THESIS REPORT

A Data Mining Approach to Study Occupant Behavior

in *De Kroeven Project*

Student: Xinyuyang Ren

Identity number: 0980573

Master Program: Sustainable Energy Technology

Department: The Department of the Built Environment

Research group: Building Physics and Services (BPS)

Thesis supervisors: prof.ir. W. (Wim) Zeiler, ir. G. (Gert) Boxem & dr. Y. Zhao

Date: 2015-06-13

Abstract

The occupants' behavior is the interactivity between the occupants and the building, which has direct influences upon both indoor environment quality and energy consumption. Although the occupants' behavior setting has become one of the essential assumptions in more and more building simulation programs, the effective methodology for studying occupants' behavior from real historical records has not been sufficiently discussed in literature. In this study, the data mining techniques are used to study the occupants' operation on the ventilation system control panel in 10 recent-renovated passive houses in the Netherlands. The whole study is performed from two perspectives in two levels. Firstly, in the single-house level, a filter-based approach is developed to calibrate from the electricity consumption signal *when* does the occupant adjust the ventilation system and a classifier-based approach is developed to find the causes behind this occupant's behavior, i.e. *why* he/she did so. Then in a higher level, the results of different occupants from 10 houses were compared and discussed, aimed at seeking the similarities as well as differences among different people. The outputs of this study could be used in the following data analysis about the impact of occupants' behaviors on indoor environment comfort and energy consumption. Also, the user profiles revealed from real records could serve as reference for the simulation program input, to improve the accuracy and reliability of simulation work.

Table of Contents

Chapter 1 Introduction

In this chapter, firstly the context of the problem addressed in this thesis would be discussed, followed by the definition of that problem. Then the outline and limitation of this work would be presented.

1.1 Project Context

The built environment contributes one of the largest shares of energy consumption in the Netherlands. Although new homes are becoming more energy-efficiently built, their contribution to the energy saving is not significant since they only slightly contribute to the total consumption compared with the existing housing stock [4].

With the expected rise in energy prices, energy efficiency of homes would become more and more important in the future. Thus, in this case comprehensive renovations were planned for the properties dating from the 60s and 70s, of which the objective is to improve the domestic energy consumption efficiency, as well as making them meet current requirement for indoor environment.

De Kroeven in Roosendaal is a housing stock built around 1964. Recently, between April 2010 and April 2011, it was completely renovated on the basis of passive house principles with comprehensive energy reduction measures. The refurbishment includes a very good insulation shell, an effective sealing of cracks and a balanced ventilation system with heat recovery for each house. As the result, in principle there should be 60%-70% less energy required by the houses compared to them before renovation [4].

In addition, it was assumed that the application of passive principle should also guarantee an improved and pleasant indoor environment. The balanced ventilation system with heat recovery should guarantee the indoor air quality and be able to keep it at desired temperature with little support from heating system.

After the finish of the renovation work, in order to test if the presumed quality of energy efficiency and indoor comfort has been reached, a monitoring program was launched. Between the year 2013 and 2015, varies of sensors were installed in 10 experimental houses (2 types, 5 for each type) and recorded the information regarding the domestic energy consumption, indoor environment as well as system running parameters etc.

During the investigation of the database, according to the *report of monitoring program Kroeven 2013* [4]*,* the indoor air quality, especially the CO2 concentration was found as one of the most severe issues in these renovated houses, for which further analysis is needed. Data shows without a distinction in the type of house, the number of hours in which the CO2 concentration excessing the requirement of 1200 ppm is far more than the design principle. In some houses, the CO2 concentration even keep at a significant higher level (>2000 ppm) in substantial amount of time. E.g. the Figure 1 below shows the CO2 concentration of house no.2 during the year 2013 and 2014, with the 1200ppm limit marked by the blue horizontal line.

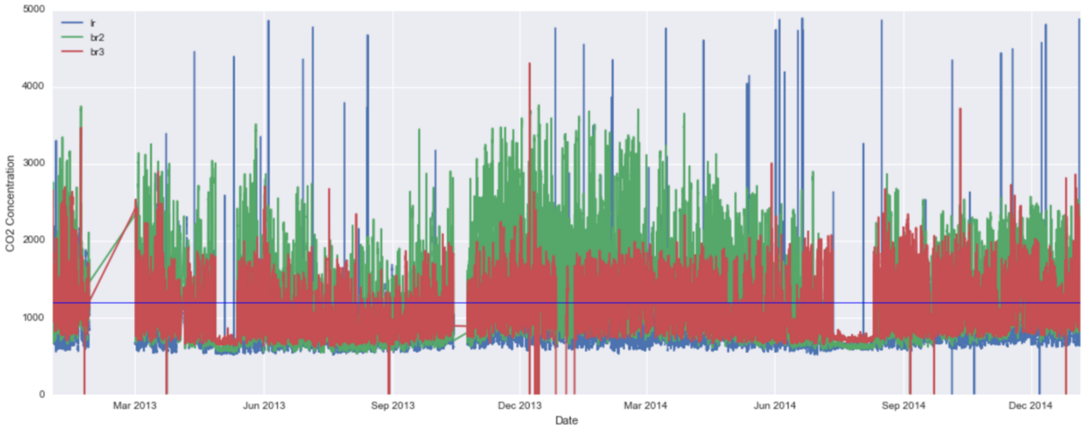


Figure 1.1 The CO2 concentration of house no.2

Obviously, the excess level of CO2 concentration comes from the insufficient ventilation rate compared to the CO2 production rate while the house is occupied. In consequence, an effective use of ventilation system could significantly ease this issue. In the report of monitoring program [4], researchers suggested that the occupants in these renovated houses should "*be more attentive to improve the indoor air quality, especially the CO2 concentration by proper use of ventilation position setting*". However, the detail of those behaviors remain undiscussed. In order to further understand the issue, this study provides detailed data analysis of occupants' ventilation-related behaviors in *De Kroeven project*.

1.2 Problem context

In this part, the motivation of performing this study, the general problem definition followed by the stepwise research questions and the limitation of this study would be elaborated.

1.2.1 The Motivation

Research shows the real energy and indoor environment performances of buildings depend not only on deterministic aspects such as building physics and design of HVAC systems, but also on stochastic aspects such as weather and occupants' behavior. So far, when calculating the expected performance of building, occupant behavior has not been adequately considered. Consequently, field test studies all over Europe have shown discrepancies between real and expected performance of buildings [17] [18].

Thus, introducing occupants' behavior models into existing building simulation software could be a key to bridge the gap between the real performance of building and the design standard. Before making behavior models, there is a need for a better understanding of occupants' behavior and in particular, the causes for their adjustments of building controls such as window opening, ventilation flow rate setting, thermostat setting, etc.

Also, in the domain of intelligent building research, one of the most important features that could indicate a building to be ‘intelligent’ is effective interaction with its occupants [21]. With a better understanding of people’s preference, the building control system could generate tailored strategies for its occupants.

In conclusion, the motivation of this research was to: 1) improve the building design procedure and 2) contribute to the intelligent building research by quantitatively analyzing occupant behavior in residential buildings, especially investigating which drivers lead occupants to interact with building controls.

1.2.2 Problem statement

Given the context of the project, the problem can be stated as follows:

Problem statement *While the excessive CO2 concentration is a severe issue in the De Kroeven project, an effective use of ventilation system could be the key to ease it. Before which it is needed to get insight about currently how do people use ventilation system, e.g. when, how and why these occupants interact with ventilation controls.*

1.2.3 Research question and solution proposal

To thoroughly analyze the topic, four sub research questions were raised. Each of these questions clearly describes one specific issue and would be discussed in a certain chapter of this report.

- **Question 1** How to find *when and how* occupants interact with their building control system without recording their behavior directly?
- **Question 2** How to find the *causes* behind occupants’ behavior, from data records?
- **Question 3** For different occupants, whether do they behave in the same way and, how similar/different are they?
- **Question 4** How to evaluate this data-based methodology and what are the potential applications?

For each research question, there is one specific solution proposed. Specifically:

- **Solution 1** Develop noise reduction and edge detection filter to mark occupants’ operation out from raw electricity consumption records. This step deals with the question “when” and “how” people interact with their ventilation system and would be discussed in *Section 3.3 noise reduction and edge detection.*

- **Solution 2** Develop a feature selection process to find the main cause for people’s behavior. This step deals with the question “why” people interact with their ventilation system and would be discussed in *Section 3.4 feature selection.*
- **Solution 3** Compare among different occupants, try to group them into several user profiles. In this step the study goes beyond individual level and investigate the similarity or diversity among different people. *Chapter 4 user profiles* would be the description of this step.
- **Solution 4** Turn the data mining results into actionable suggestions for improving indoor comfort. In this step the data analysis would finally be interpreted with expertise in built environment field. The final conclusion and discussion about the potential applications would be elaborated in *Chapter 5*.

