MSc thesis in Building Technology

Machine Learning-Assisted Analysis of Energy Consumption Profiles and Efficiency in Uilenstede Campus Buildings

– An Analysis of User Energy Profile Patterns and Clustering of Uilenstede Campus Energy Consumption Data

Yu Hsiu Tung

June 2023

A thesis submitted to the Delft University of Technology in fulfillment of the requirements for the degree of Master of Science in Building Technology

Yu Hsiu Tung: *Machine Learning-Assisted Analysis of Energy Consumption Profiles and Efficiency in Uilenstede Campus Buildings - An Analysis of User Energy Profile Patterns and Clustering of Uilenstede Campus Energy Consumption Data* (2023)

The work in this thesis was carried out in the:

Building Technology
Delft University of Technology

Research Data set provided by DUWO

Supervisors: Prof.dr. Laure Itard
Dr. Charalampos P. Andriotis

Abstract

The study of energy consumption across various building clusters offers a path to discerning intricate patterns and establishing energy efficiency metrics. However, these analyses have mostly been limited to small, controlled settings, leaving a vast potential for broader application in energy efficiency management and classification untapped. This research leverages machine learning models to determine gas consumption patterns and energy efficiency characteristics of buildings in real-life settings, based on a range of parameters including insulation properties and year of construction. The developed system was applied to a comprehensive dataset comprising eight distinct clusters of buildings, with a total of nearly 10,000 hourly gas consumption. To supplement the analysis, additional data was gathered concerning the building features. The findings indicate that the average gas consumption varies significantly across clusters, with dependencies shown for the age of the building, insulation characteristics, and building orientation The developed framework proved to be suitable for gaining insights into average gas consumption and usage patterns at a building level, non-intrusively and on a large scale. The additional data provided comparative insights between different building groups. The developed system can be easily expanded for other building characteristics and could be used to drive tailored feedback on energy efficiency improvements within buildings. This research paves the way for a more comprehensive approach to building energy efficiency, one that goes beyond the traditional parameters to include a broader set of variables such as building usage, occupant behavior, and heating system efficiency.

Keywords: smart meter data, machine learning, energy conservation, building energy management, anomaly detection, and end-user energy profiles.

Acknowledgements

Firstly, I would like to express my sincere gratitude to everyone who has been part of this journey. Each of you, in your unique ways, have contributed to this thesis.

My utmost appreciation goes to my supervisors, Prof. Dr. Laure Itard and Dr. Charalampos P. Andriotis. Their continual support, encouragement and expert insights have been instrumental in the completion of this thesis. Their commitment towards high standards of research is truly inspiring and has guided me to grow both professionally and personally.

I am deeply grateful for their immense knowledge, patience, and most importantly, their belief in me. This thesis would not have reached its present form without their enduring guidance and persistent help.

In the process of writing this thesis, I have been fortunate to be part of an academic community that values cooperation, creativity, and critical inquiry. I look forward to continuing to learn and grow in this vibrant scholarly environment.

Lastly, but not least, I wish to thank my family and friends for their unconditional love, understanding, and encouragement throughout my academic journey. Your faith in my capabilities has been a source of motivation and has pushed me to strive harder.

Thank you all, once again.

Contents

List of Figures

List of Tables

List of Algorithms

1. Introduction

Buildings represent a large portion of the world's energy consumption and associated CO2 emissions, contributing to approximately one-third of emissions and accounting for 30% of energy usage worldwide [35]. In response to the increasingly urgent need to combat global warming, energy conservation in buildings has become a central focus in EU policies and environmental initiatives. This emphasis on building efficiency has been driven by the rise in environmental awareness and the implementation of legislation to optimize building consumption for improved performance. The recognition of buildings' substantial impact on energy consumption and emissions has led to a concerted effort to address energy efficiency challenges. Significant reductions in greenhouse gas emissions can be achieved by implementing measures to optimize energy use within buildings. Such measures encompass many strategies, including adopting energy-efficient appliances and systems, improving insulation, and implementing intelligent building technologies.

In recent years, there has been a shift towards employing machine learning and data-driven approaches to address building energy conservation challenges. These approaches aim to capture the intrinsic behavior of building energy consumption by analyzing time series data and developing intelligent systems capable of understanding and optimizing energy usage patterns. By leveraging smart meter data and end-user energy profiles, these methods offer opportunities for gaining deeper insights into energy consumption dynamics on a seasonal and hourly basis.

This thesis focuses on developing a machine learning-driven framework designed to support energy conservation initiatives in the transition towards a carbon-neutral district at DUWO's Uilenstede campus. The framework utilizes smart meter data and end-user energy profiles to comprehend the energy usage patterns of building occupants. The research aims to inform strategic decision-making processes for building management and integrating renewable energy systems within the campus by establishing comprehensive user profiles and generating insights into energy consumption. Ultimately, the goal is to facilitate the campus's transition to a sustainable, carbon-free environment. The central research question guiding this study is:

"How can energy consumption profiles guide energy conservation strategies at Uilenstede campus?"

By addressing this question, the research aims to contribute to the advancement of energy conservation strategies and support the transition to a more sustainable future.

1.1. Problem outline

DUWO's Uilenstede campus, a prominent student housing complex in the Netherlands, is an ideal research setting due to its various building types and distinct energy consumption patterns. This intricate landscape offers a unique opportunity to conduct comprehensive energy consumption analysis and devise strategic energy-saving measures. Mitigating the impacts of climate change critically hinges upon enhancing building energy efficiency. Utilizing smart meter data and machine learning methods, this research aims to dive into the energy consumption patterns within the campus buildings, facilitating the detection of potential conservation zones.

1.2. Motivation

Understanding and controlling energy usage is essential for achieving sustainability goals.[47]. Buildings contribute significantly to global carbon emissions, and as such, enhancing their energy efficiency is a critical component in the quest for carbon neutrality.

The analysis of building meter data provides invaluable insights into energy usage patterns over specific periods. When associated with building features, this information can help identify potential energy conservation and efficiency improvement areas.

This research aims to improve energy efficiency in buildings by providing actionable insights through data analysis and machine learning. The goal is to contribute to creating more sustainable and energy-efficient built environments.

1.3. Knowledge gap

Despite significant advancements in smart meter technology and machine learning techniques, a substantial knowledge gap remains in applying machine learning algorithms and building data management to formulate energy conservation strategies at the district level[23]. This gap is particularly noticeable in complex environments such as DUWO's Uilenstede campus, where many building meters are operational and a diverse array of consumption data is involved.

The current comprehension and deployment of data analytics in this domain have yet to extend to energy conservation efforts fully. Thus, this study aims to bridge this knowledge gap, demonstrating how machine learning can be applied to smart meter data to enhance energy conservation strategies.

1.4. Research Objective

The primary objective of this research is to use machine learning techniques to discern energy consumption patterns in DUWO's Uilenstede campus. While significant work has been done on applying machine learning to build consumption data, a gap remains in its practical application for enhancing efficiency analysis in real-world scenarios.

This research aims to bridge that gap by developing end-user energy usage patterns and identifying the efficiency characteristics of a large-scale student campus. By integrating load disaggregation research with gas consumption data for heating systems, we aim to guide energy conservation efforts toward achieving a carbon-neutral district.

The insights from this research will provide actionable recommendations for decision-makers to implement effective energy conservation measures. This will improve energy efficiency and contribute to larger sustainability goals within the built environment.

1.4.1. Research questions

The main research question of this study is:

> **"How can the application of machine learning to smart meter data and end-user energy profiles contribute to energy conservation efforts in DUWO's Uilenstede campus?""**

To comprehensively address this overarching question, several sub-questions have been formulated:

1. Can the utilization of machine learning methods on energy consumption data offer additional insights into energy usage patterns?
2. What impact do different building characteristics, such as insulation and heating systems, have on the energy efficiency of the buildings?

2. Literature review

The literature review primarily focuses on applying Machine Learning methods for building energy profiling. Relevant research papers were accumulated using search terms including 'energy profiling,' 'clustering smart meter data for building energy,' and 'building energy usage pattern analysis.'

The literature review is further divided into several sections, each reflecting a topic pertinent to this research. These sections include:

1. Feature Extraction Techniques in Building Energy Data Analysis
2. Machine Learning Data Analytics in Building Energy Management
3. Clustering Analysis in Building Energy Pattern Recognition

Each section will delve into existing literature and research findings, shedding light on the current state of knowledge and identifying potential avenues for further exploration.

2.1. Feature Extraction Techniques in Building Energy Data Analysis

Feature extraction techniques play an essential role in energy data analysis, particularly in the context of forecasting energy consumption in buildings. They allow for dimensionality reduction, enabling the models to focus on the most crucial aspects of the data that significantly contribute to the prediction task.

There are numerous feature extraction techniques used for time series data, a few of which are outlined below [20]:

- Fourier Transform: The Fourier Transform is a method that transforms time series data from the time domain into the frequency domain. This can help to identify periodic components or trends in the data.
- Wavelet Transform: Similar to the Fourier Transform, the Wavelet Transform also transforms time series data into a different domain. However, unlike the Fourier Transform, the Wavelet Transform maintains temporal information, which can help identify trends over time.
- Autoregressive Integrated Moving Average (ARIMA) Models: ARIMA models can be used to forecast future values in a time series. The coefficients of these models can be used as features in a machine learning model.

- Autoencoders: Autoencoders are a type of neural network that are used to learn efficient encodings of input data. They can be used to compress time series data into a lower-dimensional representation, which can then be used as features for a machine learning model.

- Statistical Features: Simple statistical features such as the mean, median, standard deviation, skewness, and kurtosis of a time series can also be used as features for a machine learning model.

- Trend and Seasonality: Time series often contain trends (long-term increase or decrease) and seasonality (patterns that repeat over known, fixed periods of time). These can be extracted and used as features.

- Principal Component Analysis (PCA): PCA is a dimensionality reduction technique that can be used to reduce a large set of variables (such as time series data) to a small set that still contains most of the information in the large set.

- Segmentation and Feature Extraction: Time series data can be segmented into smaller chunks and features can be extracted from each segment. This is particularly useful when working with long time series.

Bode et al. (2019) apply an unsupervised feature extraction technique using deep convolutional auto-encoders, paired with statistical analysis of time series data. The goal is to identify and learn recurrent patterns within the data. The auto-encoder, an AI tool, first compresses the data into a lower dimensionality and then attempts to reconstruct the input, essentially learning the underlying patterns. Their approach results in effective feature extraction from time series data, crucial for later analysis or predictive modeling[12].

In the study by Zhan et al. (2021), Principal Component Analysis (PCA) was used for feature extraction and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied to generate distinct clusters, increasing between-cluster variance and decreasing within-cluster variation. These results were then used to enhance the modeling of variables using linear regression [5].

On the other hand, Xiao et al. (2020) used unsupervised learning to reduce the dimensionality of building measurement data before applying the PCA, Autoencoder, and t-Distributed Stochastic Neighbor Embedding (t-SNE). for feature extraction [69].

2.2. Similarity Measure Approaches

The field of data analysis employs numerous similarity measure approaches to determine the association between different datasets or patterns. Table 2.1 examines several methodologies, highlighting their defining characteristics and definitions.

In time-series analysis, the role of the distance measure is highly critical, particularly for clustering. A range of distance measures is employed to gauge the proximity between time series. Some of these measures are devised to be compatible with specific time-series representations.

However, other measures operate independently of the representation method and are compatible with raw time-series data. In conventional static object clustering, the distance measure is exact. In contrast, the measurement of distance in time-series clustering is often approximated. This subtle shift is due to the inherently dynamic nature of time-series data.

When comparing time series with varying sampling intervals and lengths, determining an adequate measure of time-series similarity becomes crucial. Given its irregular sampling intervals and diverse lengths, this is particularly important for time-series data. Accurately measuring similarity can significantly influence the effectiveness of time-series clustering and other related data analysis tasks.

Table 2.1.: Similarity measure approaches in the literature [7]

Distance Measure	**Characteristics**	**Defined By**
Dynamic Time Warping (DTW)	Elastic Measure; excels in temporal drift	Shape-based [52, 76]
Pearson's Correlation Coefficient	Invariant to scale and location	Compression-based [11]
Euclidean Distance (ED)	Sensitive to scaling; used in indexing, clustering, classification	Shape-based [7, 11]
Hidden Markov Models (HMM)	Captures dependencies between variables and serial correlation	Model based [20]
Autocorrelation	Measures the similarity between observations as a function of time lag	Compression-based [29]
Cosine Wavelets	Uses wavelet transformation to measure similarity	Compression-based [27]
KL Distance	Non-symmetric measure of the difference between two probability distributions	Compression-based [36]
Edit Distance with Real Penalty (ERP)	Robust to noise, shifts, scaling; uses a constant reference point	Shape-based [43]
Edit Distance on Real sequence (EDR)	Elastic measure; uses a threshold pattern	Shape-based [43]
Cross-Correlation	Measures similarity considering a time-lag, useful in signal processing	Shape-based [54]

2.3. Clustering approaches

Based on the literature research, time-series clustering algorithms can be categorized into several groups: Hierarchical, Partitioning, Model-based, Density-based, Probability-based and more.

- Hierarchical Clustering[57]: This clustering method builds a hierarchy of clusters using either agglomerative (bottom-up) or divisive (top-down) algorithms. It doesn't require the number of clusters as an initial parameter, which is a significant advantage in time-series clustering, where defining the number of clusters in real-world problems can be challenging. However, this algorithm can't adjust clusters after splitting or merging, often necessitating a combination with another algorithm to address this issue.

- Partitioning Clustering [34]: Partitioning clustering, such as k-Means, forms 'k' groups from 'n' unlabelled objects. It ensures each group contains at least one object. The method involves minimizing the total distance between all objects in a cluster from their cluster center or prototype.
- Model-based Clustering [55]: This method attempts to recover the original model from a set of data. It presumes the existence of some randomly chosen centroids to which noise is added with a normal distribution. The recovered model from the generated data defines clusters. Methods used are usually statistical or neural network-based, like COBWEB, ART, or Self-Organization Map.
- Density-based Clustering[7]: In density-based clustering, clusters are dense subspaces of objects separated by subspaces with a low density of objects. Algorithms like DB-SCAN and OPTICS operate based on this concept, where clusters are expanded if their neighbors are dense.
- Probabilistic-based Clustering [42, 57]: Probabilistic mixture models such as Gaussian Mixture (GM) models are comprised of various multivariate normal density components, each having its own mean, covariance matrix, and cluster size. GM works well with clusters that aren't roundly shaped or overlapped, as it assigns samples to clusters based on specific probabilities. After grouping the data into a set of multivariate normal density components, the mixture component controls the cluster's volume, orientation, and shape.

Some clustering algorithms use raw time-series data, while others implement reduction methods before clustering. In practice, the application of each group varies depending on the specific needs of the time-series clustering task.

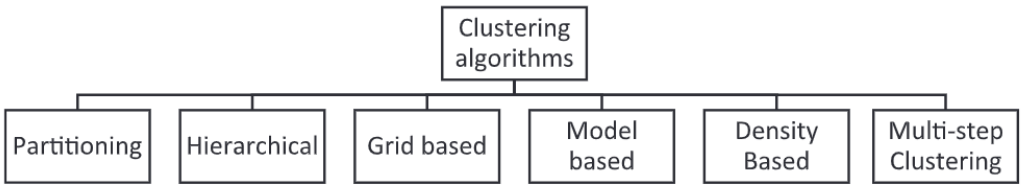


Figure 2.1.: Types of clustering approaches [7]

2.4. Machine Learning and Data Analytics in Building Energy Analysis

Machine learning techniques form the foundation of building energy prediction and classification models. They are mainly bifurcated into two primary categories: Prediction and Classification, as visualized in figure 2.2.

Several techniques are employed to predict and forecast energy consumption patterns in the' Prediction' category. These techniques include Artificial Neural Networks, which simulate the way the human brain works to 'learn' from past examples; Support Vector Machines, which are mainly used for regression and classification tasks; Statistical Regression, which is used to predict the outcome of a response variable based on one or more predictor variables;

Decision Trees, which utilize a tree-like model of decisions; and Genetic Algorithms, which are search-based algorithms inspired by the process of natural selection.

On the other hand, the 'Classification' category includes Clustering and Self-organizing Maps. Clustering methods, such as K-means and Hierarchical clustering, are unsupervised learning techniques that divide the data into groups (or clusters) based on similarities. Meanwhile, Self-organizing Maps is an artificial neural networks trained using unsupervised learning to produce low-dimensional, discretized representations of the input space, often used for classification.

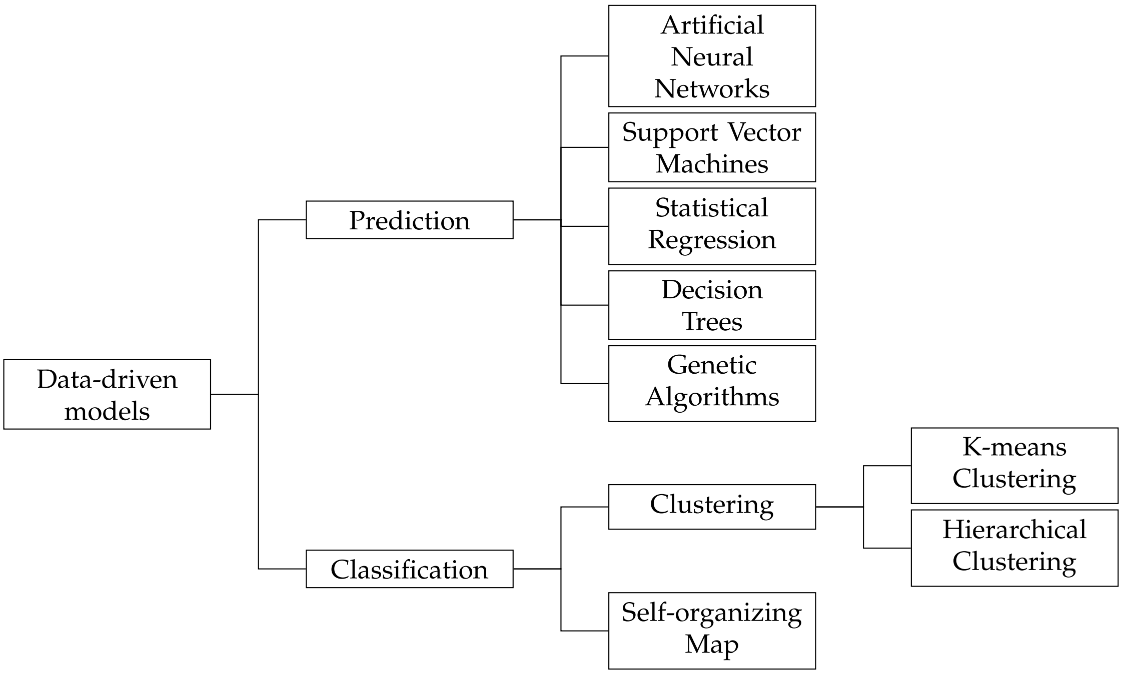


Figure 2.2.: Machine Learning Techniques used in Building Energy Analysis

2.4.1. Artificial Neural Network (ANN) in Building Energy Consumption

An Artificial Neural Network (ANN) is a machine learning methodology that takes inspiration from the neural networks found in the human brain. It involves interconnected processing elements, called neurons, which work together to solve specific problems. ANNs can process complex patterns and relationships between data, making them valuable for predicting events and making decisions[67].

Karatasou et al. [37] showcased the influential impact of choosing artificial neural networks (ANNs) alongside statistical techniques for optimizing building energy consumption models and predictions. Deep Reinforcement Learning (DRL), a blend of Deep Learning and Reinforcement Learning, is another variant of artificial neural networks. According to a review by Yu et al. (2021) [72], DRL has been used to predict the level of discomfort, thereby facilitating decision-making processes in Smart Building Energy Management (SBEM).

Originally developed for image processing, Convolutional Neural Networks (CNNs) have demonstrated remarkable versatility, extending their utility to time-series problems [41].

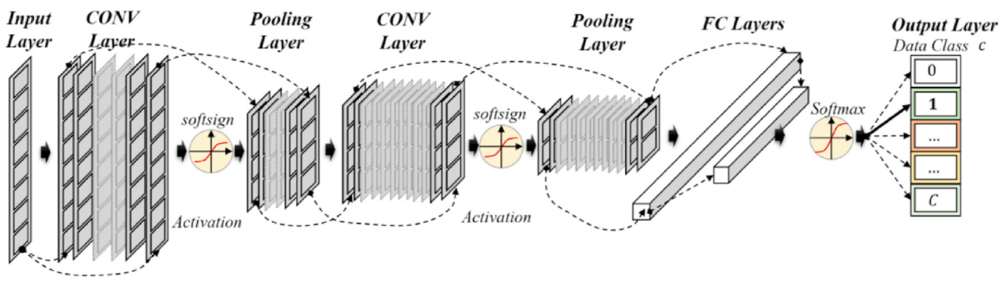


Figure 2.3.: Illustration of CNN structure for a data classification task [41]

A CNN primarily utilizes convolution and pooling operations. The convolution operation can be understood as a sliding filter applied to the energy series. Generally exhibited in a one-dimensional structure, it is often referred to as a 1D CNN [29]. This process effectively extracts local features from the input data. The implementation of CNNs in short-term load forecasting models is a significant focus in current research. For instance, Sadaei et al. [58] introduced a combined model incorporating CNNs to augment the accuracy of short-term load forecasting models. This further underlines the adaptability of CNNs beyond their conventional image-processing applications.

2.4.2. Clustering Analysis and Pattern Recognition in Energy Consumption

Analyzing time series data is crucial in energy management, with various methodologies proposed to understand building energy consumption patterns. These methodologies can be broadly categorized into three groups:

Techniques utilizing autocorrelation and probability distribution analysis Applications of Symbolic Aggregate Approximation (SAX) Clustering strategies for identifying typical daily electricity usage profiles.

Autocorrelation and Probability Distribution Analysis

Góis and Pereira (2022) proposed a technique to identify appliance usage patterns in buildings. Their process comprises a seasonality analysis using the Auto-Correlation Function (ACF), identification of significant differences, and examination of the Energy Concentration Coefficient (C) for each sliding window. This approach can potentially enhance energy efficiency [29].

Symbolic Aggregate Approximation (SAX)

Capozzoli et al.(2018) suggest a refined SAX process to identify irregular energy patterns. This method optimizes the time window width and symbol intervals in accordance with the building's energy behavior. Applying a Classification and Regression Tree (CART) regression model allows the daily energy profile to be discretized into uneven time windows, providing detailed analysis of daily energy usage variation [14].

Clustering Strategies for Identifying Typical Daily Electricity Usage Profiles

Li et al. (2018) proposed a clustering strategy to identify typical daily electricity usage profiles (TDEU) across multiple buildings. Their approach employs Gaussian mixture model-based clustering and agglomerative hierarchical clustering to identify patterns in building electricity usage. This methodology can aid energy-efficient decision-making for retrofits and performance enhancement [42].

A method for identifying fundamental load shape profiles was suggested by Park et al. (2019). The method applies rigorous clustering analysis and entropy calculation to a large dataset from a variety of buildings. The effectiveness of the method was evaluated using Clustering performance metrics, including Cohesion, Separation, and Calinski-Harabasz (CH) Score [50].

Quintana et al. (2022) provided a novel framework for outlier detection in energy load profiles using discord similarity and the Kolmogorov-Smirnov (KS) test [Mat].

Liu et al. (2021) proposed a data mining-based framework using a two-step clustering analysis. The technique uses the density-based spatial clustering application with noise (DBSCAN) and the k-means algorithm to extract TELPs [44].

A comprehensive survey by Aghabozorgi et al. (2015) identifies three distinct types of time-series clustering: whole time-series, subsequence, and time point clustering. They applied Time Series K-means to group buildings based on gas consumption patterns [**?**].

Modeling Occupancy and Occupant-Related Electric Load Profiles

Lastly, Causone et al. (2019) introduced a data-driven procedure for modeling residential buildings' occupancy and occupant-related electric load profiles. The method, which uses Self-Organizing Map (SOM) and the k-means algorithm, identifies representative electricity daily use profiles, from which daily occupant-related load profiles are generated for each cluster [15].

These unique methodologies provide insights into energy management strategies and could potentially be adapted for a variety of applications in the field of building energy consumption analysis.

2.5. Anomaly Detection in Building Energy Data

The concept of anomaly detection in building energy data involves the identification of unusual or unexpected energy usage patterns that may point to equipment malfunction, energy inefficiency, or other operational issues. Several notable studies have employed different approaches to tackle this issue, all contributing unique and valuable perspectives to this field.

Araya et al. (2016) presented a collective contextual anomaly detection (CCAD) framework, which integrates historical sensor data and additional features to train an autoencoder to recognize normal consumption patterns. Despite potential scalability and over-fitting issues, the autoencoder-based approach offered computational efficiency and improved classification accuracy. The framework's design allows the model to measure how closely the input data patterns align with the normal patterns upon which the model was initially trained [8].

The same team (Araya et al., 2017) later improved this approach by proposing an ensemble learning framework for anomaly detection (EAD). This work combines the previous model with two prediction-based anomaly classifiers using support vector regression and random forest. The framework effectively detects anomalies in building energy consumption, meeting stringent false positive requirements [9].

Chiosa et al. (2022) introduce a framework that combines a contextual matrix profile (CMP) algorithm with clustering analysis[18]. Their approach ranks anomalies with a severity score and addresses the misclassification of recurring abnormalities, known as the 'twin freak' problem.

Fan et al. (2018) explored an autoencoder-based ensemble method for anomaly detection in building energy data in a different approach. The study developed multiple autoencoders with various architectures and training schemes, with max-min normalization applied to normalize the anomaly scores generated by each autoencoder. They proposed two indirect methods to evaluate autoencoders' performance and employed a data-driven method for period identification. This study contributed significant insights into the performance of denoising autoencoders under different levels of masking noises [24].

Finally, Liu et al. (2022) conducted data mining research on office building energy patterns, employing cluster analysis and association rule discovery algorithms to analyze energy consumption in office buildings. By utilizing k-shape and apriori algorithms, the study successfully uncovered hidden energy usage patterns in the dataset. However, it also recognized the limitations of the apriori algorithm due to resource and time constraints when applied to large datasets [45].

In conclusion, these studies underline the importance of anomaly detection in building energy data. The variety of methods proposed reflects the complexity of the problem and the need for solutions that can handle different scales and types of data. Understanding and improving these techniques is crucial for achieving energy efficiency and sustainable building operation.

2.6. Conclusion

This literature review explored various aspects of data analytics and machine learning in building energy management. We highlighted various studies that applied these techniques, revealing their potential for optimizing energy consumption, recognizing patterns, extracting fractures, measuring time series similarities, and extracting salient features from energy data.

In the terms of energy management, we saw how machine learning applications can be pivotal in predicting and controlling energy consumption. Studies showed the effectiveness of these applications.

With regard to clustering analysis and pattern recognition, the literature revealed their importance in recognizing and categorizing diverse energy consumption behaviors. This information can assist in designing personalized energy efficiency strategies and detecting irregularities in energy usage.

The discussion on feature extraction demonstrated the essence of this process in distilling useful information from raw building energy data. Various techniques, such as deep convolutional auto-encoders and statistical feature extraction, have been employed to capture the crucial attributes of energy data that can improve energy management strategies.

Finally, we explored the role of anomaly detection in identifying unusual patterns or outliers in building energy data. Techniques such as the CCAD framework and Ensemble learning have been proposed to enhance the identification and handling of these anomalies, which can greatly improve the overall energy efficiency and security of building systems.

In conclusion, the literature highlights the considerable potential of machine learning and data analytics in building energy management. These techniques offer promising avenues for enhancing energy efficiency, optimizing energy usage, detecting anomalies, and uncovering useful patterns. Despite the progress made, there remains ample room for further exploration and development in this rapidly evolving field.

3. Description of case study

3.1. Building Overview

Situated on the cusp of Amstelveen and Amsterdam, the Uilenstede residential area is a vibrant student campus, constructed between 1964 and 2014. It boasts a thriving community of over 3,500 residents. It is home to various structures including more than 20 buildings with over 3,000 homes, a shopping facility, a café, the Griffioen theater, office spaces, and the VU Sports Center. The student housing organization DUWO manages the campus, which oversees the rental and maintenance of residential and utility spaces.

Each of the 20 buildings on the campus is equipped with unique metering systems, meticulously documenting the gas consumption of individual or shared heating systems hourly. This wealth of data presents a comprehensive view of the buildings' energy consumption patterns, forming the foundation for detailed analysis and subsequent optimization strategies. The energy supply for the complex is primarily in-house. DUWO centrally purchases gas and electricity from the energy company and distributes it to users and building installations. Heat for space heating and hot tap water is produced in-house using a combination of combined heat and power (CHP) units and central heating boilers. This generated heat is then disseminated to the homes via nine separate heat networks.

Building-specific information, including construction year and insulation details, is available for some buildings. Even though this information doesn't encompass all buildings, it provides valuable insights into the building characteristics and their energy efficiency. The buildings, constructed from the 1970s through the 2010s, embody the evolution of construction methodologies and insulation practices over time. Insulation levels differ across the buildings, painting a diverse picture of energy efficiency standards and revealing the potential for further improvements. Some buildings operate on a shared central heating system, while others utilize individual CHP systems, reflecting the variety of structural forms ranging from apartments to studios.

By coupling the hourly consumption data with the available building-specific information, we can conduct an in-depth analysis of energy consumption patterns. Understanding these patterns is pivotal in identifying energy-saving opportunities and devising strategies to enhance energy efficiency across the Uilenstede residential area.

The figures below provide visual representations of the building numbers and the distribution of gas consumption meters across the Uilenstede campus.

Figure 3.1 illustrates the building numbers within the campus. Figure 3.2 shows the arrangement of gas consumption meters in the campus.

For a more comprehensive list, refer to the appendix titled Building EPA information. The appendix provides detailed information about the buildings, including their construction years, complexes, and the number of homes they contain.

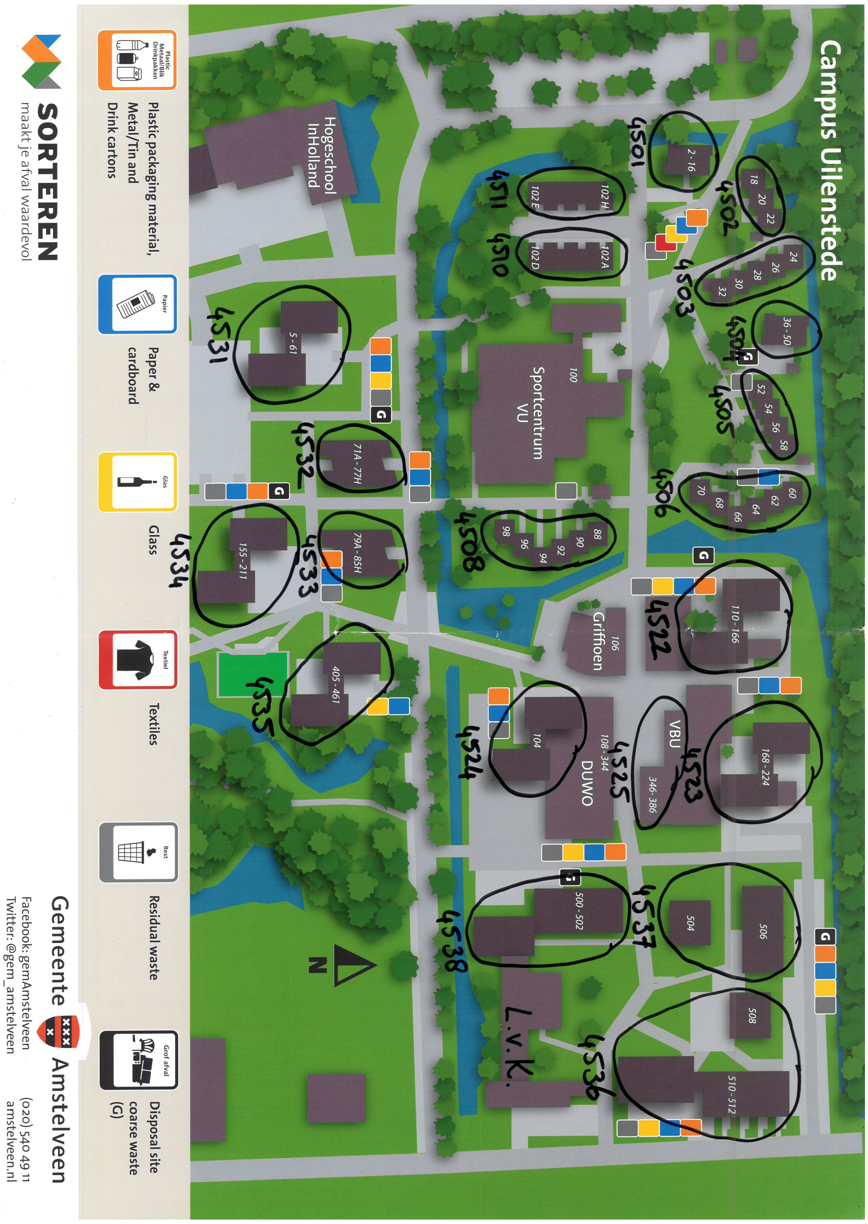


Figure 3.1.: Building numbers in Uilenstede campus

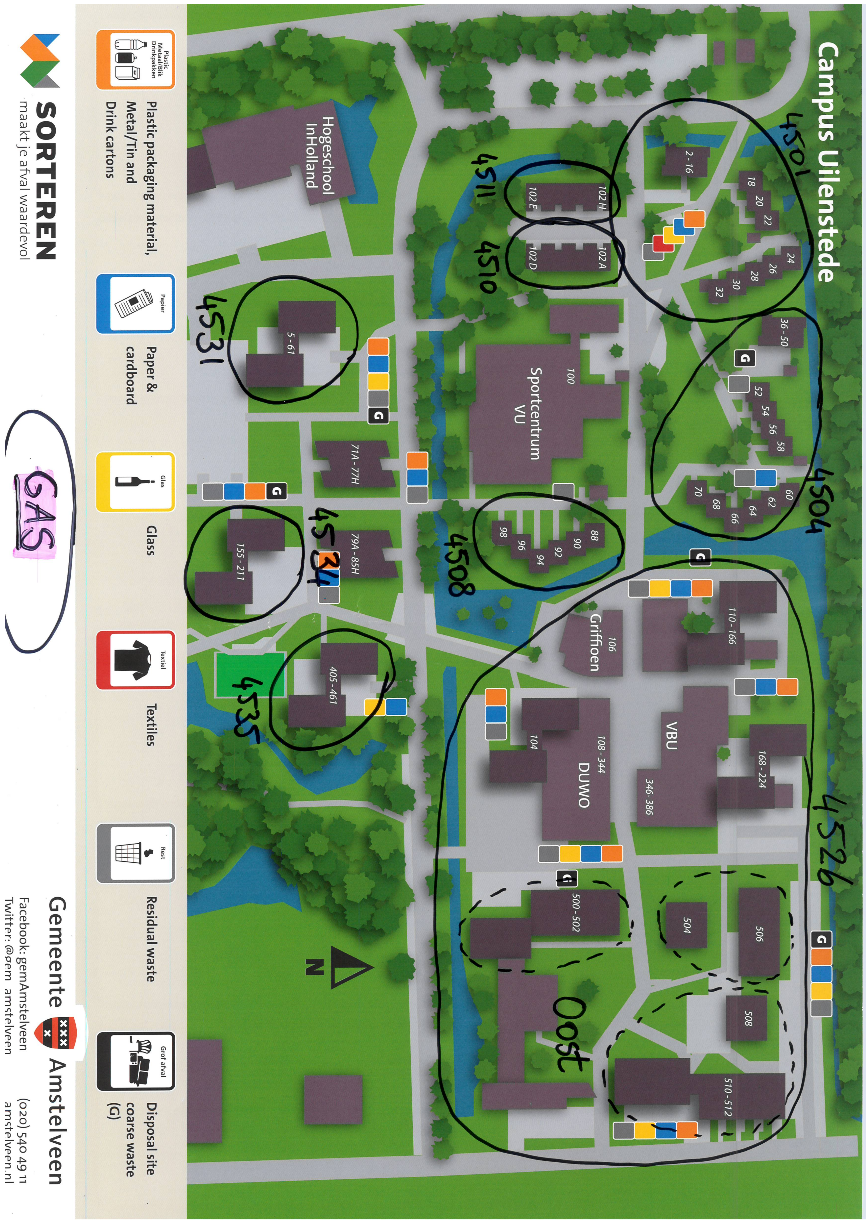


Figure 3.2.: Gas consumption meters in Uilenstede campus

3.2. Building Features

The Uilenstede residential area encompasses a variety of buildings, each with its unique features. Here is a brief overview of the buildings and their corresponding meter IDs: For a more comprehensive list, refer to the appendix titled Building EPA information.

