A systematic approach to evaluate energy behavior in residential buildings based on mining occupants’ behavioral data

Sajad Mohammadrezakhani

A Thesis in
The Department
of
Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Applied Science (Building Engineering) at
Concordia University
Montreal, Québec, Canada

August 2019

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared
By: **Sajad Mohammadrezakhani**

Entitled: **A systematic approach to evaluate energy behavior in residential buildings based on mining occupants' behavioral data**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Building Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Z. Chen __________ Chair / Examiner

Dr. A. Hammad __________ Examiner

Dr. M. Ouf __________ Examiner

Dr. F. Haghighat __________ Supervisor

Approved by Dr. F. Haghighat __________
Chair of Department or Graduate Program Director

Dr. A. Asif __________
Dean of Faculty

Date September 16, 2019

Abstract

A systematic approach to evaluate energy behavior in residential buildings based on mining occupants' behavioral data

Sajad Mohammadrezakhani

In this study, a new data mining-based methodology is developed to evaluate energy-related behavior of occupants in residential buildings. In sections 3.1 and 3.2 Occupant Activity Indicator (OAI) and Residential Energy Intensity Indicator (REII) are introduced as two new definitions which are used in this study. The proposed methodology to evaluate the energy-related behavior of the buildings' residents is based on the difference between the target REII and actual REII. The dissimilarity, which is found between the target and the actual REII, can be used to calculate the potential energy wastage/saving by occupants in different zones and different times in the building. The practicality of the proposed data mining framework is tested by applying it to a one-year dataset collected in a three-bedroom apartment in Lyon, France. The methodology applied to all zones of the apartment to evaluate the occupants' energy-related behavior in different zones. As a result, the time and location for potential energy savings by occupants is identified.

The obtained results show that occupants need to be more cautious about their energy consumption in zones 2 and 3 of the apartment. Moreover, the possible energy-wastage behavior in zones 1 and 4 is less than zones 2 and 3, even though the contribution of zone 4 to the energy consumption is significantly higher than the other zones. Besides, by the developed methodology location and time for the best and the worst energy-related behavior by the building's occupants are defined. Furthermore, the variations of occupants' energy-related behavior in the apartment, are identified by time of day, day of week, and months.

Employing the proposed methodology is beneficial for buildings' occupants to raise their awareness regarding energy consumption. Also, it gives the decision-makers a practical insight into the system behavior, enabling them to create incentives/charges for residential buildings' inhabitants to modify their energy-related behavior.

Acknowledgement

I would like to express my deepest gratitude to the following people who, in one way or another, contributed to making this study possible.

Prof. Fariborz Haghighat who, honored me with giving me the opportunity to work under his supervision and supported me in many ways to broaden my knowledge.

All my colleagues and friends in Energy and Environment Group at the Department of Building, Civil, and Environmental Engineering, especially Dr. Karthik Panchabikesan, for the practical help I received from him. Also, Dr. Benjamin C. M. Fung and Dr. Kobra Khanmohammadi for their help with data mining issues.

Last but not least my mom who devotes her life to my success, my lovely spouse and son for their part in my journey, and my dearest life coach and previous supervisor, Prof. Mehdi N. Bahadori for his practical and emotional support and paternal guidance that has made a huge difference for me in The University of Life[1].

[1] M. N. Bahadori, The University of Life. Blue Dolphin Publishing, 1993

Table of Contents

List of Figures

List of Tables

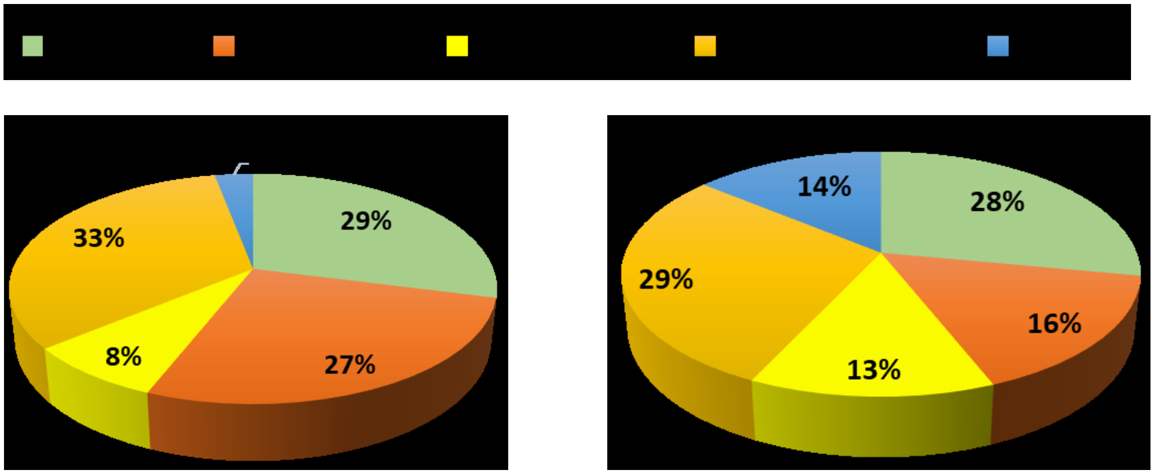
List of abbreviations

Abbreviation	Description
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ARM	Association Rule Mining
ASHRAE	American Society of Heating, Refrigerating, and Air-Conditioning Engineers
BPNN	Back Propagation Neural Network
CART	Classification And Regression Trees
DBI	Davies-Bouldin Index
DL	Deep Learning
DM	Data Mining
DT	Decision Tree
EBI	Energy-related Behavior Index
EII	Energy Intensity Indicator
ELA	Equivalent Leakage Area
EnMS	Energy Management System
EUI	Energy Use indicator
FFBPNN	Feed Forward Back Propagation Neural Network
GLM	Generalized Linear Model
GRNN	General Regression Neural Network
HLC	Heat Loss Coefficient
HVAC	Heating, Ventilating, and Air-Conditioning
IMG	Inhomogeneous Markov Chain
KDD	Knowledge Discovery in Dataset
LR	Linear Regression
MAE	Mean Absolute Error
MAE	Mean Error
MG	Multivariate Gaussian
MRE	Mean Relative Error
NCG	Normalized CO_2 Generation
NEC	Non-steady Energy Consumption
NNEC	Normalized Non-steady Energy Consumption
OAI	Occupant Activity Indicator
RBFNN	Radial Basis Function Neural Network
REII	Residential Energy Intensity Indicator
RH	Relative Humidity
RMSE	Root Mean Square Error
SVM	Support Vector Machine

1. Introduction

1.1. Energy usage awareness in buildings

Due to environmental issues, demands for energy saving and improved energy efficiency are becoming increasingly important. As it is illustrated in Figure 1, the building sector has a significant contribution to the world's total energy consumption [1], [2]. Thus, researchers are working on employing different methods to enhance buildings' energy performance.



a. Global status of residential energy consumption *b. Canada status of residential energy consumption*

Figure 1. contribution of the building's energy use to the total energy consumption [2]

Knowing the role of different influencing factors and the importance of the type of end-use in energy performance of buildings is essential to solving energy-related problems in this sector. As an example, among different energy consumers in residential buildings, the most important one is the HVAC[1] system, which uses around half of the total energy in buildings in the USA [1]. This is the reason behind the increasing interests for studies on the development of passive and low-energy air conditioning system in buildings [3]–[7].

1.2. Energy-related data

A wide range of parameters must be taken into consideration to analyze energy behavior of buildings [8]. These factors, which are influencing energy performance of buildings, can be used as the inputs in the analyses. Parameters which are influencing energy performance in building engineering can be classified into three major categories, including weather-related parameters, occupant related parameters, and physical parameters. Physical parameters group includes building-related factor and building services system parameters. Table 1 shows examples for the above mentioned three categories.

Table 1. Important variables in energy analysis in building engineering

Examples of driving factors			
Weather-related parameters	**Physical parameters**		**Occupant related parameters**
	Building services system	Building related parameters	
Outdoor air temperature & humidity	Household appliances	Conductivity of building envelope	Occupant presence
Wind speed	Lighting	ELA[2]	
Air pressure	HVAC[3]	HLC[4]	Their energy-related behavior
Solar radiation	Control system	Size and orientation	
The effects of these parameters have well recognized			The effects of these parameters are still oversimplified

The effects of weather-related factors and physical parameters are well recognized [9], [10]. However, due to the highly stochastic nature of occupants' effects on energy use in buildings, these parameters are still oversimplified. For this reason, in recent years several studies have been conducted to find a more accurate answer to the question of how occupants can affect energy consumption in buildings [3], [11]–[31].

[1] HAVC: heating, ventilating, and air conditioning

[2] Equivalent leakage area

[3] Heating, ventilating, and air conditioning

[4] Heat loss coefficient

1.3. Methods of analyzing energy performance of buildings

Generally, methods of investigating energy performance of buildings can be classified into three major categories, including engineering method, statistical techniques, and data mining. For the engineering methods, which is also called as simulation, fluid mechanics and heat transfer equations are employed to analyze the energy performance of the buildings. For the statistical methods and data mining, historical data is mostly used. Benefits and limitations of these three groups of approaches are discussed adequately in several studies [12], [30].

Employing simulation techniques enables researchers to analyze buildings' energy performance under different conditions. Also, these methods are adequately developed during the past decades, and several software tools are available for engineers [32]. However, according to the literature, due to the highly stochastic nature of occupants' behavior and their presence in buildings, the performance of these methods in dealing with occupied buildings is not as high as with unoccupied ones [12], [26], [28], [30].

Regarding statistical methods, these techniques are helpful in estimating useful energy indexes such as energy intensity index (EII) or energy use intensity (EUI). Also, they can be used to evaluate some important driving factors in energy performance of buildings such as equivalent leakage area (ELA) and heat loss coefficient (HLC). A significant advantage of these methods is their simplicity and widespread familiarity [12], [30]. For this reason, energy experts usually employ statistical techniques for energy auditing based on EnMS[1] standards.

Developing data mining frameworks to analyze energy-related data is the third method. By definition, data mining is "The analysis of large observational datasets to find unsuspected relationships and to summarize the data in novel ways so that data owners can fully understand and make use of the data" [33]. The limitations of other methods and capability of data mining paved the way for DM to become an emerging method to solve energy-related

[1] EnMS: Energy Management System

problems in building engineering. In the next section, more details on data mining and its application in building engineering is going to be discussed.

Having a higher accuracy in the estimation of energy use patterns enables building owners to take appropriate measures to reduce energy consumption in buildings. For this reason, several studies have been conducted to predict energy demand in buildings. In this regard, Zhao and Magoulès reviewed advantages and limitations of different prediction methods using in building engineering, including statistical tactics, engineering methods (both simplified simulation and detailed simulation) and the two most widely supervised machine learning techniques, artificial neural networks (ANN) and support vector machines (SVM) [12]. They reported that due to the capability of ANN and SVM in dealing with non-linear problems, these techniques are very applicable for predicting energy demand in buildings.

In Table 2, the advantages and limitations of different methods of working on energy-related data are presented. In the following, a brief introduction to the most commonly used data mining techniques in building engineering is also provided.

Table 2. Methods of working on energy-related data

Exploring building's Energy- Related Data					
	Statistical Methods		Engineering Methods		Data Mining Frameworks
	Regression analysis	Correlation analysis	Detailed simulation	Simplified simulation	
Strengths	Simplicity and widespread familiarity Comparatively efficient		Ability to analyze building energy performance under various conditions Are adequately developed in the past decades and several tools are available		Can be used to extract interesting, useful, and previously unknown and unexpected knowledge from the dataset
Limitations	Unable to exploit unexpected hidden information Not understandable and interpretable for common users		Unable to exploit unexpected useful information can only imitate some typical activities in a rigid way Comparing to unoccupied buildings, does not perform well in dealing with occupied buildings		Lack of enough experience in the application of DM in building engineering

1.4. Knowledge Discovery in Database (KDD)

Data mining or Knowledge Discovery in Database (KDD) is the process of extracting previously unknown and profitable knowledge in big data streams. In building engineering, the extracted information about energy behavior of the system can be highly beneficial for building owners and decision-makers.

Although in many research fields, such as medicine, marketing, and social science data mining have been largely used, the application of DM in building energy-related issues is still in its elementary phases [22]. Data mining frameworks are developed to find out hidden-useful information from energy-related data. The extracted practical knowledge can be used in improving energy performance of buildings and modification of occupants' energy-related behavior. In recent years, energy experts and researchers developed data mining frameworks to exploit the capability of machine learning techniques to understand hidden correlations and associations in energy-related data and predict energy performance of buildings [11], [15], [22], [27], [29], [30], [34], [35]. The application of data mining in building engineering includes four major steps that must be taken. These steps are data selection, data preprocessing, machine learning, and data interpretation and knowledge extraction.

1.4.1. Data selection

Data selection is performed based on the goal of the task. In this step, a target dataset is created in the form that is shown in Table 3. In Table 3, each row of the dataset is one example (e.g. one building), and the columns represent influencing factors[1]. In this example, energy-related data for 10 houses (y_j (X_i) and Y(X_i) j=1 to 5, i=1 to 10) is presented. Here, the goal could be developing a model, which is able to learn from the example set how to predict energy consumption of new buildings $Y(X_{new})$ (e. g. house 11), or to extract hidden relationships between attributes.

[1] also called variables, attributes, features, etc.

Table 3. Example of a dataset

		Attributes					
		y1	y2	y3	y4	y5	Y
Examples		Floor area (A)	Number of inhabitants (N)	Outside Temperature (T)	Solar Radiation (R)	Equivalent leakage area (ELA)	Annual energy consumption
House 1	X_1	y1 (X_1)	y2 (X_1)	y3 (X_1)	y4 (X_1)	y5 (X_1)	Y (X_1)
House 2	X_2	y1 (X_2)	y2 (X_2)	y3 (X_2)	y4 (X_2)	y5 (X_2)	Y (X_2)
House 3	X_3	y1 (X_3)	y2 (X_3)	y3 (X_3)	y4 (X_3)	y5 (X_3)	Y (X_3)
House 4	X_4	y1 (X_4)	y2 (X_4)	y3 (X_4)	y4 (X_4)	y5 (X_4)	Y (X_4)
House 5	X_5	y1 (X_5)	y2 (X_5)	y3 (X_5)	y4 (X_5)	y5 (X_5)	Y (X_5)
House 6	X_6	y1 (X_6)	y2 (X_6)	y3 (X_6)	y4 (X_6)	y5 (X_6)	Y (X_6)
House 7	X_7	y1 (X_7)	y2 (X_7)	y3 (X_7)	y4 (X_7)	y5 (X_7)	Y (X_7)
House 8	X_8	y1 (X_8)	y2 (X_8)	y3 (X_8)	y4 (X_8)	y5 (X_8)	Y (X_8)
House 9	X_9	y1 (X_9)	y2 (X_9)	y3 (X_9)	y4 (X_9)	y5 (X_9)	Y (X_9)
House 10	X_10	y1 (X_10)	y2 (X_10)	y3 (X_10)	y4 (X_10)	y5 (X_10)	Y (X_10)

1.4.2. Data preprocessing

Data preprocessing is the process of preparing a dataset for machine learning step and generally includes two sublevels, which are data cleaning and data transformation. Data cleaning is removing noises and outliers from the dataset. Due to incompatibility in the dataset, data transformation must be performed in many cases. For example, when we have a different range of variables, a normalization technique is usually used to homogenize the range of distribution. For instance, in the dataset mentioned above in Table 3, y1 (floor area) could be in the order of 100 square meters and y2 (number of inhabitants) is in mostly less than 5. In such cases, it is recommended to normalize the data before starting the processing step. One of the most commonly used normalization techniques is min-max normalization, which enables the user to scale values in predetermined ranges (e.g. [0,1]).

$$\frac{y'_1 - y'_{1\_min}}{y'_{1\_max} - y'_{1\_min}} = \frac{y_1 - y_{1\_min}}{y_{1\_max} - y_{1\_min}} \xRightarrow{\substack{y'_{1\_min}=0 \\ y'_{1\_max}=1}} y'_1 = \frac{y_1 - y_{1\_min}}{y_{1\_max} - y_{1\_min}}$$

Equation 1

1.4.3. Machin learning step

This step is matching one or more machine learning techniques to process transformed-data from the preprocessing step. Based on the goal of the task, the data analyst selects and combines different machine learning techniques to develop an algorithm to extract knowledge from the dataset.

Generally, machine learning techniques are classified into two categories, supervised learning and unsupervised learning.

1.4.3.1. Supervised machine learning

In supervised learning, the data analyst is aware of existing dependencies between input and output. In other words, the user knows that there might be a relationship between input and output. Generally, supervised learning techniques are aimed to do classification or regression (Figure 2), which means that by employing supervised machine learning, the data analyst can either find regression among variables or classify parameters into different predefined groups. Regression is used to predict the results in the form of continuous output, and by classification, we can deal with the categorical output. Among various supervised learning techniques, researchers in the field of building engineering prefer three techniques, which are artificial neural networks (ANN), support vector machine (SVM), and decision tree (DT).

1.4.3.2. Unsupervised machine learning

Employing unsupervised learning techniques, the data analyst can deal with unlabeled data. These techniques are employed when the user has no or limited idea about the potential results. The most popular unsupervised techniques are cluster analysis and association rule mining. In the following, a concise introduction to ANN, SVM, DT, Clustering, and ARM is presented. Besides, the strengths and limitations of each technique are going to be discussed.

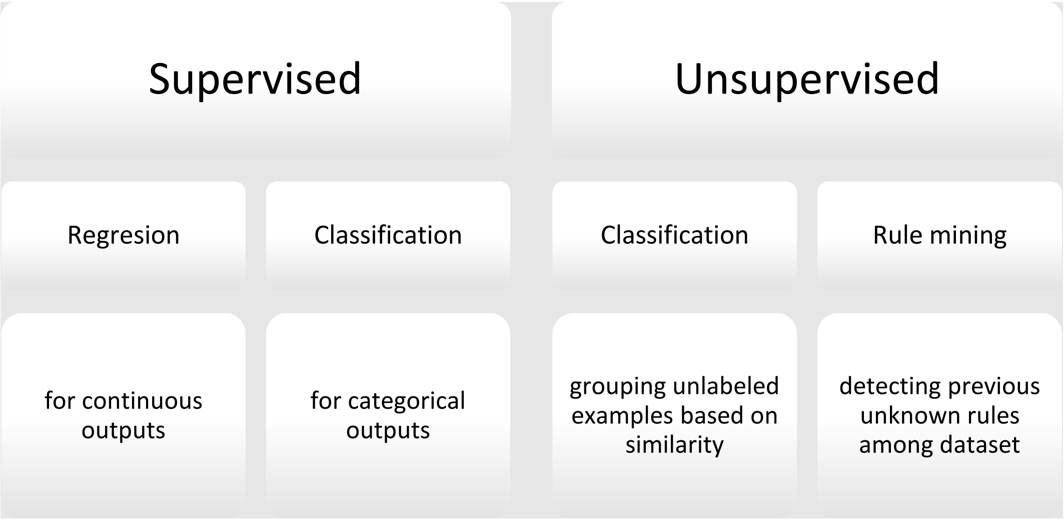


Figure 2. Supervised and unsupervised machine learning techniques [36]

1.4.3.3. Artificial neural network (ANN)

ANN is a supervised machine learning technique based on the neural structure of the brain. Three kinds of layers exist in the architecture of ANNs: they are input layer, hidden layer(s), and the output one (Figure 3). Each unit in the network is connected to all units in previous and next layers. This connection is by the matrix of weights ($\Theta^{(j)}$), which is controlling the function mapping from one layer to the next layer [37].

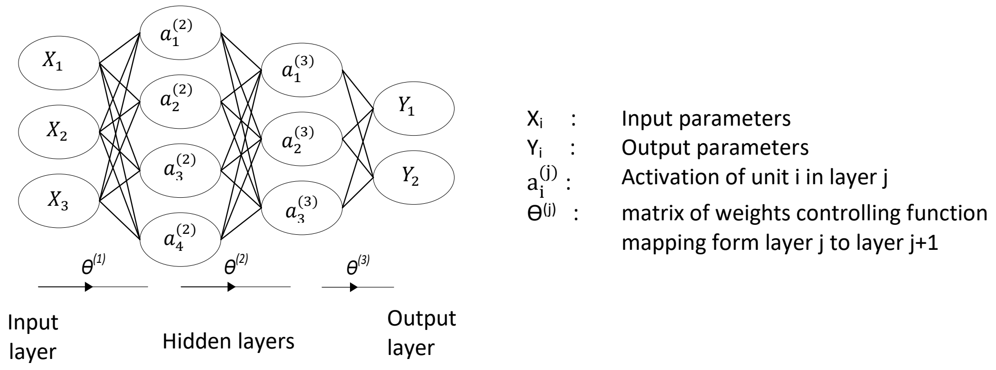


Figure 3. Artificial Neural Networks Architecture [37]

Artificial neural networks algorithms are the most widely used artificial intelligence in the application of building energy assessment. Employing ANN enables the user to model complex relationships between different influencing variables [37]. Also, in comparison to other models, ANN shows a better performance in predicting a large set of variables [38]. Commonly used ANN algorithms in building engineering are feed forward back propagation neural network (FFBPNN), General regression neural network (GRNN), and radial basis function neural network (RBFNN).

1.4.3.4. Support vector machine (SVM)

SVM is a supervised machine learning technique is used for both classification and regression. SVM is a highly effective technique in dealing with nonlinear problems, especially when the number of training examples is small. Due to the large-margin-classifier characteristic of the optimization problem of SVM, by employing support vector machine we have extra safety factor in the results of classification. In other words, by employing SVM there is no probability for the output classes [37]. Another strength of SVM is that the optimization problem is a convex problem. Therefore, the possibility of entrapping in local minimums during the optimization process can be considered zero (Figure 4). However, a serious limitation of support vector machine is its slow operation, especially when the number of attributes is more than the number of training set, and if the training set is very large.

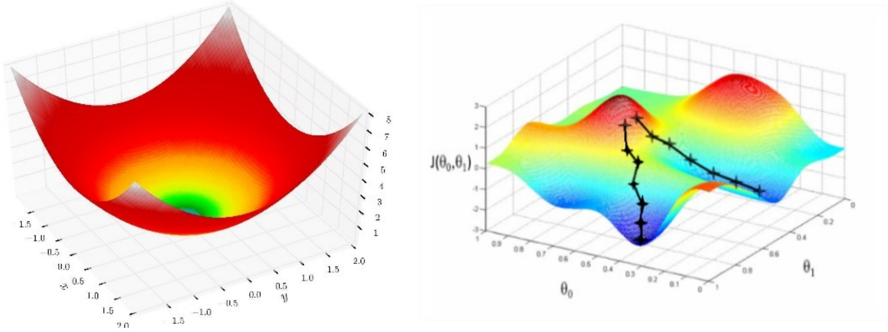


Figure 4. convex and non-convex optimization problems[1]

1.4.3.5. Decision tree

DT is a supervised machine learning model that can be used for both categorical and continuous outputs. A significant strength of decision tree lies in its understandability and interpretability, even for users without high level of knowledge [36]. Employing DT enables the

[1] Photos: https://www.coursera.org

data analyst to provide clear and useful information about the dependencies between outputs and influencing factors. Also, by DT decision rules can be easily obtained without needing high computational efforts. Researchers employ three commonly-used algorithms to generate decision tree, including ID3 [39], classification and regression trees (CART) [40], and C4.5 [41]. However, due to the ability of C4.5 algorithm in avoiding bias/variance (overfitting/ underfitting), it is the recommended algorithm in generating decision tree [42].

1.4.3.6. Cluster analysis

Clustering is an unsupervised classification technique which enables the users to group examples in previously unknown categories. The classification is done based on the similarity and dissimilarity of the instances. In the unsupervised classification (clustering), there are no predefined groups, and the machine finds the optimum number of groups. In contrary, in the supervised classification (e.g. DT, SVM, ANN, etc.) the user defines the output classes. Figure 5 shows examples of supervised and unsupervised classification.

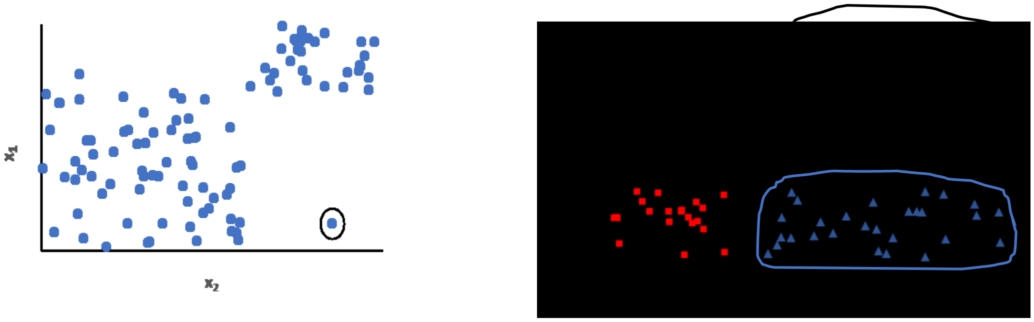


Figure 5. Supervised and unsupervised classification

K-means algorithm, along with the Euclidean distance measure is the widely used algorithm for clustering in building engineering. However, other algorithms (e.g. k-medoids, density-based, hierarchical clustering, etc.) are used in some problems as well. To find out the optimum number of clusters, usually two techniques are recommended, the Davies-Bouldin index (DBI) [36] and the elbow method [37]. Also, a combination of these two techniques can be used, as well. Interested readers are referred to the introduced references for more information.

1.4.3.7. Association rule mining (ARM)

ARM is an unsupervised machine learning technique with high performance in finding correlations and associations among different variables in the dataset. Implementation of ARM enables the data analyst to conduct the process by changing the amount for thresholds (support and confidence) to find unexpected and interesting relationships between variables. The output of ARM is a set of rules that are used to indicate patterns of variables which are frequently associated together. The ARM algorithms which are mostly used in the field of building engineering are Apriori and the FP-growth algorithms [36].

Table 4 summarizes the advantages and disadvantages of commonly used data mining techniques in building engineering.

Table 4. Popular data mining techniques used in building engineering

Technique	Commonly used algorithm	Benefits	Limitations
Decision tree	C4.5 ID3 Cart	its ease of use; able to generate accurate predictive models understandable and interpretable structures provide clear and useful information on corresponding domains perform classification and prediction tasks rapidly without requiring much computation efforts decision rules can be easily generated by traversing a path from the root node to a leaf node represent the rules visually and explicitly	is more appropriate to predict categorical variables than numerical variables the performance in dealing with non-linear problems is not as good as linear ones
ANNs	BPNN RBFNN GRNN	ability to model complex relationships between inputs and outputs is more suitable to predict a large set of parameters is the most widely used artificial intelligence in the application of building energy prediction Lots of previous works can be found good performance in solving non-linear problems	operate like a “black box” the process is not understandable and interpretable, especially for common users justifying the operation is not easy the relationship between an individual influencing factor and output cannot be observed directly
SVMs	Gaussian kernel	highly effective models in solving non-linear problems even with small quantities of training data extra safety factor by large margin classification no probability for classification the optimization problem is a convex problem	very slow operation comparing to other methods, especially if the number of attributes is more than number of training set, and the training set is very large operates like a “black box” is not understandable and interpretable especially for common users
Clustering	K-means k-Medoid CLARANS	its ease of use high performance in unsupervised classification high performance in removing the influences of some attributes on the dataset to see the effects of especial driving factors	chance of entrapping in local optimums need domain knowledge to interpret the classification results
ARM	FP-growth Apriori	data analyst is able to conduct the process by changing the thresholds (support and confidence)	needs expertise in the field of study to choose suitable thresholds

1.4.4. Data interpretation and knowledge extraction

Data interpretation is the final step. Here researchers with domain knowledge analyze the output of machine learning step to extract profitable and interesting knowledge from the framework.

1.5. Research objectives

The main objectives and specific objectives of the current work are presented in the following.

1.5.1. Main objectives

i. Developing a data mining framework to explore the patterns of energy consumption and its drivers in each zone of a residential apartment.

ii. Finding a proper indicator for evaluating occupants' energy-related behavior.

iii. Identifying the location and time for potential energy saving by occupants.

1.5.2. Specific objectives

i. Identifying the occupant activity indicator (OAI) as the driver for energy consumption by occupants.

ii. Identification of the zones in the apartment.

iii. By clustering, grouping the example set based on the similarities in energy consumption and level of activity by occupants in each zone.

iv. Finding target residential energy intensity indicator (target REII) as the baseline to evaluate energy consumption in each cluster of every zone of the apartment.

1.6. Thesis outline

Chapter 2 presents a literature review on the application of data mining in building engineering. The methodology and the developed data mining framework are explained in chapter 3. Chapter 4 discusses the results obtained for different zones of the apartment.

Finally, the conclusions of this research and recommendations for future studies are presented in Chapter 5.

2. Literature review

In this chapter, literature review is carried out on various application of data mining in building engineering. Section 2.1 reports how researchers employ data mining in buildings' energy problems. Sections 2.2 and 2.3 review the two main groups of studies on the application of data mining in building engineering. Finally, section 2.4 presents a summary of the literature review and describes the gap in the research.

2.1. Data mining in building engineering

A comprehensive classification of the application of data mining in building engineering is reported by Yu, Haghighat, and Fung in 2016 [22]. In this study, the related works are separated into two primary tasks, predictive and descriptive. In the predictive task, studies about energy demand estimation, building occupancy and occupant behavior, and fault detection diagnostics for building systems are grouped together. The descriptive task includes works on developing data mining framework, investigating the effects of occupants' energy-related behavior, building modeling and optimal control, and discovering and understanding energy use patterns.

Since in building engineering, data mining is mostly used for predicting energy consumption and evaluating energy behavior of the system, in this review, the focus is placed on the application of data mining in energy evaluation of buildings. In the estimation of energy performance of buildings, influencing parameters defined in Table 1 (section 1.2) are used as input. Considering the main goal of the reviewed studies, the studied-works are classified into two tasks. In the first task, the spotlight is put on occupant-related data as input, and in the second one, other influencing parameters are considered in the analysis (Figure 6). The first category (investigation of occupant-related data) is divided into two groups: (1) Investigating the effects of occupants' actions on the building energy consumption, and (2) discovering occupancy patterns, especially in buildings where the fluctuation rate is high. In sections 2.2 and 2.3 these two classes of research-works are studied in detail.

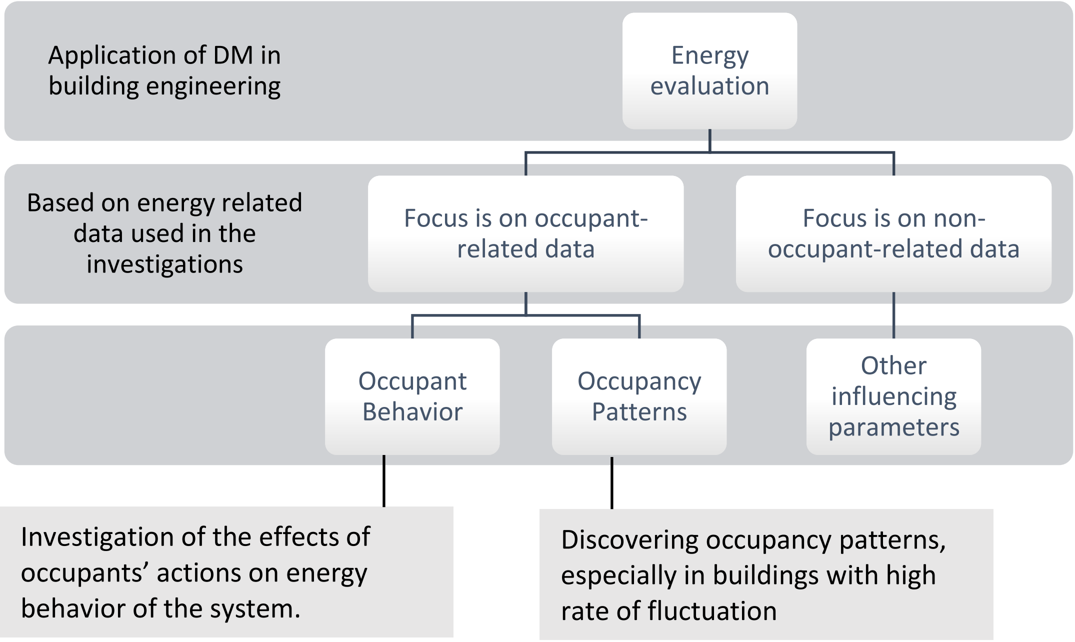

Application of DM in building engineering
Energy evaluation
Based on energy related data used in the investigations
Focus is on occupant-related data
Focus is on non-occupant-related data
Occupant Behavior
Occupancy Patterns
Other influencing parameters
Investigation of the effects of occupants' actions on energy behavior of the system.
Discovering occupancy patterns, especially in buildings with high rate of fluctuation

Figure 6. Classification of the reviewed studies

2.2. Energy evaluation without focusing on occupant-related data

2.2.1. Capability of data mining

Neto and Fiorelli compared the prediction performance of detailed simulation method, using EnergyPlus, with an artificial neural networks model for building energy consumption estimation [43]. They reported that the accuracy of prediction of both models is reasonable, while the ANNs shows a better performance in short-term predictions. They employed two

ANNs models, a simpler one which the only input is temperature, and the more complex model in which inputs are temperature, relative humidity, and solar radiation. Comparing the results to the measured data, they reported that ANN shows a fair agreement between estimated energy consumption and the actual one. They concluded that the average error for ANN is about 10% while the average error of the results of the simulation is around 13%. With regards to the accuracy of machine learning techniques in long-term prediction, credible accuracy of the artificial neural networks in annual energy consumption forecasting in energy-intensive manufacturing industries was investigated by Azadeh, Ghaderi, and Sohrabkhani [44]. They reported that ANN is highly capable of predicting long-term electricity consumption in manufacturing industries where a high rate of fluctuation exists in their energy consumption.

Comparing the accuracy of data mining with statistical models, data mining shows a better performance in predicting a large set of parameters. In this regard, Kumar et al. employed back propagation neural network (BPNN) and a conventional statistic method to evaluate energy performance and predict energy consumption in a large number of datasets [38]. The results show that the average relative percentage errors of BPNN are always lower than the average relative percentage of least square method (the statistical method). Another significant advantage of ANNs is that they are more suitable in providing predictions for multivariable problems, involving both integers and continuous variables [30].

Among different machine learning techniques, support vector machine (SVM) shows a more accurate prediction performance in comparison to others. Based on the investigation of 59 residential buildings in China, Li, Ren, and Meng compared the prediction accuracy of three common ANN algorithms to support vector machine in estimating the annual electricity consumption of the buildings [45]. In the study, Back Propagation Neural Networks (BPNN), Radial Basis Function Neural Networks (RBFNN), General Regression Neural Networks (GRNN), and SVM are employed to forecast the buildings' energy consumption. Root mean square error (RMSE) and mean relative error (MRE) were used as comparative scales. The results indicated that the prediction accuracy of SVM is significantly higher than the three ANN algorithms. It could be because of the high performance of SVM in learning, even when the size of the training dataset is small. However, it should be mentioned that despite the high efficiency of

SVM in the projection of energy behavior, in building engineering, ANN is the most commonly used artificial intelligence technique [12]. One reason could be the low-speed operation and high computational cost of SVM in comparison with ANN [37].

2.2.2. Interpreting the results

Although studies show that ANN and SVM both have good performances in energy behavior prediction, describing the operation process in these two models is not easy. In other words, because of the role of hidden layer(s) in ANN and similarity functions in SVM, the calculation process of both artificial neural networks and support vector machine is like a "black box" (Figure 7) [36], [37], meaning that justification of the process needs high domain of expertise in computer science, mathematics, and building engineering. Therefore, interpreting how individual parameters are influencing the results is not plainly understandable for many users [29].

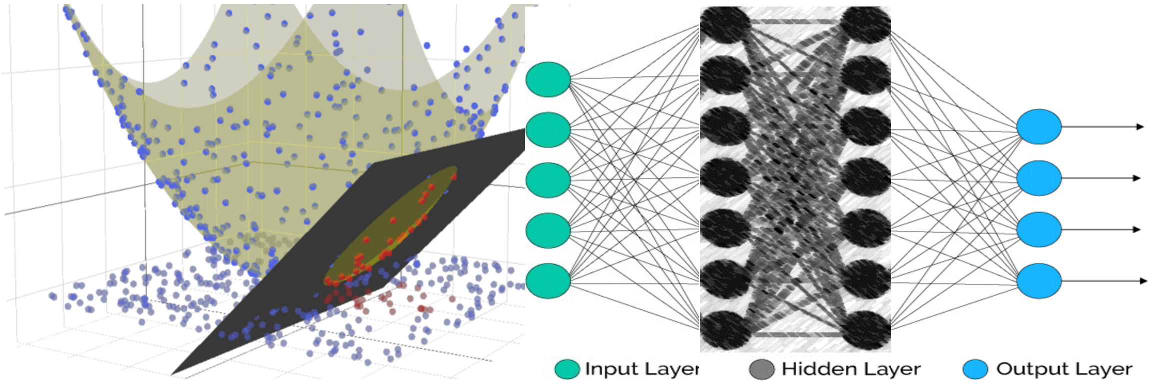


Figure 7. The operation of the hidden layers in ANN and similarity functions in SVM is not easy to interpret[1].

Considering the black-box characteristic of artificial neural networks and support vector machine, developing a more understandable data mining model seems to be reasonable. In this regard, Tso and Yau [46] employed three different models to predict electricity consumption. The models are regression, neural networks, and decision tree (Figure 8). The results demonstrated that the prediction performance of decision tree, with its more straightforward structure, is higher than other models for their case. Therefore, since DT

[1] The SVM picture: https://medium.com

procedure is significantly easier to understand and interpret than SVM and ANN, decision tree could be a good substitute for ANN and SVM in many cases.

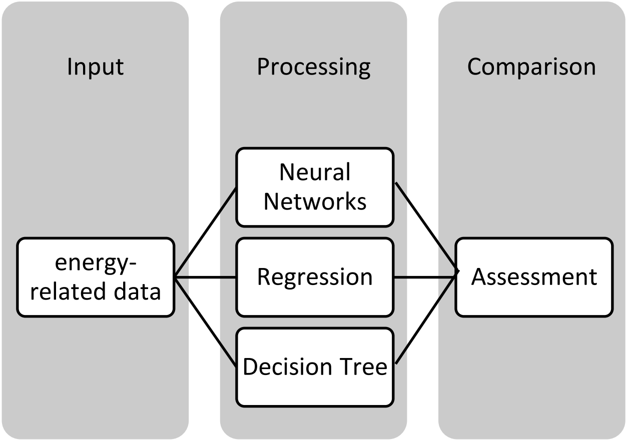


Figure 8. A comparison of regression analysis, decision tree and neural networks [46]

Considering the understandability and interpretability of decision tree Yu and Haghighat developed a DT-based model to predict EUI[1] level of buildings [28]. In their investigation, six categorical and four numerical influencing factors are grouped into four classes, climatic parameters, building characteristics, household characteristics, and energy resource. By employing these ten attributes, the decision tree is generated using C4.5 algorithm, results in obtaining eleven decision rules. The most significant advantage of the proposed model is its ability to predict and classify buildings' EUI level in an understandable way. The generated decision tree has an interpretable flowchart-like structure which enables users to extrapolate useful knowledge. The other advantage of the proposed model is that the importance of the influencing parameters on buildings energy consumption can be ranked for further decision making. The knowledge extrapolated from the study is helpful for building owners and building designers to indicate the parameters that deserve more attention. Also, it enables them to find out which energy source should be used to save energy. The proposed model provides a fast estimation of energy performance of newly constructed buildings as well.

[1] EUI: Energy Use Intensity

2.2.3. Improved models

In order to increase the accuracy of the predictive models, ensemble models are developed by researchers as well. Considering the ensemble data mining models, Fan et al. [47] examined a model which is made from a combination of eight different data mining methods to forecast the next-day energy consumption and peak power demand [47]. They concluded that the ensemble model shows a higher prediction performance than every individual one. In another study, Jovanović et al. predicted heating energy consumption of a university campus [48]. For this purpose, three different ANN algorithms are employed, including Feed Forward Back Propagation Neural Network (FFBPNN), Radial Basis Function Networks (RBFN), and Adaptive Neuro-Fuzzy Inference System (ANFIS). The estimation results illustrated that all the three individual networks have excellent agreement with measured values. Moreover, to improve the prediction performance, an ensemble model of these three neural networks is developed. Root mean square error and mean absolute percentage error are used to compare the results of the predictions by three individual networks and the ensemble model. Comparison of the results shows that the ensemble model has a higher prediction accuracy than the individual technique.

2.3. Energy evaluation by analyzing occupant-related data

To simulate a building's energy performance, modelers need inputs such as building envelope characteristics, construction materials, HVAC system size and type, interior and exterior lighting, weather information and other physical parameters, which most of them can be found in the architectural and engineering plans. However, some of these inputs are time variable and depend on the buildings' occupancy and occupants' behavior. In fact, occupants' energy-related behavior constitutes a significant portion of total discrepancies between simulated and actual energy consumption. Considering the importance of the subject, researchers developed new models to discover more accurate occupancy patterns and to analyze occupants' energy-related behavior in buildings. The results obtained from such studies are useful for prioritization of efforts of modification of occupants' behavior to reduce building energy consumption. Also, decision-makers would be able to estimate buildings' energy-saving potential by enhancing users' energy-related actions.

In a study by Wei et al. [14], the complexity of occupants' space-heating behavior in residential buildings is highlighted. They focused on 27 influencing parameters in heating behavior, which are suggested in previous studies. They reported that the influence of some factors on space-heating behavior of the occupants is well accepted by users. Outdoor temperature and dwelling type are examples of the mentioned factors which have been analyzed in many studies. However, some other parameters are considered by only a limited number of surveys and need to be investigated further. Heating price and social grade are examples of this group. It should be mentioned that based on the current body of knowledge, none of these variables can be identified as having no influence on space heating behavior.

In Figure 6, it is indicated that analyzing occupant-related data in energy evaluation of occupied buildings can be categorized into two tasks:

i. Investigating the influences of occupants' actions on energy consumption in buildings. The goal of this task is estimating energy-saving potential in buildings by improving occupants' behavior. Also, this kind of investigation helps decision-makers to acquire a real insight into energy behavior of the buildings.
ii. Studying and understanding occupancy patterns, especially in commercial and office buildings in which the fluctuation rate is usually high. The goal of the second task is acquiring more accurate energy use estimation. Therefore, decision-makers and building owners can have more practical information for further decision-making.