The Figure 1.2 shows the overall logic design, or the ‘data pipeline’ in this research. It describes generally how will the data stream ‘flow’ throughout the whole process and defines the basic blocks and their own functionalities. In the following chapters, the procedure and outputs from each step in this pipeline would be elaborated

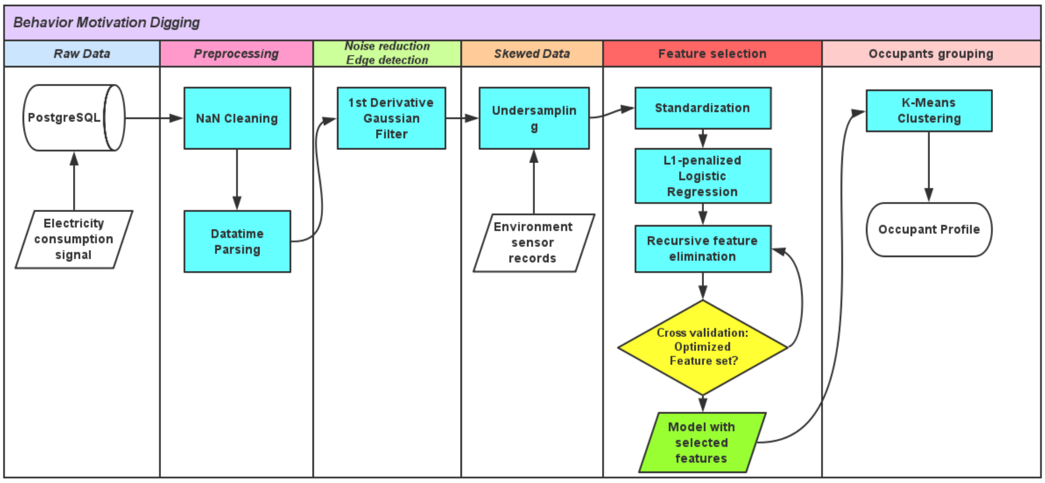

Behavior Motivation Digging
Raw Data
Preprocessing
Noise reduction Edge detection
Skewed Data
Feature selection
Occupants grouping
PostgreSQL
Electricity consumption signal
NaN Cleaning
Datatime Parsing
1st Derivative Gaussian Filter
Undersamplin g
Environment sensor records
Standardization
L1-penalized Logistic Regression
Recursive feature elimination
Cross validation: Optimized Feature set?
Model with selected features
K-Means Clustering
Occupant Profile

Figure 1.2 The data pipeline

1.2.4 Boundary

The scope of this study is limited by its methodology. Since it is a trial implementing data mining techniques in built environment research, and the student in responsibility for this study is not from the built environment background, although extra attention was paid to combine the expertise from two fields throughout the study duration, topics such as the procedure design and the result discussion are conducted more from the side of statistics and data science. Thus, all the conclusions drawn from this research are open for discussion with the experts from built environment itself.

1.3 Outline

The remainder of this report is structured as follows. In Chapter 2, the basics of passive house and the review for some relevant research upon behavioral data analysis would be presented. Also, the techniques and the dataset involved in this study would be introduced. In Chapter 3, a case study would be elaborated regarding the data analysis of operation on the ventilation system in one certain renovated house. In Chapter 4, with the case study results from 10 identical houses, we try to group their occupants into different user profiles based on their behavior patterns using clustering algorithms. Chapter 5 would contain discussions with experts from built environment about the contribution and potential application of this study, consequently the final conclusions.

Chapter 2 Preliminaries

In this chapter, firstly the basics of passive house would be presented in Section 2.1. Subsequently some relevant studies upon behavioral data analysis would be reviewed in Section 2.2, then the specific techniques and algorithms used in this research in Section 2.3. Finally, Section 2.4 contains the introduction of the datasets collected in the monitoring program of *De Kroeven project.*

2.1 Passive house concept

In the coming decades, due to the consequences of climate change and shortage of energy resources, substantial efforts should be made to use energy much more efficiently and to switch to renewable energy sources. As the energy consumption for household corresponds to about 31% of the EU15 total delivered energy [1], significant reductions in energy demand can be achieved by promoting low energy building technology such as the passive house concept, which is characterized by a holistic approach, combining several measures into a consistent framework. Buildings complying with the passive house standard are rapidly spreading across Europe [2]. Although the passive house concept is mostly applied to new buildings, it has also been used for refurbishments, as in *De Kroeven project*.

The passive house standard requires that the building fulfills the following requirements [3]:

1. The building must be designed to have an annual heating and cooling demand as calculated with the passive house *Planning Package* of no more than 15 kWh/m^2 per year or be with a peak heat load of less than 10 W/m^2.
2. Total primary energy consumption (heating, hot water and electricity etc.) must not be more than 120 kWh/m^2 per year.
3. The building must not leak more air than 0.6 times the house volume per hour at 50 Pa as tested by a blower door ($n_{50} \leqslant 0.6$ / hour).

In order to meet these requirements, the refurbishment measures in *De Kroeven project* for each house include a very good insulated and sealed shell, a balanced ventilation system with heat recovery as well as a solar thermal collector on the roof for domestic hot water demand [4].

2.2 Occupant behavior study review

With more and more researchers start to realize from practice that in addition to the study of physical refurbishment measures, understanding occupant behavior and its consequences is also essential to bridge the gap between designed and actual performance in building energy consumption and indoor environment quality. A good understanding of user behavior patterns could provide a more accurate input into building energy modeling programs in order to investigate the energy use and indoor comfort at a more precise level. Also, discerned occupant profiles enables the possibility of tailored building design plans in construction or refurbishment, to better fit the requirements of certain occupants.

Recently, more and more efforts were made to build effective methodologies to remove the impact of other variables and isolate the leverage of the human factor precisely, within which the effectiveness of statistical analysis and data mining approaches in finding meaningful correlations in data records is recognized by more and more researchers, but still remain inadequately discussed in literatures so far.

Wei, Shen, et al [16] made a statistical study monitoring occupants' window opening behavior in a mixed-mode office building in Beijing, China. Although there is no data mining concept introduced, many similar studies like this showed researchers started to turn to data-based approaches for behavior study. Cheng Fan [5] et al. pointed out the problem that currently the data in Building Automation System (BAS) can seldom be effectively utilized due to the lack of powerful tools for analyzing the large data. They introduced a general data mining framework for knowledge discovery in massive BAS data with special consider for the low quality and complexity. Based on the test result on an office building in Hong Kong, they suggest the data mining techniques could be both novel and promising in the field of built environment study. Then, Simona et al. [7] developed a framework combining statistical analysis with two data-mining techniques, cluster analysis and association rules mining, to identify valid window operational patterns in measured data. Their study demonstrated the effectiveness of logistic regression in this field and the idea of hidden cause digging as well as extracting user profiles using clustering algorithm inspired this study. Similar logistic-regression based studies were performed in different cases also by Calì, Davide, et al. [17], Andersen et al. [18] and Shi et al. [19]. Based on their results Andersen, Rune, et al [20] went further by putting the occupants' window opening behavior patterns extracted from data analysis into simulation programs. Xiaoxin Ren [9] et al. made another study investing the behavior of occupants adjusting their thermostat settings and heating system operations with data mining methodology in a 62-unit affordable housing complex in Revere, Massachusetts, USA. Beyond the two typical behavioral recognition studies mentioned above, Imran Khan [6] et al. extended the scope of data mining practice in the built environment study by developing a fault-detection framework to detect abnormal lighting energy consumption. This study is helpful for building energy management systems to reduce operating cost and time, also, their idea invokes our study to find the behavior in the ventilation functions that is not in line with the design principle. Furthermore, Fu Xiao [8] et al. developed a prediction model for next-day building energy consumption and peak power demand using data mining techniques, with considering the occupant's preference and behavior. Their study helps to understand the consequence of occupants' behavior on energy consumption and how data mining techniques could help in revealing that relations.

2.3 Techniques involved in this project

There are basically three different kinds of tasks involved in this project, respectively noise reduction/edge detection and feature selection. The theoretical solutions are based on *1st derivative Gaussian filter* and *recursive feature elimination with logistic regression kernel* respectively. In this section each technique involved would be briefly introduced.

2.3.1 First derivative Gaussian filter

1st derivative Gaussian filter is one of the most widely-used filter in signal processing domain, in this study it is used to ease the noise of the electricity consumption signal and calibrate people's adjustment from it.

Gaussian filter [11] is a common choice for noise smoothing. Mathematically, a Gaussian filter modifies the input signal by convolution with a Gaussian function, which is also known as the Weierstrass transform. In other words, a Gaussian filter replace the value at each point by the weighted average of its neighbors', with the weights come from a Gaussian distribution, then normalize the results. Below Formula 1 and Figure 2.1 show a typical Gaussian distribution.

$$g(x, \delta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

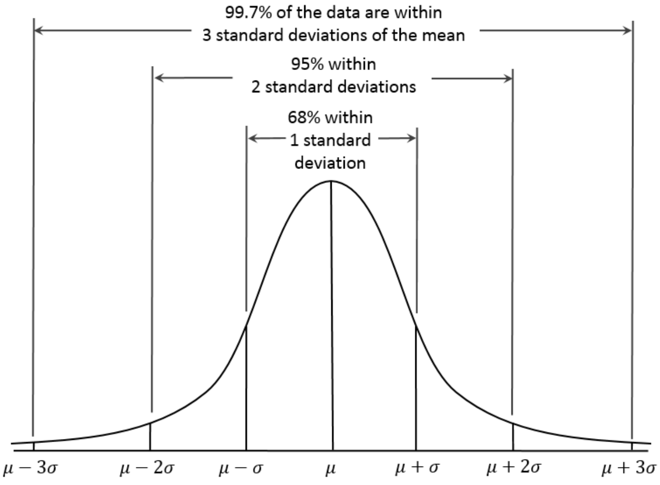


Figure 2.1 Gaussian distribution curve

After the filter, the noise, especially Gaussian noise, could be effectively smoothed. Also, it is clear from the equation that the coarseness of noise smoothing effect is controlled by the parameter *sigma* of Gaussian function: a lager sigma would bring stronger smoothing effect. Figure 2.2 and Figure 2.3 respectively show the influence of sigma and the comparison between the original signal and smoothed signal after a Gaussian filter.