The buildings, mainly from the same construction year, share similar construction attributes.

In 2001, they underwent insulation improvements to mitigate noise pollution from nearby Schiphol air traffic. The modifications primarily consisted of replacing facade glazing and panels. After consultations, insulation values were established for different parts of the residential buildings. These values are crucial for estimating the energy demand of the buildings.

3.2.1. Insulation Values

The buildings within the campus, mostly from the same time period, share similar construction attributes. In 2001, significant insulation improvements were implemented to mitigate noise pollution from Schiphol air traffic nearby. This primarily entailed the replacement of facade glazing and panels. Following consultations, insulation values were established for different parts of the residential buildings:

Table 3.1.: Building Insulation Information

Building Parts	Insulation Values
Walls	Contain uninsulated cavities, Rc = 0.36 (m2K)/W.
Ground Floor Floors	Uninsulated and constructed over crawl spaces, Rc = 0.5(m2K)/W.
Windows, Glass + Frame, and Doors	Installed with HR++ glass ('Schiphol glass'), U =1.7W/(m2K).
Tower Panels	Insulated during the application of Schiphol glass, Rc =1.75 (m2K)/W.

3.2.2. Ventilation and Energy Standards

Ventilation in the buildings is facilitated through natural supply via facade grilles and mechanical extraction. Extraction is collective for homes stacked vertically, with the fan and exhaust located on the roofs. A small portion of the homes use balanced ventilation with heat recovery. However, as this represents a minor percentage, calculations for all homes are based on natural supply and mechanical discharge. Newer homes on the east side, particularly building numbers 4536-4536, were constructed to meet a BENG insulation value of 0.75, a standard stricter than the current norm of 0.80.

3.2.3. Heating Systems

Several buildings house specific installations like combined heat and power (CHP) systems and central heating systems. For example, buildings 4531, 4534, and 4535 each have a communal CHP installation, while meter 4520 g0a serves as a central heating system for buildings 4520-4526 and 4536-4538. The buildings' usage also varies. Buildings 4501 and 4504 are apartment buildings with a shared central heating system, while buildings 4510 and 4511 were built in different periods. Buildings 4501 and 4504, constructed between 1970-1974, function as group homes with shared facilities. In contrast, buildings 4510 and 4511, erected around 1980-1985, offer multi-room homes. Figure 3.4 illustrates the energy supply within the campus.

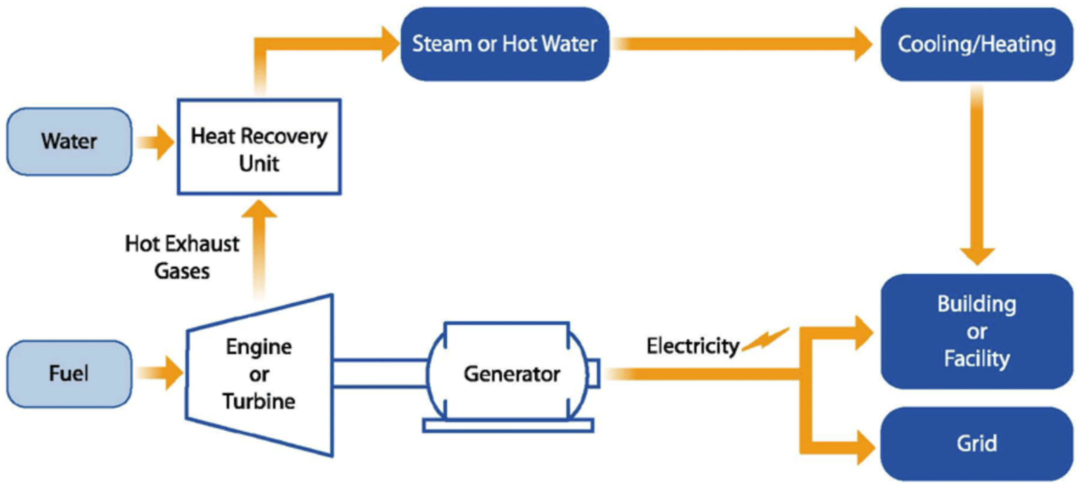


Figure 3.3.: Diagram of a CHP Installation
[10]

3.2.4. Operation of Heating Systems

A Combined Heat and Power (CHP) installation, also known as cogeneration, is a highly efficient system that simultaneously produces electricity and useful heat from the same energy source, often a fuel like natural gas. It operates by using the heat that would generally be wasted in a conventional power plant, making it a sustainable solution that reduces carbon emissions and lowers energy costs. A diagrammatic representation of a CHP installation can be seen in Figure 3.3.

In the case of the Uilenstede residential area, four complexes utilize CHP installations for their heating and hot water supply. The installations work by converting the gas into electricity through an engine. The heat generated during this process is then captured and used to heat water, which is circulated to provide heating and hot water in the residential complexes.

However, during peak loads, the CHP installations might not be sufficient to meet the demand for heat. In these instances, central heating boilers augment the CHP installations to manage the peak loads. The remaining complexes, which do not have CHP installations, rely solely on these central heating boilers for their heating needs. A detailed overview of the capacities per heat network and the distribution of CHP/central heating boilers can be found in Table 3.2 and Figure 3.4

The heating system in the complexes follows a heating curve principle, adjusting the output temperature based on the external temperature. This results in energy efficiency, as the system automatically reduces the delivery temperature as the outside temperature increases and vice versa. However, to comply with the protocols, which mandate that hot tap water temperature must not fall below 60 or 65°C, heat generation and transport for space heating and domestic hot water must be separated.

A point of improvement could be the inclusion of a cascade connection, where heat is used at different temperature levels. This would further enhance the energy efficiency and temperature management of the system.

Table 3.2.: Capacities per Heat Network and Distribution of CHP/Central Heating Boilers

Table 3.3.: Heat Network Overview

Building no	Heat network	Central heating boilers [kW]	Thermal Assets CHP [kWth]	Electric Assets CHP [kWe]	Proportion CHP thermal vermogen
4520	Centrum	5583	500	330	8%
4531	Toren 1	1315	250	165	16%
4534	Toren 2	1315	250	165	16%
4535	Toren 3	912	250	165	22%
4501	Uilenstede 2-32	1115	-	-	-
4504	Uilenstede 36-70	1115	-	-	-
4510	Uilenstede 110 ad	1115	-	-	-
4511	Uilenstede 110 eh	1115	-	-	-

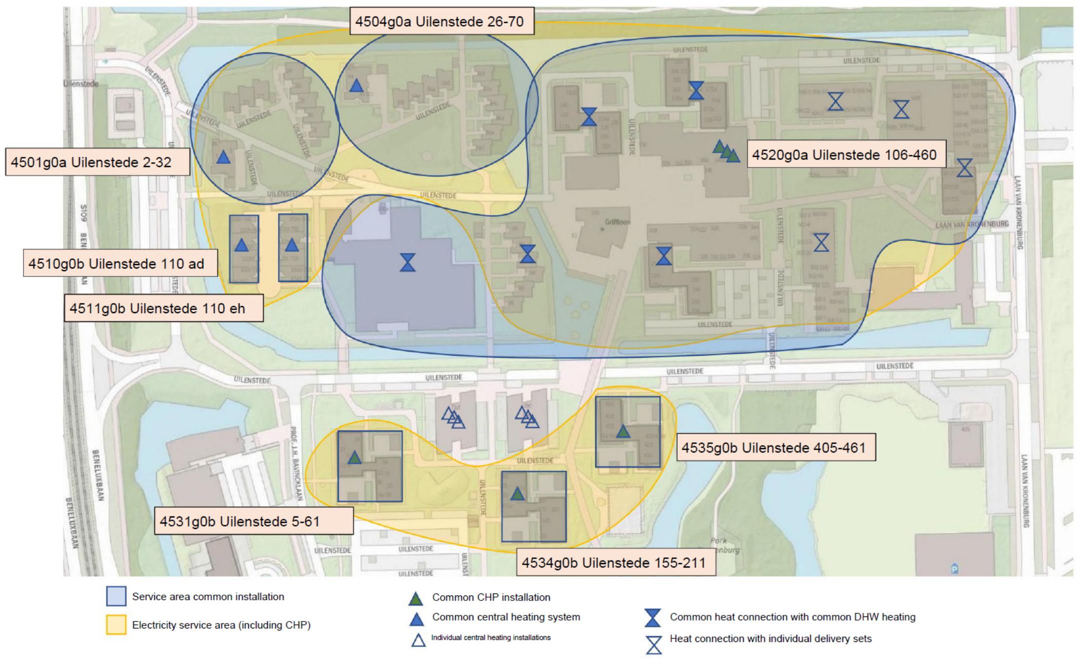


Figure 3.4.: Building heating system

3.3. Building Usage

The Uilenstede residential area exhibits a diverse range of building configurations catering to various resident profiles. Broadly, the buildings can be split into two principal categories based on their design and utilization:

1. **Group Houses:** Buildings such as 4501, 4531, 4534, and 4535, erected between 1970-1975, are conceptualized as group houses. They feature private bathrooms and shared spaces. The heating and hot water requirements of these buildings are serviced by either a central heating system or an individual Combined Heat and Power (CHP) system.

2. **Studios and Multi-Room Homes:** Buildings like 4510, 4511, and 4536-4538, this category offers a mix of multi-room homes and studio apartments. Constructed between 1980-2015, these buildings, like the group houses, are equipped with central heating or individual CHP systems.

Table 3.4 and Figure 3.5 comprehensively classify the buildings and their salient features.

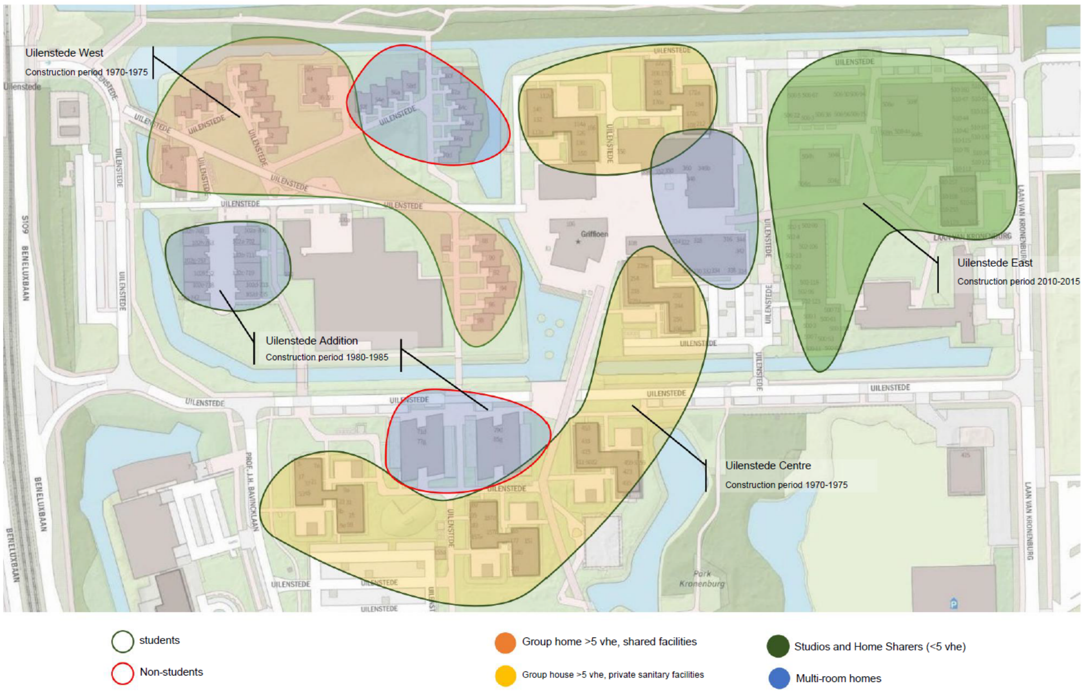


Figure 3.5.: Different housing types

Table 3.4.: Building Types and Characteristics

Building No.	Building Type	Features	Heating System
4501	Group House	Shared facilities	Central/Gas boiler
4504	Group House	Shared facilities	Central/Gas boiler
4510	Studios/Multi-Room Homes	Various configurations	Central/Gas boiler
4511	Studios/Multi-Room Homes	Various configurations	Central/Gas boiler
4531	Group House	Private bathrooms	CHP
4534	Group House	Private bathrooms	CHP
4535	Group House	Private bathrooms	CHP
4536	Studios/Multi-Room Homes	Various configurations	Central/CHP
4537	Studios/Multi-Room Homes	Various configurations	Central/CHP
4538	Studios/Multi-Room Homes	Various configurations	Central/CHP

A Combined Heat and Power (CHP) system, is an energy-efficient solution that concurrently generates electricity and useful heat from a single energy source, typically a fuel like natural gas. In the context of the Uilenstede campus, the CHP system is predominantly driven by heating demands, with electricity generation being a secondary product. During summer, the system mainly caters to hot water requirements.

On the other hand, buildings like 4501, 4504, 4510, and 4511 utilize a central heating system that operates on a gas boiler and features a hot water storage facility for water distribution. The meter data that we have collected accounts for the total gas consumption, which includes hot water, heating, and in the case of CHP, electricity production.

3.4. Metering System Overview

The dataset for Uilenstede campus is derived from a multi-building complex fitted with a Combined Heat and Power (CHP) / Photovoltaic (PV) system and a centralized heating distribution system on the southern side of the campus. Energy consumption is meticulously tracked through distributed monitors and sub-meters primarily situated in the eastern section of the buildings. The data comprises readings for both electricity and gas consumption. However, the primary focus of the analysis will be on gas consumption, especially during the winter months. This is in line with DUWO's observation that the majority of the energy consumption in the area is attributed to heating.

The table Building Number Address and Meter ID Information presents detailed information about various buildings, specifically their assigned numbers, associated Meter IDs, and locations. An important aspect to note from this table is that Meter ID 4905 is associated with the gas consumption data of eight different buildings. The table serves as an invaluable resource for understanding the allocation and usage of gas meters across different buildings, providing insights into how resources are shared and utilized within the network.

Table 3.5.: Building Number Address and Meter ID Information

Meter ID	Building No	Location
4902	4531	Uilenstede 5-61 (Toren 1)
4903	4534	Uilenstede 155-211 (Toren 2)
4904	4535	Uilenstede 405-461 (Toren 3)
	4520	Uilenstede 106-460
	4522	Uilenstede 110-166
	4523	Uilenstede 168-244
4905	4524	Uilenstede 108-268
	4525	Uilenstede 346-386
	4536	Uilenstede 508-512
	4537	Uilenstede 506-504
	4538	Uilenstede 500-502
	4504	Uilenstede 36-50
40752	4505	Uilenstede 52-58
	4506	Uilenstede 60-70
	4501	Uilenstede 2-16
40773	4502	Uilenstede 18-22
	4503	Uilenstede 24-32
41574	4510	Uilenstede 102 a-d (H-inst)
40812	4511	Uilenstede 102 e-h (trad)

	METERING TYPE
0	Central meter (connection to public grid) for total building (dwellings & amenities)
1	Metering point (connection to public grid) for domestic use
2	Metering point (connection to public grid) for non-domestic (amenities, elevator, etc)
3	Sub-metering point for a single domestic end-user
4	Sub-metering point for a shared facility/amenitie
5	Calculated domestic use (pe sum of no 3 meters in Uilenstede 500) Also used for non domestic end-metering points
6	Calculated non-domestic or shared multi user point
7	Sub-metering point for a singe installation/facility (wkk, elevator)
8	Sub-metering point, not dedicated to a single installation/facility
9	Production meter (produced or converted energy flow)

	LEVEL
a	Several buildings
b	Building level
c	Cluster (part of a building)
d	End user
e	sub metering behind end-user metering point (happens sometimes)

	E-flow	
e	Electricity	kWh
g	Natural gas	m3gas
h	domestic hot water	GJ
i	space heating only	GJ
j	heat	GJ
k	water: cold	m3h2o
l	water: hot	m3h2o

Figure 3.6.: Metering Type Classification Based on Meter ID

The metering systems in the residential buildings of Uilenstede are organized according to the type, level, and energy flow (E-flow) being measured. Each meter ID corresponds to different data types recorded, which provides a detailed understanding of energy usage patterns. Figure 3.6 presents an example of the metering type classification.

The level of measurement ranges from multiple buildings to individual sub-meters, while the E-flow captures different types of energy such as electricity, natural gas, hot water, cold

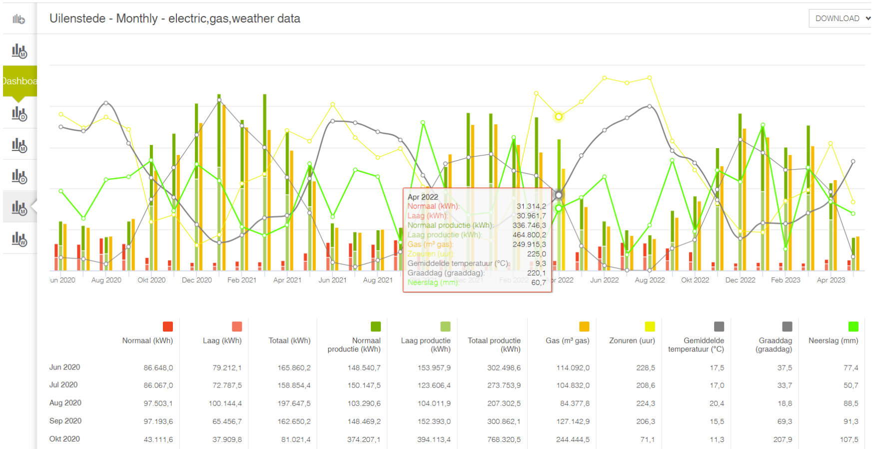


Figure 3.7.: Interface of the Joulz Dataportal

water, and heat.

Examples of meter ID representations include:

- *h3d:* Represents a sub-metering point measuring domestic hot water for an end-user.
- *k0b:* Represents a central meter measuring cold water from the public grid for the whole building.
- *l8b:* Represents a sub-metering point measuring water usage converted to domestic hot water.
- *g7a:* Represents a sub-metering point for natural gas entering a single installation, e.g., a boiler serving several buildings.
- *h9a:* Represents a production meter measuring the amount of heat produced in domestic hot water for use in multiple buildings.

The data for the analysis is obtained using an online API, which fetches both hourly and daily metered data from the Joulz Dataportal.

This allows for efficient and up-to-date data acquisition, ensuring the accuracy and relevance of the information used in the analysis, as shown in Figure 3.7.

From this time-series data, static features such as base load, peak load, and rising time can be extracted, providing valuable insights into the energy consumption patterns. This information forms the basis for understanding the various factors that contribute to the overall energy usage in the buildings and identifying opportunities for energy conservation at the Uilenstede campus.

4. Methodology

This chapter unfolds the designed methodology from the previous literature review. This approach demonstrates distinct energy usage patterns spread across different building profiles.

Initially, we present an encapsulated account of our approach, followed by an elaboration of the proposed workflow. This detailed overview delves into various clustering techniques crucial for analyzing building consumption data [28, 55, 57, 74].

Subsequently, the focus shifts to a spectrum of algorithms like dynamic time warping [5, 52, 69] and autocorrelation function, with different cluster evaluation methods including the Elbow Method, Silhouette Method, and the Bayesian Information Criterion for Gaussian Mixture Models [42, 59].

Further sections explore a range of clustering methods, user profiling [15, 50], and feature engineering [60], while also presenting a deciphering of the results with supplementary clustering methods.

In the later part of the chapter, we transition toward the practical applications and analysis of building consumption data. Here, we traverse through numerous analytical methods, including clustering based on seasonal or temporal variations [57, 74], followed by a discussion on the performance test and evaluation of the proposed strategy.

Then we illustrate the case study of the buildings under investigation and a comparative analysis of Time Series K-Means Clustering [14, 53].

Conclusively, the chapter explores the incorporation of extra information into the energy consumption analysis. Acknowledging the inherent correlation between building features and energy usage, we assimilate variables such as building age, insulation quality, and building geometry into the analysis [46]. This comprehensive and integrative approach lends an insightful understanding of energy consumption in the buildings on the Uilenstede campus, serving as a cornerstone for formulating sustainable energy conservation and efficiency strategies.

4.1. Summary of Approach

To address the main research question and its associated sub-questions, we have devised a comprehensive approach that combines the analysis of natural gas consumption patterns, the application of machine learning algorithms on smart meter data, and the examination of building attributes. The following sections outline the proposed methods for each sub-question.

4.1.1. Sub-Question 1: Can machine learning enhance the categorization of buildings based on their energy usage?

To gain novel insights into building energy efficiency via natural gas consumption patterns using machine learning methods, we propose the following strategies:

Strategy 1.1: Highlight potential energy-saving areas by analyzing deviations from average consumption trends, exploring seasonal variations, and contrasting consumption profiles across different types of buildings.

Strategy 1.2: Implement unsupervised learning techniques such as clustering to establish energy profiles for each building. This will enable us to identify common energy consumption patterns and understand disparities in gas usage among different buildings.

4.1.2. Sub-Question 2: How do building attributes influence energy consumption and a building's energy efficiency?

To comprehend how building characteristics affect energy efficiency, we propose the following:

Strategy 2: Post initial clustering analysis, incorporate more diverse data such as building features and weather records. This approach will augment our understanding of energy consumption patterns and help identify the role of building attributes like insulation and heating systems on energy efficiency.

4.2. Workflow

The proposed workflow for the analysis of gas consumption patterns and energy efficiency is as follows:

1. Data Collection
    a) Distribution map of the system
    b) Collect uilding features
2. Data Pre-processing
    a) Handling missing values
    b) Noise/outlier removal
3. Descriptive Statistics
    a) Understanding data attributes and summarizing the characteristics of a data set
4. Single and Global Building K-means Profiling
    a) Splitting time series data into daily segments
    b) Normalizing data by area (for Global Clustering)
    c) Determining the optimal number of clusters (K)

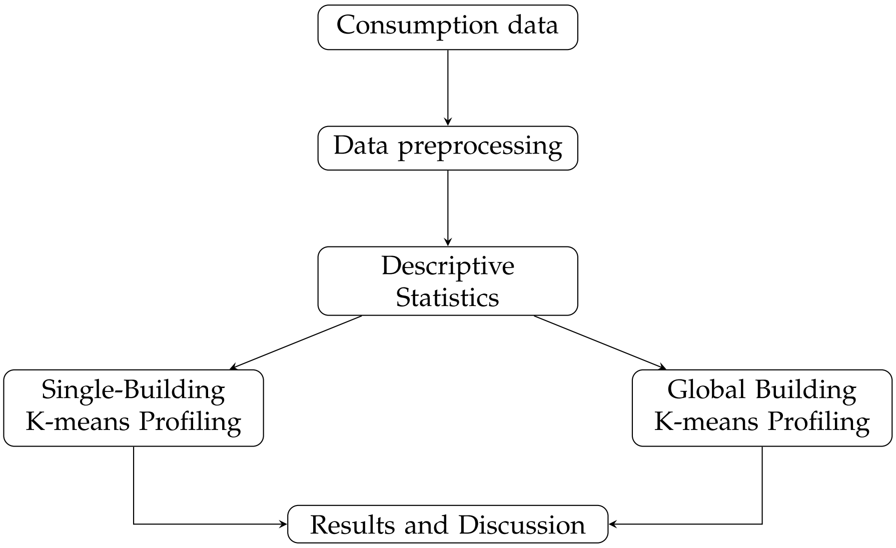


Figure 4.1.: Workflow for analyzing gas consumption patterns and energy efficiency.

   d) Applying K-means clustering
   e) Assigning typical profiles for each cluster
   f) Linking the clustered results with building characteristics

5. Results and Discussion
   a) Interpretation of the results
   b) Linking results with building characteristics

Upon completion of our detailed methodology and workflow, our objectives extend to:

- Examining variations in energy consumption across different seasons, weekdays compared to weekends, and holidays.
- Analyzing the clustering outcomes based on peak and non-peak hours of the daily cycle, specifically targeting correlations between peak periods and heating system use.
- Highlighting the differences in average consumption across different clusters, with special emphasis on seasonal variances between summer and winter.
- Providing detailed categorization of the identified clusters, including shared characteristics among buildings within each cluster. This involves examining the relationship between average gas consumption and factors such as building age, type, and the use of central heating systems.

4.3. Time-Series Analysis

The gas consumption data is a type of time-series data where the patterns of consumption can shift or warp over time. Traditional distance measures such as Euclidean distance assume synchronicity of the time series, but this might not be applicable for gas consumption patterns which could be influenced by factors like weather variations, daily routines, and seasonal cycles. The Dynamic Time Warping technique can effectively measure similarities by aligning sequences that have temporal distortions, making it more suitable for this data type [76].

4.3.1. Dynamic Time Wrapping

Dynamic Time Warping (DTW) is a method used to measure similarity between two temporal sequences, often used in time series analysis. The sequences could be of different lengths or time scales.DTW works by ”warping” the time dimension of the sequences to calculate the shortest possible path, or minimum cumulative distance, between the two sequences. This warping process allows for the sequences to be non-linearly aligned, accommodating any variations in timing or duration in the sequences. In this way, the sequences are optimally matched even when their lengths or time scales vary. Rather than calculating a direct distance between matching points in the input sequences, DTW permits a non-linear alignment between these sequences. This allows for an optimal match even when the sequences vary in length or time scales, making DTW an efficient and flexible tool for pattern recognition in temporal data [62, 76].

Regarding building energy consumption analysis, DTW is beneficial when paired with clustering algorithms like Time Series K-Means. It enables grouping buildings that display similar consumption patterns, regardless of potential variations in timing or duration across these patterns. DTW's functionality can be visually understood through 4.2. The figure reveals that while the Euclidean distance may incorrectly identify dissimilar sequences as similar due to its direct point-to-point connection, DTW ensures a reliable similarity measure by identifying the optimal alignment path.4.2 provides a binary matrix representation of the DTW path.

Each dot signifies a non-zero entry, hence, the matching of an element in one sequence with an element in the other. This representation and the corresponding optimal alignment elucidate the core functionality of DTW in comparing temporal sequences [62].

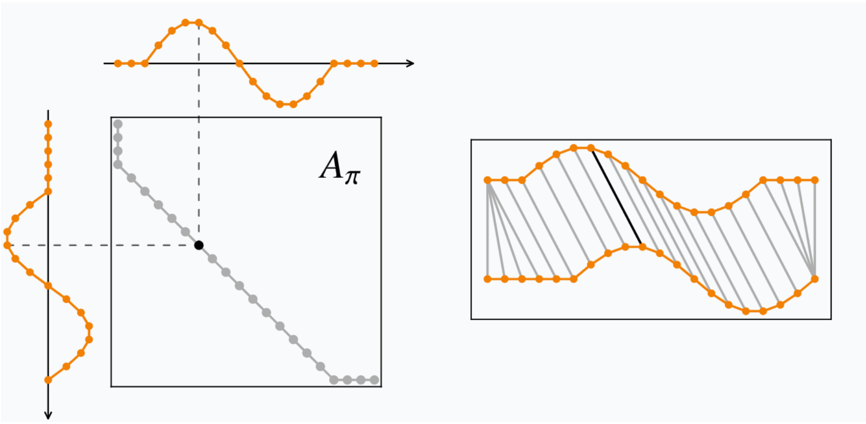


Figure 4.2.: Visual representation of Dynamic Time Warping

Alogrithm of DTW

The core principle of DTW is minimizing the distance between two time series, achieved by 'warping' the time dimension to calculate the shortest possible path or minimum cumulative distance. This calculation involves determining a cost, typically the absolute difference between the points in the two sequences. The final cell of the DTW matrix will then hold the minimum cumulative cost of aligning the sequences, as detailed in Algorithm 1.

Algorithm 1 Standard Dynamic Time Warping [3]

Data: Two sequences s and t of length n and m respectively
Result: Distance between the sequences

initialize a 2D array DTW with size (n+1) x (m+1) and set all elements to infinity DTW[0, 0] := 0 **for** *i=1 to n* **do**
    for *j=1 to m* **do**
        cost = d(s[i], t[j]) DTW[i, j] = cost + min(DTW[i-1, j], DTW[i, j-1], DTW[i-1, j-1])
    end
end
return DTW[n, m]

4.4. Clustering Methods

This section presents a comprehensive strategy for partitioning hourly gas consumption data from different building meters into distinct clusters. Initially, our proposed strategy employs Time Series K-means clustering both individually on each meter and globally across all meters. This dual-level analysis helps to identify distinct consumption patterns and outliers within and across buildings. Following the initial clustering, we explore the influence of seasonal variations on gas consumption. We particularly focus on the winter and summer seasons to elucidate the impact of weather conditions on energy usage. Subsequently, we conduct temporal clustering, dividing the consumption data based on distinct periods such as daily cycles, weekdays versus weekends, and holidays. This analysis uncovers patterns specific to different time frames, enabling more nuanced understandings of consumption behaviors. Finally, we display and interpret the results from all these clustering analyses, elucidating the various patterns and trends in building gas consumption. This comprehensive approach provides robust insights into building energy management.

4.4.1. Analyzing Hourly Gas Consumption Data with K-Means Clustering

When it comes to analyzing hourly gas consumption data collected over three years, K-Means clustering has emerged as a reliable and efficient choice due to its specific attributes. The key advantages of this algorithm include:

- Efficiency and Speed K-Means and its variant K-Medoids are known for their computational efficiency, which makes them a preferred choice over hierarchical clustering methods. They can handle large datasets, such as multi-year time-series data, and deliver prompt results.
- Representation Definition and Updating K-Means clustering relies on central representations (prototypes) to define clusters. This characteristic can significantly impact the success of the algorithm. This method is particularly beneficial for analyzing time-series data that exhibit evident trends and patterns, such as gas consumption influenced by seasonal variations and daily routines.
- Identification of Time-Series Patterns K-Means is highly effective at identifying clusters of similar time-series patterns. It can facilitate efficient detection and grouping of recurring patterns over time, which is an essential characteristic when analyzing gas consumption data.
- Compatibility with Dynamic Time Warping (DTW) K-Means can be combined with Dynamic Time Warping (DTW), a measure designed for time-series data where patterns may shift or distort over time. Incorporating DTW enhances K-Means' ability to handle temporal distortions and can improve the overall accuracy of the clustering process.

When used in conjunction with Dynamic Time Warping, K-Means clustering is a strong contender for clustering time-series data such as hourly gas consumption data spanning over three years. Its computational efficiency, effective handling of temporal distortions, and compatibility with equal-length sequences make it an ideal choice. While each clustering technique has its pros and cons, K-Means clustering offers distinct advantages for this type of task.

4.4.2. Outline of the proposed strategy

Using a clustering-based approach, the proposed strategy for identifying gas consumption profiles across various university buildings is depicted in figure 4.3. This strategy is a four-step process, encompassing data collection, single-building clustering, inter-building clustering, and result visualization and interpretation.

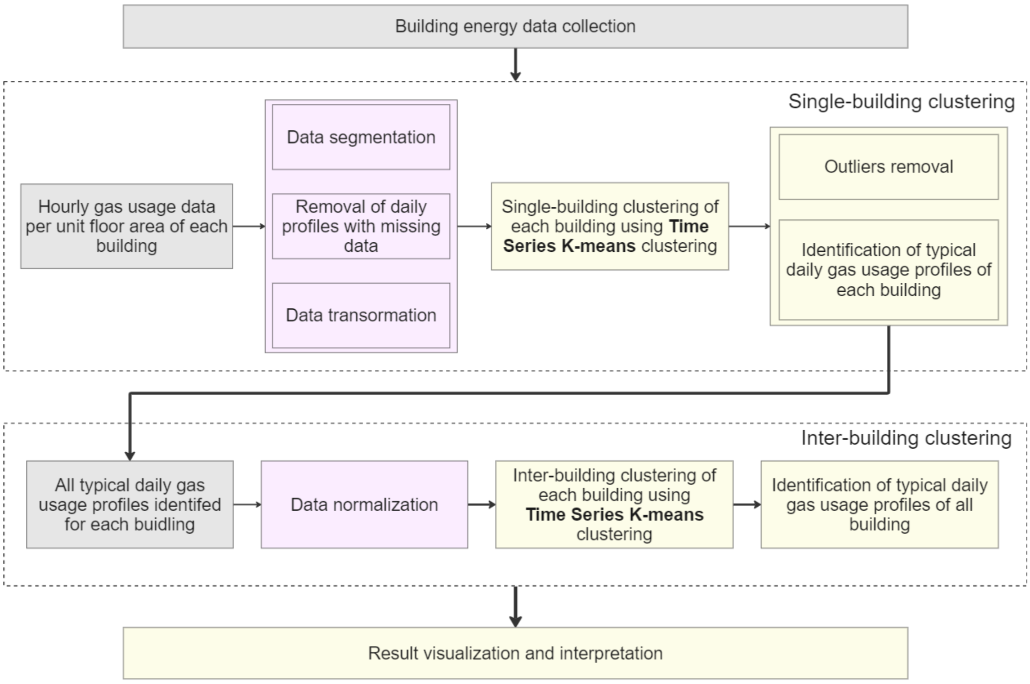


Figure 4.3.: Outline of the proposed strategy

- **Data Collection**: The first step involves gathering data on building gas consumption.

- **Single-Building Clustering**: The collected data is then segmented and normalized to provide daily usage profiles, considering hourly usage per unit floor area. Utilizing time series K-means clustering, these daily profiles are grouped based on similarity within the same group and dissimilarity from those in other groups. This process also aids in identifying and eliminating outliers. The median of all daily usage profiles within the cluster represents each cluster's typical profile. The choice of time series K-means as the clustering method comes from a comparative analysis with other clustering algorithms discussed in section 4.3. The details of this analysis are discussed in 5.3 and 5.4.

- **Inter-Building Clustering**: The identified typical profiles from each building are then inputted into the inter-building clustering phase. Here, we identify standard usage profiles across multiple buildings based on the typical daily profiles identified for each building. During this process, typical profiles from the single-building clustering phase are normalized, ensuring each profile has a mean value of 0 and a variance of 1. The dissimilarity measure between each pair of these normalized profiles is determined by calculating the Dynamic Time Warping (DTW) distance. These typical daily profiles are grouped into clusters using agglomerative time series K-means.

- **Visualization and Interpretation of Results**: Finally, the median value of all the daily usage profiles within each cluster is computed to derive the typical daily profiles for multiple buildings. These results are then visualized and interpreted to understand the gas usage behaviors across the buildings comprehensively.