2.3.1. Occupants' energy related behavior data

In order to describe the large discrepancies between estimated energy consumption and the actual one in buildings, among various factors contributed to the discrepancies, researchers found that the effect of occupants' behavior on energy consumption is a notable driving factor [21], [24]. Therefore, occupant behavior is known as the most significant source of uncertainty in estimating buildings' energy consumption by simulation tools [21]. How occupants use household appliances, how they interact with building energy and system services, and how to set the comfort criteria (thermal comfort, visual comfort, and acoustic comfort), and likewise their reaction to environmental discomfort play significant roles in the operation of buildings' energy system and occupants' energy use. With regard to office buildings, how occupants use the appliances and computers [20], how they adjust thermal and visual comfort [49], and how they leave the office in non-working-hours and weekends [24] are important occupant-related factors which significantly affect energy behavior of the system.

Hong and Lin studied building energy consumption in three different climates to understand and categorize occupant behavior impacts on the energy use of private offices. Moreover, they evaluate how different types of occupant behavior affects the buildings' energy consumption [21]. First, occupant behavior in private offices is categorized into three types of workstyle: 1) the austerity workstyle; 2) the standard workstyle; and 3) the wasteful workstyle. Then, by comparing to the standard workstyle, the obtained results show that the austerity one uses up to 50% less energy, while the wasteful workstyle consumes up to 90% more energy. Another study to show the importance of user behavior in the energy performance of buildings was conducted on six randomly selected commercial buildings in South Africa, by comparing energy consumption during working hours with non-working hours [24]. The results of energy audit surprisingly demonstrated that more than 50% of total energy is used during non-working hours than official working hours.

As an indicator of occupant-energy-related behavior, D'Oca and Hong studied the patterns of window opening in 16 office buildings [49]. Logistic regression is employed to find out the most influencing factors on window opening and closing behavior of the occupants. Then, cluster analysis was used to disaggregate occupant behavior into different patterns. By analyzing the results, the top drivers for window opening/closing in the building are found. Also, the clustering analysis results in extracting useful knowledge about occupant behavior on window opening/closing patterns, which can be used in simulation programs. As an example, the patterns of window tilting angle preferences can be used in designing and estimating energy consumption in buildings and thus can be incorporated into building simulation software.

Considering the capability of clustering in isolating the effects of some selected features on instances while the impacts of other influencing factors can be eliminated, Yu et al. developed a data mining model to investigate the influences of user-related variables on energy consumption in buildings. In the developed model, first, all influencing parameters on energy consumption are taken into consideration. Then user-related parameters are isolated from other variables, and clustering performed based on the influencing factors unrelated to user behavior. As a result, four clusters are detected which physical and weather-related parameters are roughly similar in each of them. Accordingly, comparing the energy consumption of buildings in one cluster results in identifying the effects of occupant behavior on energy consumption. The comparison is made based on five different statistical measures.

In another work, to segregate the effects of occupant behavior on building energy consumption, Yu, Haghighat et al. proposed another novel data mining framework. The proposed procedure consists of three steps, including clustering, decision tree, and association rule mining [26]. First, cluster analysis is performed to eliminate the effects of non-occupant related factors on the energy consumption of the buildings. After that, decision tree is employed to predict the cluster attribution of new buildings according to the main end-use loads. Finally, association rule mining is used, results in finding interesting rules, associations, and correlations among different user activities. According to the discovered rules, recommendations were proposed to highlight energy-saving opportunities for the buildings' occupants. The considerable advantage of the proposed framework is its high efficiency in occupant behavior modification. Moreover, the identification of energy-inefficient behavior is helpful for building owners to be aware of avoidable energy waste and motivate them to modify their activities.

Ashouri, Haghighat et al. developed a building advisory system to alert the building occupants in case of any abnormal behavior [27]. Their methodology emphasizes quantitative energy savings or losses in home appliances. In the developed framework, they used different techniques of data mining, including clustering, association rule mining, and neural networks. The system scrutinizes the historical data to find good and bad energy consumption patterns.

By examining the real stream of data, any behavior opposite to the aforementioned patterns is flagged as loss of heat or saving, respectively.

2.3.2. Discovering occupancy patterns

With respect to the occupancy schedule, many modelers use simplified occupancy patterns which are available in standards, in their energy simulation programs. Designers usually refer back to ASHRAE 90.1-2004 [50], which provides standardized hourly occupancy diversity factors for different types of buildings. The recommended occupancy diversity factor by ASHRAE is shown in Figure 9. However, the issue in office buildings is that the recommended occupancy patterns in standards do not differentiate between different types of offices (e.g. private offices and multi-tenant office types). Moreover, the last update in most of these recommended schedules occurs in 1989 [51]. These are the reasons that researchers are working on developing new methods and frameworks to discover more actual schedules to use in their simulations.

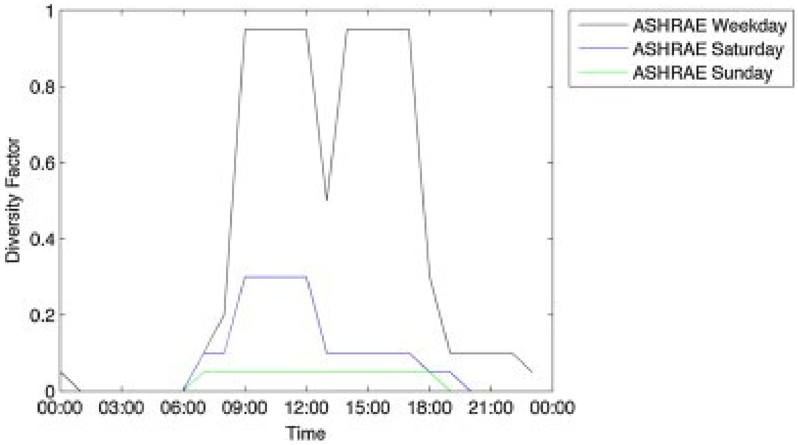


Figure 9. Recommended occupancy pattern when actual schedules are not known [50]

Studying office buildings, the arriving and leaving time, the number of occupants, the lunchtime, and the occupant presence in non-working-hours and weekends are noticeable issues in terms of energy consumption. Due to the importance of the number of users in the regulation of energy-saving after retrofit, the results of discovering more accurate occupancy schedule can facilitate energy efficiency retrofit in buildings. In this regard, Duarte et al. performed clustering on two-year collected data of a multi-tenant commercial building to find the occupancy diversity factor for different zones and office types, by time of day, day of week, month, weekdays, holidays and weekends [18]. By clustering, the results of mining occupant presence data in private offices are divided into three clusters, including high level (January, March, April, June, September, October), medium level (December, February, July), and low level (November, August). The results show that the investigated occupancy diversity factors for that building are noticeably lower than the recommended occupancy schedule in the standard.

Considering the capability of developed data mining approaches in comparison with stochastic occupancy models to build an accurate occupancy schedule, Chen and Soh analyzed performance of six different models, including a traditional baseline model (SDP), two most widely used multi-occupant models (IMC[1] and MG[2]), and three data mining approaches (ARIMA[3], ANN[4], SVR[5]) [17]. Two evaluation criteria, including Root Mean Square Error (RMSE) and Mean Error (ME), were employed to calculate the magnitude of occupancy prediction error and overestimation/underestimation of occupancy prediction, respectively. The results indicate that the accuracy of data mining approaches is significantly higher than stochastic occupancy models, which are highly limited for predicting regular occupancy in commercial buildings. In consideration of the calculated errors, ARIMA and SVR are defined as the best models for short-term and long-term predictions, respectively.

To discover more accurate occupancy schedule in buildings, Liang, Hong, and Shen developed a data mining framework, results in introducing a novel formulation to predict occupant presence in buildings [35]. The considerable advantage of the proposed framework is that only simple input data (accessing records of the building) is required for the process. In the proposed methodology, first, four occupant presence patterns are discovered by unsupervised classification. Then, decision tree is used to induce the rules of the four defined-patterns. They concluded that there is a strong relationship between the obtained patterns

[1] Inhomogeneous Markov Chain
[2] Multivariate Gaussian
[3] Autoregressive Integrated Moving Average
[4] Artificial Neural Networks
[5] Support Vector Regression

and three factors, which are outdoor temperature, daylight saving time (DST), and weekdays. Therefore, these three attributes are used to generate the decision tree. Based on the induced rules from decision tree, the occupancy schedule is predicted by a proposed equation:

$$\text{Prediction (day, t)} = \mathrm{M_{p1}(t)\,.\,P_{p1} + M_{p2}(t)\,.\,P_{p2} + M_{p3}(t)\,.\,P_{p3} + M_{p4}(t)\,.\,P_{p4}} \qquad \textit{Equation 2}$$

Where M_{pi} ($i=1,2,3,4$) denotes the mean value of the Pattern *i* and P_{pi} denotes the probability of Pattern *i*.

The prediction performance of the obtained equation is compared with two other commonly-used methods, including mean-day method and mean-week method. Root mean square error, mean absolute error, and median error are used for the comparison. According to the results of the comparison, they found that the proposed method can increase the prediction accuracy by around 30%. Also, the new model has a lower systematic tendency to overpredict or under-predict.

2.4. Summary

The classification of the reviewed literature is summarized in Figure 6. In addition, Table 5 recapitulated the reviewed literature according to the classification brought in Figure 6.

Table 5. A summary of the reviewed literature

Energy evaluation without focusing on occupant-related data		
Reference	Subject of the work	Findings
(Tso, Yau 2007)	Prediction performance of regression, neural network, and decision tree are compered in estimating energy consumption.	The precision of decision tree model, with its simpler structure, is higher than other models.
(Kumar, Aggarwal et al. 2013)	Prediction accuracy of back propagation neural network and a conventional statistic method (least square) are compared in predicting a broad set of parameters.	The average relative percentage error of BPNN is always lower than the average relative percentage of least square method.

(Zhao, Magoulès 2012)	Different methods of predicting energy consumption in buildings were reviewed and compared.[1]	ANNs and SVMs are very applicable for predicting energy demand in buildings. If the training dataset is small, SVMs is a highly effective model in dealing with non-linear problems.
(Neto, Fiorelli 2008)	Prediction performance of detailed simulation method[2] is compared with two ANNs models[3] in estimating energy consumption in buildings.	The ANN shows a better performance in short-term prediction.
(Azadeh, Ghaderi et al. 2008)	Investigating the accuracy of ANN in annual energy consumption forecasting.	ANN is highly capable of predicting long-term electricity consumption in energy-intensive cases where the fluctuation rate is high.
(Li, Ren et al. 2010)	Prediction accuracy of three common ANN models[4] is compared with SVM in estimating annual electricity consumption of buildings.	SVMs has the best performance among these four models.
(Yu, Haghighat et al. 2010)	Developing a decision tree model to predict EUI[5] level of residential buildings in different types of districts.	The proposed model is non-complex with high accuracy. The advantage of the model is its interpretability. The importance of influencing factors is ranked.
(Fan, Xiao et al. 2014) and (Jovanović, Sretenović et al. 2015)	Different data mining techniques were combined to investigate prediction performance of an ensemble model in forecasting energy consumption.	The ensemble model has more accurate prediction performance in comparison with every individual technique.

Energy evaluation by analyzing occupant-related data

Occupants' energy related behavior data

Reference	Subject of the work	Findings
(Hong, Lin 2013)	How occupant behavior can affect energy consumption in private offices by comparing different workstyles with defined-standard workstyle.	There is up to 90% potential for saving energy only by modifying occupants' workstyle.

[1] The methods are detailed simulation, simplified simulation, statistical method, artificial neural networks, and support vector machine.

[2] Using EnergyPlus

[3] A simpler model (only input is temperature) and a more complex model (inputs are weather condition)

[4] Including Back Propagation Neural Networks (BPNN), Radial Basis Function Neural Networks (RBFNN), and General Regression Neural Networks (GRNN)

[5] Energy Use Intensity

(Masoso, Grobler 2010)	Analyzing the results of energy auditing in commercial buildings and comparing energy consumption during working hours with non-working hours.	More than 50% of total energy is used during non-working hours than official working hours.
(D'Oca, Hong 2014)	Mining data of 16 private offices to discover the effects of occupant behavior in window opening/closing[1].	The top drivers for window opening/closing in the building were defined.
(Yu, Haghighat et al. 2011)	A novel methodology for mining energy-related data of residential buildings in different districts is introduced for analyzing the effects of occupant behavior on energy consumption by isolating the occupant-related factors. Discovering association and correlation among different energy-related occupant behavior.	High capability of the proposed method to improve occupant behavior to save energy. Identification of energy-inefficient behavior of occupants to avoid energy waste.
(Yu, Fung, Haghighat et al. 2011)	Mining energy-related data of residential buildings in order to investigate the contribution of occupant behavior to the buildings' EUI and analyzing five defined cases[2] to improve energy consumption in buildings.	By the proposed methodology, profitable knowledge related to energy consumption is obtained. Energy-saving potential for every building can be calculated. Practical recommendations can be made to reduce energy consumption.
(Ashouri, Haghighat, et al. 2018)	Proposing a framework for building occupants energy consumption Investigation and quantification of potential and achieved savings.	The data mining process is able to reveal potential and achieved savings that were not noticed before by conventional analysis. The model flags any abnormal energy consumption pattern and quantifies the losses. Recommendations are used to bring occupants attention to certain end-use loads that require more concern.

[1] Specifically, 1) motivational patterns, 2) opening duration patterns, 3) interactivity patterns, and 4) window position patterns were discovered.

[2] 1) End-use load shapes; 2) Variability in annual EUI of different end-use loads induced by occupant behavior; 3) Reference building and energy-saving potential; 4) Monthly variations of end-use loads induced by occupant behavior; 5) Monthly average indoor temperature of air-conditioned room

Energy evaluation by analyzing occupant-related data

Discovering occupancy patterns

Reference	Subject of the work	Findings
(Duarte, Van Den Wymelenberg et al. 2013)	Analyzing occupant presence data to find the occupancy diversity factors for different zones and office types, by time of day, day of week, month, weekdays, holidays, and weekends.	The investigated occupancy diversity factors are considerably lower than the recommended diversity factors in standards. Thus, significant potential for saving energy is existing.
(Chen, Soh 2017)	Comparing the accuracy of three data mining models[1] with two stochastic occupancy models[2] in different time horizons to discover the most accurate occupancy schedule.	The accuracy of data mining models is significantly higher than stochastic occupancy models. Regular stochastic occupancy models are highly limited for predicting occupancy in commercial buildings. ARIMA and SVR are the best models for short-term and long-term predictions, respectively.
(Liang, Hong et al. 2016)	A data mining framework and a novel formulation are developed to discover occupancy patterns.	The newly proposed formulation to predict occupant presence in buildings increases the prediction accuracy by around 30%. Also, it has a lower systematic tendency to overpredict or underpredict.

2.4.1. Inference from the literature

- There is a high potential for understanding the correlation between occupants’ behavior and energy consumption in buildings.
- There is a high interest in acquiring practical insight into the system behavior.
- There is a high interest in evaluating behavior of the system and modifying energy-related behavior of buildings’ occupants.

[1] Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Networks (ANN), and Support Vector Regression (SVR)

[2] Inhomogeneous Markov Chain (IMC) and Multivariate Gaussian (MG)

2.4.2. Summary of the reviewed literature and challenges found in the previous works

- The objective of the previous works is either analyzing occupancy patterns or energy prediction.
- Regarding occupancy patterns, most of the works are done in office buildings where the occupancy schedule is routine, and a few numbers of studies are done in residential buildings [52]. Also, the studies which are done in residential buildings mostly address either occupancy patterns or energy consumption separately [53]. Alternatively, they wanted to predict energy consumption based on the occupancy patterns [54], [55].
- However, except the paper published recently by Li et.al. [31], no study is done to identify the energy wastage in residential buildings by considering the dynamic occupancy patterns.

2.4.3. Explanation of research gap to support the objectives of current work

To evaluate energy behavior of a residential building, it is clear that only using energy consumption amount, without considering the drivers for energy consumption is not enough. As a simple example, comparing energy consumption of a house in Montreal with an apartment in Vancouver is not rational, and from an energy-expert point of view, no one can say that the inhabitants of the house in Montreal are wasting energy and occupants of the apartment in Vancouver are saving energy, only because of considering the amount of energy consumption in these two cases.

Therefore, firstly, we need to distinguish the driving factors for energy consumption in residential buildings from other parameters. After considering the influencing parameters, we are able to make logical comments on the energy performance of the buildings.

To evaluate energy behavior of a system, having a *target energy consumption* is essential. In many cases, the target energy use is called *baseline* for energy consumption. To draw a proper baseline for energy consumption, we need to know what the influencing parameters in energy consumption are. In many cases, three groups of parameters are considered as drivers for energy consumption in residential buildings, including weather-related features, building's physical parameters, and number of occupants (Figure 10) [9], [21], [28].

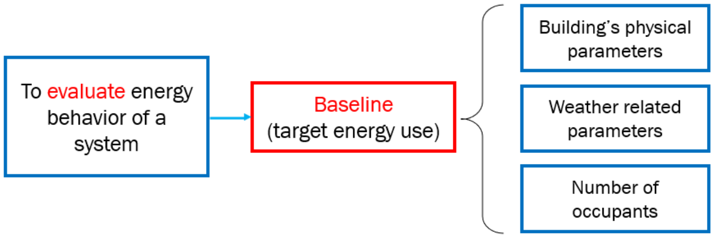

To evaluate energy behavior of a system
Baseline
(target energy use)
Building's physical parameters
Weather related parameters
Number of occupants

Figure 10. Parameters to draw a baseline for energy consumption in residential buildings. Building physical parameters including HLC[1], ELA[2], conductivity of the building envelope, building's size and orientation, operation of HVAC[3] system, household appliances, building's control system, etc. Weather-related group includes temperature, humidity, wind speed, air pressure, solar radiation, etc.

Therefore, if we consider the effects of the three groups of parameters (i.e. physical parameters, weather-related parameters, no. of occupants) on energy consumption in residential buildings, the results of comparing energy behavior of buildings are less controversial (Figure 11).

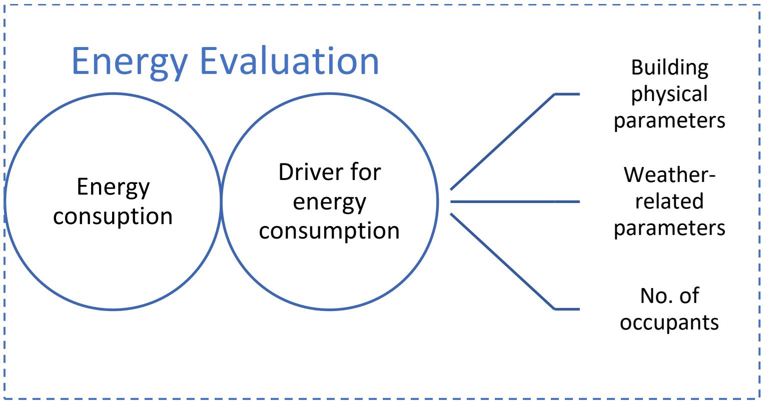

Energy Evaluation
Energy consuption
Driver for energy consumption
Building physical parameters
Weather-related parameters
No. of occupants

Figure 11.Energy evaluation in residential buildings

Now consider another scenario: two apartments with the same plan in the same building with two inhabitants each. In this case, considering building physical parameters and weather-related parameters as drivers for energy consumption to evaluate energy behavior of the two apartments results in *the same baseline* for them.

Now suppose that in apt.1, an aged retired couple is living and the occupants of apt.2 are a single parent working at home with his/her kid (Figure 12).

[1] Heat Loss Coefficient

[2] Equivalent Leakage Area

[3] Heating, Ventilating, and Air Conditioning

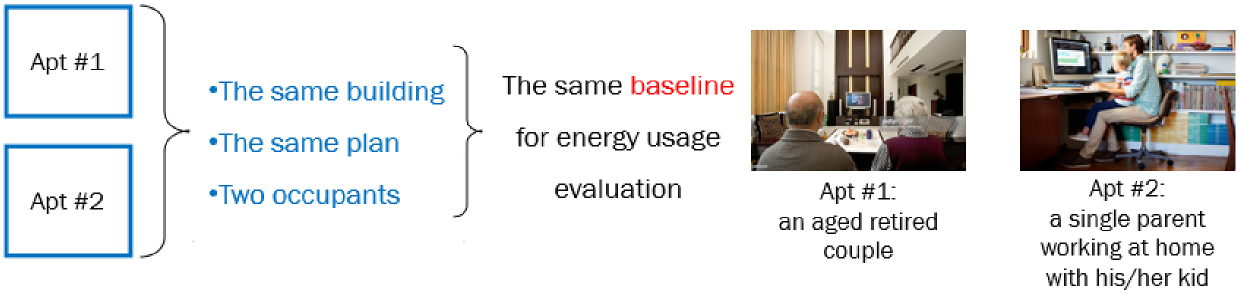

Apt #1
Apt #2
•The same building
•The same plan
•Two occupants
The same baseline for energy usage evaluation
Apt #1: an aged retired couple
Apt #2: a single parent working at home with his/her kid

Figure 12. Problem to be solved

Based on daily needs, aged-retired people usually consume less energy than kids and adults. However, since these two apartments seem to have the same baseline for energy evaluation, in this scenario, energy consumption in apt.2 is mostly more than energy consumption in apt.1. However, it is not fair to say occupants in apt.1 are saving energy, and people in apt.2 are wasting energy, without perceiving occupants' behavioral patterns, which comes from their needs.

Therefore, for such cases, we need to define a driver for energy consumption, based on the occupants' behavioral patterns and their daily needs. Two new definitions, OAI and REII, are introduced in this study to approach this problem. In the next chapter, the developed methodology to fill the mentioned gap, and the two new definitions are going to be described.

3. Methodology

To solve the mentioned problem in section 2.4.3, a novel data mining framework is developed. The purpose of this study is to evaluate energy behavior in residential buildings based on the available data. This process will be done by identifying the location and time for potential energy-wastage/saving. The results of this work can be useful in providing the inhabitants with practical advice to save energy.

In the following, two new definitions (OAI and REII) are introduced in sections 3.1 and 3.2. Then the whole methodology is shown in section 3.3, and sections 3.4 up to 3.8 present the developed data mining framework in detail. The practicality of the proposed data mining framework is by applying to a one-year dataset collected in a three-bedroom apartment in Lyon, France. The case study dataset is presented in section 3.9.

3.1. Occupant Activity Indicator (OAI)

Occupant Activity Indicator (OAI) is defined to represent the level of activity by occupants in a building. To take the occupancy patterns into account while evaluating energy consumption in residential buildings, OAI can be employed as an important driver for energy use.

When OAI is high, it is indicated that inhabitants are highly active, or there are more people. Low OAI shows that people are relatively inactive, or there is less number of people at home.

Based on the available data, occupant activity indicator is built by parameters which are indicating what the level of activeness of people is (e.g. motion detected data and indoor CO_2 level).

Before generating OAI, it is essential to find the weights of its components. For this reason, we can find how the selected features for OAI are contributing to energy consumption. Also, since the range of the selected features can be different, to avoid skewed distribution, it is necessary to normalize the data before generating OAI.

The suggested formulation for OAI is shown by equation 3:

$$\mathrm{OAI} = \sqrt{\mathrm{a_1 . D_1}^2 + \mathrm{a_2 . D_2}^2 + \cdots + \mathrm{a_n . D_n}^2} \qquad \textit{Equation 3}$$

where

D_i : the normalized occupant behavioral drivers which are contributing to energy consumption in residential buildings (e.g. motion and CO_2 generation) $(0 \leq D_i \leq 1)$

a_i : the weight of driver i in energy consumption $(0 \leq a_i \leq 1)$

3.2. Residential Energy Intensity Indicator (REII)

Energy Intensity Indicator or EII is a measure to compare energy consumption by two systems and usually considered as energy use per driver (for the energy use). The driver is determined by the subject and can be Gross Domestic Product (GDP) and Gross Product (for industries).

Following this intuition, Residential Energy Intensity Indicator or REII is defined to provide a fair evaluation of energy behavior for a residential building (as a system). REII can be used to find the time and location for potential energy wastage in residential buildings. REII is formulated as it is shown in equation 4:

$$\mathrm{REII} = \frac{\mathrm{NNEC}}{\mathrm{OAI}}$$ *Equation 4*

where

NNEC is the Normalized Non-steady Energy Consumption and OAI is the Occupant Activity Indicator.

NNEC is the energy use by appliances and other end-uses, which are following a non-steady pattern. For some end-uses, the pattern of energy consumption follows a constant trend with negligible fluctuations, even when there is no one at home (no motion is detected). For this reason, in the REII formulation, non-steady energy consumption is used instead of total energy consumption. Therefore, we can emphasize what kinds of behavior can be modified to reduce energy consumption. However, if all the end-use appliances are non-steady, NNEC = total energy consumption.

OAI shows the level of activity by people in the house. Here OAI plays the role of the driver for energy consumption in residential buildings.

High residential energy intensity indicator represents high energy consumption with a relatively low level of activity by people. Therefore, if we could find times that actual REII is higher than target REII (baseline), we can mention these moments as potential energy wastage. Contrarily, if the target REII is less than the actual one, it shows that we have a potential energy saving at that moment.

3.3. The developed framework

The developed data mining framework includes the following steps (Figure 13):

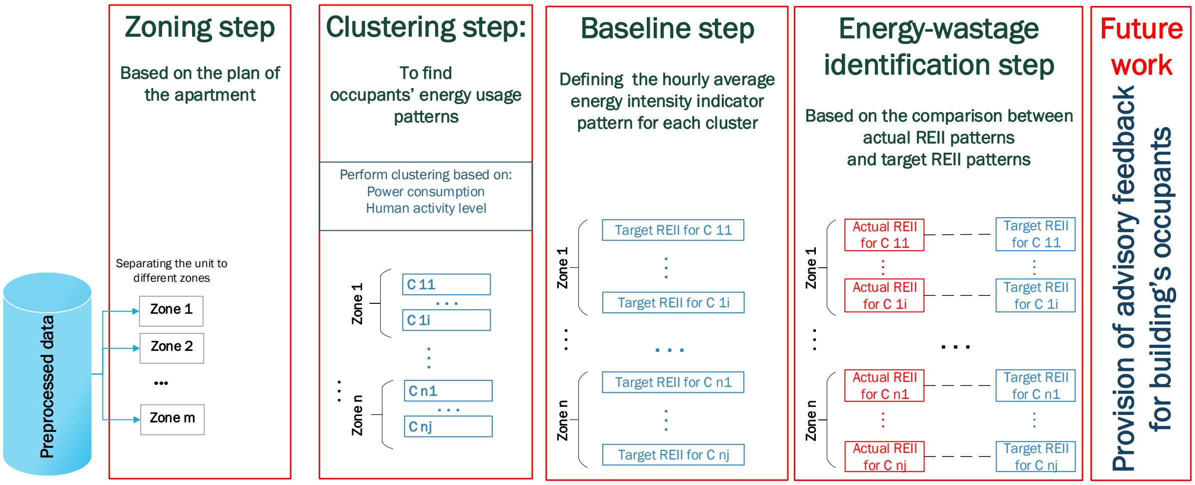


Figure 13. The whole data mining framework

i. **Preprocessing step**, includes missing value prediction, feature selection, data aggregation, normalization, and feature generation;

ii. **Zoning step**, based on the plan of the apartment, separating it to different zones;

iii. **Clustering step**, to find energy use patterns for each zone of the building;

iv. **Baseline step**, to fine target residential energy intensity indicator (REII) or baseline for each cluster in every zone of the building;

v. **Energy wastage identification step**, to find potential energy-wastage by comparing the actual residential energy intensity indicator (actual REII) to the target REII (baseline) for each cluster in every zone.

3.4. Data Preprocessing

Data preprocessing is an essential step in the process of data mining. The importance of data preprocessing is emphasized when we know that the phrase "garbage in garbage out" is particularly applicable to data mining projects. The data format must be in a proper manner to achieve better results from the applied model in data-mining problems [36].

In this study, the following data-preprocessing tasks are implemented:

3.4.1. Missing values prediction

Missing values, which can have significant impacts on the expected outcomes of the data mining, are a common phenomenon in data collection. In statistics, missing values happen when, in an observation, no information is recorded for the feature. Two kinds of missing values can be found in the datasets:

i. Sparse missing values, where the missing values happen sparsely. In this case, statistical techniques can be employed to estimate the unknown values (e.g. average of neighboring values).

ii. Continuous missing values, where we have missing data for a feature during a considerable period (e.g. one-third of total data recording period). Here, if the feature with missing values is crucial that it can not be ignored, the data analyst can deal with it as a label feature. Then, try to find a proper machine learning procedure to predict the missing values. In this study, we have the same problem. For one important sensor, three months of data is missed. Thus, different machine learning models are tested to find the missing values with the lowest possible error. In section 4.1, the procedure is provided in detail.

3.4.2. Data aggregation

Data aggregation[1] is a process in data mining where data is searched, gathered, and presented in a report-based summarized format, to achieve specific objectives [56]. Aggregation is done by climbing up hierarchically or by dimension reduction. An example of data aggregation is provided in Appendix F. Data aggregation helps the user have a good insight into the data and understand how the trends are changing.

3.4.3. Normalization:

Before performing cluster analysis, it should be noted that the parameters of our dataset in this study have different ranges. Also, in this work, the features were considered to be of equal

[1] Also called roll-up or drill-up in some references

importance. To prevent the features with large ranges (e.g. CO_2) from outweighing those with comparatively smaller ranges (Motion), min-max normalization is applied to the selected parameters before the clustering step. Because the min-max normalization performs linear normalization, its most key advantage is the ability to reserve the relationships between the initial data.

3.5. Zoning step

Knowing the location of energy wastage helps the building occupants to have a better insight into their energy-related behavior. Since one of the objectives of this research is investigating the place where the energy wastage has occurred, separating the building into zones is essential. Zoning the apartment is done based on its plan, and the investigations are done for each zone separately. As a result, occupants can identify what kind of behavior needs to be modified to reduce energy consumption.

3.6. Clustering step

The goal of this step is eliminating the effects of time, date, and season on energy consumption by buildings' occupants. Since the main purpose of this work is studying energy use patterns and occupants' activity patterns to analyze energy-related behavior of inhabitants in residential buildings, finding the time and location for potential energy wastage is necessary.

Therefore, two groups of attributes are chosen to perform clustering. The variables which represent how much the occupants are active in the building (e.g. motion and CO2) are gathered together in the first group, and features which are showing the amount of non-steady energy consumption (plug power consumption and lighting power consumption) forming the second group.

The k-means algorithm is used to perform clustering. k-means determines a set of k clusters and assigns each example to one cluster. The clusters consist of similar examples. The similarity between examples is based on a distance measure between them. The position of the center in n-dimensional space of the n attributes of an example set determines a cluster in the k-means algorithm. This position is called centroid. It can, but do not have to be the position of an example of the example set. The k-means algorithm starts with k points which are treated as the centroid of k potential clusters. These start-points are either the position of k randomly drawn examples of the input example set or are determined by the k-means. All examples are assigned to their nearest cluster (the measure type defines the nearest). Next, the centroids of the clusters are recalculated by averaging over all examples of one cluster. The previous steps are repeated for the new centroids until the centroids no longer move or max optimization step is reached. The procedure is repeated with different sets of start-points. The set of clusters that is delivered has the minimal sum of squared distances of all examples to their corresponding centroids [36], [57].

After performing min-max normalization to the selected features for clustering, a combination of Davies-Bouldin index (DBI) and Elbow method is used to find the optimum number of clusters for each zone. By definition, DBI is “the ratio of the sum of average distance inside clusters to distance between clusters” and calculated by equation 5 [58]:

$$\mathrm{DBI} = \frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{max}_{\mathrm{i}\neq\mathrm{j}}\left[\frac{\mathrm{d_i} + \mathrm{d_j}}{\mathrm{C_{i,j}}}\right] \qquad \textit{Equation 5}$$

where:

k: the number of clusters;

d_i : the average distance inside the clusters i, in other words, it is the average distance between each object in the cluster i and the centroid of cluster i;

d_j : the average distance inside the clusters j, in other words, it is the average distance between each object in the cluster j and the centroid of cluster j;

C_{ij} : the distance between the cluster centroids.

As a consequence, a smaller DBI represents a higher quality of the clustering. The k = n algorithm that carries out clusters with high intra-class similarity and low inter-class similarity has a small Davies–Bouldin index and can be considered the optimum number of clusters for the data set.

After finding the optimum number of clusters, we can perform the clustering. As a result, items in each cluster have close similarity in energy consumption, total motion, and CO2 concentration.

3.7. Baseline step

In the baseline step, a target hourly REII[1] is generated for each cluster and is used later as the baseline for that cluster. As it is shown in equation 4 (section 3.2), REII is calculated by a fraction which the numerator is NNEC[2], and the denominator is OAI[3].

To calculate NNEC, the energy use patterns of each end-use load (different plug powers and lights) must be taken into consideration. Then, those with non-steady energy consumption pattern are selected to add together. After performing min-max normalization, NNEC is ready to be used in equation 4. The process of calculating NNEC is shown in Figure 14.

Base on the available data, OAI is built by its components, which are representing the human activity (e.g. motion, CO2 level, window change). Before calculating OAI, it is necessary to find the weight of each variable contributing to energy consumption. Also, the values must be normalized before using in OAI formulation.

The hourly average of NNEC and OAI are utilized to calculate the target residential energy intensity indicator. In the next step (energy wastage identification step), target REII is employed to estimate the potential energy wastage/saving.

[1] Residential Energy Intensity Indicator
[2] Non-steady Normalized Energy Consumption
[3] Occupant Activity Indicator

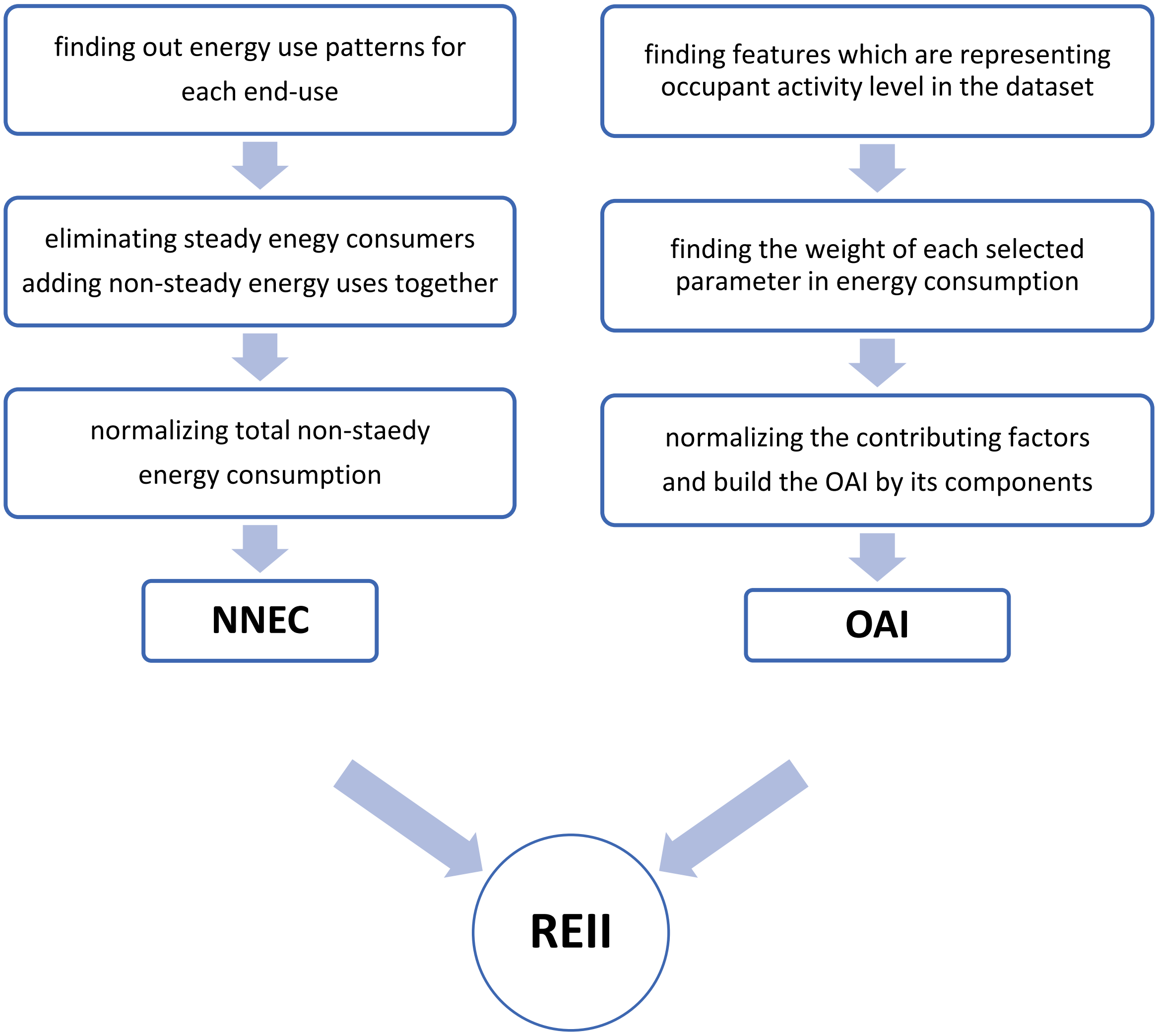

finding out energy use patterns for each end-use
eliminating steady enegy consumers adding non-steady energy uses together
normalizing total non-staedy energy consumption
NNEC
finding features which are representing occupant activity level in the dataset
finding the weight of each selected parameter in energy consumption
normalizing the contributing factors and build the OAI by its components
OAI
REII

Figure 14. The Process to calculate NNEC

3.8. Energy wastage identification step

The actual residential energy intensity indicator (actual REII) for every ten-minute is calculated by the same process, as illustrated in Figure 14. Then, a comparison is made between the target REII and the actual one.

Equation 6 shows the formulation which is used to evaluate occupants' energy-related behavior in this study.

$$EBI = \sum(REII_{actual} - REII_{target})$$ *Equation 6.*

where:

EBI is Energy-related Behavior Index.

When EBI > 0 (actual REII > target REII) it is assumed that there is possible energy-wastage behavior by occupants. Contrarily, if EBI < 0 (actual REII < target REII) it indicates efficient energy-related behavior of occupants. The procedure is shown in figure 15.

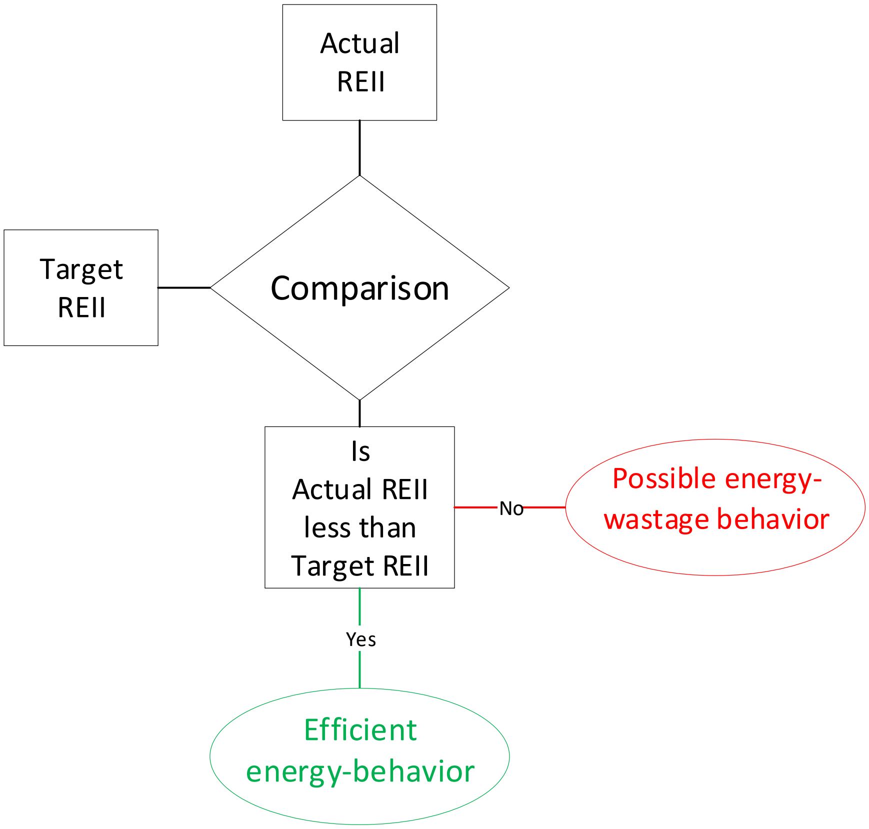

Actual REII
Target REII
Comparison
Is Actual REII less than Target REII
No
Possible energy-wastage behavior
Yes
Efficient energy-behavior

Figure 15. Energy wastage identification step

3.9. Case study

This research uses the data set collected from a high-performance building in Lyon, France. The summers are warm, and the winters are very cold in Lyon. Also, the sky is partly cloudy year-round. The temperature typically ranges from 1 °C to 28 °C over the course of the year and is rarely below -5 °C or above 34 °C [59].

Home energy management system (HEMS) is available for all apartments in this building, monitoring the indoor environment and energy efficiency, and data on both occupancy motion, plug power consumption and lighting power usage for every one-minute data row. A one-year data of a three-bedroom apartment is used in this study. Although individual power sensors have been installed in this apartment, because of the privacy issues, the information regarding their connected appliances and the place that the sensors are located is unknown.

Table 6 shows the available parameters in the dataset. Also, the sensors' information located in each part of the apartment is presented in appendix B.

Table 6. all attributes available in the dataset

	Parameter name	Type	Range
Time and Date	Hour	integer	[0 – 23]
	Day	integer	[1 – 31]
	Day of Week	Categorical	[Mon - Sun]
	Weekday / weekend	Categorical	[Weekday - Weekend]
	Month	integer	[1 – 12]
	index 10 min	integer	[1 – 144]
	Time	date_time	[Jan 1, 12:00:00 AM EST – Dec 31, 11:50:00 PM EST]
Indoor environmen t	4 CO_2 sensors	integer	[249 – 3205]
	4 Temperature sensors	integer	[12 – 28]
	4 Relative Humidity sensors	integer	[27 – 81]
	14 LUX sensors	integer	[0 – 987]
Occupant behavior	14 Motion sensors	Binary	[0 – 1]
	6 Thermostat setpoints	integer	[0 – 50]
	10 window-blinds	integer	[0 – 100]
	10 window-shades	integer	[0 – 100]
	7 Window Open/Close	Binary	[0 – 1]
	14 Light on/off	Binary	[0 – 1]
Energy Consum ption	14 Lighting Power sensors	integer	[0 – 33]
	17 Plug Power sensors	integer	[0 – 546]

The academic version of RapidMiner Studio[1] is used to mine data in this study. A brief introduction about RapidMiner Studio and its advantages are presented in Appendix E.

