Smart Building Lighting Inefficiencies Detection through Time Series Analysis

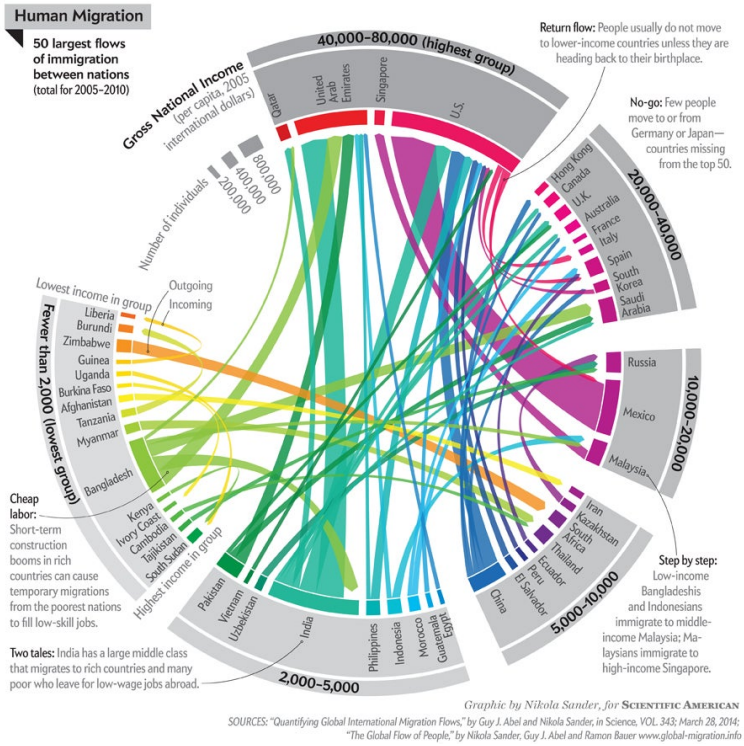
Case study: LAB42 UvA

submitted in partial fulfillment for the degree of master of science

Karim Anwar
13994565

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 15.04.2023



	UvA Supervisor
Title, Name	Dr. Hamed Seiied Alavi PhD
Affiliation	UvA Supervisor
Email	h.alavi@uva.nl

Amsterdam Data Science

The Awesome Company ™

ABSTRACT

Buildings consume a major share of global energy consumption and contribute significantly to overall carbon emissions. Furthermore, buildings offer a lot of potential for helping to reach energy efficiency goals. Smart buildings leverage advanced automation, real-time data analytics, and integrated systems to optimize energy usage and improve overall performance. As a result, energy-saving goals aimed at buildings can substantially contribute to lowering the environmental impact. Intensive time series analysis and clustering methods that organize data points into groups based on their similarity are used to find behavioral patterns in the building to improve energy efficiency. The sample data analyzed is collected from the two topmost floors from 21 different rooms facing different directions in LAB42, the specific source for each room in specific are four sensors, passive infrared (PIR), artificial light level, automated shading level, and solar irradiance amount.

KEYWORDS

keywords, belong, here, with, commas, like, this

GITHUB REPOSITORY

https://github.com/Karim-Anwar/masterProject

1 INTRODUCTION

"The current energy crisis has put the spotlight on energy demand for both governments and the public, not least of all in Europe. By dampening energy demand, energy efficiency plays an indispensable role in lowering energy bills for households and businesses and shielding consumers from volatile fuel prices. It also brings energy security benefits, especially at a time when the world is moving towards a decarbonized energy system, by reducing strains on fuel markets and the need for costly and uncertain investments in new supply."[9] Buildings play a significant role in global energy consumption and contribute extensively to carbon emissions, making them a crucial focus for energy efficiency efforts. The potential for improving energy efficiency within buildings is immense, presenting a promising pathway toward reducing their environmental impact. In this context, the emergence of smart building technologies has opened up new avenues for achieving greater energy efficiency and sustainability.

Smart buildings leverage advanced automation, real-time data analytics, and integrated systems to optimize energy usage and enhance overall performance. One particular aspect that plays a vital role in smart buildings is the effective management of lighting systems. Efficient light management not only leads to energy savings but also contributes to improved occupant comfort and productivity.

With the presence of multiple factors that can influence data collection, work in the field of pattern recognition is still in progress for improvement to detect inefficiencies in time series data. This research paper focuses on the analysis of time series data and the application of clustering methods to identify behavioral patterns related to light management within smart buildings. By exploring a rich dataset collected from LAB42, a state-of-the-art smart building, our study aims to uncover insights that drive effective light optimization strategies and contribute to sustainable building practices. With our findings, we try to answer the following question: *"How can intensive time series analysis and clustering methods be utilized to identify behavioral patterns within smart buildings and improve energy efficiency?"*

To answer that question we will break it down into the following sub-questions:

- How do the identified behavioral patterns within smart buildings contribute to improving energy efficiency, specifically in terms of lighting management?
- To what extent do the data collected from different sensors within the observed area provide insights into the interdependencies and potential causal relationships among variables?
- How do environmental factors in adjacent areas or surroundings impact the analysis of time series data collected from smart building sensors in the observed area of analysis?
- What are the potential challenges and limitations encountered during the analysis of time series data and clustering methods, and how do they impact the accuracy and reliability of the findings?

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of relevant literature, highlighting advancements in smart building technologies and the importance of light management for energy efficiency. Section 3 describes the methodology employed, including the data collection process, time series analysis techniques, and clustering methods used to identify the building patterns. Section 4 presents the results and analysis of the identified patterns related to light management. Finally, Section 5 concludes the paper, discussing the implications of our findings, potential applications, and avenues for future research.

