plot will show a decreasing trend in WCSS as K increases, but at some point the rate of decrease slows down, creating an “elbow" in the plot. The optimal value of K is often chosen at this elbow point.

In most cases, the “elbow” point is determined by visual analysis. However, in cases where visually the “elbow” is not significant, a mathematical method can be used. This can be done first by plotting the inertia vs. the number of clusters. Then, calculating the first derivative of the curve (Slope) as shown in Eq. (5), where I is the inertia and K is the number of clusters.

$$Slope = \frac{dI}{dK} \tag{5}$$

Subsequently, the second derivative (Curvature) is calculated (Eq. (6)). The second derivative allows us to detect where the slope of the inertia curve changes most dramatically, which corresponds to the elbow point. The maximum value of the second derivative indicates the “elbow” point. For more information on the “elbow” method, readers are referred to this book [55].

$$Curvature = \frac{d^2 I}{dK^2} \tag{6}$$

3. Method and setup

3.1. Data description

3.1.1. Case study building introduction

The case study building is part of Linköping University (LiU) campus in Linköping, Sweden. The building was constructed as a student coworking space ’Studenthuset’ with 1000 new working spaces in 2019. Studenthuset is an 8-storey building that includes the basement where all the main building systems equipment is located, six main floors with coworking/meeting spaces, kitchenettes, library, canteen and chill spaces for students, and the upper floor dedicated to administration offices. The façade of the building is combined from wooden panels with a U value of $0.13W/(m^2K)$ and windows with a U value of $0.8W/(m^2K)$. The structural elements are reinforced concrete insulated elements with a U value of $0.13W/(m^2K)$ for the outer walls, $0.2W/(m^2K)$ for the basement wall, and $0.1W/(m^2K)$ for the roof. The building has its digital representation 3D BIM model enriched with semantic information of the building elements, the models are available in an open IFC format as well as proprietary Audotesk RVT and DWG formats. The architectural, structural, electrical, and HVAC disciplines of the building are present and can be viewed as a complete model or as separate parts (Fig. 2). The building is connected to district heating (DH), the city’s electricity grid, water supply, and wastewater treatment systems. Electricity is also generated by installed PV solar panels that are expected to produce $9kWh/m^2$ of the useful area of the building per year.

The building is equipped with 3 air handling units (AHU) that are placed on the roof and 1 air circulation unit in the basement. Two of the air handling units placed on the roof with an airflow rate of $8m^3/s$ are responsible for the ventilation needs throughout the entire building, except the kitchen located on the second floor. The third AHU on the roof with an airflow rate of $2.8m^3/s$ ensures the required air quality in the kitchen. The air circulation unit with an airflow rate of $0.43m^3/s$ located in the basement is connected to the main 2 AHUs on the roof and controls air quality in the sensitive space for the book archive. The building has implemented variable air volume (VAV) control. VAV dampers are controlled on the basis of CO2 sensors with maximum set point values ranging between 600–900 ppm. When the set point is reached, the damper is fully opened, and all air is exchanged in the space. The projected supply air temperature is 16° C (the kitchen 18° C), while the comfort set point of the indoor air temperature ranges between 21° - 23°C. The building’s heating system is combined from hydronic based floor heating on the second floor and radiators on all other floors. The heating system is controlled by BMS-controlled valves that are activated on the basis of the room temperature set point. The heating system is also responsible for maintaining the required supply air temperature in the AHUs.

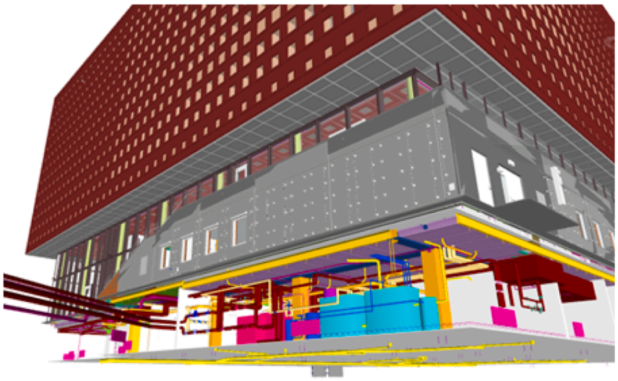


Fig. 2. Case study building services room and DH connection.

3.1.2. Input data set

Historical time-series data for indoor CO2 levels and indoor air temperature in different rooms of the case study building are used as input. CO2 levels are available for 29 rooms in the building, where in some larger spaces such as the library there is more than 1 sensor, resulting in 44 total CO2 sensors in the building. Similarly, data are available from 218 indoor air temperature sensors, deployed in a total of 165 rooms. Time period for the data obtained - 1 year and 4 months (2022/01/01 2023/05/05), with 10 minute time stamps measured in ppm for CO2 and degrees Celsius for indoor air temperature.

3.2. Data preparation

From the available CO2 and indoor air temperature data, 3 data sets were combined. This was necessary as there are more than five times more sensor data available for indoor air temperature than CO2. Furthermore, not for all rooms where CO2 data are available, a temperature reading is available, and vice versa. To account for this, a data set was prepared only for rooms where both CO2 and indoor air temperature are present, resulting in a total of 35 sensors included for each of the parameters in the data set. The second and third data sets separately included the sensor readings of CO2 (44 total sensors) and indoor air temperature (218 total sensors), including all available data for distinct parameters.

3.3. Proposed algorithm for thermal zoning

Fig. 3 shows the algorithm developed for the thermal zoning of buildings. The algorithm is divided into 3 main stages including data preparation, principal component analysis, and clustering. The process employs available historical time series data that are relevant to define the thermal performance of a zone. Before applying PCA, the input data undergoes common data preparation procedures, including data cleaning and structuring. In this stage, the mean and STD values are extracted for separate months of a year for each parameter and combined to a single data set that is further used in PCA.

In the PCA stage, the data are first standardised based on the Eq. (1). Subsequently, PCA is conducted (Eqs. (2) and (3)) and the resulting principal components (PCs) are displayed in a bar plot that illustrates both individual and cumulative explained variances. This visualisation facilitates the determination of the optimal number of PCs to include in further analysis. It is commonly accepted that the cumulative explained variance of the chosen PCs should be at least 80 % [56]. However, this number highly depends on the criticality of the data accuracy. Keeping the cumulative variance of the PCs between 90–95 % allows one to preserve the most significant information in the data set while eliminating noise or less meaningful variation.

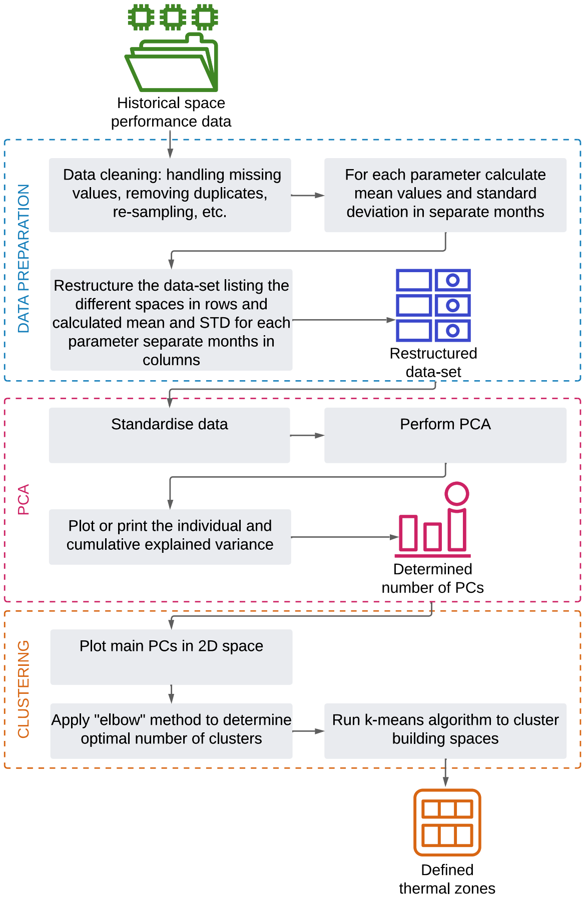

Historical space performance data
DATA PREPARATION
Data cleaning: handling missing values, removing duplicates, re-sampling, etc.
For each parameter calculate mean values and standard deviation in separate months
Restructure the data-set listing the different spaces in rows and calculated mean and STD for each parameter separate months in columns
Restructured data-set
PCA
Standardise data
Perform PCA
Plot or print the individual and cumulative explained variance
Determined number of PCs
CLUSTERING
Plot main PCs in 2D space
Apply "elbow" method to determine optimal number of clusters
Run k-means algorithm to cluster building spaces
Defined thermal zones

Fig. 3. Data-driven building's thermal zoning algorithm.

In the clustering stage, the PCA results are shown in a 2D plot and the k-means clustering algorithm is used to identify the thermal zones of the building. The optimal number of clusters in K-means clustering is obtained by using the elbow method introduced in Section 2.3.

3.4. Statistical validation approach

Fig. 4 illustrates the proposed statistical validation algorithm for thermal zoning of buildings. The first part of the algorithm explains the method for splitting the analysed data set into two equal parts. Splitting the data set into 2 parts is suggested based on the simplicity of finding the most distant base points. However, for very large data sets it can be considered splitting the set into more parts. The algorithm starts by computing all the distances among all samples in the 2D PCA space. Then, the two most distant samples are placed in the first set, and the other two most distant samples are placed in the second set. Further, the two sets are populated one after another with the furthest sample from the set until all samples are distributed among the two sets.

This allows us to create two equal sets of data that further undergo the same procedure defined above for PCA and k-means clustering. The final validation step includes comparison of the sensor labels assigned to distinct clusters with the original data set. The validation process shows how much randomness is involved in assigning the labels for each of the cluster. If after splitting the data set most of the sensor labels went into the original cluster, it proves that the thermal zones were assigned meaningfully.

3.5. Qualitative criteria for thermal zones validation

In this work, after applying data-driven methods to determine thermal zones, the resulting zones are evaluated based on the proposed qualitative criteria (Fig. 5). The six aggregated criteria include: envelope exposure and orientation, location of the space within the building, occupancy, schedules, HVAC distribution type & control and space function.

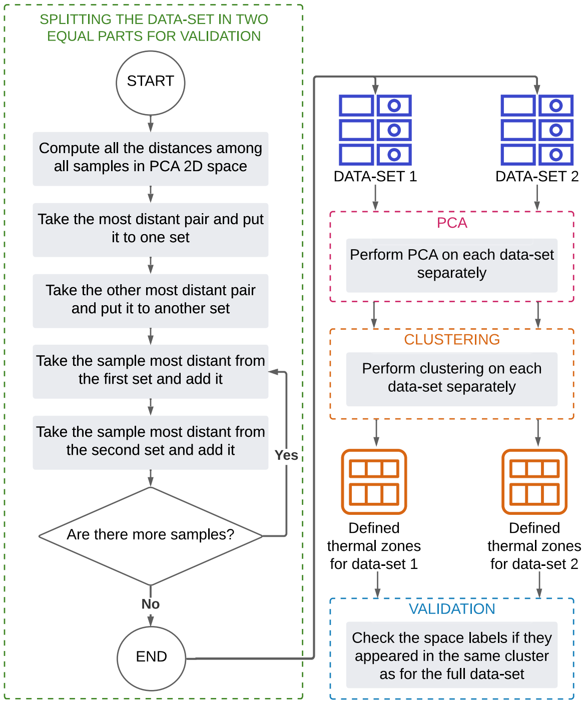

SPLITTING THE DATA-SET IN TWO EQUAL PARTS FOR VALIDATION
START
Compute all the distances among all samples in PCA 2D space
Take the most distant pair and put it to one set
Take the other most distant pair and put it to another set
Take the sample most distant from the first set and add it
Take the sample most distant from the second set and add it
Yes
Are there more samples?
No
END
DATA-SET 1
DATA-SET 2
PCA
Perform PCA on each data-set separately
CLUSTERING
Perform clustering on each data-set separately
Defined thermal zones for data-set 1
Defined thermal zones for data-set 2
VALIDATION
Check the space labels if they appeared in the same cluster as for the full data-set

Fig. 4. Data-driven building's thermal zoning validation algorithm.

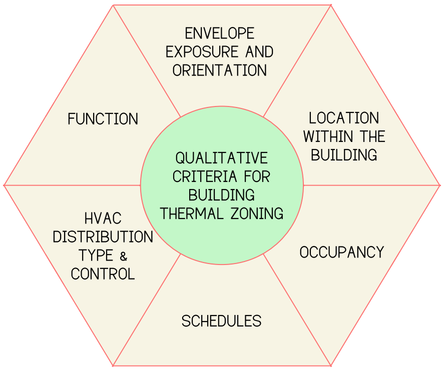

ENVELOPE EXPOSURE AND ORIENTATION
FUNCTION
LOCATION WITHIN THE BUILDING
QUALITATIVE CRITERIA FOR BUILDING THERMAL ZONING
HVAC DISTRIBUTION TYPE & CONTROL
OCCUPANCY
SCHEDULES

Fig. 5. Qualitative criteria for building's thermal zoning.