[1] https://rapidminer.com/

4. Results and Discussion

This chapter presents the results of applying the proposed methodology to the introduced dataset in the previous chapter (section 3.9). Preprocessing and missing value prediction is presented in Section 4.1. In section 4.2, zoning step is described. Sections 4.3 to 4.6 represent the implementation of the developed methodology on different zones of the apartment. Finally, a summary of the obtained results from the investigation is presented in section 4.7.

4.1. Preprocessing

i. *Roll-up:* In the initial dataset, the time step is one-minute. Before starting the process of analyzing data, the time step is changed from 1-minute to 10-minute for this study.

ii. *Missing values:* according to the types of missing values discussed in section 3.4.1, in this step, both groups of missing values are estimated. For the sparse missing data, the average of neighboring values is used.

Continuous missing values: for one of the plug power sensors in the kitchen (sensor code: ZTPG001E5E09020469D3) 12673 values are missed. From August 1st to November 1st. This sensor is considered as the label feature to calculate the missing data. First, a subset of data is created by removing the rows (examples) with missing values. Then different machine learning techniques are employed to find the most proper method to predict the class label missing values. For this reason, artificial neural networks (ANN), support vector machine (SVM), decision tree (DT), deep learning (DL), and generalized linear model (GLM) are tested. Cross validation is employed to have a more accurate model for predicting the missing values. RMSE[1] is used to compare the prediction performance of the techniques. Comparing the performance of the techniques results in choosing deep learning as the best model for our case. The calculated RMSE for the results of deep learning is 0.647 +/- 0.158. Figure 16 shows the whole procedure of predicting missing values, and Figure 17 shows this process in RapidMiner Studio. Also, in appendix A, the detailed-information about the performance of deep learning is presented. The results of cross validation for predicting values for sensor ZTPG001E5E09020469D3 (the class label) is shown in Figure 18 (visualization of prediction performance of deep learning).

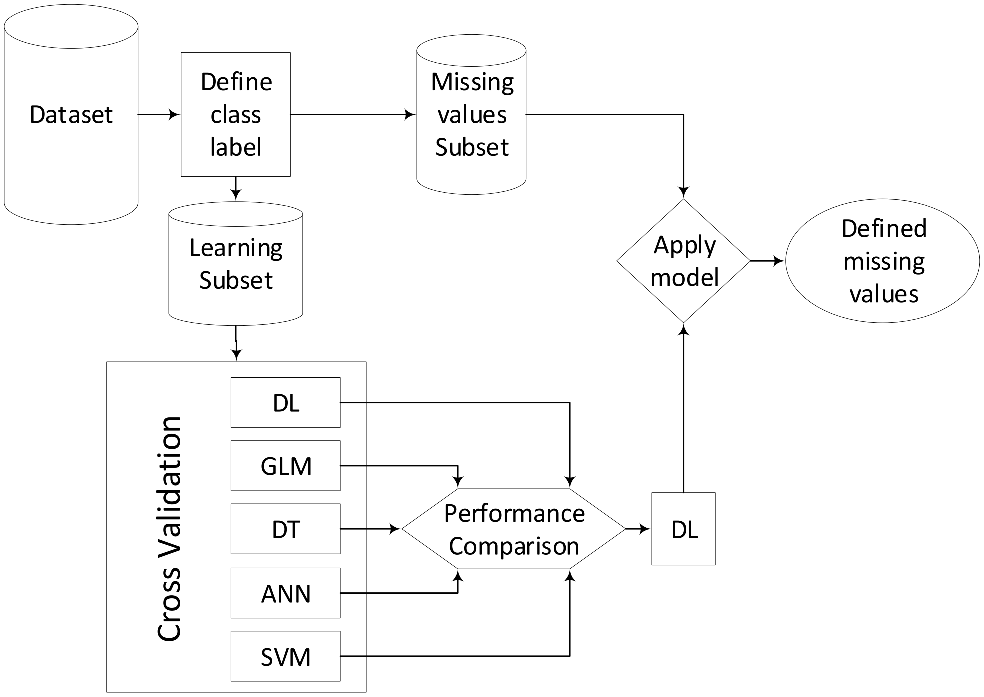
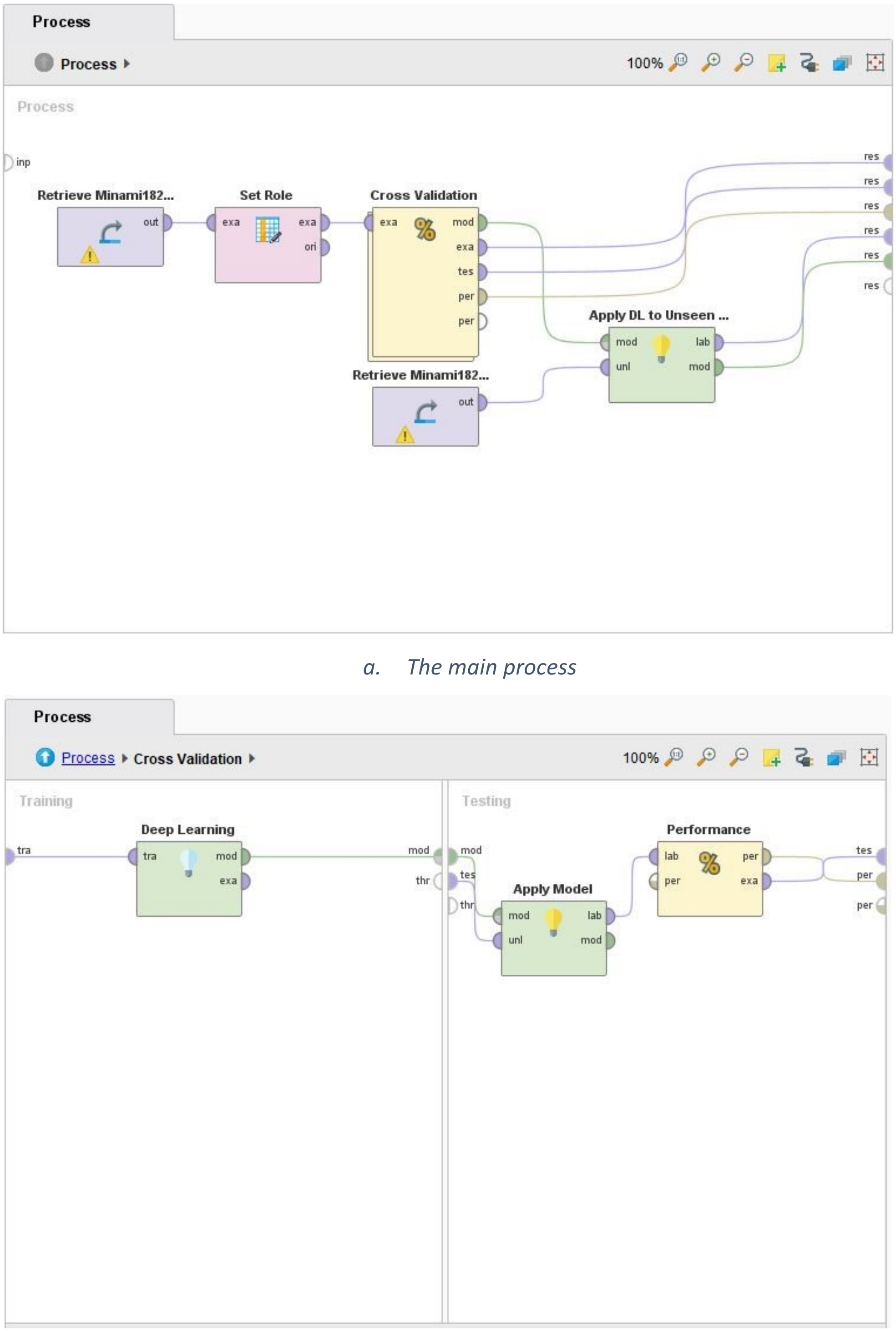


Figure 16. The procedure of predicting missing values

[1] RMSE: root mean square error



a. The main process

b. The sub-process: cross validation

Figure 17. The Procedure to predict missing values in RapidMiner

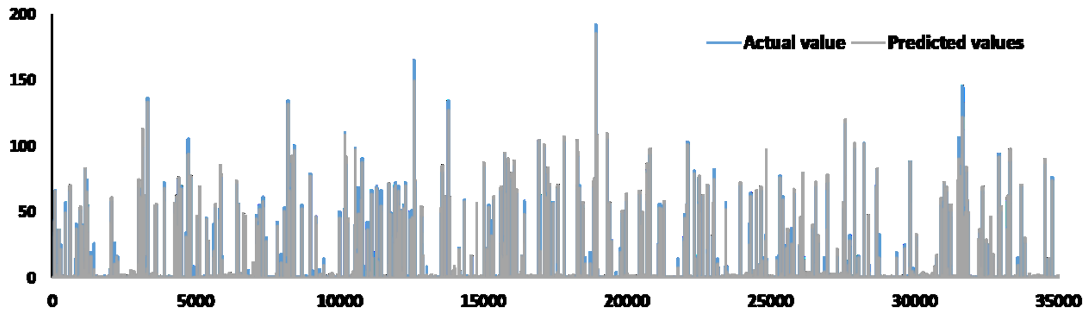


Figure 18. Visualization of prediction performance of deep learning

iii. *Attribute selection:* In this study, among all available parameters which are shown in Table 6, some features are selected to use in the investigation. In this dataset, we have 125 columns (attributes). To understand the influence of each attribute on the lighting power consumption and plug power use, the attributes are analyzed separately. In addition, 20 new attributes were generated in the dataset:

- 10 Blind change: to see if the number of changing the window-blinds influence energy consumption.
- 10 Shade change: to see if the number of changing the window-shades influence energy consumption.

Then, based on the goal of the task, among all 145 attributes, 55 parameters are chosen to use for this study. These attributes are shown in Table 7.

Table 7. the selected features

Parameter name	Type	Range
Hour	integer	[0 – 23]
Day	integer	[1 – 31]
Day of Week	Categorical	[Mon - Sun]
Weekday / weekend	Categorical	[Weekday - Weekend]
Month	integer	[1 – 12]
index 10 min	integer	[1 – 144]
4 CO_2 sensors	integer	[249 – 3205]
14 Motion sensors	Binary	[0 – 1]
14 Lighting Power sensors	integer	[0 – 33]
17 Plug Power sensors	integer	[0 – 546]

4.2. Zones of the apartment

To be able to inform the occupants about the location for potential energy wastage/saving, before the energy analysis step, the apartment is separated into four zones. Zoning the apartment is done based on its plan. Figure 19 shows the plan of the case-study apartment of this research. Zones are illustrated in different colors. Also, in Appendix B more details about the installed sensors are provided.

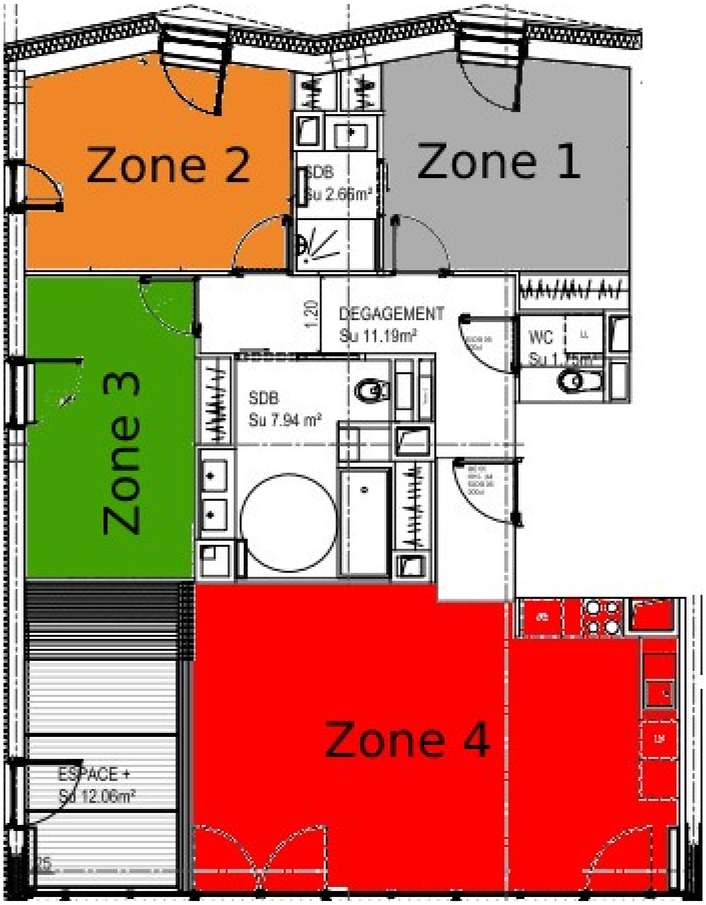


Figure 19. The plan of the case-study apartment.

4.3. Zone one (bedroom 1)

As it is mentioned before, the first zone we are going to work on is bedroom one. In this zone, we have data for 14 sensors which are introduced in Appendix B. Six sensors have been chosen to use in this study, which are CO_2, motion, lighting power, plug power 1, plug power 2, and plug power 3.

4.3.1. Energy use patterns investigation - Zone one (bedroom 1)

The following images illustrate the variation of energy consumption and its drivers for zone one of the apartment. The data is normalized and aggregated hourly. Also, to understand the effects of time on the trends, the whole year data is separated into weekdays/weekends, months, and day of the week. Also, the RapidMiner procedure for this investigation is provided in appendix C.

Figure 20 shows how the hourly plug power consumption is following occupants' motion in zone one. Figure 21 illustrates how hourly lighting power consumption follows occupants' motion.

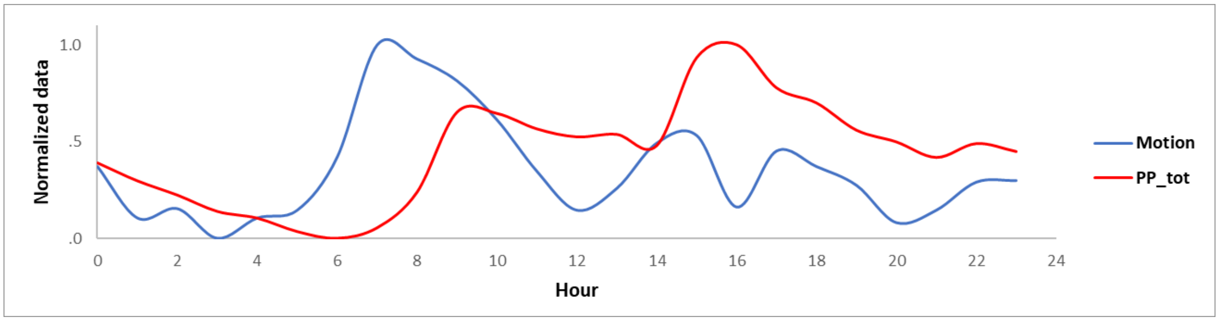


Figure 20. Aggregated-hourly variation of total plug power consumption and total motion detected in zone 1

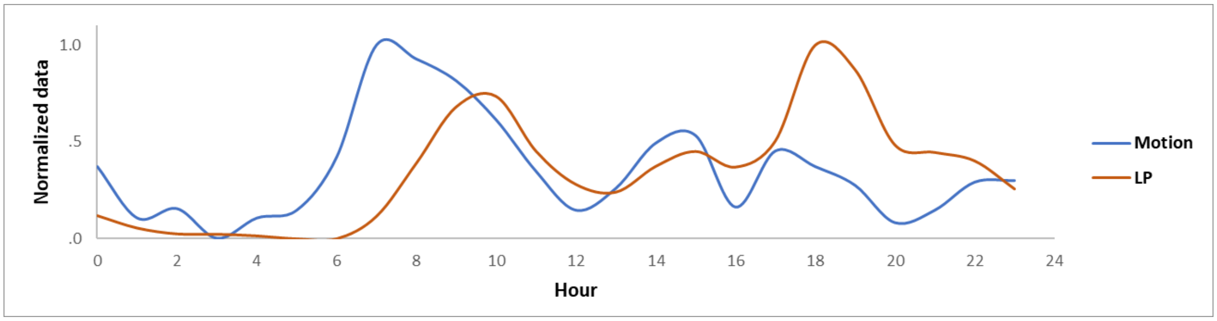


Figure 21. Aggregated-hourly variation of total lighting power consumption and total motion detected in zone 1

Figures 22 to 25 show the hourly variations for different days of the week (Sunday to Saturday). Figure 22 shows the aggregated hourly occupant's motion detected by the motion detector in zone one.

As it is shown in Figure 23, which is illustrating the hourly variation of normalized CO_2 concentration in zone one, the carbon dioxide amount in this bedroom is significantly higher during nights comparing to day times. One reason is that the window was mostly kept closed during night times. Because of this inconstancy, we cannot consider CO₂ as an indicator of occupants' activity-level in this zone. Therefore, for bedroom one, equation 3 (section 3.1) is simplified to equation 7:

$$\mathrm{OAI}_{\mathrm{zone}_1} = \sqrt{(0 \times NCG^2) + (1 \times \mathrm{RoM}^2)} = \mathrm{RoM}_{zone1} \qquad \textit{Equation 7}$$

where RoM is the rate of motion detected by motion detector in bedroom one (normalized-aggregated data).

Figures 24 and 25 show the daily variation of hourly energy consumption in zone one.

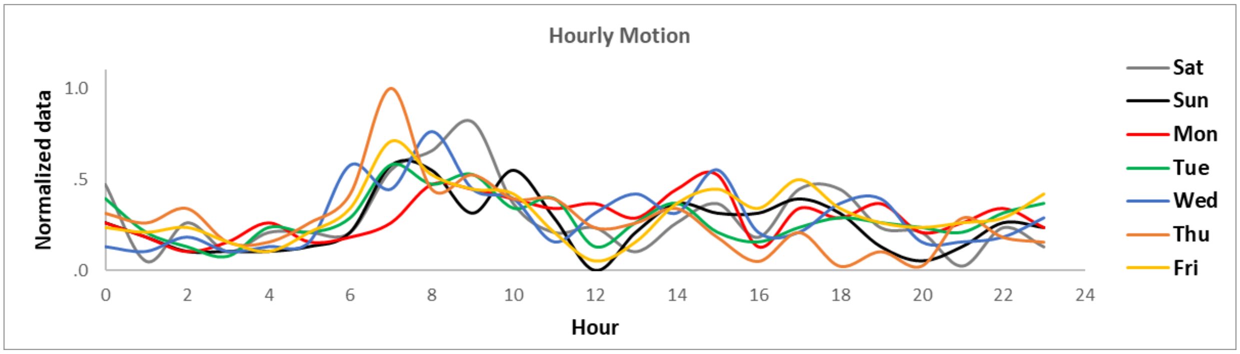


Figure 22. Aggregated-hourly variation of total motion detected in zone 1 per weekday

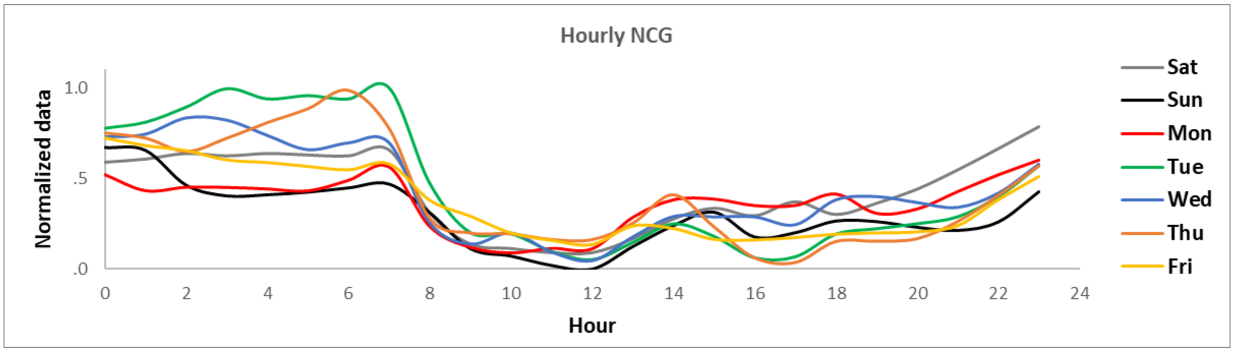


Figure 23. Aggregated-hourly variation of total CO₂ concentration in zone 1 per weekday

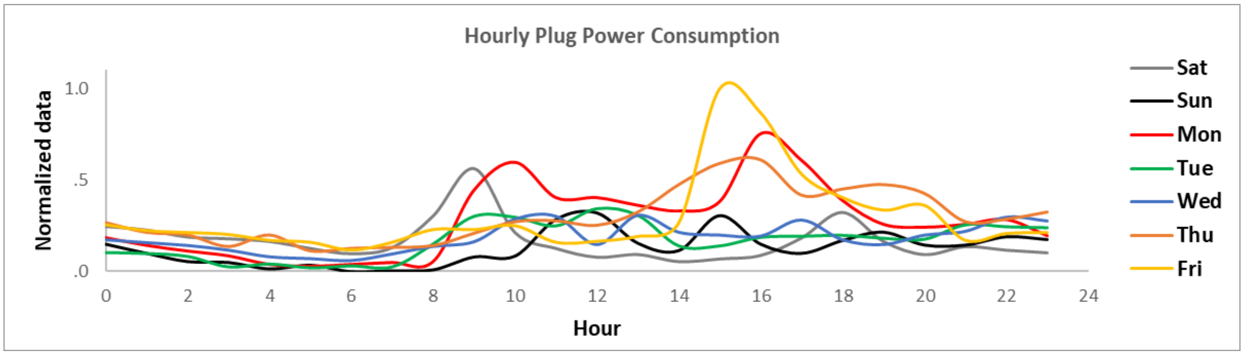


Figure 24. Aggregated-hourly variation of total plug power consumption in zone 1 per weekday

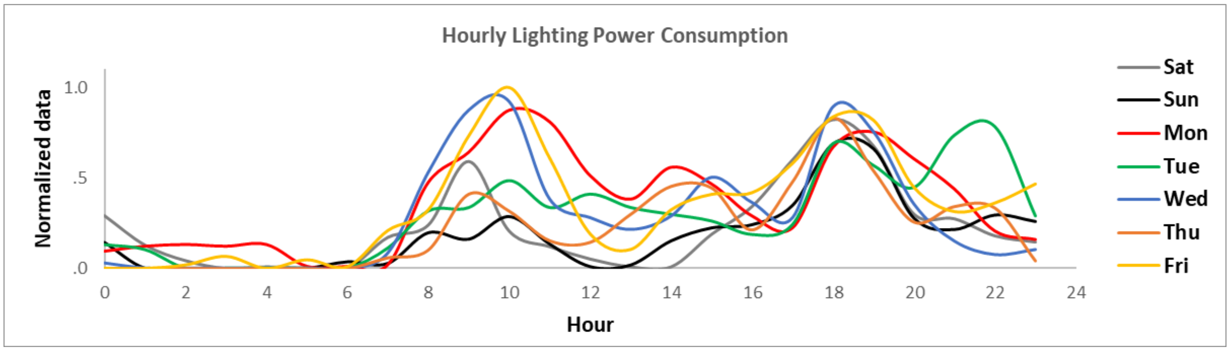


Figure 25. Aggregated-hourly variation of total lighting power consumption in zone 1 per weekday

Figures 26 to 29 show the hourly variations for all months (January to December). Figure 26 shows the aggregated hourly occupant's motion detected by the motion detector in zone one.

Similar to Figure 23, as it is shown in Figure 27, the carbon dioxide amount in bedroom one is higher during nights comparing to day times. Figures 28 and 29 show the monthly variation of hourly energy consumption in zone one. Referring to the monthly aggregated data, it seems that this bedroom was unoccupied in most of the days in July and August.

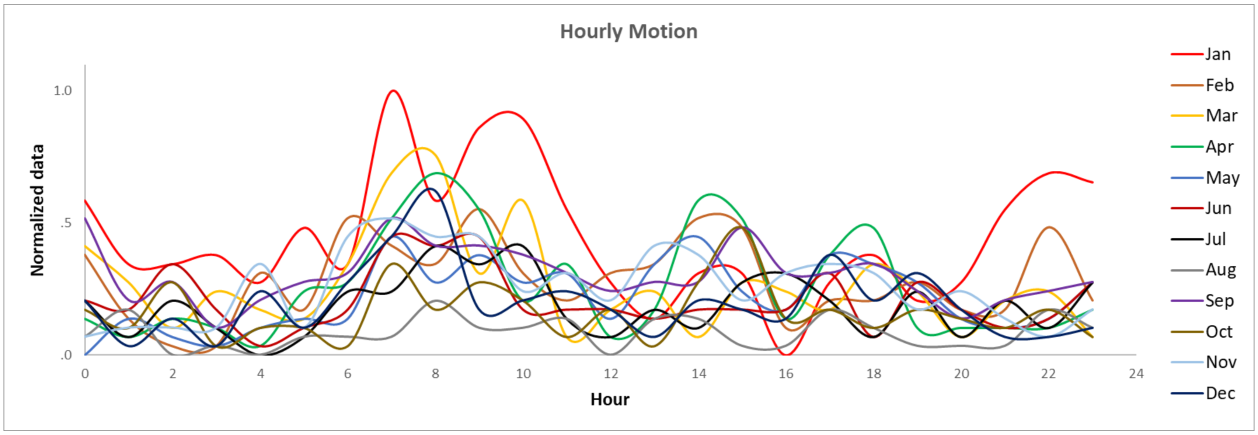


Figure 26. Aggregated-hourly variation of total motion detected in zone 1 per month

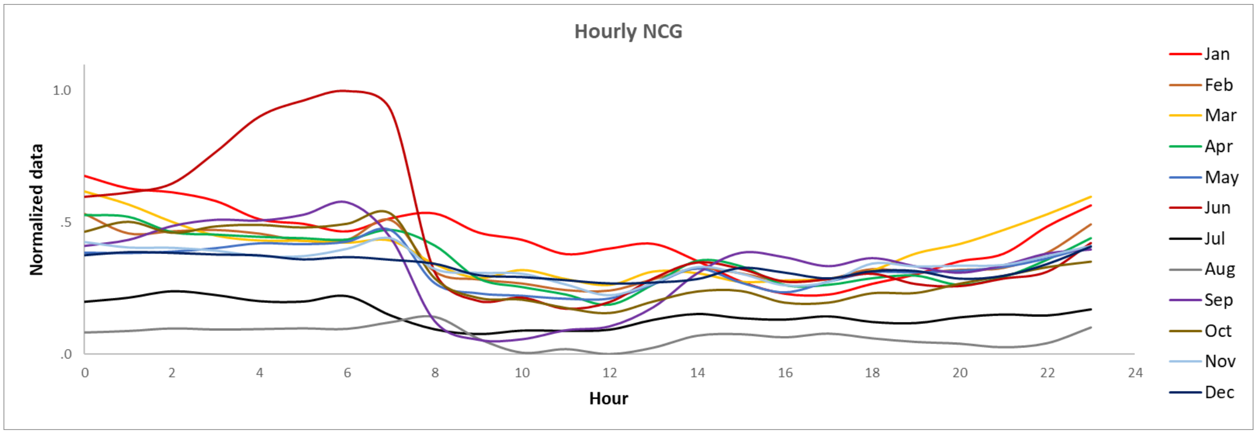


Figure 27. Aggregated-hourly variation of total CO_2 concentration in zone 1 per month

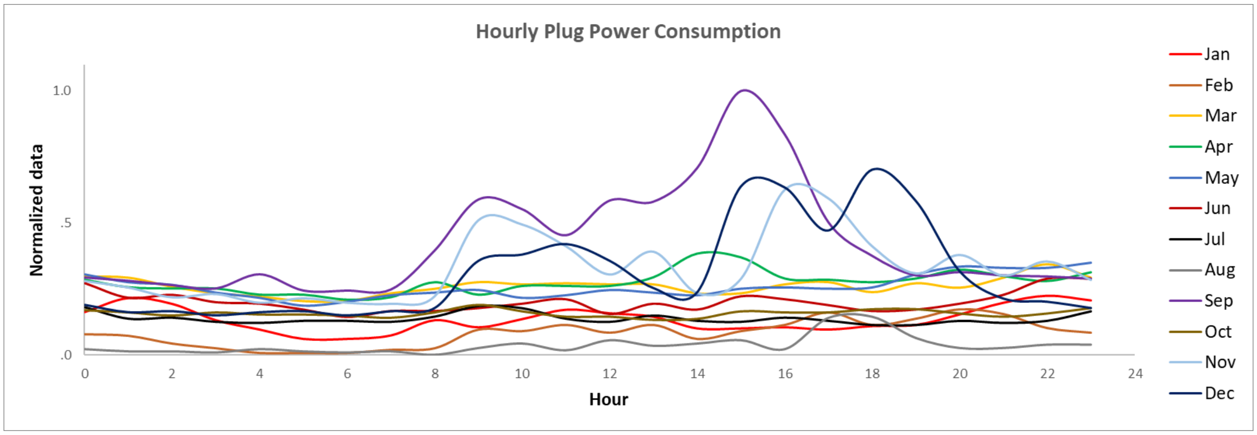


Figure 28. Aggregated-hourly variation of total plug power consumption in zone 1 per month

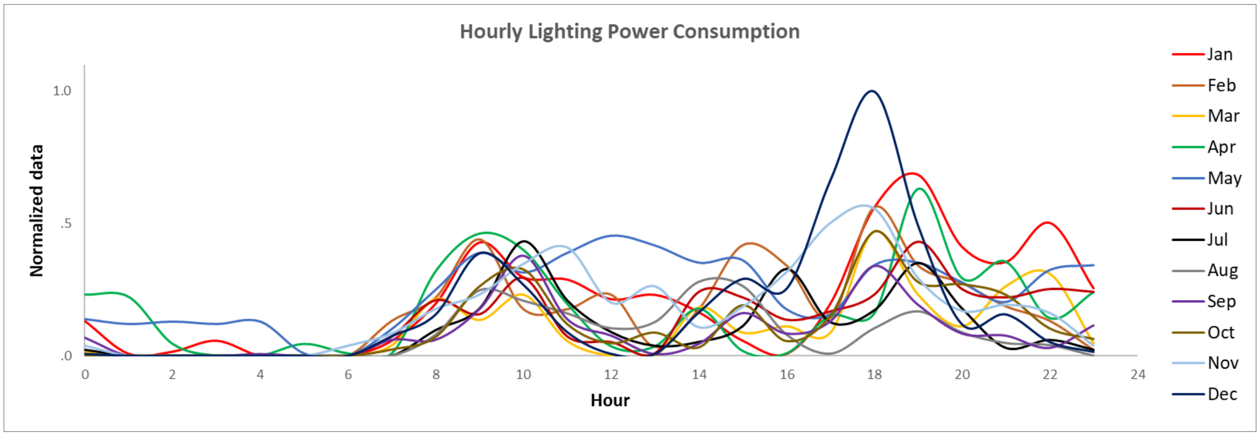


Figure 29. Aggregated-hourly variation of total lighting power consumption in zone 1 per month

Figures 30 to 33 show the hourly variations for weekdays/weekends. Figure 30 shows the aggregated hourly occupant's motion detected by the motion detector in zone one. Similar to Figures 23 and 27, as it is shown in Figure 31, the carbon dioxide amount in the bedroom is significantly higher during nights comparing to day times. Figures 32 and 33 show the variation of hourly energy consumption in zone one.

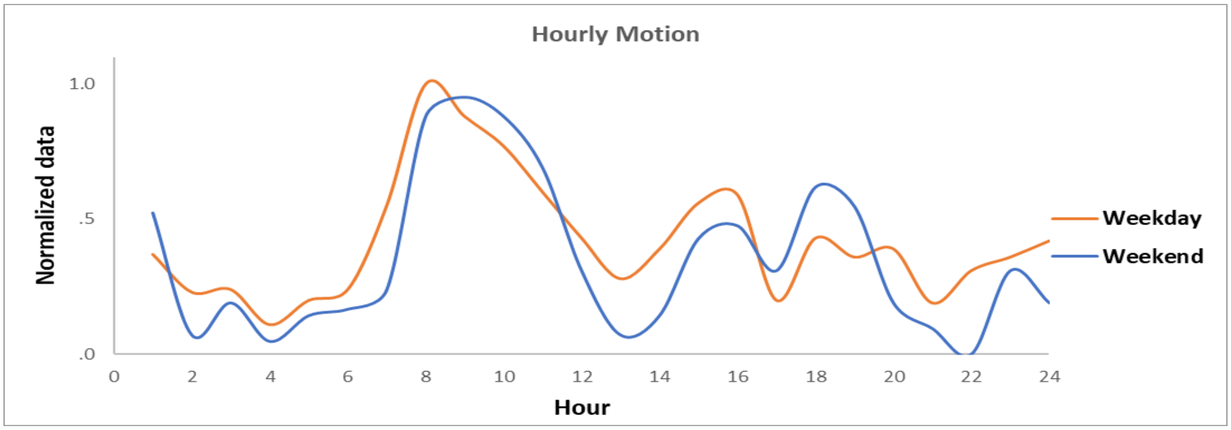


Figure 30. Aggregated-hourly variation of total occupants' motion detected in zone 1, separated for weekdays/weekends

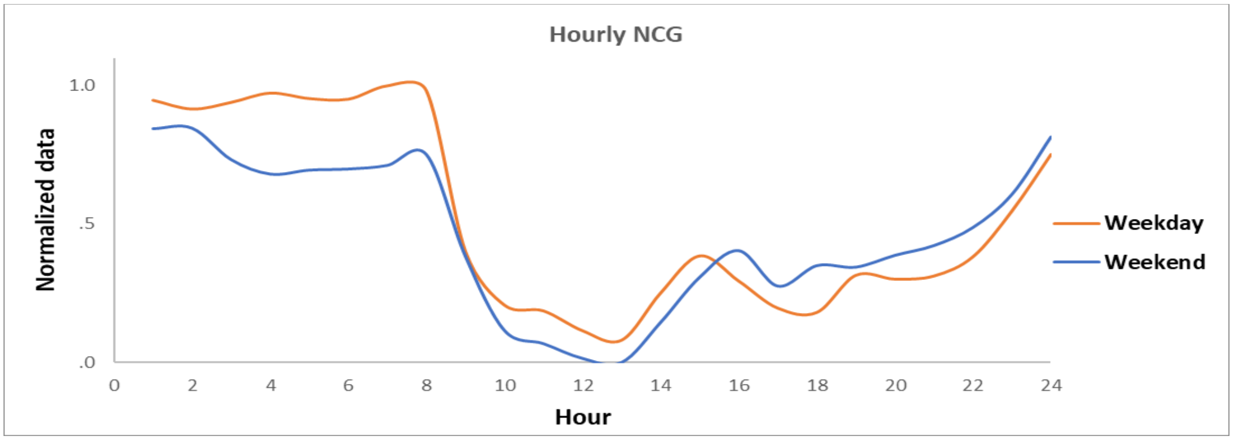


Figure 31. Aggregated-hourly variation of total CO_2 concentration in zone 1, separated for weekdays/weekends

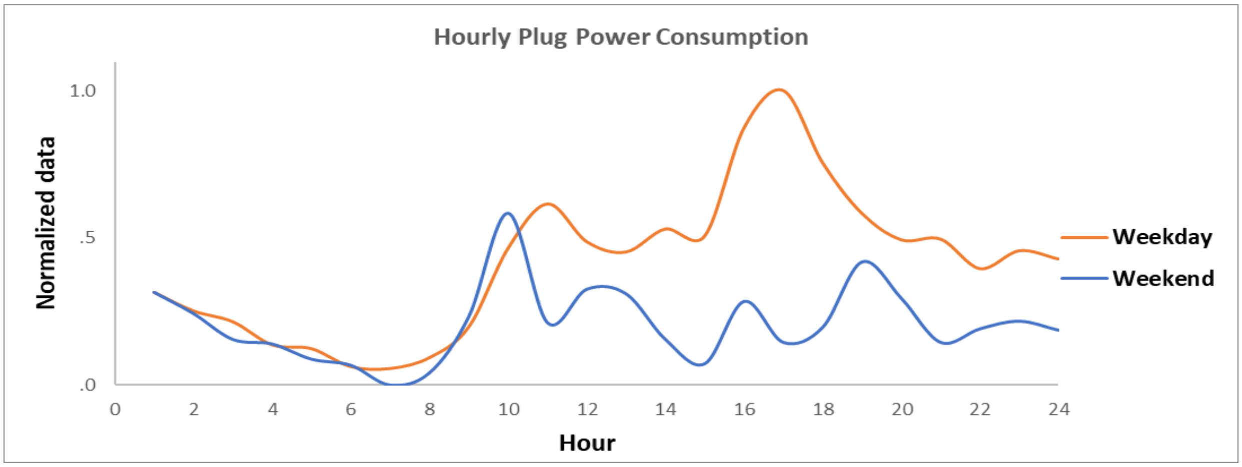


Figure 32. Aggregated-hourly variation of total plug power consumption in zone 1, separated for weekdays/weekends

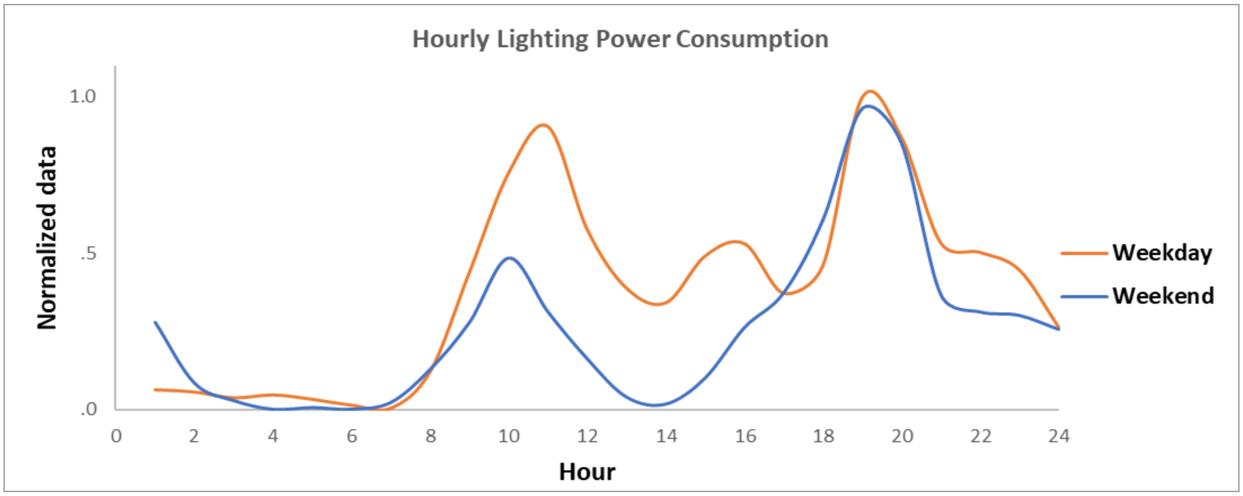


Figure 33. Aggregated-hourly variation of total lighting power consumption in zone 1, separated for weekdays/weekends

4.3.2. Clustering step - Zone one (bedroom 1)

The clustering step includes four sub-steps, which are illustrated in Figure 34.

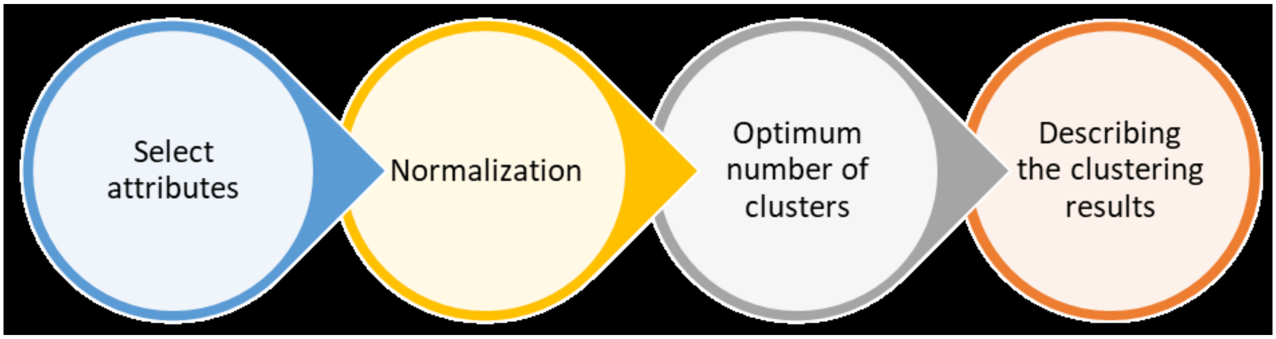


Figure 34. Clustering sup-steps

i) Feature selection:

Based on the goal of the task, the contributing attributes for performing clustering must be chosen. Accordingly, aggregated-monthly data for five parameters, including RoM[1], LP[2], PP_1, PP_2, and PP_3[3] are selected.

[1] Rate of motion

[2] Lighting power consumption

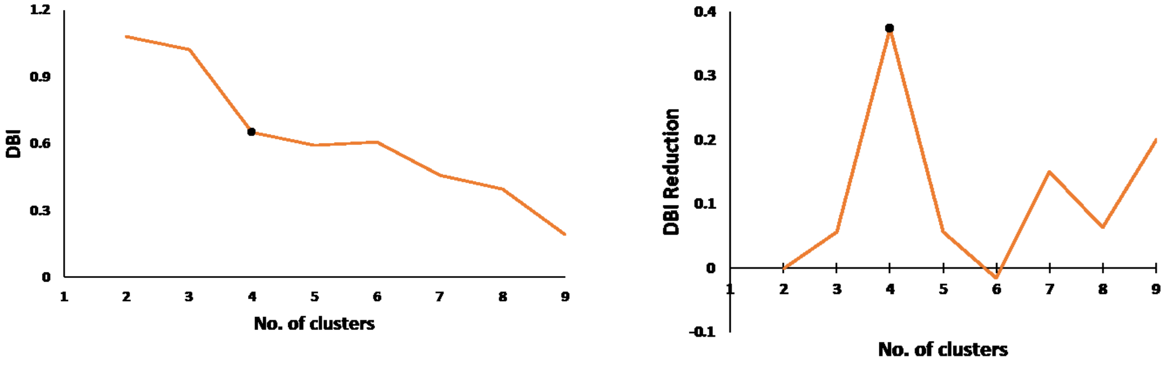
[3] Plug power consumption, no. 1, 2, and 3

ii) Normalization:

Next, the data must be normalized to prevent the features with large ranges from outweighing those with comparatively smaller ranges. Min-max normalization is used to normalize the data.

iii) Finding k_{opt}:

We need to find the optimum number of clusters (k). For this reason, clustering is performed for k=2 to k=9, and DBI[1] calculated for each k. Elbow method is employed to choose the optimum number of clusters (k_{opt}). As it is illustrated in figure 35.a, $k_{opt} = 4$ for zone one. Also, Figure 35.b shows that the amount of DBI reduction is the highest when the number of clusters is 4.



a. *b.*

Figure 35. Performance of clustering, a. DBI for different number of clusters, b. reduction in DBI by increasing the number of clusters

iv) Implementation:

After finding the optimum number for clusters, the defined clusters must be analyzed and interpreted. The visualization of the clusters' centroids can be seen in Figure 36.