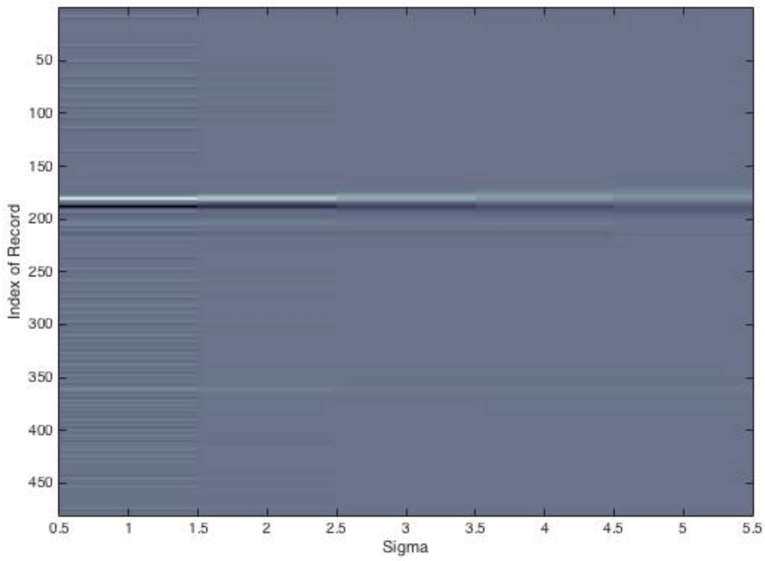


Figure 2.2 The scale space of Gaussian filter

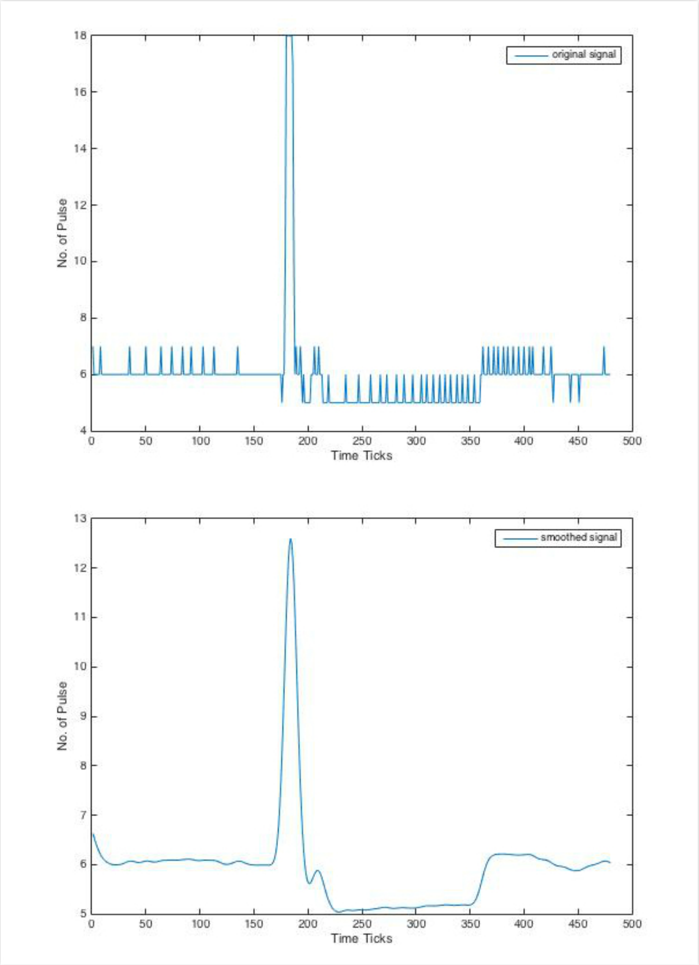


Figure 2.3 Comparison between the original signal and smoothed signal.

In the Figure 2.2 *the scale space of Gaussian filter,* the horizontal axis represents the value of *sigma*, which is increasing from left to right while the vertical axis is the filtered signal after the Gaussian filter with respective *sigma*. It could be clearly observed that a filter with larger *sigma* tends to filter more details out, since in the left part we can still observe many tiny ripples while in the right side, the biggest main edge is the only thing readable. Figure 2.3 shows the comparison between the original signal and smoothed signal, it could be observed that the tiny saw-tooth like noise is inhibited, while the main edge remains at its place.

1st derivative Gaussian filter [12] [15] is similar to an original *Gaussian filter* but the Gaussian distribution kernel is substituted by its 1st derivative formula. A finely-tuned 1st derivative Gaussian filter could be used to detect the sharp edges of signal or images if a *threshold* parameter is set*,* indicating above what fraction of the global extreme a local extreme could be recognized as an effective edge: apparently a kernel with a lager threshold tends to be less sensitive. Formula 2.2 and Figure 2.4 below show the 1st derivative Gaussian kernel equation and its effect on the demo signal mentioned above.

$$\frac{\partial g(x,\sigma)}{\partial x} = \frac{x}{\sigma^3\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

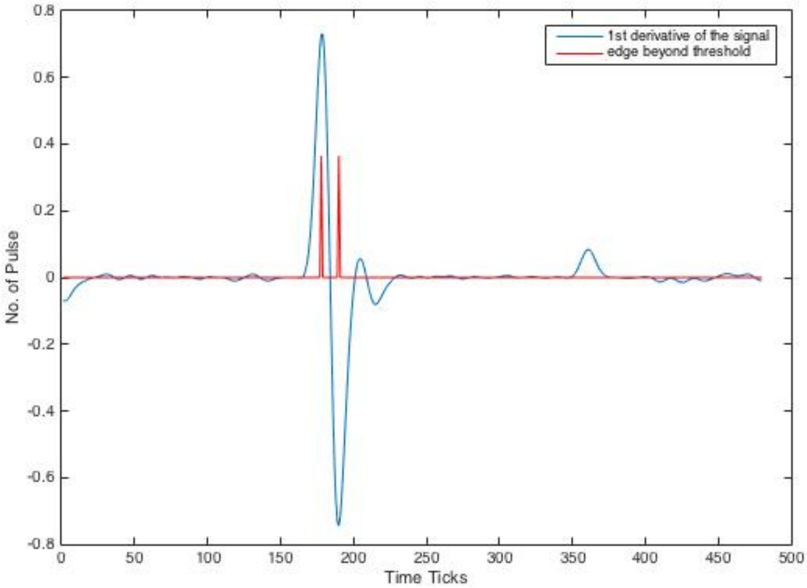


Figure 2.4 Demo signal after 1st derivative Gaussian kernel

2.3.2 Logistic regression

Logistic regression [13], despite its name, is a linear model for classification rather than regression. It is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier.

Based on its linear nature, in this study the coefficient of each feature in a trained logistic regression model is used to evaluate the importance of this feature. The effectiveness, interpretability and robustness of this approach have been validated by many peer researchers [7] [17] [18] [19] [20].

When using a regular linear regression, we did

$$h_\theta(x) = \theta^t x$$

where x is a series of features, θ is a vector containing coefficients for each feature and $h_\theta(x)$ represents the regression result. While in logistic regression, since we want to do a classification instead of regression, the linear regression equation is fitted in to a sigmoid function

$$g(z) = \frac{1}{1+e^{-z}}$$

Finally, the equation of logistic regression becomes

$$h_\theta(x) = \frac{1}{1+e^{-\theta^t x}}$$

The function is plotted in below Figure 2.5, from which it could be observed that the range of logistic regression is between 0 and 1, if we choose a threshold, say 0.5 to divide two different categories (i.e. if $h_\theta(x)$ < 0.5 predict the case to be in category 1, else if $h_\theta(x)$ >= 0.5 predict it to be in category 2), the decision boundary could be determined as $\theta^t x = 0$. After training, the classifier model is adjusted to minimize the prediction error based on the training set and the coefficients of each feature, i.e. the θ^t could then be used to evaluate the relative importance of each feature in its classification process.

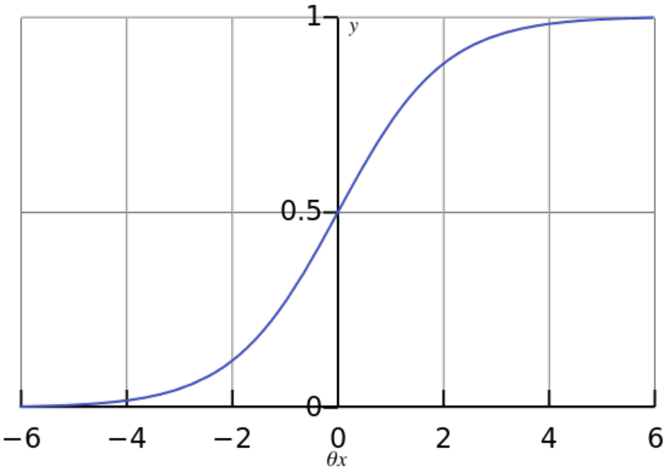


Figure 2.5 Logistic regression curve

In addition, in this project the logistic regression kernel we use is with *L1-norm regularization*, which means when calculating error in the *cost function*, there is an extra penalty factor coming from the L1-norm of the coefficient vector.

$$L1 = \lambda \sum_{i=1}^{n} |\vartheta_i|$$

Since linear model penalized with L1 norm tends to give sparse solutions i.e. many of its estimated coefficients would be zero, it could be used for feature selection purpose [10].

The logistic regression runs repeatedly to make a *recursive feature elimination*. First, the estimator is trained on the initial set of features and weights are assigned to each one of them. Then, features whose absolute weights are the smallest are pruned from the current set features. At last, the most informative feature combination (judged by cross-validation accuracy) in this case could be determined, which could imply the cause of occupant behavior.

2.3.3 K-means clustering

K-means clustering [14] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem with good interpretability. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The clustering partition with high intra-cluster similarity and low inter-cluster similarity would be considered as good performance.

In this study, the K-means clustering is used to group occupants from 10 different houses into several types. This approach has been validated also by the research from Simona et al. [7] and Andersen, Rune, et al [20].