The exposure and orientation of the envelope includes parameters such as solar irradiation, wind speed, outdoor temperature, façade orientation, and materiality (especially important for glazed façades). The location within the building criterion separates the spaces on the upper and lower floors of the building, as well as the spaces adjacent to the building façades and inner spaces. The occupancy criterion defines the number of users within the spaces. In some cases, additional parameters can be added specifying the type of user or any special needs. Schedules identifies the space usage patterns that could be considered in different scales such as daily, weekly, monthly, etc. The type of HVAC distribution and control criterion allows one to evaluate the practical application of zoning in terms of available hardware and software to control the defined zones. Lastly, the function of the space enables grouping of areas based on their intended use and helps in recognizing spaces that may need particular consideration.

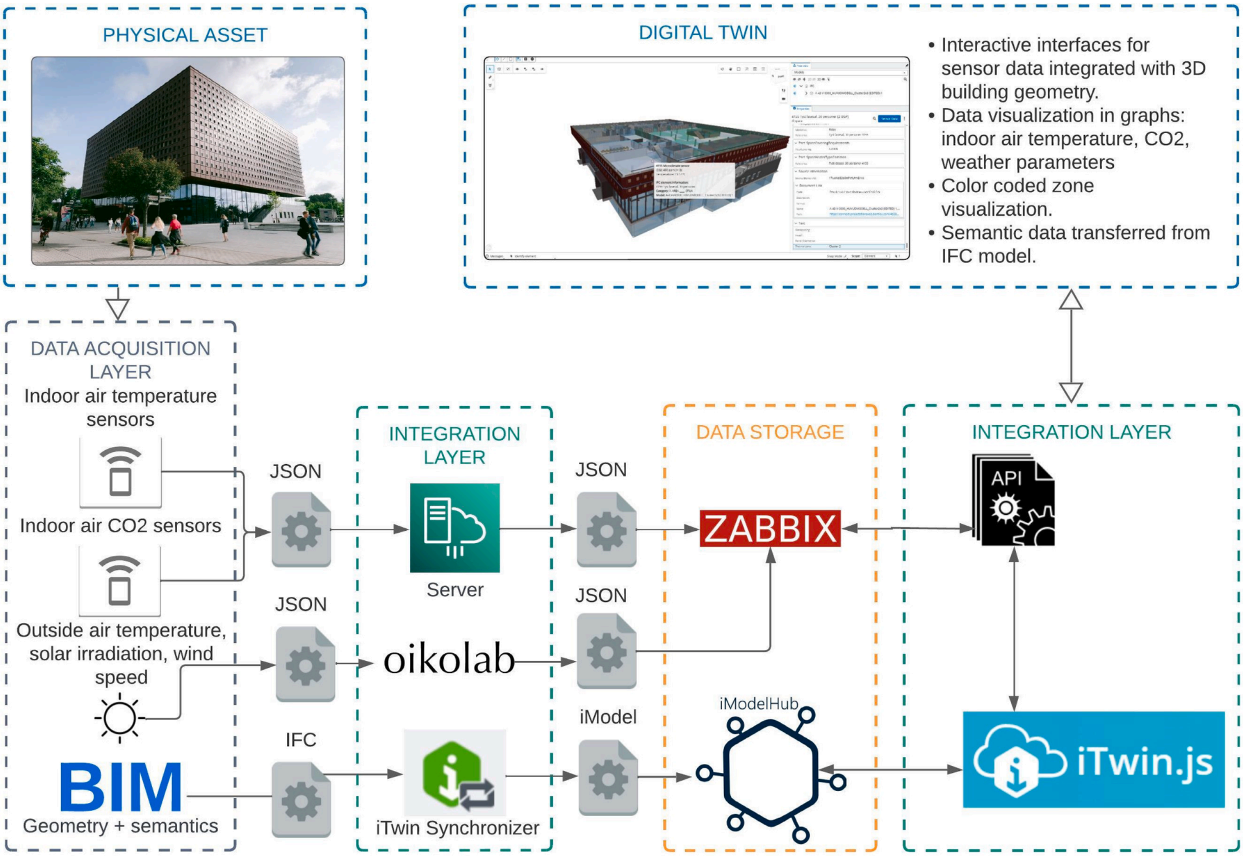

PHYSICAL ASSET
DIGITAL TWIN
• Interactive interfaces for sensor data integrated with 3D building geometry.
• Data visualization in graphs: indoor air temperature, CO2, weather parameters
• Color coded zone visualization.
• Semantic data transferred from IFC model.
DATA ACQUISITION LAYER
Indoor air temperature sensors
Indoor air CO2 sensors
Outside air temperature, solar irradiation, wind speed
BIM
Geometry + semantics
JSON
JSON
IFC
INTEGRATION LAYER
Server
oikolab
iTwin Synchronizer
JSON
JSON
iModel
DATA STORAGE
ZABBIX
iModelHub
INTEGRATION LAYER
API
iTwin.js

Fig. 6. Digital Twin integration architecture.

To enhance the evaluation of the defined thermal zones based on the qualitative criteria, some additional parameters are retrieved. Solar irradiation, wind speed, and outdoor air temperature for Linköping city are collected from the open Oikolab repository [57].

3.6. Digital twin integration architecture

For this research, a DT was developed for the case study building, integrating real-time IoT data with a BIM model containing 3D building geometry and semantic information (Fig. 6). The IoT data was transmitted via an internal server in JSON format to the Zabbix platform, which functions as dynamic data storage. Additionally, weather dataincluding solar irradiation, outdoor temperature, and wind speedwas retrieved from Oikolab in JSON format, stored in Zabbix, and integrated into Bentleys iTwin platform using iTwin.js. Simultaneously, the buildings BIM model, including 3D geometry and semantic data in open IFC format, was imported into iTwin Synchronizer, which converted it to the iModel format and stored it in iModelHub. The model was then integrated into the iTwin platform via iTwin.js, enabling the DT platform to aggregate and visualize real-time sensor data, weather information, and the BIM model.

The user interface (UI) was developed using iTwin.js, enabling interactive visualization of real-time sensor data integrated with 3D building geometry. Sensor readings and weather parameters, are displayed through interactive graphs. The defined thermal zones are represented in different colours and cluster numbers are available on side panel. Semantic data from the BIM model, transferred via IFC, is accessible within the UI, allowing users to retrieve additional building information alongside real-time monitoring. The integrated DT falls under capability level 2 that allow the asset monitoring in real-time, but misses the predictive, prescriptive capabilities and the link between the DT and real asset closing the automation loop.

4. Results and discussion

4.1. Data preparation

4.1.1. Data exploration and cleaning: CO2 data

CO2 data are available for 29 different rooms in the building, including coworking spaces, library, canteen, and offices. Some rooms are

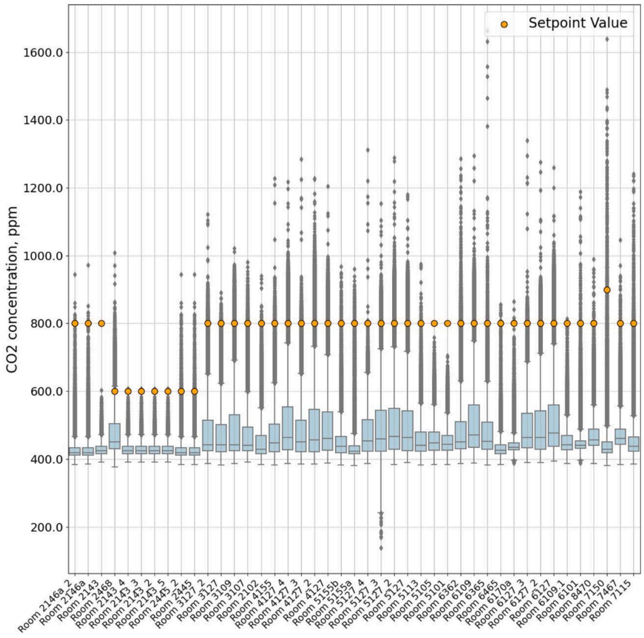


Fig. 7. CO2 concentration in rooms and respective set-points.

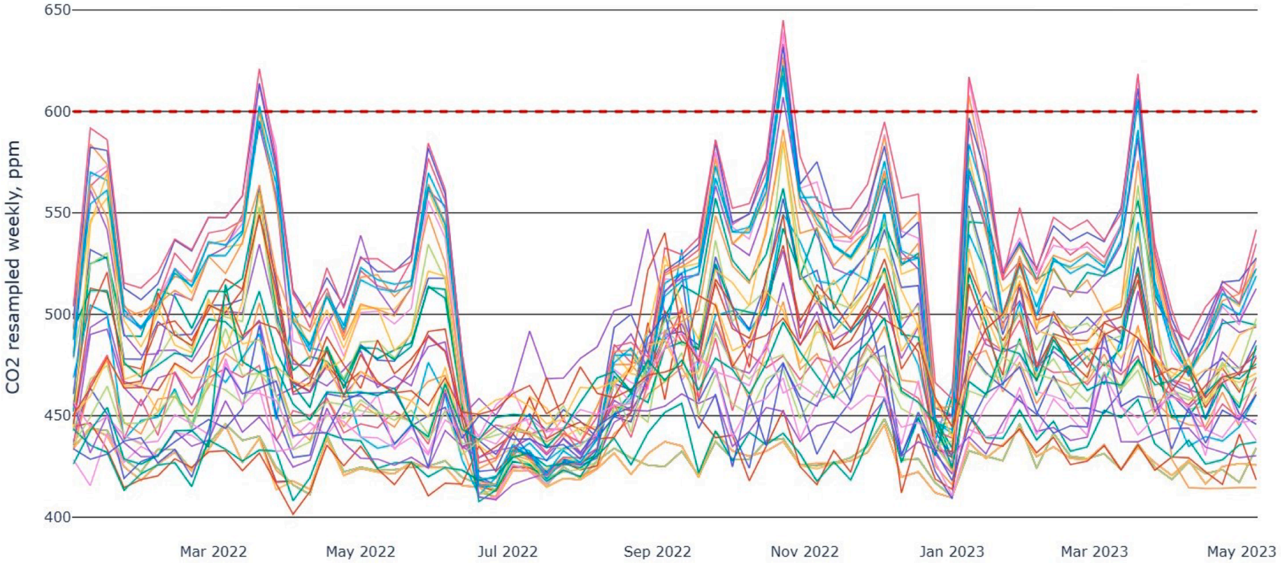


Fig. 8. CO2 concentration re-sampled weekly for separate sensors in rooms. The dashed line indicates the lowest CO2 set-point.

equipped with multiple sensors that expand the data set to a total of 44 separate sensor readings for CO2 levels in various rooms. Before further data analysis, the dataset was subjected to typical cleaning procedures, including resampling and filling the missing values by linear interpolation. The complete CO2 data set included a total of 3,030,812 values, of which 26,377 were missing (approximately 600 per sensor). Some data points were collected 1 minute later than the set 10 minute interval, offsetting the subsequent measurements by the same interval. Resampling the entire data set for a 10-minute interval allowed easier data handling.

The box plots plotted for all the rooms where CO2 data are available show that most values are between 400 and 550 ppm, indicating good indoor air quality (IAQ) in the rooms (Fig. 7). Higher values and a larger interquartile range are mostly observed in larger rooms. The yellow dots in the plot indicate the set point values for separate rooms (for most rooms, the set point for CO2 is 800 ppm). Some extreme values are also visible. For example, room 6365 reaches a maximum value of 1664 ppm. In contrast, in room 5127, one of the four sensors shows exceptionally low values, with a minimum value of 139 ppm. Room 6365 is a classroom with a capacity of 40 persons, which explains the higher CO2 values when the room is fully occupied. On the other hand, room 5127 is a library room with a large open space. This space is equipped with four sensors, only one of which measures very low values. This is most likely due to a faulty sensor measurement. For the case study building, regular sensor calibration is not implemented; however, the CO2 sensors in the VAV system have automatic baseline correction (ABC).

A clear influence of occupancy on air quality in rooms is represented in the weekly re-sampled CO2 concentration plot (Fig. 8). In this plot, the CO2 values were re-sampled taking the mean value of the 10-minute readings for the target week. During the summer holidays, re-sampled weekly CO2 values almost never exceed 500 ppm. The same can be observed during the winter break. It is also interesting to observe that although CO2 varies from room to room, trends are moving almost parallel. This indicates a highly controlled environment that is similarly influenced by external factors.