[1] Davies-Bouldin index

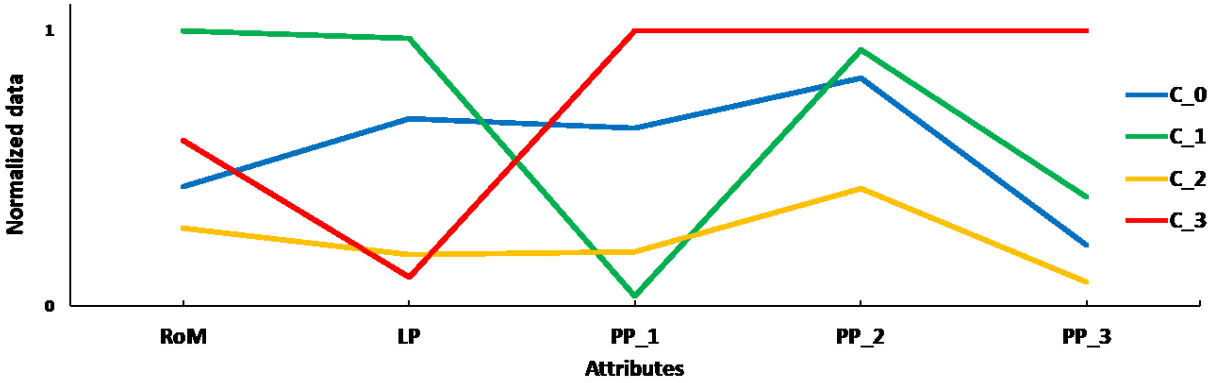


Figure 36. Visualization of cluster centroids

- **Clusters' specifications:**

Cluster_0: This cluster represents the moments that the level of activity by occupants in zone one is comparable low. Lighting power consumption and usage of plug powers 1 and 2 are above average but plug power 3 usage is much less.

Cluster_1: This cluster represents the moments that the level of activity and lighting power consumption by occupants in zone one are the highest. Plug power 1 consumption is negligible, plug power 2 usage is very high, and plug power 3 is around average.

Cluster_2: This cluster represents the moments that the level of activity by occupants in zone one is the lowest. Lighting power consumption is much less than average and usage of plug powers 2 and 3 are the lowest.

Cluster_3: This cluster represents the moments that the level of activity by occupants in zone one is around average. Lighting power consumption is the lowest, but plug power usage is the highest.

The RapidMiner procedure for performing the clustering step is provided in appendix D.

4.3.3. Baseline step - Zone one (bedroom 1)

Here we are going to calculate target REII[1] for zone one. The process is based on the procedure which was described in Figure 14 (section 3.7):

i) Calculation of *NNEC*[2]:

Following the presented introduction about NNEC in section 3.2, all the four energy consumers in zone one are considered as non-steady end-use. Therefore, a new attribute is generated by summing up the lighting power consumption and all plug power consumptions (e.g. LP, PP_1, PP_2, and PP_3). This new attribute is called Non-steady-Energy-Consumption (*NEC*):

$$\mathrm{NEC} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{LP_i} + \sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{PP_j} \qquad \textit{Equation 8}$$

where:

- *NEC* is non-steady energy consumption (before normalization)
- LP_i is i^{th} lighting power consumption
- PP_j is j^{th} plug power consumption
- *n* and *m* are the number of non-steady lighting powers and plug powers, respectively.

Therefore, for zone one equation 8 become:

$$\mathrm{NEC_{zone_1}} = \mathrm{LP} + \mathrm{PP_1} + \mathrm{PP_2} + \mathrm{PP_3} \qquad \textit{Equation 9}$$

$$\mathrm{NNEC_{zone_1}} = \mathrm{normalized}\ (\mathrm{NEC_{zone_1}}) \qquad \textit{Equation 10}$$

ii) Calculation of *OAI*[3]:

[1] REII: Residential Energy Intensity Indicator

[2] Normalized-Non-Steady Energy Consumption

[3] Occupant Activity Indicator

According the equation 7 (section 4.3.1), for zone one *OAI = RoM.*

iii) Calculation of *REII*:

The *target REII* attribute for zone one for each cluster is generated by equation 11:

$$\mathrm{REII}_{\mathrm{zone}_1} = \frac{\mathrm{NNEC}_{\mathrm{zone}_1}}{\mathrm{OAI}_{\mathrm{zone}_1}} \qquad \textit{Equation 11}$$

Then the hourly REII is used as the baseline (target REII) to evaluate occupants' energy behavioral patterns and investigate the amount of potential energy wastage/saving by occupants in zone one.

iv) Target REII for zone one:

Figure 37 shows the baseline (target hourly REII) for the four defined clusters in zone one. This chart is used in the next step to identify the potential energy wastage/saving in zone one.

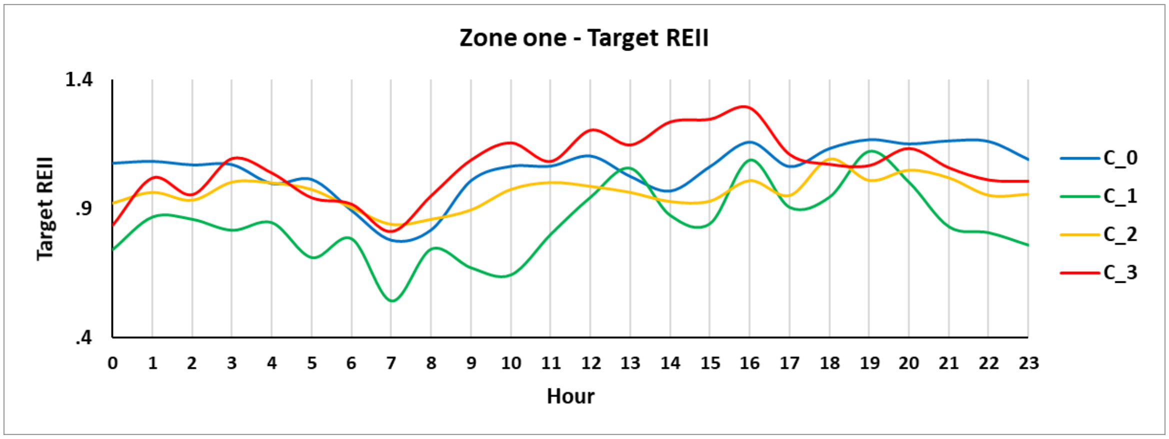
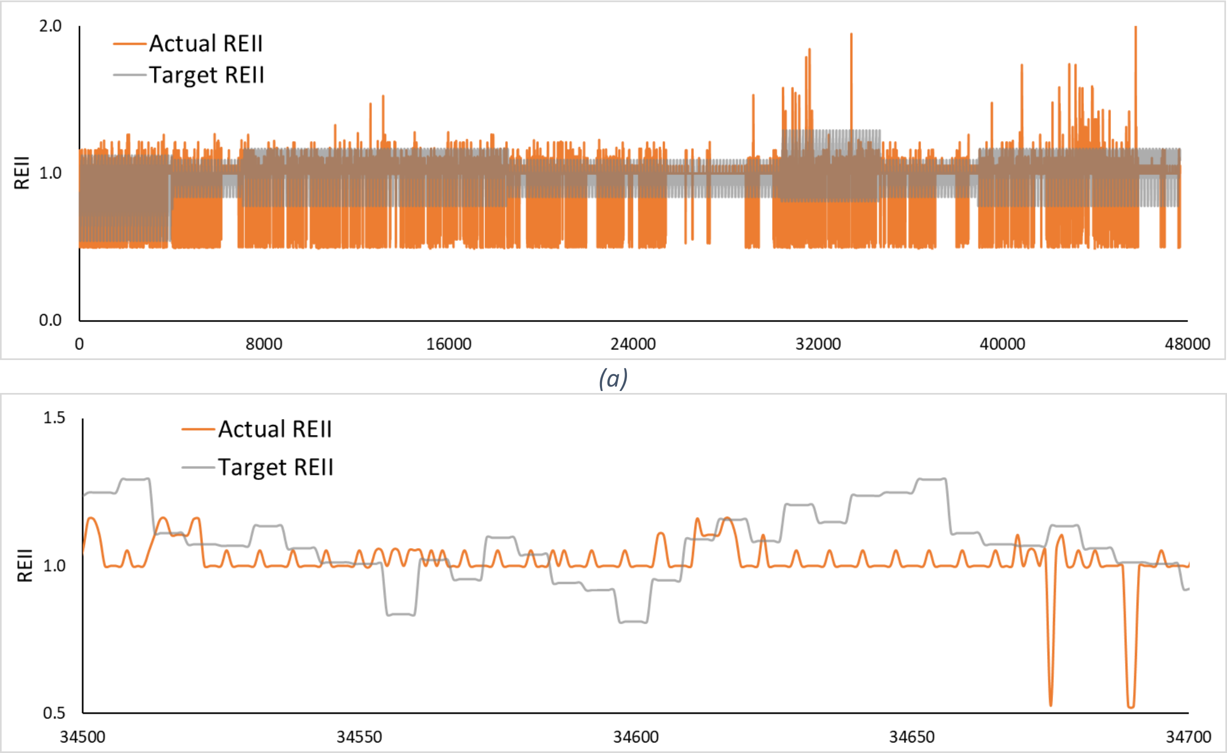


Figure 37. Target hourly REII for the four defined clusters

4.3.4. Energy wastage identification step - Zone one (bedroom 1)

Employing the results of the previous section as the baseline for evaluation of occupants' energy-related behavior in zone one enables us to calculate the potential energy wastage/saving by occupants in this zone. Figure 38.a shows both the actual REII and the target one (baseline). Considering the concept behind REII definition, which is energy consumption

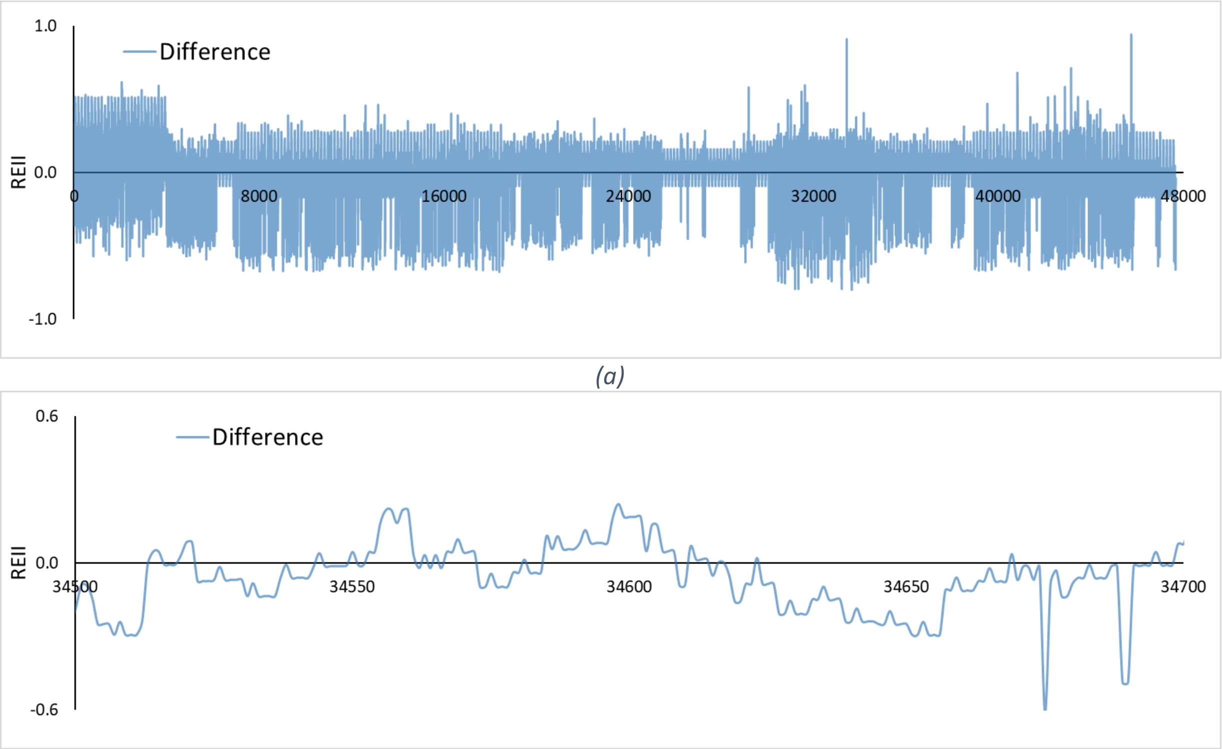
(NNEC) over occupants' activity (OAI), the moments that the actual REII is higher than the baseline can be defined as the possible energy wastage behavior. Similarly, when the actual REII is lower than the target one, it indicates efficient energy-related behavior. Figure 38.b highlights a part of Figure 38.a for better understanding. The potential energy wastage/saving is calculated by the difference between the actual REII and the baseline. Figure 39 indicates this difference.



(a)

(b)

Figure 38. Actual REII and target REII for the whole year. (a) whole year data, (b) zoom a part of the chart for better understanding



(a)

(b)

Figure 39. Difference between the actual REII and target REII for the whole year – potential energy wastage/saving. (a) whole year data, (b) zoom a part of the chart for better understanding

4.4. Zone two (bedroom 2)

The second zone is bedroom two. In this zone, we have data for 15 sensors which are introduced in Appendix B. Four sensors have been chosen to use in this study, which are CO_2, motion, lighting power, and plug power.

4.4.1. Energy use patterns investigation - Zone two (bedroom 2)

The following images illustrate the variation of energy consumption and its drivers for zone two of the apartment. Similar to zone 1, the data is normalized and aggregated hourly and is separated into weekdays/weekends, months, and day of the week. The RapidMiner procedure for this investigation is the same as zone one (appendix C).

Figure 40 shows how the hourly plug power consumption is following occupants' motion in zone two. Figure 41 illustrates how hourly lighting power consumption follows occupants' motion.

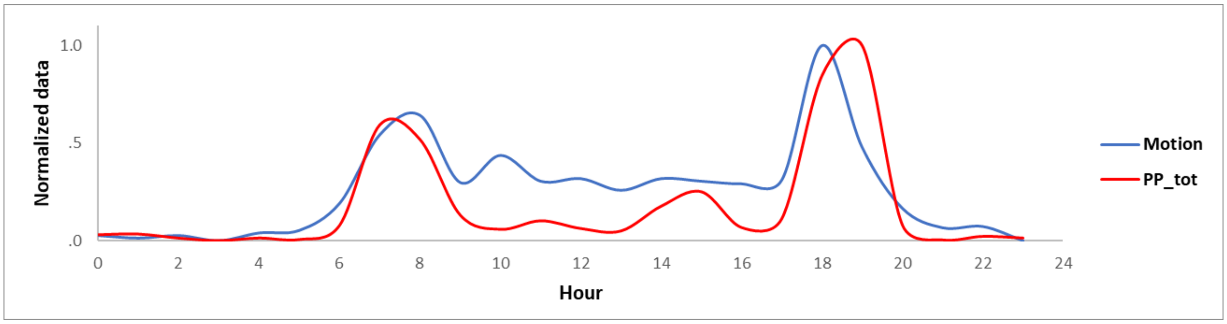


Figure 40. Aggregated-hourly variation of total plug power consumption and total motion detected in zone 2

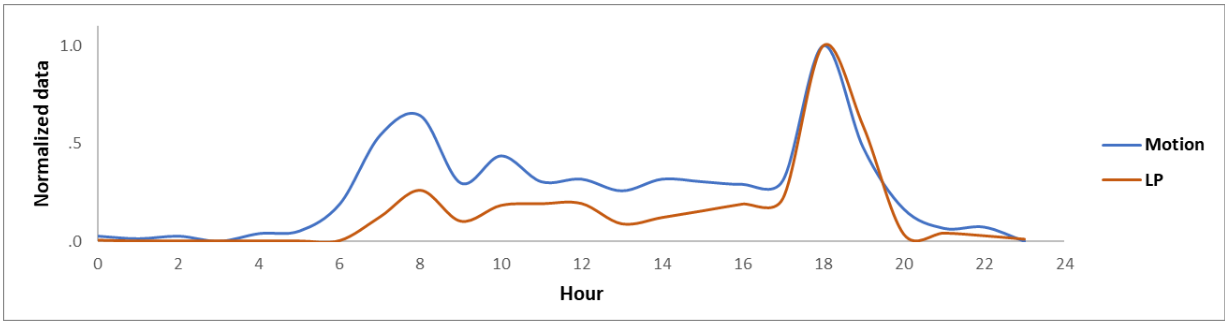


Figure 41. Aggregated-hourly variation of total lighting power consumption and total motion detected in zone 2

Figures 42 to 45 show the hourly variations for different days of the week (Sunday to Saturday). Figure 42 shows the aggregated hourly occupant's motion detected by the motion detector in zone two.

As it is shown in Figure 43, which is illustrating the hourly variation of normalized CO_2 concentration in zone two, again, the carbon dioxide amount in the bedroom is significantly higher during nights comparing to day times. Similar to zone one, because of this inconstancy, we cannot consider CO_2 as an indicator of occupants' activity-level in this zone. Therefore, for bedroom tow, equation 3 (section 3.1) is simplified to equation 12:

$$\mathrm{OAI}_{\mathrm{zone}_2} = \sqrt{(0 \times NCG^2) + (1 \times \mathrm{RoM}^2)} = \mathrm{RoM}_{zone2}$$

Equation 12

where RoM is rate of motion detected by motion detector in bedroom two (normalized-aggregated data.

Figures 44 and 45 show the daily variation of hourly energy consumption in zone two.

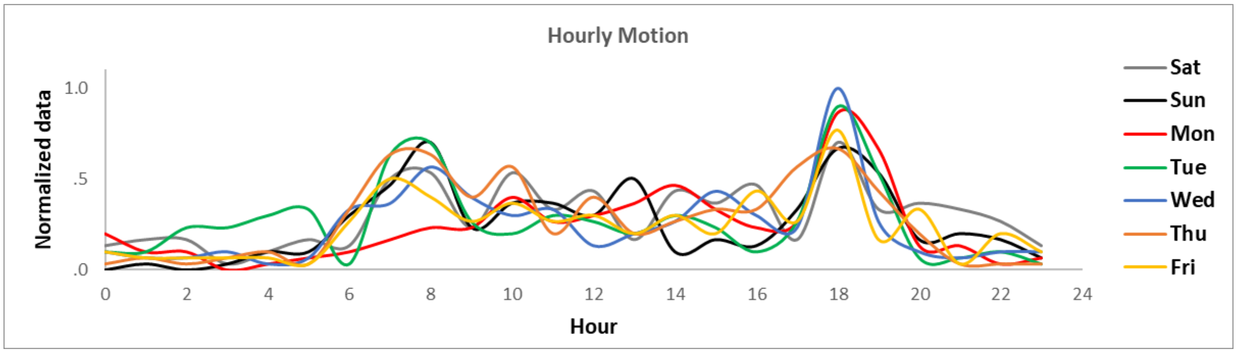


Figure 42. Aggregated-hourly variation of total motion detected in zone two per weekday

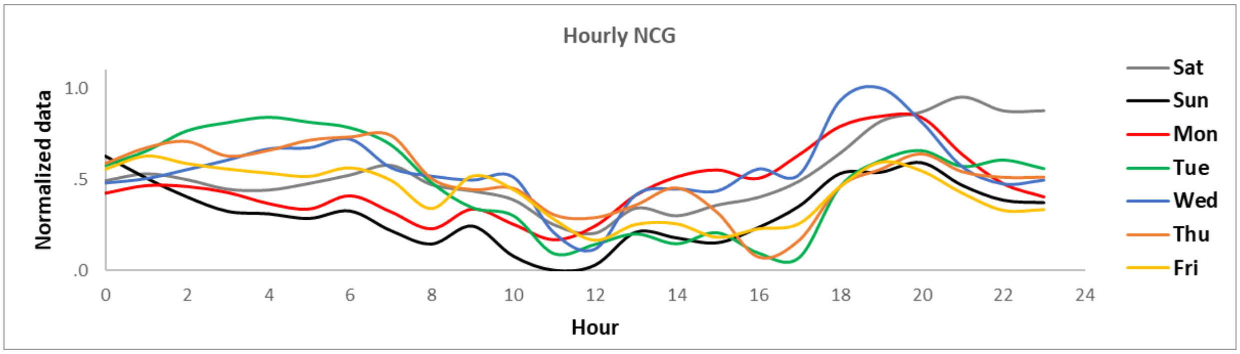


Figure 43. Aggregated-hourly variation of total CO$_2$ concentration in zone two per weekday

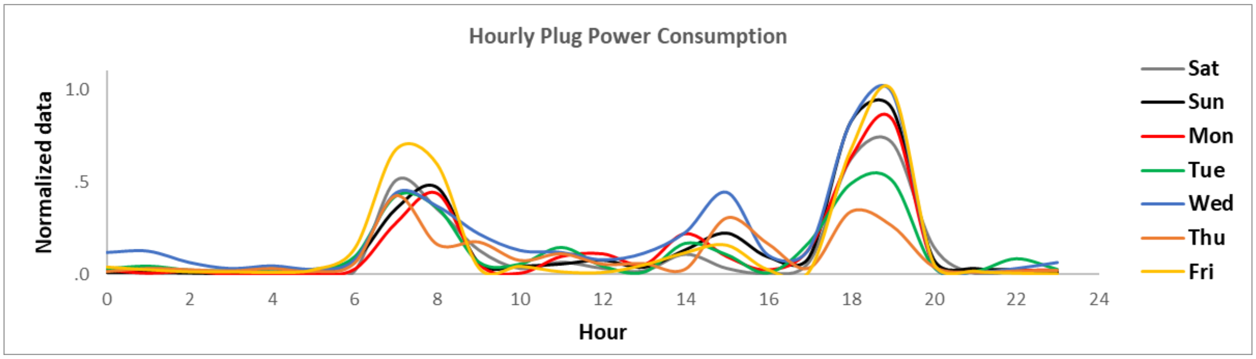


Figure 44. Aggregated-hourly variation of total plug power consumption in zone two per weekday

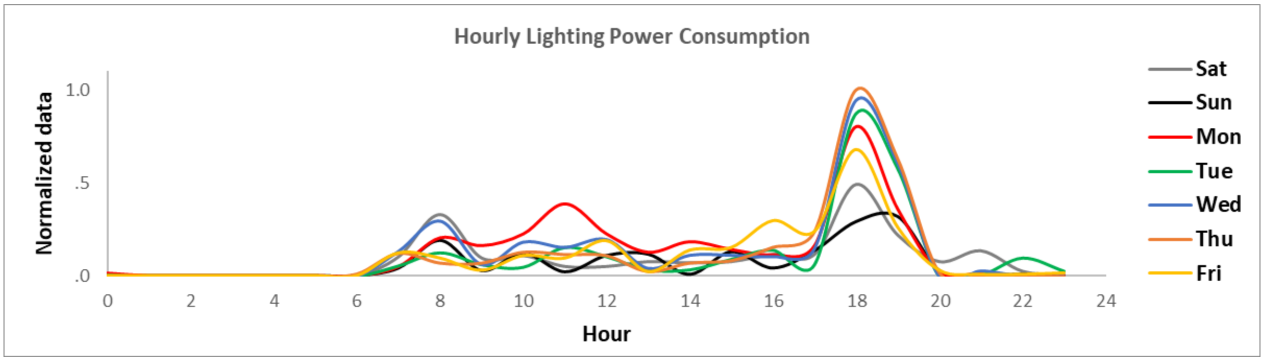


Figure 45. Aggregated-hourly variation of total lighting power consumption in zone two per weekday

Figures 46 to 49 show the hourly variations for all months (January to December). Figure 46 shows the aggregated hourly occupant's motion detected by the motion detector in zone two.

Similar to Figure 43, as it is shown in Figure 47, the carbon dioxide amount in bedroom two is higher during nights comparing to day times. Figures 48 and 49 show the monthly variation of hourly energy consumption in zone two. Referring to the monthly aggregated data, it seems that this bedroom was also unoccupied in most of the days in July and August.

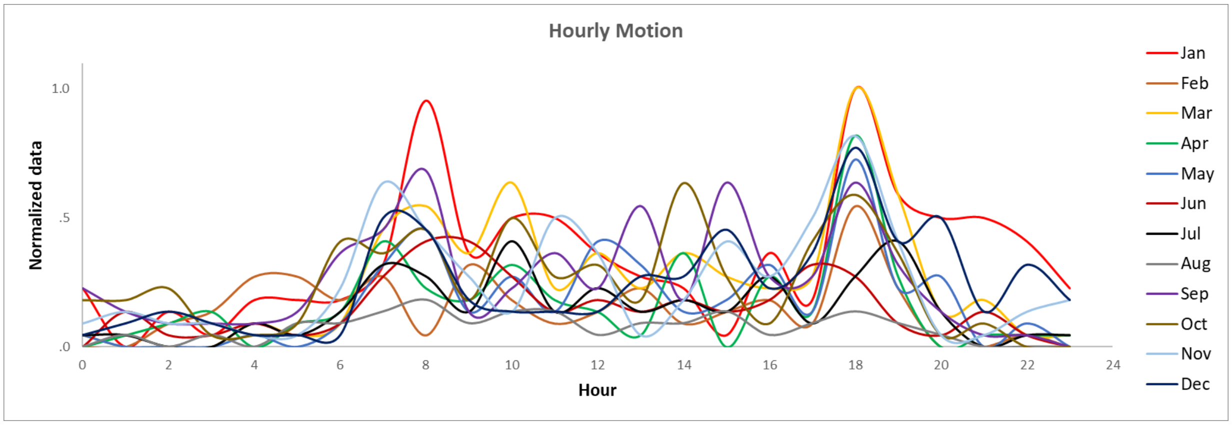


Figure 46. Aggregated-hourly variation of total motion detected in zone two per month

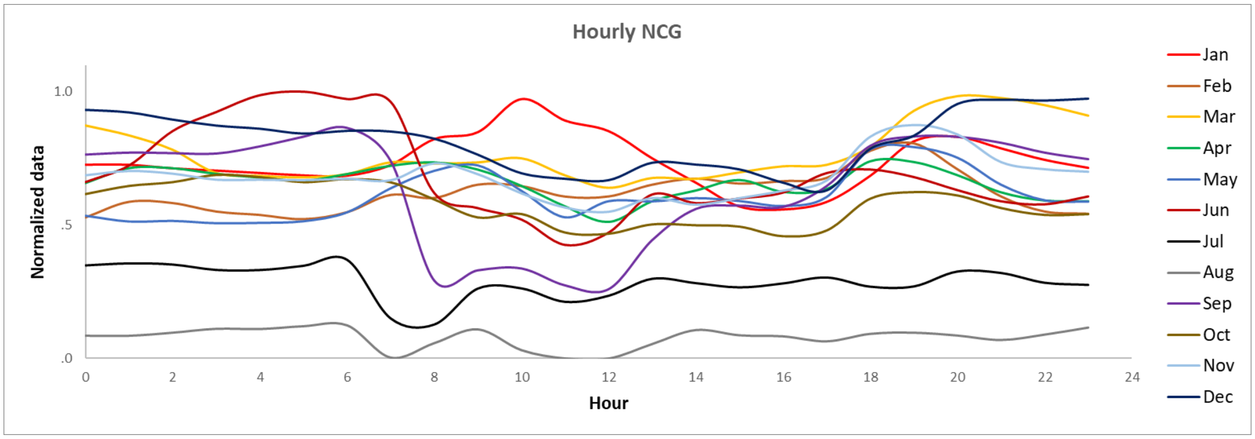


Figure 47. Aggregated-hourly variation of total CO_2 concentration in zone two per month

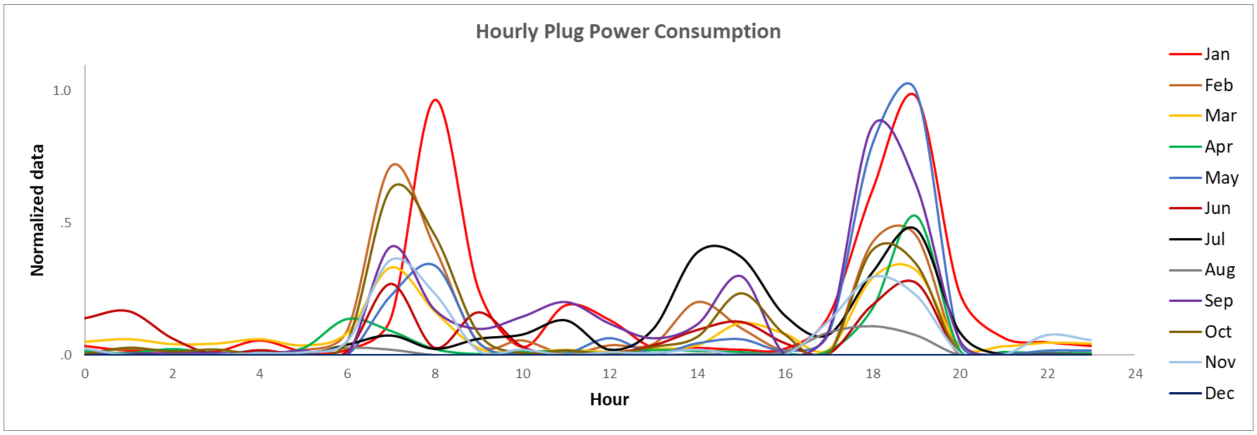


Figure 48. Aggregated-hourly variation of total plug power consumption in zone two per month

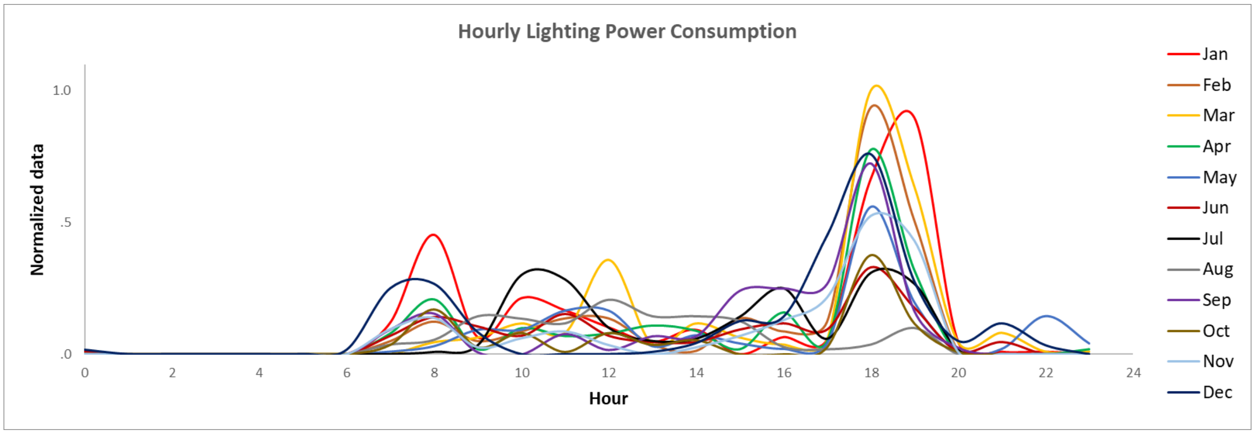


Figure 49. Aggregated-hourly variation of total lighting power consumption in zone two per month

Figures 50 to 53 show the hourly variations for weekdays/weekends. Figure 50 shows the aggregated hourly occupant's motion detected by the motion detector in zone two.

Similar to Figures 43 and 47, as it is shown in figure 51, the carbon dioxide amount in this bedroom is significantly higher during nights comparing to day times. Figures 52 and 53 show the variation of hourly energy consumption in zone two.

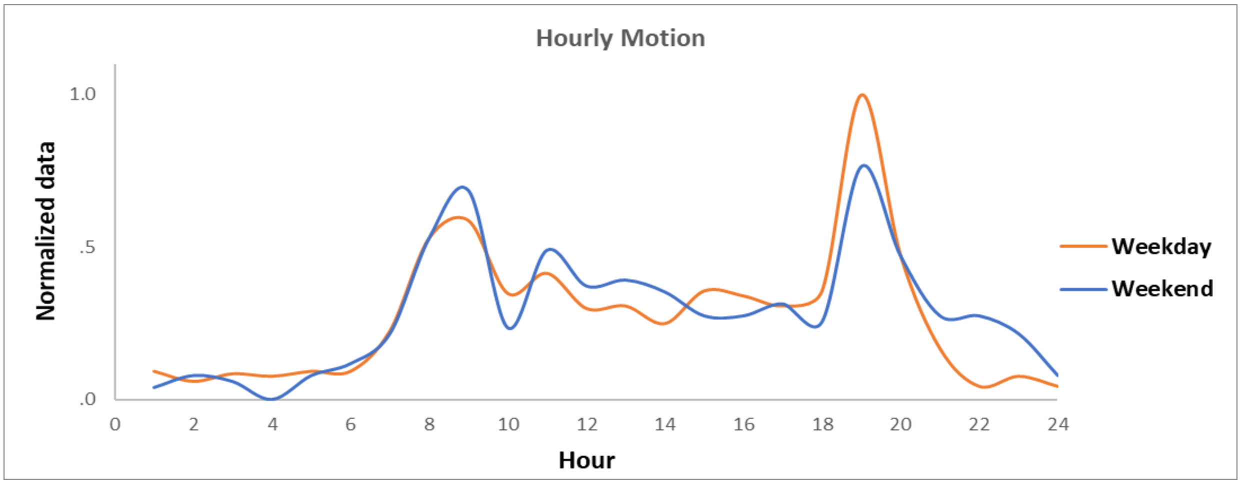


Figure 50. Aggregated-hourly variation of total occupants' motion detected in zone two, separated for weekdays/weekends

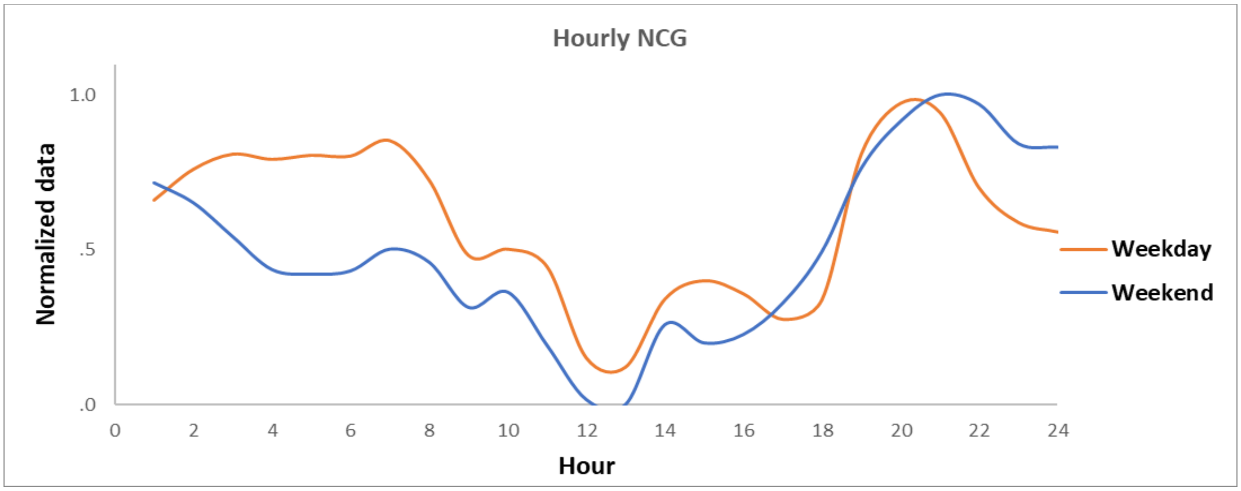


Figure 51. Aggregated-hourly variation of total CO_2 concentration in zone two, separated for weekdays/weekends

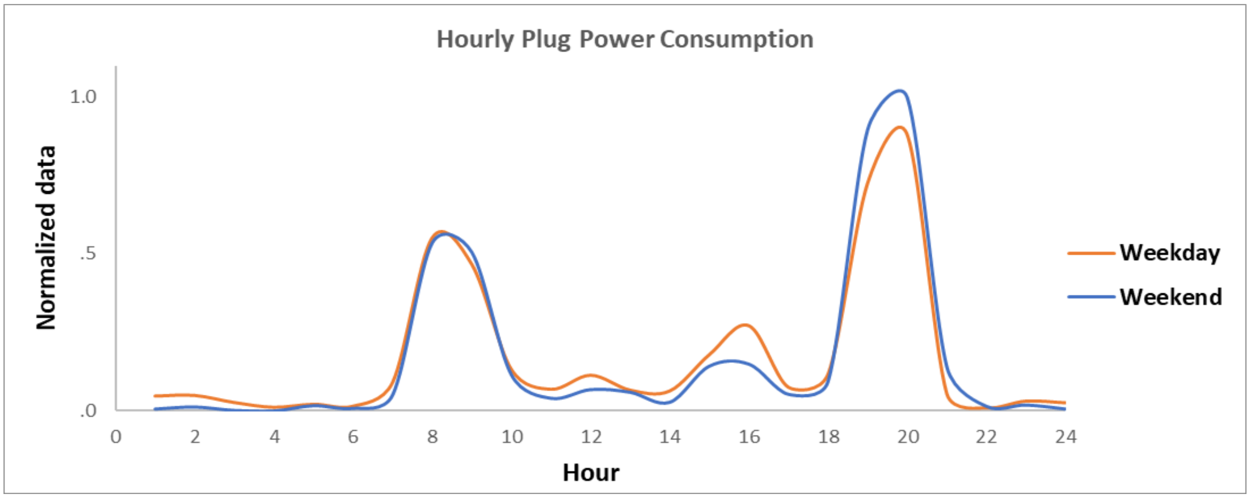


Figure 52. Aggregated-hourly variation of total plug power consumption in zone two, separated for weekdays/weekends

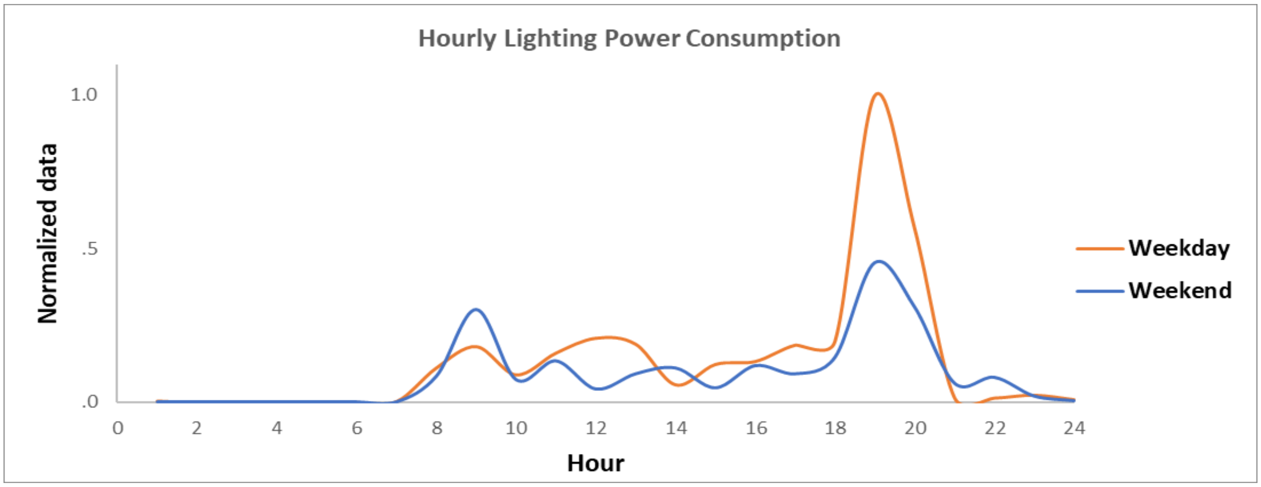


Figure 53. Aggregated-hourly variation of total lighting power consumption in zone two, separated for weekdays/weekends

4.4.2. Clustering step - Zone two (bedroom 2)

The clustering step includes four sub-steps, similar to zone one (Figure 34):

v) Feature selection:

The contributing attributes to perform clustering in this zone are aggregated-monthly data of three parameters, including Rom[1], LP[2], and PP[3].

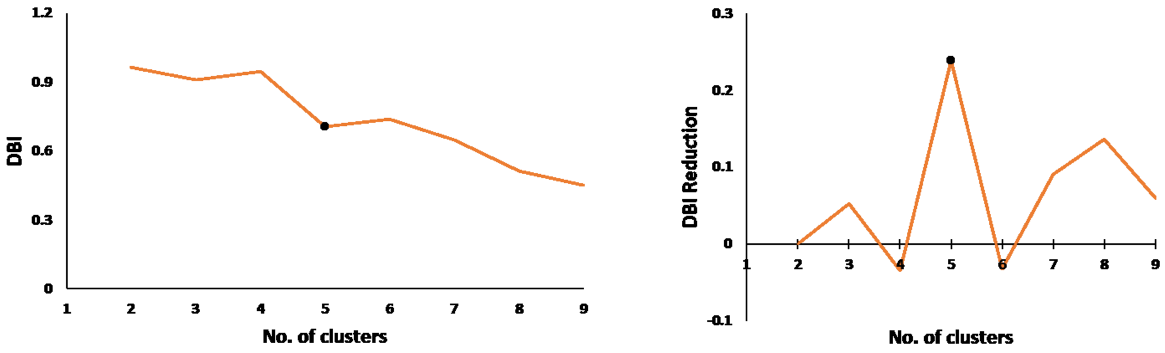
[1] Rate of motion
[2] Lighting power consumption
[3] Plug power consumption

vi) Normalization:

Similar to zone one, here the data is normalized by min-max normalization.

vii) Finding *k_{opt}*:

Figure 54 shows the results of DBI calculation. As it is illustrated in Figure 54.a, $k_{opt} = 5$ for zone two. Also, Figure 54.b shows that the amount of DBI reduction is the highest when the number of clusters is 5.



a. *b.*

Figure 54. Performance of clustering, a. DBI for different number of clusters, b. reduction in DBI by increasing the number of clusters

viii) Implementation:

The characteristics of clusters' centroids can be seen in Figure 55.