2 RELATED WORK

Understanding how light management can impact energy efficiency is crucial, the lack of information specifically light management encourages the direction of this research. The following literature will support the decisions of how to approach this problem and illustrate why specific approaches and techniques were decided to answer the research question.

Cesar Benavente-Peces describes the energy efficiency for the next generation of smart buildings [2], his paper focuses on the key characteristics and contribution to obtain higher energy efficiencies in smart buildings, where he supports Artificial Intelligence and Data Analytics as well as Big Data for Energy Efficiency Optimization in Smart Buildings to extract patterns and improve energy savings and reduce their impact on the environment.

Zhao et al. present an argument for the relationship between building energy consumption (BEC) and the effect of reducing BEC on the economic development of China[15], where they use the Granger causality test suggesting that current BEC intensity is the Granger cause of China's economic development in specific cycles of the future.

For contextual anomaly detection, Araya et al. [1] take a look at what effect context has on identifying normal and abnormal building consumption behavior from the data.

Liu et al. show the importance of visualizing inefficiencies and why it is crucial for multiple industries[10], this paper investigates current practices used to detect and investigate anomalies in time series data in industrial contexts and identifies corresponding needs and will inspire later visualizations for our analysis.

Talei et al. describe in their paper on unsupervised clustering method using K-mean clustering to find patterns of energy inefficiency for the heating, ventilation, and air conditioning (HVAC) systems in the city of Huston as well as evaluate the amount of energy reduction that will be made using their model on an already highly efficient EMS system [12]. Do and Cetin takes an approach to energy efficiency evaluation using predictive data to estimate the electricity demand of HVAC systems while using environmental data to determine the occurrences of faults[8]. Even if this approach was successful taking into consideration the limitations of k-means clustering, Berndt et al. [3] describe Dynamic Time Warping (DTW) to evaluate distance for similarity adapted to time series data better than Euclidean distance since it does not allow temporal shifts to be taken into account.

Chandola et al. [6] use principal component analysis (PCA) which can easily identify multivariate anomalies by evaluating how far a data point deviates from the principal component and then assigning an anomaly score. PCA is another method analysis distance between time series to find similarities by applying dimensionality reduction to capture the underlying patterns and structure in the temporal domain.

Cook et al. [7] try to give the definition of what an anomaly is within the context of the Internet of Things (IoT), in a few words, it is the measurable consequences of an unexpected change in the state of a system which is outside of its local or global norm. They show as well that univariate time series, which is a sequence of observations taken by a single sensor have different handling compared to multivariate data, multivariate data is a sequence of observations taken by multiple sensors. Unsupervised Multivariate Time Series Anomaly Detection (MTSAD) is motivated by the continuous, rapid progress in the development of new machine learning techniques comparable to the conventional statistical which are deemed sub-optimal for the massive multivariate sequence datasets [5].

3 METHODOLOGY

3.1 Data description

The data collected comes from UvA's LAB42 building. What was collected was sensor data from 21 different rooms on the 5th and 6th floor of the building which represent different rooms facing different directions towards the outside of the building. The are seven sensors in particular that we use for our analysis, three of them being part of the rooms, the artificial lighting level, passive infrared, and the automatic shades level. The remaining 4 sensors are solar irradiance sensors from the North, East, South, and West side of the building. The datasets observed, in particular, have been collected since 2023-03-22 00:20:00 until today, but we cut out short the collection till 2023-05-19 12:10:00. For all datasets excluding the PIR datasets, each data point was collected in intervals of one minute during the collection time. The PIR sensors have a detection aspect to them so the data were collected at random periods of the day where there was an interaction with the sensor so it was collected to the second between two collection times of the other sensors. The data before being preprocessed included up to 84131 different timestamps, in table 1 you can see the range of the units of measurement for the different sensors.

	Artificial Light Level	Automated Shading Level	Solar Irradiance Level	PIR Indicator
Values Range	0-100%	0-100%	0-141870 Units	0 or 1

Table 1: The range of the sensor values

3.2 Data preprocessing and Exploratory Analysis

Most of the data has been collected every minute on the dot on the hour while the sensors are running, except for the PIR data. To have all our data aligned and to understand what kind of patterns we can observe, some preprocessing is required since the most precise time we have is for the PIR data which is in seconds. To bring all other datasets to be in seconds we have to resample the frequency rate, to take care of the missing values, each type of dataset had a different approach to it. For the artificial light level after resampling we backfill the data with the presumption that between two moments it is most likely that the next event to occur in the middle of it rather than it being the prior, the same assumptions can be said for the shading levels and the solar irradiance sensors but instead of doing backfill we used forward linear interpolation since compared to the light levels which were regular, both shading and solar irradiance are related and vary between two time periods either increasing or decreasing regularly between two points. Other datasets were derived from the final versions of the prior datasets, where the seconds datasets were downsampled to minutes, hours, days, and weekday versions of them to get more insight into the behaviors. After collecting all the data and handling the missing data we get for each room a dataset of 1048576 entries with 5 features: time, pir, shade, alight, and sunlight. Since the data is too granular to work with or to represent we visualize how a week of data activity looks like in Figure 1 after being resampled to be timestamped every one minute and averaging the values within those minutes.

For analyzing the data and the patterns, some descriptive analysis will be done through graph visualization. The visualization of data in the form of graphs and getting insight from them will be crucial as well for the early stages of the analysis. In Figure 2 we can observe unexpectedly, that the correlation between PIR, shade, and alight doesn't have a strong correlation between them, we shouldn't expect a perfect correlation either since the activities in the building start as early as 8 am as seen in Figure 3 and ends around 8 pm. If we compare that graph with the proportion of artificial light used as seen in Figure 4 we can see that there is light in the room there is a presence.