4.1.2. Data exploration and cleaning: Indoor air temperature data

Indoor air temperature data consists of measurements from 218 sensors, some of which are located in the same room. Similarly to the CO2

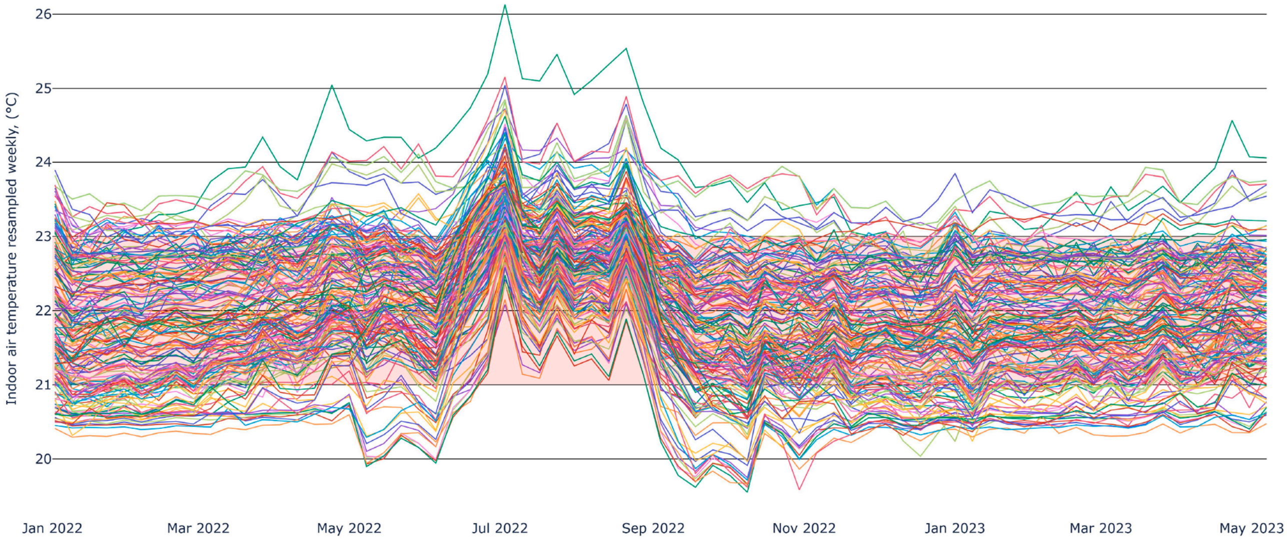


Fig. 9. Indoor air temperature re-sampled weekly for separate sensors in rooms. The red background indicates the set-point range. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data, the time stamps were resampled every 10 minutes and missing values were filled in using linear interpolation. In the case of the indoor air temperature dataset, a total of 15,222,816 data points were collected, of which 154,724 were missing (approximately 700 per sensor).

The box plots show that the temperature in all rooms is on average distributed between 20°C and 24°C (Fig. 10), i.e. close to the set-point range. However, there are quite significant number of rooms in which the majority of measurements do not satisfy the set point condition. For example, for the rooms on the second floor, rooms 2109 - 2119 in Fig. 10 indicated by a red-dashed rectangle), most of the values are always below the 22°C set point. Similarly, there are a number of rooms, including some large library spaces (in Fig. 10 indicated by a green-dashed rectangle), where the temperature is most of the time above the set point, most likely due to high user concentration.

The lowest temperature observed in room 4171 (in Fig. 10 indicated by a purple-dashed rectangle), which is a group room for 8 persons on the western façade of the building, is 13.99°C. Interestingly, the same room also has one of the highest temperature values in the data set, 27.61°C. Extreme values, among others, can be explained by the influence of the user, such as leaving an open window in the cold or hot seasons. The maximum indoor air temperature in this data set is 31.84°C in room 7126 (in Fig. 10 indicated by blue-dashed rectangle). Room 7126 is a resource room for 2–4 persons on the south elevation of the building.

The weekly resampled indoor air temperature data shows a clear seasonality pattern (Fig. 9). In this plot, the indoor air temperature values were resampled by taking the mean value of the 10 min readings for the target day. During the summer period, the temperature in all rooms is around 2°C higher than in other seasons. In addition, a strong control influence can be observed for measurements that show similar patterns in indoor air temperature change for all rooms. A clear start and end of the heating season can also be observed. Here, indoor air temperature drops to exceptionally low values, even compared to the winter period. In Fig. 9 this can be seen between the months May-July and September-November, where the indoor temperature drops below 20°C. In contrast, between July and September, the indoor air temperature is exceptionally high due to the warmer weather outside. The exploratory analysis of the indoor air temperature indicates that in many cases RBC cannot meet the indoor thermal comfort requirements.

4.2. PCA and k-means clustering for thermal zoning

4.2.1. Joint indoor air temperature and CO2 data set

The merged indoor air temperature and CO2 data set consists of 35 rows with separate sensors in rooms and 48 columns (CO2 and indoor air temperature mean and STD values for each month) for each month of 2022 with calculated mean and standard deviation values. Once the values are standardised, PCA is run. The results show that the first five principal components can describe 90.3 % of the data (Fig. 11). The first principal component accounts for the 50.3 % individual explained variance ratio, the second for 19.6 %, the third for 12.1 %, the fourth for 4.6 % and the fifth for 3.7 %. How the first 3 components explain the individual parts of the data set can be seen in Fig. 12.

Fig. 12 shows the loadings for each of the features for PC1, PC2 and PC3. In other words, the loadings represent how much each original variable contributes to each principal component. PC1 captures the most variation in the data. Significant positive and negative loadings across different months indicate that PC1 represents the general contrast between different seasonal patterns, such as the differences in indoor conditions between the cooler and warmer months. Further, PC2 captures additional variance not explained by PC1, associated with fluctuations in temperature and CO2 that are orthogonal (independent) to those captured by PC1. The alternating positive and negative loadings suggest that PC2 represents transitions between states, possibly capturing changes like switching between heating and cooling in a building's HVAC system or changes in occupancy patterns. Finally, PC3 captures smaller, more nuanced variations in the data. The loadings show smaller

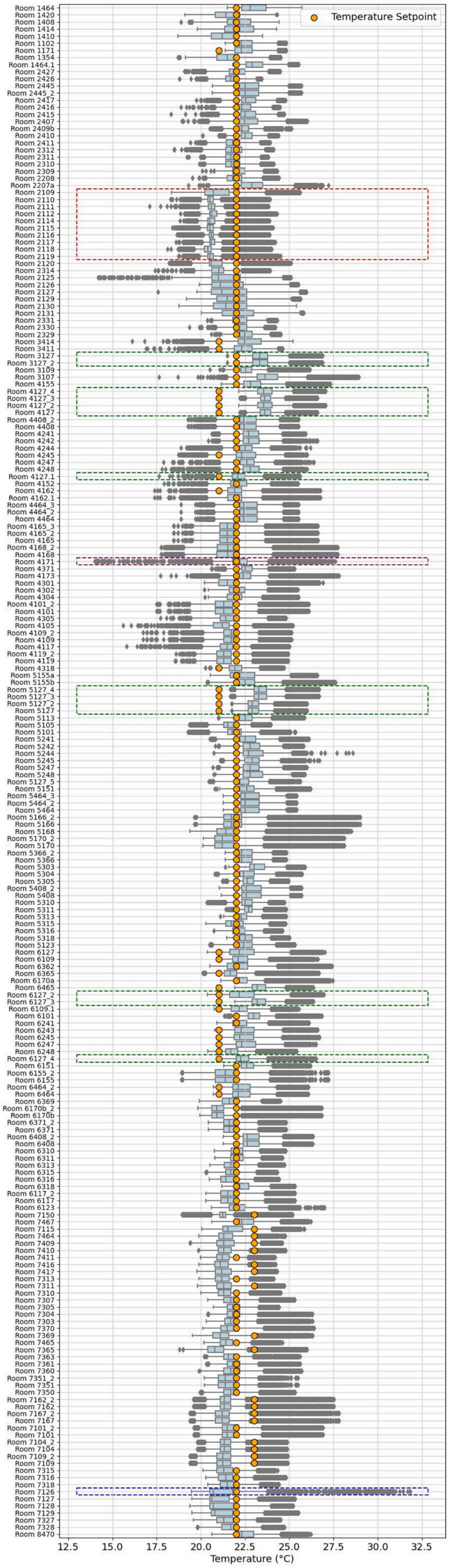


Fig. 10. Indoor air temperature in different Rooms (°C).

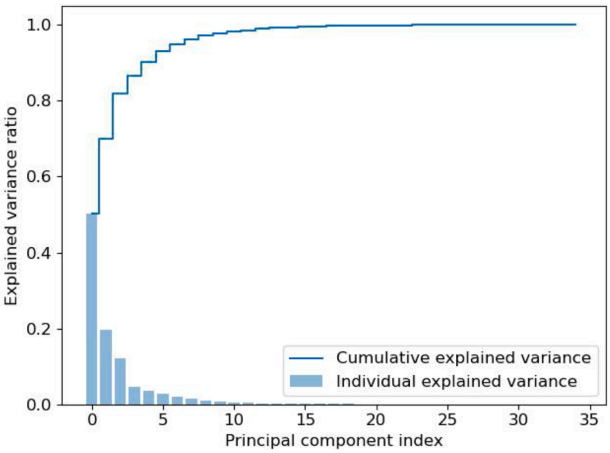


Fig. 11. Individual and cumulative explained variance for the joint CO2 and indoor air temperature data set.

peaks (except for temperature STD) and less consistency, which could indicate specific events or anomalies in the data such as equipment malfunctions, unusual weather conditions, or irregular occupancy patterns that do not follow the larger seasonal trends.

Based on recommendations provided in Section 3.3, for this case study, 5 PCs were used in further analysis as their cumulative explained variance covers more than 90 % of the total variance of the data set. Before running the k-means clustering algorithm, an “elbow” method is used to determine the optimal number of clusters. However, for this data set, the method did not yield significant results. Therefore, as described in Section 2.4, a mathematical approach was used to identify the “elbow” points. The most significant “elbow” was found at cluster No. 3 (with a second derivative value of 14.87). Since this data set includes the fewest number of rooms and the most distinct variables (CO2 and indoor air temperature), the decision was made to retain 3 clusters, as indicated by the “elbow” method (Fig. 13). However, from the plotted curve, it is evident that five clusters (with a second derivative value of 16.69) could also be a viable option, allowing for the capture of more subtle patterns in the data. At seven clusters, the curve shifts from upward to downward as the second derivative value turns negative (-9.69), indicating that seven clusters would likely introduce unnecessary complexity.

Therefore, given the number of rooms analysed, 3 clusters are selected and the k-means clustering algorithm is run. Rooms 3127, 4127, 5127 and 6127 fell into the same green cluster (Fig. 14). These rooms are larger open library spaces spread over four floors. Here, CO2 concentrations and average indoor air temperatures reach their highest values, probably due to the high concentration of occupants. However, high CO2 concentration values could also appear due to faulty ventilation system operation.

The second big cluster in red (Fig. 14) mostly includes large meeting rooms and classrooms, such as rooms 6362, 6365 or 5155, where the standard deviation of the CO2 and temperature values is slightly lower than in the green cluster. Finally, the two rooms in the purple cluster are the recording room, which can accommodate five persons, and the main entrance hall. These rooms have very stable variations in CO2 and temperature. For example, the lowest mean temperature value in room 6465 is 23.15° and the highest is 23.51°. Similarly, the CO2 concentration varies only from the minimum mean value of 433.23 ppm to the maximum value of 443.45 ppm. These 3 clusters are shown in the floor plans in Fig. 15.

4.2.2. Indoor air temperature data set

As the indoor air temperature data set contains 218 individual sensor readings, it was interesting to see the results of the proposed thermal zoning algorithm in a larger data set. The final indoor air temperature data set was combined from 218 rows with sensors in rooms and 24 dimensions, including mean and standard deviation values for each month of 2022. Similarly to the merged CO2 and indoor air temperature data set, the PCA results show that the first 5 principal components can describe 94 % of the data (Fig. 16). The first principal component accounts for 51.4 % of the individual explained variance, the second for another 25.3 %, the third for 8,8 %, the fourth for 5.7 % and the fifth for 3.8 %. How the first 3 components explain the individual parts of the data set can be seen in Fig. 17.

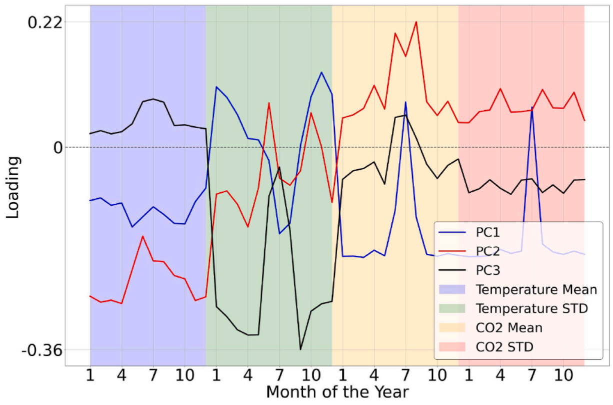


Fig. 12. The first 3 principal components of the joint data set.