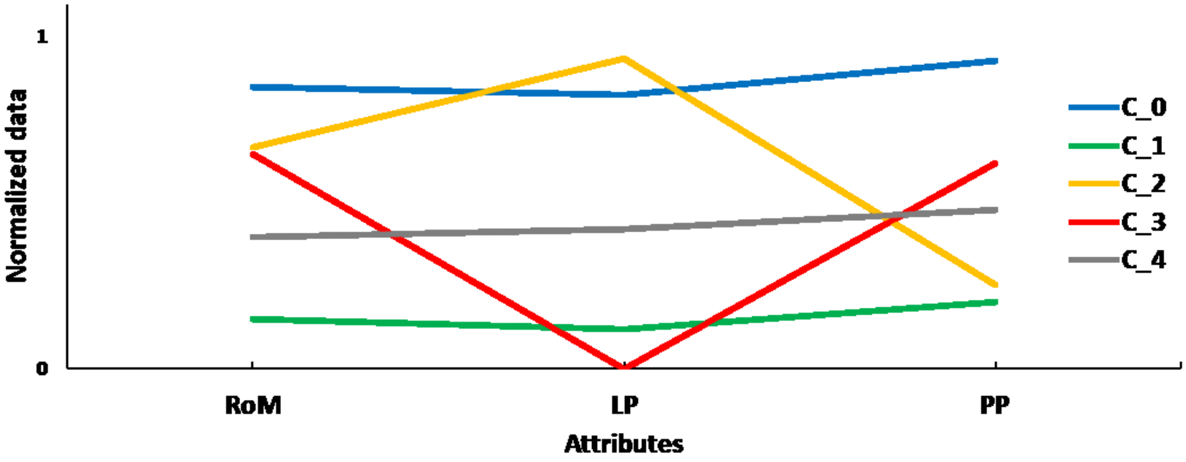


Figure 55. Visualization of cluster centroids

- **Clusters' specifications:**

Cluster_0: This cluster represents the moments that the level of activity and energy consumption by occupants in zone two is very high. In fact, RoM and plug power usage are the highest in the dataset.

Cluster_1: This cluster represents the moments that the level of activity and energy consumption by occupants in zone two is very low. Here, RoM and plug power usage are the lowest in the dataset.

Cluster_2: This cluster represents the moments that the level of activity by occupants in zone two is fairly high. Lighting power consumption is at the highest level and plug power usage near the lowest amount in the dataset.

Cluster_3: This cluster represents the moments that the level of activity and plug power usage by occupants in zone two are fairly high. Lighting power consumption is at the lowest level.

Cluster_4: This cluster represents the moments that both energy consumption and level of activity by occupants in zone two are around average.

The RapidMiner procedure for performing the clustering step is similar to zone one and presented in appendix D.

4.4.3. Baseline step - Zone two (bedroom 2)

Similar to zone one, the steps are as follow:

i) Calculation of *NNEC*:

The two energy consumers in zone two are considered as non-steady end-use. Therefore, for zone two, equation 8 become:

$$\mathrm{NEC_{zone_2}} = \mathrm{LP} + \mathrm{PP}$$ Equation 13

$$\mathrm{NNEC_{zone_2}} = \mathrm{normalized}\ (\mathrm{NEC_{zone_2}})$$ Equation 14

ii) Calculation of *OAI*:

According the equation 12 (section 4.4.1), for zone two OAI = RoM.

iii) Calculation of *REII*:

The *target REII* attribute for zone two for each cluster is generated by equation 15:

$$\mathrm{REII_{zone_2}} = \frac{\mathrm{NNEC_{zone_2}}}{\mathrm{OAI_{zone_2}}}$$ Equation 15

iv) Target REII for zone two:

Figure 56 shows the baseline (target hourly REII) for the five defined clusters in zone two.

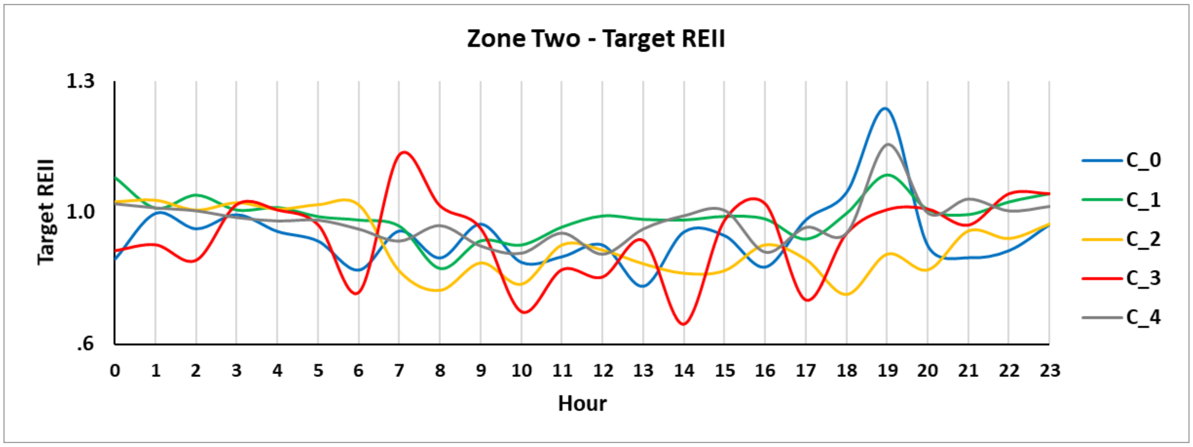
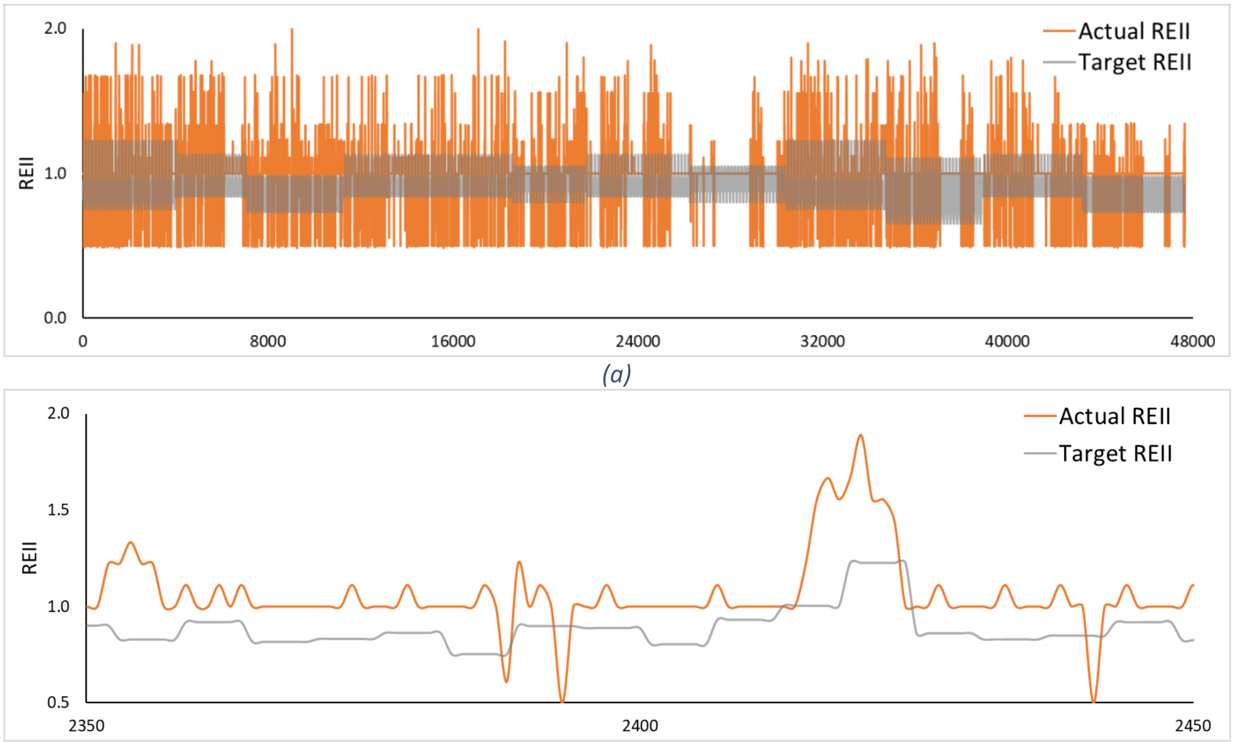


Figure 56. Target hourly REII for the five defined clusters

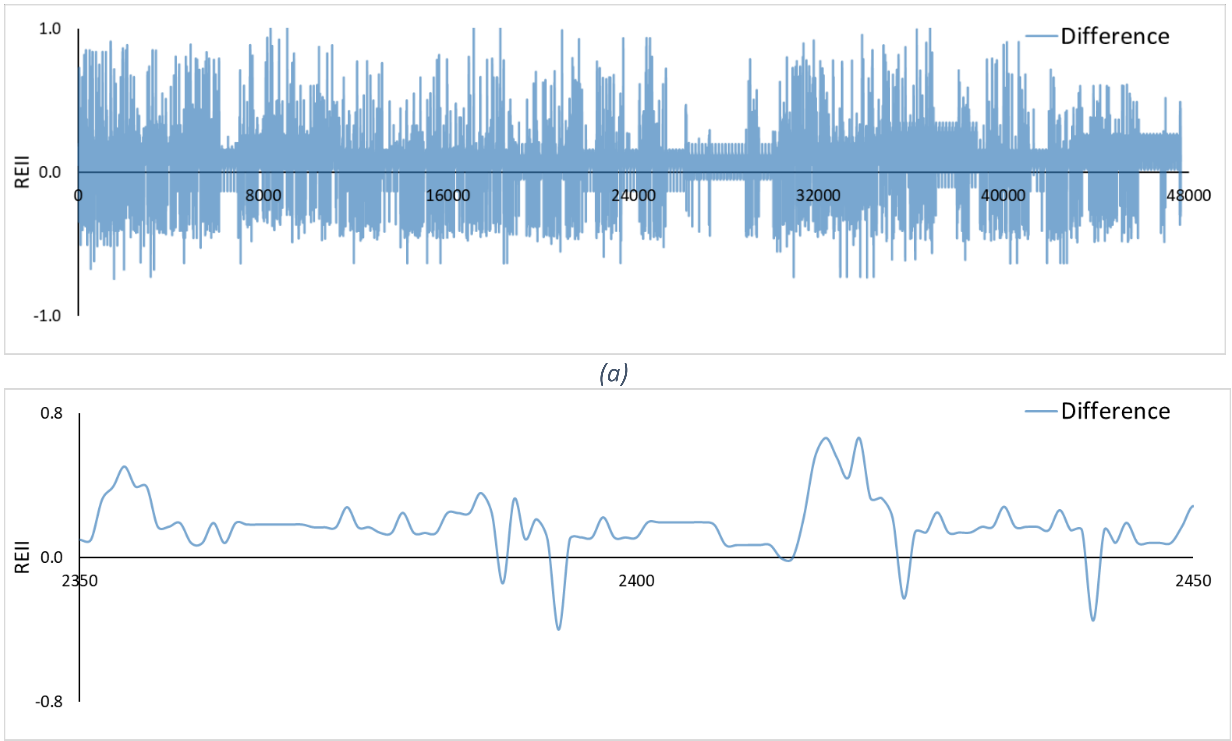
4.4.4. Energy wastage identification step - Zone two (bedroom 2)

Figure 57.a shows both the actual REII and the target one (baseline). Figure 57.b highlights a part of Figure 57.a for better understanding. Figure 58 shows this difference between the actual REII and the baseline.



(b)

Figure 57. Actual REII and target REII for the whole year. (a) whole year data, (b) zoom a part of the chart for better understanding



(a)

(b)

Figure 58. Difference between the actual REII and target REII for the whole year – Potential energy wastage/saving. (a) whole year data, (b) zoom a part of the chart for better understanding

4.5. Zone three (bedroom 3)

The third zone is bedroom three. In this zone, we have data for 11 sensors which are introduced in Appendix B. three sensors have been chosen to use in this study, which are CO_2, motion, and lighting power.

4.5.1. Energy use patterns investigation - Zone three (bedroom 3)

The following images illustrate the variation of energy consumption and its drivers for zone three of the apartment. Similar to zone 1 and 2, the data is normalized and aggregated hourly and is separated into weekdays/weekends, months, and day of the week. The RapidMiner procedure for this investigation is the same as the other zones (appendix C).

Figure 59 shows how hourly lighting power consumption is following CO_2 concentration in zone three. Figure 60 illustrates how hourly lighting power consumption follows occupants' motion.

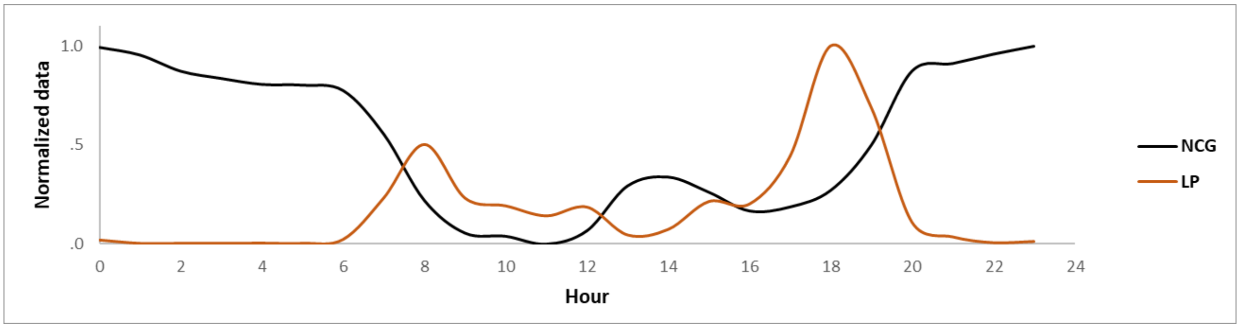


Figure 59. Aggregated-hourly variation of total lighting power consumption and CO2 concentration in zone 3

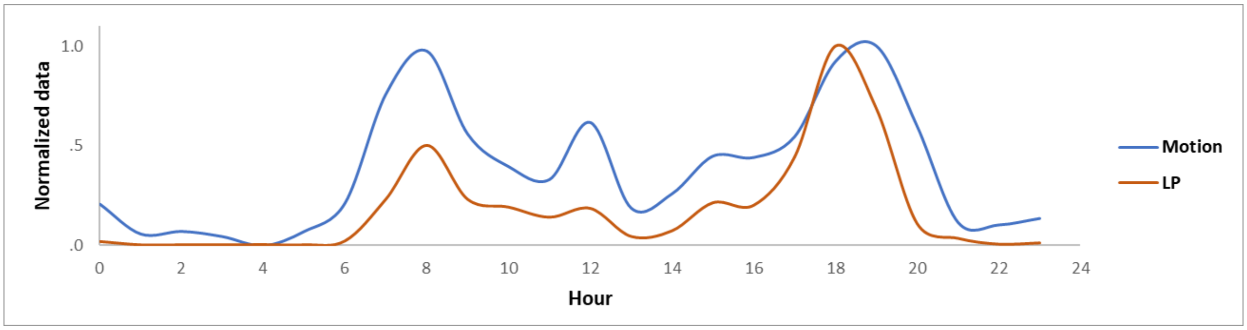


Figure 60. Aggregated-hourly variation of total lighting power consumption and total motion detected in zone 3

Figures 61 to 62 show the hourly variations for different days of the week (Sunday to Saturday). Figure 63 shows the aggregated hourly occupant's motion detected by the motion detector in zone three.

As it is shown in Figures 59 and 62, which are illustrating the hourly variation of normalized CO_2 concentration in zone three, the carbon dioxide amount in this bedroom is significantly higher during nights comparing to day times. The same as zone one and two, because of this inconstancy, we cannot consider CO_2 as an indicator of occupants' activity-level in this zone. Therefore, for bedroom three, equation 3 in section 3.1 is simplified to equation 16:

$$\mathrm{OAI}_{\mathrm{zone}_3} = \sqrt{(0 \times \mathrm{NCG})^2 + (1 \times \mathrm{RoM})^2} = \mathrm{RoM}_{zone3}$$ *Equation 16*

where RoM is rate of motion detected by motion detector in bedroom two (normalized-aggregated data.

Figure 63 shows the daily variation of hourly lighting power consumption in zone three.

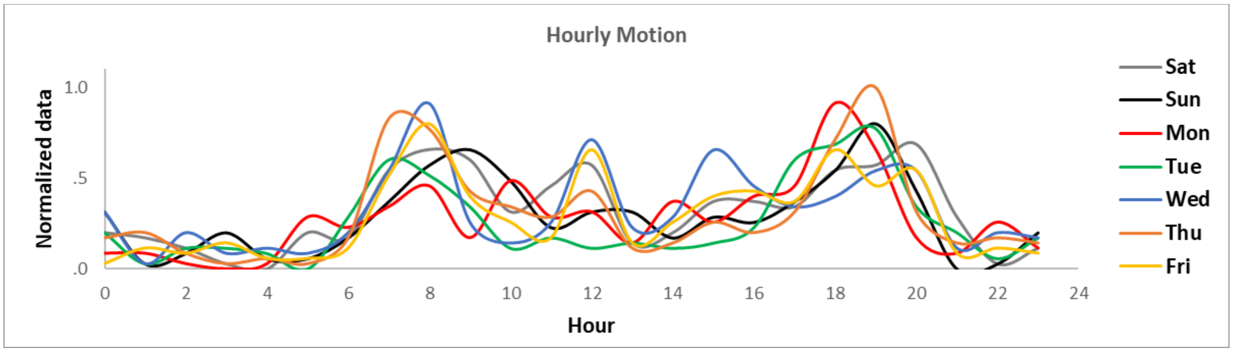


Figure 61. Aggregated-hourly variation of total motion detected in zone three per weekday

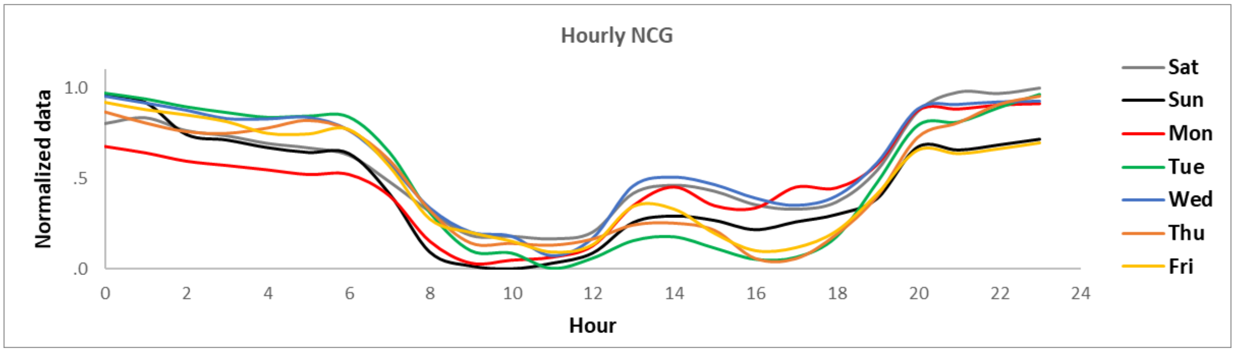


Figure 62. Aggregated-hourly variation of total CO_2 concentration in zone three per weekday

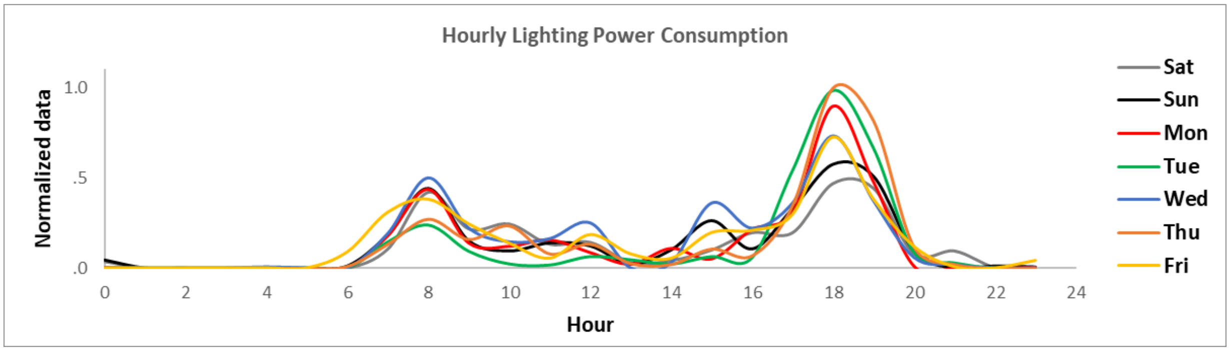


Figure 63. Aggregated-hourly variation of total lighting power consumption in zone three per weekday

Figures 64 to 66 show the hourly variations for all months (January to December). Figure 64 shows the aggregated hourly occupants' motion detected by the motion detector in zone three. Similar to Figures 59 and 62, as it is shown in Figure 65, the carbon dioxide amount in bedroom three is higher during nights comparing to day times. Figure 66 shows the monthly variation of hourly energy consumption in zone three. Referring to the monthly aggregated data, it seems that this bedroom was also unoccupied in most of the days in July and August.

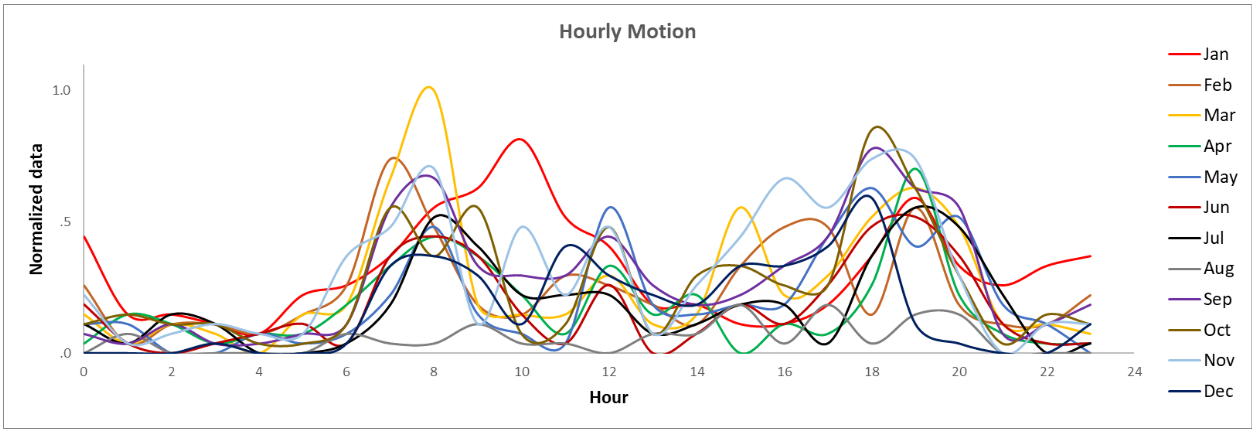


Figure 64. Aggregated-hourly variation of total motion detected in zone three per month

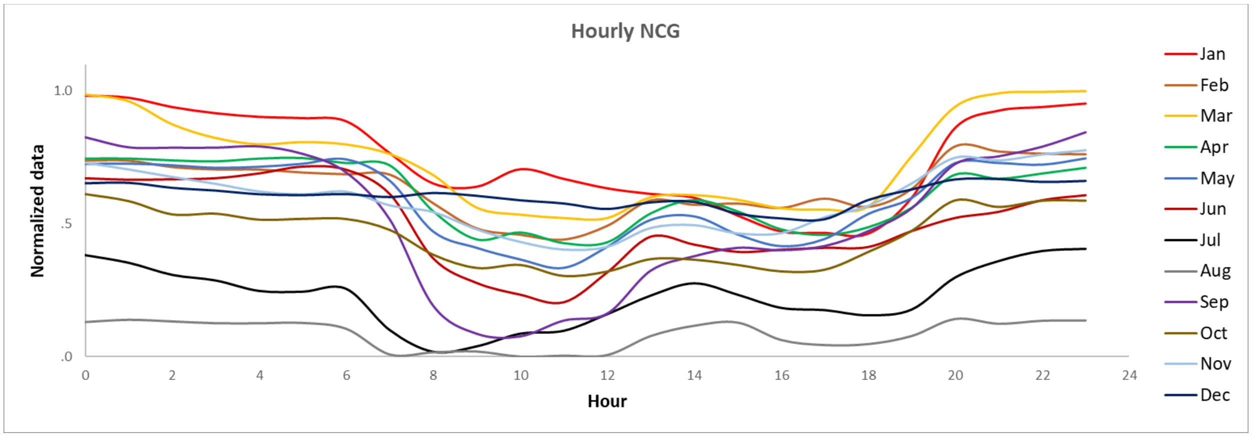


Figure 65. Aggregated-hourly variation of total CO_2 concentration in zone three per month

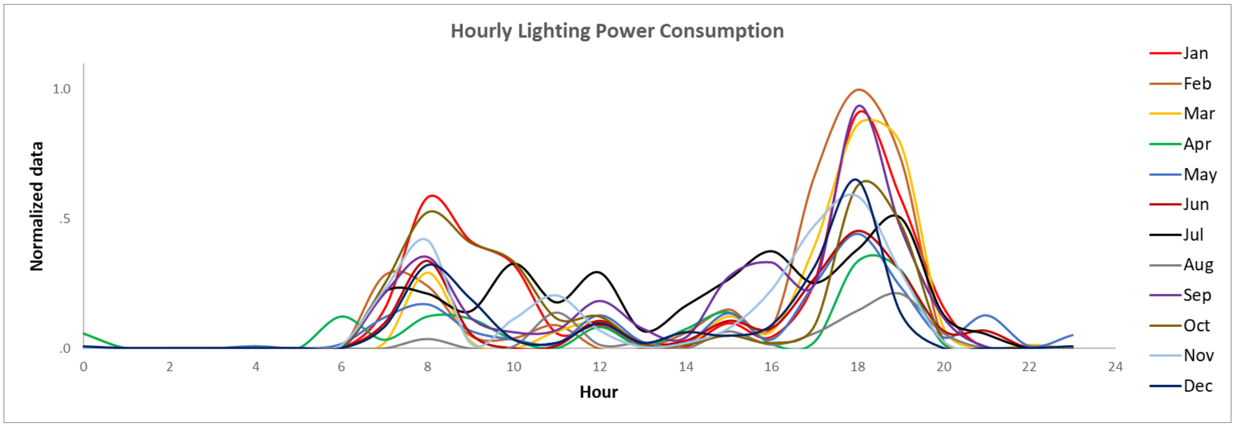


Figure 66. Aggregated-hourly variation of total lighting power consumption in zone three per month

Figures 67 to 69 show the hourly variations for weekdays/weekends. Figure 67 shows the aggregated hourly occupant's motion detected by the motion detector in zone three. Similar to Figures 59, 62, and 65, as it is shown in Figure 68, the carbon dioxide amount in this bedroom is significantly higher during nights comparing to day times. Figure 69 shows the variation of hourly lighting power consumption in zone three.

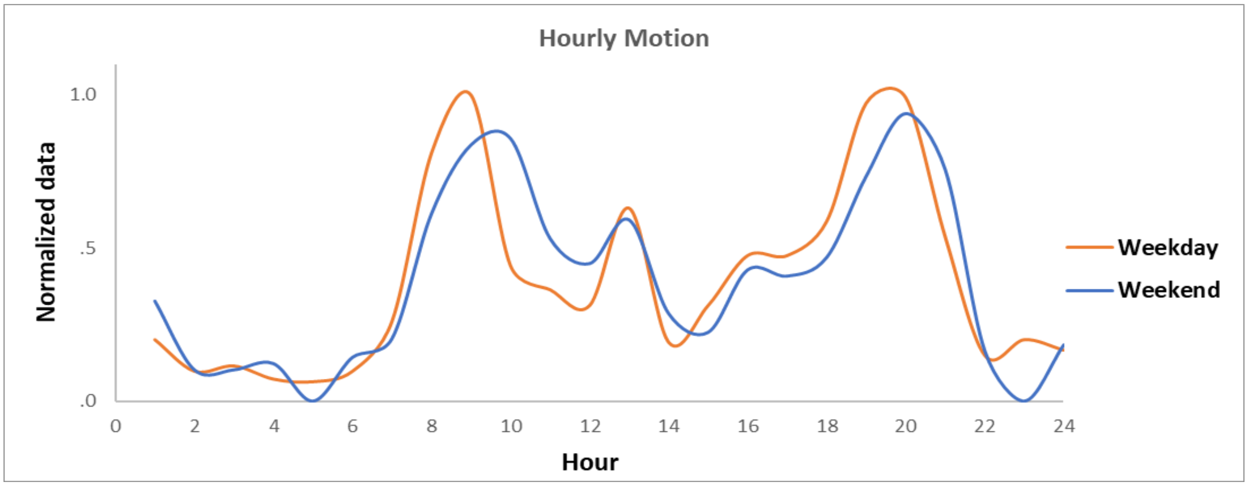


Figure 67. Aggregated-hourly variation of total occupants' motion detected in zone three, separated for weekdays/weekends

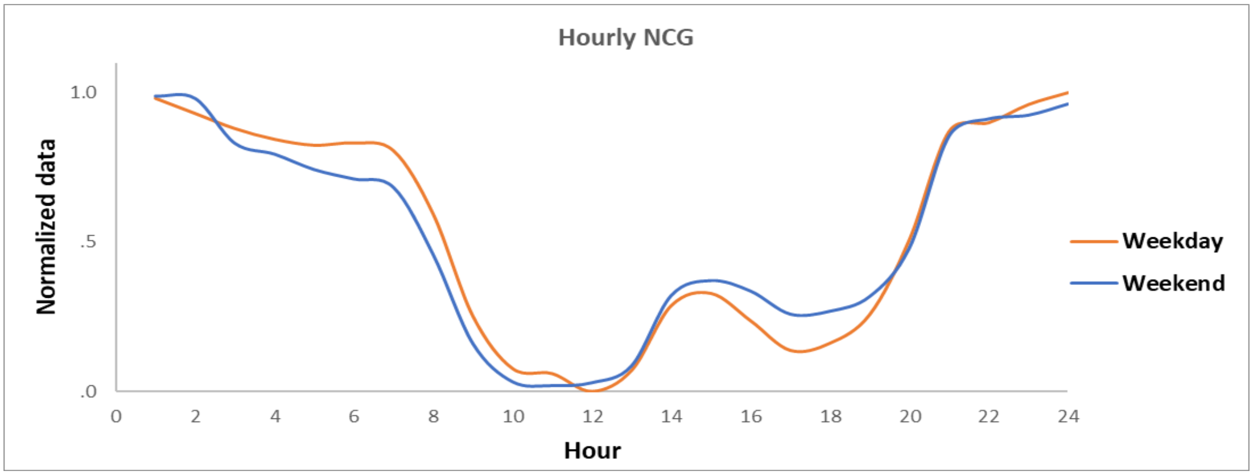


Figure 68. Aggregated-hourly variation of total CO_2 concentration in zone three, separated for weekdays/weekends

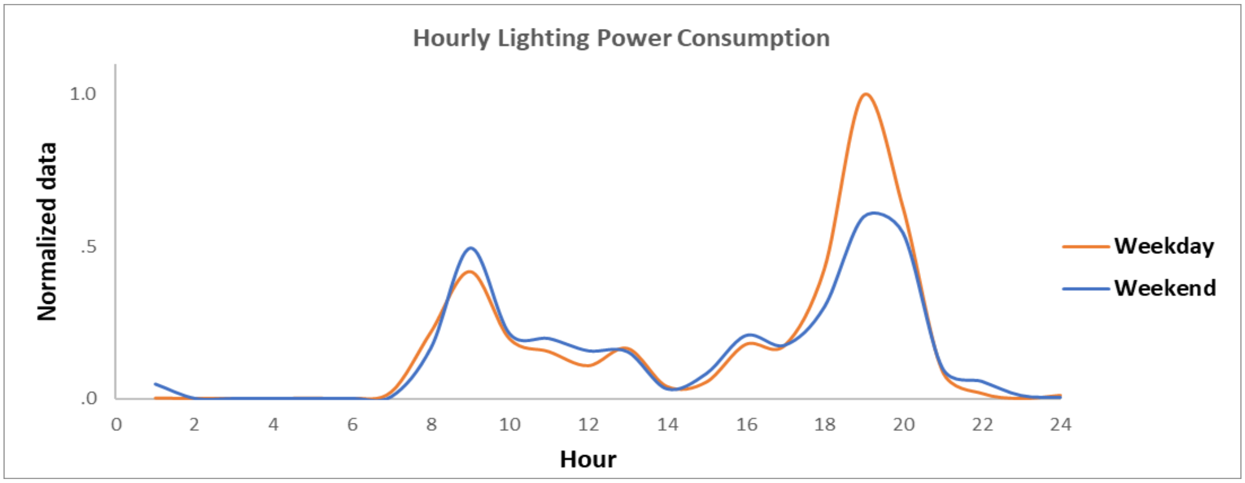


Figure 69. Aggregated-hourly variation of total lighting power consumption in zone three, separated for weekdays/weekends

4.5.2. Clustering step - Zone three (bedroom 3)

The clustering step includes four sub-steps, similar to other zones (Figure 34):

i) Feature selection:

The contributing attributes to perform clustering in this zone are aggregated-monthly data of Rom[1] and LP[2].

ii) Normalization:

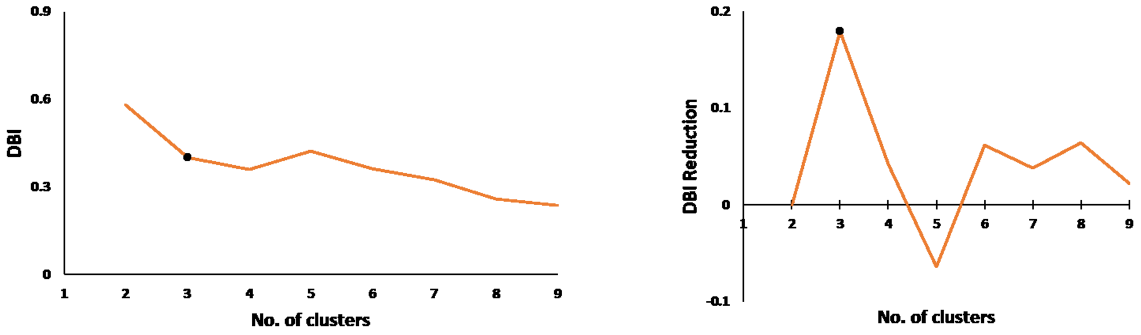
Similar to other zones, here the data is normalized by min-max normalization.

iii) Finding k_{opt}:

Figure 70 shows the results of DBI calculation. As it is illustrated in Figure 70.a, $k_{opt} = 3$ for zone three. Also, Figure 70.b shows that the amount of DBI reduction is the highest when the number of clusters is 3.

[1] Rate of motion

[2] Lighting power consumption



a.

b.

Figure 70. Performance of clustering, a. DBI for different number of clusters, b. reduction in DBI by increasing the number of clusters

iv) Implementation:

The characteristics of clusters' centroids can be seen in Figure 71.

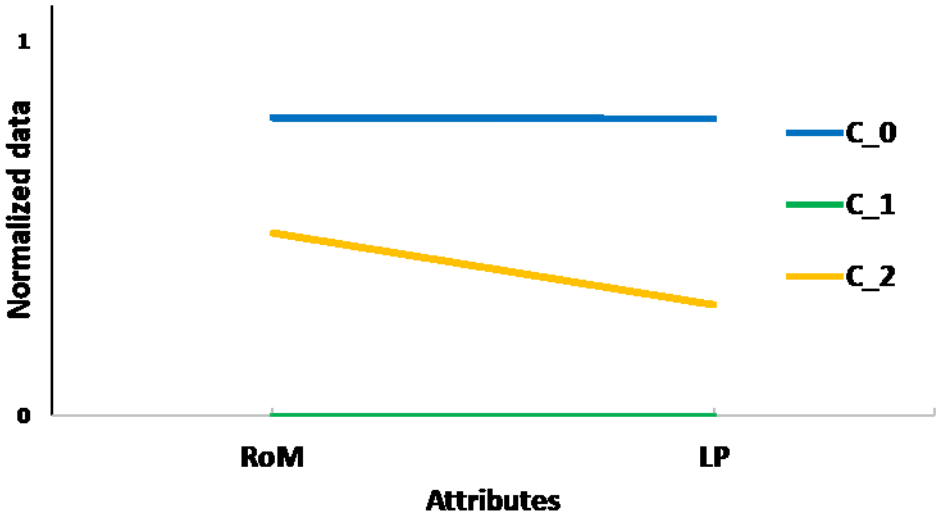


Figure 71. Visualization of cluster centroids

- **Clusters' specifications:**

Cluster_0: This cluster represents the moments that the level of activity and energy consumption by occupants in zone three are at the highest level.

Cluster_1: This cluster represents the moments that the level of activity and energy consumption by occupants in zone three are at the lowest level.

Cluster_2: This cluster represents the moments that both energy consumption and level of activity by occupants in zone three are around average.

The RapidMiner procedure for performing the clustering step is similar to zones one and two and presented in appendix D.

4.5.3. Baseline step - Zone three (bedroom 3)

Similar to zone one, the steps are as follow:

i) **Calculation of *NNEC*:**

In zone three the only energy consumer is lighting power. Therefore, equation 8 becomes:

$$\mathrm{NEC_{zone_3}} = \mathrm{LP} \qquad \textit{Equation 17}$$

$$\mathrm{NNEC_{zone_3}} = \mathrm{normalized}\ (\mathrm{NEC_{zone_3}}) \qquad \textit{Equation 18}$$

ii) **Calculation of *OAI*[1]:**

According the equation 16 (section 4.5.1), for zone three OAI = RoM.

iii) **Calculation of *REII*:**

The *target REII* attribute for zone three for each cluster is generated by equation 19:

$$\mathrm{REII_{zone_3}} = \frac{\mathrm{NNEC_{zone_3}}}{\mathrm{OAI_{zone_3}}} \qquad \textit{Equation 19}$$

iv) **Target REII for zone three:**

Figure 72 shows the baseline (target hourly REII) for the three defined clusters in zone three.