After we explored the proportion of artificial light usage and the activity proportion we have a look at the behavior of the usage of light during the week in Figure 5 as well as the activity behavior

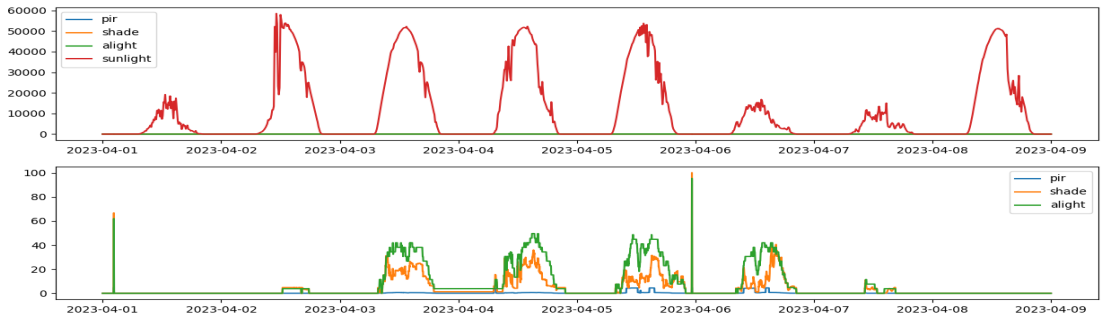


Figure 1: Representation of a week in April 2023 of the data with 1-minute intervals: (top) With all features, (bottom) Without the solar irradiance feature

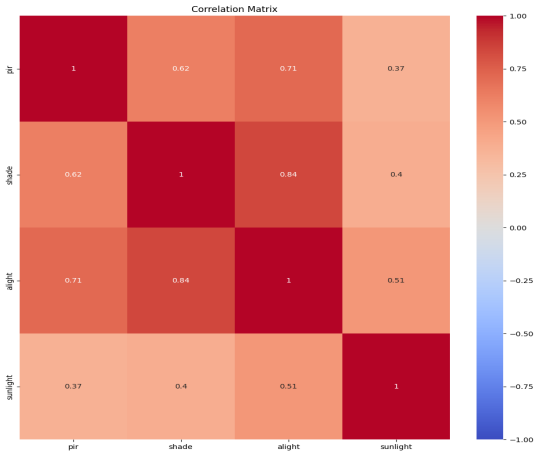


Figure 2: Correlation Matrix of the PIR, Shading, Artificial Light, and Sunlight sensors representing both of the observed floors

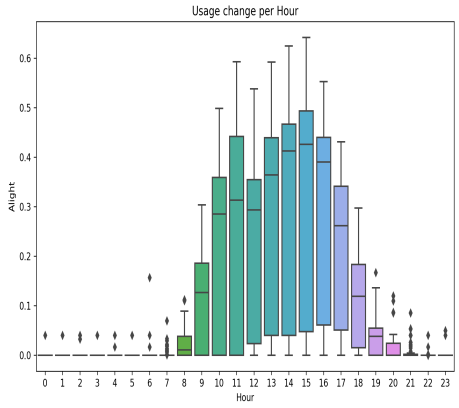


Figure 4: Light usage proportions for each hour of the day

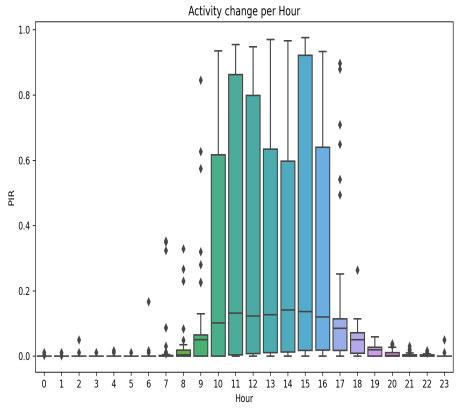


Figure 3: Activity proportions for each hour of the day recorded by the PIR sensor

in Figure 6 we can see for in the activity behavior a peak around Wednesday and a slight decrease towards Friday, and the same trend for the weekend day with little activity. For the light usage a similar behavior from Monday to Thursday and a bigger variation on Friday with the same level for maximum usage. For the weekend days, there is much less usage but Saturday has a bigger variation than Sunday. This shows us that there is a relationship between the presence of activity and light usage. We can then consider that if there is no activity there shouldn't have artificial light usage but the opposite isn't true.

If we look at the correlation matrix after downsampling our data to the hour Figure 7 and to the day of the week Figure 8 we can see the increase in the relationship between PIR (pir) and Artificial light (alight) usage. We can still in Figure 7 a good result in the relationship between the usage of Artificial Light and the amount of sunlight. Therefore, we decided to use the K-means clustering algorithm to understand with more depth the behaviors of light usage in the building.