4.4.3. Cluster Evaluation Methods

cluster evaluation methods were used, such as the elbow method, and silhouette score, to optimize the number of clusters used in our analysis. This step is critical as it influences the robustness and interpretability of our findings. The validation of these methods is important, as the reliability of the results relies on the accurate functioning of these algorithms. To optimally partition our data into clusters, we apply various evaluation techniques:

- **Elbow Method (for K-means)**: This method is commonly used to decide on the optimal number of clusters in K-means clustering [70]. It entails plotting the explained variance—the sum of squared distances of each data point to its cluster centroid—against the number of clusters. As we increase the clusters, the explained variance declines. We select the number of clusters at the point where further increase in clusters results in a marginal decrease in explained variance. This point resembles an "elbow" on the plot. Formally, this technique seeks to minimize the within-cluster sum of squares (WCSS):

$$WCSS = \sum_i (x_i - c_j)^2$$

where x_i denotes a data point, and c_j represents the centroid of the cluster to which x_i is assigned. The summation covers all data points and cluster centroids, as detailed in Algorithm 2

Algorithm 2 Elbow Method [38]

Data: Data set X and maximum number of clusters $maxClusters$
Result: Plot of WCSS values against number of clusters

initialize an empty list $wcssValues$ **for** $k = 1$ *to* $maxClusters$ **do**
  Initialize $kmeans$ with k clusters Fit $kmeans$ on X Append $kmeans.inertia\_$ to $wcssValues$
end
Plot $wcssValues$

- **Silhouette Method**: This method involves calculating the silhouette coefficient for each data point, which reflects the contrast between its average distance to points in its cluster (a) and its average distance to points in the nearest neighboring cluster (b). The silhouette coefficient ranges between -1 and 1, with higher values indicating better clustering [56]. It is computed as follows and shown in Algorithm 3:

$$s(i) = (b(i) - a(i))/max(a(i), b(i))$$

Here, $s(i)$ is the silhouette coefficient for data point i, $a(i)$ represents the average intra-cluster distance, and $b(i)$ is the average nearest-cluster distance. High silhouette coefficients indicate that data points are well-clustered.

Algorithm 3 Silhouette Method [38]

Data: Data set X and maximum number of clusters $maxClusters$
Result: Plot of Silhouette scores against number of clusters

initialize an empty list $silhouetteValues$ **for** $k = 2$ *to* $maxClusters$ **do**
    Initialize $kmeans$ with k clusters $preds \leftarrow kmeans.fit\_predict(X)$ Append SilhouetteScore(X, $preds$) to $silhouetteValues$
end
Plot $silhouetteValues$

4.4.4. Time Series K-Means Clustering

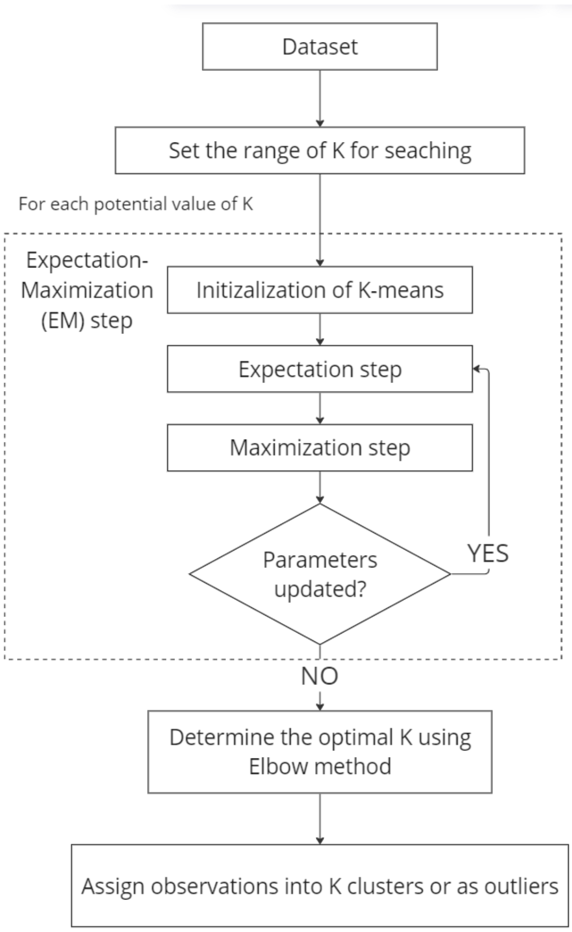


Figure 4.4.: Flowchart of Clustering

The application of Time Series K-means, depicted in Fig.4.4, takes advantage of the inherent K-means methodology to categorize numerous observations in a dataset into distinct clusters. To perform time series K-means-based clustering, a model is initially fitted to consumption data over specific time periods, which are transformed via different time segmentation. Here, each observation signifies an original time period usage profile. Following the fitting process, similar observations are grouped into the same cluster. The median of the associated original usage profiles within a cluster is subsequently taken as the characteristic usage profile of that cluster.

Time Series K-means is a specialized variant of the traditional K-means clustering algorithm to handle time-dependent data. The fundamental objective of Time Series K-means, similar to its traditional counterpart, is to minimize the sum of squared distances between data points and their corresponding cluster centroids. However, in the context of time series data, these centroids represent the average shape or pattern of the time series within each cluster rather than mere arithmetic means [34].

Traditional K-means clustering operates by partitioning n observations into k clusters, where each observation is assigned to the cluster with the nearest mean, which serves as the cluster's prototype. This traditional approach, however, does not account for the sequential nature of time-series data. It treats each time point as an independent dimension, leading to the possibility of misleading results when applied to time-series data.

On the other hand, Time Series K-means (TSKmeans) clustering offers a novel extension of the K-means approach, specifically tailored for time series data. The salient feature of TSKmeans lies in its innovative objective function that takes into account the temporal structure inherent in the data. Differing significantly from traditional K-means, TSKmeans introduces weights assigned to different time stamps, thereby facilitating the recognition of patterns across time, not just across observations.

The objective function of TSKmeans is designed to minimize the within cluster scatter while simultaneously smoothing the weights of adjacent time stamps. The objective function also incorporates a parameter α which is utilized to balance the influence between the scatter of objects within clusters and the smoothness of the weights of the time stamps.

In mathematical terms, the objective function can be described as follows[34]:

$$P(U,Z,W) = \sum_{p=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ip}w_{pj}(x_{ij}-z_{pj})^2 + \frac{1}{2}\alpha\sum_{p=1}^{k}\sum_{j=1}^{m-1}(w_{pj}-w_{p,j+1})^2 \tag{4.1}$$

subject to

$$\sum_{p=1}^{k} u_{ip} = 1, \quad u_{ip} \in 0,1, \quad \sum_{j=1}^{m} w_{pj} = 1, 0 \leq w_{pj} \leq 1. \tag{4.2}$$

where α is a balancing parameter.

This feature makes Time Series K-means particularly suitable for analyzing building consumption patterns. By identifying clusters of buildings with similar temporal patterns in gas usage, this method can provide insights into common patterns and anomalies, help identify factors driving these patterns, and inform strategies for management and intervention. For instance, it could help identify groups of buildings that show unusually high consumption at certain times of the day or year, suggesting opportunities for targeted energy efficiency measures.

A crucial aspect of K-means-based clustering is determining the optimal number of mixture components, K. The Elbow method is employed to ascertain this parameter 2. The K value

Algorithm 4 TSkmeans Algorithm [34]

Data: Time series data set $X = X_1, X_2, \ldots, X_n$, number of clusters k, smoothing parameter α
Result: Membership matrix U, centroid vectors Z, weight vectors W

Randomly choose an initial $Z^0 = Z_1, Z_2, \ldots, Z_k$ and weight $W = wp, j$
repeat
    Z and W fixed Solve the membership matrix U with Eq.4.1 U and W fixed Solve the centroids Z with Eq.4.2 U and Z fixed Solve the weight W with quadratic programming
until *Convergence*;

that minimizes the Elbow score is selected for each building as the optimal K for the K-means fitting. Prior research indicates that the number of typical profiles for individual buildings typically ranges between 2 to 8 [14, 71]. Therefore, the optimal K value for the single-building clustering in this study is derived within a scope of 2 to 14.

Algorithm 5 K-means Clustering [30]

procedure KMEANSCLUSTERING(*data*, *maxClusters*)
    Initialize *silhouetteValues* as an empty list **for** $k = 2$ *to maxClusters* **do**
        Initialize *clusters* randomly **while** *clusters change* **do**
        clusters ← AssignPoints(*data*, *clusters*) ▷ Assign each point to nearest centroid
    clusters ← UpdateCentroids(*clusters*) ▷ Recompute new centroid of each cluster
    preds ← AssignPoints(*data*, *clusters*) ▷ Final assignment of points to clusters
    Append SilhouetteScore(*data*, *preds*) to *silhouetteValues* ▷ Compute silhouette score
    optimalK ← FindElbowPoint(*silhouetteValues*) ▷ Determine optimal k using the Elbow method2
    clusters ← RunKMeans(*data*, *optimalK*) ▷ Run k-means with optimal number of clusters
    VisualizeClusters(*data*, *clusters*) ▷ Visualize the resulting clusters

These steps are reiterated until either the assignments no longer change or the algorithm reaches a predefined maximum number of iterations. This is a specific implementation of the Expectation-Maximization (EM) algorithm, where the 'expectation' step corresponds to the assignment of data, and the 'maximization' step corresponds to the recalculation of centroids.

5. Results

5.1. Summarizing the characteristics of data set

This section delves into the analytical aspects of the collected building gas consumption data. Firstly, we offer a comprehensive overview of the consumption data, discussing its general characteristics and initial observations. This overview forms the basis of our subsequent in-depth analyses. Next, we perform an Interquartile Range (IQR) analysis across different meters. By focusing on the spread of the middle 50% of the data, the IQR analysis allows us to understand the dispersion and detect potential outliers within each meter's consumption data. Finally, we investigate the periodicity in the gas consumption data using autocorrelation. This analysis reveals the correlation of the data with its past values, enabling us to discern patterns and cycles in the time series data.

5.1.1. Overview of consumption data

The energy consumption pattern of a building can be significantly influenced by seasonal changes, as illustrated in Figure 5.1 representing a year's worth of hourly gas consumption data for building id 4902.

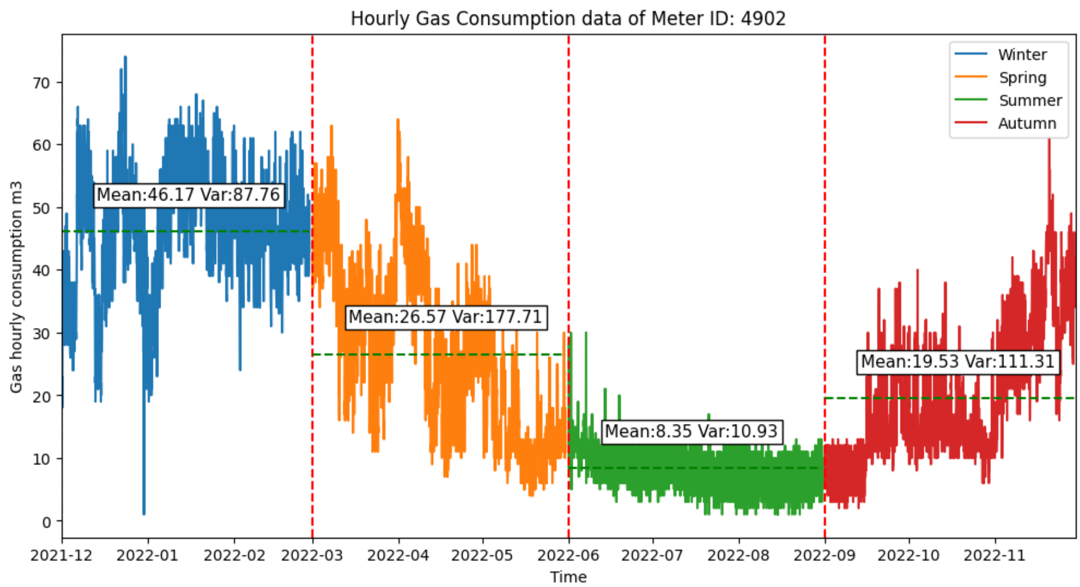


Figure 5.1.: Hourly consumption of one-year data for building 4902

In the figure, the transitions between the seasons are represented by vertical red lines. Green horizontal lines correspond to the average consumption throughout each respective season. A trend observed in the data is the relatively high gas consumption during winter months, contrasted by the lowest consumption during summer.

However, the most interesting aspect lies in the degree of variability or inconsistency in consumption, represented by the variance in the data. The summer season records the smallest variance, suggesting a stable and predictable consumption pattern during these months. This could be attributed to relatively uniform weather conditions and daylight hours. The spring and autumn seasons, also known as the transitional seasons, demonstrate the highest variances - 179 for spring and 130 for autumn, compared to a mere 12 during summer. Spring's variance is approximately 18 times greater than summer's, indicative of a more unpredictable and fluctuating consumption pattern.

The unpredictability of weather conditions can explain this heightened variability during these seasons. While winter and summer present fairly consistent cold and hot conditions respectively, spring and autumn alternate between a broader range of temperatures. These erratic weather patterns lead to inconsistent usage of heating systems in buildings, reflected in the unstable gas consumption data. Variations in daylight hours with the lengthening and shortening of days could also impact gas consumption during these transitional seasons.

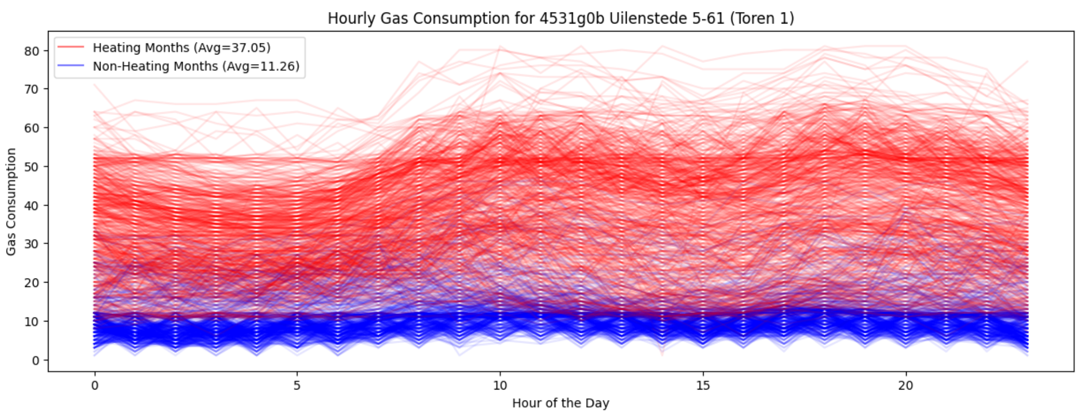


Figure 5.2.: Hourly consumption of one-year data for building 4902

To delve deeper into the disparities in gas consumption, we've examined and visualized the consumption data during heating and non-heating months in Figure 5.2, with the Netherlands' climate as our frame of reference. Typically, the heating season extends from October 1st through April 30th, a period when the chilly weather necessitates the use of heating systems. Contrastingly, the non-heating season runs from May 1st to September 30th, during which warmer conditions usually eliminate the need for active heating measures.

Taking the example of Building 4531, a discernible difference is evident in gas consumption between the heating and non-heating months. In fact, the average gas usage during the heating months exceeds that of the non-heating months by more than three times. This stark contrast underscores the significant role that climate control, particularly heating, plays in determining gas consumption. Therefore, as the temperatures dip during the colder months, the necessity for heating escalates, leading to a corresponding surge in gas consumption.

5.1.2. Interquartile Range (IQR) Analysis

The Interquartile Range (IQR), also known as the midspread, middle 50%, or H-spread, is a measure of statistical dispersion, providing insight into a dataset's variability [65]. It is defined as the difference between the upper quartile (Q3, 75th percentile) and the lower quartile (Q1, 25th percentile), i.e., IQR = Q3 - Q1. Figure **??** illustrates a boxplot with an interquartile range and a probability density function of a Normal Population. The IQR values for the gas consumption data per floor area for different buildings are presented in Figure 5.3.

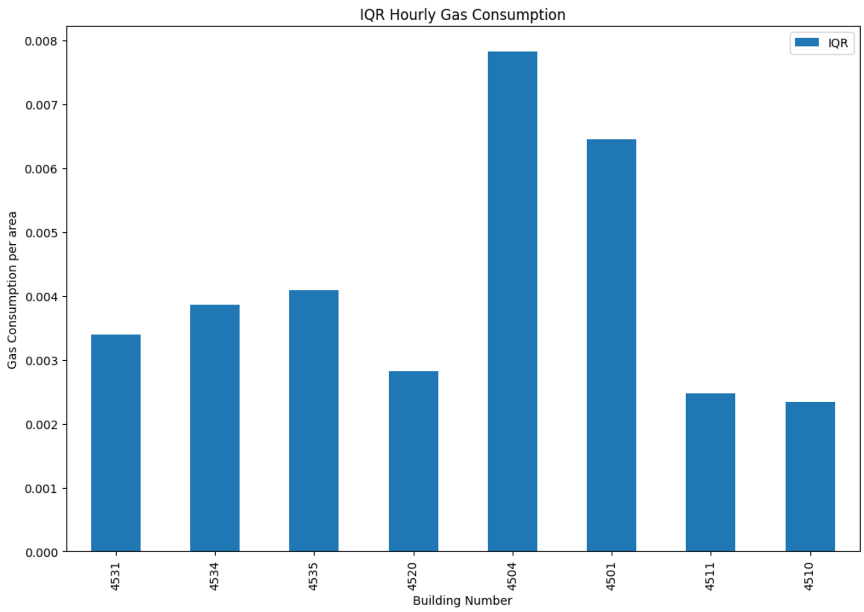


Figure 5.3.: IQR bar plot of each building

From these results from Figure 5.3 5.4, we can draw several conclusions:

- **Variability in gas consumption**: Buildings 4504 and 4520, which have larger IQR values, show greater variability in gas consumption per floor area. Conversely, buildings 4510 and 4511, with smaller IQR values, exhibit a more consistent gas consumption per floor area.
- **Energy efficiency**: Buildings 4510 and 4511, with smaller IQRs, might be more energy efficient, as their gas consumption per floor area varies less. However, additional data, such as the types of heating systems and insulation used, would be required to confirm this.
- **Comparison across buildings**: The IQR allows us to compare gas consumption variability across different buildings. For instance, building 4520, which has a larger IQR, suggests higher variability in gas consumption per floor area than the other buildings.
- **Potential for optimization**: Buildings with higher IQR values, such as 4504 and 4520, may benefit the most from energy consumption optimization measures, as they exhibit greater variability in their gas consumption.

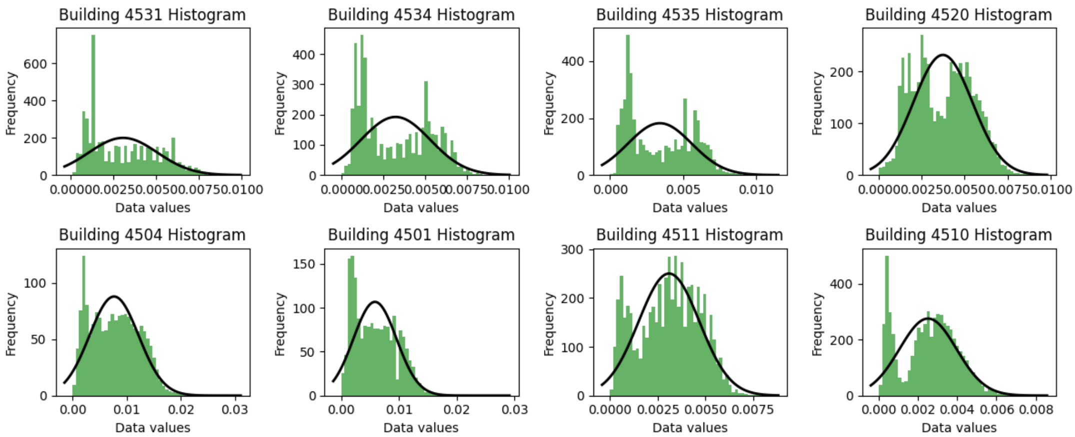


Figure 5.4.: Histogram of each building

- **Skewness** Skewness values are relatively low (close to 0) for all buildings. This suggests that the gas consumption per area in each building is roughly symmetrical.

5.1.3. Periodicity Analysis via Autocorrelation

Autocorrelation function (ACF) is an integral analytical tool for determining recurring patterns in time-series data. Autocorrelation in a discrete time frame assesses the correlation between data points spaced apart by a distinct time lag. As a mathematical tool, autocorrelation unveils repetitive patterns within the data. Formally, the autocorrelation of a real, discrete-time signal $x[n]$ for a delay of m is defined as

$$R[m] = \sum_{n=0}^{N-m-1} x[n]x[n+m],$$

where $R[m]$ is the autocorrelation at lag m, and N is the total number of points in the signal. The normalized autocorrelation at lag m, denoted as $r[m]$, is calculated by dividing the autocorrelation by the zero-lag autocorrelation, i.e., the energy of the signal:

$$r[m] = \frac{R[m]}{R[0]}.$$

This normalization ensures that the autocorrelation function has a value of 1 at zero lag. The autocorrelation function measures the correlation between data points separated by a specific lag, providing insights into the presence of recurring patterns in the data. In the discrete time case, autocorrelation is also referred to as serial correlation, as it quantifies the correlation of a signal with a delayed copy of itself as a function of delay. Thus, autocorrelation analysis serves as a mathematical tool for detecting repeating patterns.

In the context of building energy consumption, the autocorrelation function proves particularly valuable. Patterns in energy usage over time can be effectively identified using ACF [25, 57]. For instance, a strong autocorrelation at a specific lag may suggest a daily or weekly pattern in energy consumption, such as elevated usage during daytime hours or workdays.

Recognizing these patterns can inform energy management strategies, facilitating the identification of potential opportunities for energy conservation. The following pseudo-code outlines the steps involved in this process:

Algorithm 6 Autocorrelation Analysis of Energy Consumption Data

```
procedure AutocorrelationAnalysis(data)
    meter_data ← ExtractMeterData(data, meter_id)
    acf ← ComputeAutocorrelation(meter_data, max_lag)
    filtered_points ← FilterPoints(acf, min_lag_difference)    ▷ At least six intervals apart
    VisualizeACF(acf, filtered_points)
    PrintPoints(filtered_points)
```

By executing the steps outlined in this pseudo code, we can evaluate the auto-correlation at different lags, allowing us to identify recurring patterns in the energy consumption data. Figures5.5, and 5.7 display these patterns.

5.1.4. Hourly autocorrelation analysis

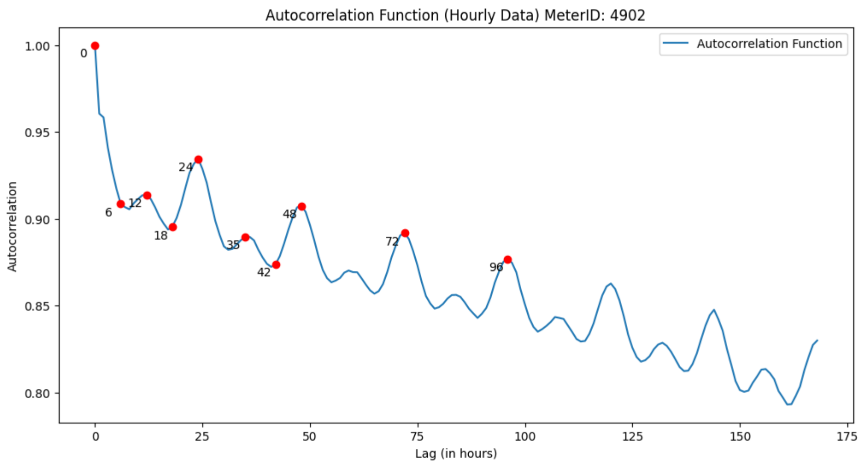


Figure 5.5.: Autocorrelation of Hourly Consumption of Meter ID 4902

Based on the autocorrelation values of the hourly gas consumption for MeterID 4902 in figure 5.5, clear patterns emerge that suggest a daily cycle in energy usage.

The highest autocorrelation values are observed at 0, 24, 12, 6, and 48 hours lags. The strong autocorrelation at a lag of 24 hours underscores a daily cycle in energy usage. This aligns with typical patterns of building energy usage, which often show daily fluctuations due to human activities and environmental factors.

Furthermore, the significant autocorrelations at 12 and 6 hours indicate smaller repeating patterns within each day. These may reflect patterns such as morning and evening peaks in energy usage.

Additional autocorrelations are observed at 48, 72, and 96 hours, suggesting a repeating pattern every two to four days. However, the physical interpretation of these patterns may require more context or additional analysis.

Also of interest is the autocorrelation at lags of 18, 35, and 42 hours. These lags do not align neatly with a 24-hour cycle and suggest there may be additional patterns of usage at these intervals. These could potentially be tied to certain activities or operational schedules specific to this building.

In conclusion, the analysis of autocorrelation values clearly indicates a daily cycle in the gas consumption of MeterID 4902, with additional smaller patterns repeating within each day. The findings also suggest possible multi-day patterns. Further analysis and additional contextual information about the building could provide more insights and confirm these observations.

5.1.5. Hourly autocorrelation analysis across all meters

Several patterns and trends can be observed based on the autocorrelation values of the hourly gas consumption for various meter IDs.

In the case of MeterID 4902 and 4903, strong autocorrelation values are observed at lags of 24, 12, 6, and 48 hours, underscoring a robust daily cycle in energy usage. Significant autocorrelation values are also observed at lags of 72 and 96 hours, hinting at potentially repeating patterns every 3-4 days.

Interestingly, MeterID 4903 shows a high autocorrelation at a lag of 11 hours, indicating some specific energy usage pattern during the day. For MeterID 4904, the highest autocorrelations are observed at 6, 23, and 12 hours. These values, deviating from the full 24-hour cycle, may be indicative of specific, less common operational or usage patterns.

MeterID 4905 demonstrates significant autocorrelations at 24-hour lags and multiples thereof (48 and 72 hours), revealing a strong daily usage cycle. High correlation values are also noted at 35 and 60-hour lags, suggesting additional operational patterns for this meter.

MeterID 40752 shows exceptionally strong autocorrelations at 24-hour intervals up to a 168-hour (7-day) lag, suggesting a highly regular weekly pattern of gas usage. Similarly, MeterID 40812 demonstrates high autocorrelations at lags of 24, 48, 72, and 96 hours, again indicating a steady daily cycle in energy use.

MeterID 40773, much like MeterID 4904, presents high autocorrelations at non-24-hour intervals, such as 11, 18, and 35 hours, indicating unique usage patterns.

Finally, MeterIDs 41574 and 40812 show strong autocorrelation values at lags of 6, 12, 18, and 24 hours, again emphasizing the importance of daily cycles. These meters, however, also show strong autocorrelation at non-24-hour intervals, possibly indicating unique daily usage patterns.

In conclusion, while most meters show strong daily and weekly usage patterns, some present unique operational patterns at non-standard intervals. These patterns could be the result of specific building operations or usage requirements, highlighting the need for customized energy management strategies. Further context about each meter's usage environment would help to refine these observations.

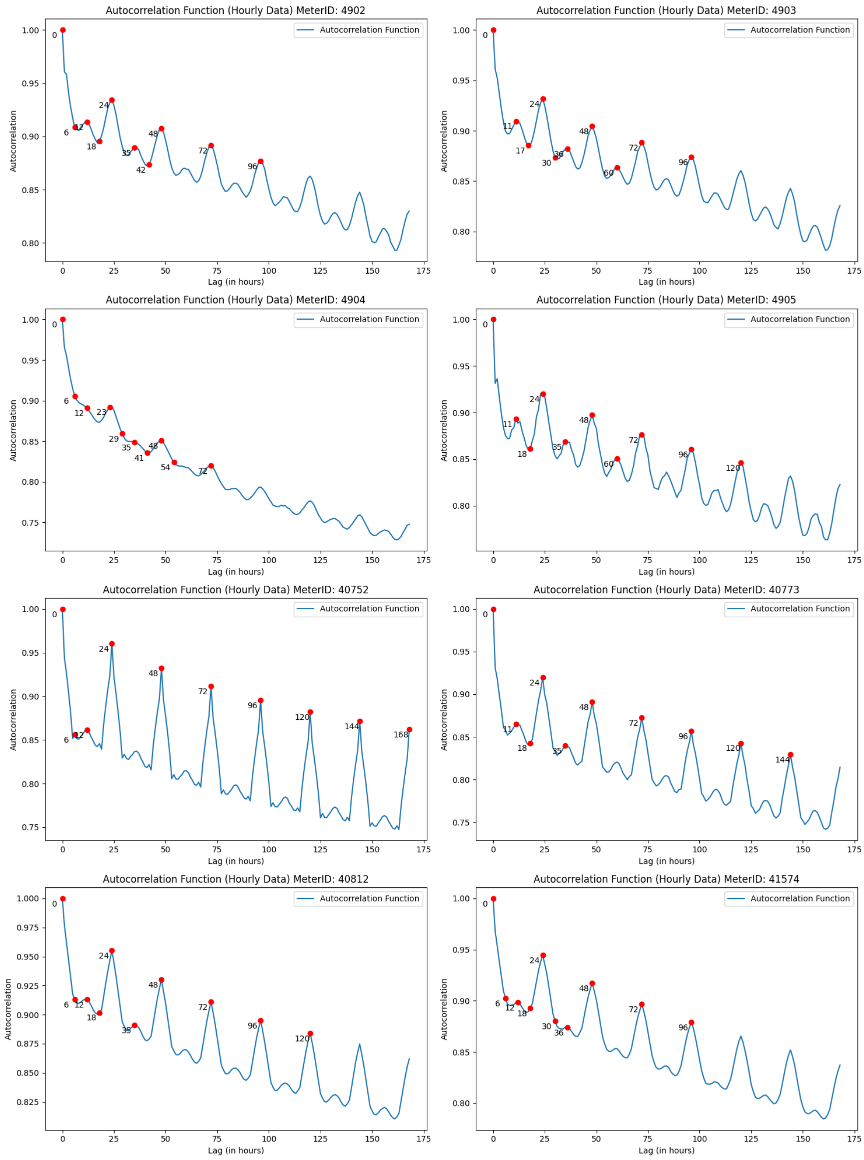


Figure 5.6.: Autocorrelation of Daily Consumption of all Meter ID

5.1.6. Daily autocorrelation analysis

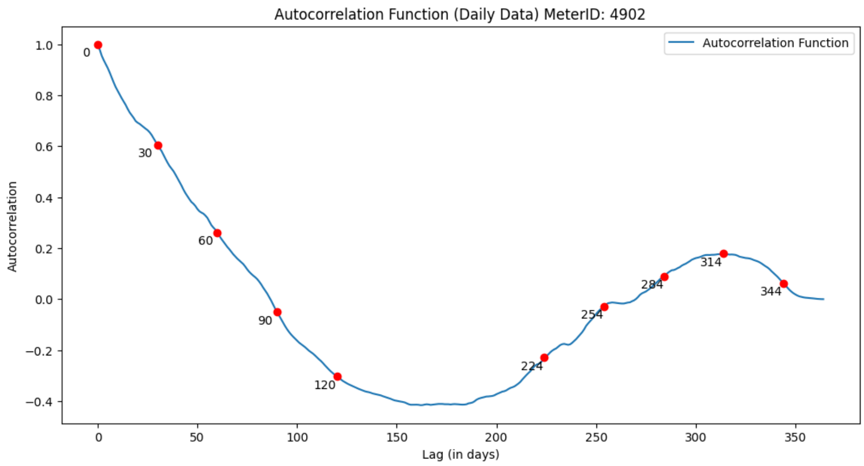


Figure 5.7.: Autocorrelation of Daily Consumption of Meter ID 4902

Based on the daily data presented in Figure 5.7, the autocorrelation values do not display a clear weekly or monthly data pattern. However, there is evidence of yearly data. Based on the autocorrelation values of the daily gas consumption for MeterID 4902, a clear cyclical pattern is not immediately evident at the weekly or monthly level. The first significant autocorrelation occurs at a lag of 30 days, which might suggest some monthly pattern. However, this should be further investigated, as it could be a coincidental finding rather than a proper monthly cycle.

Autocorrelation values at lags of 60, 90, and 120 days show a weaker correlation and even a negative correlation for the 120-day lag, suggesting there is less similarity in the gas consumption pattern at these intervals.

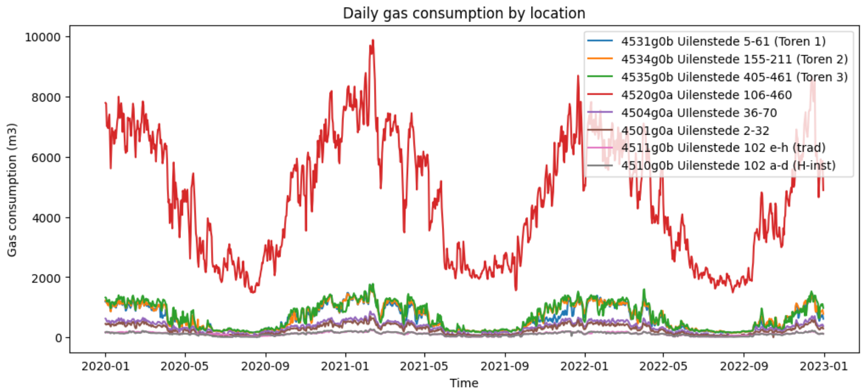
Interestingly, we see noticeable autocorrelation values at lags of approximately 300 days (314, 284, and 344 days), indicating some degree of year-long periodicity. These findings suggest that the gas consumption pattern may be related to annual cycles, possibly reflecting seasonal variations in usage. The negative autocorrelation at a lag of 224 days could indicate a seasonal inversion where the consumption pattern changes direction.

In conclusion, while there's no clear monthly cyclical pattern in the gas consumption of MeterID 4902, there's an indication of a possible yearly cycle. Further analysis could reveal more insights and confirm these observations. Additional context about gas usage, such as geographical location, usage type (residential, commercial, etc.), could provide insights when interpreting these results.

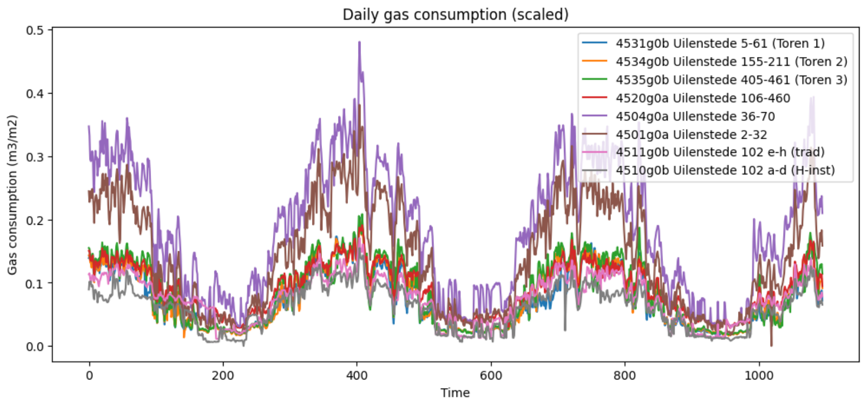
5.2. Performance test and evaluation of the proposed strategy

This section provides a brief outline of the methodology implemented in our research. The strategic approach proposed in this study was developed using Python. The time-series K-means-based clustering, a crucial component of our strategy, was implemented utilizing the scikit-learn library [51].

5.2.1. Insights from the Case Study Buildings



(a) Daily consumption data 2020-2023 across locations



(b) Scaled Daily consumption data 2020-2023 across locations

Figure 5.8.: Usage profiles for Building 4531

Figure 5.8 provides a thorough breakdown of daily gas usage across different meter locations, exhibiting distinct consumption trends. Specifically, Meter 4905, which is centrally located within the campus, demonstrates the highest consumption level. This pronounced consumption can be attributed to its extensive servicing role; while most buildings service 2-3 buildings or have standalone meters (as detailed in Table 3.5), Meter 4905 supports a network of approximately ten buildings. This wider reach naturally incurs higher energy demands, which accounts for the observed elevated consumption. Apart from Meter 4905, the remaining buildings present similar gas consumption levels. Building 4504 and Building 4501, built in 1991, notably exhibit the highest gas consumption per floor area. These buildings are low-rise and high-rise structures, respectively, accommodating 150 to 200 group units each, with each unit equipped with a separate central heating system.