[1] Occupant Activity Indicator

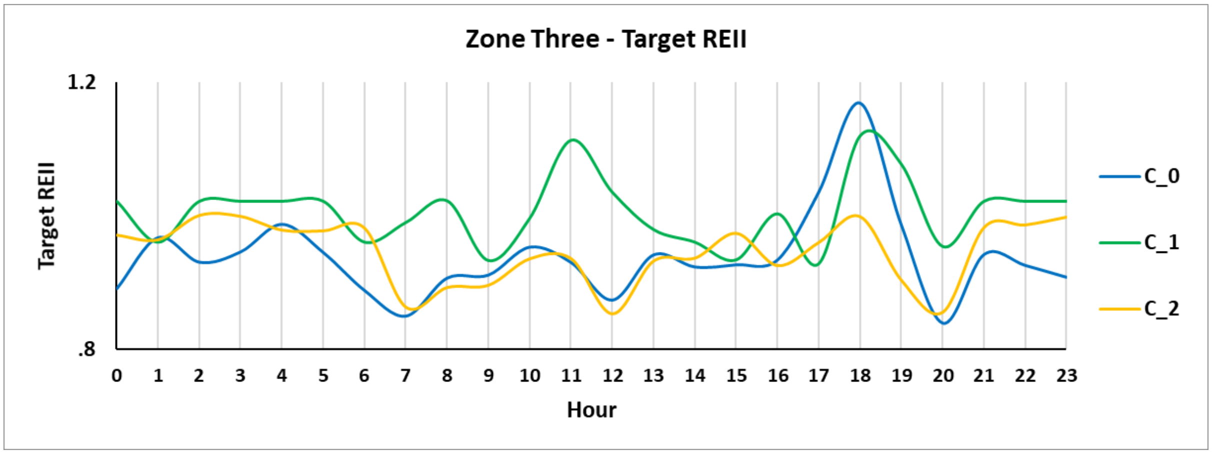
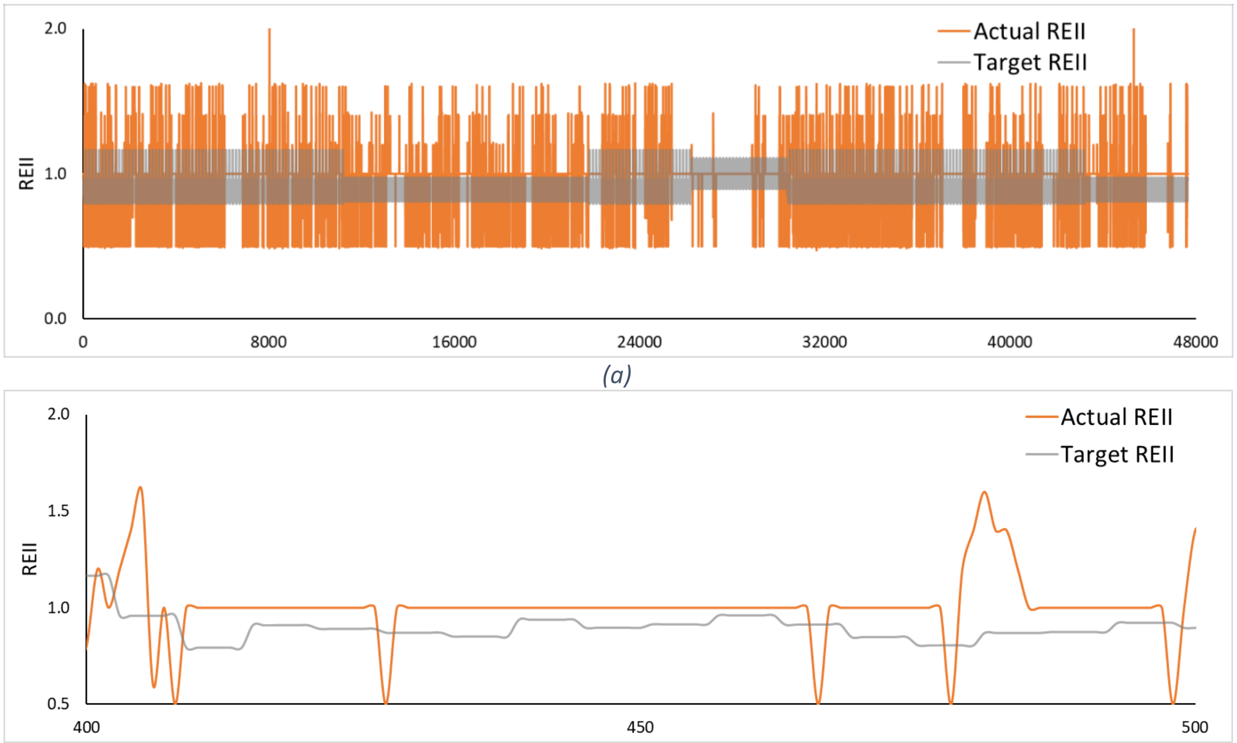


Figure 72. Target hourly REII for the three defined clusters

4.5.4. Energy wastage identification step - Zone three (bedroom 3)

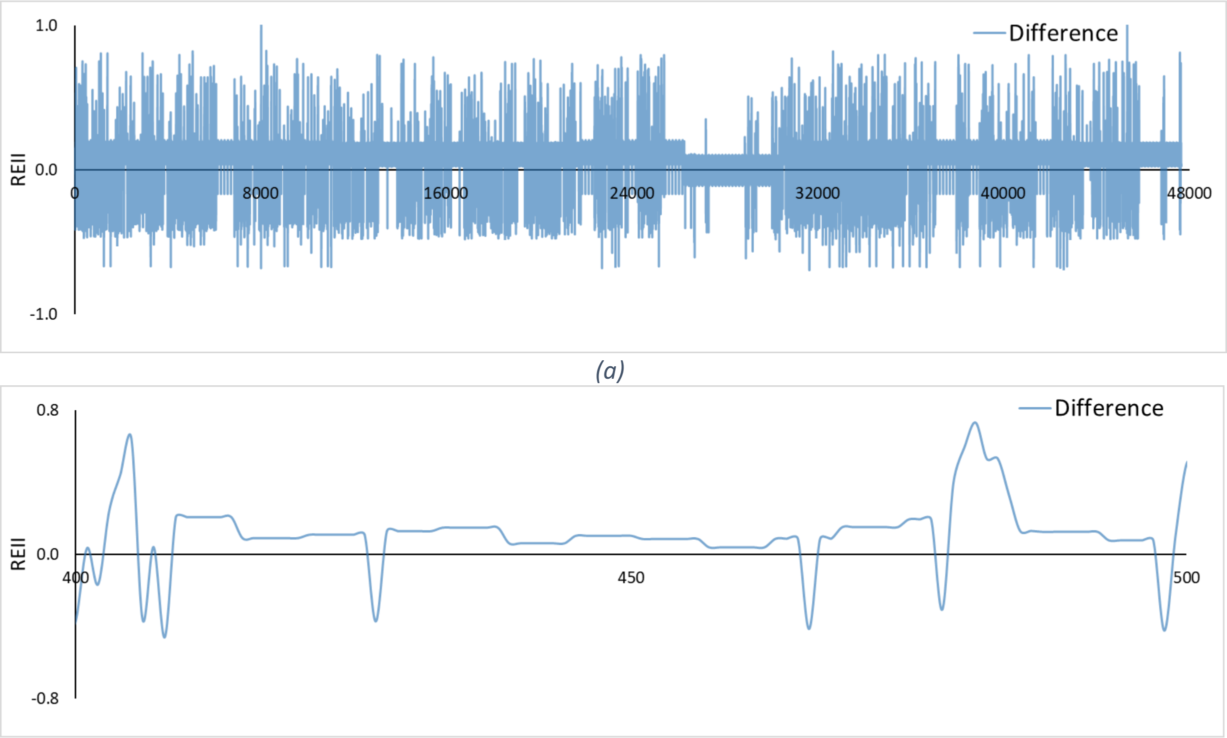
Figure 73.a shows both the actual REII and the target one (baseline). Figure 73.b highlights a part of Figure 73.a for better understanding. Figure 74 shows this difference between the actual REII and the baseline.



(a)

(b)

Figure 73. Actual REII and target REII for the whole year. (a) whole year data, (b) zoom a part of the chart for better understanding



(a)

(b)

Figure 74. Difference between the actual REII and target REII for the whole year – Potential energy wastage/saving. (a) whole year data, (b) zoom a part of the chart for better understanding

4.6. Zone four (Kitchen & living room)

The fourth zone is kitchen and living room. In this zone, we have data for 48 sensors which are introduced in Appendix B. Twenty-two sensors have been chosen to use in this study, which are one CO_2 sensor, three motion sensors, five lighting power sensors, and 13 plug power sensors.

4.6.1. Energy use patterns investigation - Zone four (Kitchen & living room)

The following images illustrate the variation of energy consumption and its drivers for zone four of the apartment. Similar to other zones, the data is normalized and aggregated hourly and is separated into weekdays/weekends, months, and day of the week. The RapidMiner procedure for this investigation is the same as the other zones (appendix C).

Figure 75 and 76 show how the hourly plug power consumption is following occupants' motion and CO_2 concentration in zone four, respectively. Figure 77 and 78 show how the hourly lighting power consumption is following occupants' motion and CO_2 concentration, respectively.

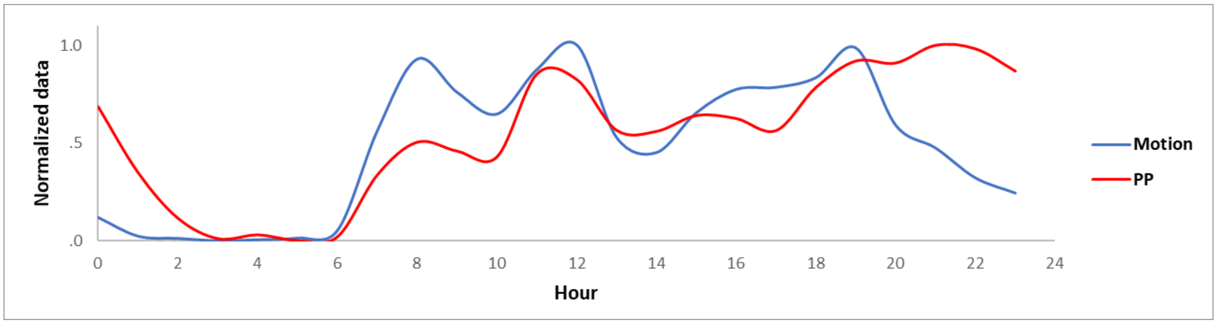


Figure 75. Aggregated-hourly variation of total plug power consumption and total motion detected in zone four

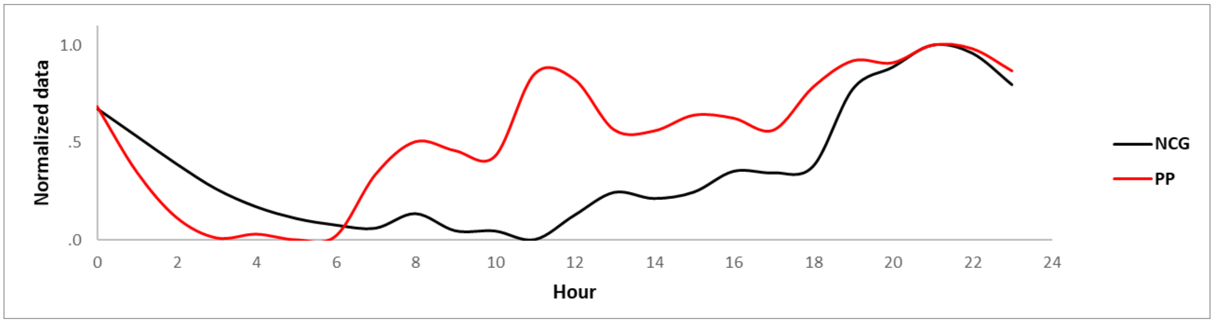


Figure 76. Aggregated-hourly variation of total plug power consumption and CO_2 concentration in zone four

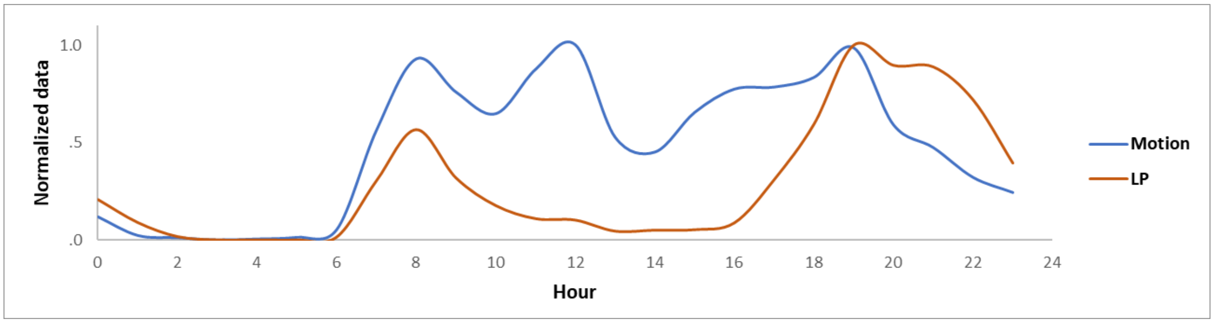


Figure 77. Aggregated-hourly variation of total lighting power consumption and total motion detected in zone four

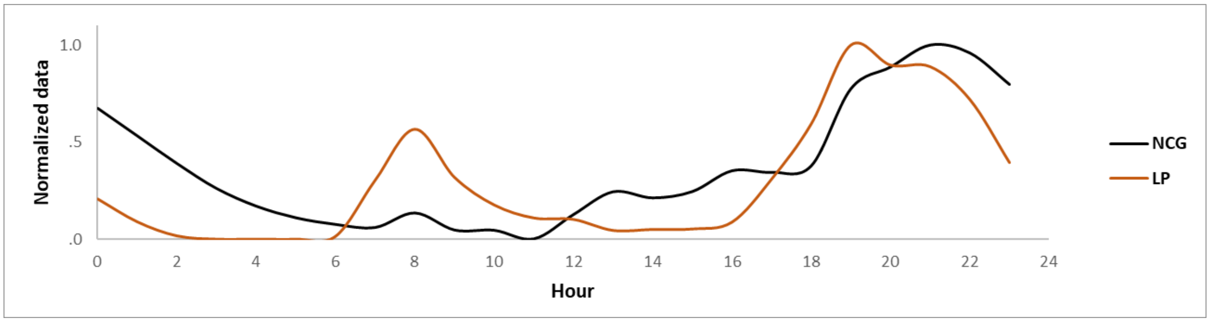


Figure 78. Aggregated-hourly variation of total lighting power consumption and CO_2 concentration in zone four

Figures 79 to 82 show the hourly variations for different days of the week (Sunday to Saturday). Figure 79 shows the aggregated hourly occupant's motion detected by the motion detector in zone four. Figure 80 illustrates the hourly variation of normalized CO_2 concentration in zone four. Figures 81 and 82 show the daily variation of hourly energy consumption in zone four.

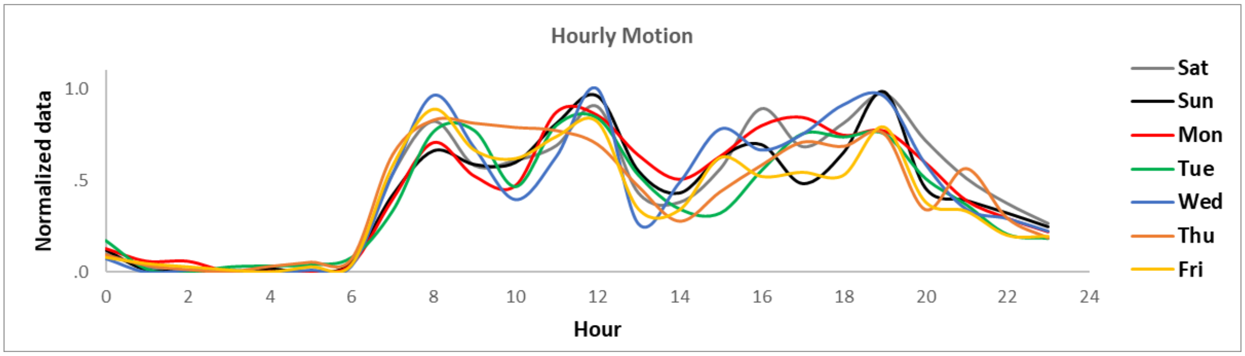


Figure 79. Aggregated-hourly variation of total motion detected in zone four per weekday

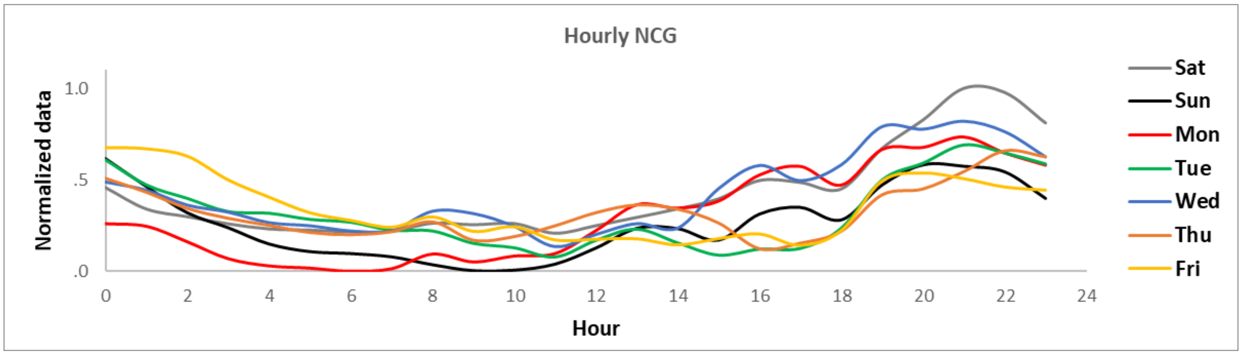


Figure 80. Aggregated-hourly variation of total CO_2 concentration in zone four per weekday

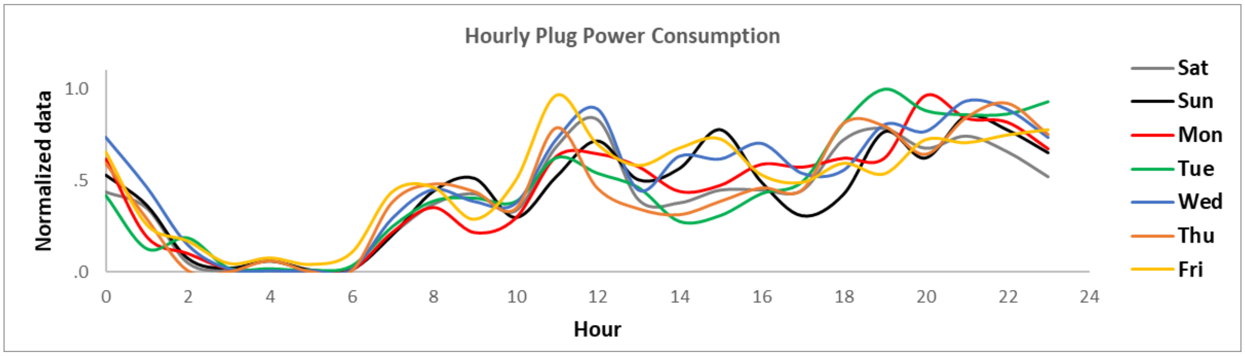


Figure 81. Aggregated-hourly variation of total plug power consumption in zone four per weekday

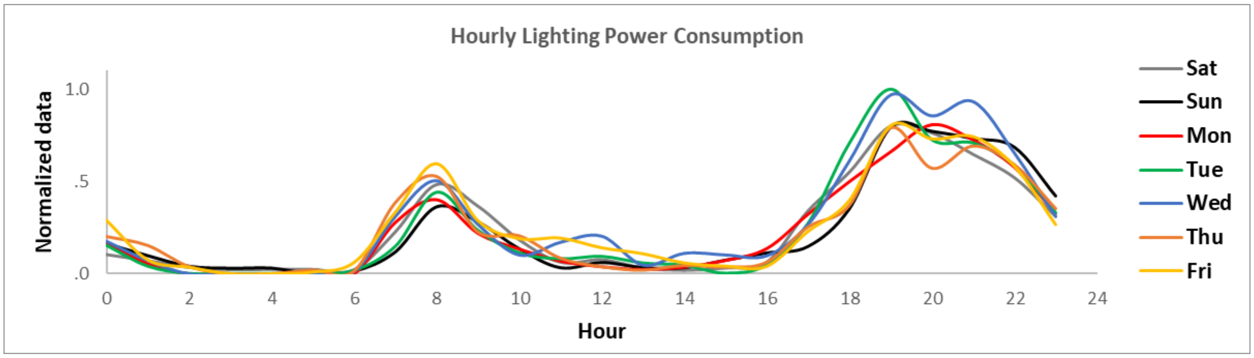


Figure 82. Aggregated-hourly variation of total lighting power consumption in zone four per weekday

Figures 83 to 86 show the hourly variations for all months (January to December). Figure 83 shows the aggregated hourly occupant's motion detected by the motion detector in zone four. Figure 84 shows the hourly CO_2 concentration. Figures 85 and 86 show the monthly variation of hourly energy consumption in zone four. Referring to the monthly aggregated data, it seems that this zone was also unoccupied in most of the days in August.

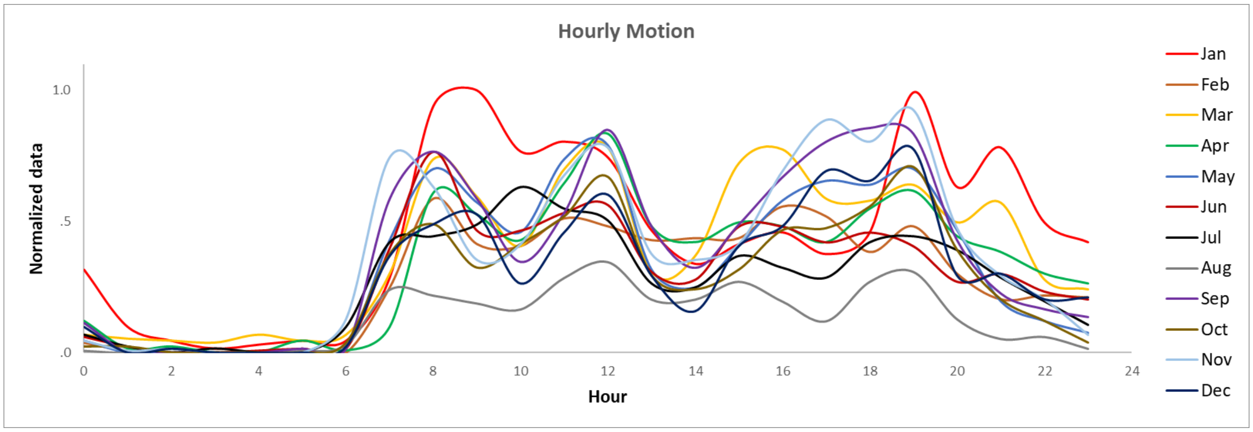


Figure 83. Aggregated-hourly variation of total motion detected in zone four per month

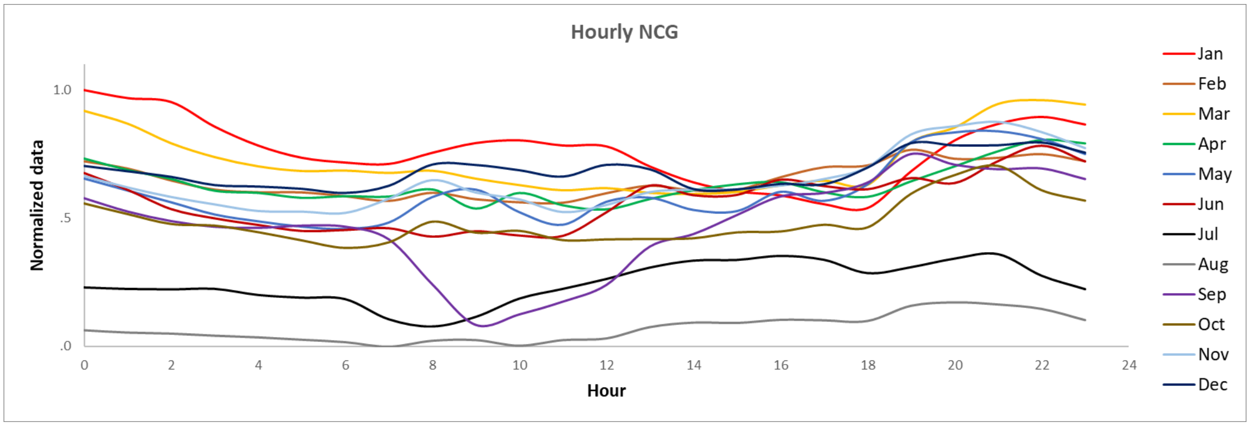


Figure 84. Aggregated-hourly variation of total CO₂ concentration in zone four per month

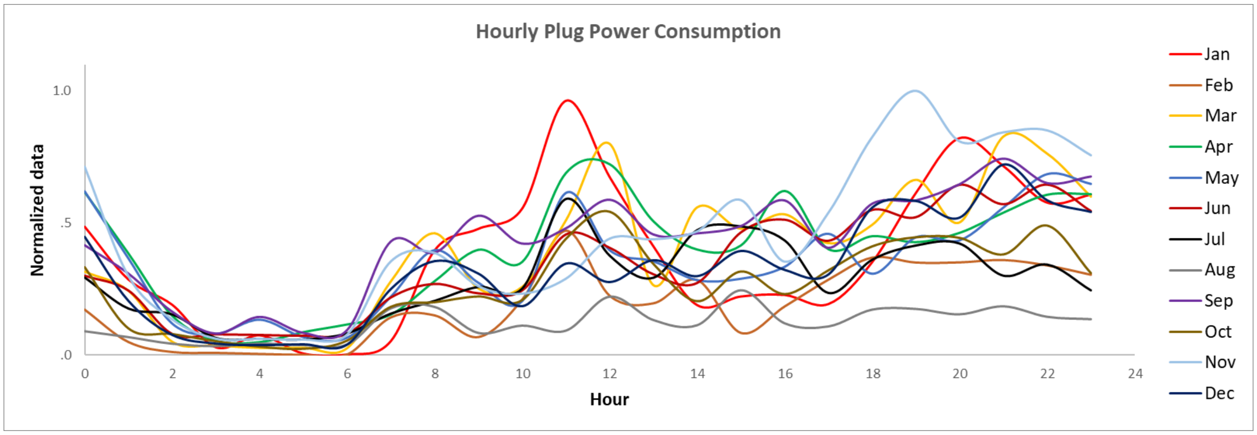


Figure 85. Aggregated-hourly variation of total plug power consumption in zone four per month

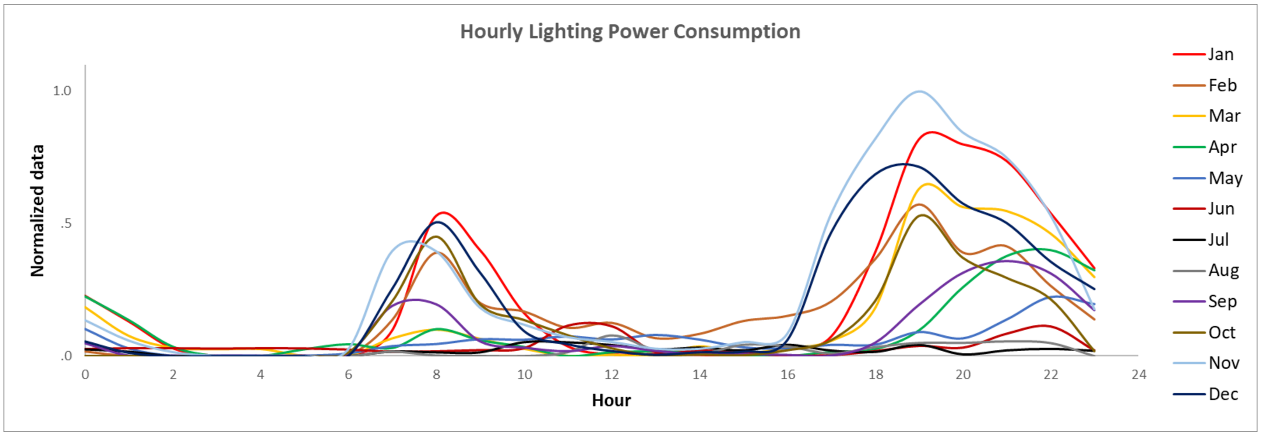


Figure 86. Aggregated-hourly variation of total lighting power consumption in zone four per month

Figures 87 to 98 show the hourly variations for weekdays/weekends. Figures 87 to 90 show the aggregated hourly occupant's motion detected by the motion detector, hourly CO_2

concentration, hourly plug power consumption, and hourly lighting power consumption separated for weekends and weekdays, respectively.

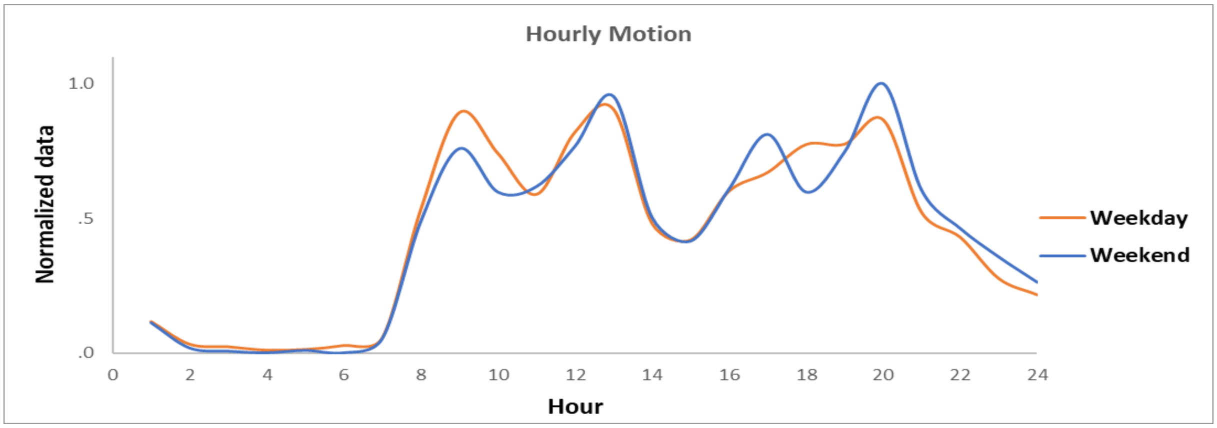


Figure 87. Aggregated-hourly variation of total occupants' motion detected in zone four, separated for weekdays/weekends

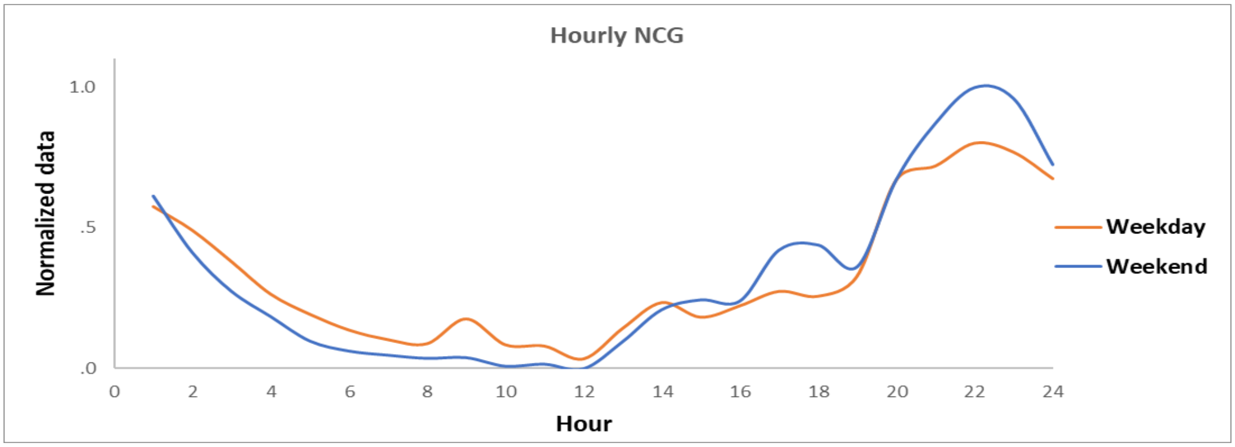


Figure 88. Aggregated-hourly variation of total CO_2 concentration in zone four, separated for weekdays/weekends

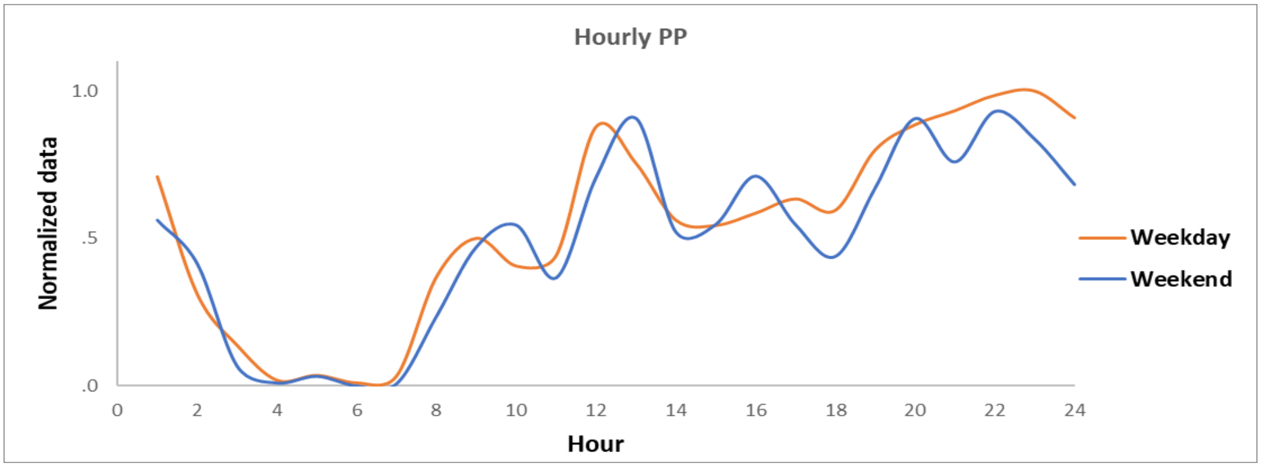


Figure 89. Aggregated-hourly variation of total plug power consumption in zone four, separated for weekdays/weekends

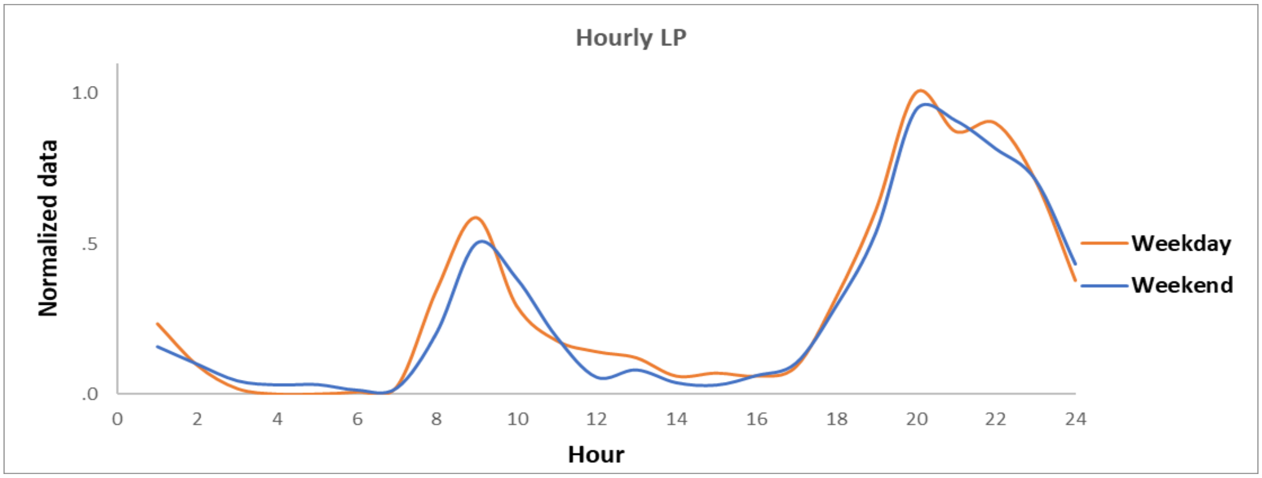


Figure 90. Aggregated-hourly variation of total lighting power consumption in zone four, separated for weekdays/weekends

Figure 91 and 92 show how the plug power consumption follows the total motion detected by occupants in zone four for weekdays and weekends, respectively. Figures 93 and 94 show how the lighting power consumption follows the occupants' motion for weekdays and weekends, respectively.

Figure 95 and 96 shows how the plug power consumption follows the CO_2 concentration in zone four for weekdays and weekends, respectively. Figure 97 and 98 shows how the lighting power consumption follows the CO_2 concentration in zone four for weekdays and weekends, respectively.

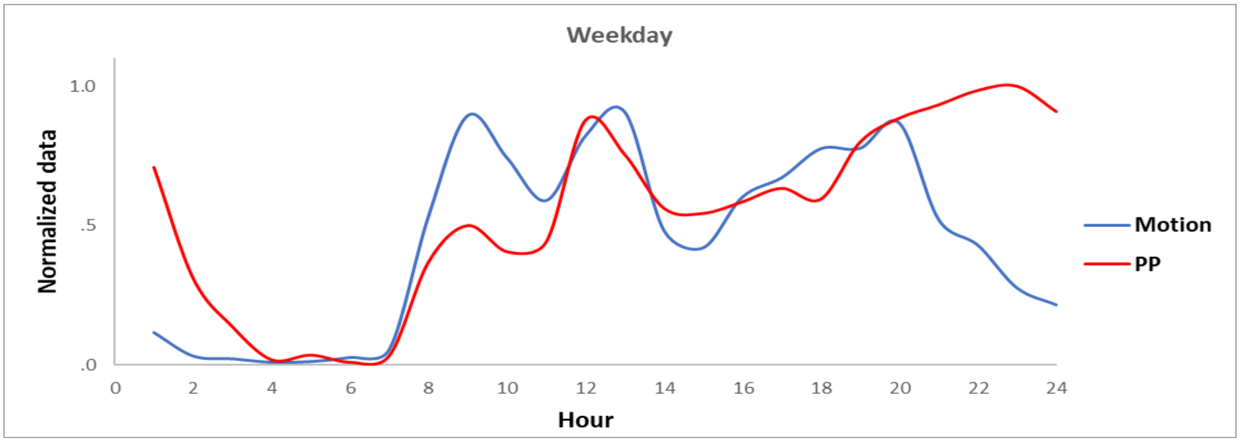


Figure 91. Aggregated-hourly variation of total plug power consumption and occupants' motion detected in zone four, separated for weekdays

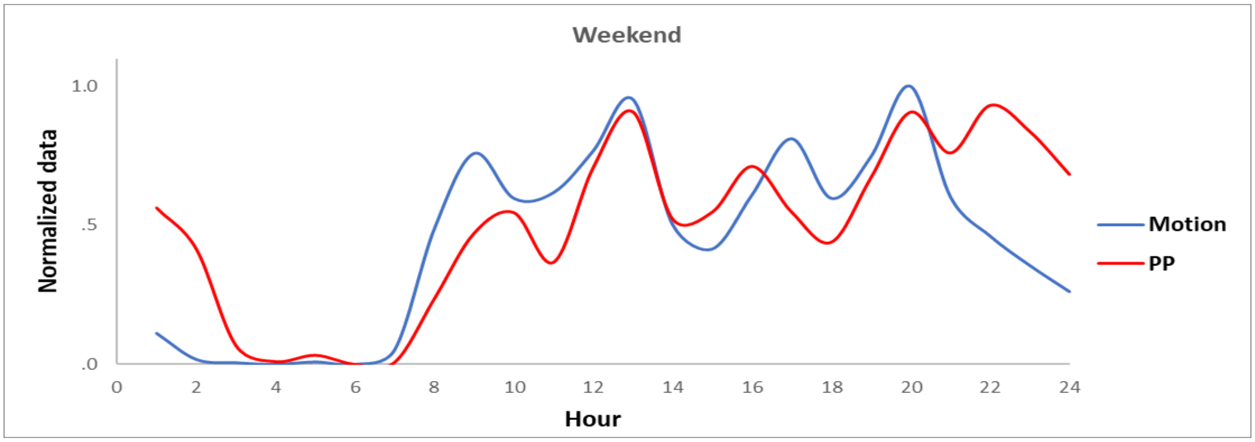


Figure 92. Aggregated-hourly variation of total plug power consumption and occupants' motion detected in zone four, separated for weekends

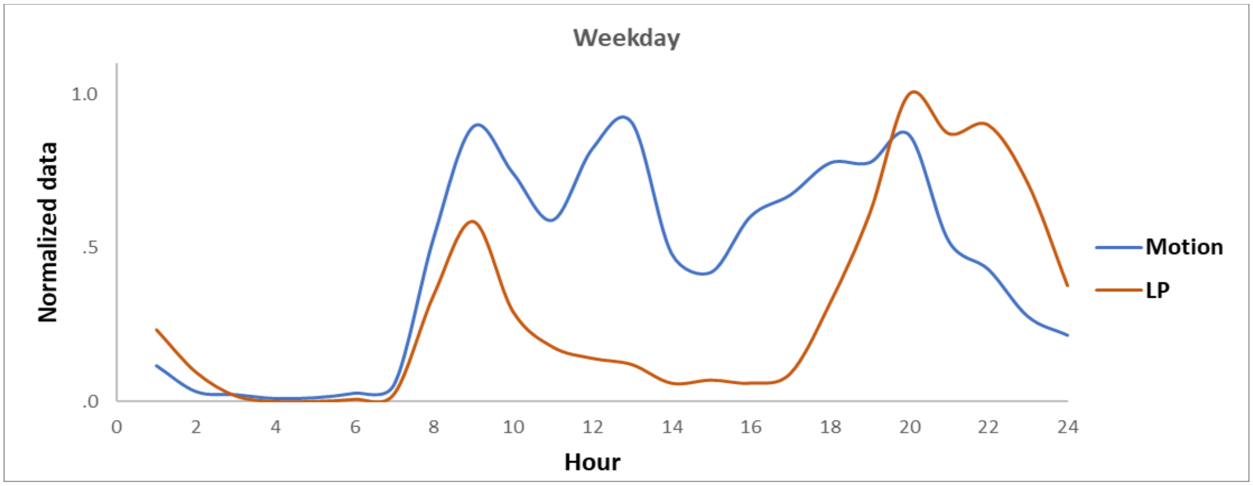


Figure 93. Aggregated-hourly variation of total lighting power consumption and occupants' motion detected in zone four, separated for weekdays

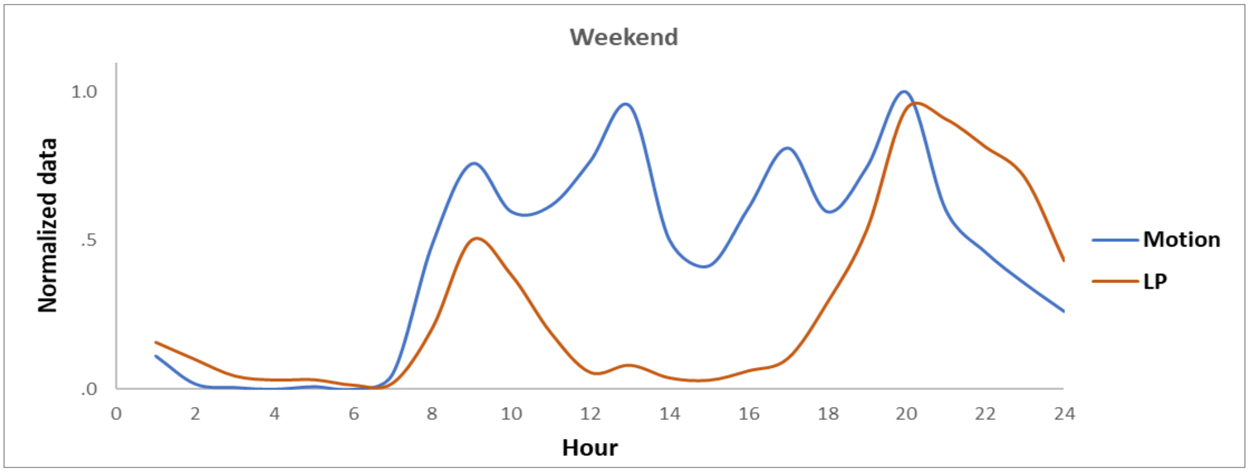


Figure 94. Aggregated-hourly variation of total plug lighting consumption and occupants' motion detected in zone four, separated for weekends

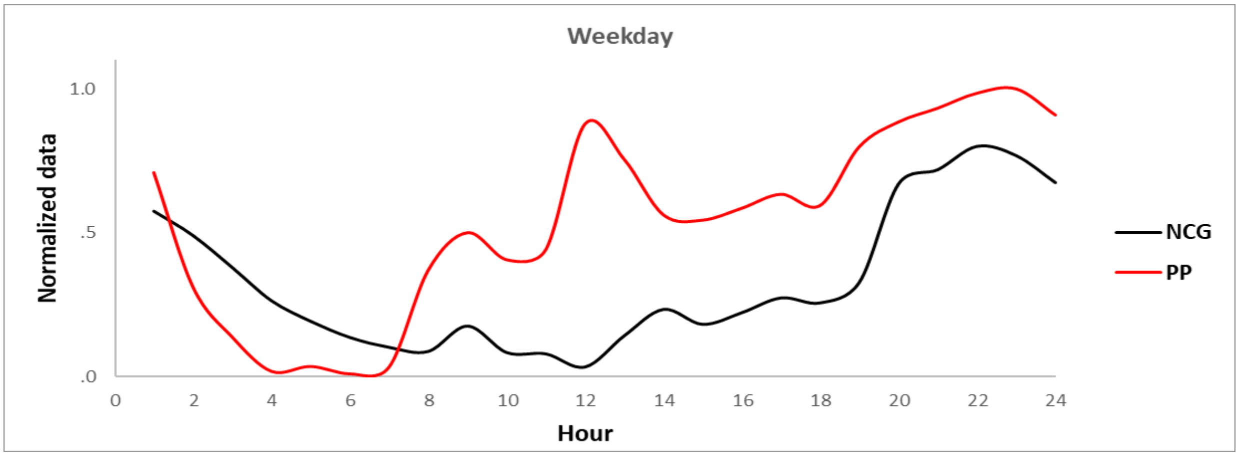


Figure 95. Aggregated-hourly variation of total plug power consumption and CO₂ concentration in zone four, separated for weekdays