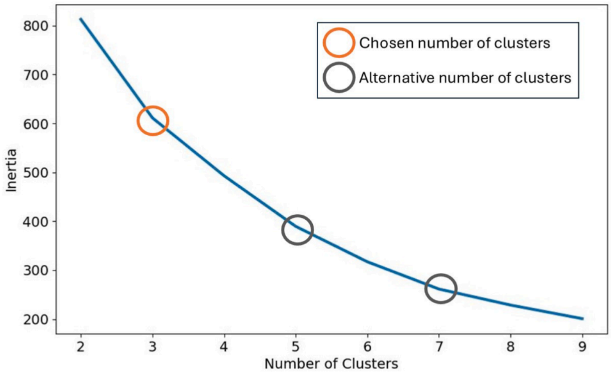


Fig. 13. The “elbow” method results for joint dataset.

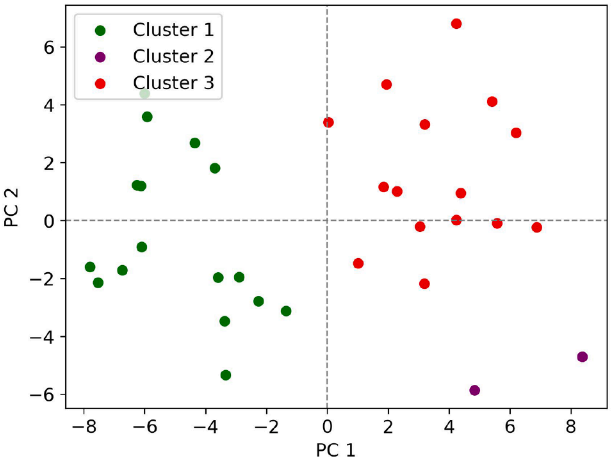


Fig. 14. K-means clustering results for joint data set.

PC1 reflects the overall average temperature trends throughout the year, with more significant loadings around the colder months and less significant around the warmer months. This suggests that PC1 could

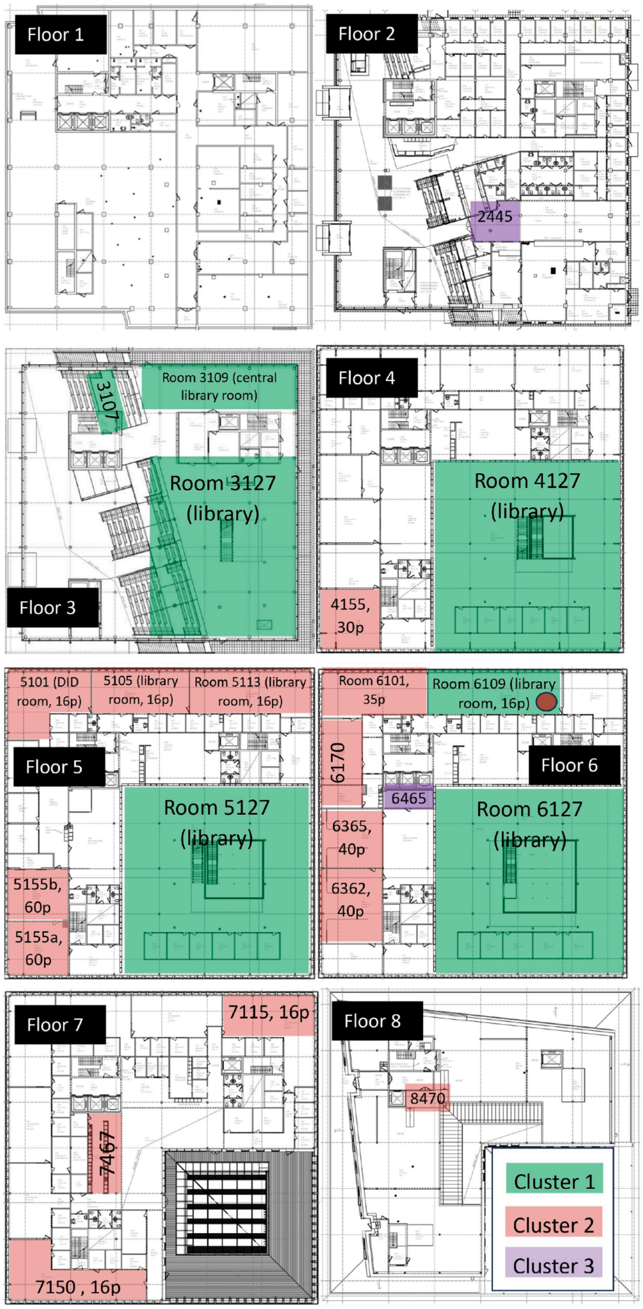


Fig. 15. Joint temperature and CO2 data set clusters in plan view.

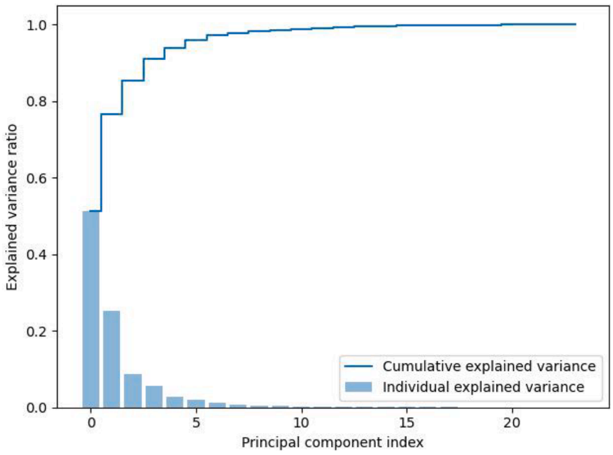


Fig. 16. Individual and cumulative explained variance for the temperature data set.

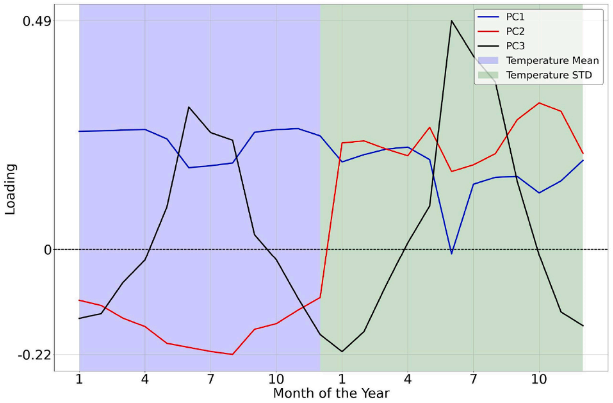


Fig. 17. The first 3 principal components of the temperature data set.

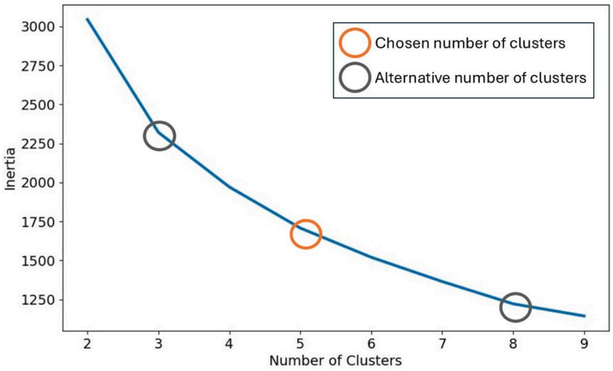


Fig. 18. The “elbow" method results for indoor air temperature dataset.

be capturing the basic seasonal temperature shift. Peaks in the colder months suggest that this component might capture when and how temperatures are generally lower, impacting building heating demands. PC2 shows an alternating pattern of loadings and may represent changes in temperature variability that are not explained by the general seasonal trend. Negative loadings for mean indoor air temperature values represent adjustments in indoor climate control or differences in external temperature fluctuations. As the most significant (lowest negative values) are during the summer months, this could indicate the cooling season. PC3 loadings range from negative to positive, indicating variable influences throughout the year, with less consistency compared to PC1 and PC2. PC3 exhibits its highest loadings for STD values, especially notable from 0.092 to 0.486, suggesting that PC3 is particularly sensitive to changes in temperature variability rather than average temperature.

The same as for the merged CO2 and indoor air temperature data set “elbow” method did not show the optimal number of clusters. Similarly to the CO2 dataset, the most significant “elbow” point is observed at 3 clusters (with a second derivative value of 83.79). At 5 clusters, the “elbow” is less pronounced (second derivative value of 21.73); however, as previously discussed, there is still considerable variation to capture, as the curve does not flatten immediately (Fig. 18).

As there are more rooms to analyse in this case, five clusters were selected (Fig. 19). Interestingly, the small offices (4 persons) on the north façade of the second floor were clustered in the same group on the left (dark green cluster). These rooms are the only ones in the data set where the average indoor air temperature values decreased to 19°. Looking at the blue cluster, 2 rooms drop out at the very top of the graph. These are room 1410 (women’s changing room) and room 2109 (small office

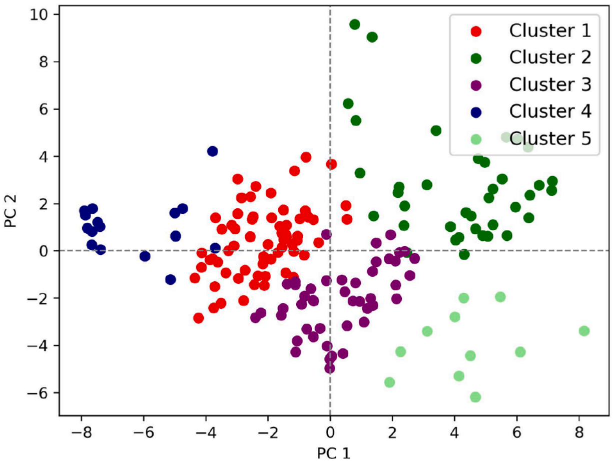


Fig. 19. K-means clustering results for temperature data set.

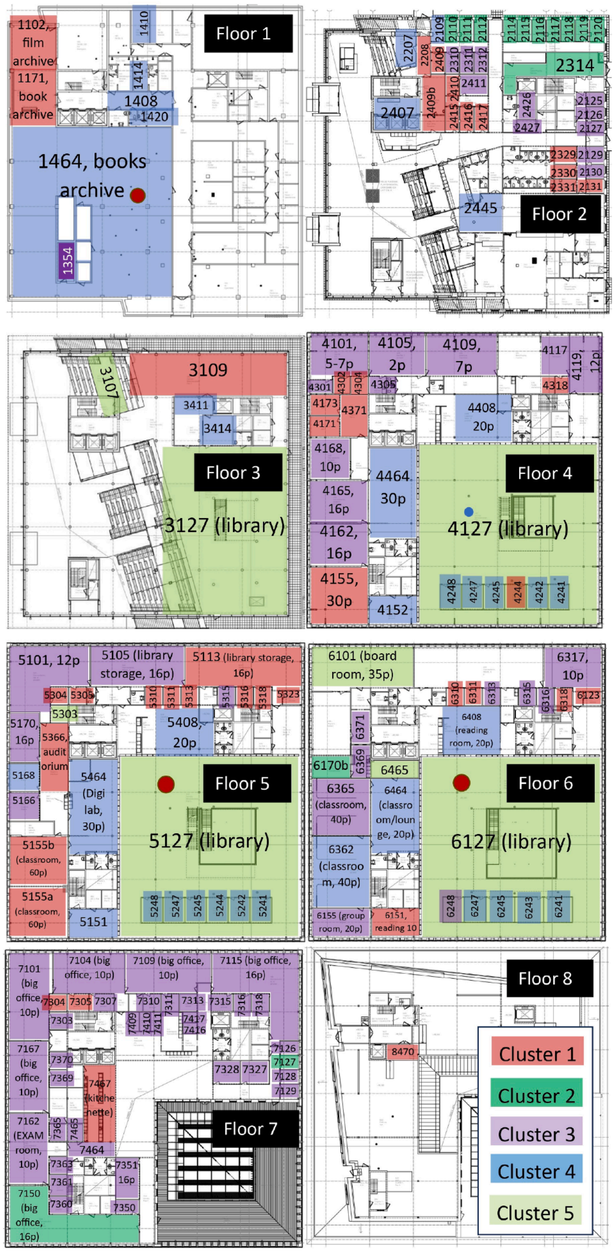


Fig. 20. Indoor air temperature data set clusters in plan view.