Specifically, the procedure follows a simple and easy way to cluster a given data set through a certain number of clusters. The basic idea is to first define k centroids, one for each cluster, which should be placed in a cunning way because different location causes different result. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as the barycenter of the data points belonging to a certain cluster resulting from the previous step. After we have these k new centroids, a new binding could be done in a similar way, between the same data set points and the nearest new centroid. So far the loop has been generated. As a result of this loop, we may notice that the k centroids change their location step by step until no more change. In other words, centroids do not move any more after a certain number of loops.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is the chosen distance measure between a data point and the cluster center it belongs to. In this case we choose Euclidean distance as the distance measure method.

2.4 The data set

Aramis AlleeWonen is a housing stock in the district Kroeven, Roosendaal, Netherlands. It is built around 1964 and is completely renovated during April 2010 to April 2011 on the basis of passive house principles, with comprehensive energy reduction measures including heat recovery ventilation and solar collector.

Figure 2.6 Aramis AlleeWonen dwellings

After the renovation, to test if the presumed quality of the renovated dwellings is achieved, a monitoring program was started. A series of sensors were installed in 10 dwellings with the records stored in a SQL database. The monitoring program lasts for 2 years, from Jan. 2013 to Jan.2015. The data relevant in this study is listed in table 2.1.

Table 2.1 Data characteristics

Category	Items	Interval
Weather Condition	Average Temperature [℃]	1 hour
	Average Relative humidity [%]	1 hour
	Average Irradiation [W/m²]	1 hour
	Average Wind speed [m/s]	1 hour

Indoor Environment (living room and 2 bedrooms)	Indoor Temperature [℃]	3 min
	Relative humidity [%]	3 min
	CO_2 Concentration [ppm]	3 min
	Ventilation System Supply Air Temperature [℃]	3 min
Occupant Behavior	Window state [0/1] (living room, 2 bedrooms, attic)	3 min
Energy Consumption	Gas Consumption [m^3]	1 hour
	Electricity Consumption [Kwh]	1 hour

Chapter 3 Data analysis for ventilation system operation

This chapter elaborates a data analysis case study for one house regarding its occupants' operation on the ventilation system, based on the system electricity consumption signal and other relevant sensor records.

3.1 Context

As mentioned in the project introduction, there is a balanced ventilation system with heat recovery installed inside every renovated house in *De Kroeven project*, of which the ventilation flow rate is controlled by a fan system, and adjustable by occupants. However, due to some reason it is not possible to directly record people's operation on this ventilation system.

The electricity consumption of the fan system is recorded every 3 minutes by a smart meter in terms of pulse. Obviously, occupants' flow rate setting could put significant influence on the electricity consumption (e.g. once the occupant turns the flow rate into a higher option there should be a steep increasing edge on the electricity consumption signal), which could provide possibility to calibrate when and how people adjust their ventilation system based on the electricity consumption.

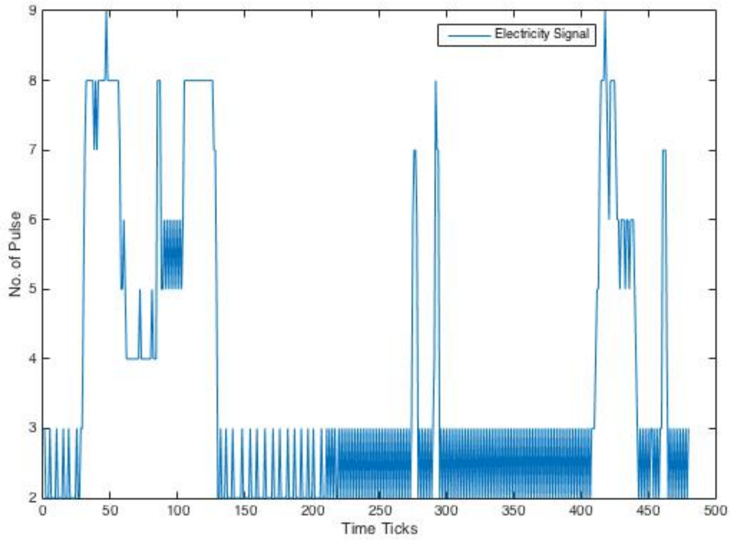


Figure 3.1 One-day electricity consumption signal of the fan system

However, on the one hand, as shown in Figure 3.1, with the influences from back pressure, wind speed etc., the record is not an ideal clean square wave, as we originally thought, indicating the adjustment only. Instead, it is quite noisy. On the other hand, there are many different houses in the project with similar structure but the scales of records may vary. As the consequence, it is not practical to calibrate the ventilation position change by presuming fixed intervals manually (like *the number of pulse less than 3 infers to position 1 while between 3 and 5 infers to position 2 etc.*), for which it is difficult to determine the dividing boundaries and also could result in considerable miscounts from the noises. A new algorithmic method is needed to ease the noise and detect the changes, finally calibrate the occupants' operations out from electricity consumption signal automatically.

After the calibration we should have known 'when' and 'how' occupants interact with their ventilation system, then it is needed to answer 'why' they are doing so, according to the problem description. As introduced in the previous chapters, the monitoring program also includes the records of indoor environment parameters and outside weather information etc., thus the task could be described as *find the most informative/influential features in occupants' decision making process.* Technically, it could be solved by developing a machine learning classifier trying to predict occupants' reaction to a certain circumstance, and evaluate the importance of each feature mathematically inside the algorithm.

3.3 Noise reduction and edge detection

After essential preprocessing and cleaning steps, in which the missing values are backfilled by their closest subsequent neighbors and the date & time string is compiled to the system-recognizable timestamp, we get a system electricity consumption signal like shown in Figure 3.2.

As mentioned above, a sudden change in the signal could imply the occupants' interaction with the system as long as the noise (caused by wind etc. or system itself) and "fake operation" (status change with too-short duration) could be effectively filtered out.

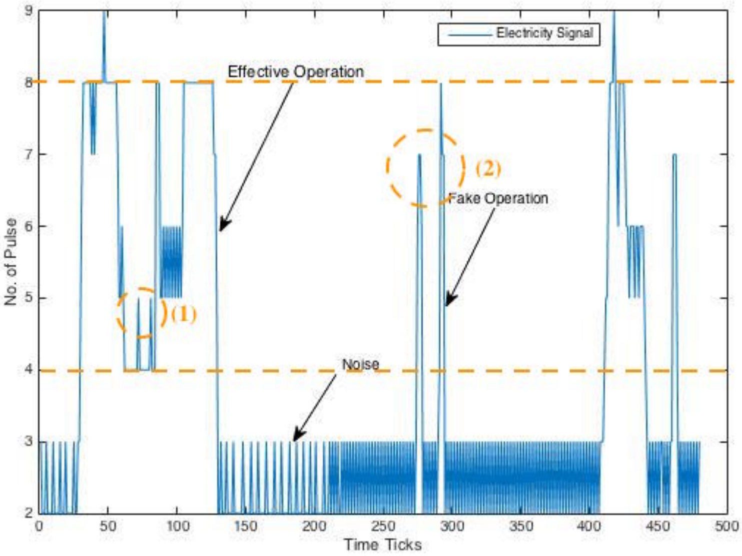
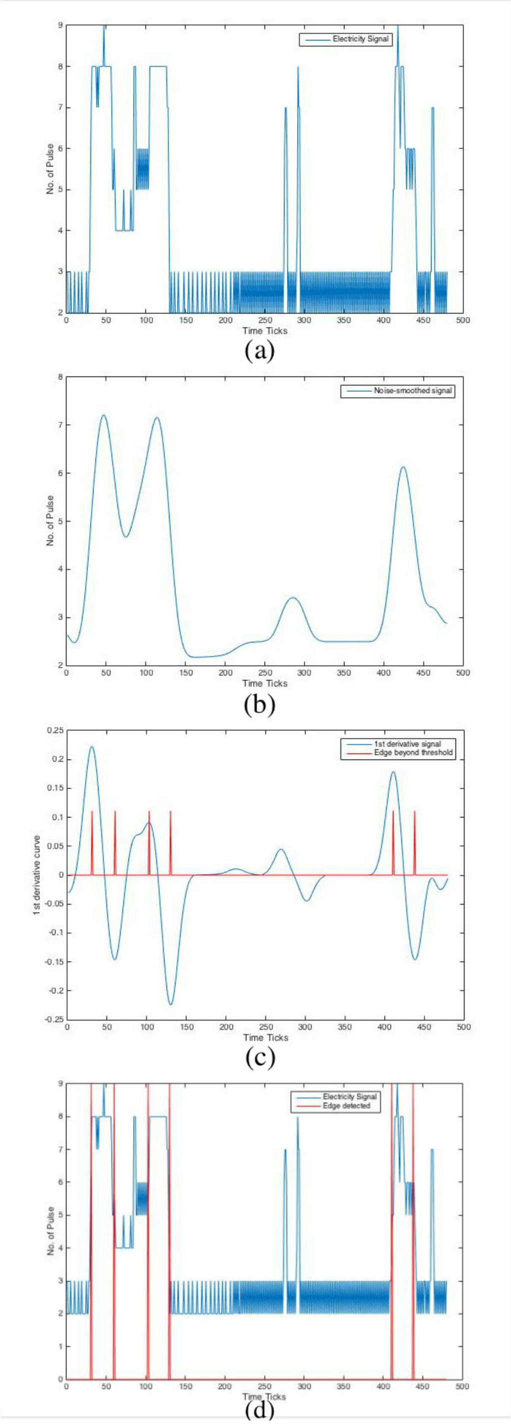


Figure 3.2 Operation, noise and fake interaction

In the previous research work before this study, researchers used to calibrate each position with fixed interval. E.g. in this case, positions with pulse no. fallen in [0,4] are assigned as ‘position 1’, while (4,8) for ‘position 2’ and pulse no. larger than 8 represents ‘position 3’. Follow this approach, the user operation frequency could be seriously over-estimated since both the noise (e.g. in circle 1) and ‘fake operations’ (e.g. in circle 2) are counted as effective user operation. In fact, in the previous report the researchers estimated this house with over 1,000 operations per year, which is apparently too much for a regular ventilation system controller. To make things worse, with the fixed-interval approach, for each house the intervals need to be decided case by case since the scope of the no. of pulse in different house may vary. In the next paragraphs, I will show how does the filter-based approach developed in this study solve all the issues mentioned above by automatically marking the effective operational edges and filtering out the noises and ‘fake operations’.