5.2.2. Assessment of Average Yearly Gas Consumption per Building Floor Area

The insulation characteristics of each building (refer to Table B.1) suggest potential factors contributing to the elevated gas consumption per floor area in Buildings 4501 and 4504. Although these buildings are not the oldest ones on campus, their insulation standards fall short in comparison to newer buildings like Uilenstede 500-502, Uilenstede 504-508, and Uilenstede 510-1 (built between 2012 and 2014). With their lower insulation values, Buildings 4501 and 4504 require more energy to maintain an optimal indoor temperature, thereby increasing gas consumption.

Our proposed strategy's effectiveness was assessed using yearly gas usage data from 2020 to 2023, recorded by 8 meters located at the Uilenstede Campus, South of Amsterdam. These meters are predominantly found in student accommodations and communal areas (as shown in Tables A.3 and A.2). Figure 5.9 presents the average yearly gas usage in relation to the building floor area. There is substantial variation in the average yearly gas consumption across the meters, with values ranging from 21.99 m^3/m^2 (Meter 41574) to 67.22 m^3/m^2 (Meter 40752). Notably, buildings with comparable functions exhibit substantial consumption differences. For instance, Meter 40752 and Meter 41574, both servicing similar floor areas, demonstrate a nearly three times difference in their mean yearly gas usage, with Meter 40752 recording the higher value.

5.2.3. Estimation of Summer Hot Tap Water Consumption and Space Heating Demand

An iterative process was adopted to calculate the summer hot tap water consumption percentage. First, the gas data was sourced from meters connected to the Combined Heat and Power (CHP) system and the boiler, which track the gas consumption of the heating system. For each meter, the data was filtered for the non-heating months (May to September) and averaged to estimate the gas consumption exclusively for hot tap water during summer.

The hot tap water percentage can be calculated using the equation:

$$\text{Hot tap water percentage} = \frac{\text{Averaged summer months consumption (May-Sep)}}{\text{Averaged total consumption}}$$

The total gas consumption for each meter was then computed by averaging the consumption values over the year. The proportion of gas used for hot tap water was obtained by dividing the average consumption during summer months by the total intermediate consumption,

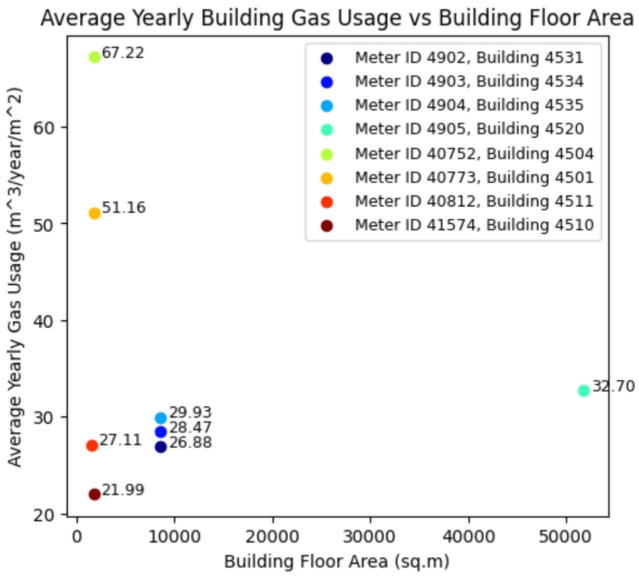


Figure 5.9.: Average yearly building gas usage per building floor area

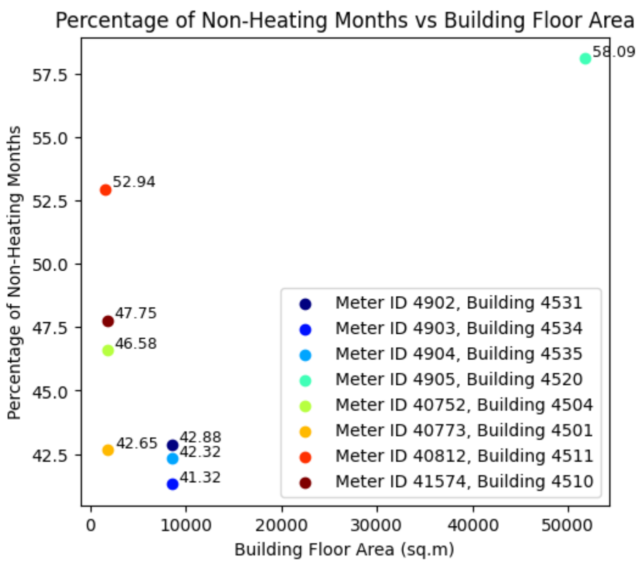


Figure 5.10.: Percentage of Hot Tap Water v.s Building Floor Area

yielding a percentage indicative of the gas consumption specifically allocated to hot tap water.

To estimate space heating demand, the gas consumption for hot tap water was reduced by the total gas consumption, including contributions from both the CHP system and the boiler. The resulting value indicates the gas consumption dedicated to space heating. This methodology aids in understanding the allocation of the heating system's gas consumption towards hot tap water and space heating. The report from DUWO clearly indicates a significant consumption of gas for hot tap water, as evidenced by reference [REFERENCE].

The chart in Figure 5.10 displays the distribution of hot tap water usage across different meters in the building floor area. The data reveals that hot tap water accounts for about 45% of overall consumption across all meters, with the highest usage percentage being 58%. This information highlights significant disparities in hot tap water usage among the meters, with Meter 4905 having the highest usage. This could potentially be due to the central distribution system serving more than 10 buildings, which may result in heat losses during the transmission process. Meters 40812 and 41574, located in building 4510 and 4511 respectively, followed closely behind.

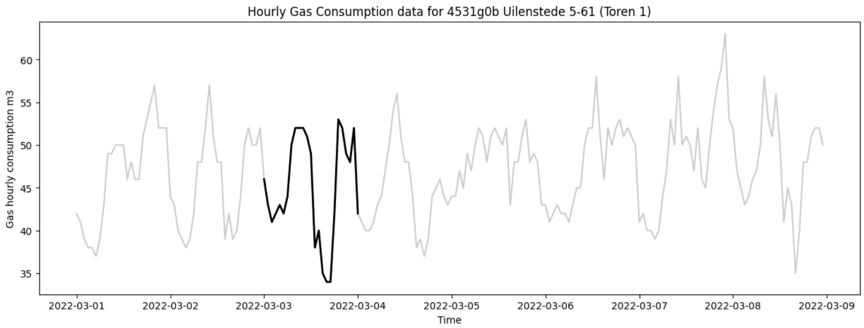
Interestingly, the hot tap water usage did not consistently correlate with the building floor area. For example, Meter 40812, despite a smaller floor area, exhibited higher hot tap water usage than Meter 4903. In contrast, Meter 4905, with the largest floor area due to its function as a central heating system, showed the highest hot tap water usage.

These observations underscore the importance of considering factors other than floor area when examining consumption patterns. Potential influencing factors might include central heating distribution systems or unique building characteristics. A more comprehensive investigation is necessary to uncover the root causes of these disparities in hot tap water usage.

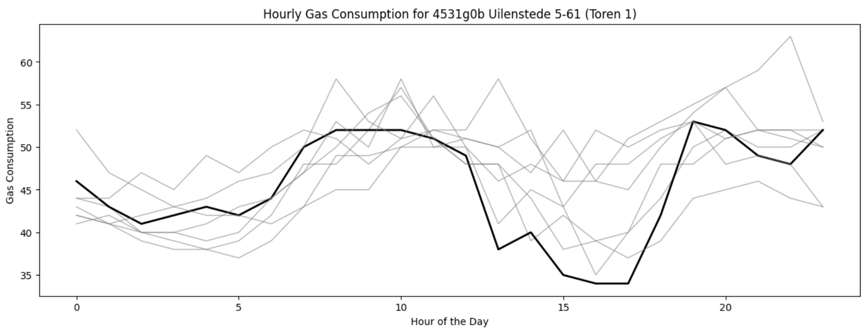
5.3. Results from Single Building Time Series K-Means Clustering

The segmentation of the data is visually depicted in Figure 5.11b, where the hourly gas usage data for an entire week Figure5.11a is divided into daily profiles Figure5.11b. Daily gas usage profiles containing missing data were excluded from subsequent analysis during the segmentation process. Each of these daily profiles comprises 24 data points, each corresponding to an hour of the day. Figure5.11a and Figure5.11b highlight the representation of the same set of data points through bold curves.

The Elbow Method 4.3, utilized to ascertain the optimal count of clusters, suggested two as the most appropriate number for most buildings. Buildings 4510 and 4511, however, presented an exception to this trend by showcasing three distinct clusters. The ideal cluster count, representing typical daily gas consumption profiles for each building, was determined through time series K-means clustering. The clustered distribution of daily usage profiles is observed to be relatively balanced across the clusters. These findings are summarized in Table A.4. For a more in-depth analysis, please refer to 'Single Building Clustering Result' and 'Typical Usage Profiles' in the appendix C and appendix D.



(a) Usage profiles before the data segmentation of building 4531

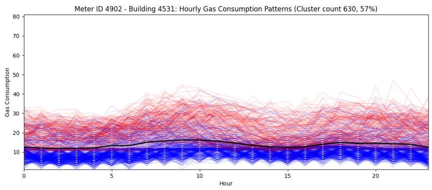


(b) Usage profiles after the data segmentation of building 4531

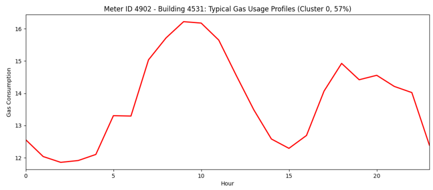
Figure 5.11.: Usage profiles for Building 4531

5.3.1. Typical Daily Usage Profiles of Building 4531

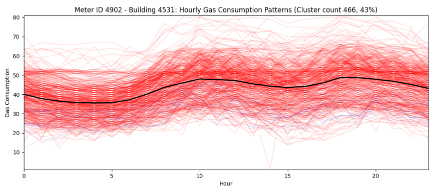
Constructed in 1970, Building 4531 is a high-rise structure housing approximately 360 units, predominantly utilized for group housing. This building also features an in-house Combined Heat and Power (CHP) system. For illustrative purposes, the clustering results for Building 4531 are demonstrated graphically, with the heating months distinguished by red lines and the non-heating months represented by blue lines. In these graphs, a single curve indicates the typical profiles in each cluster—calculated by averaging all the usage profiles within that cluster. The colored curves, on the other hand, represent all individual daily usage profiles within the cluster, as seen in Figure 5.12.



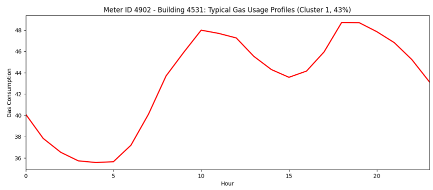
(a) Building 4531 Cluster Result Cluster 0



(b) Building 4531 Typical Usage Profile 0



(c) Building 4531 Cluster Result Cluster 1



(d) Building 4531 Typical Usage Profile 1

Figure 5.12.: Visualisation of Cluster Results for Building 4531

These graphical representations underline the seasonal variations in gas consumption, highlighting noticeable differences in both usage patterns and volumes between heating and non-heating months. As expected, heating months generally show higher gas consumption than summer months, reaffirming the significant impact of seasonal changes on gas usage.

Specifically, for Building 4531, the data suggests the presence of two distinct usage profiles, comprising 630 and 466 individual profiles, respectively, accounting for 57% and 43% of the data (see TableA.4 for more details). These clusters are predominantly evenly distributed, as visualized in Figure 5.12.

It's also observable that each cluster exhibits two prominent peaks of gas consumption. For Cluster 0, representing the non-heating months, the peaks occur around 9 am and 9 pm. In contrast, Cluster 1, representing the heating months, shows peaks around 10 am and 7 pm. This may be attributed to the longer daylight hours during the summer, leading to a wider interval between peak usage times. Notwithstanding the season, usage dips can be observed around 3 am and 3 pm, although the exact timing varies.

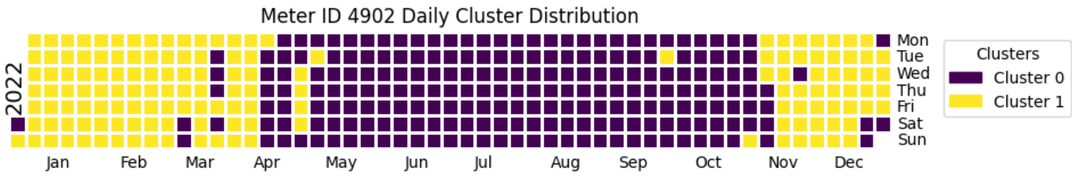


Figure 5.13.: Building 4531 Daily Cluster Distribution

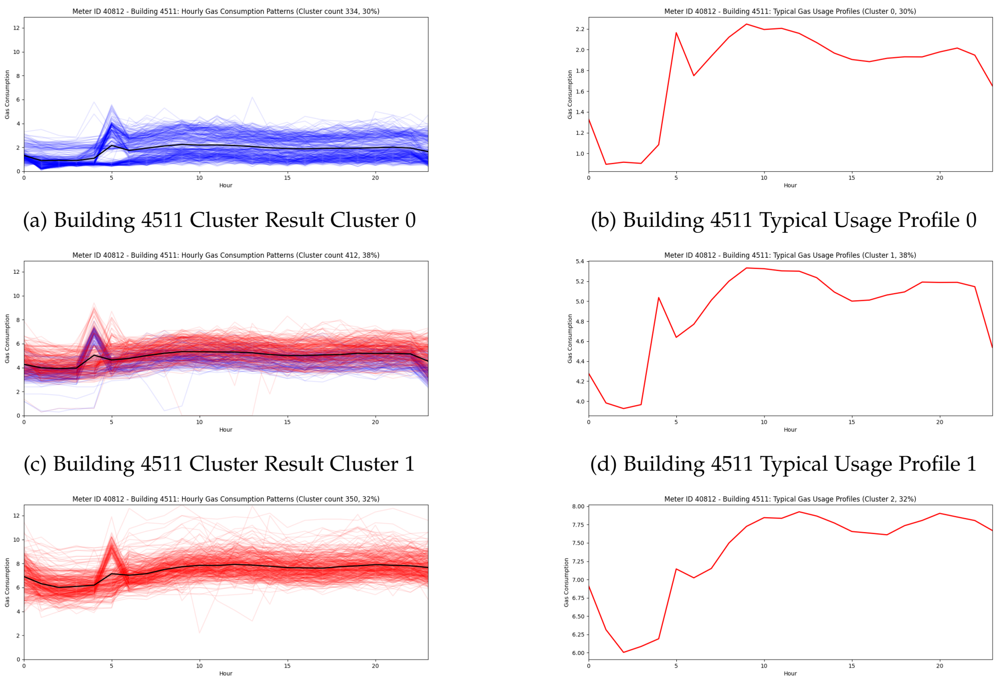
Figure 5.13 offers a clear visualization of the calendar distribution of daily gas usage profiles for Building 4531. In this figure, each block represents a distinct cluster, allowing the viewer to easily discern the prevalence of each cluster throughout 2022. This graphic illustration decorates the division of the two clusters, particularly noticeable between April and October.

Cluster 0, corresponding to non-heating months, exhibits dominance from April to October. During this interval, the weather generally remains mild, lessening the necessity for heating and consequently lowering gas consumption. Conversely, Cluster 1 symbolizes heating months where the chilling weather conditions necessitate amplified heating demands, thus escalating gas usage. The blocks in this cluster represent days where the gas consumption resonates with the distinct heating-month usage profile.

This illustrative representation underscores the temporal pattern in gas consumption, revealing a distinct seasonality in the usage profiles for Building 4531. The influence of seasonal climatic variations on these patterns is clear, further emphasizing the close-knit relationship between weather conditions and heating requirements.

5.3.2. Typical Daily Usage Profiles of Building 4511

Established in 1982, Building 4511 is a low-rise edifice housing 60 units predominantly comprised of studios and multi-room homes. The building boasts its own independent central heating system. The gas consumption patterns of buildings 4510, 4511, and 4504 in Figure 5.14 exhibit a distinctive trend observable during heating and non-heating months, as discussed in greater detail in Appendix D.



(a) Building 4511 Cluster Result Cluster 0

(b) Building 4511 Typical Usage Profile 0

(c) Building 4511 Cluster Result Cluster 1

(d) Building 4511 Typical Usage Profile 1

(e) Building 4511 Cluster Result Cluster 2

(f) Building 4511 Typical Usage Profile 2

Figure 5.14.: Visualisation of Cluster Results for Building 4511

These structures show a marked consumption spike around 6 am, coinciding with the period when boilers are typically heated for thermal disinfection. This process, crucial for the control of Legionella bacteria and other waterborne microbial contaminants, necessitates an early morning surge in gas use to maintain water temperatures between 60-70°C for a given duration.

For Building 4511, the peak usage moment surfaces around 10 am and 9 pm. Three distinct clusters were identified (see TableA.4 for more details): Cluster 0, representing non-heating months, accounts for 30% of the daily usage profile; Cluster 1, a mix of heating and non-heating months, contributes to 38% of the daily usage; and finally, Cluster 2, symbolizing the heating months, makes up 32% of the daily usage profile (see appendix C for reference). The peak moment in Cluster 2 displays a relatively lower spike at 6 am compared to other buildings. This is due to the relatively high overall gas consumption during the heating season, making the 6 am increase less pronounced. As with other buildings, the lowest gas usage consistently occurs around 3 am.

Interestingly, Building 4511 exhibits a minor fluctuation in gas consumption around 3 pm compared to other buildings. This variance can be attributed to the smaller number of units (60 in total, see Table A.3) and the presence of its independent central heating system, both factors that potentially minimize heat losses compared to centralized systems. Consequently, this leads to lower gas consumption when adjusting to changes in heating demand.

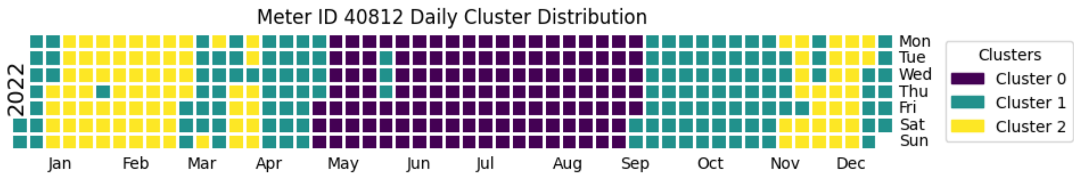


Figure 5.15.: Building 4511 Daily Cluster Distribution

Figure 5.15 showcases the calendar distribution of the daily gas usage profiles for building 4511, with each distinct block signifying a different cluster. The clusters visually depict various periods of gas usage throughout the year, each corresponding to different seasonal demands.

Cluster 0 represents the summer months, during which gas consumption is at its lowest. This lower demand for gas in the summer can be attributed to warmer temperatures, which reduce the need for space heating. The blocks representing this cluster indicate the days when daily gas consumption aligns with the typical summer usage profile.

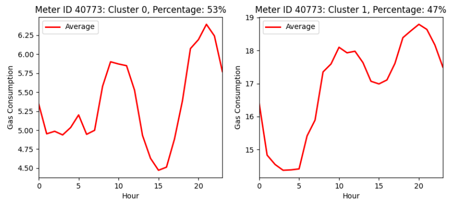
Cluster 2, on the other hand, represents the winter months, typically spanning from late November to late February. Gas consumption during this period is relatively higher due to the increased demand for heating in the colder weather. This cluster shows days with a usage pattern consistent with the typical winter profile, where heating needs drive the consumption to its peak.

Lastly, cluster 1 embodies the transitional periods, notably the spring and autumn months and the holiday period around Christmas. The days represented in this cluster are interspersed between the heating and non-heating months, leading to a mix of high (red) and low (blue) usage profiles in Figure5.14. This mixed profile is due to the varying weather conditions during these transitional seasons, which can fluctuate between warm and cold days. This variability could also lead to a minor increase in fluctuation around 3 pm, primarily because these periods of the year exhibit more variance in gas usage due to changing weather patterns.

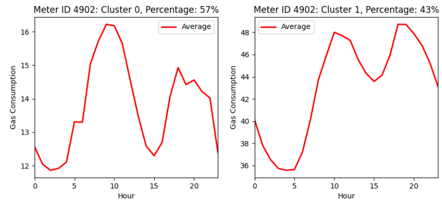
5.3.3. Variation of Single Building Typical Gas Usage Profiles

The following figures 5.16 present typical usage profiles for the specified meter IDs: 40773, 4501, 4902, 4531, 4905, 4520, and 40812, which correspond to Buildings 4511, 4501, 4520, and 4531, respectively.

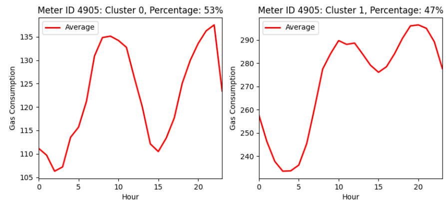
Building 4520 employs a central Combined Heat and Power (CHP) system that distributes energy to more than ten building units while Building 4531 operates its own CHP system within its premises. Building 4501, on the other hand, is a low-rise group house.



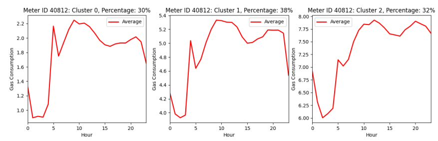
(a) Building 4501



(b) Building 4531



(c) Building 4520



(d) Building 4511

Figure 5.16.: Gas usage profiles

All the buildings house student accommodations. The distribution of daily usage profiles within each building reveals a common pattern among the top three profiles, characterized by clear peaks and dips. The clusters' distribution is roughly balanced, with each representing around 50% of the instances. These clusters illustrate gas usage patterns during both non-heating (Cluster 0) and heating (Cluster 1) months. The profile for Building 4511, which displays three similar clusters, deviates slightly due to the distinct operation of its central system and heating schedule.

In conclusion, the usage patterns across all buildings are quite similar, presumably because they all serve as student accommodations. The slight differences in each building's typical profile might be attributed to variations in their heating systems or the types of housing units present. However, the overall differences among these meters are minimal.

5.4. Results from Inter-Building Time Series K-Means Clustering

To cluster the gas consumption data across all meters, each consumption data point was scaled by its respective floor area. This normalization ensures that variations in building size do not affect the analysis, allowing for a focused investigation into potential commonalities or patterns across different meters. A more detailed explanation of this methodology is provided in the appendix under 'K-Means Global Building Clustering Result.' [to be finished appendix]

The employment of intra-building clustering analysis led to identifying 8706 unique daily usage profiles in a 24-dimensional space across 8 meters (refer to Table3.5 for more details). The choice of forming 8 clusters reflects the number of building meter ids, facilitating an exploration of whether global clustering could differentiate between different meter ids across all buildings through inter-building clustering. The representative profiles for each cluster were deduced by averaging the daily usage profiles categorized into each cluster. Considering the significant computational demands of clustering three years' worth of hourly building data, potential future investigations could look into alternative methods for clustering daily usage profiles across different buildings.

Each cluster presented its unique characteristics:

- **Cluster 0:** Building 4504 makes up most of the smallest cluster. The energy consumption of this cluster shows a significant increase at around 6 am, which is the usual time for boiler heat-up for thermal disinfection. This trend is also observed in buildings 4510 and 4511. Additionally, energy consumption peaks again at 9 pm, with a slight decrease seen at 3 pm in the afternoon.
- **Cluster 1:** The main group consists of meter readings from buildings 4531, 4534, and 4535. These readings are only during non-heating months and show a low average hourly consumption compared to other groups. There is a slight increase around 6 am for thermal disinfection.
- **Cluster 2:** The majority of this cluster is made up of just two buildings - 4504 (80%) and 4501 (20%). The fact that they have similar gas usage patterns may indicate that they are similar in terms of their building type and the people who occupy them.
- **Cluster 3:** This group, primarily consisting of Meter ID 4501 and 4504, has a consistent daily consumption pattern with slight increases at approximately 11 am and 7 pm. The higher usage during colder months may be due to shorter daylight hours, which explains the 8-hour gap between the two peaks.
- **Cluster 4:**This group of buildings, specifically 4510, 4511, and 4520, tend to use more energy at 9 am and 9 pm. This trend is more noticeable during the months when heating is not required. The longer daylight hours during these months may be a contributing factor to the 12-hour gap between the peaks.
- **Cluster 5:** This group of readings includes a fair mix of data from all the buildings, and Meter IDs 4511 and 4510 account for one-third of the daily profiles. The consumption pattern is similar to that of clusters 3 and 4, which are mainly during the heating months.

- **Cluster 6:** This cluster has an even distribution of readings from Meter IDs 4531, 4534, 4535, and 4520. The consumption pattern is similar to clusters 3, 4, and 5, which are mostly during heating months.
- **Cluster 7:** The data in this cluster comes mostly from buildings 4504 and 4501 and accounts for 99% of the readings. The way energy is being used is similar to the pattern seen in cluster 1.

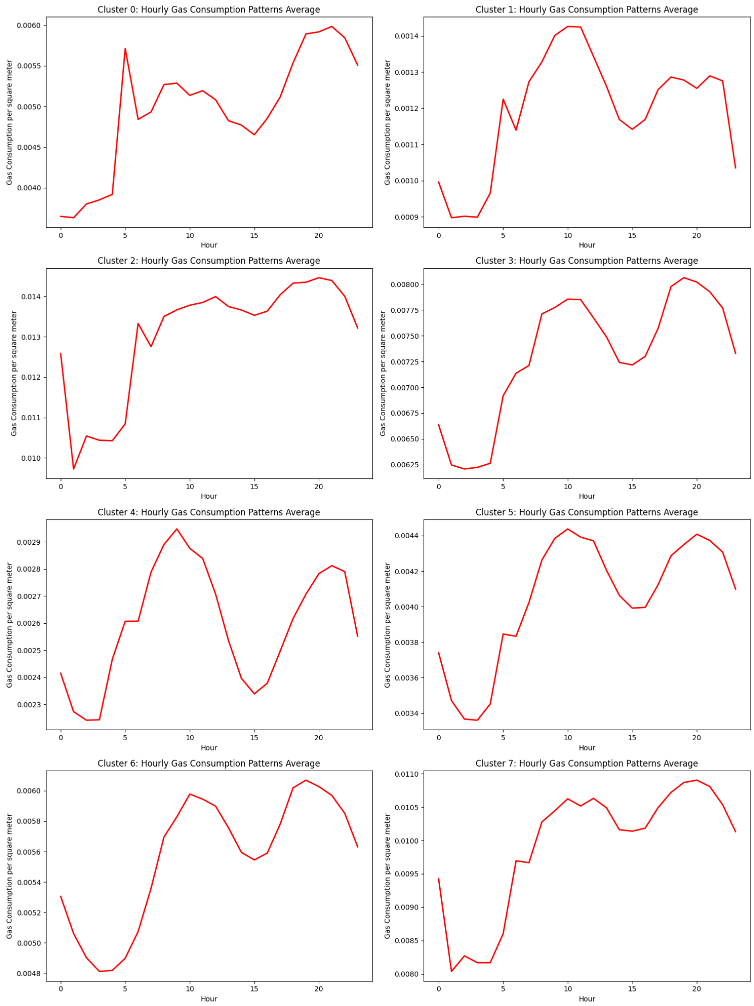


Figure 5.17.: Daily usage profiles across all buildings

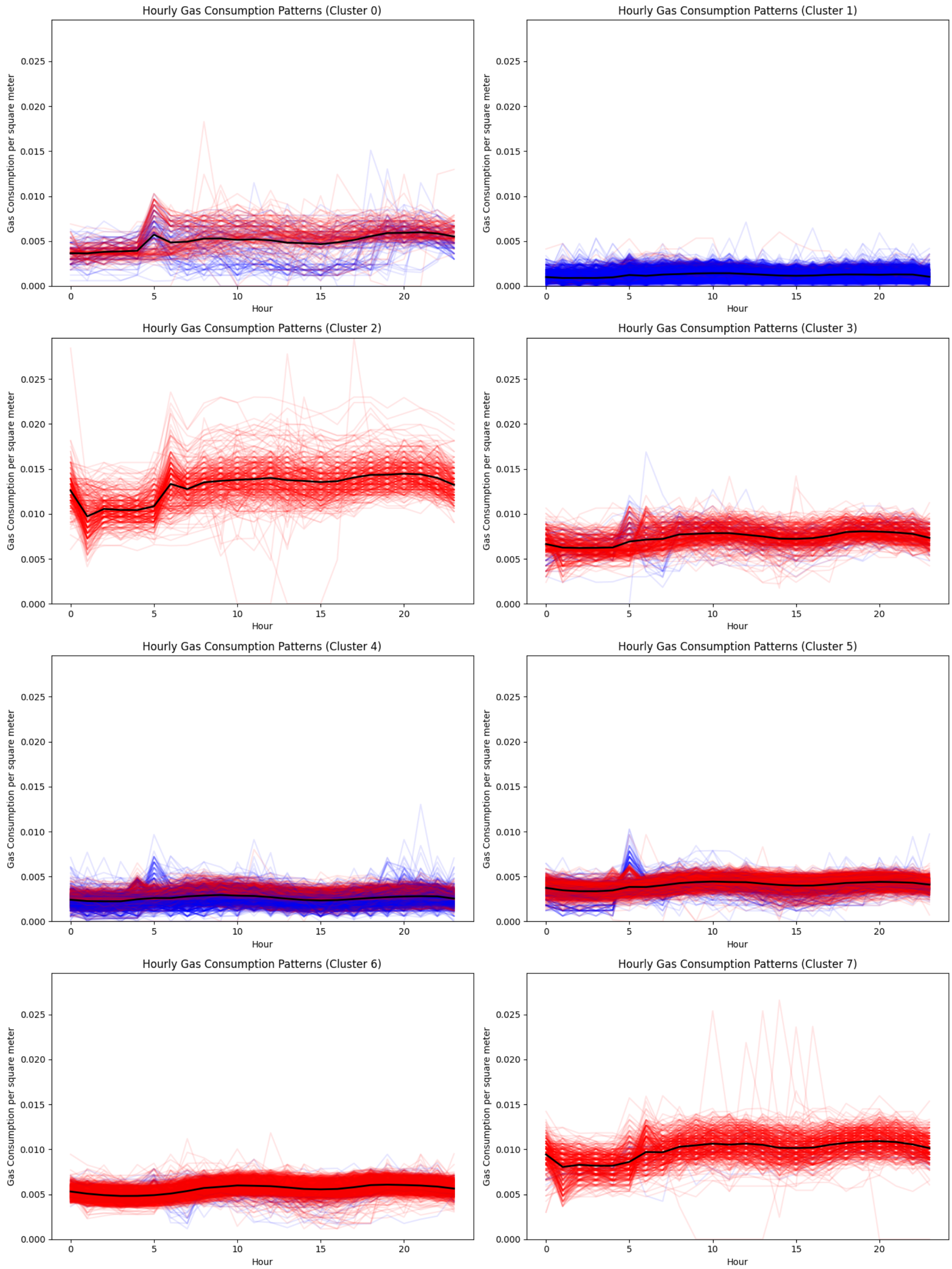


Figure 5.18.: Cluster result of all buildings

5.4.1. Building feature in each cluster

The data indicate a broad distribution of buildings across eight clusters, each characterized by distinct building features and gas consumption levels. This report aims to dissect these clusters, comprehensively understanding their characteristics and drawing crucial conclusions from an energy efficiency standpoint.

Cluster 1 - This cluster, comprising 3012 buildings, has the lowest average gas consumption. It features a diverse mix of buildings constructed between 1970 and 2014. The prevalent orientation is westward, and the buildings demonstrate relatively high insulation values for the floor, roof, and facades. This suggests effective insulation could be a contributing factor to their energy efficiency. The insulation level of the windows is also satisfactory.

Cluster 4 - With the lowest average gas consumption, this cluster is primarily composed of buildings constructed in 2014. These buildings exhibit high insulation values for the floor, roof, and facades, which may account for their energy efficiency. Like Cluster 1, the buildings in this cluster are predominantly oriented towards the west.

Contrastingly, clusters with elevated average gas consumption, namely **Cluster 2** and **Cluster 7**, feature buildings with relatively lower insulation levels. This implies that inadequate insulation may be a contributing factor to their higher energy use for heating.

To understand these clusters further, we delve into a detailed analysis of each cluster:

Cluster 0 (265 buildings)	
Year	Mostly constructed in 1991
Insulation	Floor: 1.36 m^2 K/W Roof: 1.69 m^2 K/W Facades (excluding AOR): 1.02 m^2 K/W Windows: 1.85 W/(m^2 K)
Consumption (per m2)	Lower average gas consumption at 0.00496
Cluster 1 (3012 buildings)	
Year	Primarily built in 1972
Insulation	Floor: 1.75 m^2 K/W Roof: 2.06 m^2 K/W Facades (excluding AOR): 1.53 m^2 K/W Windows: 2.11 W/(m^2 K)
Consumption (per m2)	Less gas consumption on average, with a value of 0.00128
Cluster 2 (339 buildings)	
Year	Mainly built in 1991
Insulation	Floor: 1.3 m^2 K/W Roof: 2.22 m^2 K/W Facades (excluding AOR): 2.0 m^2 K/W Windows: 1.80 W/(m^2 K)
Consumption (per m2)	Highest average gas consumption among the clusters at 0.01303

Cluster 3 (524 buildings)	
Year	Constructed primarily in 1991
Insulation	Floor: 1.66 m² K/W Roof: 2.29 m² K/W Facades (excluding AOR): 1.76 m² K/W Windows: 2.01 W/(m² K)
Consumption (per m2)	Average gas consumption, with a value of 0.00729
Cluster 4 (2660 buildings)	
Year	Mostly built in 1976
Insulation	Floor: 1.85 m² K/W Roof: 2.30 m² K/W Facades (excluding AOR): 1.81 m² K/W Windows: 2.07 W/(m² K)
Consumption (per m2)	Lower average gas consumption, with a value of 0.00256
Cluster 5 (2302 buildings)	
Year	Constructed primarily in 1977
Insulation	Floor: 1.85 m² K/W Roof: 2.36 m² K/W Facades (excluding AOR): 1.85 m² K/W Windows: 2.10 W/(m² K)
Consumption (per m2)	Average gas consumption of 0.00406
Cluster 6 (2491 buildings)	
Year	Constructed mainly in 1978
Insulation	Floor: 1.74 m² K/W Roof: 2.16 m² K/W Facades (excluding AOR): 1.70 m² K/W Windows: 2.08 W/(m² K)
Consumption (per m2)	Higher average gas consumption at 0.00555
Cluster 7 (463 buildings)	
Year	Mainly built in 1991
Insulation	Floor: 1.3 m² K/W Roof: 2.22 m² K/W Facades (excluding AOR): 2.0 m² K/W Windows: 1.80 W/(m² K)

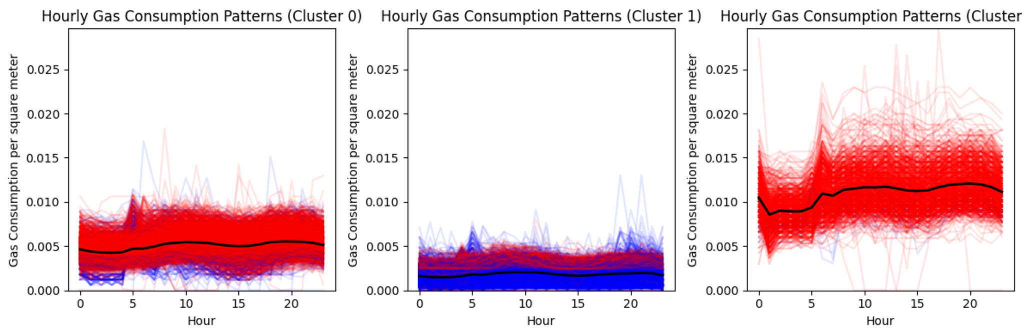
In conclusion, despite the variance in high and low consumption values across clusters, consumption patterns have a degree of uniformity. This suggests a typical usage pattern prevalent across all buildings, regardless of factors such as construction year, insulation values, or building size. See appendix F

5.5. Comparing Proposed Clustering Strategy with Elbow Method for Inter-Building Clustering

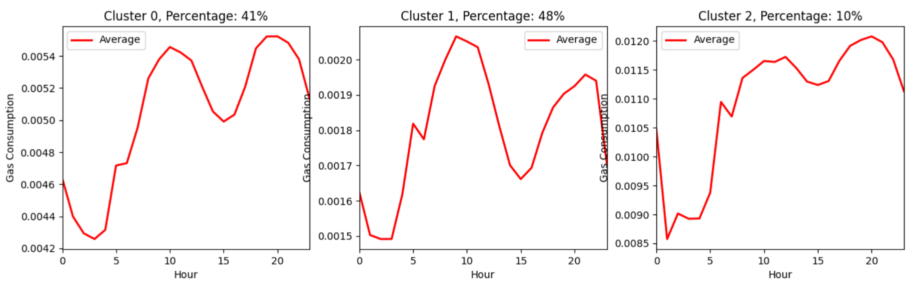
After reviewing the data, we have identified some areas where our clustering approach could be improved. Specifically, we will focus on refining our use of Dynamic Time Warping (DTW) to evaluate the similarity between time-series data more accurately. We will also explore alternative clustering methods to determine if they could produce better results. Our goal is to find the optimal number of clusters for inter-building clustering and improve our approach's overall accuracy and effectiveness.