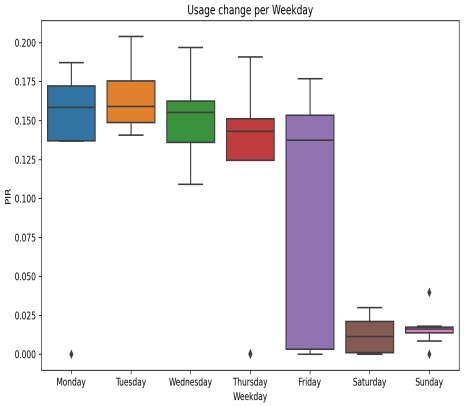


Figure 5: Artificial light usage proportion per week day

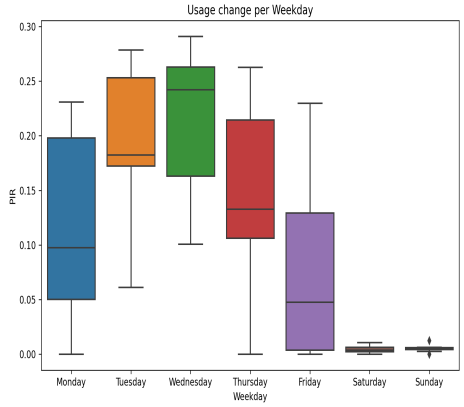


Figure 6: Activity proportion per week day

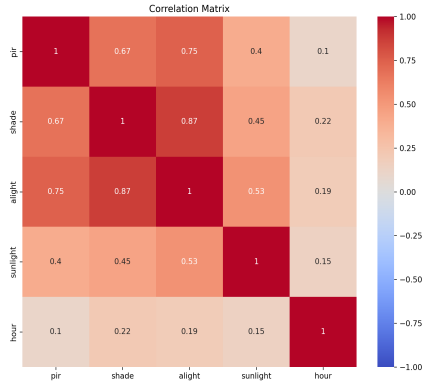


Figure 7: Correlation Matrix for the hour behavior

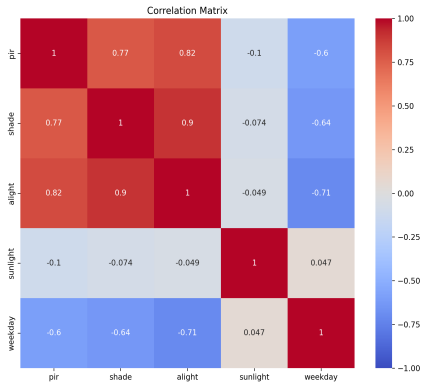


Figure 8: Correlation Matrix for the weekdays' behavior

3.3 K-means clustering

Clustering is a popular technique for grouping similar data points together based on their characteristics. When it comes to time series data, clustering algorithms can be valuable in identifying patterns and similarities among different time series. K-means is an unsupervised machine learning algorithm used to group data into K groups using the mean (average) computation, where the number of clusters has to be defined in advance. The algorithm requires one input parameter to be specified - the number of clusters (k). Given k the algorithm iterates over two phases: calculate the centroids, and assign the data point to their closest centroids until the termination condition (ex. number of iterations).

While k-means clustering is a widely used algorithm for clustering data, including time series data, there are several downsides to consider when applying it specifically to time series data:

Firstly, the performance of the algorithm can be highly sensitive to the initial choice of the cluster centroids. Different initialization may lead to different cluster assignments and potentially suboptimal results. This sensitivity is particularly pronounced in time series data, where the temporal order of the data points is crucial and may lead to locally optimal solutions that are not globally supported.

Secondly, in contrast to supervised learning, where we have the ground truth to evaluate the model's performance, clustering analysis lacks a solid evaluation metric that can be used to compare the results of various clustering algorithms. Furthermore, because K-means requires k as input and does not learn it from data, there is no correct answer in terms of the number of clusters in any problem. Domain knowledge and intuition can be useful at times, but this is not always the case.

Lastly, in addition to choosing the right number of clusters, feature scaling is an important step for machine-learning algorithms that use the distance between data.

To avoid these the first one we use the default Scikit-Learn K-means implementation, also known as k-means++. K-means++ has the characteristic of selecting cluster centroid using sampling based on an empirical probability distribution. This technique speeds

up convergence by making several trials at each sampling step and choosing the best centroid among them. Regarding the second downside, to choose the right number of cluster different metrics are used to asses the clustering results using different numbers of clusters. The ideal results occur when the inter-cluster is minimized and the intra-cluster is maximized, in other words when dissimilarity is minimized and the distance or similarity is maximized[13]. In clustering, the goal is to group similar data points together in clusters while keeping dissimilar points in separate clusters. These objectives are evaluated using metrics such as the within-cluster sum of squares (WCSS) or silhouette score[11]. The optimal number of cluster was based on the location of a bend in the generated graphs, we used four different metrics in this paper to get the best decision on which k to pick. In Table 2 the four metrics where described.

In addition to choosing the right number of clusters, feature scaling is important for k-means clustering because it helps ensure that all features contribute equally to the clustering process. K-means clustering is based on distances between data points. If the features have different scales or units, those with larger scales will dominate the distance calculations, and when features have different measurement units, scaling them to a common range eliminates the discrepancy in units and ensures comparability. Since we work with time series data to distance measurements could be used:

The Euclidean distance is commonly used as a dissimilarity measure between time series [4].

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

When working with time series data, it's important to consider other techniques such as dynamic time warping (DTW) that can account for temporal shifts and distortions in the data. Euclidean distance may not be suitable for all types of time series data, especially when they exhibit variations in time alignment.

$$\text{DTW(x, y)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - y_j)^2} \tag{2}$$

DTW requires calculating the pairwise distances between all points in the time series, resulting in a quadratic time complexity. This can make DTW computationally expensive, especially for long timeseries or large datasets. It is more efficient for datasets where the time series are not aligned and have different densities, in the preprocessing steps we have already realigned the datasets so using Euclidean Distance is still viable and uses less computation time.
In Table 3 we describe some feature scaling methods that are used to remove the volume difference in the data. We use three different methods and chose the method that led to balanced clusters.

Following Table 1 four features have been used in our analysis, with different range to their corresponding units. The effect of scaling are shown in Figure 9 over a sample date of the original data (April 4th 2023) and the impact of scaling, the latter removed the differences in volumes and revealed some initial data patterns.