Through a finely-tuned *1st derivative Gaussian filte*r, with a proper parameter combination, in this case containing a *sigma* value of 10 and a *threshold* value of 0.35, the noise and fake operations last less than 3 minutes could be effectively filtered and the valid operations could be marked out. Figure 3.4 below shows the comparison among the noise-reduced smooth signal, edge-detected 1st derivative signal and the original electricity consumption signal with the detected valid operation marked out by red vertical lines.



(a)

(b)

(c)

(d)

Figure 3.3 The working process of noise reduction and edge detection

In Figure 3.3 above, (a) is the raw signal of fan electricity consumption, with noise and fake operations; (b) shows the signal after the Gaussian filter, with which the signal is smoothed and the noise is reduced; (c) is the 1st derivative signal of (b), each peak here could imply an edge in (b), with a proper threshold, we can filter out the real operation edge we want in a certain sensitivity. (d) is the output of this filter, which is the original signal with operation edge marked out according to the positions indicated from (c). It could be observed that the finely-tuned algorithm could automatically ignore the noise and fake operation, only mark the real operation edge we want.

With a lager scale in2 years, the operation detected could be presented in the Figure 3.4, in which +1 represents increasing operation while -1 represents decreasing operation.

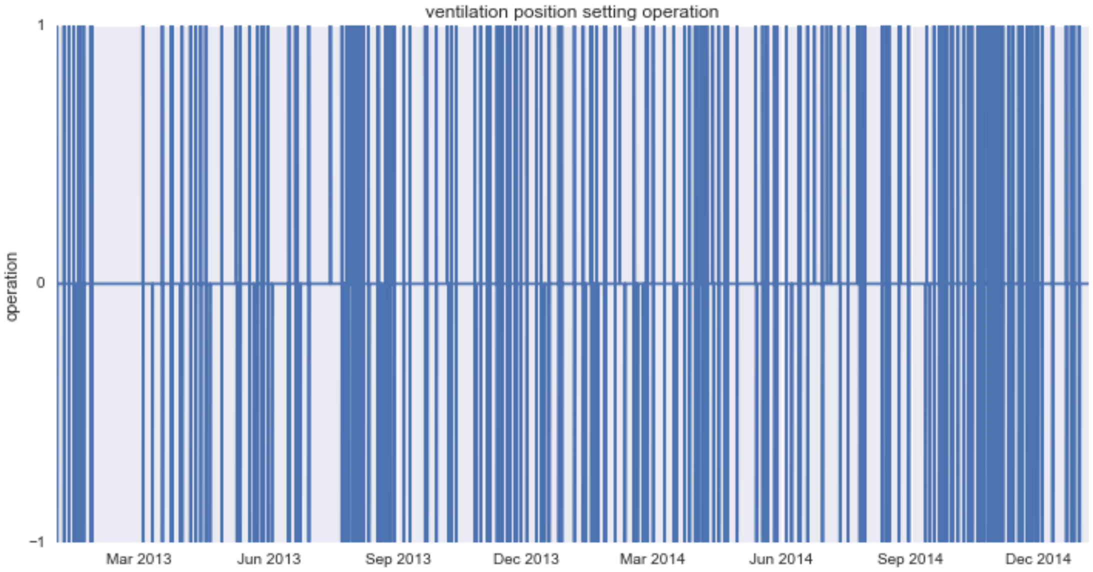


Figure 3.4 ventilation position setting operation

3.4 Feature selection

Then the marked data set would undergo an *undersampling* process since the dataset is now skewed i.e. the no. of records marked with 'no operation' is far more than ones with operation, either flowrate increase or decrease. In the undersampling process, we randomly picked a fraction of the records with 'no operation' to ensure the data set has balanced scales with each class, which is essential to guarantee the effectiveness of the statistical classification algorithm.

After undersampling, the training set would be *normalized,* which means the values are re-scaled into a zero-mean and unit-variance distribution, so that the coefficient of the linear model we used could indicate the relative importance of each feature.

Then the normalized dataset is fed into a L1-penalized logistic regression classifier. As mentioned in Chapter 2, a linear model penalized with L1 norm has *sparse solutions* i.e. many of its estimated coefficients would be zero, thus could be used for feature selection purpose. Figure 10 below shows an example of the coefficients output for the classier distinguish 'no operation' and 'increase operation'.

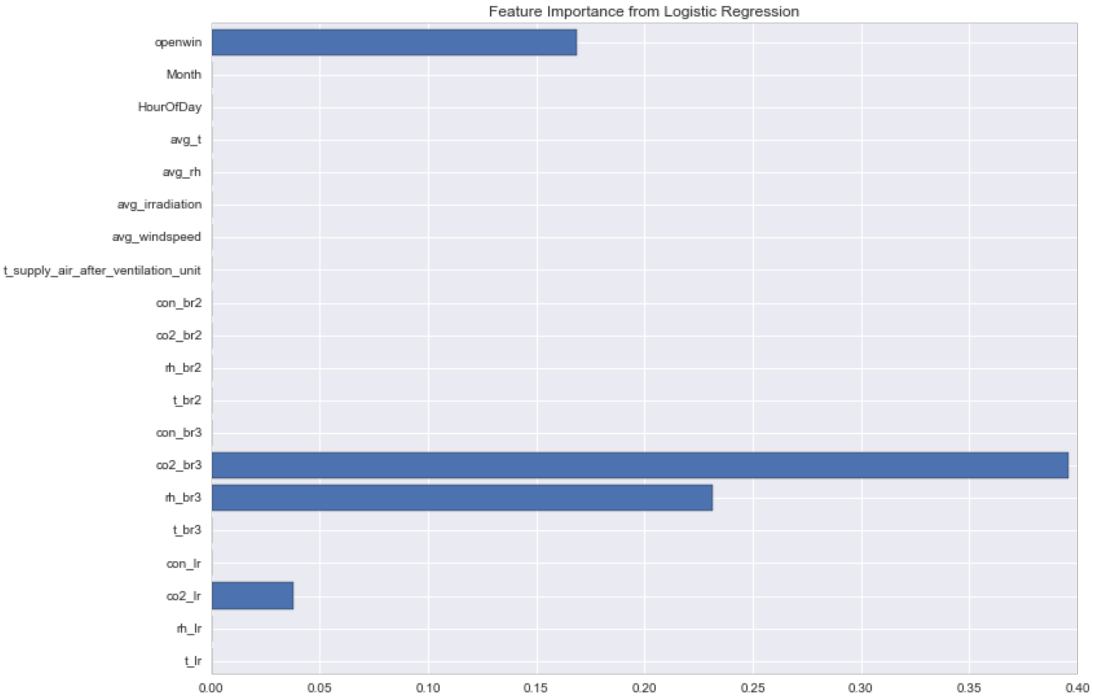


Figure 3.5 Feature importance output of increase operation

As shown in figure 3.5, each feature is assigned with a value that indicate its relative importance in people's decision-making process. The explanation of each abbreviation is shown in table 3.1.

Table 3.1 Abbreviation Explanation

Abbreviation	**It stands for**
openwin	No. of opened windows in the house
Month	Month of year [1,12]
HourofDay	Hour of day [0,23]
avg	Average
t	Temperature (℃)
rh	Relative humidity (%)
irradiation	Irradiation outside
windspeed	Wind speed outside (m/s)
CO2	Indoor CO2 concentration (ppm)
br1/2/3	Bedroom 1/2/3
lr	Living room

It could be observed that most less-informative features for this occupant have been filtered out with zero-coefficients, while the remaining, the number of opened windows, CO2 concentration in bedroom 3 and living room, as well as the relative humidity in bedroom 3, seem to be the most important motivations for this occupant to increase the ventilation flowrate.

Finally, the logistic regression runs repeatedly to make a *recursive feature elimination*. As the result the top 3 informative features are selected in this case, namely the number of opened windows, CO2 concentration in bedroom 3 and the relative humidity in bedroom 3 respectively. Based on these three features the algorithm could reach an accuracy of 87.5%.

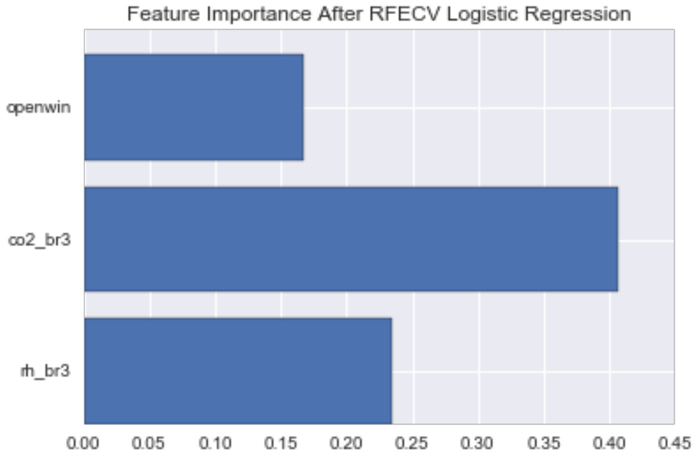


Figure 3.6 The RFECV result of increase operation

Similarly, the analysis could be made regarding distinguish “no operation” and “decrease operation” in order to find the cause for this occupant to decrease ventilation flowrate.