Upon applying the Elbow Method to our dataset, it was determined that the optimal number of clusters was three, representing most of the daily profiles at 41% and 49%, respectively. In addition, a smaller cluster was identified, which constituted roughly 10% of the daily profiles primarily composed of data collected from buildings 4501 and 4502.

Figure 5.19 presents the clustering results and the typical daily gas usage profiles. Comparing clusters 0 and 1, both have peak consumption times around 9-10 am and 8-9 pm. However, their crucial difference lies in their average consumption per floor area. As shown in Figure 5.19a, cluster 1 mainly consists of data from non-heating months, resulting in lower overall gas consumption. Cluster 2 represents buildings 4504 and 4501 due to their gas consumption spike at 6 am for thermal disinfection.



(a) Daily profiles of each cluster



(b) Averaged daily profiles of each cluster

Figure 5.19.: Result of the Inter-Building Clustering using Elbow Method

Table 5.2.: Cluster Statistics

Cluster	Year Highest Proportion	Avg. Roof Insulation (m^2 K/W)	Avg. EP1 (kWh/m^2)	Avg. EP2 (kWh/m^2)	Avg. EP2 EMG forf. (kWh/m^2)	Average Insulation Floor [m^2K/W]
0	2014	2.91	100.64	180.85	180.85	2.28
1	1970	2.52	122.47	195.02	195.02	1.91
2	1991	2.21	122.22	201.30	201.30	1.30

Cluster	Avg. Insulation Facades Excl. AOR (m^2 K/W)	Avg. Facades Ex AOR (m^2)	Avg. Insulation Windows (W/m^2 K)	Avg. Until. Window(m^2)	Non-Heating Month Percentage
0	2.18	13.83	2.10	10.10	0.51
1	1.84	14.26	2.10	10.59	0.50
2	2.00	29.73	1.80	9.28	0.45

In Table G.4, we present each cluster's most common attribute values, which provide a representative snapshot of the characteristics within each cluster. The detailed information on each cluster's features is in appendix G.0.3. The relative proportion of each cluster in the dataset is illustrated in Figure 5.21. Furthermore, Figure 5.20 outlines the consumption per unit of floor area for each respective cluster.

EP1 and EP2 are commonly used energy performance indicators in the field of building energy analysis.

- EP1 (Energy Performance Indicator 1): EP1 is a metric representing the primary energy consumption per unit area of a building. It quantifies the amount of primary energy required to meet the energy demands of a building, including heating, cooling, lighting, and other energy uses. EP1 is typically measured in kilowatt-hours per square meter (kWh/m^2) and indicates the overall energy efficiency of a building.
- EP2 (Energy Performance Indicator 2): EP2 is another energy performance indicator that focuses on the energy performance related to heating demand in a building. It specifically measures the heating energy consumption per unit area. Like EP1, EP2 is also measured in kilowatt-hours per square meter (kWh/m^2) and helps assess the heating system's energy efficiency and insulation measures in a building.

Both EP1 and EP2 are useful metrics for evaluating buildings' energy efficiency and performance, allowing for comparisons between different buildings or analyzing the effectiveness of energy-saving measures.

5.5.1. Cluster Summaries

Cluster 0:

This cluster demonstrates moderate gas consumption, with an average of 0.005 per hour per floor area($m^3/hr/m^2$). This cluster includes predominantly recently constructed buildings from 2014 (33.6%) and older buildings from 1970 (30.7%). These structures show high insulation values and lower Energy Performance (EP) values across multiple components. Such characteristics suggest that the buildings in this cluster were designed with a strong focus on energy efficiency.

Cluster 1:

The buildings in Cluster 1 have the lowest gas consumption among all clusters, with an average gas consumption of 0.0018 per hour per floor area($m^3/\mathrm{hr}/m^2$). Cluster 1 is characterized by buildings primarily built-in 1970 (29.3%) and 2014 (26.4%). However, this cluster has higher average EP1 and EP2 values than Cluster 0, suggesting a higher energy consumption. Despite the lower insulation levels for floors and facades (excluding AOR) compared to Cluster 0, Cluster 1 still records the lowest energy consumption per floor area. This could be attributed to the dominance of non-heating months in these buildings.

Cluster 2:

The buildings in Cluster 2 consume the most gas on average, with a value of 0.0109 per floor area($m^3/\mathrm{hr}/m^2$). The buildings in this cluster were predominantly built around 1991. The insulation levels for various elements like floors, roofs, facades, and windows are lower than the other clusters. This implies these buildings may be less well-insulated, leading to higher energy consumption per floor area.

Additional Observations:

Scheduled heating systems operating year-round can significantly contribute to gas consumption, depending on the system specifics. A deeper investigation into heating schedules and disinfection strategies can offer further insights into energy conservation measures. It's also worth exploring why the energy consumption per floor area in Cluster 2 is almost double that of Cluster 0 to identify potential areas for improvement in energy performance.

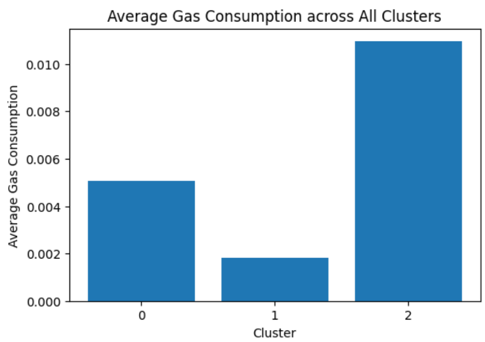


Figure 5.20.: Gas consumption per floor area Inter-Building Clusters

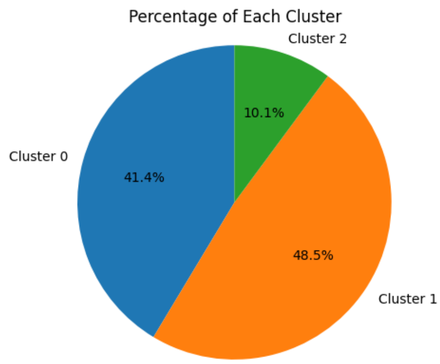


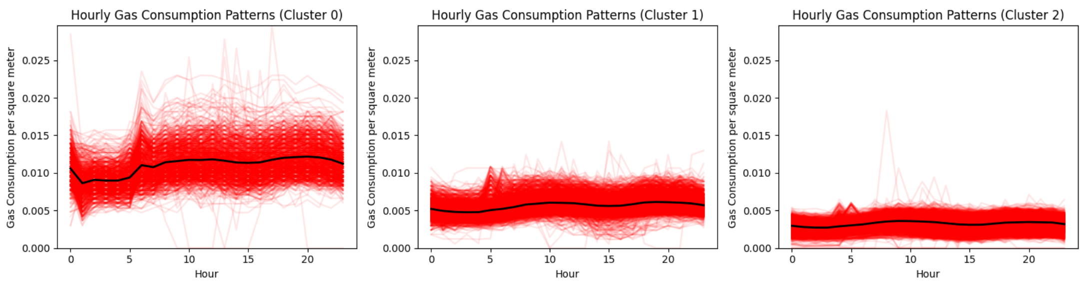
Figure 5.21.: Persentage of Daily usage profiles for each Inter-Building Cluster

5.6. Seasonal Clustering: Building features in each inter-building cluster for winter months

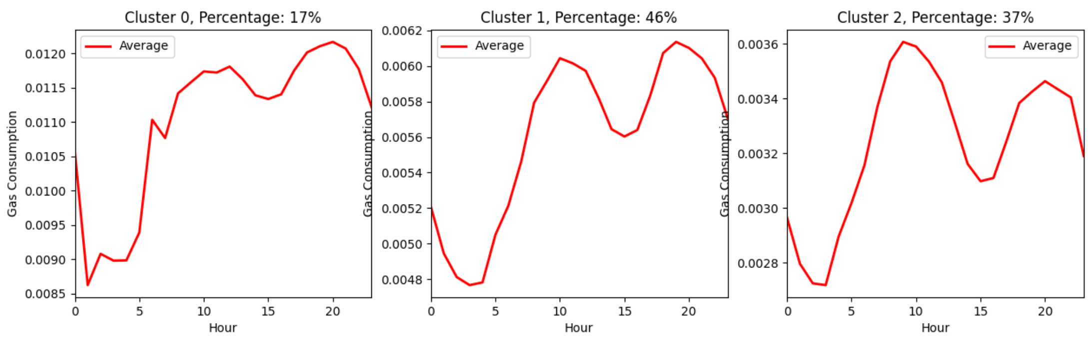
This subsection concentrates on the gas consumption data collected during heating months across all meters. Our primary objective is to explore the influence of heating month variations on the patterns and energy usage trends. A comprehensive explanation can be found in the appendix H.

The Elbow Method was employed to determine the optimal number of clusters, leading to three as the ideal number for heating months. As depicted in Figure5.22a, clusters 1 and 2 constitute similar proportions of total daily profiles (46% and 37% respectively, see Figure5.23a), with meter ids evenly spread among them. Conversely, cluster 0 represents 17% of total daily profiles and is predominantly associated with buildings 4501 and 4504, which exhibit the highest construction per floor area.

Figure5.22 exhibits the clustering outcomes and corresponding typical daily gas usage profiles. Clusters 1 and 2 reveal peak consumption periods around 9-10 am and 8-9 pm, but they differ significantly in average consumption per floor area. Figure5.23b shows that cluster 1 is characterized by higher overall gas consumption, while cluster 2 represents data with lower overall gas consumption. As noted in the prior section5.22, cluster 0 experiences a consumption spike at 6 am, attributed to the thermal disinfection schedule of the central heating system in buildings 4501 and 4504.

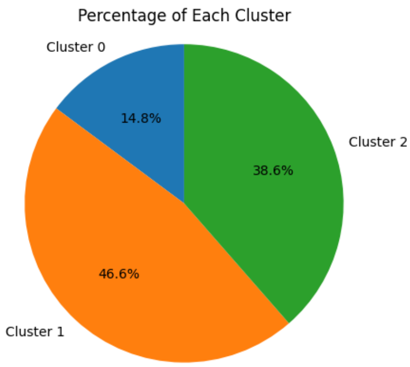


(a) Daily profiles of each cluster

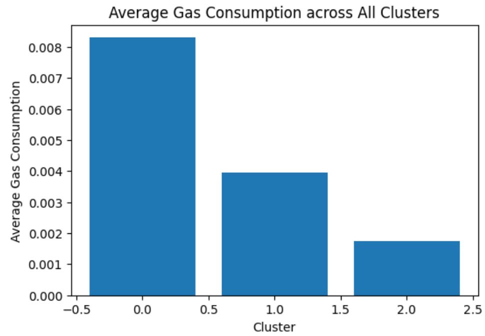


(b) Averaged daily profiles of each cluster

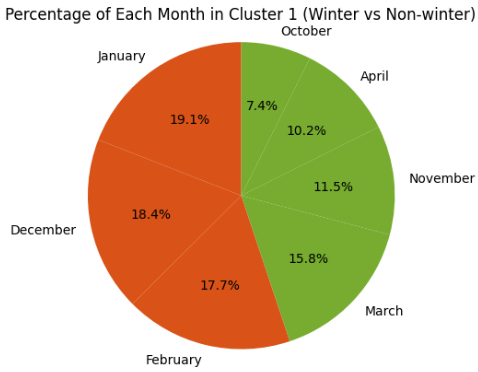
Figure 5.22.: Result of Inter-Building Clustering for winter months using Elbow Method



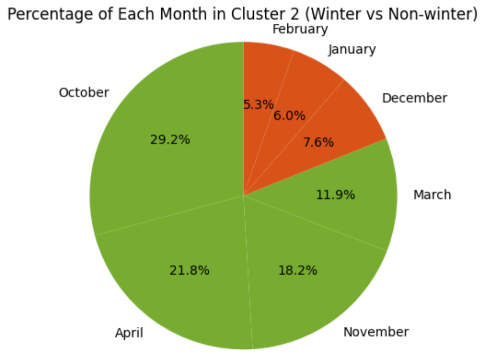
(a) Percentage of all clusters



(b) Gas consumption per floor area Inter-Building Clusters



(a) Percentage of Each Month in cluster 1



(b) Percentage of Each Month in cluster 2

Figure 5.24.

5.6.1. Cluster Summaries

Table H.4 outlines the most frequent values for each attribute within the clusters, thereby creating a characteristic profile for each cluster. The consumption patterns across the three clusters, namely 0, 1, and 2, indicate different levels of gas usage, with respective values of 0.011, 0.005, and 0.003 ($m^3/\text{hr}/m^2$).

Cluster 0

Cluster 0 encompasses buildings that exhibit the highest average gas consumption, clocking in at 0.011 ($m^3/\text{hr}/m^2$). Most buildings within this cluster were built around 1991. The insulation measures across various structural elements such as roofs, floors, facades, and windows are noticeably lower compared to the other clusters, suggesting that these buildings might not be as well-insulated, leading to elevated energy consumption per floor area. Key specifications for this cluster include:

Table 5.3.: Cluster Statistics

Cluster	Year Highest Proportion	Avg. Roof Insulation (m^2 K/W)	Avg. EP1 (kWh/m^2)	Avg. EP2 (kWh/m^2)	Avg. EP2 EMG forf. (kWh/m^2)	Average Insulation Floor [m^2K/W]
0	1991	2.21	122.22	201.3	201.3	1.30
1	2014	2.98	91.17	174.96	174.96	2.41
2	1982	2.33	142.19	207.43	207.43	1.66

Cluster	Avg. Insulation Facades Excl. AOR (m^2 K/W)	Avg. Facades Ex AOR (m^2)	Avg. Insulation Windows (W/m^2 K)	Avg. Until. Window(m^2)
0	2.00	29.73	9.28	1.8
1	2.28	13.71	10.06	2.08
2	1.58	13.10	10.92	2.15

The average roof insulation level stands at 2.22 m² K/W. EP1 and EP2 remain consistent across all buildings, amounting to 122.22 kWh/m² and 201.29 kWh/m², respectively. The average insulation level for floors is approximately 1.3 m² K/W, and for facades, it is about 2 m² K/W. Window insulation averages at about 1.79 W/(m² K).

Cluster 1

Cluster 1 records moderate gas consumption, with an average value of 0.005 (m^3/hr/m^2). This cluster primarily consists of newly constructed buildings from 2014 (35.7%) and older ones from 1970 (33.7%). Key specifications for this cluster include:

Buildings display significant variations in roof insulation, with an average value of around 2.98 m² K/W. EP1 and EP2 also display variations, with average values being 91.17 kWh/m² and 174.95 kWh/m², respectively. The average floor insulation level is around 2.4 m² K/W, and for facades, it is about 2.28 m² K/W. Window insulation averages around 2.08 W/(m² K).

Cluster 2

Buildings in Cluster 2 register the lowest gas consumption among all clusters, averaging at 0.003 (m^3/hr/m^2). Predominantly, buildings in this cluster were constructed in 1982 (35.1%). Interestingly, this cluster displays higher average EP1 and EP2 values than Cluster 1, indicating more energy consumption. Despite the lower insulation measures for floors and facades in comparison to Cluster 1, Cluster 2 still records the lowest energy consumption per floor area. Key specifications for this cluster include:

1. There is a noticeable variation in roof insulation, with an average level of around 2.33 m² K/W.
2. EP1 and EP2 vary across buildings, with average values being 142.18 kWh/m² and 207.43 kWh/m², respectively.
3. The average floor insulation level is around 1.65 m² K/W, and it is about 1.57 m² K/W for facades. Window insulation averages around 2.15 W/(m² K).

Additional Observations

Despite Cluster 1 comprising a higher percentage of newer buildings than Cluster 2 (Cluster 1: 35.7% from 2014, Cluster 2: 23.7% from 2014), and having lower EP values in general along with higher insulation measures, its gas consumption remains higher than that of Cluster 2.

This situation could be due to various factors, such as:

- Central heating: The schedule of heating systems operating year-round can significantly contribute to gas consumption. The specifics of these systems need further investigation to identify potential areas for improvement in energy performance.
- Building age: The age of a building is often used as a predictor of energy consumption. However, it has been observed that buildings constructed in 1982 consume less gas than those built in 1991, despite having lower insulation values. This observation may be attributed to the cluster results that capture months with lower consumption. Further investigation is needed to understand this anomaly and its implications for energy efficiency in building construction.
- Winter months: In the following figure5.24, the distribution of months for heating and non-heating periods can be observed for both cluster 1 and cluster 2.

  Cluster 2, despite having a higher overall gas consumption, seems to be more active during the transitional periods between seasons, specifically in October, April, and November, which often feature milder temperatures. These months collectively constitute about 69.2% of the total months for Cluster 2.

  On the other hand, Cluster 1 has a higher concentration of winter months, with January, December, and February collectively constituting about 55.2% of the total months. These months are traditionally colder and, hence, require more heating. The increased heating demand during these months could explain the higher gas consumption observed for Cluster 1, even though it generally features more recent constructions and better insulation levels.

  Therefore, the difference in heating schedule and the concentration of heating demand in different months between these clusters could be a significant factor explaining the lower gas consumption in Cluster 2 compared to Cluster 1. These insights highlight the importance of considering the building characteristics and insulation levels, the heating schedule, and the specific months during which heating is most required when evaluating a building's energy performance.

5.7. Conclusion

This research has undertaken a comprehensive exploration of gas usage patterns within and across multiple buildings with the utilization of K-Means Clustering. Several findings have been made through evaluating and testing the proposed strategy.

After conducting Single Building Time Series K-Means Clustering, it was found that all buildings considered, which serve as student accommodations, have similar gas usage patterns. However, slight variations in their profiles were observed due to differences in heating systems, consumption levels in different months, and housing unit types. Despite these differences, they were minimal overall.

In Inter-Building Time Series K-Means Clustering, eight distinct clusters were identified, each exhibiting unique characteristics regarding gas usage patterns and building features.

These clusters varied in size, structure, heating methods, and insulation levels, thus influencing their gas consumption levels.

The comparative analysis of our proposed strategy with the Elbow Method led us to an optimal number of three clusters. Despite fewer clusters, this method revealed distinct usage patterns, particularly in peak consumption times and average consumption per floor area, emphasizing the behavior pattern consistency across buildings.

During winter, there were similar trends observed in seasonal clustering as before. The impact of thermal disinfection schedules on gas consumption spikes was especially significant in Cluster 0, highlighting the importance of heating schedules in overall energy usage. The differences in consumption between clusters 1 and 2 are relatively large and may contribute to the overall proportion of winter months resulting in higher consumption in cluster 1. However, further analysis is necessary to determine the extent of this impact on consumption.

In conclusion, this study offers a multifaceted perspective on building gas consumption patterns, which are influenced by various factors, from insulation levels and building orientation to heating schedules and building types. The knowledge from this research can be instrumental in guiding energy efficiency initiatives, thereby encouraging more sustainable building practices. However, future investigations may explore alternative methods for clustering daily usage profiles across different buildings.

6. Conclusion Discussion and Reflection

6.1. Conclusion

This study proposed an efficient clustering-based method to discern the daily utility consumption profiles across multiple academic structures. This approach initially determined the average daily usage profiles of each building via time series k-means clustering. The dissimilarity measure was gauged using dynamic time wrapping (DTW) distance to collate the daily usage profiles of all individual buildings.

This method's efficacy was assessed using three years' worth of hourly power consumption data sourced from 20 structures on a university campus in Amsterdam. The findings revealed that this approach could effectively discern pertinent details relating to the behavioral patterns of gas usage. This was noted in the averaged daily gas consumption profiles identified via single-building clustering and the average of original daily usage profiles corresponding to the inter-building typical profiles.

The energy usage patterns identified by this proposed strategy offer insights that can be harnessed to categorize structures exhibiting similar gas usage behaviors. This, in turn, can guide decision-making processes for implementing efficient retrofits and enhancing performance.

Furthermore, the information is valuable in establishing advanced energy management protocols and devising fault detection and diagnosis strategies for building structures. It suggests that the proposed method holds the potential to be adapted and employed in the energy planning of campus buildings.

6.2. Discussion

This result also outlines the research method and its questions, focusing on three key aspects.

- **Understanding Natural Gas Consumption Patterns:** The data collated aims to comprehend gas consumption within the Uilenstede campus and understand the energy behavior in student accommodations.
- **Influence of Building Features on Energy Efficiency:** As discussed in section 5.4.1, examining the building features within each cluster provides insights into their potential impact on consumption levels. Factors such as seasonality, occupancy schedules, and specific features of each building (e.g., insulation, occupancy schedules, heating system efficiency, and central heat losses), can be considered. Additionally, providing detailed and standardized features of each building can help refine the clustering results and provide a more comprehensive understanding of each cluster.

The evaluation of potential energy savings considers several factors, including indoor temperature, ventilation system, and heating inertia of dwellings. However, there remain some factors that are not currently considered. These include the actual adjustments of consumption practices, which could directly impact potential savings. As part of future research, it would be interesting to delve deeper into these factors, further refining the connection between consumption practices elements and their influence on energy consumption efficiency and potential savings.

6.3. Reflection

In the context of our project, the research methodology facilitated the identification of distinct gas consumption patterns in university accommodations, allowed us to spot potential irregularities, and helped determine the factors influencing energy efficiency. These insights from our research underpinned the design of our clustering-based strategy, providing a broader understanding of potential approaches for enhancing building consumption.

Our proposed strategy offers a systematic framework for profiling the daily usage of buildings. This framework can support informed decision-making processes for the implementation of energy-efficient retrofits and performance improvements. Moreover, it could serve as an instrumental tool in developing advanced strategies for building energy management, as well as fault detection and diagnosis.

In sum, the project achieved its targeted innovation, demonstrating the practical application of consumption data within a student campus, while showcasing the iterative process inherent in research and design, particularly within the sphere of machine learning and clustering methods tied to building features.

Future strategies might consider incorporating additional information such as indoor temperature, building heating schedules, heat storage systems, and their efficiencies to optimize the heating system. As occupant schedules can be challenging to obtain, one alternative approach could be comparing summer and winter consumption data to discern the differences between indoor heating demand and hot water demand.

Understanding the utility of different types of data is crucial. While additional data can be valuable, discerning which data is impactful and which is not, is vital. Heating demand could potentially be calculated more accurately using a physics-based model that leverages simulation software and information about building insulation and temperature differences. However, the trade-off between the level of accuracy and efficiency should be considered, especially in a complex setting such as a student campus.

6.3.1. Additional data sources

Integrating additional data sources, such as building features and weather information, can potentially provide more detailed and meaningful clusters regarding energy consumption. This multi-faceted approach would allow us to consider the interplay of various factors that can impact energy use.

- Building Features: This includes information such as the year of construction, building type (residential, commercial, industrial), size (floor area, number of floors), occupancy, equipment installed (heating, cooling, lighting systems), and insulation levels. These features can affect how a building uses energy.
- Weather Data: Temperature, humidity, solar radiation, wind speed, and precipitation are all weather factors that can greatly influence a building's energy consumption, particularly for heating and cooling. By incorporating weather data, you can better understand fluctuations in energy usage and isolate weather-dependent trends.

- Time-Series Data Analysis: Time-series data, like hourly gas consumption data, is rich in information. Statistical properties such as mean, median, standard deviation, seasonality, and trend can provide insights into consumption patterns. Additionally, advanced techniques like Fourier analysis can extract periodic patterns from the data.

However, it's important to note that integrating more data also increases the complexity of the analysis and may require more sophisticated machine learning or statistical techniques to interpret the data accurately. Furthermore, data quality, availability, and the need for data pre-processing can also be challenges in this approach.

It's often the case that not all additional information is readily available. Focusing on effectively analyzing and extracting insights from the available data is crucial in such cases. This could involve exploring different ways of structuring the data, applying more advanced data analysis techniques, or creating new derived variables that capture essential data characteristics.

A. Appendix

Table A.4.: Number of clusters and their percentage in each building

Building Number	Number of Clusters	Cluster ID	Percentage of Each Cluster
4531	2	0	57.49%
		1	42.51%
4534	2	0	53.34%
		1	46.66%
4535	2	0	52.72%
		1	47.28%
4520	2	0	52.65%
		1	47.35%
4504	2	0	52.07%
		1	47.93%
4501	2	0	53.14%
		1	46.86%
4511	3	0	30.51%
		1	37.69%
		2	31.80%
4510	3	0	32.52%
		1	28.05%
		2	39.43%

Table A.1.: Building information from EPA

Building no.	Construction year	Complex	Buildings	Number of homes
4520	1970	Center	Combination buildings	1,080
	2010		New high-rise buildings	724
	1970		VU guest house	320
4531	1970	Tower 1	High-rise building No.5-61	360
4534	1970	Tower 2	High-rise building No.155-211	360
4535	1970	Tower 3	High-rise building No.405-461	360
4501	1991	Uilenstede 2-32	High-rise building No. 2-16	92
			Low-rise no. 18-22	48
			Low-rise no. 24-32	80
4504	1991	Uilenstede 36-70	High-rise no. 36-50	92
			Low-rise no. 52-58	24
			Low-rise no 60-70	36
4510	1982	Uilenstede 102 ad	Low-rise no. 102	60
4511	1982	Uilenstede 102 eh	Low-rise no. 102	60
4508	1991	Uilenstede 72-98	High-rise building nr 72-86	92
			Low-rise no. 88-98	102
Total				3,890

Table A.2.: Building Numbers, ID Descriptions, and Floor Areas

Building no.	Complex Number	Meter ID	Floor Area (sq.m)
4531	4902	Uilenstede 5-61 (Toren 1)	8555
4534	4903	Uilenstede 155-211 (Toren 2)	8553
4535	4904	Uilenstede 405-461 (Toren 3)	8552
4520	4905	Uilenstede 106-460	51766
4504	40752	Uilenstede 36-70	1801
4501	40773	Uilenstede 2-32	1843
4511	40812	Uilenstede 102 e-h (trad)	1534
4510	41574	Uilenstede 102 a-d (H-inst)	1798

Table 7 provides an overview of the calculated consumption for the three types of residential buildings. Note that hot tap water accounts for approximately 45% of consumption.

Table 7. Calculated energy demand for hot tap water and space heating for 3 types of residential buildings on Uilenstede.

Building type	Tower	High-rise 92	Low-rise
number of homes	360		building 36
m² GFA	8,220	1,910	1,754
domestic hot water energy demand [GJ]	2,340	600	230
space heating energy demand [GJ]	3,150	710	520
Total energy demand [GJ]	5,490	1,310	750
Energy demand per dwelling [GJ]	15.3	14.2	20.8

Figure A.1.: Calculated energy demand from DUWO

Table A.3.: Building information from EPA

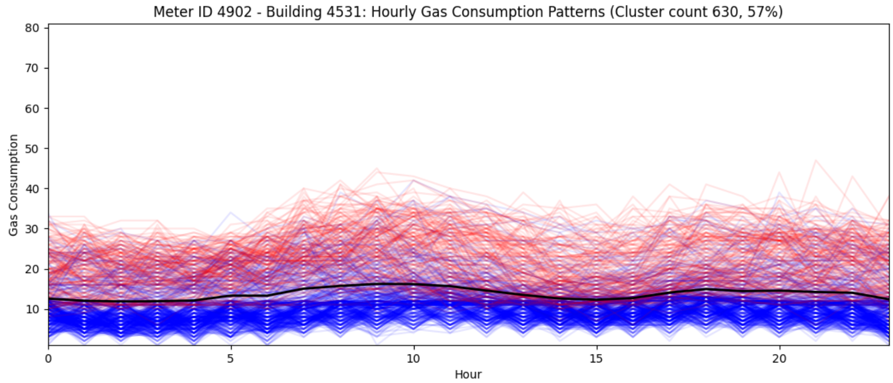
Building no.	Construction year	Complex	Buildings	Number of homes
4520	1970	Center	Combination buildings	1,080
	2010		New high-rise buildings	724
	1970		VU guest house	320
4531	1970	Tower 1	High-rise building No.5-61	360
4534	1970	Tower 2	High-rise building No.155-211	360
4535	1970	Tower 3	High-rise building No.405-461	360
4501	1991	Uilenstede 2-32	High-rise building No. 2-16	92
			Low-rise no. 18-22	48
			Low-rise no. 24-32	80
4504	1991	Uilenstede 36-70	High-rise no. 36-50	92
			Low-rise no. 52-58	24
			Low-rise no 60-70	36
4510	1982	Uilenstede 102 ad	Low-rise no. 102	60
4511	1982	Uilenstede 102 eh	Low-rise no. 102	60
4508	1991	Uilenstede 72-98	High-rise building nr 72-86	92
			Low-rise no. 88-98	102
Total				3,890

B. Building EPA information

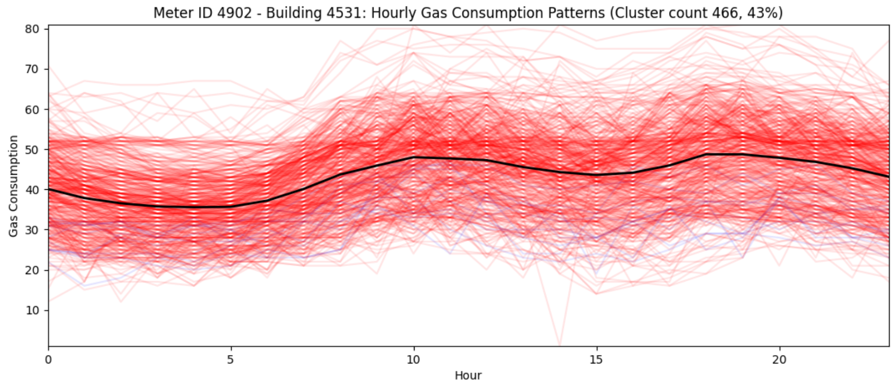
Table B.1.: Building Information

Location	Building no.	meter id	Construction year	Normative orientation [-]	EP1 [kWh/m²]	EP2 [kWh/m²]	EP2 EMG forf. [kWh/m²]	Avg. insulation floor [m² K/W]	Avg. roof insulation [m² K/W]	Avg. insulation facades excl. AOR [m² K/W]	Until. facades ex AOR [m²]	Avg. insulation windows [W/(m² K)]	Until. window [m²]
Uilenstede 002-100	4501	40773	1991	North				1.30	2.22	2.00	29.76819672	1.795936909	9.267868852
Uilenstede 102	4510	41574	1982	West	238.7209722	268.0051389	268.0051389	0.15	1.30	0.35	13.52972222	2.30	11.86986111
Uilenstede 005-455	4534	4903	1970	West	122.2210526	201.2968421	201.2968421	1.225412186	0	0.38924552	14.40741935	1.8	15.45354839
Uilenstede 108-268	4524	4905	1970	North	121.3045455	206.845	206.845	1.93	1.320560606	0.920899181	11.8060396	1.800448918	12.33574257
Uilenstede 500-502	4538	4905	2014	East	65.52405172	167.1018103	167.1018103	3.261111	4.664444	3.471111	10.398836	2.3	6.521164
Uilenstede 504-508	4537	4905	2014	South	61.443462	139.998308	139.998308	3.261111	4.664444	3.471111	12.836846	2.3	9.501077
Uilenstede 510-1	4536	4905	2012	West	59.681133	144.089933	144.089933	3.261111	4.664444	3.471111	10.873533	2.3	7.116467

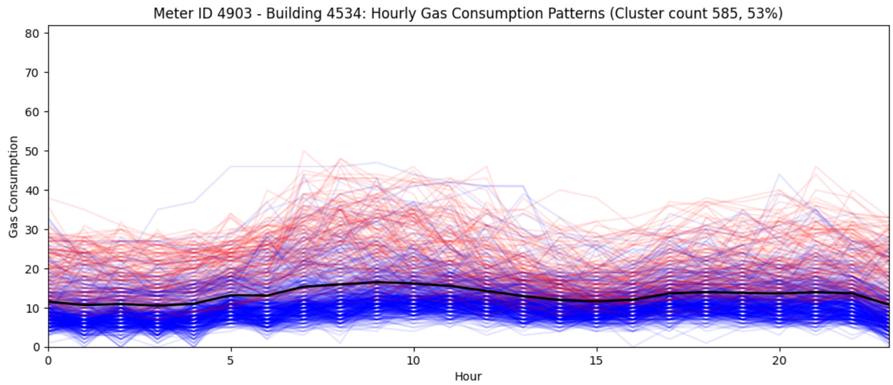
C. Single Building Cluster Result



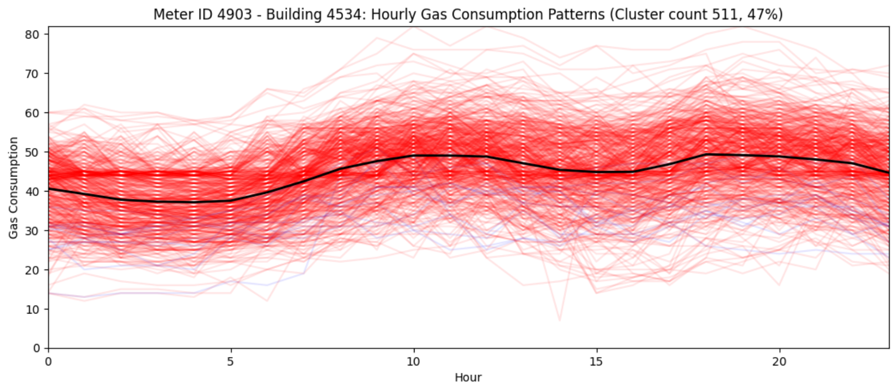
(a) Building 4531 Uilenstede 5-61 Cluster Result - Cluster 0