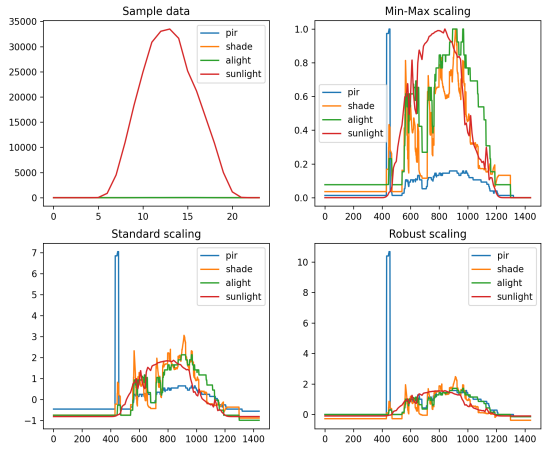


Figure 9: Data scaling effect on the data

In the next section we represent the results using the the pre-processed data and the K-means algorithm to identify possible inefficiencies of light usage in the building.

4 RESULTS

The K-means algorithm requires initialization of the k-value of clusters given in advance to directly affect the convergence result [14], using the metrics explained in Table 2 we plotted the graphs in Figure 10 to distinct that k-value. Using these metrics we look to watch where changes in the graph as the number of clusters increases, we pick the number where the first elbow formation curve occurs, it is where there is a sharp change in the rate, it's the trade-off between capturing sufficient structure in the data and avoiding overfitting.

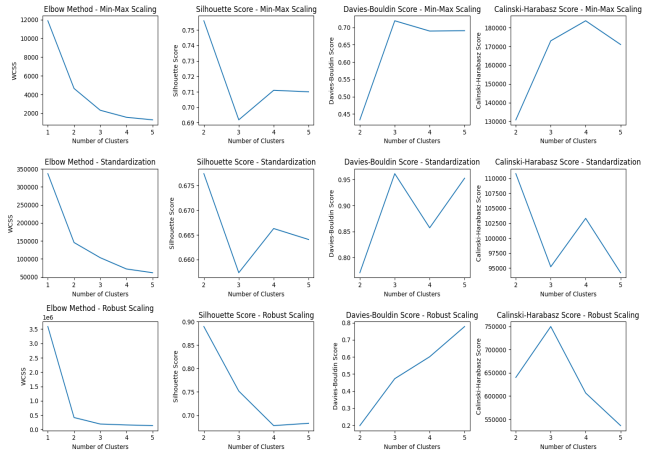


Figure 10: Metrics change for the different scaling methods on the data

Metric name	Description	Mathematical Formula	Interpretation of score
Elbow Method	It looks at the point of diminishing returns in terms of clustering improvement by analyzing the relationship between the number of clusters and the within-cluster sum of squares (WCSS)	WCSS = $\sum_{i=1}^{n} \sum_{x \in C_i} \|x - c_i\|^2$	The lower the score the better, as a higher score represents more heterogeneous clusters
Silhouette Score	The silhouette of a data instance is a measure of how closely it is matched to data within its cluster and how loosely it is matched to data of the neighboring cluster	Silhouette Score = $\frac{1}{n} \sum_{i=1}^{n} \left(\frac{b_i - a_i}{\max\{a_i, b_i\}} \right)$	A silhouette close to 1 implies the datum is in an appropriate cluster, while a silhouette close to -1 implies the datum is in the wrong cluster.
Davies-Bouldin Score	The average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances	DBI = $\frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d(c_i, c_j)} \right)$	A lower DBI score indicates better clustering, with well-separated and distinct clusters
Calinski-Harabasz Score	The score is defined as the ratio of the sum of between clusters dispersion and of inter-cluster dispersion	Calinski-Harabasz Index = $\frac{B(k)}{W(k)} \times \frac{N-k}{k-1}$	A higher score refers to better-defined clusters

Table 2: Different clustering metric techniques used in the analysis

Scaler name	Description	Mathematical Formula
Min-Max Scaler	This scaler transforms the data to a fixed range, typically between 0 and 1. It preserves the original shape and distribution of the data while maintaining the range.	$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$
Standard Scaler	This scaler transforms the data to have zero mean and unit variance. It centers the data around zero and scales it to have a standard deviation of 1.	$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$
Robust Scaler	This scaler is similar to standard scaling but uses robust statistics, such as the median and interquartile range, instead of the mean and standard deviation. It is less sensitive to outliers and extreme values, making it suitable when the data contains outliers or has a non-normal distribution.	$X_{\text{scaled}} = \frac{X - \text{median}(X)}{\text{IQR}(X)}$

Table 3: Data scaling methods used

For most metrics, we can see a sharp change around k = 3, whereas some others are around k = 2 or k = 4. Since the majority of the graphs happen to be k = 3, we pick that as we go forward. In Table 4 we have a look at the distribution of the data points over the clusters.

Scaling Method	Cluster Size		
	Cluster 0	Cluster 1	Cluster 2
Min-Max Scaler	16841	59117	8273
Standard Scaler	8255	60200	15776
Robust Scaler	11718	8255	64258

Table 4: Cluster distribution according to the different scaling methods

Looking at the cluster distributions, all three scalers have a similar distribution where around 65% of data points are collected in the same cluster. For both the min-max scaler and the standard scaler, the majority of the data were grouped in cluster 1, whereas for the robust scaler, they are found in cluster 2. To get more insight into the impact of the scaling method, we decide to go forward by exploring clustering using specific variables that may impact each other decided from both our domain knowledge and the correlation matrices observed in the previous section.