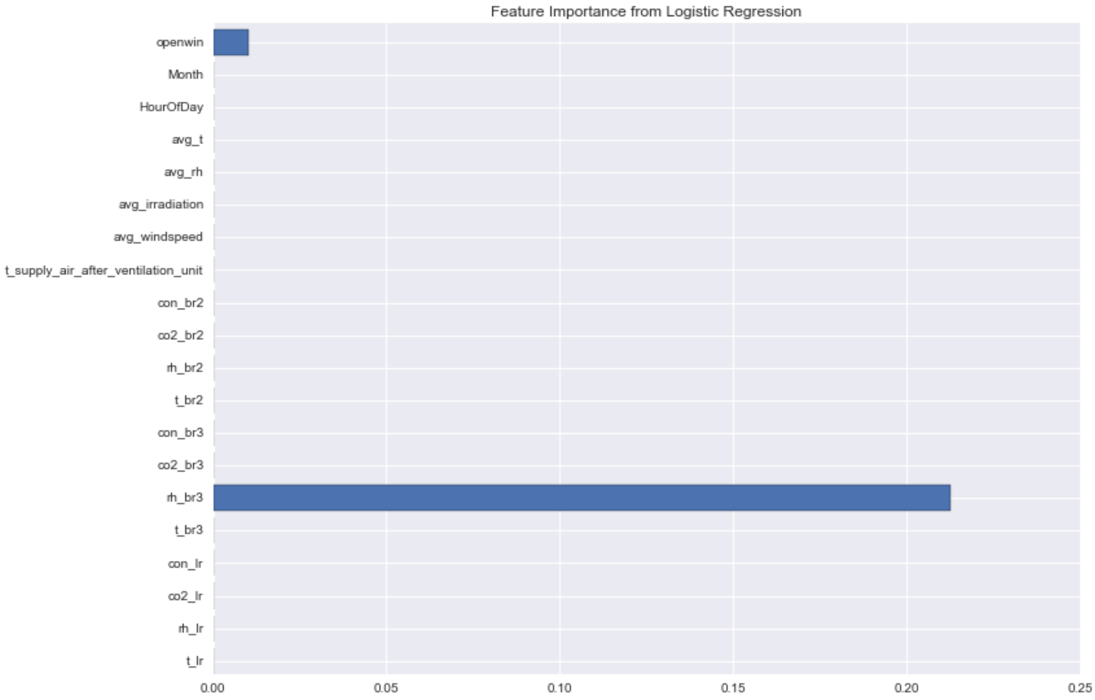


Figure 3.7 Feature importance output of decrease operation

It could be observed that compared with increase operation, the causes of decrease operation for this occupant seems to be simpler, dominated by the relative humidity in bedroom 3, while the number of opened window still contributes a little bit.

From the previous analysis, this occupant cares about his/her indoor environment, especially in the bedroom and he/she is aware of improving the indoor environment quality by interacting with the building control system.

3.5 Conclusion for the ventilation system analysis

In this chapter, the analysis process of ventilation system operation is elaborated with a case study from house no.2. The data pipeline designed is proven to be effective in answering “when” and “how” occupants interact with the system by the Gaussian-filter-based noise reduction and edge detection. Also, the logistic regression based feature selection could explain “why” people were perfuming those interactions. In the next chapter, the study will go beyond one house case, look into different houses and different occupants and try to group people into several typical user profiles.

Chapter 4 User Profiles

In the previous chapter we performed the data mining framework in one house to analyze the occupants' operation on their ventilation system. During the monitoring program of *De Kroeven project,* there are in total 10 houses have been monitored. Thus the comparison among these 10 houses would be interesting and could lead our research into next level: find behavior patterns among different people. In this chapter we will try to answer the research question *do the occupants from different houses behave similarly or differently, and how?*

4.1 Positon change frequency

The operation frequency represents how often do the occupants operate the ventilation system and is counted by the edges detected by our algorithm mentioned in the previous chapter. It implies the awareness of occupants to control their indoor environment themselves by interacting with building control system. The statistics is collected for each house during 2 years.

Generally speaking, people started to get more proactivity upon the ventilation system in the year of 2014, compared to the previous year of 2013. More operations were performed in almost all houses (except for house no.4 and 5). That could because of the weather condition difference in 2 years since there is no man-made interference during the whole monitoring duration. Based on their operation frequency, those 10 houses could be clustered into 3 different types:

- Low frequency: houses no. 1,5,10 – less than 50 times per year
- Medium frequency: houses no. 4,7,8,9 - 100 ~ 200 times per year
- High frequency: houses no. 2, 3 ,6 - 300 ~ 450 times per year

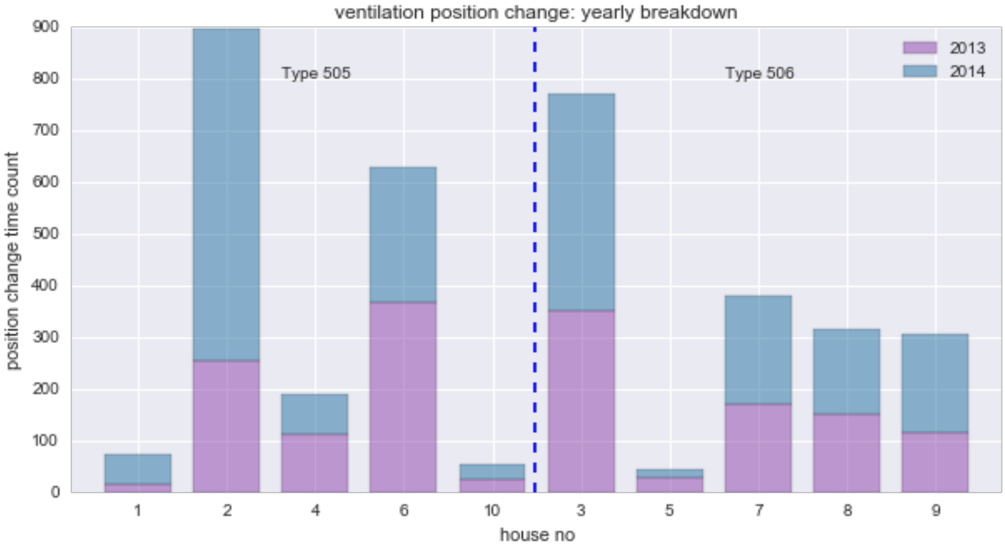


Figure 4.1 Ventilation position change frequency

4.2 Ventilation position distribution.

There are three possible setting positions for the ventilation system installed in each houses studied. Besides the operation frequency, on what position did the ventilation system stayed for the most of time is another important issue. It implies the average ventilation capacity throughout the monitoring period.

Before this study, it was difficult to compare ventilation position among different houses since the lack of effective method to calibrate the position from electricity consumption record, also, the different scales of electricity consumption between the house type 505 and type 506 brought extra barrier. With the approach based on the 1^{st} *derivative Gaussian filte*r developed in this study, the position information could be isolated and extracted from electricity consumption, represented simply by position 1,2,3 and ready for comparing.

Below Figure 4.2 shows the ventilation position distribution of 10 houses.

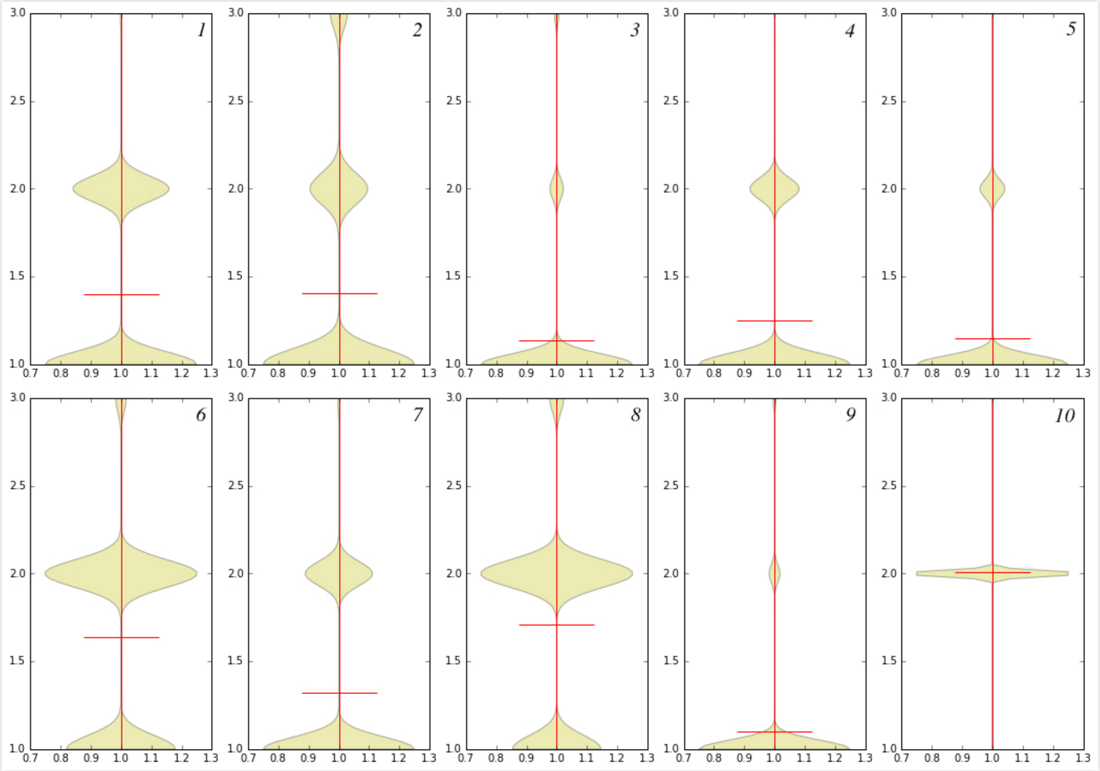


Figure 4.2 ventilation position distribution

In every single violin plot above, the vertical axis represents the ventilation position (namely 1, 2 and 3) and horizontal axis represents the distribution density (note that it is the smoothed kernel density and the area of each part is used to deliver an intuition about its fraction within the whole, so the value is not exactly integer 1, 2, 3 only) while the read horizontal line indicates the average position. E.g. it could be observed that the occupant in house no.10 rarely touch their ventilation system and just keep it at position 2 for the most of time, thus they also got a relative low operation frequency in the previous Figure 4.1.