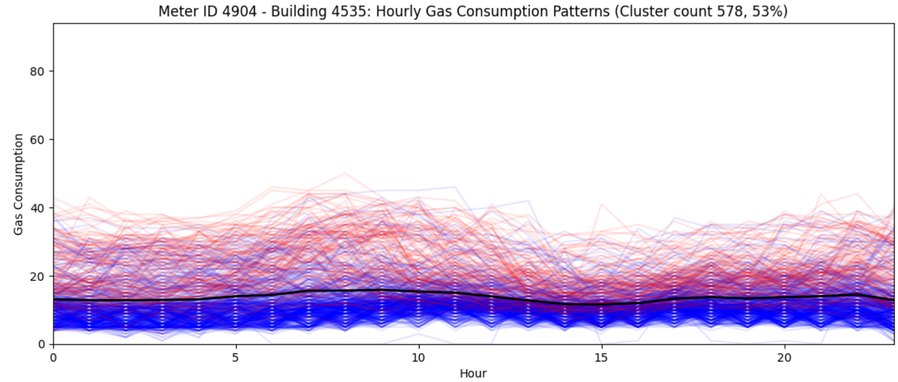
(b) Building 4531 Uilenstede 5-61 Cluster Result - Cluster 1



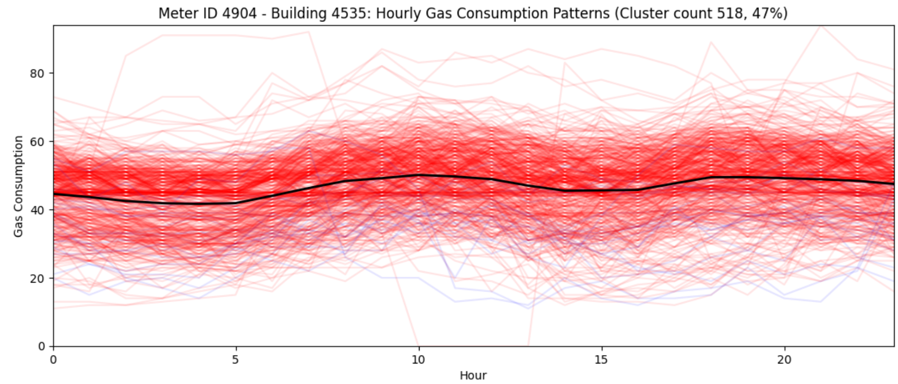
(c) Building 4534 Uilenstede 155-211 Cluster Result - Cluster 0



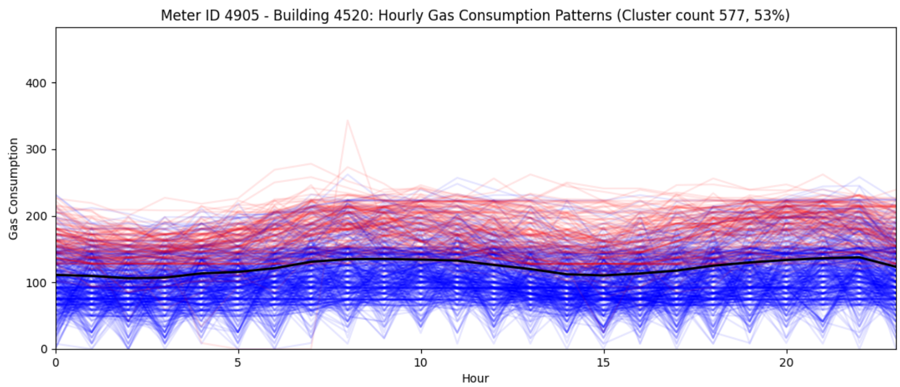
(d) Building 4534 Uilenstede 155-211 Cluster Result - Cluster 1



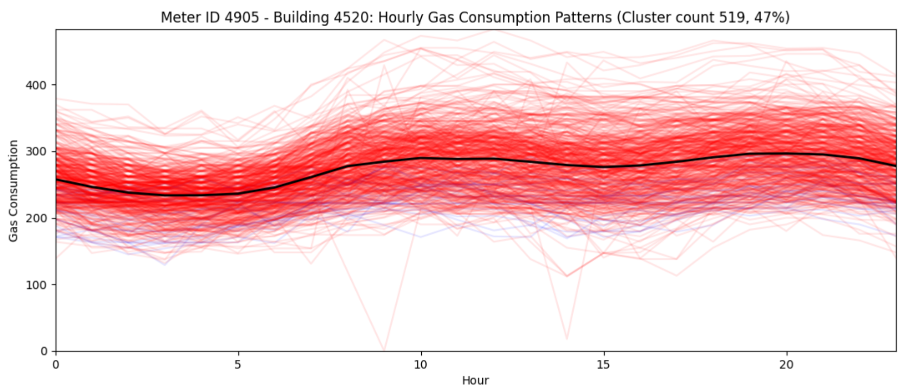
(a) Building 4535 Uilenstede 405-461 Cluster Result - Cluster 0



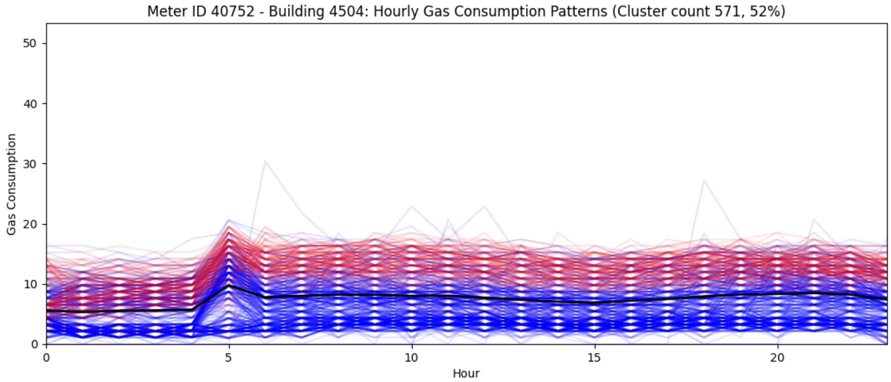
(b) Building 4535 Uilenstede 405-461 Cluster Result - Cluster 1



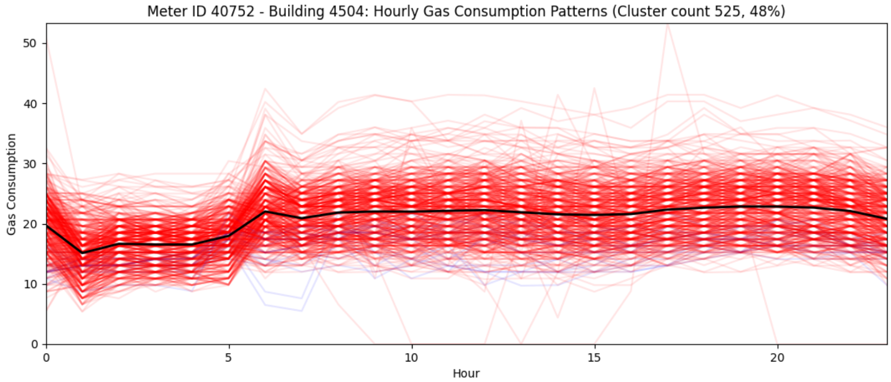
(c) Building 4520 Uilenstede 106-460 Cluster Result - Cluster 0



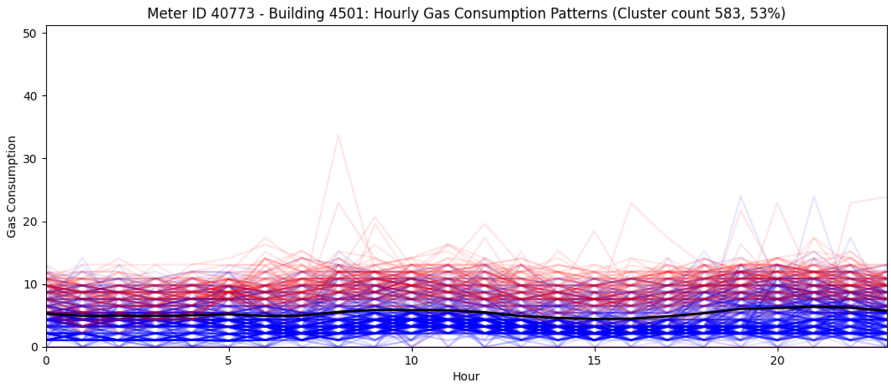
(d) Building 4520 Uilenstede 106-460 Cluster Result - Cluster 1



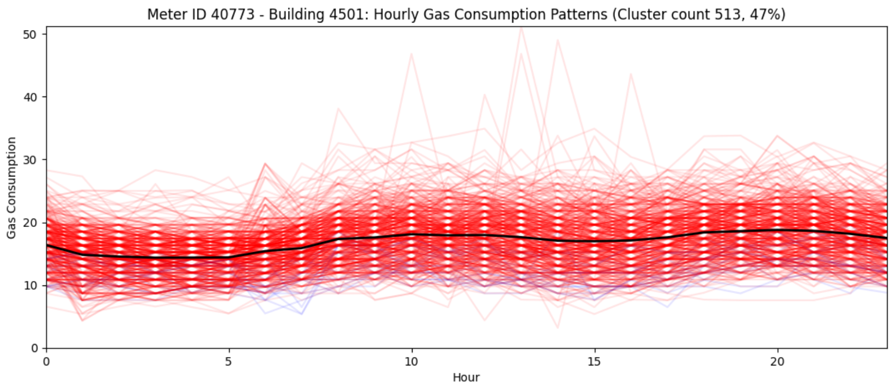
(a) Building 4504 UIlenstede 36-70 Cluster Result - Cluster 0



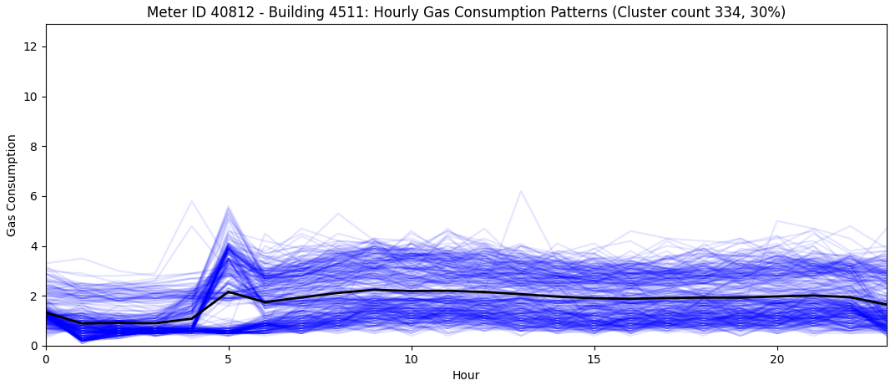
(b) Building 4504 UIlenstede 36-70 Cluster Result - Cluster 1



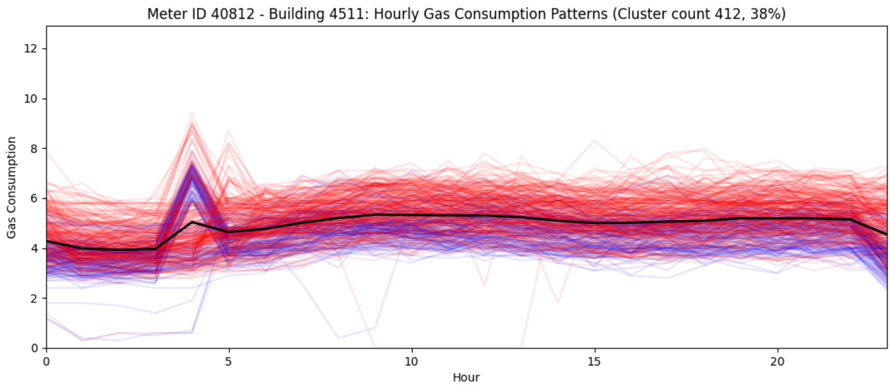
(c) Building 4501 Uilenstede 2-32 Cluster Result - Cluster 0



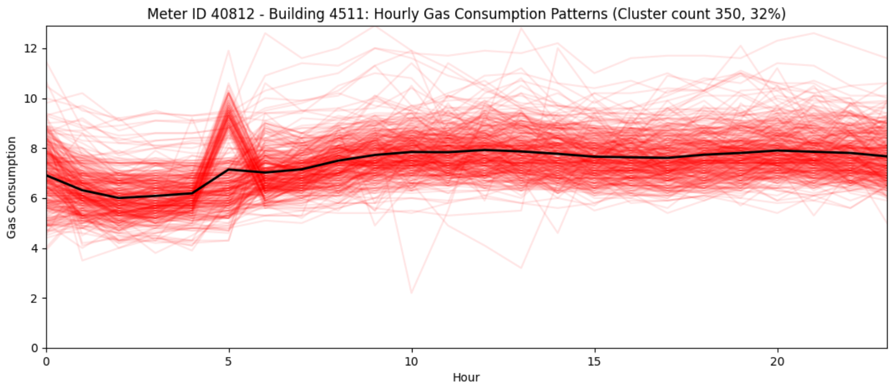
(d) Building 4501 Uilenstede 2-32 Cluster Result - Cluster 1



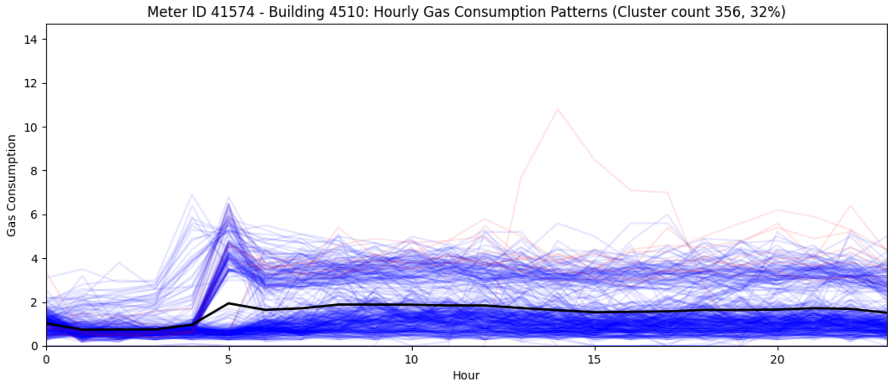
(a) Building 4511 Uilenstede 102 Cluster Result - Cluster 0



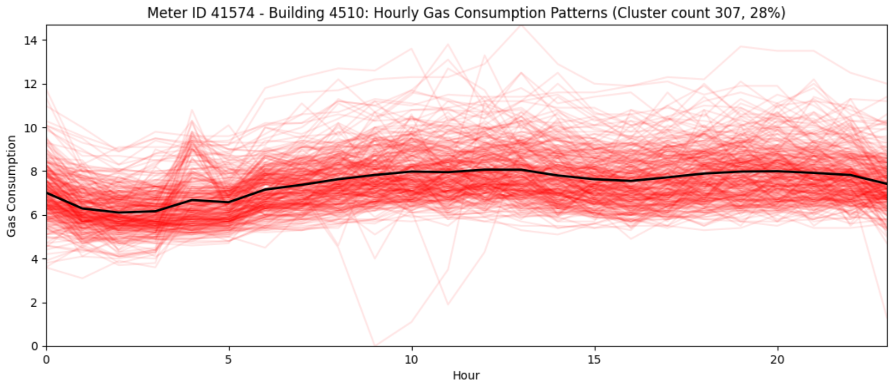
(b) Building 4511 Uilenstede 102 Cluster Result - Cluster 1



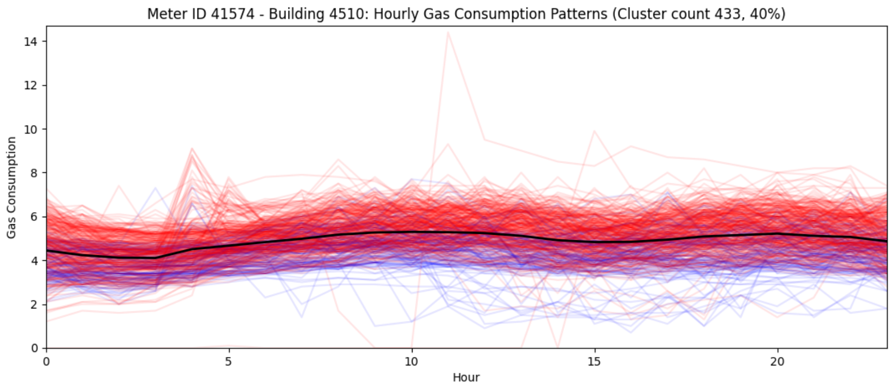
(c) Building 4511 Uilenstede 102 Cluster Result - Cluster 2



(a) Building 4510 Uilenstede 102 Cluster Result - Cluster 0

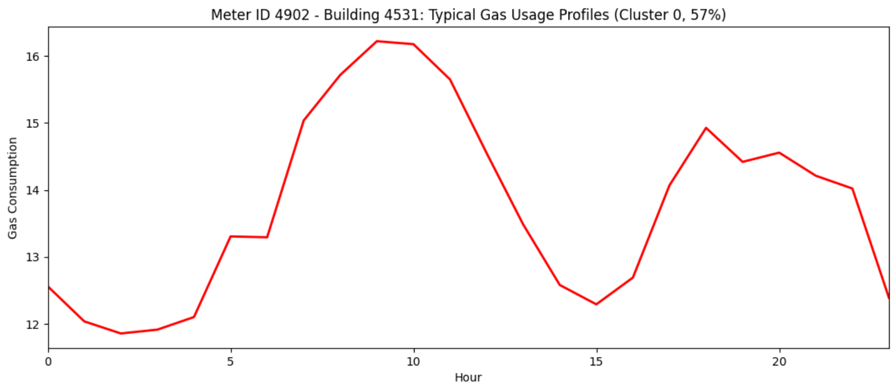


(b) Building 4510 Uilenstede 102 Cluster Result - Cluster 1

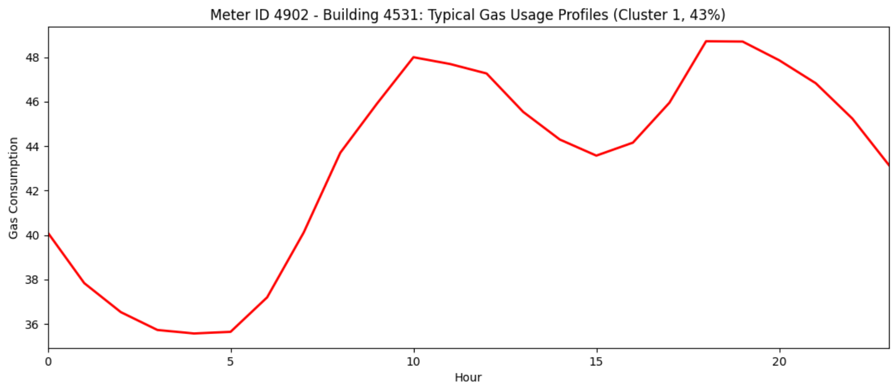


(c) Building 4510 Uilenstede 102 Cluster Result - Cluster 2

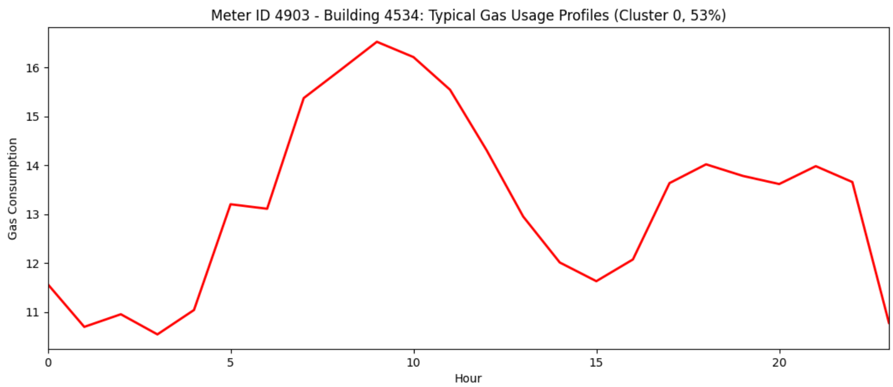
D. Typical Usage Profiles of each buildings



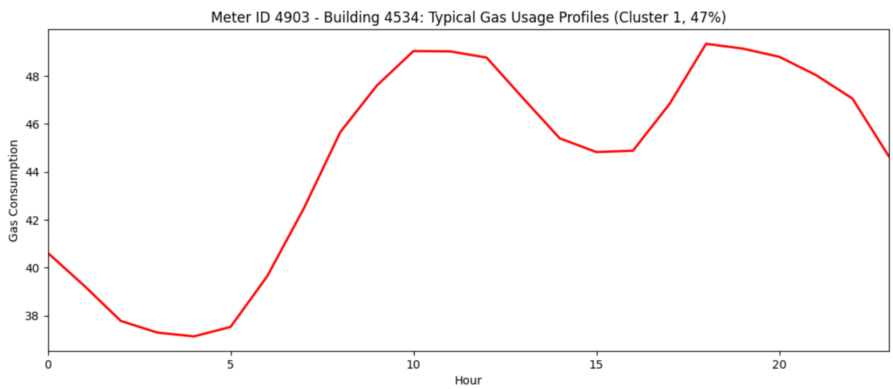
(a) Building 4531 Uilenstede 5-61 Typical Gas Usage Profiles - Cluster 0



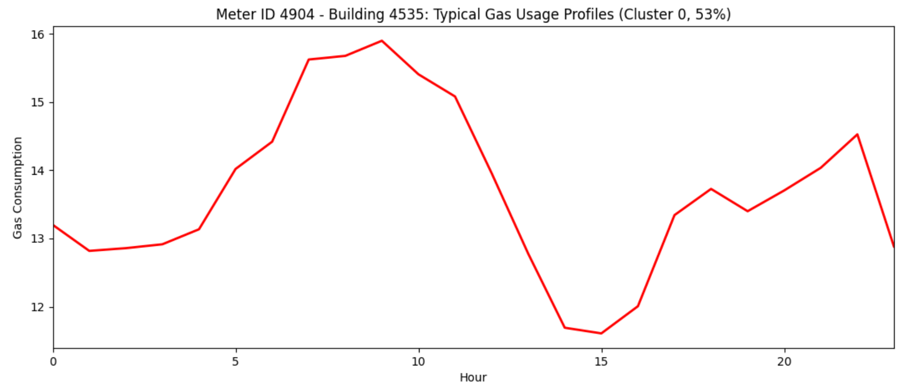
(b) Building 4531 Uilenstede 5-61 Typical Gas Usage Profiles - Cluster 1



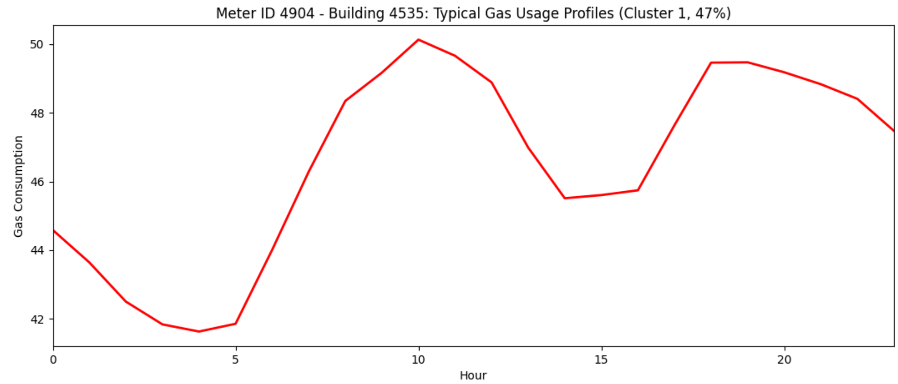
(c) Building 4534 Uilenstede 155-211 Typical Gas Usage Profiles - Cluster 0



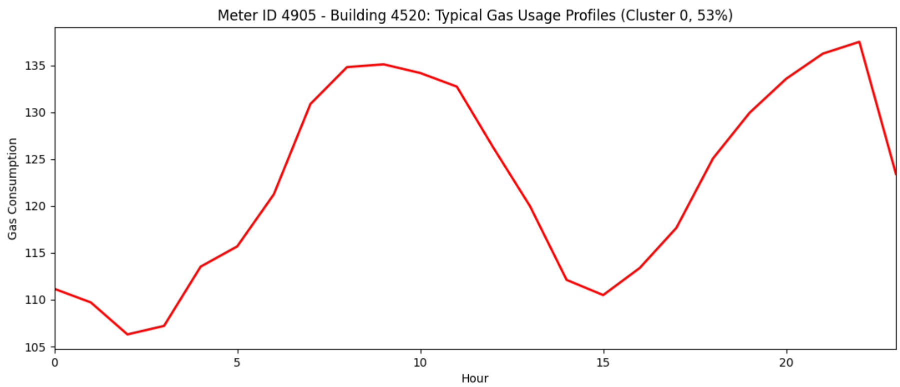
(d) Building 4534 Uilenstede 155-211 Typical Gas Usage Profiles - Cluster 1



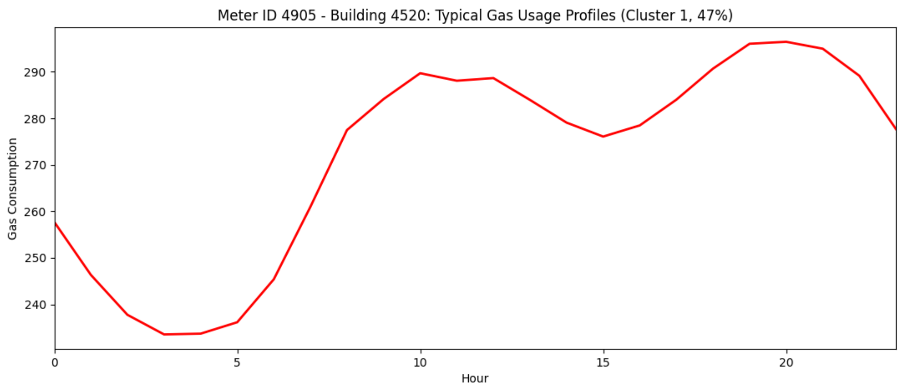
(a) Building 4535 Uilenstede 405-461 Typical Gas Usage Profiles - Cluster 0



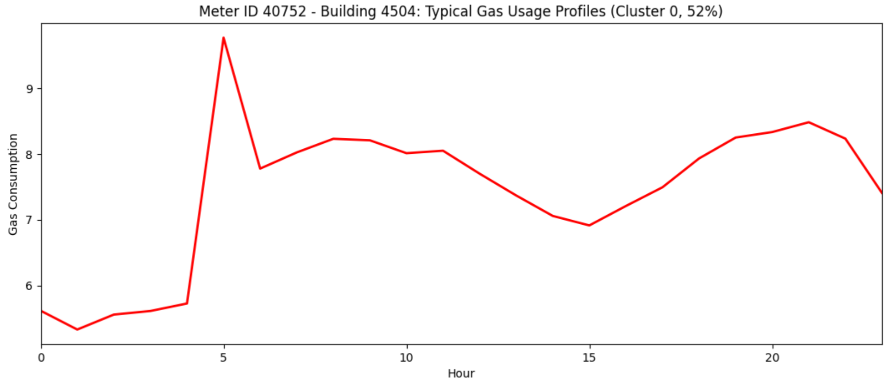
(b) Building 4535 Uilenstede 405-461 Typical Gas Usage Profiles - Cluster 1



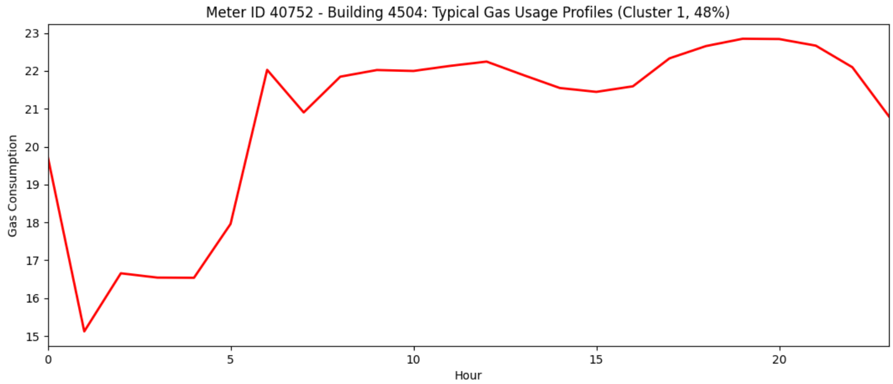
(c) Building 4520 Uilenstede 106-460 Typical Gas Usage Profiles - Cluster 0



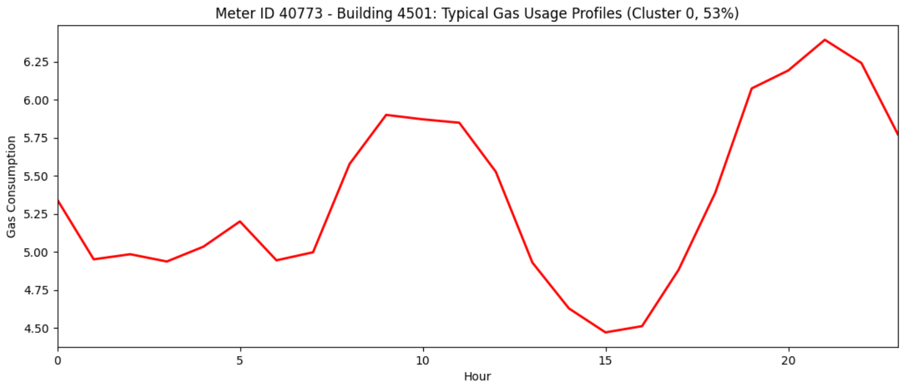
(d) Building 4520 Uilenstede 106-460 Typical Gas Usage Profiles - Cluster 1



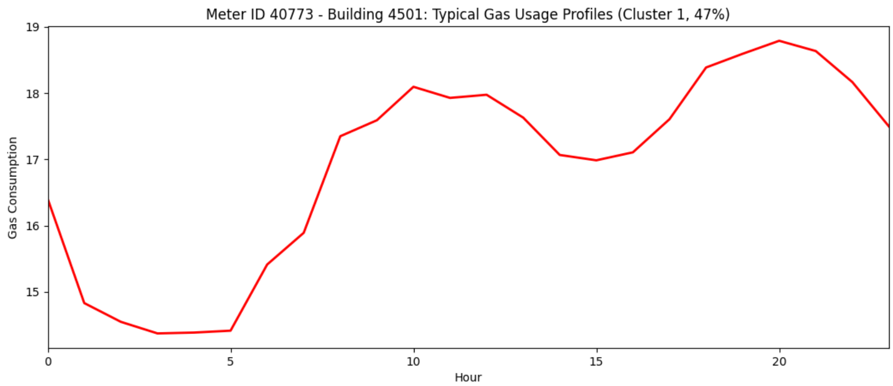
(a) Building 4504 Uilenstede 36-70 Typical Gas Usage Profiles - Cluster 0



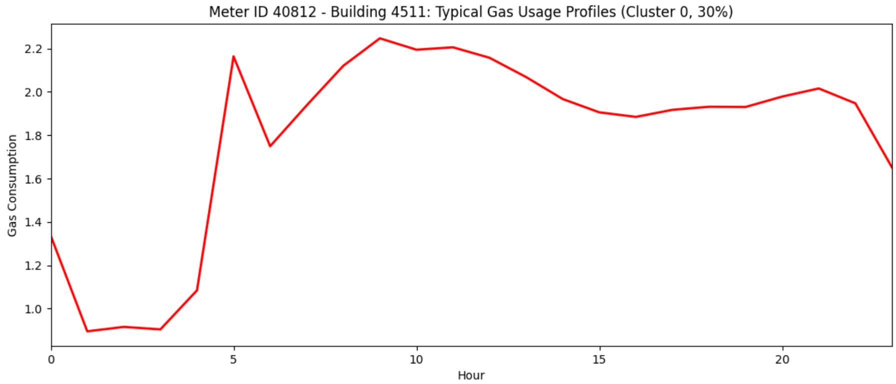
(b) Building 4504 Uilenstede 36-70 Typical Gas Usage Profiles - Cluster 1



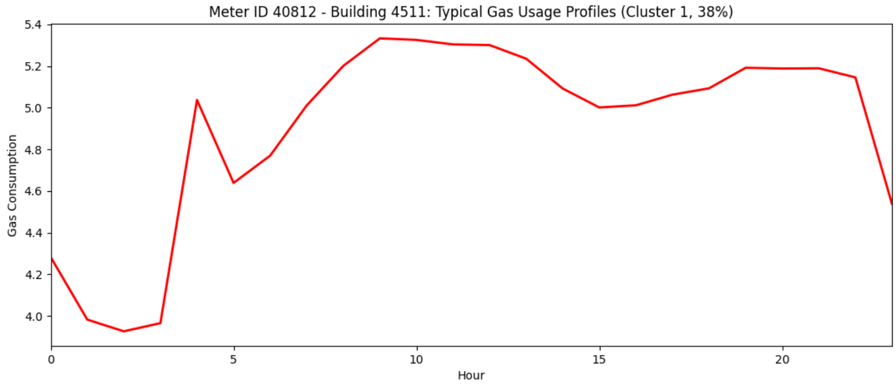
(c) Building 4501 Uilenstede 2-32 Typical Gas Usage Profiles - Cluster 0



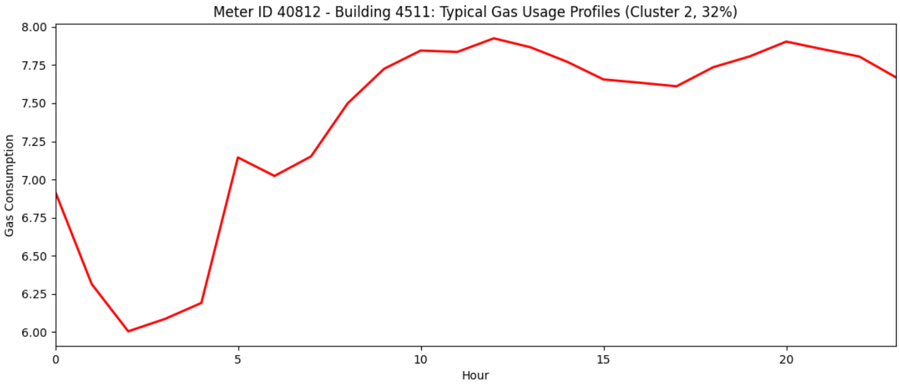
(d) Building 4501 Uilenstede 2-32 Typical Gas Usage Profiles - Cluster 1



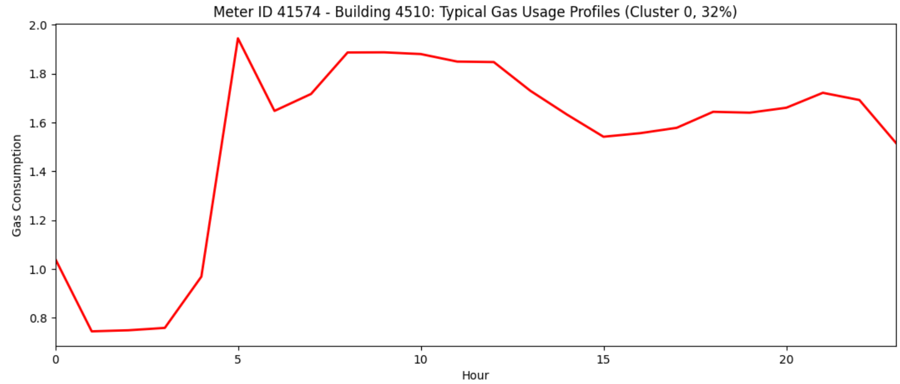
(a) Building 4511 Uilenstede 102 Typical Gas Usage Profiles - Cluster 0