4.1 Clustering Artificial light level with PIR activity proportions

We presume that there shouldn't be artificial light turned on in the building, or in a room, without there being any presence in the theme, where it would be considered as waste or inefficiency. We can see the results of clustering in Table 5 using all three scaling methods.

Looking at the results, we can see that all scaling methods resulted in a nearly similar distribution of data points for each scaling method. Cluster 0 grouped data points with little to no activity (average of 0.02/5) and little to no artificial light usage (average of

	Min-Max Scaler			Standard Scaler			Robust Scaler		
Cluster	# of data points	activity proportion	artificial light proportion	# of data points	activity proportion	artificial light proportion	# of data points	activity proportion	artificial light proportion
0	64905	0.02	1.35	64905	0.02	1.35	64893	0.02	1.35
1	8247	4.34	39.33	8247	4.34	39.33	8247	4.34	39.33
2	11079	0.55	35.65	11079	0.55	35.65	11091	0.55	35.61

Table 5: Results of clustering PIR activity proportion and artificial light levels proportion usage

1.35%). Cluster 1 assembled the opposite where there is high activity (average of 4.34/5) and a higher artificial light usage (average of 39.33%). In Cluster 2 the activity proportion is small (average of 0.55/5) compared to the amount of light usage recorded (average of 35.65). When we look at the count of the occurrence of unique weekdays in Cluster 2 on observe its distribution in Table 6 and when we observe both Figure 6 and 5 and compare it to the distribution of Table 6, we see a similar distribution for the workdays, but in contrast, for the weekend, there is an increase of inefficiency. Using the clustering results, it was possible to identify which days of the week the building manager must observe more closely.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
# Unique Weekday	9	8	5	1	1	7	8

Table 6: Weekdays distributions collected from Cluster 2 Min-Max Scaler

In Figure 11 we show the hourly distribution and we can observe as well there where cutoffs could be used to reduce inefficiencies. We know that on weekdays the building is open from 7 am to 10 pm and on weekends from 10 am to 6 pm. A stricter cutoff of the usage in office rooms can be argued for when seeing where inefficiencies can be found outside of work hours.

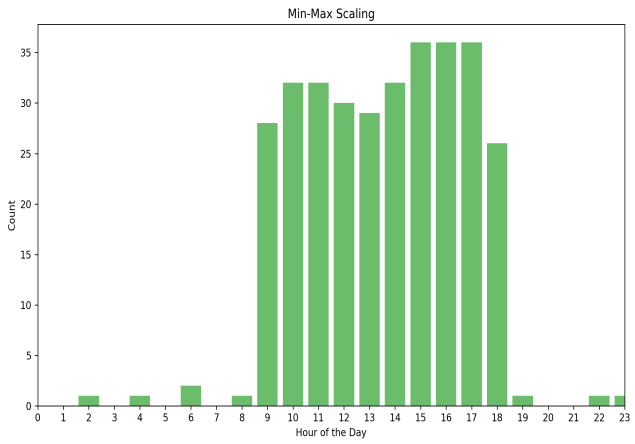


Figure 11: Distribution of the unique hour counts from the Cluster 2 produced using the Min-Max Scaler

Following steps that just focus on human activity and light usage our research can push up to 13% in energy savings concerning just lights, it is important that building managers consider these situations as a potential waste of energy.

4.2 Clustering Artificial Light, Daylight, and Shading used

There are other factors that can be considered as inefficiencies, it is important as well to not overuse artificial light and shading when there is enough sunlight to compensate, that is why we cluster these three features to understand their inefficiencies of working together. Similar to the previous section we observe the data point distribution as seen in Table 7 under the three methods of scaling. We still can see that the distribution is similar for all three methods. To decide which cluster represents the inefficiencies we have to first understand the relationship of the three features. We consider a building automated system to be efficient if there is direct daylight that hits the rooms, the minimum artificial light to be used, and where shading may vary. It is more efficient to adjust shading to reduce direct sunlight before activating the artificial lights. This assumption may not be 100% accurate since based on specific circumstances and the needs of the people using a room may vary a lot from person to person. So we observe this analysis from a statistical matter without any specifications taken into consideration.

We can start by observing that in Cluster 0, the algorithm grouped data points where low light was used (average of 1.5), with relatively low direct sunlight (average of 2706) and low shading level. This could be data points where there is no presence in the building and it is nighttime. In Cluster 1, the light proportion used is the highest (average of 39.52), sunlight at mid-range (average of 24857), and shading is the highest (average of 17.05). This could be interpreted as moments where there is a transition of phases between strong daylight to nightfall. And lastly Cluster 2, with low artificial light usage (average of 6.73), strong daylight (average of 41720), and low shading (average of 3.2). We will observe Cluster 2 again to understand if more interpretation can be done to understand the requirements of light when the sunlight is at its highest. We can start by describing when in the week is the data points distributed from Table 8. We observe that the distribution is almost uniform.

We will have to have a look then at the hourly distribution of the data points to get a better understanding. According to Figure 12 it is mainly focused on the active hours of the building. This can give insight to the building manager that tuning the sensors that represent these features might be needed to improve the building activities.