4.3 Cause pattern

The cause analysis, which is discussed in the previous Chapter 3.4, is to reveal the main driven factors of occupants' behavior in the level of one single house. It could be expected that with more houses included, different people should hold different preferences and not behaving in the same way. Thus in this chapter the analysis is continued by a clustering process, trying to segment different occupants into several different cause types.

After the data pipeline discussed in the previous chapter, the relative importance of each feature could be extracted, the results from different occupants showed that the main driven factors for the occupants to interact with the ventilation system fall into two categories: time-related factors, including month, weekday/weekend, hour of day and indoor-environment-related factors, including indoor temperature, temperature of supplied ventilation air, relative humidity or CO2 concentration etc. Then with essential re-scaling, the occupants of those ten houses could be clustered into three different types by K-Means algorithm, as shown in the Figure.4.3.

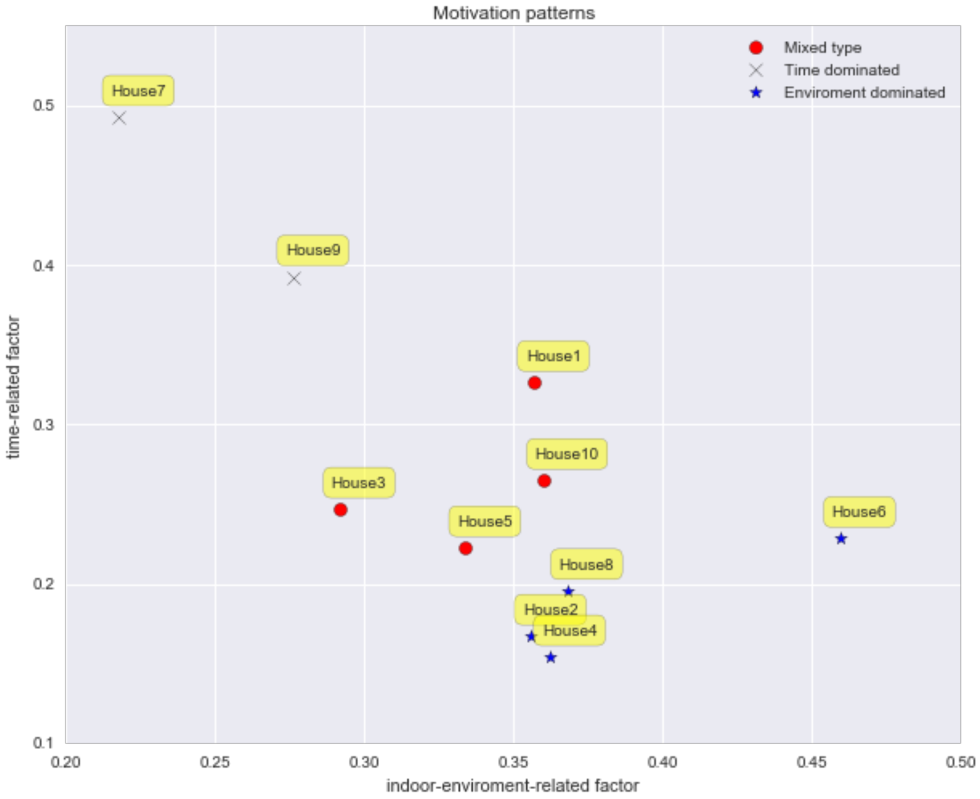


Figure 4.3 Cause pattern of ventilation system operation

In Figure 4.3, the horizontal axis represents the importance of indoor environment in determining occupants' behavior, while the vertical axis represents the importance of time-related factors. Three different types of occupants could be observed:

- Indoor environment sensitive occupants: house no. 2, 4, 6, 8
- Time sensitive occupants: house no.7, 9
- Mixed type occupants: houses no. 1, 3, 5, 10

The complexity of occupants' behavioral pattern is demonstrated by the data analysis result. The Indoor environment sensitive occupants are more likely to interact with their ventilation control panel when they feel unsatisfied about the indoor comfort, while the time sensitive occupants are more likely to have fixed timetables for their behavior (e.g., as soon as they wake up or come back from work etc.) and there are also some people in between, as mixed-type occupants their behaviors are effected considerably by both factors in the same time.

4.4 Conclusion from the comparison of occupants

In this chapter, we take the study further, from the single house level to the comparison among ten different houses. The comparison is made upon the perspectives of operation frequency, position setting distribution as well as driven factors, from which a user profile could be generated for each group of occupants regarding their behavioral habits and preferences. The result could be used further to explain the impact of their behavior on indoor environment quality or energy consumption etc., also it provides us the possibility for incoming refurbishment with tailored design according to its occupants' individual preference.

Chapter 5 Final Conclusion and Discussion

In the previous four chapters, details of the data mining framework developed in this research was elaborated with a case study *De Kroeven Project*. This chapter will contain a short summarization of all the key issues, final conclusions and some potential future plans for this topic.

5.1 Summarization and Conclusion

In this research, a framework combining statistical analysis with signal processing and data mining techniques was employed to study the occupant behavior of adjusting the ventilation control panel in ten recently-renovated passive houses in the Netherlands.

Due to the fact that the operation is not directly recorded in the database, for the first step the goal of this research was to identify 1) when, 2) how, and then 3) why each individual occupant interacts with the ventilation system from the electricity consumption signal and other indoor environmental data. Also, since there are 10 different houses in total included in the database of the monitoring program, later a comparison among them was also performed to identify the similarities as well as differences in their behavioral patterns.

For this goal, a *1^{st} derivative Gaussian filter* was developed and tuned to filter the noisy electricity consumption signal and automatically detect the edge, which could indicate the time and direction of occupant's adjustment. Then, a *logistic regression* classifier was developed and tuned to predict occupant's possible reaction to a certain circumstance, during which it also evaluates the relative importance of each feature in the decision-making process mathematically. In addition, with the help of *L1-penalization* and *recursive feature elimination model*, we could eliminate the uninformative features and remain the top 3 influential features for this occupant, which is the best parameters to explain the causes for this occupant's behavior. So far, with these techniques, we could answer the question 1) when, 2) how, and 3) why for a certain occupant's behavior. In a bigger picture, in the comparison among occupants from ten different houses, some unique user profiles were summarized, i.e. based on certain criteria, occupants could be grouped into several types by K-means clustering model.

As the result of the comparison work, the occupants from 10 different houses were grouped into several types according to the frequency and causes of their adjusting operation. The frequency of operation implies occupant's awareness of control the indoor environment proactively via the ventilation system, which actually varies a lot from less than 50 times up to 450 times per year, we classified them in to three categories with low, medium and high operation frequency respectively.

The cause pattern implies the preference of occupant. Some of them were found to be more sensitive in indoor temperature or CO2 concentration than others and tend to react proactively to the change of these indoor environment parameters. For some other occupants, their behaviors tend to follow a regular agenda every day. Respectively we call those two kinds of occupants as *environment-driven* occupants and *time-driven* occupants. Of course, there are also people in between, whose behaviors are more complicated, influenced by both factors in the same time, and in this study they are assigned as *mixed-type* occupants.

In general, the data-based approach built in this research is able to effectively analyze people's behavior and the causes behind. Instead of doing a survey or interview, the algorithmic method is more efficient and the result drawn from real database records is more reliable with less man-made disturbances.

5.2 Contribution of the study

This study makes its contribution in two perspectives, both for the *De Kroeven project* itself and the data-based behavior study methodology also.

For *De Kroeven project*: The previous researchers suggested in their report the awareness upon occupant behavior but related study is not fully performed. This study provided quantitative behavior analysis for De Kroeven project in a proven approach (logistic regression-based method [7] [17] [18] [19] [20]). Also, this approach could be demonstrated an improvement compared to the method previous researchers adopted.

For the *data-based behavior study methodology*: This study also contributed several novel parts technically, which are not contained in other related papers, including

a) Operation detection from indirect resources: Instead of use direct records of occupants' operation (like open/close window), a filter-based approach is built to automatically detect user's operation from indirect sources (like electricity consumption in this case), which added up extra possibilities and extended the potential scope of behavior study. E.g. for some operations that is missing from database/difficult or costly to record, this approach could be effective.

b) Recursive feature elimination (RFE) module to filter the most informative feature mathematically from massive amount of them. In other related researches, in order to avoid statistical difficulties, the features to monitor are either presumed by interview [18] or intentionally constrained into a small scope [19], which could be one-sided and introduce unnecessary man-made influence. Instead of study 4 or 5 features, this research studied 30 features in the same time, and was able to filter out the top 3 using L1-regularization and RFE, which is a proven method for feature selection in data science practice.

c) User profile generation. This research is beyond the scope of study the behavior habits, preferences, motivations etc. of one certain occupant, it also contains a comparison among occupants from 10 different houses and try to find their similarities as well as differences in behavioral pattern. These people are clustered in to different "types" based on their behavior, which could provide a better understanding of the complexity of people.