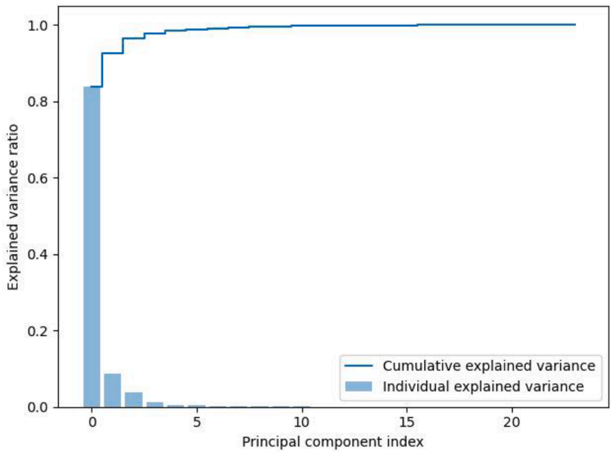


Fig. 21. Individual and cumulative explained variance for the CO2 data set.

on the second floor north façade). In both rooms, the temperature varies little (about 1.5°) with respect to the minimum and maximum values. Interestingly, room 2109 adjacent to one of the rooms on the 2nd floor north façade shows a different performance. Here, the average temperature never drops below 20°. Finally, the light green cluster also shows some values farther away from most points in the data set. These values belong to the large spaces in the building, i.e., rooms 4127, 5127, 6127, which is the library, and 3107, which is the amphitheatre space. These 5 clusters are shown in the floor plans in Fig. 20.

4.2.3. *CO2 data set*

Finally, the CO2 data set is also analysed separately. This data set is combined from 44 rows with sensors in rooms and 12 dimensions with mean and standard deviation for each month of 2022. The PCA on this data set revealed slightly different results compared to the first two in terms of the identified principal components. In this case, the first PC alone explains 83.8 % of the data, the second explains 8.8 %, and the third explains 3.9 % (Fig. 21). This means that only three PCs can explain 96.5 % of the variation in the data set. The way these three components explain separate parts of the data set can be seen in Fig. 22.

PC1 represents the overall average level of CO2 concentration, capturing baseline fluctuations in different months. Fluctuations are influenced by seasonal changes that affect indoor CO2 levels, such as varying occupancy and ventilation rates. PC2 exhibits more pronounced swings, with sharp negative loadings around the 7th month and smaller positive loadings at other points. PC2 seems to capture the most significant deviations from the average CO2 levels, possibly reflecting specific periods of high variability or exceptional events, such as sudden changes in building occupancy or HVAC system performance. The sharp negative peak in the middle of the year indicates a period with unusually low CO2 variability during low occupancy in the summer months.

The “elbow” method for the CO2 data set produced clearer results, with a distinct “elbow” point appearing at three clusters. However, as with the other datasets, the plot (Fig. 23), suggests that adding more clusters could increase precision in defining the thermal performance of each zone. Since this data set contains only CO2 values, 5 clusters were selected for further analysis in order to capture more variability.

Once again, the large areas with the library and other large rooms were grouped in the blue cluster on the right-hand side (Fig. 24). The left side (purple cluster) shows a large cooking area with several sensors. Here, the mean CO2 values range from a low of 419.04 ppm to a high of 433.26 ppm. The red cluster shows rooms 5155a and 5155b, which are classrooms with a capacity of 60 persons. Here, the range between the average minimum and maximum values is larger: min. 418.97 ppm and max. 461.53 ppm. These 5 clusters are shown in the floor plans 25.

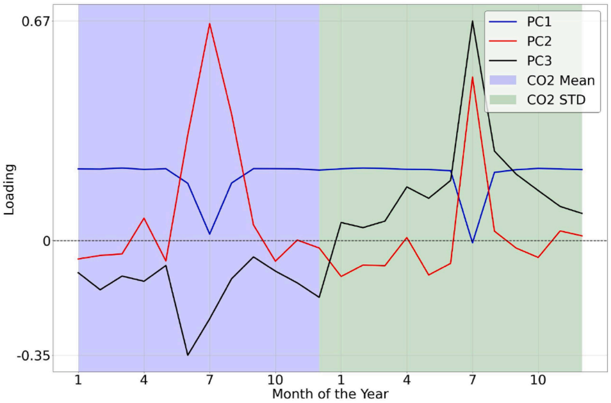


Fig. 22. The first 3 principal components of the CO2 data set.

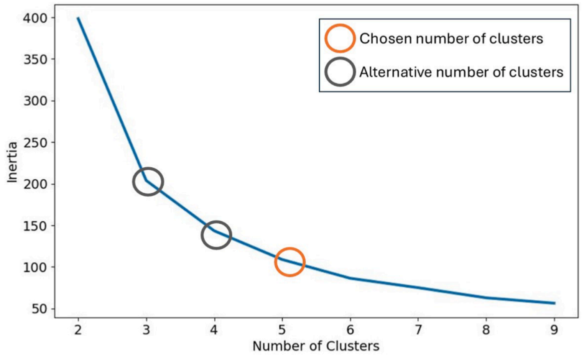


Fig. 23. The “elbow" method results for CO2 dataset.

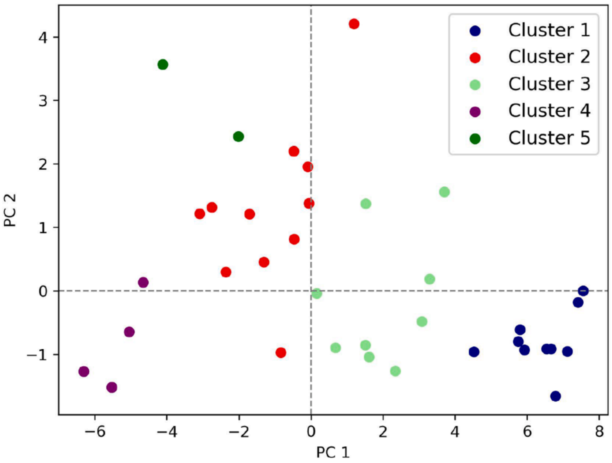


Fig. 24. K-means clustering results for CO2 data set.

4.3. Statistical validation results

The thermal zones identified using three separate data sets discussed above are further validated using the proposed data-driven validation algorithm (Fig. 4). The split of the two parts of the first data set, including CO2 and indoor air temperature measurements, is shown in Fig. 26. The figure shows that the data points equally cover the space for both data sets. Since the combined CO2 and temperature data set contained a total of 35 measurements, the split resulted in 18 measurements in one part and 17 measurements in the other. The other two data sets, which contain only indoor air temperature measurements and only CO2 measurements, were split in a similar way, resulting in two data sets with 109 data points each for indoor air temperature and 22 measurements each for CO2.

Further, the thermal zoning algorithm is applied to both parts of each data set using PCA and k-means clustering. The results of the statistical validation of each data set are presented in Table 1. To check whether

Table 1
Statistical thermal zoning validation results for case study building.

Data set	**Combined CO2 + indoor air temperature data set**									
Total data-points	35									
Data-points in splitted parts	18					17				
Cluster number	1	2	3	4	5	1	2	3	4	5
Total data-points in cluster	8	2	8	–	–	7	1	9	–	–
Data-points from wrong cluster	0	0	0	–	–	0	0	0	–	–
Cluster consistency score, %	100	100	100	–	–	100	100	100	–	–
Total consistency score for data set, %	100									
Data set	**Indoor air temperature data set**									
Total data-points	218									
Data-points in splitted parts	109					109				
Cluster number	1	2	3	4	5	1	2	3	4	5
Total data-points in cluster	25	38	23	15	8	25	43	25	8	8
Data-points from wrong cluster	0	1	3	2	0	0	10	0	3	0
Cluster consistency score, %	100	97	91	87	100	100	77	100	63	100
Total consistency score for data set, %	91.5									
Data set	**CO2 data set**									
Total data-points	44									
Data-points in splitted parts	22					22				
Cluster number	1	2	3	4	5	1	2	3	4	5
Total data-points in cluster	5	7	4	5	1	5	6	5	5	1
Data-points from wrong cluster	0	2	0	0	0	0	0	0	0	0
Cluster consistency score, %	100	71	100	100	100	100	100	100	100	100
Total consistency score for data set, %	97.1									

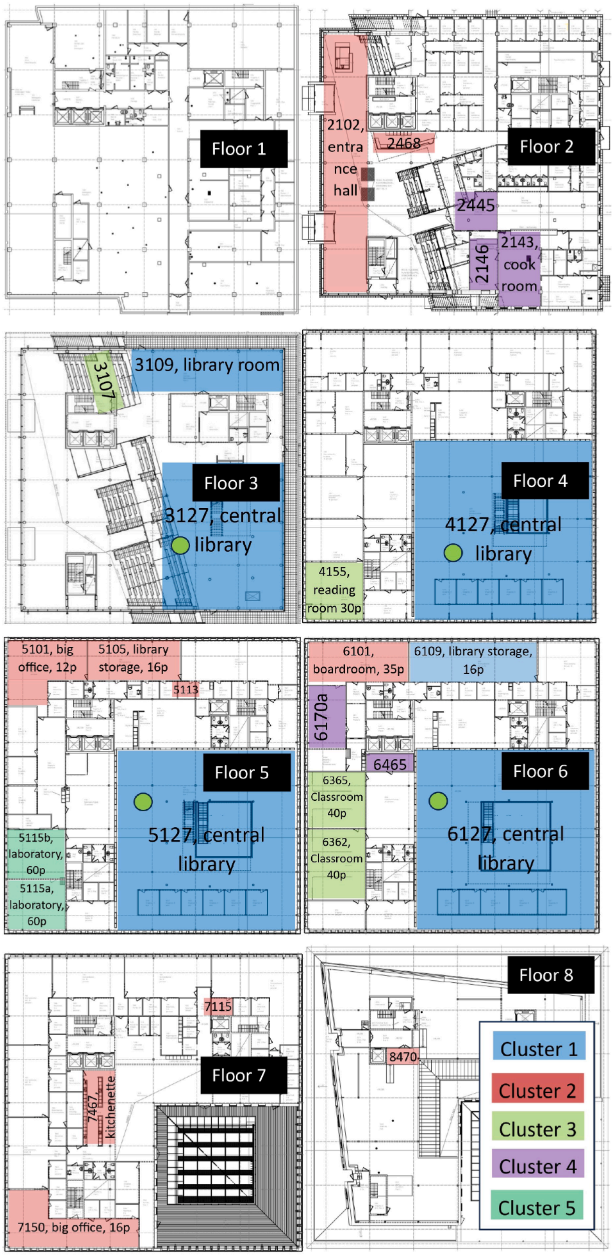


Fig. 25. CO2 data set clusters in plan view.

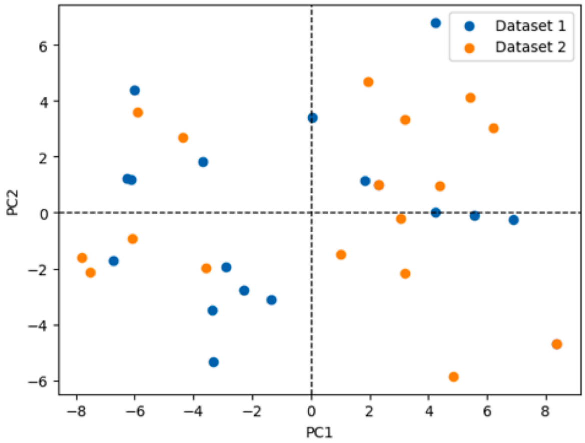


Fig. 26. CO2 and indoor air temperature data set split.

the same rooms fall into the original clusters after performing the PCA and k-means analysis on the split data sets, sensor labels were extracted. In the table, cluster consistency score (%) refers to the percentage of data points in the split data set that matched the same cluster in the original data set. This percentage reflects how well the clustering algorithm was able to replicate the cluster assignment in both halves of the split data. Total consistency score for data set (%) represents the overall percentage of data points that were correctly classified into their respective clusters across both parts of the split data set.

The joint CO2 and indoor air temperature data set was initially clustered into 3 clusters, so the same number of clusters was also used for validation. The validation results for this data set showed that 100 % of the relevant labels went into the original clusters, indicating that the number of PCs, clusters, and input data was optimal to define the thermal zones. The results for the larger data set, which included 218 indoor air temperature sensors, were slightly less accurate. Some original cluster labels were mixed up. For example, in one of the split-data set clusters, 10 out of 43 labels were included from the wrong cluster, resulting in consistency score of only 63 % for that cluster. This shows that in larger data sets, a single parameter may not be sufficient to define the thermal zones of a building. This data set was initially divided into 5 clusters, so 5 clusters were also used for validation. Several other combinations of clusters and the number of PCs to be included in them were also tested during the analysis. However, the 5 clusters with 3 main PCs showed the best results, achieving 91.5 % consistency score for the whole data set.

For the data set that contains only CO2 data, only one of the clusters in the split data set had 2 misplaced labels, reducing the consistency score of that cluster to 71 %. However, the other clusters contained the same labels as the original cluster, resulting in an overall data set consistency score of 97.1 %. The validation results showed that for the proposed data-driven thermal zoning algorithm, the number of input parameters, the number of clusters chosen, and the PCs to be included can significantly influence the consistency score of thermal zoning.

4.4. *Qualitative validation results*

4.4.1. *Envelope exposure and orientation*

The results of the data-driven thermal zoning are further put for qualitative evaluation based on the criteria defined in Fig. 5. Considering envelope exposure and orientation criteria, several factors, such as solar irradiation, wind speed, wind direction, façade materiality, and available shading, play an important role.