4.3 Clustering all features

Finally, we have to look at the clustering results when considering all the features of the dataset (PIR, shade, alight, and sunlight). The results are represented in Table 9. We look to identify the inefficiencies of light management from theses features. We look at to find were in the data there is high artificial light usage, with low activity in the building, with enough sunlight, and high shading. Looking at theses conditions should be what the building managers

Cluster	Min-Max Scaler				Standard Scaler				Robust Scaler			
	# of data points	artificial light proportion	sunlight level	shading level	# of data points	artificial light proportion	sunlight level	shading level	# of data points	artificial light proportion	sunlight level	shading level
0	58829	1.50	2706	1.11	58631	1.45	2712	1.02	58132	1.32	2706	0.88
1	16707	39.52	24857	17.05	16949	39.25	24694	17.18	17508	38.14	24922	17.38
2	8695	6.73	41720	3.20	8651	6.58	41580	3.10	8591	3.80	39997	2.59

Table 7: Results of clustering the three features

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
# Unique Weekday	7	7	6	7	7	7	7

Table 8: Weekdays distributions collected from Cluster 2 for the three features

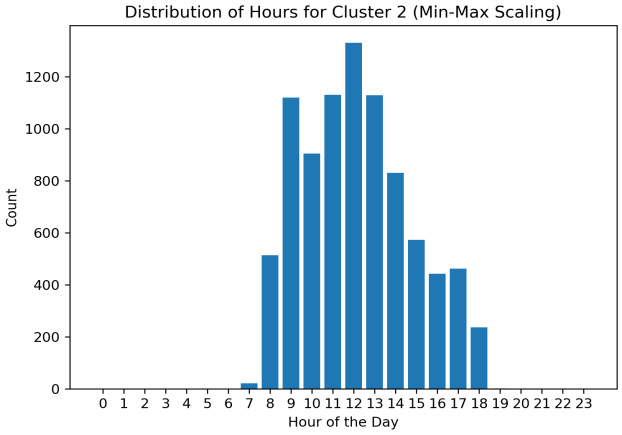


Figure 12: Distribution of the unique hour counts from Cluster 2

to change the setting of these features to a more energy-efficient condition.

All three scaling method produced a good clustering results. For every scaling method used, the k-means algorithm resulted in a cluster with high activity, relative to the high artificial light usage, when there is relative amount of sunlight and a good amount of shading used. It produced additionally, a second cluster, with low activity proportional to the artificial light used, and with low amount of natural light, and required shading. It produced as well a thirds cluster with still relatively low activity proportion level, but with high artificial light usage, and a high level on natural light with a strong usage of the shades. Using the information from that last cluster we can calculate the amount of inefficiency produced by the building. We measure that from the collected sample data an up to 20% energy inefficiency in regards of light sensors and their usage to manage energy efficiency in the building. Our analysis shows that an extra savings from depending on which condition the energy manager applies from the results collected in this paper can save from 13-20% on energy costs yearly if followed.

5 DISCUSSION

This paper had for objective to respond the the following research question "How can intensive time series analysis and clustering methods be utilized to identify behavioral patterns within smart buildings and improve energy efficiency?", in this section we discuss we will discuss a our findings in relation to it.

5.1 Contribution on Identifying Behavioral Patterns in Smart Buildings to Improving Energy Efficiency

In this study, intensive time series analysis and clustering methods were utilized to identify behavioral patterns of a smart building and how they contributed to improve energy efficiency, particularly in terms of lighting management. The analysis revealed several significant findings. Firstly, the identified behavioral patterns allowed for a deeper understanding of occupant behavior and preferences regarding lighting usage. By capturing and analyzing high-resolution data from smart sensors, we were able to identify patterns of lighting utilization, including peak hours, periods of low activity, and weekly patterns. These identified patterns serve as a foundation for developing intelligent lighting management strategies. By implementing adaptive lighting control systems that align with these patterns, energy consumption can be optimized, reducing unnecessary lighting usage during periods of low occupancy or when natural light is available. This targeted approach to lighting management has the potential to significantly improve energy efficiency within smart buildings.

5.2 Insights from Data Collected from Different Sensors and the Identification of Interdependencies and Causal Relationships

Data collected from a variety of sensors, including PIR sensors, artificial light intensity sensors, shading usage patterns, and sunlight intensity sensors, provided valuable insights into the interdependence and potential causal relationships among variables. We integrated and analyzed the data from these sensors using cluster analysis, resulting in a comprehensive understanding of the factors influencing energy consumption. Our investigation uncovered correlations between various variables, allowing us to identify relationships that contribute to energy inefficiency. For example, we discovered strong correlations between PIR sensor-detected occupancy patterns and artificial light usage, indicating the potential for occupancy-driven lighting control strategies. We also discovered that the interaction of shading usage patterns, sunlight intensity levels, and artificial light intensity levels had a significant impact on energy consumption. These findings highlight the significance

	Min-Max Scaler					Standard Scaler					Robust Scaler				
Cluster	# of data points	activity proportion	artificial light proportion	sunlight level	shading level	# of data points	activity proportion	artificial light proportion	sunlight level	shading level	# of data points	activity proportion	artificial light proportion	sunlight level	shading level
0	16841	0.37	23.84	32284	10.02	59032	0.02	1.33	2952	1.00	11727	0.53	34.37	25330	14.91
1	59117	0.02	1.36	2952	1.03	16926	0.37	23.84	32163	10.10	8257	4.33	11.15	17819	17.89
2	8273	4.33	39.33	26477	17.88	8273	4.33	39.33	26477	17.88	64247	0.02	1.23	6559	0.86

Table 9: Results of clustering all features

of considering multiple factors in energy management strategies within smart buildings.

5.3 Challenges and Limitations in Analyzing Time Series Data and Cluster Analysis Methods

Throughout our analysis, we encountered challenges and limitations stemming from potential data inaccuracies resulting from preprocessing. While we aimed to enhance data completeness, compromises in accuracy may impact the reliability of our findings.

Data incompleteness and potential inaccuracies can arise due to sensor malfunctions, intermittent connectivity issues, or limitations in data collection methods. As a result, missing data points and gaps in the analysis may occur. Additionally, sensor calibration issues, such as inaccurate readings or inconsistencies among sensors, may introduce noise and further compromise data accuracy.