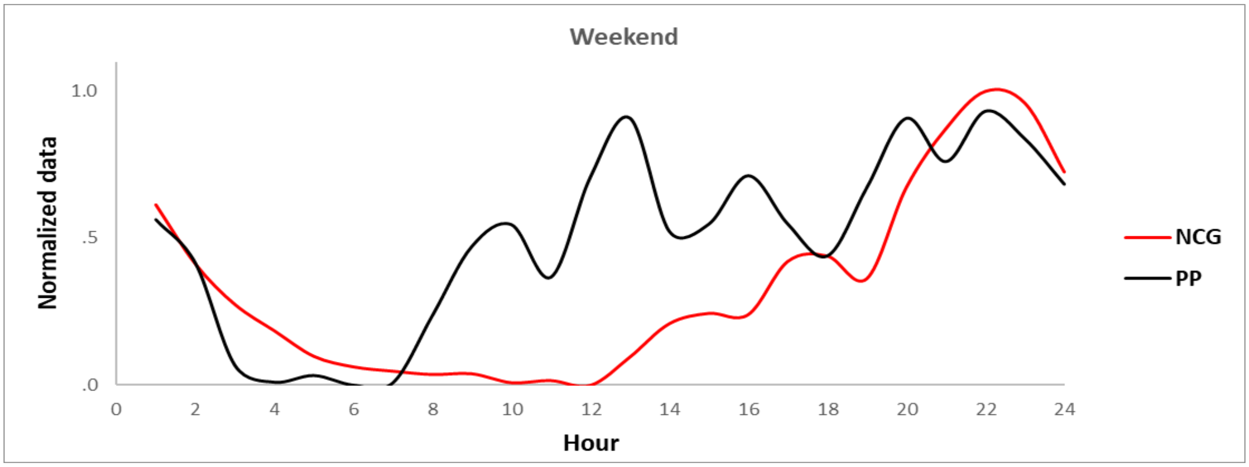


Figure 96. Aggregated-hourly variation of total plug power consumption and CO₂ concentration in zone four, separated for weekends

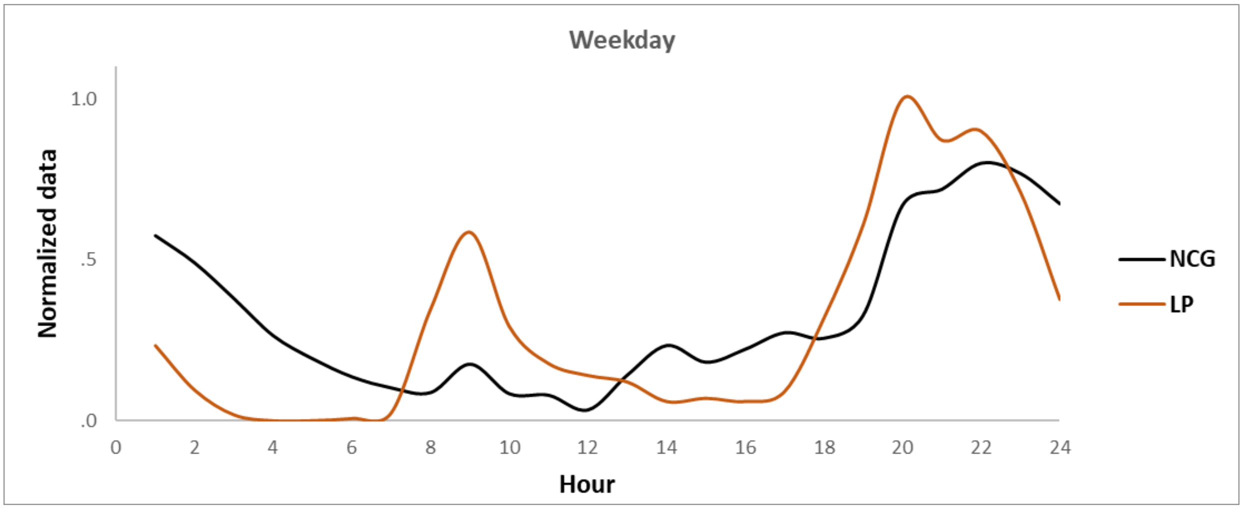


Figure 97. Aggregated-hourly variation of total lighting power consumption and CO₂ concentration in zone four, separated for weekdays

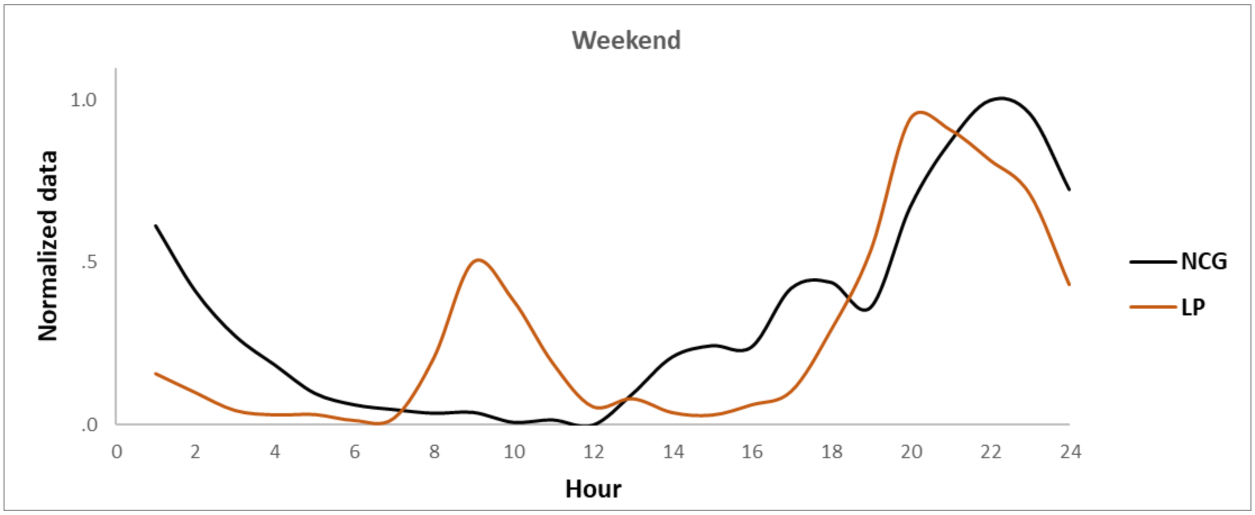


Figure 98. Aggregated-hourly variation of total plug lighting consumption and CO$_2$ concentration in zone four, separated for weekends

4.6.2. Clustering step - Zone four (Kitchen & living room)

Similar to the other zones, the clustering step includes four sub-steps.

i) Feature selection:

The contributing attributes to perform clustering for zone four are aggregated-monthly data of seven parameters, including NCG[1], Rom_{kit}[2], Rom_{liv}[3], LP_{kit}[4], LP_{liv}[5], PP_{kit}[6] and PP_{liv}[7]. Out of these seven features, three of them are components of OAI[8] (e.g. NCG, RoM_{kit}, and RoM_{liv}), and the rest represent the energy consumption in this zone.

ii) Normalization:

Here the data is normalized by min-max normalization, the same as other zones.

[1] Normalized CO$_2$ Generation
[2] Rate of motion (Kitchen)
[3] Rate of motion (living room)
[4] Lighting power consumption (Kitchen)
[5] Lighting power consumption (living room)
[6] Plug power consumption (Kitchen)
[7] Plug power consumption (living room)
[8] Occupant Activity Indicator

iii) Finding k_{opt}:

Figure 99 shows the results of DBI calculation. As it is illustrated in Figure 99.a, $k_{opt} = 5$ for zone four. Also, Figure 99.b shows that the amount of DBI reduction is the highest when the number of clusters is 5.

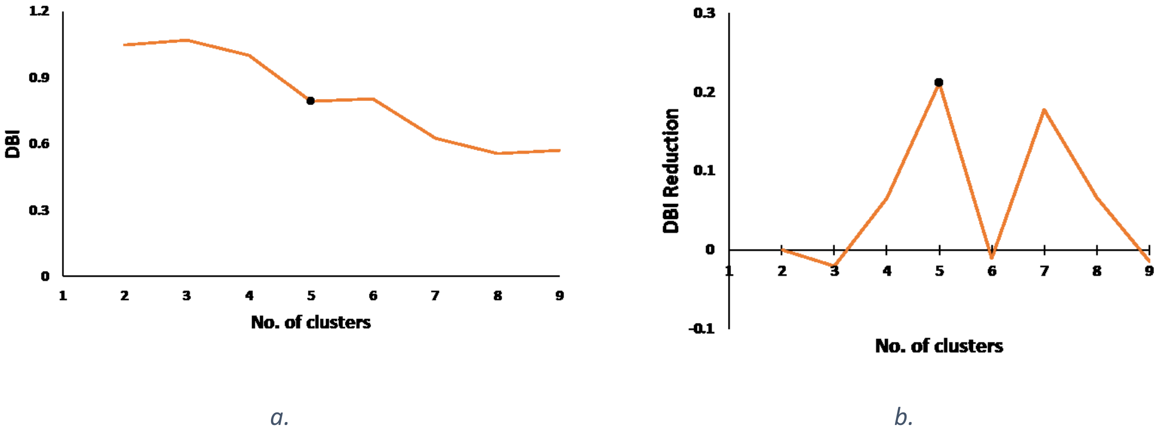


Figure 99. Performance of clustering, a. DBI for different number of clusters, b. reduction in DBI by increasing the number of clusters

iv) Implementation:

Figure 100 represents the characteristics of clusters.

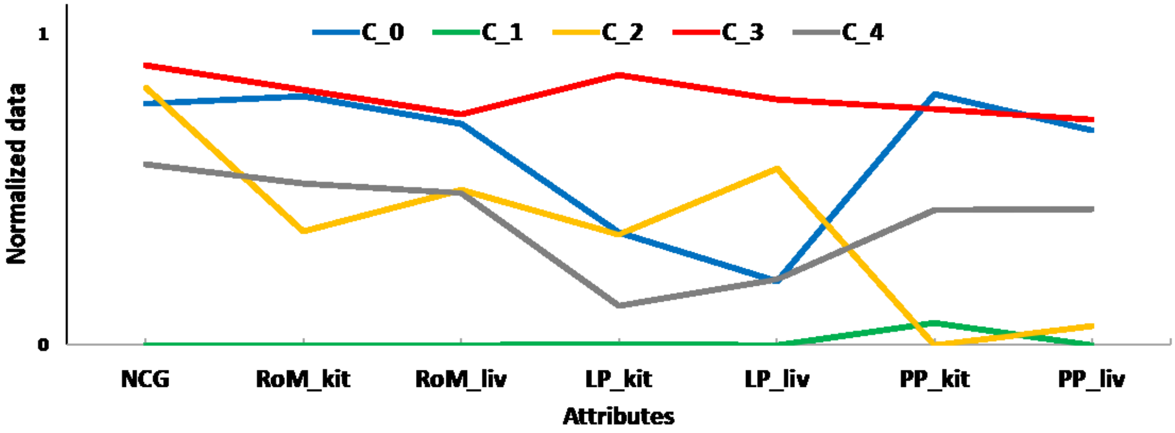


Figure 100. Visualization of cluster centroids

- **Clusters' specifications:**

Cluster_0: This cluster represents the moments that the level of activity and plug power consumption by occupants in zone four is very high. However, the lighting power consumption is below the average.

Cluster_1: This cluster represents the moments that the level of activity and energy consumption by occupants are at the lowest level. In fact, except for the plug power consumption of the kitchen, which is very low, all other attributes are the lowest in the dataset.

Cluster_2: This cluster represents the moments that the people produce CO_2 much more than average. Lighting power consumption is average, and plug power usage is fairly at the lowest place.

Cluster_3: This cluster represents the moments that the level of activity and energy consumption by occupants are at the highest level. In fact, except for the plug power consumption of the kitchen, which is very high, all other attributes are the highest in the dataset.

Cluster_4: This cluster represents the moments that both energy consumption and level of activity by occupants in zone four are around average. However, energy consumption in the living room is a little more than the kitchen's energy usage.

The RapidMiner procedure for performing the clustering step is similar to other zones and presented in appendix D.

4.6.3. Baseline step - Zone four (Kitchen & living room)

Similar to other zones, the steps are as follow:

i) **Calculation of *NNEC*:**

Out of 18 end-uses in zone four, 17 of them are considered as non-steady end-use. Therefore, another time equation 8 is employed to generate *NEC* in the dataset by summing up lightings power consumption and plugs power consumption.

Thus, for zone four equation 8 become:

$$\begin{aligned}\mathrm{NEC}_{\mathrm{zone}_4} = \mathrm{LP}_{\mathrm{kit1}} + \mathrm{LP}_{\mathrm{kit2}} + \mathrm{LP}_{\mathrm{kit3}} + \mathrm{LP}_{\mathrm{liv1}} + \mathrm{LP}_{\mathrm{liv2}} + \mathrm{PP}_{\mathrm{kit2}} \\ + \mathrm{PP}_{\mathrm{kit3}} + \mathrm{PP}_{\mathrm{kit4}} + \mathrm{PP}_{\mathrm{kit5}} + \mathrm{PP}_{\mathrm{kit6}} + \mathrm{PP}_{\mathrm{kit7}} + \mathrm{PP}_{\mathrm{liv1}} \\ + \mathrm{PP}_{\mathrm{liv2}} + \mathrm{PP}_{\mathrm{liv3}} + \mathrm{PP}_{\mathrm{liv4}} + \mathrm{PP}_{\mathrm{liv5}} + \mathrm{PP}_{\mathrm{liv6}}\end{aligned}$$

Equation 20

where:

LP_{kit1_3} are the three lightings power consumption in the kitchen

LP_{liv1_2} are the two lightings power consumption in the living room

PP_{kit2_7} are the six plugs power consumption in the kitchen

PP_{liv1_6} are the six plugs power consumption in the living room

$$\mathrm{NNEC}_{\mathrm{zone}_4} = \mathrm{normalized}\ (\mathrm{NEC}_{\mathrm{zone}_4})$$

Equation 21

ii) Calculation of *OAI*:

Equation 3 (section 3.1) is used to calculate $\mathrm{OAI}_{\mathrm{zone4}}$:

$$\mathrm{OAI}_{\mathrm{zone4}} = \sqrt{\mathrm{RoM}^2 + \mathrm{NCG}^2}$$

Equation 22

iii) Calculation of *REII*:

For each cluster, the *REII* attribute for zone four is generated by equation 23:

$$\mathrm{REII}_{\mathrm{zone}_4} = \frac{\mathrm{NNEC}_{\mathrm{zone}_4}}{\mathrm{OAI}_{\mathrm{zone}_4}}$$

Equation 23

iv) Target REII for zone four:

Figure 101 represent the baseline (target hourly REII) for the five defined clusters in zone four.

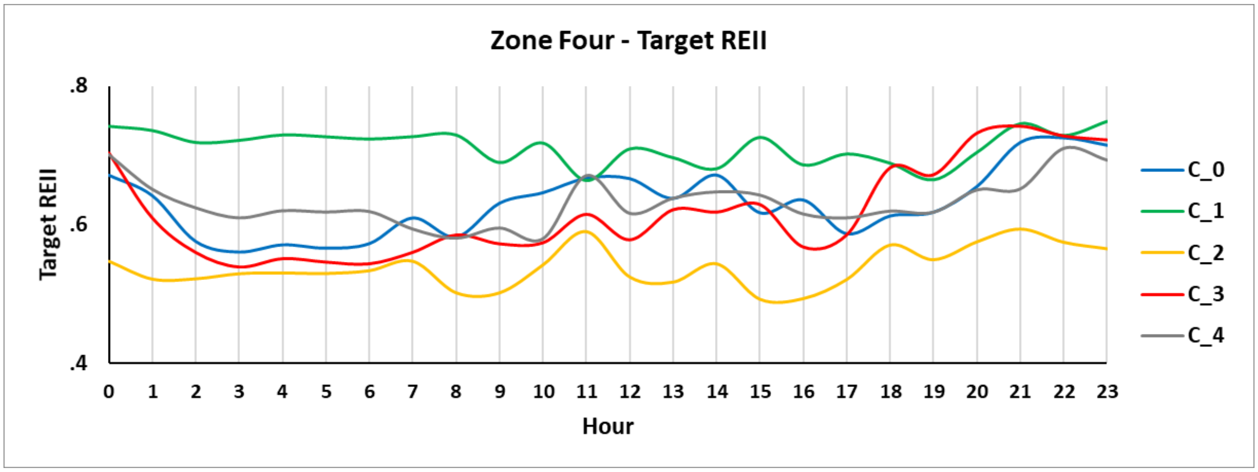
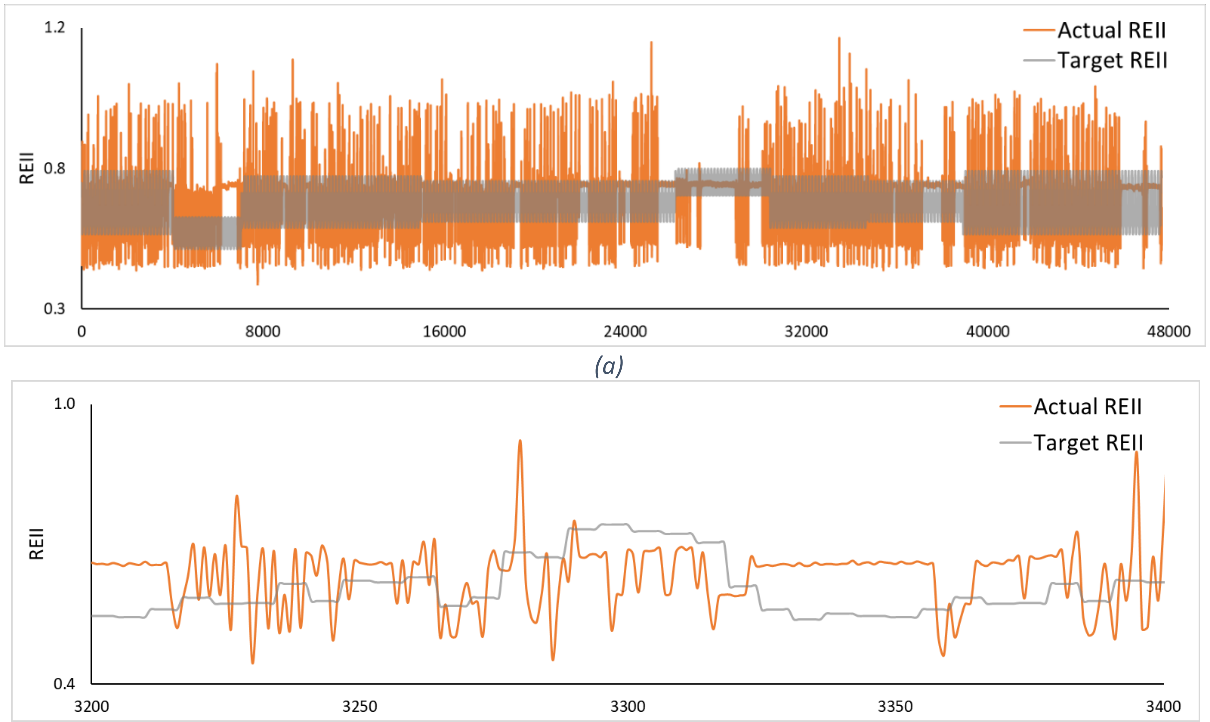


Figure 101. Target hourly REII for the five defined clusters

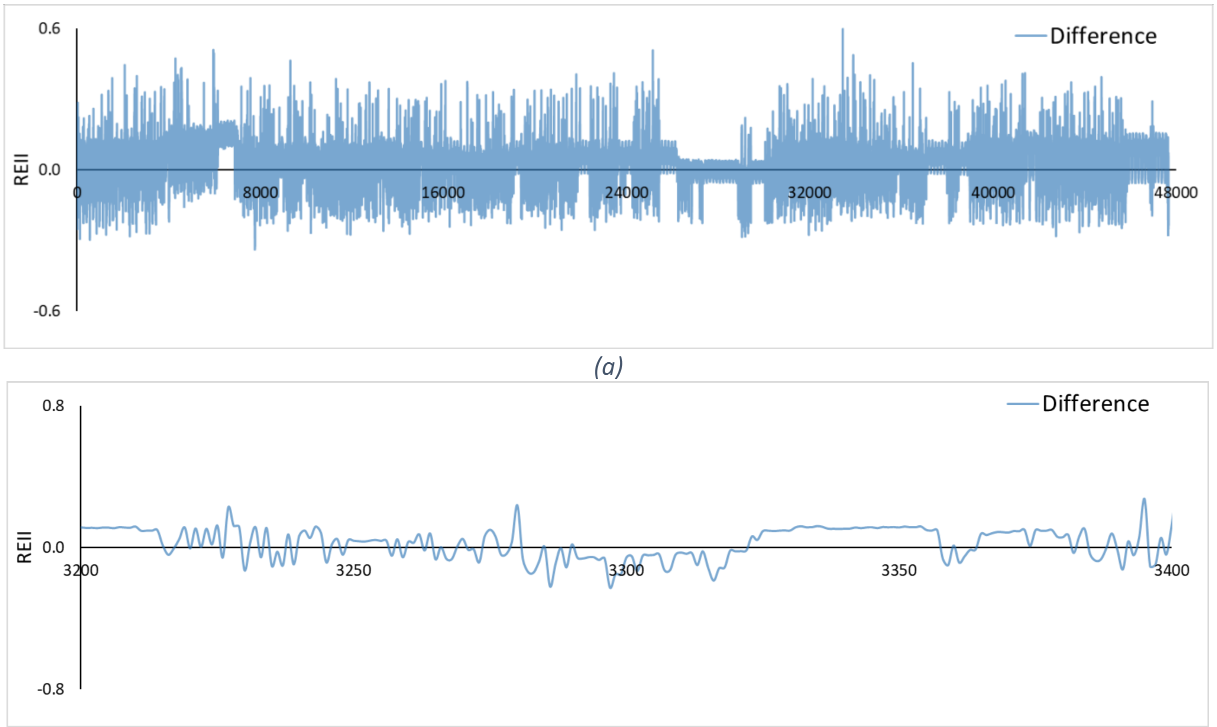
4.6.4. Energy wastage identification step - Zone four (Kitchen & living room)

Figure 102.a shows both the actual REII and the target one (baseline). Figure 102.b highlights a part of Figure 102.a for better understanding. Figure 103 shows this difference between the actual REII and the baseline.



(b)

Figure 102. Actual REII and target REII for the whole year. (a) whole year data, (b) zoom a part of the chart for better understanding



(a)

(b)

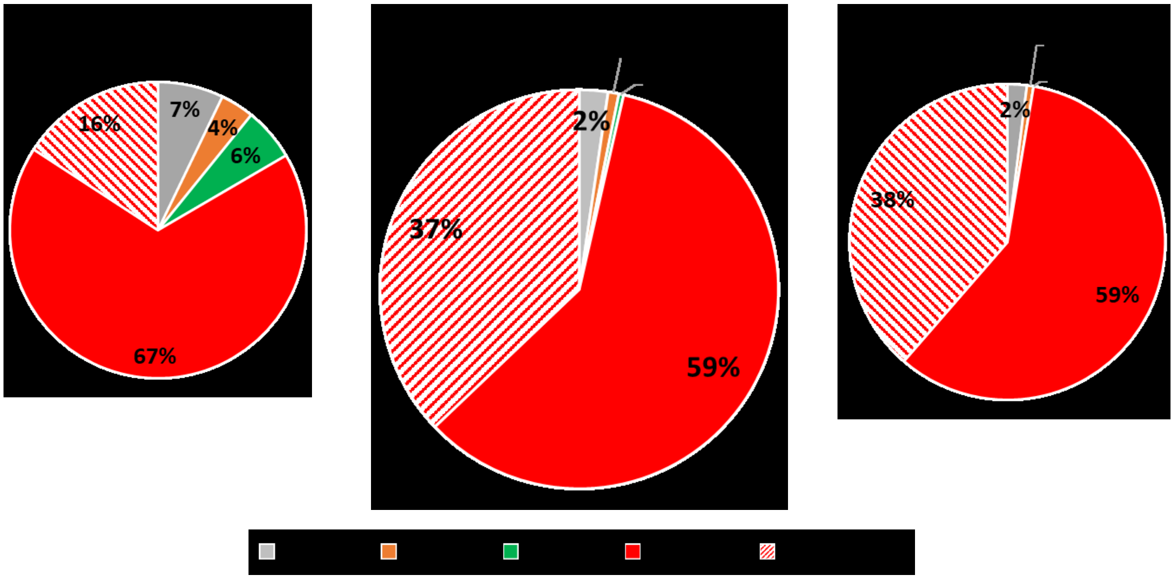
Figure 103. Difference between the actual REII and target REII for the whole year – Potential energy wastage/saving. (a) whole year data, (b) zoom a part of the chart for better understanding

4.7. Summary

Figure 104 summarize the energy consumption by occupants in each zone of the apartment. As it is illustrated in the figure, the contribution of zone 4 to the total energy consumption in the apartment is considerably higher than in other zones. However, as it is shown in Figures 105, 107, and 108 and Table 8, occupants' energy-related behavior in this zone is more efficient than zones 2 and 3.

Figure 105 shows the monthly variation of normalized EBI[1] for all zones of the apartment. When EBI is more than zero, it shows that there is a possible wasteful-behavior by occupants. Also, Table 8 shows the location of the worst and the best energy-related behavior by occupants in different months.

[1] EBI: Energy-related Behavior Index



a. lighting power consumption *c. Non-steady Energy consumption* *b. Plug power consumption*

Figure 104. contribution of different zones in total energy consumption in the apartment

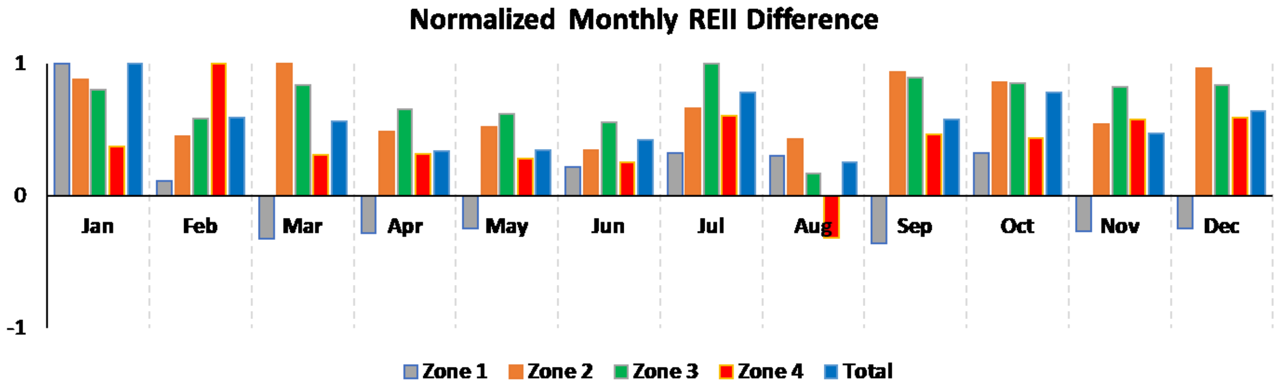


Figure 105. Normalized monthly energy-wastage/saving in different zones

Table 8. The monthly place of the worst and the best energy-related behavior by occupants

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
the worst energy-related behavior	Zone 1	Zone 4	Zone 2	Zone 3				Zone 2			Zone 3	Zone 2
the best energy-	Zone 4	Zone 1						Zone 4	Zone 1			

related behavior	🟥	▨	🟥	▨

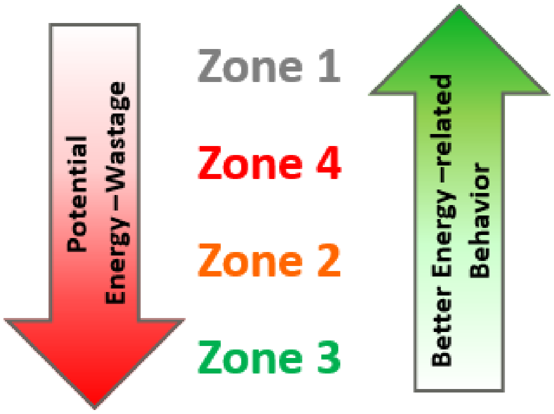


Figure 106. Variation of energy-related behavior of the occupants in different zones of the apartment

As it is illustrated in Figure 106, occupants need to pay more attention to their energy-related behavior in zone 2 and zone 3. Also, occupants behave more efficient in zone 1 than in other zones.

Figures 107 and 108 show the daily variation and hourly variation of EBI in every zone of the apartment. Regarding Figure 107, occupants wasteful-behavior on Thursdays and Weekends is more than other weekdays. Also, as it is illustrated in Figure 108, occupants show more efficient energy-related behavior in the evenings than other times of the day.

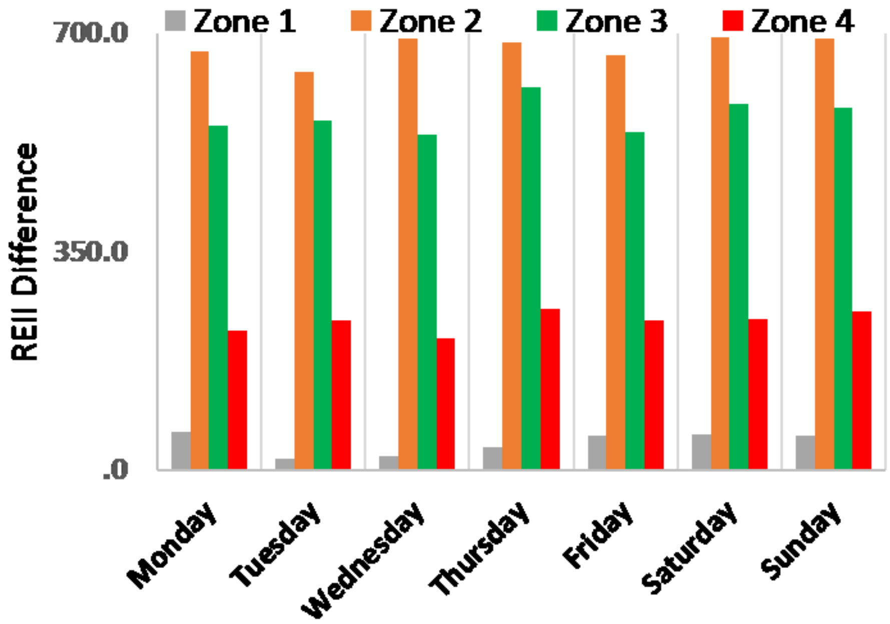


Figure 107. Total daily REII difference in different zones

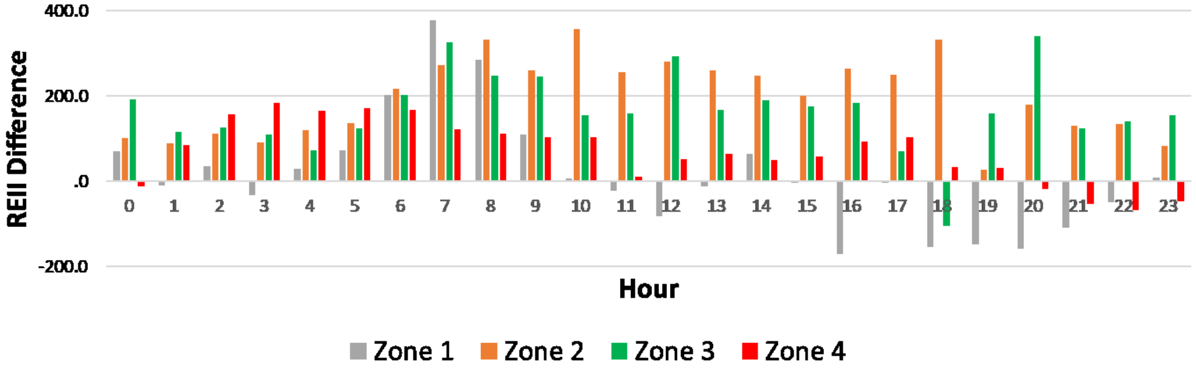


Figure 108. Total hourly REII difference in different zones

EBI is calculated for the whole year data for all zones. The monthly, daily, and hourly variations of EBI are illustrated in figures 109, 110, and 111, respectively. As a result, we are able to give occupants practical feedback about their energy-related behavior.

Aug Apr May Jun Nov Mar Sep Feb Dec Oct Jul Jan

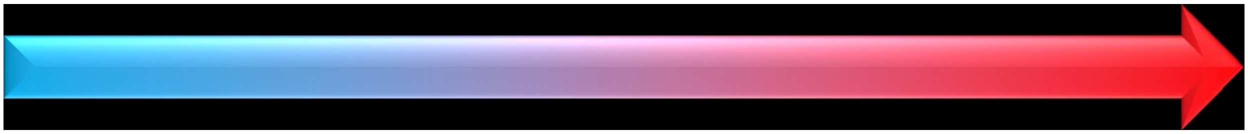


Figure 109. Monthly variation of EBI

(Blue: the best energy-related behavior – Red: the worst energy-related behavior

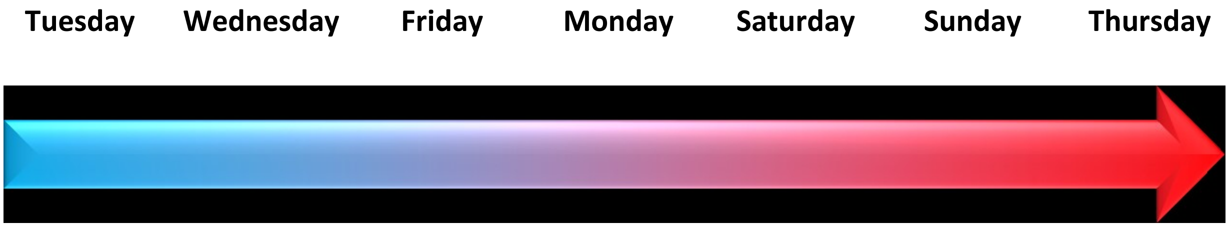


Figure 110. Daily variation of EBI

(Blue: the best energy-related behavior – Red: the worst energy-related behavior

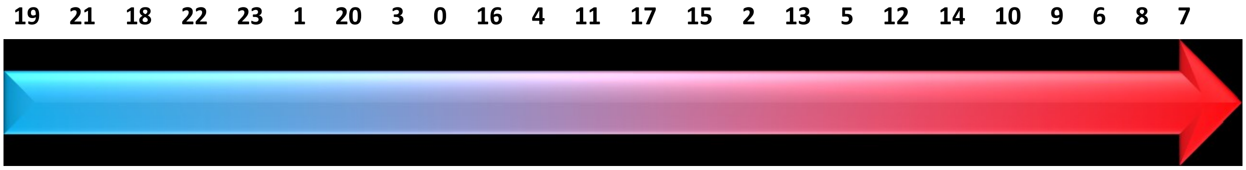


Figure 111. Hourly variation of EBI

(Blue: the best energy-related behavior – Red: the worst energy-related behavior

5. Conclusion

In this study, a new data mining-based methodology is developed to evaluate energy-related behavior of occupants in residential buildings. An introduction about energy-related data, as well as the research objectives, are presented in chapter one. Chapter two reviewed the literature on the application of data mining in building energy. In section 2.4, a summary of the literature review, as well as the challenges and the gap in research, are discussed.

Chapter three represents the developed methodology in detail. In sections 3.1 and 3.2 Occupant Activity Indicator (OAI) and Residential Energy Intensity Indicator (REII) are introduced as two new definitions which are used in this study. The proposed methodology to evaluate the energy-related behavior of the buildings' residents is based on the difference between the target REII and actual REII. The dissimilarity, which is found between the target and the actual REII, can be used to calculate the potential energy wastage/saving by occupants in different zones and different times in the building.

The practicality of the proposed data mining framework is evaluated in chapter four. In this chapter, the developed methodology is applied to a one-year dataset collected in a three-bedroom apartment in Lyon, France. The methodology applied to all zones of the apartment to evaluate the occupants' energy-related behavior. As a result, the time and location for potential energy savings by occupants is identified. The obtained results are summarized in section 4.7.

The results show that occupants need to be more cautious about their energy consumption in zones 2 and 3. Moreover, the possible energy-wastage behavior in zones 1 and 4 is less than zones 2 and 3, even though the contribution of zone 4 to the energy consumption is significantly higher than the other zones. Besides, by the developed methodology location and time for the best and the worst energy-related behavior by the building's occupants are defined. Furthermore, the variations of occupants' energy-related behavior in the apartment, are identified by time of day, day of week, and months.

Employing the proposed methodology is beneficial for buildings’ occupants to raise their awareness regarding energy consumption. Also, it gives the decision-makers a practical insight into the system behavior, enabling them to create incentives/charges for residential buildings’ inhabitants to modify their energy-related behavior.

References

[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, “A review on buildings energy consumption information,” *Energy Build.*, vol. 40, no. 3, pp. 394–398, Jan. 2008.

[2] P. Nejat, F. Jomehzadeh, M. M. Taheri, M. Gohari, and M. Z. Abd. Majid, “A global review of energy consumption, CO2 emissions and policy in the residential sector (with an overview of the top ten CO2 emitting countries),” *Renew. Sustain. Energy Rev.*, vol. 43, pp. 843–862, Mar. 2015.

[3] M. Dardir, K. Panchabikesan, F. Haghighat, M. El Mankibi, and Y. Yuan, “Opportunities and challenges of PCM-to-air heat exchangers (PAHXs) for building free cooling applications—A comprehensive review,” *J. Energy Storage*, vol. 22, pp. 157–175, Apr. 2019.

[4] S. M.R. Khani, M. N. Bahadori, A. Dehghani-Sanij, and A. Nourbakhsh, “Performance evaluation of a modular design ofwind tower withwetted surfaces,” *Energies*, vol. 10, no. 7, 2017.

[5] M. N. Bahadori, A. Dehghani-sanij, and A. Sayigh, *Wind Towers: Architecture, Climate and Sustainability*. Springer International Publishing, 2014.

[6] S. M.R.Khani, M. N. Bahadori, and A. R. Dehghani-Sanij, “Experimental investigation of a modular wind tower in hot and dry regions,” *Energy Sustain. Dev.*, vol. 39, pp. 21–28, Aug. 2017.

[7] A. R. Dehghani-sanij, M. Soltani, and K. Raahemifar, “A new design of wind tower for passive ventilation in buildings to reduce energy consumption in windy regions,” *Renew. Sustain. Energy Rev.*, vol. 42, pp. 182–195, Feb. 2015.

[8] N. B. Hutcheon and G. O. P. Handegord, *Building science for a cold climate PB*, Third. Ottawa, Ontario: Institute for Research in Construction, 1995.

[9] F. C. McQuiston, J. D. Parker, and J. D. Spitler, *Heating, ventilating, and air conditioning : analysis and design*, Sixth. John Wiley & Sons, 2005.

[10] N. B. Hutcheon and G. O. Handegord, *Building science for a cold climat*. Toronto: Institute for Research in Construction (Canada), 1995.

[11] J. Li, K. Panchabikesan, Z. Yu, F. Haghighat, M. El Mankibi, and D. Corgier, “Systematic data mining-based framework to discover potential energy waste patterns in residential buildings,” *Energy Build.*, vol. 199, pp. 562–578, Sep. 2019.

[12] H. Zhao and F. Magoulès, “A review on the prediction of building energy consumption,” *Renew. Sustain. Energy Rev.*, vol. 16, no. 6, pp. 3586–3592, Aug. 2012.

[13] Y. Zhang and P. Barrett, “Factors influencing the occupants’ window opening behaviour in a naturally ventilated office building,” *Build. Environ.*, vol. 50, pp. 125–134, Apr. 2012.

[14] S. Wei, R. Jones, and P. De Wilde, “Driving factors for occupant-controlled space heating in residential buildings,” *Energy Build.*, vol. 70, pp. 36–44, 2014.

[15] S. D’Oca and T. Hong, “Occupancy schedules learning process through a data mining framework,” *Energy Build.*, vol. 88, pp. 395–408, Feb. 2015.

[16] T. Hong, S. C. Taylor-Lange, S. D’Oca, D. Yan, and S. P. Corgnati, “Advances in research and applications of energy-related occupant behavior in buildings,” *Energy Build.*, vol. 116, pp. 694–702, Mar. 2016.

[17] Z. Chen and Y. C. Soh, “Comparing occupancy models and data mining approaches for regular occupancy prediction in commercial buildings,” *J. Build. Perform. Simul.*, vol. 10, no. 5–6, pp. 545–553, 2017.

[18] C. Duarte, K. Van Den Wymelenberg, and C. Rieger, “Revealing occupancy patterns in an office building through the use of occupancy sensor data,” *Energy Build.*, vol. 67, pp. 587–595, Dec. 2013.

[19] A. Al-Mumin, O. Khattab, and G. Sridhar, “Occupants’ behavior and activity patterns influencing the energy consumption in the Kuwaiti residences,” *Energy Build.*, vol. 35, no. 6, pp. 549–559, Jul. 2003.

[20] J. Zhao, B. Lasternas, K. P. Lam, R. Yun, and V. Loftness, “Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining,” *Energy Build.*, vol. 82, pp. 341–355, Oct. 2014.

[21] T. Hong and H. Lin, “Occupant Behavior : Impact on Energy Use of Private Offices,” *Asim IBSPA Asia Conf.*, no. January, 2013.

[22] Z. Yu, F. Haghighat, and B. C. M. Fung, “Advances and challenges in building engineering and data mining applications for energy-efficient communities,” *Sustain. Cities Soc.*, vol. 25, pp. 33–38, 2016.