Typical solar radiation to the horizontal surface on a clear day can reach $1W/m^2$. The data from Linköping city show very similar values, with solar radiation reaching above $0.8W/m^2$ during the summer months (Fig. 27). However, the radiation is proportionally decreasing towards the winter, where it barley reaches $0.1kW/m^2$. This indicates that solar radiation can significantly influence the thermal zoning of buildings during the summer, depending on the available shading options.

The analysed “Studenthuset” building has several shading options installed. On the 2nd floor, where the main entrance is located, only manual shading is available. Up to floor 7, the shading options vary depending on the façade orientation. Automatic shading is installed for most of the west façade and about for half of the south and east façades. All rooms on the upper floor also have automatic shading. The remaining rooms are equipped with fixed sun protection, which leaves between 53 % and 83 % of the glazed area uncovered, depending on the time of day and the cloud cover.

As a large part of the building’s façade is glazed, solar irradiation can be influential for the thermal zoning. Considering the thermal zones defined by a proposed data-driven method, the influence of solar irradiation, orientation, and shading is best seen in the zones defined by the indoor air temperature data set. Fig. 20 shows that the rooms on the north façade floor 2, where only manual shading is available, fall

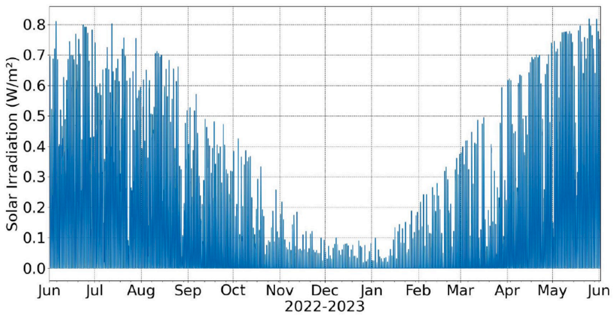


Fig. 27. Solar irradiation for year 2022–2023 in Linköping city.

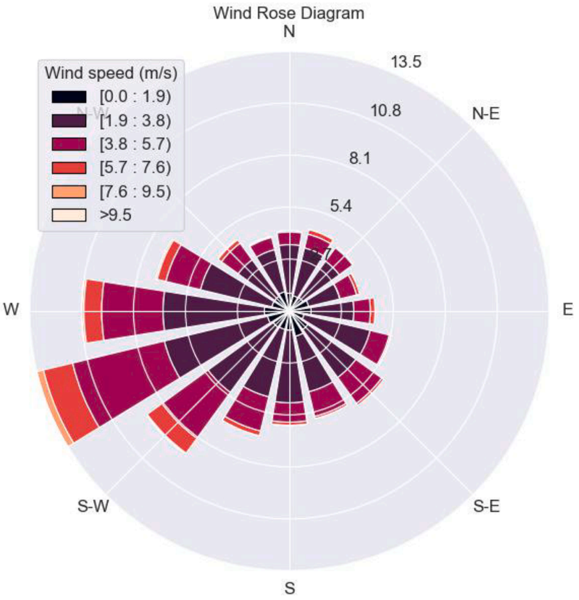


Fig. 28. Wind rose diagram for year 2022–2023 in Linköping city.

into a separate cluster. These rooms show the lowest single temperature values throughout the whole data set. On floors 4 and 5, part of the rooms on the west façade fell into red and the other part into purple clusters. This could also be partly influenced by the available shading (automatic or fixed) and the orientation of the building. The large library spaces fell into one thermal zone (light green). However, a coloured dot on the plans indicates that one of the sensors installed in the specific space fell into a different zone. For example, on floors 5 and 6, one of the sensors in each of the two library spaces has been placed in the red cluster. This may also have been influenced by the positioning of the sensors in a large space where they may be more influenced by external factors.

The wind rose diagram (Fig. 28) shows that the wind is mostly from the west, where it reaches its highest speed. The wind infiltration rate of the building is assumed to be $0.25l/s$ and the air leakage rate is assumed to be around 50 Pa, which means that the buildings are considerably airtight. As the case study building is a large educational building, such an infiltration rate at a wind speed of 10 m/s could even be useful for air exchange, reducing the load for ventilation systems. Therefore, in this case, the wind can be neglected for thermal zoning.

The insights from qualitative envelope exposure and orientation criteria analysis show that the thermal zones of the case study building are influenced by solar irradiation, available shading options, and overall orientation of the façades. Compared with the quantitative results of the thermal zoning, it can be seen that the thermal zones defined using the proposed method reflect this criteria.

4.4.2. *Location within the building*

Space location within the building criteria is closely related to the environmental exposure and orientation criteria. In most cases, it is considered that if a space is not adjacent to the building's façade or located at the bottom or top floors, the outdoor environmental influence on the space can be neglected. This is also highlighted in the ASHRAE standard [25], which suggests assigning perimeter spaces that have exposure to the outdoor environment, the upper / lower floors, and the core of the building as separate zones. However, according to the data, this type of thermal zoning seems to be too generalised.

Taking into account the results of the data-driven thermal zoning approach on the indoor air temperature data set (Fig. 20) it is seen that the spaces of the 2nd floor north façade fall primarily into the green cluster and the spaces of the east façade into the purple. However, the core spaces that do not have any of the bounding walls facing the exterior are still divided between 4 separate clusters. This is also seen while investigating the division between the clusters among the other building's floors. Another point to consider is the corner spaces. According to the standard, the corner space should be cut diagonally, dividing the thermal zones based on the exterior façade. However, this approach is not practical, as the HVAC control cannot be applied in such a way. Performance-based thermal zoning allows to determine the appropriate thermal zone for corner spaces. For example, the large library space distributed throughout floors 3,4,5 and 6, includes more than half of both south and east façades. However, based on the data-driven method, all these spaces are assigned to the same cluster.

However, some patterns are still seen in the data-driven clustering approach that complies with the ASHRAE standard suggestions. For example, almost all spaces on the top 7th floor fell into the same purple cluster, as also recommended by the standard. Similarly, spaces adjacent to the façades of the building are normally split into separate groups. For some façades, the exterior facing spaces are divided into two clusters, which is likely due to the different shading options available that were discussed in the previous Section.

4.4.3. *Occupancy*

The occupancy of the building, as a qualitative criterion, involves several assumptions. The maximum space capacity is normally defined during the building design phase. However, the actual occupancy during the operational phase can only be determined on the basis of the measured data and is highly dependent on the schedule criteria, discussed in the next section. Another important consideration is that the capacity of the room does not always match the area, meaning that a room with larger useful area could have lower occupancy and vice versa. For example, the case study building, on the fifth floor room 5101 with projected capacity of only 12 persons, has a larger total area than room 5155b with capacity of 60 persons. This distinction is often related to the room's function, which is also connected to the density of occupants in a room. As in this case, room 5101 is a large office space, while room 5155b is a classroom.

Although CO2 levels in rooms are determined not only by the number of people in a room, but also by user activity and HVAC control, it is often used to evaluate occupancy. In the case of the analysed building, the data set that includes only CO2 data (Fig. 25) proves these assumptions. In the determined thermal zones based only on CO2 data, it is clearly seen that rooms with similar capacity went to the same clusters, although some rooms with higher capacity have a smaller useful area.

4.4.4. *Schedules*

Knowing the building's operational schedules is paramount for optimal control and energy balancing. With increasing penetration of renewable energy sources (RES) into the grid, the fluctuation in energy prices became even more prominent. This, on the one hand, poses a challenge for current energy supply-demand models and, on the other, reveals an opportunity for improved control mechanisms taking advantage of cheap energy prices at a certain time. That being said, it is important to take this criterion into consideration while defining building's thermal zones.

As the case study building is an educational building, there are two main usage schedules defined depending on the space function. For office spaces, the building is considered operational between 7AM and 6PM during work-days; for all other spaces, the operational hours are between 7AM and 10PM on weekdays, and 8AM to 5PM on weekends. In this study, the data for the thermal zoning were not separated according to the schedules. However, since schedules are closely related to the occupancy of the building, the distinction between clusters is clearly visible in the CO2 data set (Fig. 25).

4.4.5. *HVAC distribution type and control*

An in-depth understanding of existing HVAC systems and available control applications is crucial before defining the thermal zones of the building. Normally, there are many different systems and their combinations employed to ensure the comfort of building occupants. It is common that not all systems allow highly distributed control options for separate spaces or rooms. In such cases, there is no use in combining separate rooms into a zone if the installed control mechanisms are not able to implement defined algorithms for the zone's control model.

For "Studenthuset" case study, mechanical ventilation with 3 AHUs and 1 air circulation unit installed in the building ensures the air quality in rooms. The AHUs allow manipulation of the air flow rate in separate rooms via installed dampers that are controlled based on the CO2 set-points. However, the implemented VAV control only allows the air exchange, but is not able to adjust the supply air temperature. In this case, the supply air temperature can be manipulated only centrally at the AHUs. The projected set point for the supply air temperature is 18°C for the kitchen areas and 16°C for all the other rooms. These values are set to low to ensure cooling needs during the summer. However, according to the operational data retrieved from the building BMS, the duct temperature in specific rooms varies by +/-2°C. This can be influenced by several parameters including the damper position, air flow rate, or even the current temperature in the room.

Furthermore, a larger central extraction of exhaust air is installed on floor 6 of the roof lantern, which is balanced against the supply air in public library areas. In addition, supply air ducts are thermally insulated against unwanted heating in view of low supply air temperature, except the ducts at level 1 which are not thermally insulated. These factors also influence the temperature of the supply air, meaning that even if the supply air from the AHUs is set to 16° C, it is likely that the actual temperature will vary for separate rooms in the building.

DHS is used for building heating and hot water preparation needs. In all rooms, radiators are installed. In addition, on the second floor, for the main entrance and large amphitheatre space, floor heating is used. The floor heating system is divided into 4 branches that can be controlled separately via installed valves. Similarly, the radiators installed in separate rooms can be individually adjusted while manipulating the valves by the BMS system or manually via installed thermostats.

The synthesis of ventilation and heating systems ensures the required IAQ in building's spaces. The ventilation system, controlled based on the CO2 levels, allows exchanging the needed amount of air in rooms. However, the exact supply air temperature cannot be controlled for separate rooms. Therefore, during the cold season, the heating system must compensate for the possible temperature drop.

These insights are especially important when developing the dynamic thermal models for separate zones control. In this case, since dampers and heaters can be controlled separately in different spaces, thermal zoning can be versatile, and a single model can be adapted for the defined zones. However, for different HVAC systems this could not be the case. Therefore, before starting the thermal zoning, it is very important to analyse existing HVAC systems and available control measures to ensure that the developed model can be implemented for the zone.