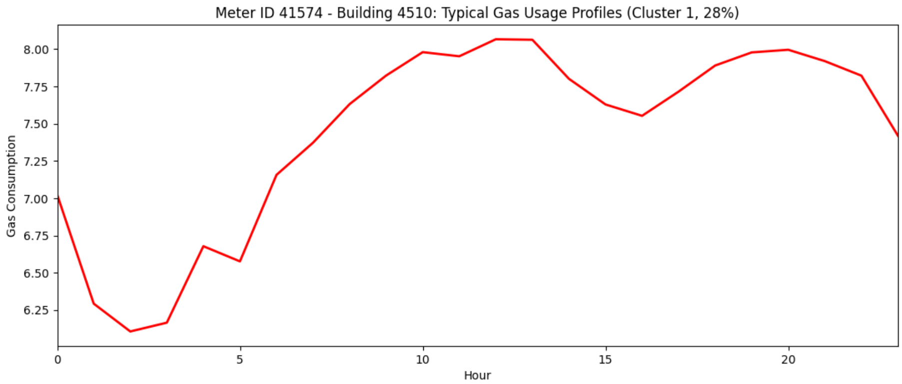
(b) Building 4511 Uilenstede 102 Typical Gas Usage Profiles - Cluster 1



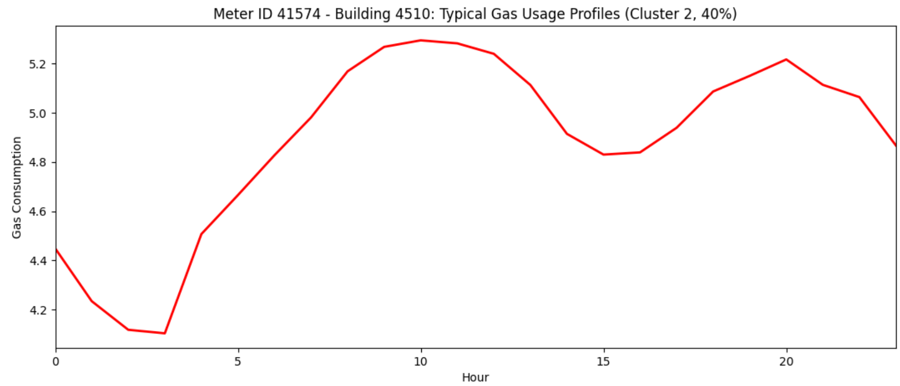
(c) Building 4511 Uilenstede 102 Typical Gas Usage Profiles - Cluster 2



(a) Building 4510 Uilenstede 102 Typical Gas Usage Profiles - Cluster 0

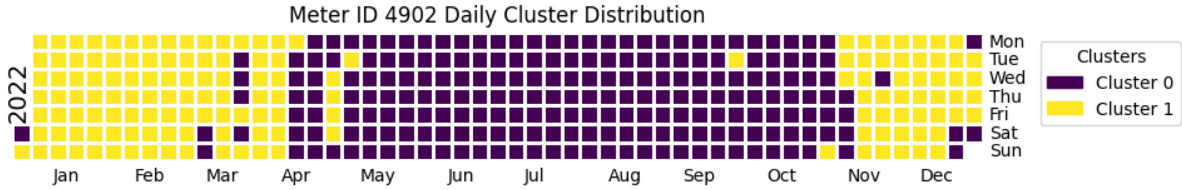


(b) Building 4510 Uilenstede 102 Typical Gas Usage Profiles - Cluster 1

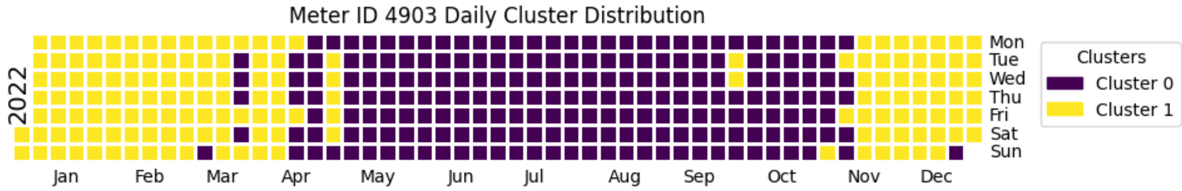


(c) Building 4510 Uilenstede 102 Typical Gas Usage Profiles - Cluster 2

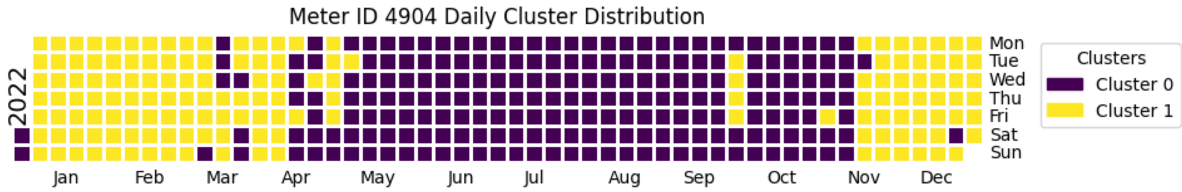
E. Daily Cluster Distribution in colander view for each cluster



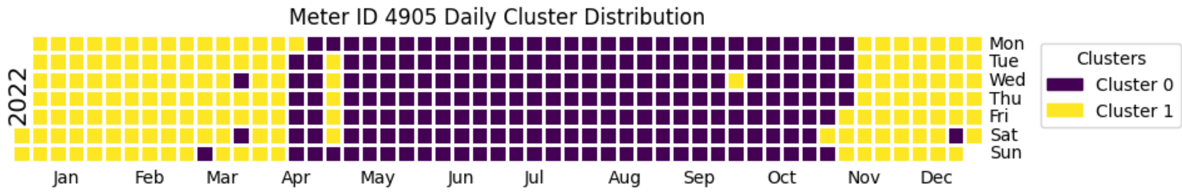
(a) Building 4531 Uilenstede 5-61 Meter ID 4902 Daily Cluster Distribution



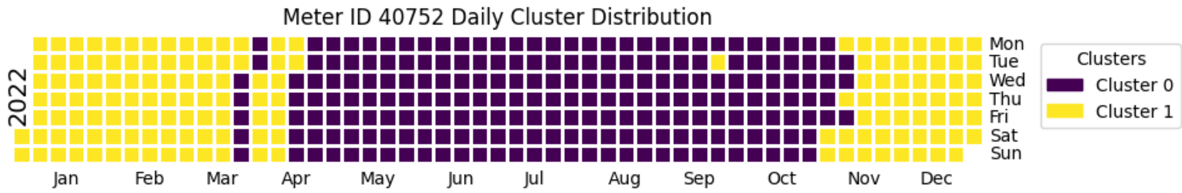
(b) Building 4534 Uilenstede 155-211 Meter ID 4903 Daily Cluster Distribution



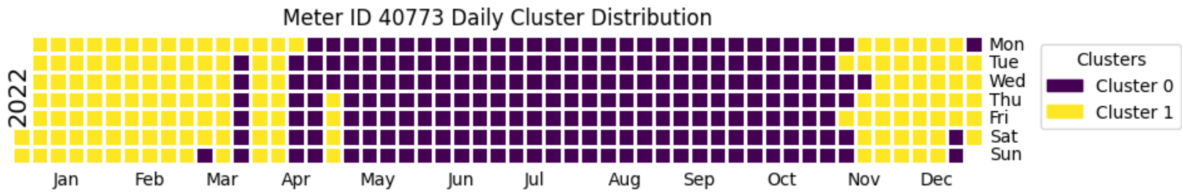
(c) Building 4535 Uilenstede 405-461 Meter ID 4904 Daily Cluster Distribution



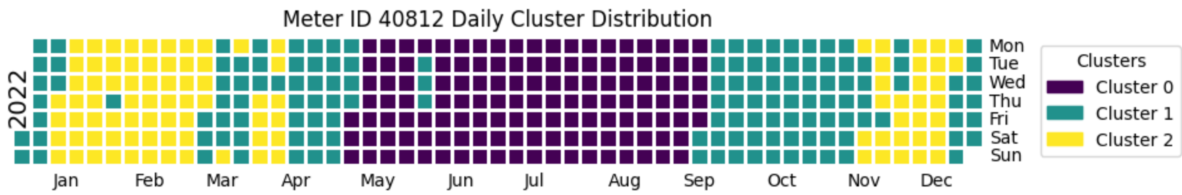
(d) Building 4520 Uilenstede 106-460 Meter ID 4905 Daily Cluster Distribution



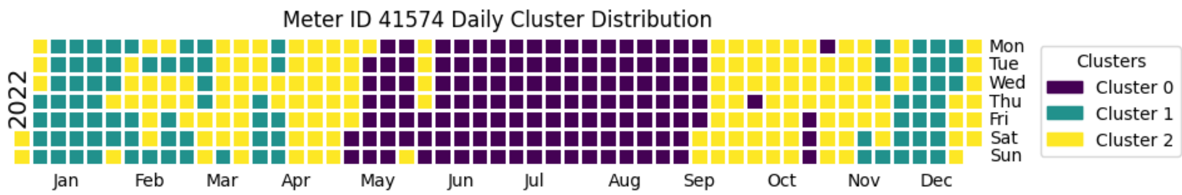
(a) Building 4504 UIlenstede 36-70 Meter ID 40752 Daily Cluster Distribution



(b) Building 4501 Uilenstede 2-32 Meter ID 40773 Daily Cluster Distribution

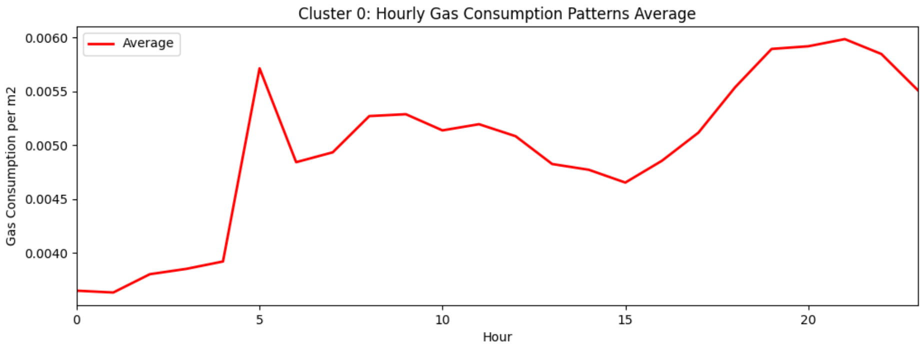


(c) Building 4511 Uilenstede 102 Meter ID 40812 Daily Cluster Distribution

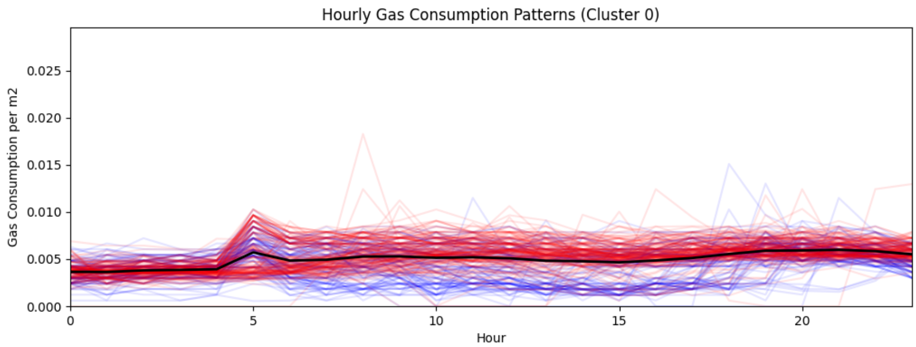


(d) Building 4510 Uilenstede 102 Meter ID 41574 Daily Cluster Distribution

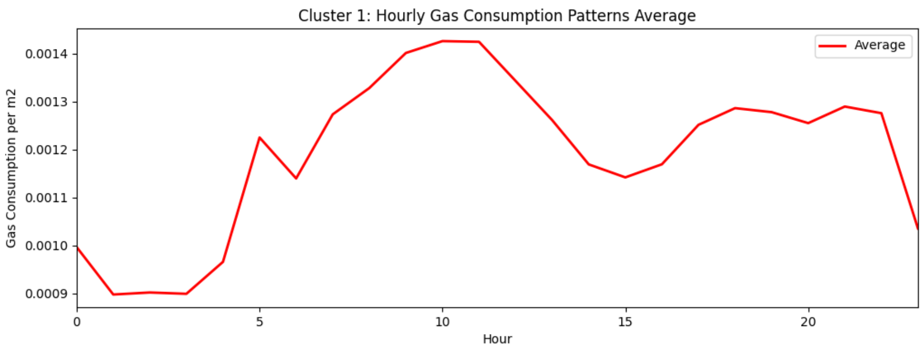
F. Inter Buildings Cluster Results



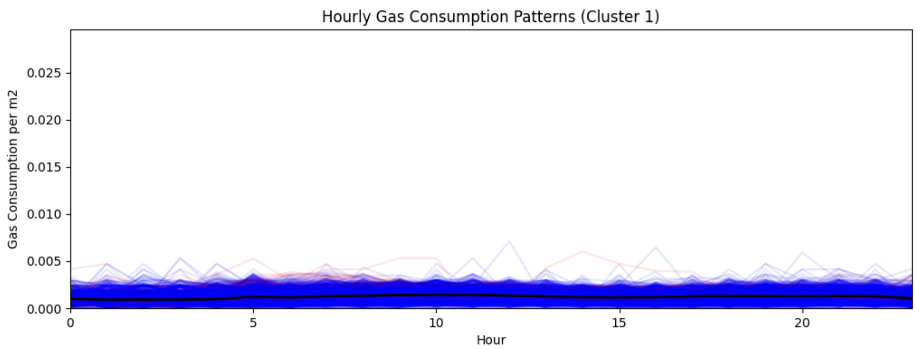
(a) Inter Building Cluster Result Cluster 0



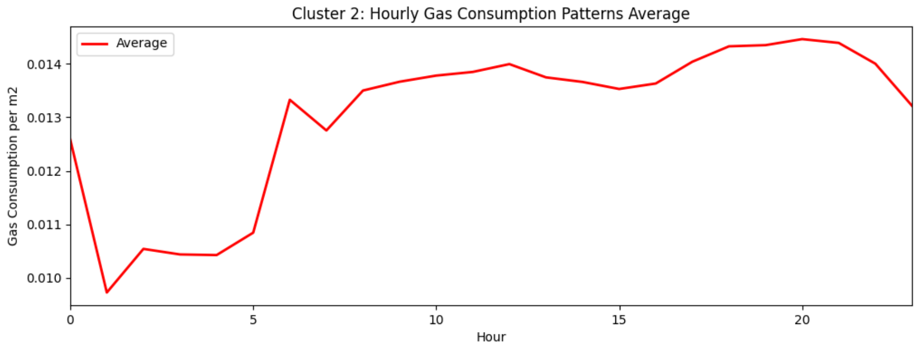
(b) Inter Building Typical Usage Profile 0



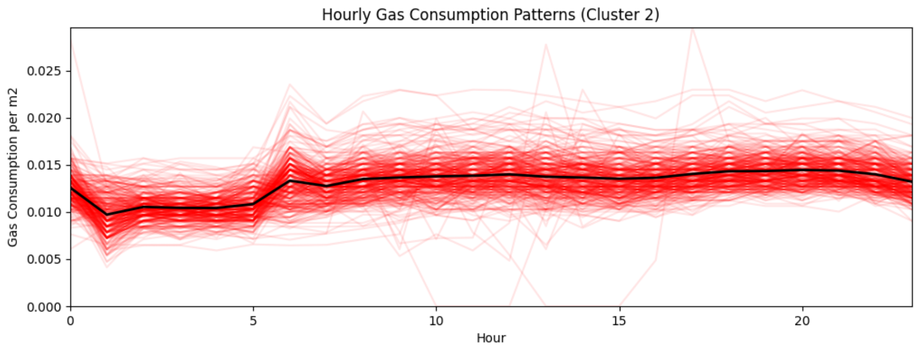
(c) Inter Building Cluster Result Cluster 1



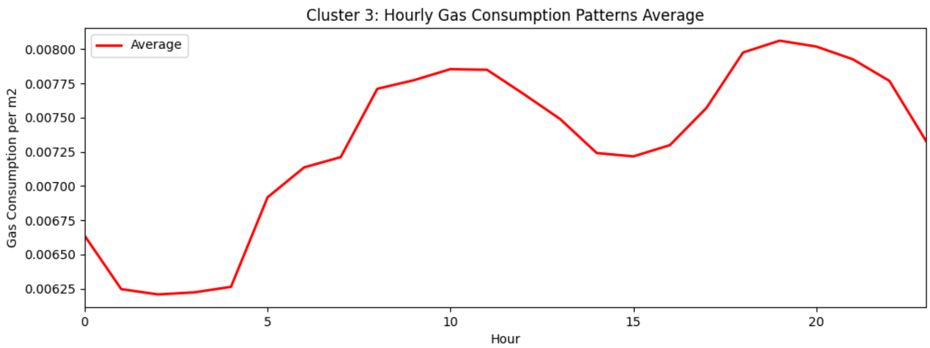
(d) Inter Building Typical Usage Profile 1



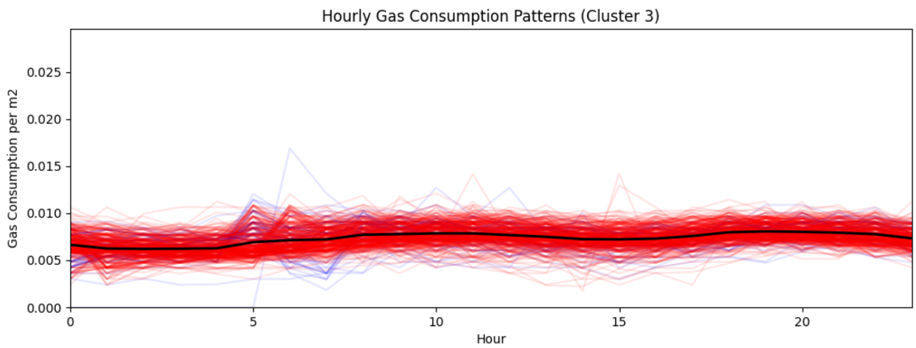
(a) Inter Building Cluster Result Cluster 2



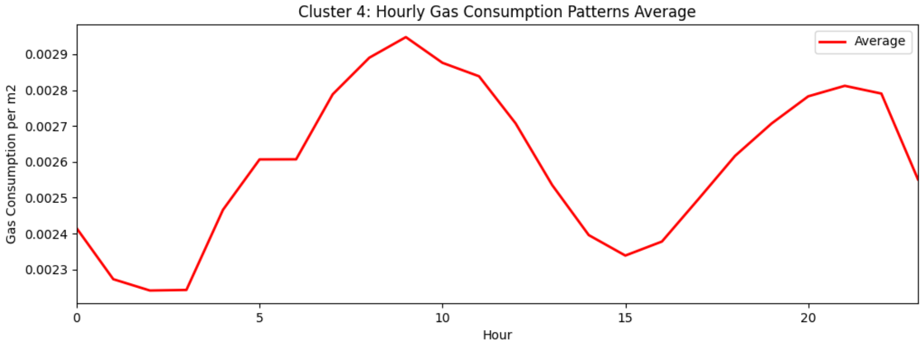
(b) Inter Building Typical Usage Profile 2



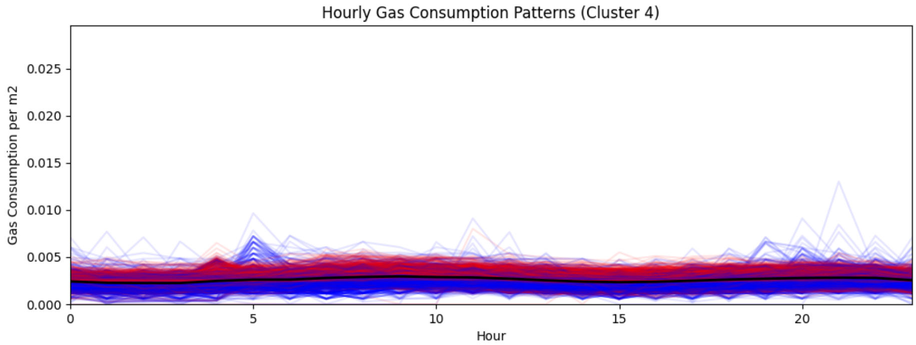
(c) Inter Building Cluster Result Cluster 3



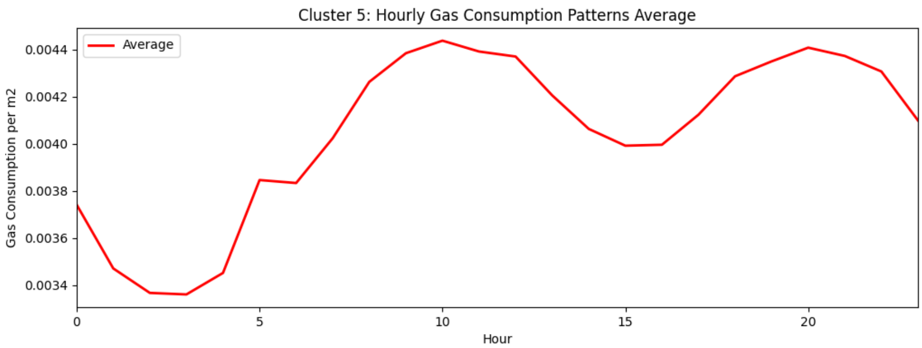
(d) Inter Building Typical Usage Profile 3



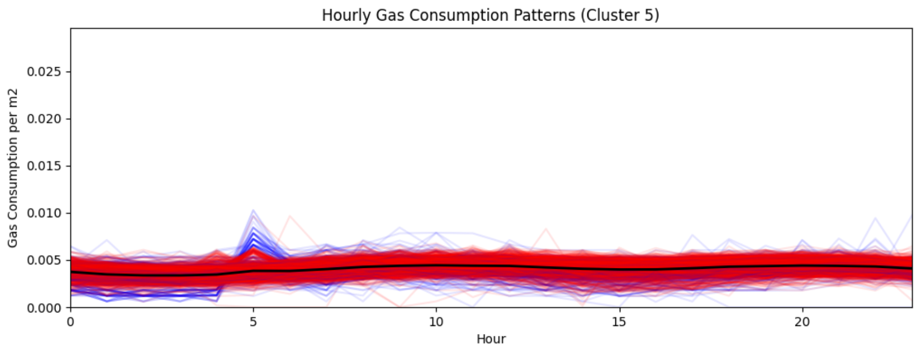
(a) Inter Building Cluster Result Cluster 4



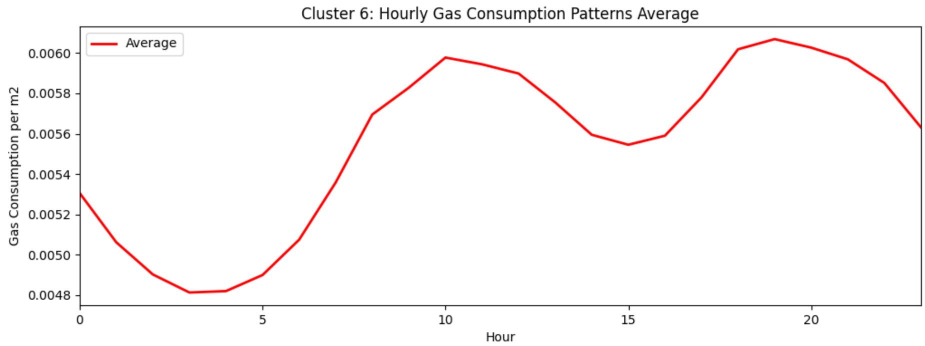
(b) Inter Building Typical Usage Profile 4



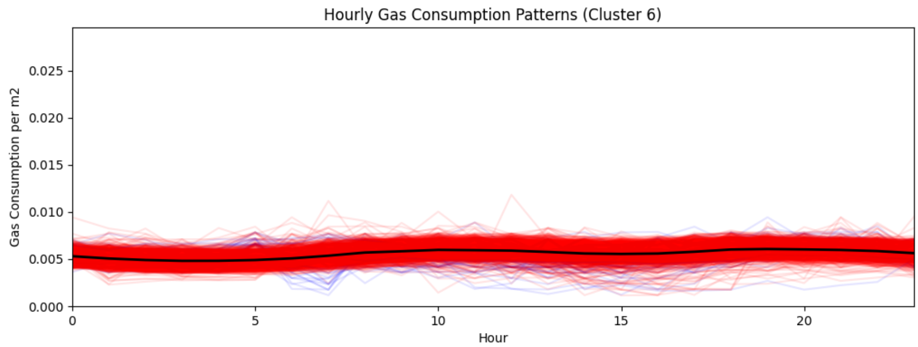
(c) Inter Building Cluster Result Cluster 5



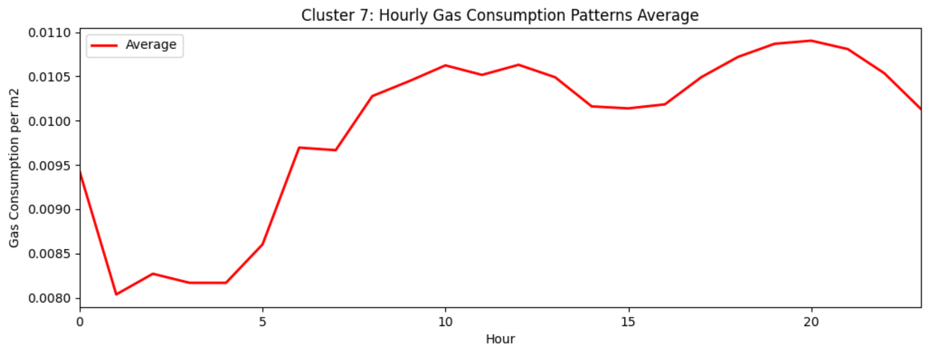
(d) Inter Building Typical Usage Profile 5



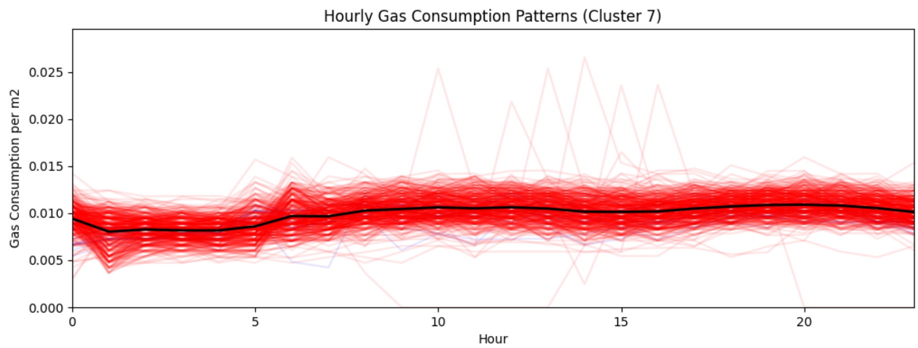
(a) Inter Building Cluster Result Cluster 6



(b) Inter Building Typical Usage Profile 6

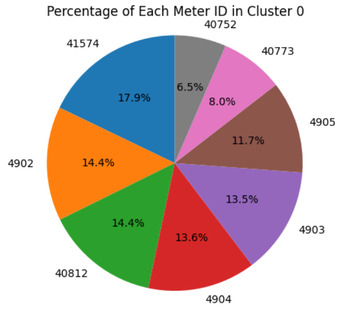


(c) Inter Building Cluster Result Cluster 7

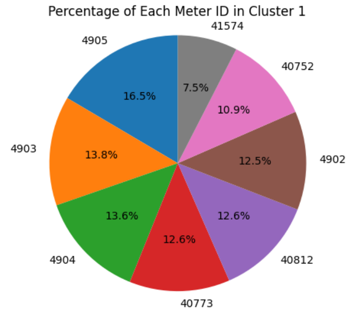


(d) Inter Building Typical Usage Profile 7

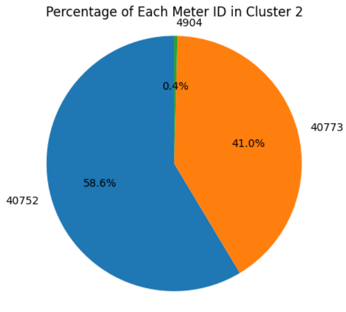
F.0.1. The percentage of meter ID in each cluster



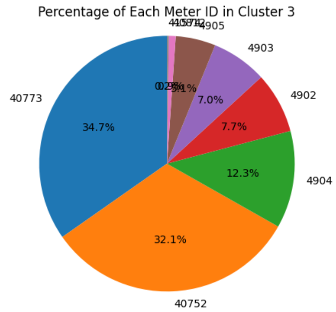
(a) Percentage of meter id in cluster 0



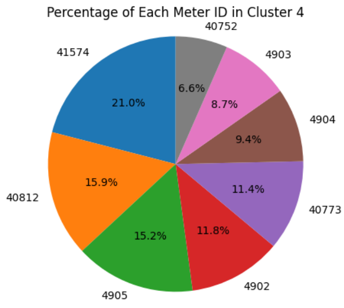
(b) Percentage of meter id in cluster 1



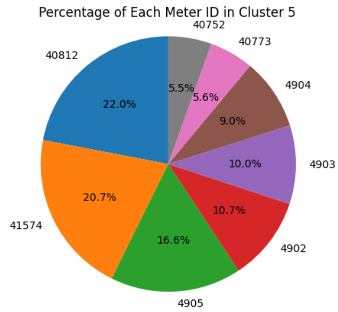
(c) Percentage of meter id in cluster 2



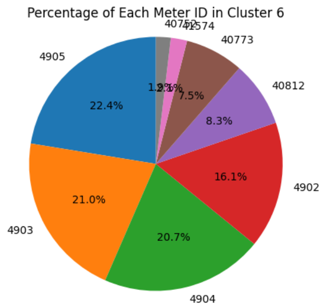
(d) Percentage of meter id in cluster 3



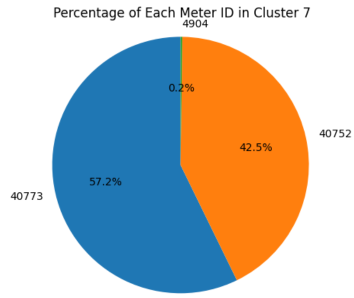
(e) Percentage of meter id in cluster 4



(f) Percentage of meter id in cluster 5



(g) Percentage of meter id in cluster 6



(h) Percentage of meter id in cluster 7

Figure F.5.: Percentage of meter id in each cluster

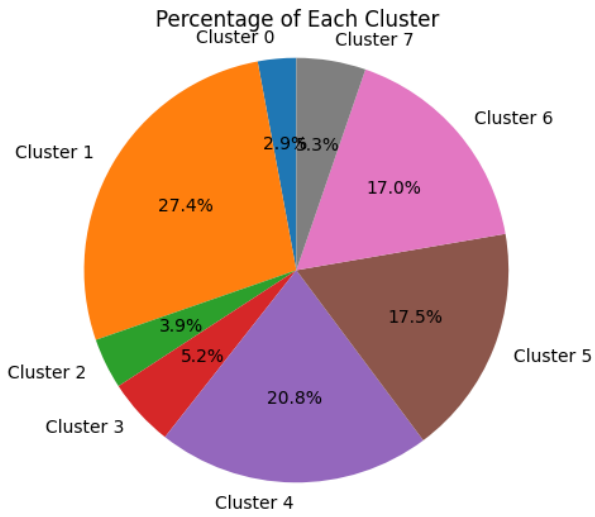


Figure F.6.: Percentage of all clusters

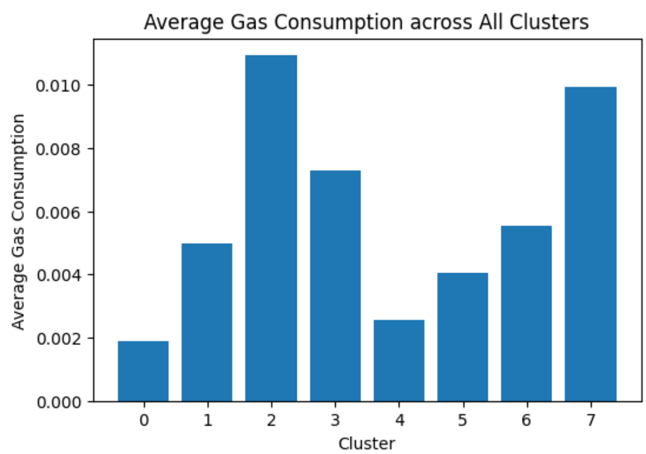
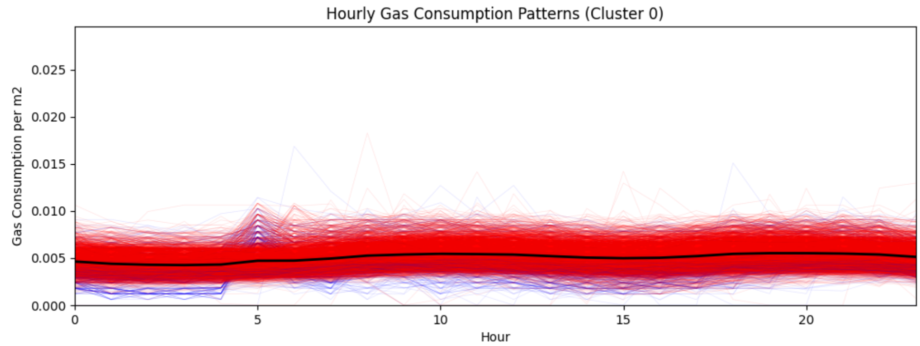
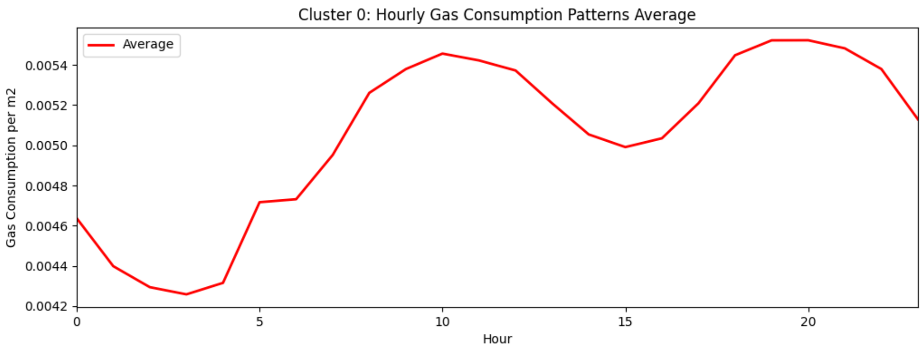


Figure F.7.: Percentage of all clusters

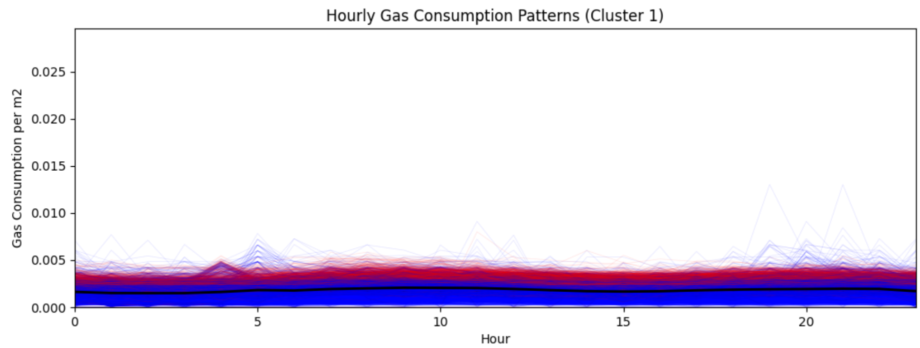
G. Inter Buildings Cluster Results using Elbow Method