[23] J. Ouyang and K. Hokao, “Energy-saving potential by improving occupants’ behavior in urban residential sector in Hangzhou City, China,” *Energy Build.*, vol. 41, no. 7, pp. 711–720, Jul. 2009.

[24] O. T. Masoso and L. J. Grobler, “The dark side of occupants’ behaviour on building energy use,” *Energy Build.*, vol. 42, no. 2, pp. 173–177, Feb. 2010.

[25] Z. (Jerry) Yu, F. Haghighat, B. C. M. Fung, and L. Zhou, “A novel methodology for knowledge discovery through mining associations between building operational data,” *Energy Build.*, vol. 47, pp. 430–440, Apr. 2012.

[26] Z. (Jerry) Yu, F. Haghighat, B. C. M. Fung, E. Morofsky, and H. Yoshino, “A methodology for identifying and improving occupant behavior in residential buildings,” *Energy*, vol. 36, no. 11, pp. 6596–6608, Nov. 2011.

[27] M. Ashouri, F. Haghighat, B. C. M. Fung, A. Lazrak, and H. Yoshino, “Development of building energy saving advisory: A data mining approach,” *Energy Build.*, vol. 172, pp. 139–151, Aug. 2018.

[28] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, “A decision tree method for building

energy demand modeling," *Energy Build.*, vol. 42, no. 10, pp. 1637–1646, 2010.

[29] Z. Yu, B. C. M. Fung, F. Haghighat, H. Yoshino, and E. Morofsky, "A systematic procedure to study the influence of occupant behavior on building energy consumption," *Energy Build.*, vol. 43, no. 6, pp. 1409–1417, Jun. 2011.

[30] Z. Yu, B. C. M. Fung, and F. Haghighat, "Extracting knowledge from building-related data - A data mining framework," *Build. Simul.*, vol. 6, no. 2, pp. 207–222, 2013.

[31] J. Li, K. Panchabikesan, Z. Yu, F. Haghighat, M. El Mankibi, and D. Corgier, "Systematic data mining-based framework to discover potential energy waste patterns in residential buildings," *Energy Build.*, vol. 199, pp. 562–578, Sep. 2019.

[32] US Department of Energy, "Building Energy Software Tools Directory." [Online]. Available: https://www.energy.gov/node/3028339.

[33] D. J. Hand, *Principles of Data Mining*, vol. 30, no. 7. 2007.

[34] C. Fan, F. Xiao, and C. Yan, "A framework for knowledge discovery in massive building automation data and its application in building diagnostics," *Autom. Constr.*, vol. 50, pp. 81–90, Feb. 2015.

[35] X. Liang, T. Hong, and G. Q. Shen, "Occupancy data analytics and prediction: A case study," *Build. Environ.*, vol. 102, pp. 179–192, Jun. 2016.

[36] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[37] A. Ng, "Machine Learning by Stanford University," 2012. .

[38] R. Kumar, R. K. Aggarwal, and J. D. Sharma, "Energy analysis of a building using artificial neural network: A review," *Energy Build.*, vol. 65, pp. 352–358, Oct. 2013.

[39] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[40] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees," *Ecology*, vol. 81, pp. 3178–3192, 1984.

[41] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[42] S. L. Salzberg, “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993,” *Mach. Learn.*, vol. 16, no. 3, pp. 235–240, Sep. 1994.

[43] A. H. Neto and F. A. S. Fiorelli, “Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption,” *Energy Build.*, vol. 40, no. 12, pp. 2169–2176, Jan. 2008.

[44] A. Azadeh, S. F. Ghaderi, and S. Sohrabkhani, “Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors,” *Energy Convers. Manag.*, vol. 49, no. 8, pp. 2272–2278, Aug. 2008.

[45] Q. Li, P. Ren, and Q. Meng, “Prediction Model of Annual Energy Consumption of Residential Buildings Qiong Li, Peng Ren, Qinglin Meng,” *2010 Int. Conf. Adv. Energy Eng. Predict.*, pp. 223–226, 2010.

[46] G. K. F. Tso and K. K. W. Yau, “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks,” *Energy*, vol. 32, no. 9, pp. 1761–1768, Sep. 2007.

[47] C. Fan, F. Xiao, and S. Wang, “Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques,” *Appl. Energy*, vol. 127, pp. 1–10, 2014.

[48] R. Jovanović, A. A. Sretenović, and B. D. Živković, “Ensemble of various neural networks for prediction of heating energy consumption,” *Energy Build.*, vol. 94, pp. 189–199, 2015.

[49] S. D’Oca and T. Hong, “A data-mining approach to discover patterns of window opening and closing behavior in offices,” *Build. Environ.*, vol. 82, pp. 726–739, Dec. 2014.

[50] ASHRAE 90.1, “Standard 90.1-2004,” 2004.

[51] ASHRAE 90.1, “Standard 90.1-2004,” 1989.

[52] B. Huchuk, S. Sanner, and W. O’Brien, “Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data,” *Build. Environ.*, vol. 160, no. March, p. 106177, 2019.

[53] Z. Li and B. Dong, “A new modeling approach for short-term prediction of occupancy in residential buildings,” *Build. Environ.*, vol. 121, pp. 277–290, 2017.

[54] F. Causone, S. Carlucci, M. Ferrando, A. Marchenko, and S. Erba, “A data-driven procedure to model occupancy and occupant-related electric load profiles in residential buildings for energy simulation,” *Energy Build.*, vol. 202, p. 109342, 2019.

[55] J. Jazaeri, R. L. Gordon, and T. Alpcan, “Influence of building envelopes, climates, and occupancy patterns on residential HVAC demand,” *J. Build. Eng.*, vol. 22, no. July 2018, pp. 33–47, 2019.

[56] B. C. M. Fung, *Data mining by McGill University*. 2019.

[57] RapidMiner Docs, “RapidMiner Docs.” [Online]. Available: https://docs.rapidminer.com.

[58] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[59] Ministry of Ecology, “Department of Sustainable Development and Energy Website.” [Online]. Available: http://www.meteofrance.com/climat.

Appendixes

Appendix A.

```
root_mean_squared_error: 0.647 +/- 0.158 (micro average: 0.664 +/- 0.000)
```

DeepLearning

```
Model Metrics Type: Regression
 Description: Metrics reported on temporary training frame with 10083
samples
 model id: rm-h2o-model-deep_learning-562295
 frame id: rm-h2o-frame-deep_learning-653599.temporary.sample.28.58%
 MSE: 0.25526237
 R^2: 0.98830163
 mean residual deviance: 0.25526237
Status of Neuron Layers (predicting PP_KITCHEN_ZTPG001E5E09020469D3,
regression, gaussian distribution, Quadratic loss, 5,651 weights/biases,
78.6 KB, 361,911 training samples, mini-batch size 1):
 Layer Units      Type Dropout       L1       L2 Mean Rate Rate RMS
Momentum Mean Weight Weight RMS Mean Bias Bias RMS
     1    60      Input  0.00 %
     2    50 Rectifier  0.00 % 0.000010 0.000000  0.029524 0.068504
0.000000   -0.003628   0.118159  0.525673 0.179313
     3    50 Rectifier  0.00 % 0.000010 0.000000  0.073157 0.176276
0.000000   -0.003057   0.134128  0.940528 0.115274
     4     1    Linear         0.000010 0.000000  0.003273 0.003571
0.000000    0.046880   0.208012  0.004606 0.000000
Scoring History:
           Timestamp   Duration Training Speed   Epochs Iterations
Samples Training MSE Training Deviance Training R^2
 2019-04-17 14:45:33  0.000 sec                 0.00000          0
0.000000          NaN               NaN          NaN
 2019-04-17 14:45:35  2.739 sec  7738 rows/sec  0.51199          1
17915.000000     18.34218          18.34218      0.15940
 2019-04-17 14:45:42  9.833 sec  7798 rows/sec  2.06262          4
72173.000000      7.73185           7.73185      0.64566
 2019-04-17 14:45:49 16.831 sec  7873 rows/sec  3.62053          7
126686.000000      1.25117           1.25117      0.94266
 2019-04-17 14:45:56 23.877 sec  7870 rows/sec  5.16930         10
180879.000000      0.79514           0.79514      0.96356
```

```
 2019-04-17 14:46:04 31.125 sec  7820 rows/sec  6.72119         13
235181.000000      0.98987           0.98987     0.95464
 2019-04-17 14:46:11 38.478 sec  7768 rows/sec  8.27364         16
289503.000000      0.33297           0.33297     0.98474
 2019-04-17 14:46:16 43.550 sec  7721 rows/sec  9.30705         18
325663.000000      0.25526           0.25526     0.98830
 2019-04-17 14:46:21 48.668 sec  7678 rows/sec 10.34297         20
361911.000000      0.28923           0.28923     0.98674
 2019-04-17 14:46:21 48.829 sec  7677 rows/sec 10.34297         20
361911.000000      0.25526           0.25526     0.98830

H2O version: 3.8.2.6-rm9.0.0
```

Appendix B.

Zone 1: Bedroom 1

	Sensor code		Sensor code
CO2	KTCO1145	WINDOW SHADE	KBLD1107.00_1
TEMPERATURE	KTCO1145_1	WINDOW OPEN/CLOSE	KOCL1140.00
RELATIVE HUMIDITY	KTCO1145_2	LIGHTING POWER	KLGT1102.00
THERMOSTAT SET POINT	KTMS1148	LIGHT ON/OFF	KLGT1102.00_2
LUX	KMVL1209	PLUG POWER	ZTPG001E5E090200468F
MOTION	KMVL1209_1		ZTPG001E5E0902004697
WINDOW BLIND	KBLD1107.00		ZTPG001E5E09020048D4

Zone 2: Bedroom 2

	Sensor code		Sensor code
CO2	KTCO1146	WINDOW SHADE	KBLD1107.01_1
TEMPERATURE	KTCO1146_1		KBLD1107.02_1
RELATIVE HUMIDITY	KTCO1146_2	WINDOW OPEN/CLOSE	KOCL1141.00
THERMOSTAT SET POINT	KTMS1149		KOCL1142.00
LUX	KMVL120A	LIGHTING POWER	KLGT1103.01
MOTION	KMVL120A_1	LIGHT ON/OFF	KLGT1103.01_2
WINDOW BLIND	KBLD1107.01	PLUG POWER	ZTPG001E5E0902004B03
	KBLD1107.02		

Zone 3: Bedroom 3

	Sensor code		Sensor code
CO2	KTCO1147	WINDOW BLIND	KBLD1107.03
TEMPERATURE	KTCO1147_1	WINDOW SHADE	KBLD1107.03_1
RELATIVE HUMIDITY	KTCO1147_2	WINDOW OPEN/CLOSE	KOCL1143.00
THERMOSTAT SET POINT	KTMS114A	LIGHTING POWER	KLGT1104.00
LUX	KMVL120B	LIGHT ON/OFF	KLGT1104.00_2
MOTION	KMVL120B_1		

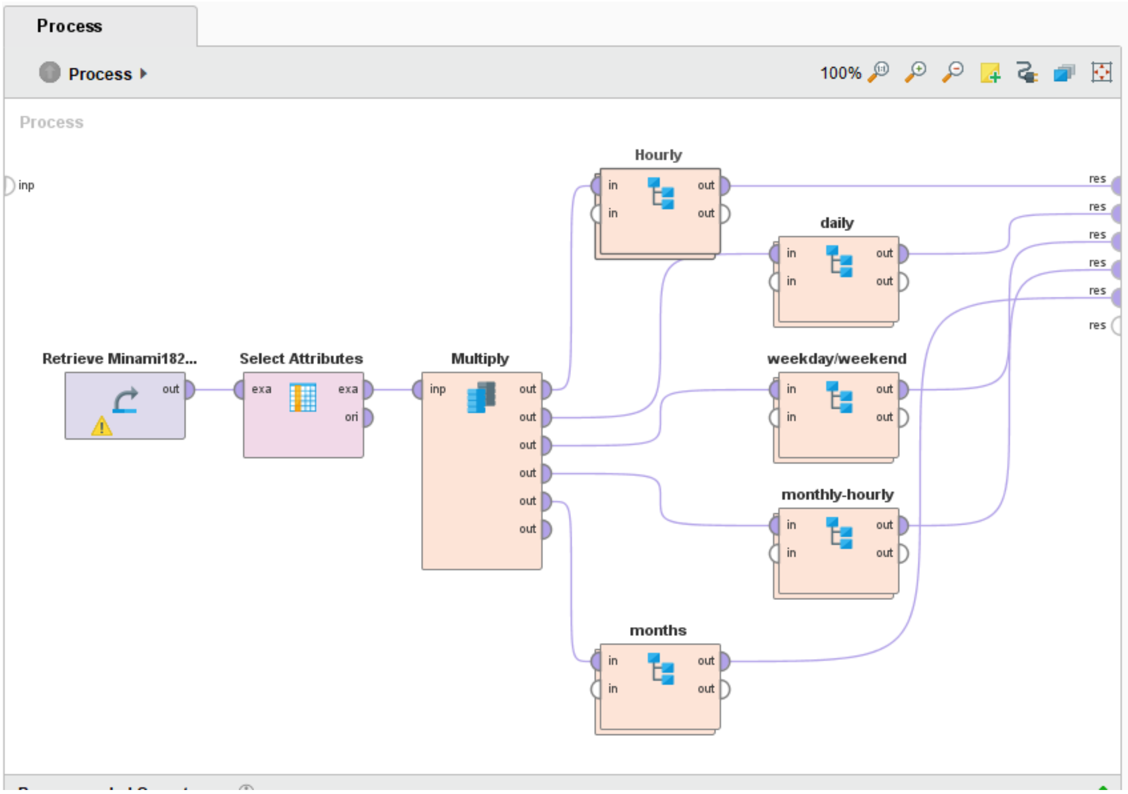
Zone 4: Kitchen and Living room

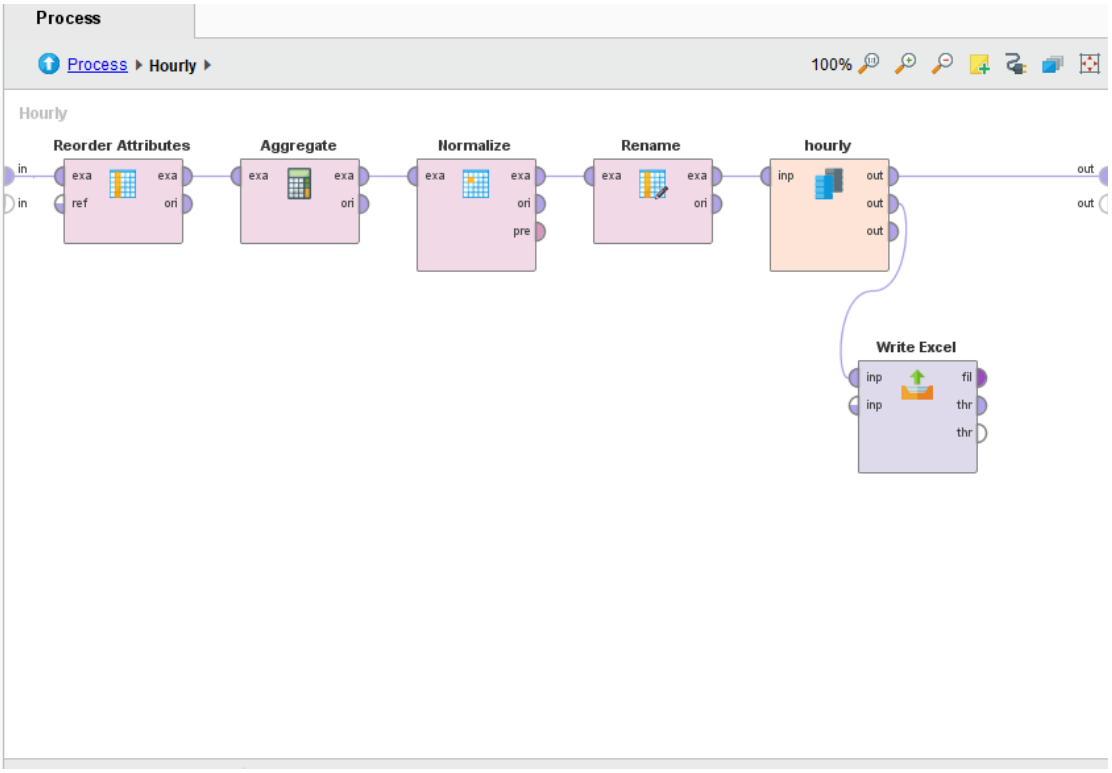
	Sensor code		Sensor code
CO2	KTCO1144	THERMOSTAT SET POINT	KTMS1152
RELATIVE HUMIDITY	KTCO1144_2	LIGHTING POWER	KLGT1101.01
Temperature	TEMPERATURE KTCO1144_1		KLGT1101.02
LUX	KMVL1203		KLGT1105.01
	KMVL1201		KLGT1101.00
	KMVL1202		KLGT1105.00
MOTION	KMVL1203_1	LIGHT ON/OFF	KLGT1101.01_2
	KMVL1201_1		KLGT1101.02_2
	KMVL1202_1		KLGT1105.01_2
WINDOW BLIND	KBLD1106.04		KLGT1101.00_2
	KBLD1106.05		KLGT1105.00_2

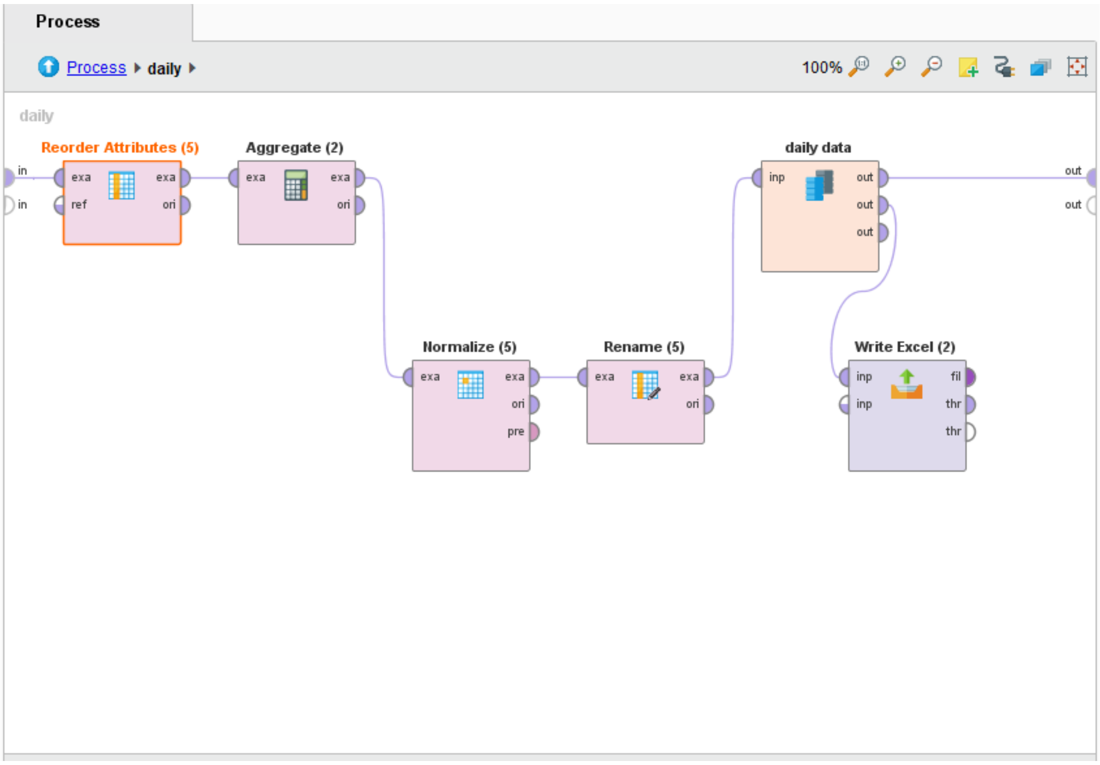
	KBLD1106.00	PLUG POWER	ZTPG001E5E0902004755
	KBLD1106.01		ZTPG001E5E09020047A7
	KBLD1106.02		ZTPG001E5E09020047F9
	KBLD1106.03		ZTPG001E5E0902004999
WINDOW SHADE	KBLD1106.04_1		ZTPG001E5E0902004B26
	KBLD1106.05_1		ZTPG001E5E09020469D3
	KBLD1106.00_1		ZTPG001E5E0902046DD5
	KBLD1106.01_1		ZTPG001E5E0902004885
	KBLD1106.02_1		ZTPG001E5E0902004AFC
	KBLD1106.03_1		ZTPG001E5E0902004B00
Window Open/Close	KOCL113D.00		ZTPG001E5E0902004B02
	KOCL113E.00		ZTPG001E5E0902004B1A
	KOCL113F.00		ZTPG001E5E0902046A02

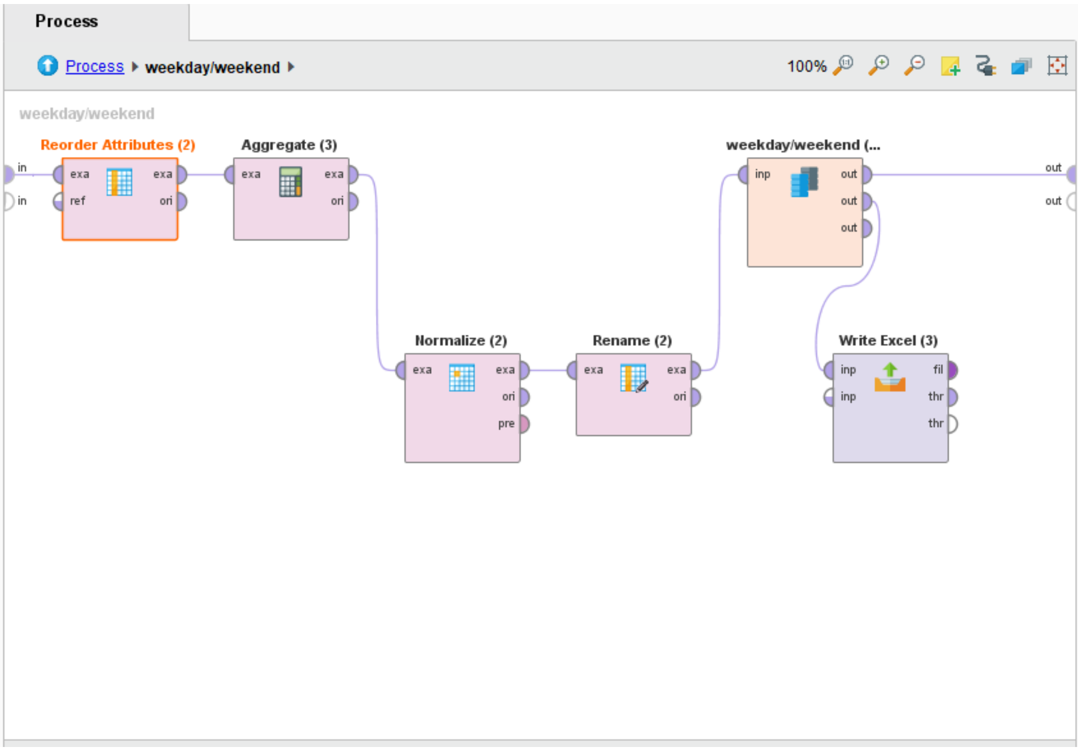
Appendix C.

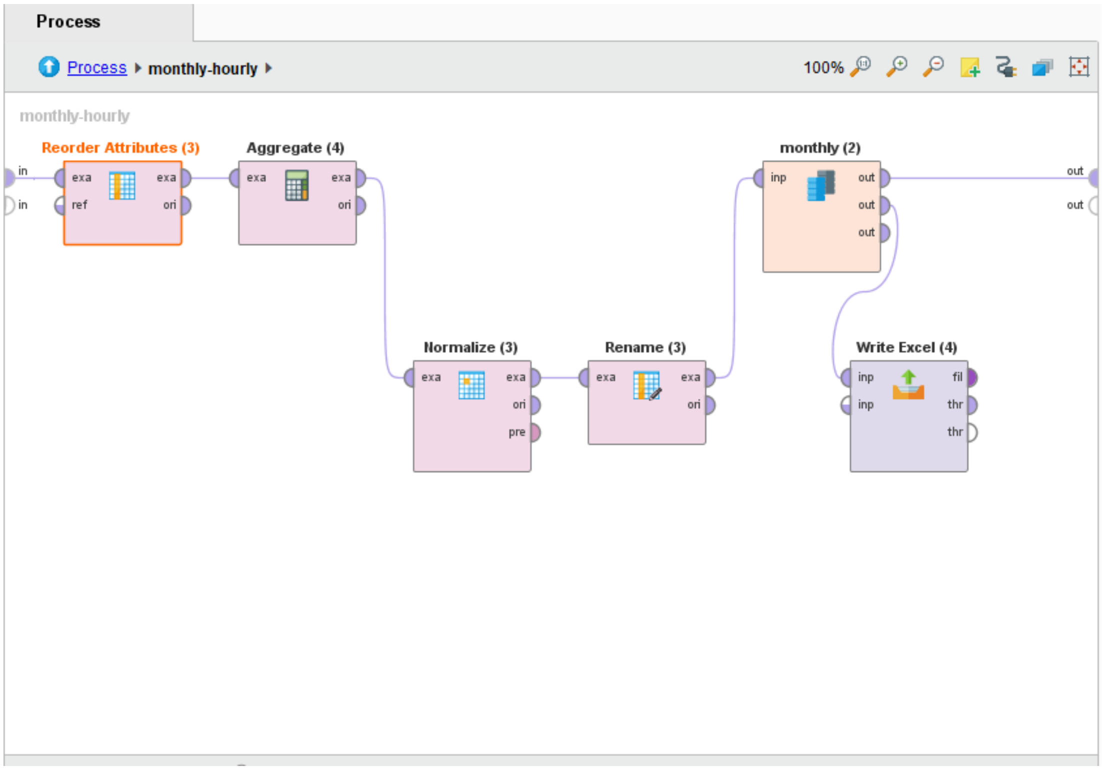
RapidMiner procedure to investigate the patterns of occupant activity and energy consumption in the building. The main process and all sub-process are presented in the following images.

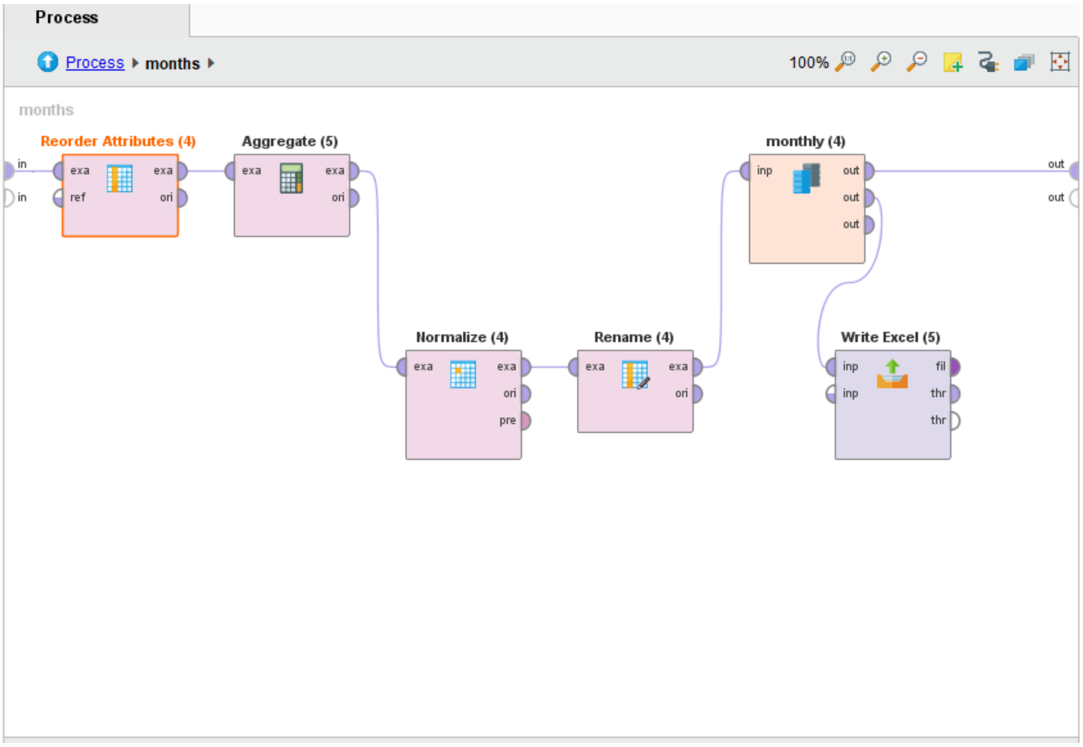

Process
Process
100%
Process
inp
Hourly
daily
weekday/weekend
monthly-hourly
months
Retrieve Minami182...
Select Attributes
Multiply
in
out
exa
ori
res


Process
Process
Hourly
100%
Hourly
Reorder Attributes
Aggregate
Normalize
Rename
hourly
Write Excel
in
out
exa
ref
ori
pre
inp
fil
thr


Process
Process ▸ daily ▸
100%
daily
Reorder Attributes (5)
in
in
exa
ref
exa
ori
Aggregate (2)
exa
exa
ori
daily data
inp
out
out
out
out
out
Normalize (5)
exa
exa
ori
pre
Rename (5)
exa
exa
ori
Write Excel (2)
inp
inp
fil
thr
thr

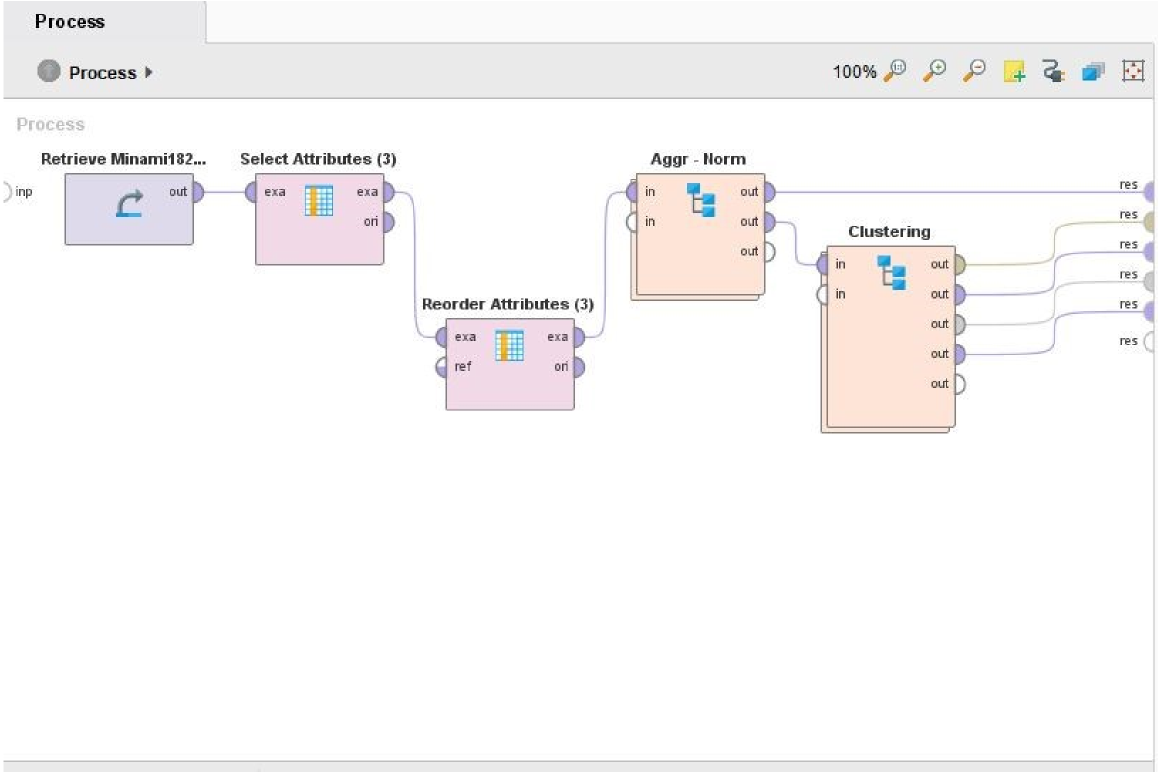

Process
Process ▸ weekday/weekend ▸
100%
weekday/weekend
Reorder Attributes (2)
in
in
exa
ref
exa
ori
Aggregate (3)
exa
exa
ori
weekday/weekend (...
inp
out
out
out
out
out
Normalize (2)
exa
exa
ori
pre
Rename (2)
exa
exa
ori
Write Excel (3)
inp
inp
fil
thr
thr

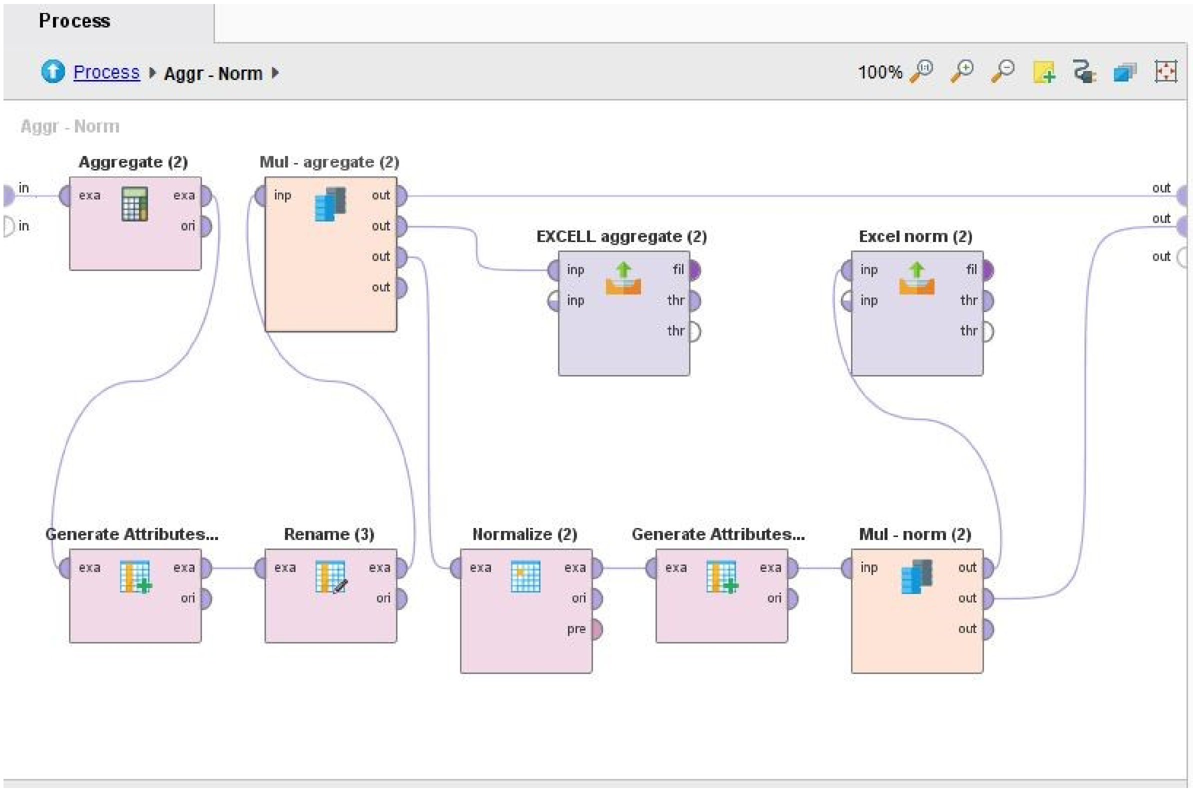

Process
Process ▸ monthly-hourly ▸
100%
monthly-hourly
Reorder Attributes (3)
Aggregate (4)
monthly (2)
Normalize (3)
Rename (3)
Write Excel (4)
in
in
exa
ref
exa
ori
exa
exa
ori
inp
out
out
out
out
out
exa
exa
ori
pre
exa
exa
ori
inp
inp
fil
thr
thr

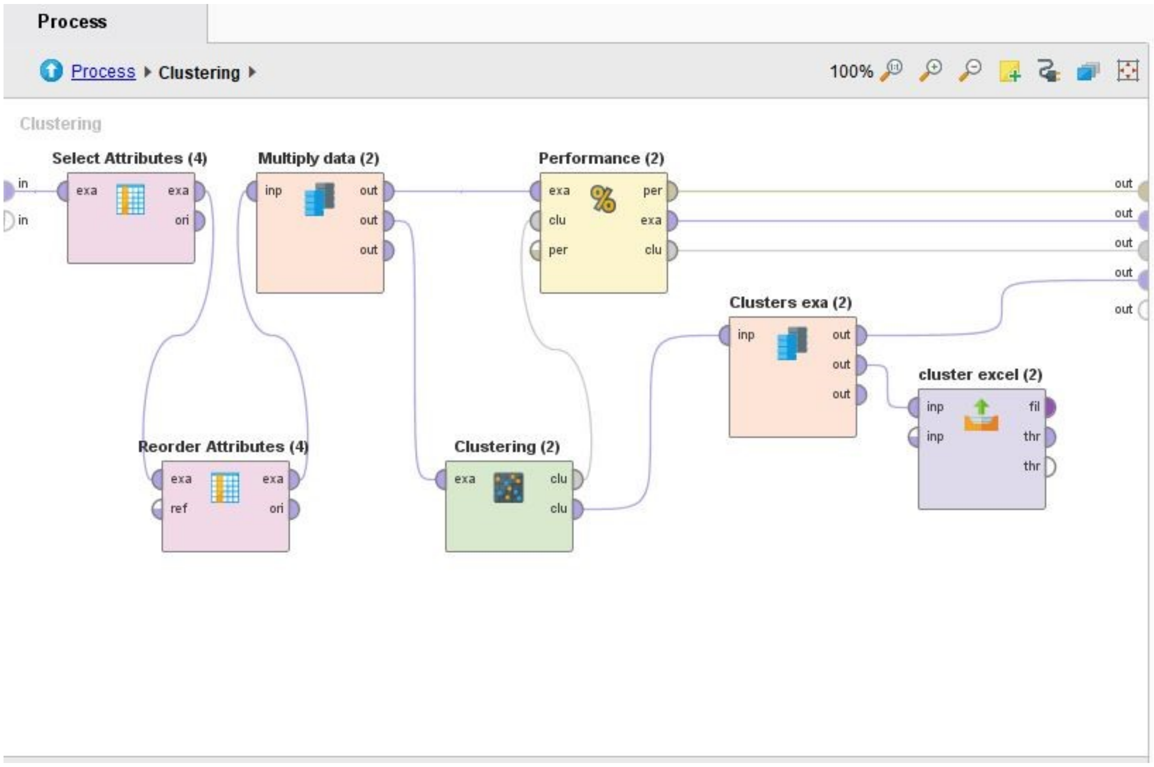

Process
Process ▸ months ▸
100%
months
Reorder Attributes (4)
Aggregate (5)
monthly (4)
Normalize (4)
Rename (4)
Write Excel (5)
in
in
exa
ref
exa
ori
exa
exa
ori
inp
out
out
out
out
out
exa
exa
ori
pre
exa
exa
ori
inp
inp
fil
thr
thr

Appendix D.

RapidMiner procedure to perform the clustering step. The main process and all sub-process are presented in the following images.




Process
Process ▸ Aggr - Norm ▸
100%
Aggr - Norm
Aggregate (2)
Mul - agregate (2)
EXCELL aggregate (2)
Excel norm (2)
Generate Attributes...
Rename (3)
Normalize (2)
Generate Attributes...
Mul - norm (2)
in
in
exa
ori
inp
out
fil
thr
pre
out


Process
Process ▸ Clustering ▸
100%
Clustering
Select Attributes (4)
Multiply data (2)
Performance (2)
Clusters exa (2)
cluster excel (2)
Reorder Attributes (4)
Clustering (2)
in
in
exa
ori
inp
out
clu
per
ref
fil
thr
out

Appendix E.

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation, and optimization. RapidMiner is developed on an open core model [57].

Advantages of RapidMiner:

i. **Visualization**: Easy to use visual environment for building analytics processes;

   Every analysis is a process, each transformation or analysis step is an operator, making design fast, easy to understand, and fully reusable;

   Convenient set of data exploration tools and visualizations;

   Wizards for Microsoft Excel & Access, CSV, and database connections;

   Repository-based data management on local systems or central servers via RapidMiner Server;

   Immediately understand and create a plan to prepare the data automatically to extract statistics and key information.

ii. **Data Preprocessing:** Basics, Transformations, Data Partitioning, aggregation, Binning, Weighting and Selection, Attribute Generation.

iii. **Modeling:** Similarity Calculation, Clustering, Market Basket Analysis, Decision Trees, Rule Induction, Bayesian Modeling, Regression, Neural networks, Support Vector Machines, Memory-Based Reasoning, Model Ensembles.

iv. **Validation.** RapidMiner Studio provides the means to accurately and appropriately estimate model performance. Where other tools tend to tie modeling and model validation closely, RapidMiner Studio follows a stringent modular approach which prevents information used in pre-processing steps from leaking from model training into the application of the model. This unique approach is the only guarantee that no overfitting is introduced, and no overestimation of prediction performances can occur.

Appendix F.

week	Day	Lighting power (kWh)	CO2	Plug power (kWh)
1	Sunday	93	973	90
1	Monday	11	419	91
1	Tuesday	63	949	69
1	Wednesday	3	766	25
1	Thursday	51	742	50
1	Friday	78	941	54
1	Saturday	36	561	46
2	Sunday	6	530	23
2	Monday	68	802	4
2	Tuesday	56	660	0
2	Wednesday	51	625	68
2	Thursday	31	944	57
2	Friday	49	935	3
2	Saturday	90	671	59
3	Sunday	37	617	13
3	Monday	62	764	74
3	Tuesday	93	736	59
3	Wednesday	28	935	46
3	Thursday	42	461	29
3	Friday	57	822	46
3	Saturday	74	859	25

Aggregate by Day

Day	Average (CO2)	SUM (Lighting Power-kWh)	SUM (Plug power-kWh)
Sunday	706.7	136	126
Monday	661.7	141	169
Tuesday	781.7	212	128
Wednesday	775.3	82	139
Thursday	715.7	124	136
Friday	899.3	184	103
Saturday	697.0	200	130

Aggregate by week

week	Average (CO2)	SUM (Lighting Power-kWh)	SUM (Plug power-kWh)
1	764.4	335	425
2	738.1	351	214
3	742.0	393	292