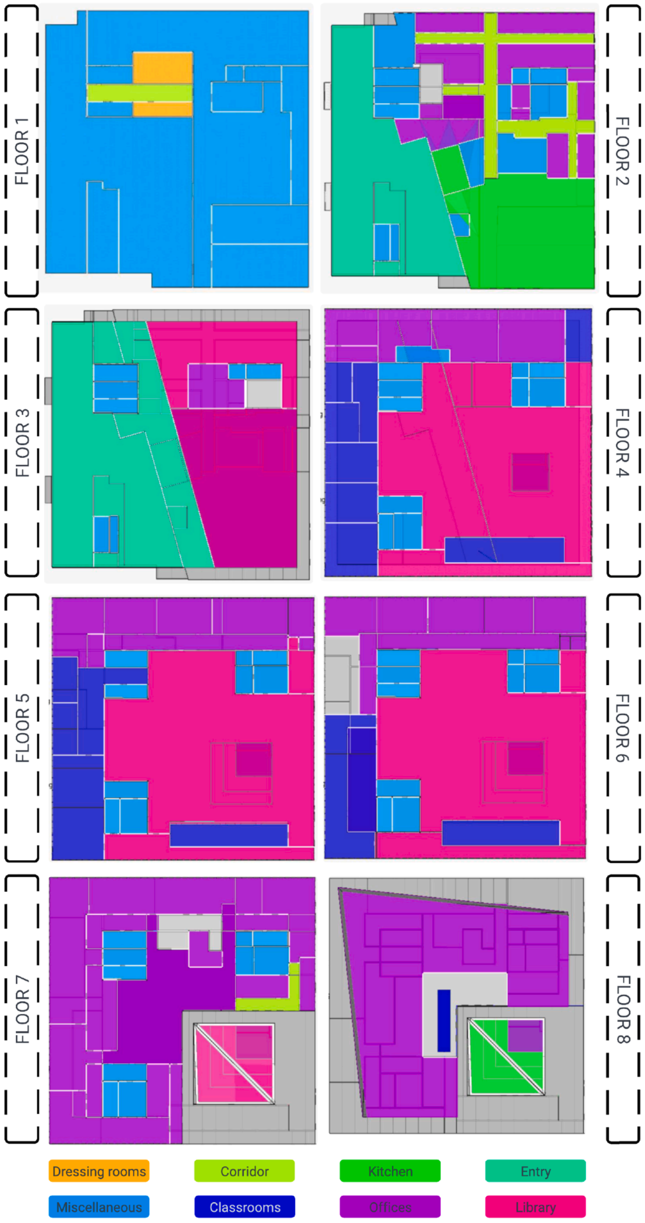
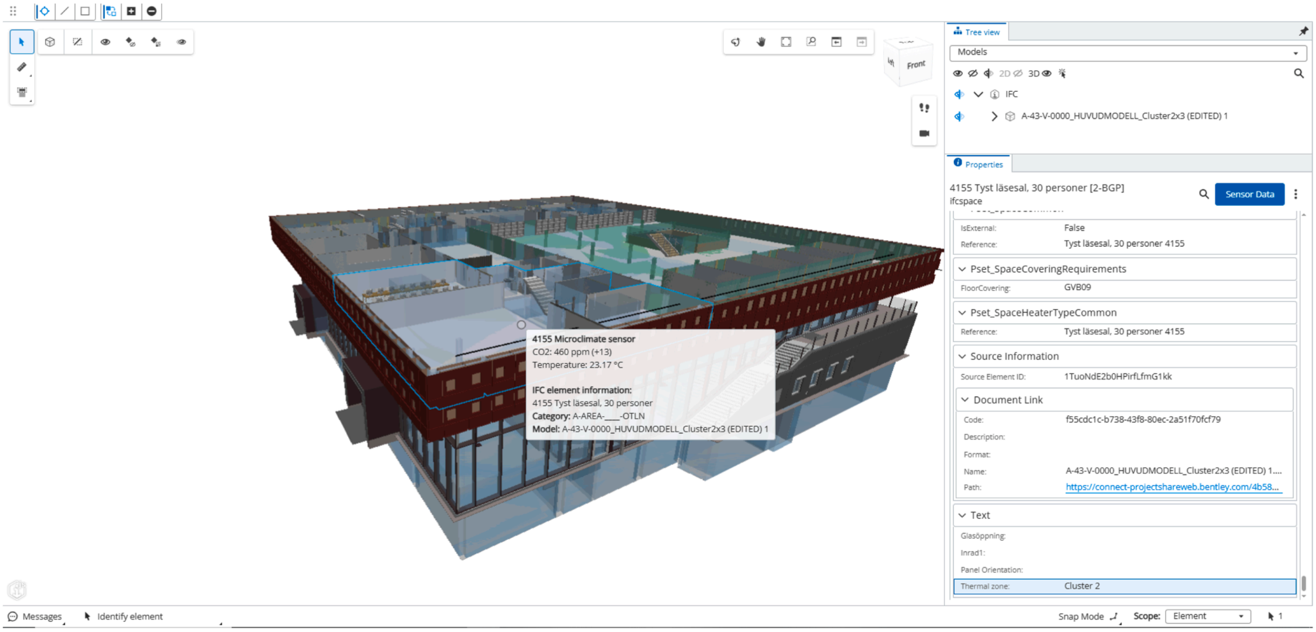


Fig. 29. Functional zones.

4.4.6. *Space function*

Finally, a common criterion used to define thermal zones is the space function. This criterion is often considered in the simulation software as it allows one to easily generalise the occupancy and scheduling criteria as well. Normally, the same function rooms have similar projected occupant density and schedules. For example, the functional zones for the case study building can be seen in Fig. 29. The zones in Fig. 29 were identified during the design stage of the building. Here, the building is divided into 8 zones depending on their function: (1) Dressing rooms, (2) Corridor, (3) Kitchen, (4) Entry, (5) Miscellaneous, (6) Classroom, (7) Office and (8) Library. In Table 2 they are listed with the respective area.

Considering the data-driven thermal zoning approach, there are some similarities between the zones defined by the proposed method and the functional zones. All data sets analysed resulted in the separate cluster for the large library areas (Figs. 15, 20 and 25) the same as it is also seen in the functional zones. In addition, the distinction is visible in most cases for office spaces and classrooms. However, depending on



A-43-V-0000_HUVUDMODELL_Cluster2x3 (EDITED) 1

4155 Microclimate sensor
CO2: 460 ppm (+13)
Temperature: 23.17 °C

IFC element information:
4155 Tyst läsesal, 30 personer
Category: A-AREA-____-OTLN
Model: A-43-V-0000_HUVUDMODELL_Cluster2x3 (EDITED) 1

Properties

4155 Tyst läsesal, 30 personer [2-BGP]
ifcspace

Sensor Data

IsExternal: False
Reference: Tyst läsesal, 30 personer 4155

Pset_SpaceCoveringRequirements
FloorCovering: GVB09

Pset_SpaceHeaterTypeCommon
Reference: Tyst läsesal, 30 personer 4155

Source Information
Source Element ID: 1TuoNdE2b0HPirfLfmG1kk

Document Link
Code: f55cdc1c-b738-43f8-80ec-2a51f70fcf79
Description:
Format:
Name: A-43-V-0000_HUVUDMODELL_Cluster2x3 (EDITED) 1....
Path: https://connect-projectshareweb.bentley.com/4b58...

Text
Glasöppning:
Inrad1:
Panel Orientation:
Thermal zone: Cluster 2

Fig. 30. Integrated Digital Twin.

Table 2
Area of building's thermal zones defined in IDA-ICE.

Zone type	Area, m^2
Classrooms	1319
Offices	3445
Library	4930
Kitchen	495
Corridor	553
Entry	1992
Miscellaneous	1750
Dressing rooms	74
Total	**14732**

the input parameters, the data-driven method allows us to distinguish rooms that perform differently than other rooms of the same function. At the same time, the functional approach places all offices in the same cluster, regardless of their location within the building. However, spaces adjacent to the façade or located on the upper/lower floors are likely to perform differently.

Almost all first floor spaces fall into the same "Miscellaneous" function cluster according to the functional zones. This is also likely to be too generalized, as on the 1st floor a large book archive is located where specific indoor air conditions are required to maintain the integrity of the books. For this, a separate air circulation unit is installed. However, according to the function, this space is grouped together with very different function spaces, such as HVAC equipment rooms on the 1st floor, stairs, and bathrooms located on all floors of the building.

Finally, it is important to examine how many zones are economically viable to distinguish and develop separate control models for. According to the function, there are 8 zones identified. However, it is important to maintain the balance between the effort and costs needed to develop and maintain the control mechanisms within the building and the benefits that come from them.

4.5. Case study building digital twin

Thermal zones defined based on real data were visualized in the DT environment (Fig. 30). The developed DT allows users to manipulate the 3D model for a customizable view. Through the UI, users can select colour-coded thermal zone clusters, distinguishing different spaces by zone colours. The cluster number is displayed in the right panel alongside other semantic data retrieved from the IFC file. When hovering over a selected space, an interactive window pops up, displaying the current CO2 concentration, its change since the last reading, indoor air temperature, room number, and occupant capacity. To access historical time-series data for a selected room, users can click the *"Sensor Data"* button in the right panel.

The current DT, operating at capability level 2, not only provides a strong foundation for advancing predictive capabilities at level 3 but also serves as a data aggregation tool for implementing the proposed methodology for thermal zoning qualitative validation as within the DT environment, users can inspect the qualitative criteria. Additionally, the defined thermal zones establish the granularity of dynamic thermal models, which can be further leveraged for advanced control mechanisms, ultimately progressing toward capability level 5 of autonomous control.

5. Conclusions and future work

5.1. Conclusions

This study investigated the current practices and gaps in defining building thermal zones, emphasizing the lack of quantitative methods. To address this, a novel data-driven approach was proposed, integrating Principal Component Analysis (PCA) and k-means clustering to define and validate thermal zones based on real sensor data. Additionally, a Digital Twin (DT) was developed to visualize, interact with, and qualitatively validate the defined zones using the proposed criteria. They key contributions of this study:

- Proposed a data-driven thermal zoning algorithm, consisting of data cleaning, PCA-based dimensionality reduction, and k-means clustering, enhanced with the elbow method for optimal cluster selection.
- Introduced a quantitative validation algorithm to assess the consistency of the defined data-driven thermal zones across different datasets.
- Developed a DT environment to visualize the thermal zoning results, integrating 3D building geometry, semantic data, weather information, and IoT sensor data into a single platform.

The key findings of the study are summarized as follows:

- Exploratory analysis of CO2 and indoor air temperature data from the case study building revealed that existing rule-based control (RBC) strategies often fail to maintain indoor thermal comfort requirements.
- PCA proved to be a suitable method for defining thermal zones, explaining over 85 % of the variance in all three tested datasets. Additionally, it enables the inclusion of multiple parameters to refine the zoning process.
- Applying k-means clustering after PCA-based dimensionality reduction allowed for the identification of similarly performing clusters. The elbow method, combined with statistical calculations of slope and intersection, facilitated the optimal selection of clusters.
- The quantitative validation approach confirmed a consistency score of over 91 %, depending on the selected number of principal components (PCs), clusters, and input parameters.
- Qualitative validation demonstrated that different parameters contribute to different aspects of thermal zoning: CO2 data better captured occupancy and schedules, while temperature data more effectively reflected envelope exposure and orientation. In general, all qualitative criteria were reflected in thermal zones defined by the proposed data-driven algorithm.
- DT was implemented to provide a centralized platform for thermal zone visualization and validation. The DT integrates 3D building geometry, semantic information, real-time weather data, and IoT sensor readings, enabling users to interactively explore zone distributions and environmental conditions.

The proposed method provides a structured, data-driven approach to thermal zoning and validation, supporting future development of advanced control strategies. The current DT implementation (Capability Level 2) serves as a foundation for further development toward Capability Level 5 and closing the control loop.

5.2. Future work

In future work, additional parameters could be incorporated into the thermal zone definition process to enhance both consistency score and robustness. For example, parameters such as indoor noise levels, lighting conditions, or movement detection could improve the reliability of occupancy criteria. Similarly, outdoor environmental factors, including wind speed, solar irradiation, temperature, humidity, and CO2 levels, could be integrated into the quantitative evaluation of thermal zoning. Furthermore, to demonstrate the benefits of the data-driven approach for control applications, the study should be expanded to test the defined thermal zones by developing zone-specific dynamic thermal models. Finally, the DT should be further developed, integrating dynamic thermal models and predictive components.

Limitations

This study has several key limitations. First, the methodology was applied to a single building, and future studies should explore its application across various building types, which may have different HVAC systems, ventilation rates, and usage patterns. The experimental design relied on controlled test cases, such as varying ventilation rates, which may limit the approach's flexibility in real-world settings where such controlled conditions are not feasible. A comparative study in environments where test data can be collected under less controlled conditions could provide new insights. Additionally, while the CO2 sensors in the VAV system employ Automatic Baseline Correction (ABC), a test case with regular sensor calibration could ensure more reliable data. Furthermore, data were not available for all rooms in the building, which may have led to fewer identified thermal zones; with more comprehensive data, the number of zones might have increased. Finally, as the case study building is located in a cold-climate region, indoor air humidity was not considered.

Data availability

The data that has been used is confidential.

CRediT authorship contribution statement

Lina Morkunaite: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization; **Adil Rasheed:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization; **Darius Pupeikis:** Writing – review & editing, Validation, Supervision, Software, Conceptualization; **Vangelis Angelakis:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition; **Tobias Davidsson:** Writing – review & editing, Resources, Project administration, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study is part of the dissemination activities of the research projects “Demonstrating Holistic Data-driven Co-Creative Approaches in Nature-Based Solutions towards Climate Adaptation and Mitigation (GreenIn Cities)” (Grant ID Number 101139730), funded by Horizon Europe call HORIZON-MISS-2023-CLIMA-CITIES-01 and “Twin AIR - Digital Twins Enabled Indoor Air Quality Management for Healthy Living,” funded by Horizon Europe Research and Innovation Programme GA no. 101057779, HORIZON-HLTH-2021-ENVHLTH-02.

References

[1] International Energy Agency, Buildings: a source of enormous untapped efficiency potential, 2024, Accessed: March 13, 2025, https://www.iea.org/energy-system/buildings.

[2] European Commission, The European Green Deal, 2024, Accessed: 2024-10-06, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/story-von-der-leyen-commission/european-green-deal_en.

[3] Y. Song, M. Xia, Q. Chen, F. Chen, A data-model fusion dispatch strategy for the building energy flexibility based on the digital twin, Appl. Energy 332 (2023) 120496. https://doi.org/10.1016/j.apenergy.2022.120496

[4] Electricity Maps, Electricity maps - live CO2 emissions & electricity mix, 2025, Accessed: 25 Feb. 2025, https://app.electricitymaps.com/zone/SE-SE3/72h/hourly/2025-02-23T08:00:00.000Z.