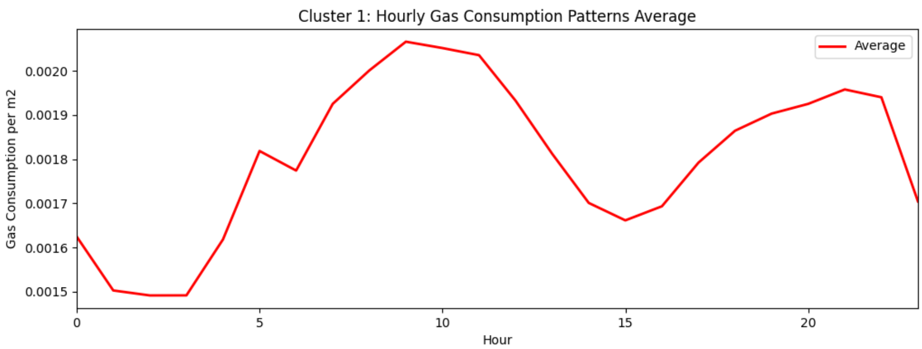
(a) Inter Building Cluster Result Cluster 0 using Elbow Method



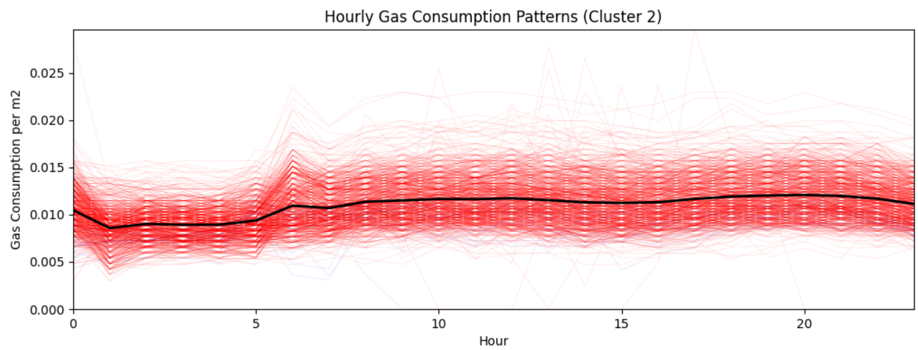
(b) Inter Building Typical Usage Profile 0 using Elbow Method



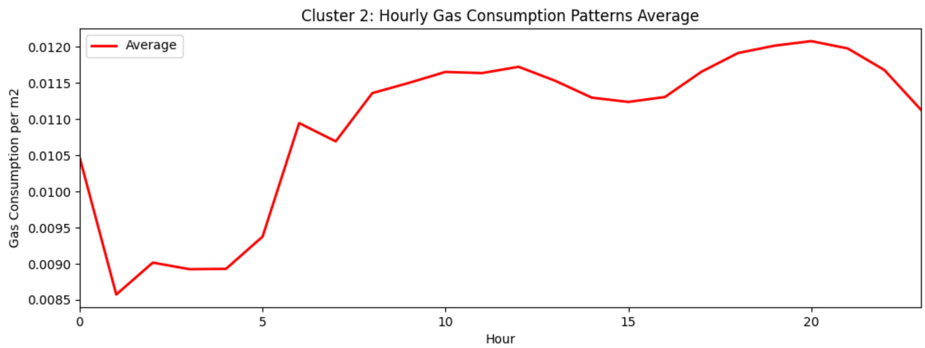
(c) Inter Building Cluster Result Cluster 1 using Elbow Method



(d) Inter Building Typical Usage Profile 1 using Elbow Method



(a) Inter Building Cluster Result Cluster 2 using Elbow Method



(b) Inter Building Typical Usage Profile 2 using Elbow Method

G.0.1. The percentage and consumption of each cluster

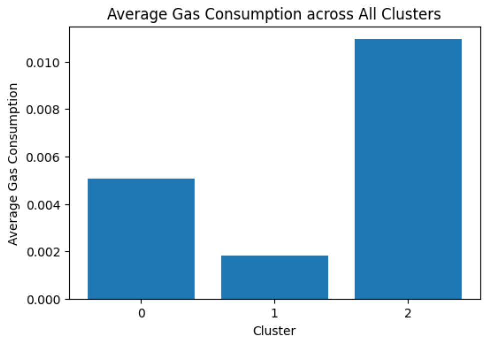


Figure G.3.: Gas consumption per floor area Inter-Building Clusters

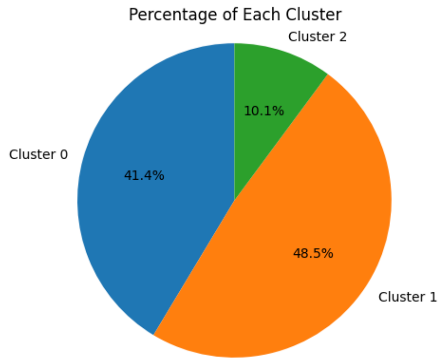
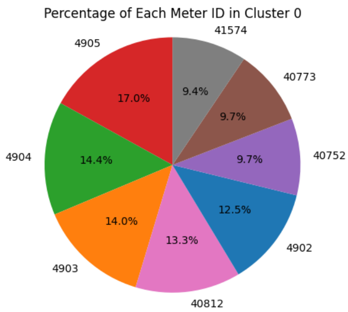
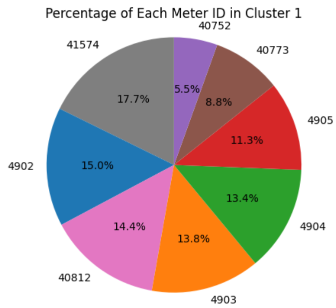


Figure G.4.: Percentage of Daily usage profiles for each Inter-Building Cluster

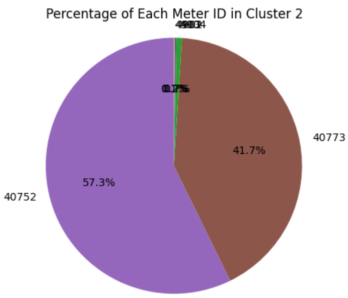
G.0.2. The percentage of meter ID in each cluster



(a) Percentage of meter id in cluster 0



(b) Percentage of meter id in cluster 1



(c) Percentage of meter id in cluster 2

Figure G.5.: Percentage of meter id in each cluster

Table G.1.: Cluster 0 Information

Feature	Statistics
Construction Year Composition	2014 - 33.6%, 1970 - 30.7%, 2012 - 16.8%, 1991 - 9.6%, 1982 - 9.3%
Average Roof Insulation [m² K/W]	Mean: 2.91, Std: 1.85
EP1 [kWh/m²]	Mean: 100.64, Std: 54.30
EP2 [kWh/m²]	Mean: 180.85, Std: 39.64
EP2 EMG forf. [kWh/m²]	Mean: 180.85, Std: 39.64
Average Insulation Floor [m² K/W]	Mean: 2.28, Std: 1.08
Average Insulation Facades Excl. AOR [m² K/W]	Mean: 2.18, Std: 1.37
Until. Facades Ex AOR [m²]	Mean: 13.83, Std: 5.36
Average Insulation Windows [W/(m² K)]	Mean: 2.10, Std: 0.25
Until. Window [m²]	Mean: 10.10, Std: 2.98
Non-heating Percentage	Mean: 0.51, Std: 0.07

Table G.2.: Cluster 1 Information

Feature	Statistics
Construction Year Composition	1970 - 29.3%, 2014 - 26.4%, 1982 - 20.7%, 2012 - 13.2%, 1991 - 10.3%
Average Roof Insulation [m² K/W]	Mean: 2.52, Std: 1.83
EP1 [kWh/m²]	Mean: 122.47, Std: 68.77
EP2 [kWh/m²]	Mean: 195.02, Std: 46.92
EP2 EMG forf. [kWh/m²]	Mean: 195.02, Std: 46.92
Average Insulation Floor [m² K/W]	Mean: 1.91, Std: 1.22
Average Insulation Facades Excl. AOR [m² K/W]	Mean: 1.84, Std: 1.40
Until. Facades Ex AOR [m²]	Mean: 14.26, Std: 5.42
Average Insulation Windows [W/(m² K)]	Mean: 2.10, Std: 0.25
Until. Window [m²]	Mean: 10.59, Std: 2.95
Non-heating Percentage	Mean: 0.50, Std: 0.07

G.0.3. Statistics features in each cluster

Table G.3.: Cluster 2 Information

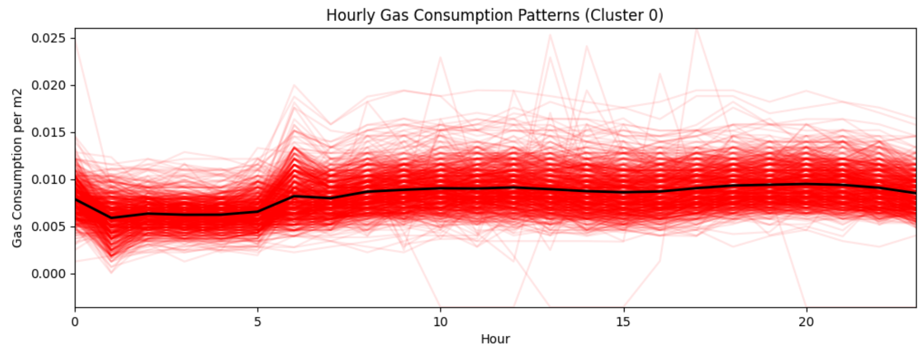
Feature	Statistics
Construction Year Composition	1991 - 99.7%, 1970 - 0.3%
Average Roof Insulation [m² K/W]	Mean: 2.21, Std: 0.12
EP1 [kWh/m²]	Mean: 122.22, Std: NaN
EP2 [kWh/m²]	Mean: 201.3, Std: NaN
EP2 EMG forf. [kWh/m²]	Mean: 201.3, Std: NaN
Average Insulation Floor [m² K/W]	Mean: 1.30, Std: 0.00
Average Insulation Facades Excl. AOR [m² K/W]	Mean: 2.00, Std: 0.08
Until. Facades Ex AOR [m²]	Mean: 29.73, Std: 0.80
Average Insulation Windows [W/(m² K)]	Mean: 1.8, Std: 0.0
Until. Window [m²]	Mean: 9.28, Std: 0.32
Non-heating Percentage	Mean: 0.45, Std: 0.02

Table G.4.: Cluster Statistics

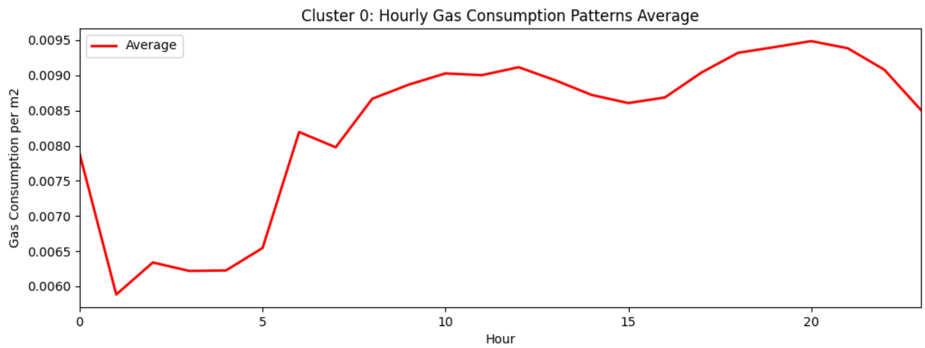
Cluster	Year Highest Proportion	Avg. Roof Insulation (m^2 K/W)	Avg. EP1 (kWh/m^2)	Avg. EP2 (kWh/m^2)	Avg. EP2 EMG forf. (kWh/m^2)	Average Insulation Floor [m^2K/W]
0	2014	2.91	100.64	180.85	180.85	2.28
1	1970	2.52	122.47	195.02	195.02	1.91
2	1991	2.21	122.22	201.30	201.30	1.30

Cluster	Avg. Insulation Facades Excl. AOR (m^2 K/W)	Avg. Facades Ex AOR (m^2)	Avg. Insulation Windows (W/m^2 K)	Avg. Until. Window(m^2)	Non-Heating Month Percentage
0	2.18	13.83	2.10	10.10	0.51
1	1.84	14.26	2.10	10.59	0.50
2	2.00	29.73	1.80	9.28	0.45

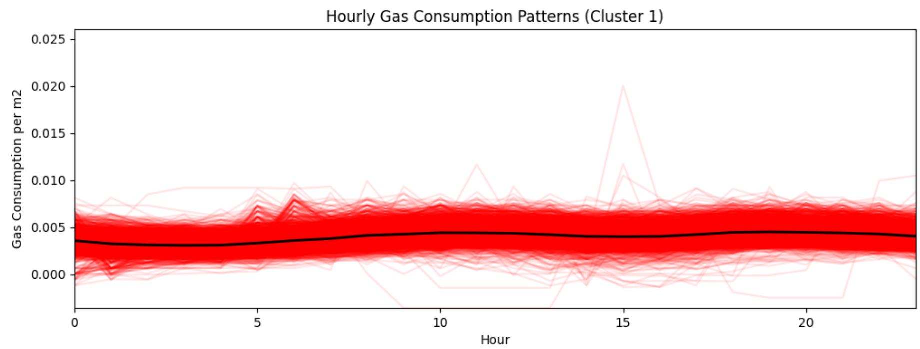
H. Inter Buildings Cluster Results for the winter months using Elbow Method



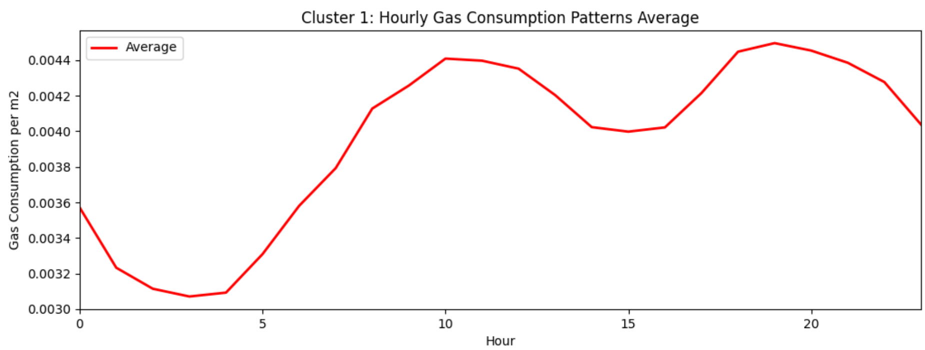
(a) Inter Building Cluster Result Cluster 0 using Elbow Method



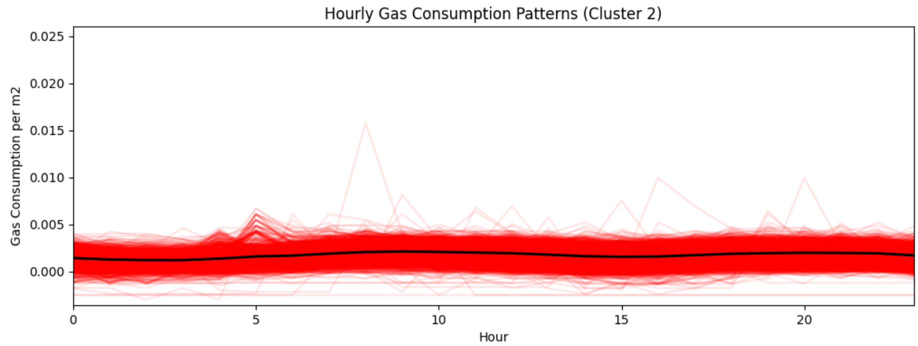
(b) Inter Building Typical Usage Profile 0 using Elbow Method



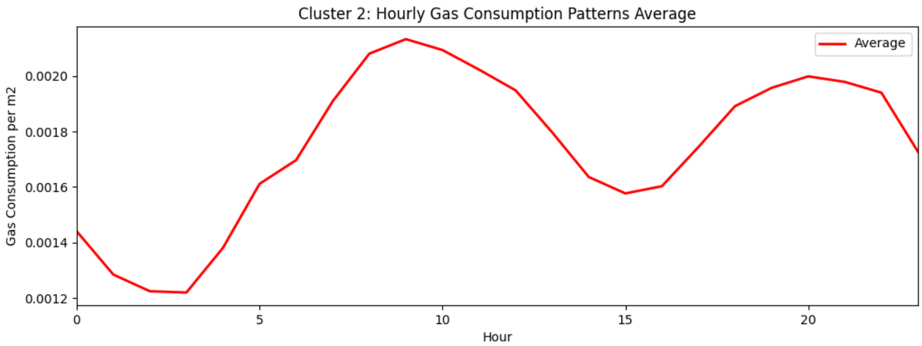
(c) Inter Building Cluster Result Cluster 1 using Elbow Method



(d) Inter Building Typical Usage Profile 1 using Elbow Method



(a) Inter Building Cluster Result Cluster 2 using Elbow Method



(b) Inter Building Typical Usage Profile 2 using Elbow Method

H.0.1. The percentage and consumption of each cluster

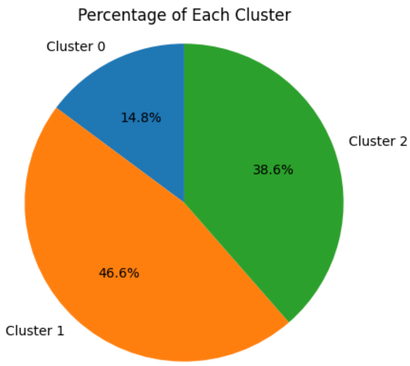


Figure H.3.: Percentage of all clusters

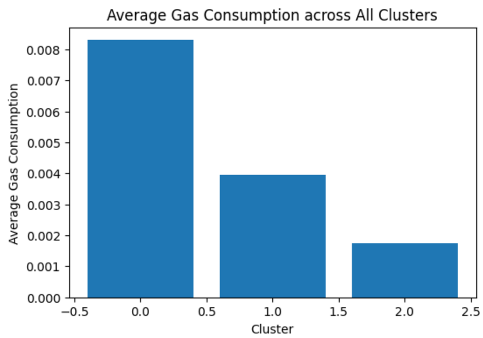
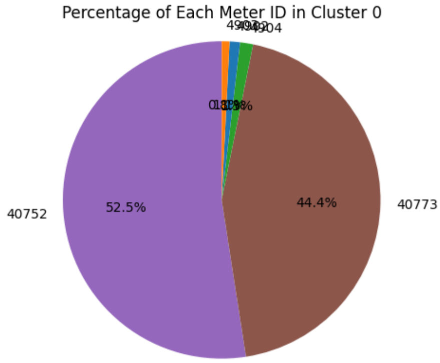
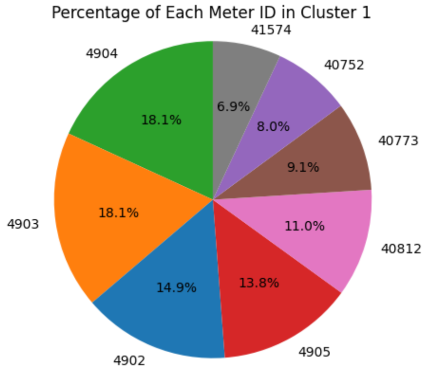


Figure H.4.: Gas consumption per floor area Inter-Building Clusters

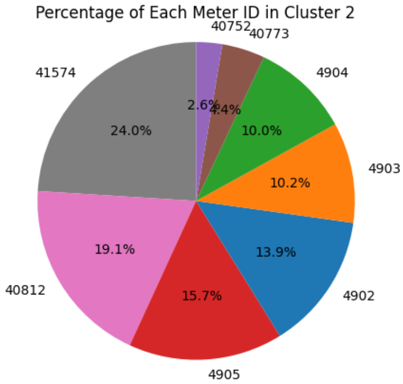
H.0.2. The percentage of meter ID in each cluster



(a) Percentage of meter id in cluster 0



(b) Percentage of meter id in cluster 1



(c) Percentage of meter id in cluster 2

Figure H.5.: Percentage of meter id in each cluster

H.0.3. Statistics features in each cluster

Table H.1.: Cluster 0 Information

Feature	Statistics
Construction Year Composition	1991 - 99.7%, 1970 - 0.3%
Average Roof Insulation [m² K/W]	Mean: 2.21, Std: 0.12
EP1 [kWh/m²]	Mean: 122.22, Std: NaN
EP2 [kWh/m²]	Mean: 201.3, Std: NaN
EP2 EMG forf. [kWh/m²]	Mean: 201.3, Std: NaN
Average Insulation Floor [m² K/W]	Mean: 1.30, Std: 0.00
Average Insulation Facades Excl. AOR [m² K/W]	Mean: 2.00, Std: 0.08
Until. Facades Ex AOR [m²]	Mean: 29.73, Std: 0.81
Average Insulation Windows [W/(m² K)]	Mean: 1.8, Std: 0.0
Until. Window [m²]	Mean: 9.28, Std: 0.32

Table H.2.: Cluster 1 Information

Feature	Statistics
Construction Year Composition	2014 - 35.7%, 1970 - 33.7%, 2012 - 17.8%, 1991 - 9.2%, 1982 - 3.5%
Average Roof Insulation [m² K/W]	Mean: 2.98, Std: 1.89
EP1 [kWh/m²]	Mean: 91.17, Std: 41.09
EP2 [kWh/m²]	Mean: 174.96, Std: 33.11
EP2 EMG forf. [kWh/m²]	Mean: 174.96, Std: 33.11
Average Insulation Floor [m² K/W]	Mean: 2.41, Std: 0.97
Average Insulation Facades Excl. AOR [m² K/W]	Mean: 2.28, Std: 1.34
Until. Facades Ex AOR [m²]	Mean: 13.71, Std: 5.30
Average Insulation Windows [W/(m² K)]	Mean: 2.08, Std: 0.25
Until. Window [m²]	Mean: 10.06, Std: 3.09

Table H.3.: Cluster 2 Information

Feature	Statistics
Construction Year Composition	1970 - 26.8%, 1982 - 35.1%, 2014 - 23.7%, 2012 - 11.9%, 1991 - 2.6%
Average Roof Insulation [m² K/W]	Mean: 2.33, Std: 1.80
EP1 [kWh/m²]	Mean: 142.19, Std: 76.23
EP2 [kWh/m²]	Mean: 207.43, Std: 50.68
EP2 EMG forf. [kWh/m²]	Mean: 207.43, Std: 50.68
Average Insulation Floor [m² K/W]	Mean: 1.66, Std: 1.32
Average Insulation Facades Excl. AOR [m² K/W]	Mean: 1.58, Std: 1.44
Until. Facades Ex AOR [m²]	Mean: 13.10, Std: 3.02
Average Insulation Windows [W/(m² K)]	Mean: 2.15, Std: 0.23
Until. Window [m²]	Mean: 10.92, Std: 2.82

Table H.4.: Cluster Statistics

Cluster	Year Highest Proportion	Avg. Roof Insulation (m^2 K/W)	Avg. EP1 (kWh/m^2)	Avg. EP2 (kWh/m^2)	Avg. EP2 EMG forf. (kWh/m^2)	Average Insulation Floor [m^2K/W]
0	1991	2.21	122.22	201.3	201.3	1.30
1	2014	2.98	91.17	174.96	174.96	2.41
2	1982	2.33	142.19	207.43	207.43	1.66

Cluster	Avg. Insulation Facades Excl. AOR (m^2 K/W)	Avg. Facades Ex AOR (m^2)	Avg. Insulation Windows (W/m^2 K)	Avg. Until. Window(m^2)
0	2.00	29.73	9.28	1.8
1	2.28	13.71	10.06	2.08
2	1.58	13.10	10.92	2.15

Bibliography

[Mat] Aldi++: Automatic and parameter-less discord and outlier detection for building energy load profiles. Alogirthem for outlier detection.

[0] Forecasting the space heating demand of dutch households.

[3] (2007). page 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg.

[4] (2014). Real-time detection of anomalous power consumption. 33:400–411.

[5] (2021). An anomaly detection framework for time series data: An interval-based approach. *Knowledge-Based Systems*, 228.

[6] (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109:756–795.

[7] Aghabozorgi, S., Seyed Shirkhorshidi, A., and Ying Wah, T. (2015). Time-series clustering - a decade review. *Information Systems*, 53:16–38.

[8] Araya, D. B., Grolinger, K., Elyamany, H. F., Capretz, M. A., and Bitsuamlak, G. (2016). Collective contextual anomaly detection framework for smart buildings. volume 2016-October, pages 511–518. Institute of Electrical and Electronics Engineers Inc.

[9] Araya, D. B., Grolinger, K., ElYamany, H. F., Capretz, M. A., and Bitsuamlak, G. (2017). An ensemble learning framework for anomaly detection in building energy consumption. *Energy and Buildings*, 144:191–206.

[10] Beretta, G. P., Iora, P., and Ghoniem, A. F. (2014). Allocating resources and products in multi-hybrid multi-cogeneration what fractions of heat and power are renewable in hybrid fossil-solar chp. *Energy (Oxford, England)*, 78:587–603.

[11] Berthold, M. R. and Höppner, F. (2016). On clustering time series using euclidean distance and pearson correlation. *arXiv [cs.LG]*.

[12] Bode, G., Schreiber, T., Baranski, M., and Müller, D. (2019). A time series clustering approach for building automation and control systems. *Applied energy*, 238:1337–1345.

[13] Boers, M. (2022). Smart home energy eciency application of machine learning on smart meter power data to improve ecient electric energy usage in households.

[14] Capozzoli, A., Piscitelli, M. S., Brandi, S., Grassi, D., and Chicco, G. (2018). Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings. *Energy*, 157:336–352.

[15] Causone, F., Carlucci, S., Ferrando, M., Marchenko, A., and Erba, S. (2019). A data-driven procedure to model occupancy and occupant-related electric load profiles in residential buildings for energy simulation. *Energy and Buildings*, 202.

[16] Chahla, C., Snoussi, H., Merghem, L., Esseghir, M., and Merghem, L. (2053). A deep learning approach for anomaly detection and prediction in power consumption data.

[17] Chen, J., Zhang, L., Li, Y., Shi, Y., Gao, X., and Hu, Y. (2022). A review of computing-based automated fault detection and diagnosis of heating, ventilation and air conditioning systems. *Renewable and Sustainable Energy Reviews*, 161.

[18] Chiosa, R., Piscitelli, M. S., Fan, C., and Capozzoli, A. (2022). Towards a self-tuned data analytics-based process for an automatic context-aware detection and diagnosis of anomalies in building energy consumption timeseries. *Energy and Buildings*, 270:112302.

[19] contributors, W. (2023). *Interquartile range*.

[20] Deb, C., Zhang, F., Yang, J., Lee, S. E., and Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924.

[21] Delaunay, B. N. (1934). Sur la sphère vide. *Izvestia Akademia Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7:793–800.

[22] Descartes, R. (1637). *Discours de la méthode*. Jan Maire, Leyde.

[23] Fan, C., Sun, Y., Zhao, Y., Song, M., and Wang, J. (2019a). Deep learning-based feature engineering methods for improved building energy prediction. *Applied energy*, 240:35–45.

[24] Fan, C., Xiao, F., Li, Z., and Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy and Buildings*, 159:296–308.

[25] Fan, C., Xiao, F., and Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied energy*, 127:1–10.

[26] Fan, C., Xiao, F., Yan, C., Liu, C., Li, Z., and Wang, J. (2019b). A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. *Applied Energy*, 235:1551–1560.

[27] Graps, A. (1995). An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2):50–61.

[28] Gunay, H. B. and Shi, Z. (2020). Cluster analysis-based anomaly detection in building automation systems. *Energy and Buildings*, 228.

[29] Góis, J. and Pereira, L. (2022). A novel methodology for identifying appliance usage patterns in buildings based on auto-correlation and probability distribution analysis. *Energy and Buildings*, 256. usage patteren reconginzation, sliding window analysis of time series,.

[30] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 28(1):100.

[31] Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F., and Amira, A. (2021). Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *Applied Energy*, 287.

[32] Hong, T., Wang, Z., Luo, X., and Zhang, W. (2020). State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212(109831):109831.

[33] Hsu, C.-Y. and Liu, W.-C. (2021). Multiple time-series convolutional neural network for fault detection and diagnosis and empirical study in semiconductor manufacturing. 32:823–836.

[34] Huang, X., Ye, Y., Xiong, L., Lau, R. Y. K., Jiang, N., and Wang, S. (2016). Time series k -means: A new k -means type smooth subspace clustering for time series data. *Information sciences*, 367–368:1–13.

[35] Kahn, M. E., Kok, N., and Quigley, J. M. (2014). Carbon emissions from the commercial building sector: The role of climate, quality, and incentives. *Journal of public economics*, 113:1–12.

[36] Kakizawa, Y., Shumway, R. H., and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441):328–340.

[37] Karatasou, S., Santamouris, M., and Geros, V. (2006). Modeling and predicting building's energy use with artificial neural networks: Methods and results. *Energy and buildings*, 38(8):949–958.

[38] Kodinariya, T. (2016). Review on determining of cluster in k-means clustering. `https://www.researchgate.net/profile/Trupti-Kodinariya/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering/links/5789fda408ae59aa667931d2/Review-on-Determining-of-Cluster-in-K-means-Clustering.pdf`. Accessed: June 16, 2023.

[39] Koniari, A. M. (2022). Development of a topology optimization algorithm for a mass-optimized cast glass component.

[40] Li, C., Ding, Z., Zhao, D., Yi, J., and Zhang, G. (2017). Building energy consumption prediction: An extreme deep learning approach. *Energies*, 10(10):1525.

[41] Li, G., Yao, Q., Fan, C., Zhou, C., Wu, G., Zhou, Z., and Fang, X. (2021). An explainable one-dimensional convolutional neural networks based fault diagnosis method for building heating, ventilation and air conditioning systems. *Building and Environment*, 203.

[42] Li, K., Ma, Z., Robinson, D., and Ma, J. (2018). Identification of typical building daily electricity usage profiles using gaussian mixture model-based clustering and hierarchical clustering. *Applied energy*, 231:331–342.

[43] Lines, J. and Bagnall, A. (2015). Time series classification with ensembles of elastic distance measures. *Data mining and knowledge discovery*, 29(3):565–592.

[44] Liu, X., Ding, Y., Tang, H., and Xiao, F. (2021). A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy and Buildings*, 231.

[45] Liu, X., Sun, H., Han, S., Han, S., Niu, S., Qin, W., Sun, P., and Song, D. (2022). A data mining research on office building energy pattern based on time-series energy consumption data. *Energy and Buildings*, 259.

[46] Luo, X. J., Oyedele, L. O., Ajayi, A. O., aOlugbenga O. Akinade, Owolabi, H. A., and Ahmed, A. (2020). Feature extraction and genetic algorithm enhanced adaptive deep neural network for energy consumption prediction in buildings. *Renewable and Sustainable Energy Reviews*, 131.

[47] Mariano-Hernández, D., Hernández-Callejo, L., Zorita-Lamadrid, A., Duque-Pérez, O., and Santos García, F. (2021). A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis. *Journal of building engineering*, 33(101692):101692.

[48] Name, A. (2007). *Information Retrieval for Music and Motion*. Springer Berlin Heidelberg.

[49] Park, J. Y., Wilson, E., Parker, A., and Nagy, Z. (2020). The good, the bad, and the ugly: Data-driven load profile discord identification in a large building portfolio. *Energy and Buildings*, 215.

[50] Park, J. Y., Yang, X., Miller, C., Arjunan, P., and Nagy, Z. (2019). Apples or oranges? identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. *Applied Energy*, 236:1280–1295.

[51] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[52] Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, 44(3):678–693.

[53] Piscitelli, M. S., Brandi, S., Capozzoli, A., and Xiao, F. (2021). A data analytics-based tool for the detection and diagnosis of anomalous daily energy patterns in buildings. *Building Simulation*, 14:131–147.

[54] Podobnik, B. and Stanley, H. E. (2008). Detrended cross-correlation analysis: a new method for analyzing two nonstationary time series. *Physical review letters*, 100(8):084102.

[55] Rajabi, A., Eskandari, M., Ghadi, M. J., Li, L., Zhang, J., and Siano, P. (2020). A comparative study of clustering techniques for electrical load pattern segmentation. *Renewable and Sustainable Energy Reviews*, 120.

[56] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

[57] Ruiz, L. G., Pegalajar, M. C., Arcucci, R., and Molina-Solana, M. (2020). A time-series clustering methodology for knowledge extraction in energy consumption data. *Expert Systems with Applications*, 160.

[58] Sadaei, H. J., de Lima e Silva, P. C., Guimarães, F. G., and Lee, M. H. (2019). Short-term load forecasting by using a combined method of convolutional neural networks and fuzzy time series. *Energy (Oxford, England)*, 175:365–377.

[59] Seyedzadeh, S., Rahimian, F. P., Glesk, I., and Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. *Visualization in engineering*, 6(1).

[60] Sun, Y., Haghighat, F., and Fung, B. C. M. (2020). A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and buildings*, 221(110022):110022.

[61] Talei, H., Benhaddou, D., Gamarra, C., Benbrahim, H., and Essaaidi, M. (2021). Smart building energy inefficiencies detection through time series analysis and unsupervised machine learning. *Energies*, 14.

[62] Tavenard, R. (2017). An introduction to dynamic time warping.

[63] Voronoi, G. M. (1908). Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik*, 134:198–287.

[64] Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological methods*, 17(2):228–243.

[65] Wan, X., Wang, W., Liu, J., and Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology*, 14(1):135.

[66] Wang, F., Xuan, Z., Zhen, Z., Li, K., Wang, T., and Shi, M. (2020). A day-ahead pv power forecasting method based on lstm-rnn model and time correlation modification under partial daily pattern prediction framework. *Energy Conversion and Management*, 212.

[67] Wang, Z. and Srinivasan, R. S. (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, 75:796–808.

[68] Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., and Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, 82:1027–1047.

[69] Xiao, C., Khayatian, F., and Dall'O', G. (2020). Unsupervised learning for feature projection: Extracting patterns from multidimensional building measurements. *Energy and Buildings*, 224.

[70] Xu, S., Qiao, X., Zhu, L., Zhang, Y., Xue, C., and Li, L. (2016). Reviews on determining the number of clusters. *Applied Mathematics Information Sciences. An International Journal*, 10(4):1493–1512.

[71] Yang, J., Ning, C., Deb, C., Zhang, F., Cheong, D., Lee, S. E., Sekhar, C., and Tham, K. W. (2017). k-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy and buildings*, 146:27–37.

[72] Yu, L., Qin, S., Zhang, M., Shen, C., Jiang, T., and Guan, X. (2021). A review of deep reinforcement learning for smart building energy management. *IEEE Internet of Things Journal*, 8(15):12046–12063.

[73] Zhan, S. and Chong, A. (2021). Building occupancy and energy consumption: Case studies across building types. *Energy and Built Environment*, 2(2):167–174.

[74] Zhan, S., Liu, Z., Chong, A., and Yan, D. (2020). Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking. *Applied Energy*, 269.

[75] Zhang, C., Cao, L., and Romagnoli, A. (2018). On the feature engineering of building energy data mining. *Sustainable cities and society*, 39:508–518.

[76] Zhang, S., Zhang, L., and Zhang, X. (2022). Clustering based on dynamic time warping to extract typical daily patterns from long-term operation data of a ground source heat pump system. *Energy (Oxford, England)*, 249(123767):123767.

[77] Zhao, Y., Li, T., Zhang, X., and Zhang, C. (2019). Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. *Renewable and Sustainable Energy Reviews*, 109:85–101. artificial neural network regression model is a continuous variablesuch as energy consumption, temperature and so onfuzzy inference system??¡br/¿fault detection: regression based, bayesian based, clustering¡br/¿Unsupervised learning-based method¡br/¿Association rule algorithm is powerful in discovering relations be-tween various measurements.¡br/¿Bayesian networks-based methods.

[78] Zhao, Y., Zhang, C., Zhang, Y., Wang, Z., and Li, J. (2020). A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. *Energy and Built Environment*, 1:149–164. an overview of data mining methods¡br/¿superviesed data in building energy system¡br/¿-regression based load prediction¡br/¿-ANN -DNN -SVR¡br/¿-fault detection¡br/¿-regression based¡br/¿-classification based¡br/¿-unsupervised¡br/¿-patterens identification¡br/¿.

Colophon

This document was typeset using LaTeX, using the KOMA-Script class scrbook. The main font is Palatino.