Furthermore, accurate time series analysis and cluster analysis methods necessitate expertise in handling large volumes of high-resolution data. Choosing appropriate clustering algorithms, tuning parameters, and identifying meaningful clusters require careful consideration and domain expertise especially when the dataset's accuracy may get jeopardized.

These limitations in our study must be acknowledged since they may affect the interpretation and generalization of our findings. Future research should try to solve these problems by developing better data gathering and preprocessing approaches, as well as improving sensor measurement accuracy. Additionally, the development of robust methodologies and tools for analyzing imperfect datasets can contribute to more accurate and reliable results in the field of smart building energy efficiency.

6 CONCLUSION

The research question of this study was to explore how intensive time series analysis and clustering methods can be utilized to identify behavioral patterns within smart buildings and improve energy efficiency. By analyzing the energy usage data collected from sensors in the LAB42 building at UvA, we aimed to gain insights into energy consumption patterns and identify areas where energy efficiency measures could be implemented.

Through the application of intensive time series analysis and clustering techniques, we were able extract valuable information. The use of K-means clustering allowed us to group similar data points based on their characteristics, particularly focusing on the correlation between artificial light levels and passive infrared (PIR) activity proportions. By identifying these behavioral patterns, our study provides insights into areas where energy efficiency improvements can be made within the smart building.

The findings highlight the importance of monitoring and adjusting energy management strategies during weekends, as Cluster 2 showed low activity but continued high artificial light usage during those periods, indicating potential inefficiencies and energy waste.

In this study, we focused primarily on energy usage data collected from sensors within the smart building. However, future work could explore the integration of additional data sources, such as weather data, occupancy data. By incorporating a wider range of data, researchers can gain a more comprehensive understanding of energy consumption patterns and identify additional factors that influence energy efficiency.

While K-means clustering provided valuable insight on the data, there is more advanced clustering algorithms that can have better investigative power for time series data.

The present study focused on a specific smart building, and future work should aim to validate the findings in different settings and across a larger sample size. Conducting similar analyses in various smart buildings can help assess the generalizability of the identified behavioral patterns and explore their scalability in different contexts. It would be possible as well to add to the current building for domain based features, like the orientation of building the get better understanding of lest impactful features in this study.

By pursuing these avenues for future work, researchers can continue to advance the field of intensive time series analysis and clustering methods in smart buildings, leading to improved energy efficiency and sustainability in the built environment.

REFERENCES

[1] Daniel B. Araya, Katarina Grolinger, Hany F. ElYamany, Miriam A. M. Capretz, and Girma T. Bitsuamlak. 2016. Collective contextual anomaly detection framework for smart buildings. *2016 International Joint Conference on Neural Networks (IJCNN)* (2016), 511–518.

[2] César Benavente-Peces. 2019. On the Energy Efficiency in the Next Generation of Smart Buildings—Supporting Technologies and Techniques. *Energies* 12, 22 (2019). https://doi.org/10.3390/en12224399

[3] Donald J. Berndt and James Clifford. 1994. Using Dynamic Time Warping to Find Patterns in Time Series. In *KDD Workshop*.

[4] Michael Berthold and Frank Höppner. 2016. On Clustering Time Series Using Euclidean Distance and Pearson Correlation. (01 2016).

[5] Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep Learning for Anomaly Detection: A Survey. arXiv:1901.03407 [cs.LG]

[6] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.* 41, 3, Article 15 (jul 2009), 58 pages. https://doi.org/10.1145/1541880.1541882

[7] Andrew A. Cook, Göksel Mısırlı, and Zhong Fan. 2020. Anomaly Detection for IoT Time-Series Data: A Survey. *IEEE Internet of Things Journal* 7, 7 (2020), 6481–6494. https://doi.org/10.1109/JIOT.2019.2958185

[8] Huyen Do and Kristen Cetin. 2019. Data-Driven Evaluation of Residential HVAC System Efficiency Using Energy and Environmental Data. *Energies* 12 (01 2019), 188. https://doi.org/10.3390/en12010188

[9] Iea. 2022. World energy outlook 2022 – analysis. https://www.iea.org/reports/world-energy-outlook-2022

[10] Dongyu Liu, Sarah Alnegheimish, Alexandra Zytek, and Kalyan Veeramachaneni. 2022. MTV: Visual Analytics for Detecting, Investigating, and Annotating Anomalies in Multivariate Time Series. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 103 (apr 2022), 30 pages. https://doi.org/10.1145/3512950

[11] Ketan Rajshekhar Shahapure and Charles K. Nicholas. 2020. Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (2020), 747–748.

[12] Hanaa Talei, Driss Benhaddou, Carlos Gamarra, Houda Benbrahim, and Mohamed Essaaidi. 2021. Smart Building Energy Inefficiencies Detection through Time Series Analysis and Unsupervised Machine Learning. *Energies* 14, 19 (2021). https://doi.org/10.3390/en14196042

[13] Alexander Tureczek, Per Sieverts Nielsen, and Henrik Madsen. 2018. Electricity Consumption Clustering Using Smart Meter Data. *Energies* 11, 4 (2018). https://doi.org/10.3390/en11040859

[14] Chunhui Yuan and Haitao Yang. 2019. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J* 2, 2 (2019), 226–235. https://doi.org/10.3390/j2020016

[15] Guodang Zhao, Xin Wang, Dezhi Zheng, and Changde Yang. 2023. Analysis of the Sustainable Driving Effect of Building Energy Consumption on Economic Development Based on the Sustainable Driving Force Model. *Buildings* 13, 5 (2023). https://doi.org/10.3390/buildings13051180