[5] J.C. Ketterer, The impact of wind power generation on the electricity price in Germany, Energy Econ. 44 (2014) 270–280. https://doi.org/10.1016/j.eneco.2014.04.003

[6] A. Rauf, D.E. Attoye, R. Crawford, Embodied and operational energy of a case study villa in UAE with sensitivity analysis, Buildings 12 (9) (2022) 1469. https://doi.org/10.3390/buildings12091469

[7] Y. Bae, S. Bhattacharya, B. Cui, S. Lee, Y. Li, L. Zhang, P. Im, V. Adetola, D. Vrabie, M. Leach, T. Kuruganti, Sensor impacts on building and HVAC controls: a critical review for building energy performance, Adv. Appl. Energy 4 (2021) 100068. https://doi.org/10.1016/j.adapen.2021.100068

[8] S.H. Khajavi, N.H. Motlagh, A. Jaribion, L.C. Werner, J. Holmström, Digital twin: vision, benefits, boundaries, and creation for buildings, IEEE Access 7 (2019) 147406–147419. https://doi.org/10.1109/ACCESS.2019.2946515

[9] S. Yang, M.P. Wan, W. Chen, B.F. Ng, S. Dubey, Model predictive control with adaptive machine-learning-based model for building energy efficiency and comfort optimization, Appl. Energy 271 (2020) 115147. https://doi.org/10.1016/j.apenergy.2020.115147

[10] A. Afram, F. Janabi-Sharifi, Review of modeling methods for HVAC systems, Appl. Therm. Eng. 67 (1) (2014) 507–519. https://doi.org/10.1016/j.applthermaleng.2014.03.055

[11] L. Vandenbogaerde, S. Verbeke, A. Audenaert, Optimizing building energy consumption in office buildings: a review of building automation and control systems and factors influencing energy savings, J. Build. Eng. 76 (2023) 107233. https://doi.org/10.1016/j.jobe.2023.107233

[12] G. Serale, M. Fiorentini, A. Capozzoli, D. Bernardini, A. Bemporad, Model predictive control (MPC) for enhancing building and HVAC system energy efficiency: problem formulation, applications and opportunities, Energies 11 (3) (2018) 631. https://doi.org/10.3390/en11030631

[13] I. Ajifowowe, H. Chang, C.S. Lee, S. Chang, Prospects and challenges of reinforcement learning- based HVAC control, J. Build. Eng. 98 (2024) 111080. https://doi.org/10.1016/j.jobe.2024.111080

[14] E. Khanmirza, A. Esmaeilzadeh, A.H.D. Markazi, Predictive control of a building hybrid heating system for energy cost reduction, Appl. Soft Comput. 46 (2016) 407–423. https://doi.org/10.1016/j.asoc.2016.03.008

[15] R. Halvgaard, N.K. Poulsen, H. Madsen, J.B. Jørgensen, Economic model predictive control for building climate control in a smart grid, in: 2012 IEEE PES Innovative Smart Grid Technologies (ISGT), IEEE, 2012, pp. 1–6. https://doi.org/10.1109/ISGT.2012.6175653

[16] Y. Gao, S. Miyata, Y. Akashi, Energy saving and indoor temperature control for an office building using tube-based robust model predictive control, Appl. Energy 341 (2023) 121106. https://doi.org/10.1016/j.apenergy.2023.121106

[17] V.A. Arowoiya, R.C. Moehler, Y. Fang, Digital twin technology for thermal comfort and energy efficiency in buildings: a state-of-the-art and future directions, Energy Built Environ. 5 (5) (2024) 641–656. https://doi.org/10.1016/j.enbenv.2023.05.004

[18] A. Rasheed, O. San, T. Kvamsdal, Digital twin: values, challenges and enablers from a modeling perspective, IEEE Access 8 (2020) 21980–22012. https://doi.org/10.1109/ACCESS.2020.2970143

[19] R. Bortolini, R. Rodrigues, H. Alavi, L.F.D. Vecchia, N. Forcada, Digital twins applications for building energy efficiency: a review, Energies 15 (19) (2022) 7002. https://doi.org/10.3390/en15197002

[20] Y. Balali, A. Chong, A. Busch, S. OKeefe, Energy modelling and control of building heating and cooling systems with data-driven and hybrid models–a review, Renew. Sustain. Energy Rev. 183 (2023) 113496. https://doi.org/10.1016/j.rser.2023.113496

[21] E. Atam, L. Helsen, Control-oriented thermal modeling of multizone buildings: methods and issues: intelligent control of a building system, IEEE Control Syst. Mag. 36 (3) (2016) 86–111. https://doi.org/10.1109/MCS.2016.2535913

[22] M. Shin, J.S. Haberl, Thermal zoning for building HVAC design and energy simulation: a literature review, Energy Build. 203 (2019) 109429. https://doi.org/10.1016/j.enbuild.2019.109429

[23] D. Mazzeo, N. Matera, C. Cornaro, G. Oliveti, P. Romagnoli, L. De Santoli, Energyplus, IDA ICE and TRNSYS predictive simulation accuracy for building thermal behaviour evaluation by using an experimental campaign in solar test boxes with and without a PCM module, Energy Build. 212 (2020) 109812. https://doi.org/10.1016/j.enbuild.2020.109812

[24] F.S. Javier Arroyo, L. Helsen, Identification of multi-zone grey-box building models for use in model predictive control, J. Build. Perform. Simul. 13 (4) (2020) 472–486. https://doi.org/10.1080/19401493.2020.1770861

[25] ASHRAE, ANSI/ASHRAE/IESNA 90.1-2022, Energy Standard for Buildings Except Low-Rise Residential Buildings, 2022.

[26] C.R. Timur Dogan, P. Michalatos, Autozoner: an algorithm for automatic thermal zoning of buildings with unknown interior space definitions, J. Build. Perform. Simul. 9 (2) (2016) 176–189. https://doi.org/10.1080/19401493.2015.1006527

[27] B. Rajasekhar, W. Tushar, C. Lork, Y. Zhou, C. Yuen, N.M. Pindoriya, K.L. Wood, A survey of computational intelligence techniques for air-Conditioners energy management, IEEE Trans. Emerg. Top. Comput. Intell. 4 (4) (2020) 555–570. https://doi.org/10.1109/TETCI.2020.2991728

[28] M. Shin, J.S. Haberl, A procedure for automating thermal zoning for building energy simulation, J. Build. Eng. 46 (2022) 103780. https://doi.org/10.1016/j.jobe.2021.103780

[29] L. Wan, F. Rossa, T. Welfonder, E. Petrova, P. Pauwels, Enabling scalable model predictive control design for building HVAC systems using semantic data modelling, Autom. Constr. 170 (2025) 105929. https://doi.org/10.1016/j.autcon.2024.105929

[30] J. Drgoa, J. Arroyo, I. Cupeiro Figueroa, D. Blum, K. Arendt, D. Kim, E.P. Ollé, J. Oravec, M. Wetter, D.L. Vrabie, L. Helsen, All you need to know about model predictive control for buildings, Annu. Rev. Control 50 (2020) 190–232. https://doi.org/10.1016/j.arcontrol.2020.09.001

[31] D. Ruch, L. Chen, J.S. Haberl, D.E. Claridge, A change-point principal component analysis (CP/PCA) method for predicting energy usage in commercial buildings: the PCA model, J. Sol. Energy Eng. 115 (2) (1993) 77–84. https://doi.org/10.1115/1.2930035

[32] E. Wang, Z. Shen, K. Grosskopf, Benchmarking energy performance of building envelopes through a selective residual-clustering approach using high dimensional dataset, Energy Build. 75 (2014) 10–22. https://doi.org/10.1016/j.enbuild.2013.12.055

[33] R. Platon, V.R. Dehkordi, J. Martel, Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis, Energy Build. 92 (2015) 10–18. https://doi.org/10.1016/j.enbuild.2015.01.047

[34] T. Parhizkar, E. Rafieipour, A. Parhizkar, Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction, J. Clean. Prod. 279 (2021) 123866. https://doi.org/10.1016/j.jclepro.2020.123866

[35] A. Wickramasinghe, S. Muthukumarana, D. Loewen, M. Schaubroeck, Temperature clusters in commercial buildings using k-means and time series clustering, Energy Inf. 5 (1) (2022) 1. https://doi.org/10.1186/s42162-022-00186-8

[36] X. Deng, Z. Tan, M. Tan, W. Chen, A clustering-based climatic zoning method for office buildings in China, J. Build. Eng. 42 (2021) 102778. https://doi.org/10.1016/j.jobe.2021.102778

[37] J. Rodriguez, N. Fumo, Zoned heating, ventilation, and air-conditioning residential systems: a systematic review, J. Build. Eng. 43 (2021) 102925. https://doi.org/10.1016/j.jobe.2021.102925

[38] O. San, A. Rasheed, T. Kvamsdal, Hybrid analysis and modeling, eclecticism, and multifidelity computing toward digital twin revolution, GAMM-Mitteilungen 44 (2) (2021) e202100007. https://doi.org/10.1002/gamm.202100007

[39] A. Clausen, K. Arendt, A. Johansen, F.C. Sangogboye, M.B. Kjærgaard, C.T. Veje, B.N. Jørgensen, A digital twin framework for improving energy efficiency and occupant comfort in public and commercial buildings, Energy Inf. 4 (S2) (2021) 40. https://doi.org/10.1186/s42162-021-00153-9

[40] C. Vering, P. Mehrfeld, M. Nürenberg, D. Coakley, M. Lauster, D. Müller, Unlocking potentials of building energy systems operational efficiency: application of digital twin design for HVAC systems, in: 16th International Building Performance Simulation Association (IBPSA), University College Dublin, 2019, pp. 1304–1310.

[41] DNV, DNV-RP-A204: assurance of digital twins, 2024, Accessed: 2024-10-06, https://www.dnv.com/oilgas/download/dnv-rp-a204-assurance-of-digital-twins/.

[42] F.F. Hernández, J.M.P. Suárez, J.A.B. Cantalejo, M.C.G. Muriano, Impact of zoning heating and air conditioning control systems in users comfort and energy efficiency in residential buildings, Energy Convers. Manag. 267 (2022) 115954. https://doi.org/10.1016/j.enconman.2022.115954

[43] D. Blum, Z. Wang, C. Weyandt, D. Kim, M. Wetter, T. Hong, M.A. Piette, Field demonstration and implementation analysis of model predictive control in an office HVAC system, Appl. Energy 318 (2022) 119104. https://doi.org/10.1016/j.apenergy.2022.119104

[44] M.F. Ibrahim, M. Mohamed, B.H. Far, Measuring the effectiveness of zonal heating control for energy saving, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2016, pp. 000132–000136. https://doi.org/10.1109/SMC.2016.7844231

[45] V. Kumar, R. Kumar, D. Patkar, A.S. Bopardikar, A method to identify dynamic zones for efficient control of HVAC systems, in: 2017 IEEE International Symposium on Circuits and Systems (ISCAS), 2017, pp. 1–4. https://doi.org/10.1109/ISCAS.2017.8050507

[46] J. Reynolds, Y. Rezgui, A. Kwan, S. Piriou, A zone-level, building energy optimisation combining an artificial neural network, a genetic algorithm, and model predictive control, Energy 151 (2018) 729–739. https://doi.org/10.1016/j.energy.2018.03.113

[47] W.T. Grondzik, Mechanical and Electrical Equipment for Buildings, Wiley, Hoboken [N.J.], 11th ed., 2010.

[48] H.E. Bovay, Jr, Handbook of Mechanical and Electrical Systems for Buildings, McGraw-Hill, Inc., New York, NY, United States, United States, 1981. https://www.osti.gov/biblio/6166280.

[49] S. Alsaadani, Thermal Zoning in Speculative Office Buildings: Discussing the Connections between Space Layout and Inside Thermal Control, 2012.

[50] Chartered Institution of Building Services Engineers, Applications manual 11: building performance modelling, 2015, Accessed: 2024-10-06, https://www.cibse.org/knowledge-research/knowledge-portal/applications-manual-11-building-performance-modelling-2015.

[51] L.R. Bachman, Integrated Buildings: The Systems Basis of Architecture, John Wiley & Sons, 2004.

[52] R. McDowall, Fundamentals of HVAC Systems: SI Edition, Academic Press, 2007.

[53] W.T. Grondzik, A.G. Kwok, Mechanical and Electrical Equipment for Buildings, John Wiley & Sons, 2019.

[54] G. Strang, Linear Algebra and Learning from Data, SIAM, 2019.

[55] P.D. Harrison, Statistics for Machine Learning: Techniques for Exploring Supervised, Unsupervised, and Reinforcement Learning Models with Python and R, Packt Publishing, Birmingham, UK, 1st edition, Birmingham, UK, 2018. Available through O'Reilly Media, https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/.

[56] J.C. Lam, K.K.W. Wan, T.N.T. Lam, S.L. Wong, An analysis of future building energy use in subtropical Hong Kong, Energy 35 (3) (2010) 1482–1490.

[57] Oikolab, Oikolab weather downloader, 2024, Accessed: 2024-10-06, https://weatherdownloader.oikolab.com/app.