5.3 Potential Future Plan

This research is a trial to combine the novel data science technique with a research topic in built environment field. The result implies the approach developed to be effective and since this kind of combined study still remains largely undiscussed, there are many more potential fields for this research to go further:

- The approach built is essentially a generic framework of data-based behavioral study. In the case study contained in this report it is used to study the ventilation panel adjustment, but it could also be used to study other behavior like window opening/closing, thermostat setting etc. A comprehensive understanding of occupants' behavior could help bridge the gap between designed and actual building performance;
- This kind of behavior study provides a more accurate assumption of actual ventilation scenarios that may serve as reference for the simulation work in design phase, also, it allows building designers and operating manager to tailor more efficient and robust control strategies;
- The quantitative analysis of occupants' behavior provides the possibility to study its consequence. E.g. researchers could further study the link between the occupants' behavior and indoor energy consumption.
- In this research the causes of occupants' behavior are analyzed by machine learning algorithms, which implied the capability of making building control systems to 'understand' its occupants. E.g. in this case the predictor built could predict a certain occupant's reaction to a certain circumstance, then in the future the building control system may be able to do it for him/her automatically. The effective interaction between the building and occupants would play an important role in future *intelligent building* design.

Chapter 6 Business Potential[1]

This chapter contains the discussion about the business potential of machine-learning-based products in the built environment industry.

In the previous chapters, it has been pointed out that the potential applications for the state-of-art machine learning technology in the built environment domain mainly fall into two categories: energy saving and intelligent building design. The former concentrates on improving building energy consuming strategy to use energy more efficiently while the latter is more about developing adaptive indoor control agents to satisfy the unique preferences or habits of occupants, in order to improve indoor comfort. In the following paragraphs, one case application that has been already pushed into market will be presented for each application.

Energy Disaggregation

Today, more and more building designers seek for the possibility to further save energy with the help from artificial intelligence. Energy disaggregation is one of their efforts, it allows us to take a whole building aggregated energy signal and separate it into appliance specific data via a set of statistical and machine learning approaches. E.g. with this technique, you could take a single measure of household power use -- say, smart meter data, or measurements from a single home sensor -- and get to know exactly how much power your air conditioning, water heating, kitchen appliances, household lighting and all other consumer electronics are using respectively.

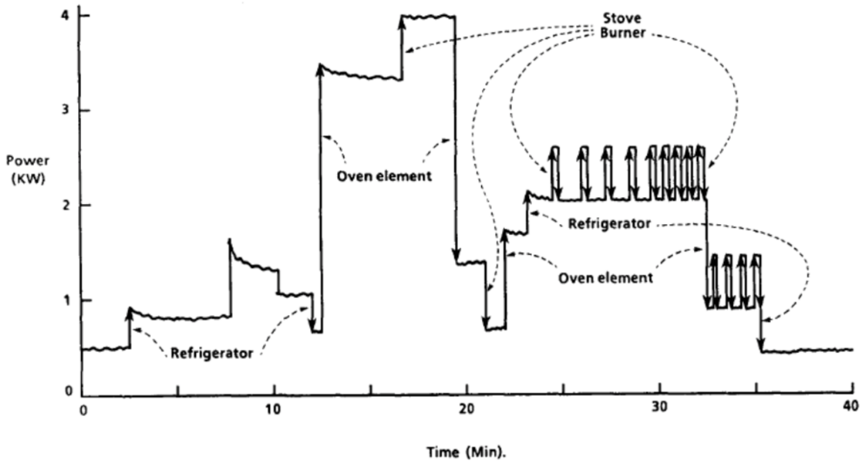


Figure 6-1 An energy disaggregation example [22]

This technology captures value from being able to break out discrete energy loads to adapt to occupants' using patterns, diagnose inefficiencies or imminent equipment

[1] This attachment is for the EIT-KIC Innoenergy program

failures, through algorithms and analytics.

In the recent years, many startups focusing on this technology emerged worldwide, such as Bidgely, which has closed a $16.6 million Series B round in 2015 to expand its HomeBeat energy management platform [2] based on energy disaggregation technology. With this energy management platform, the company claims a peak load reduction of 25% and energy efficiency increase of 7.7% [23]. Other relevant startups include PlotWatt, Navetas, Energy Aware and also a KIC-Innoenergy owned Enervalis. In addition, some big companies including Belkin[3] and Intel[4] are also involved in this field since 2010.

By accomplishing peak load reduction and energy saving, the potential profitability of energy disaggregation has been validated by its investors, especially when extra sensor for each appliance is not needed, but the classification is done with the help of machine learning algorithms. The basic principle of this today's so-called software-based energy disaggregation is the same as the methodology we used for this thesis project: machine-learning classifier and clustering algorithms.

Smart windows

The buildings today are often equipped with sophisticated sensing and control systems. However, the performance of such systems is often hindered by reactive control systems based on simple thresh holding or scheduling thus fail to adapt to the preferences of individual users.

For tasks such as heating, cooling, lighting or ventilation, information does exist that can be used to enhance performance and efficiency by predicting future states (outdoor temperature, occupancy history, for etc.) and learning occupants' personal preference, but it is not extensively used today.

E.g., via the procedures of feature selection and clustering stated in Chapter 4, the building control systems are able to 'understand' its occupants. In this case the predictor built could predict a certain occupant's reaction to a certain circumstance, then combined with the predicted future states, the building control system may be able to react automatically in advance, in the way its occupants prefer. The effective interaction between the building and occupants would play an important role in future intelligent building design.

Relevant startups are emerging, such as View[5], a California-based startup who makes electro-chromic windows, which are sometimes called "smart windows" because they

[2] http://www.greentechmedia.com/articles/read/bidgely-raises-16m-for-energy-disaggregation
[3] https://gigaom.com/2010/11/16/belkin-and-power-map-bet-on-one-plug-approach/
[4] https://www.greentechmedia.com/articles/read/intel-tests-whole-home-smart-power-sensors-in-texas
[5] http://viewglass.com

can be programmed to absorb a different amount of light throughout the day in order to cool or warm up a room according to both the environment condition and occupants' preference.

View Dynamic Glass is powered by a sophisticated intelligence engine that adapts to your environment and you. It automatically responds to outdoor conditions and people's preferences, or can be controlled individually from a mobile device.

Figure 6-2 A demo smart building by View

In the perspective of finance, the round of funding, from Madrone Capital Partners, is a mix of equity and debt for this Milpitas-based startup, which focuses primary on the U.S. market but is expanding its effort to Europe and Asia. Since inception in 2007, View has raised over $300 million in total.

As for the projects, the company counts the January 2012 installation at the Palo Alto office of software developer SAP as its first commercial-scale project. Since then, the company has seen roughly 50 installations of its smart windows in North America, at locations that include hotels, hospitals, government buildings and colleges[6].

Conclusion

From the previous cases introduced, it could be concluded that both the engineers and investors are optimistic upon the effectiveness and profitability of machine learning based products within the built environment industry. Since it is basically a software-based business model, the cost would be diluted via the effect of scale while the profit, which could come from the energy saving, living environment improvement and incentives, would continue increase with the popularization of this novel technology.

[6] http://www.forbes.com/sites/uciliawang/2014/01/07/a-startups-100m-plan-to-make-dumb-windows-smart/#56eba98e106c

Acknowledgement

This work is a master thesis project conducted in the Department of the Built Environment, Eindhoven University of Technology, the Netherlands. The author very appreciated dr.ir. M.G.L.C. (Marcel) Loomans of TU/e for sharing the dataset and providing explanation about the monitoring program. During the project phase, the author received enlightening suggestions from prof.ir. W. (Wim) Zeiler, ir. G. (Gert) Boxem and dr. Y. (Yang) Zhao. Also, ir. Hussain Kazmi, data scientist from Enervalis, helped with improving the data pipeline structure. This work could not be finished without the contribution from each of them.

References

[1] Eurostat, Energy consumption in households, 1995 survey, European Commission, 1999.

[2] W. Feist, et al., Passivhaus Projektierungs Paket 2002, Anforderungen an qualita¨tsgepru¨fte Passivha¨user, Passivhaus Institut, Darmstadt, 2002.

[3] Cost Efficient Passive Houses as European Standards http://www.cepheus.de/eng/index.html

[4] The report of monitoring program Kroeven 2013

[5] Fan, Cheng, Fu Xiao, and Chengchu Yan. "A framework for knowledge discovery in massive building automation data and its application in building diagnostics." Automation in Construction 50 (2015): 81-90.

[6] Khan, Imran, et al. "Fault detection analysis of building energy consumption using data mining techniques." Energy Procedia 42 (2013): 557-566.

[7] D'Oca, Simona, and Tianzhen Hong. "A data-mining approach to discover patterns of window opening and closing behavior in offices." Building and Environment 82 (2014): 726-739.

[8] Fan, Cheng, Fu Xiao, and Shengwei Wang. "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques." Applied Energy 127 (2014): 1-10.

[9] Ren, Xiaoxin, Da Yan, and Tianzhen Hong. "Data mining of space heating system performance in affordable housing." Building and Environment 89 (2015): 1-13.

[10] Dodge, Yadolah, ed. Statistical data analysis based on the L1-norm and related methods. Birkhäuser, 2012.

[11] Endrenyl, L. (1978), Statistics. by David Freedman, Robert Pisani, Roger Purves. W. W. Norton & Co., Inc., New York, 1978. xv + 506 + A83 pp., U.S. $13.95. ISBN 0-393-09076-0. [Instructor's Manual, 135 pp. ISBN 0-393-09041-8.]. Can J Statistics, 6: 137. doi: 10.2307/3314838

[12] Van Vliet, Lucas J., Ian T. Young, and Piet W. Verbeek. "Recursive Gaussian derivative filters." Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on. Vol. 1. IEEE, 1998.

[13] Hosmer Jr, David W., and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.

[14] Andrew Moore: “K-means and Hierarchical Clustering - Tutorial Slides” http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html

[15] UNC Edge Detector 1D. University of North Carolina at Chapel Hill http://www.cs.unc.edu/~nanowork/cismm/download/edgedetector/

[16] Wei, Shen, et al. "Analysis of factors influencing the modelling of occupant window opening behavior in an office building in Beijing, China." (2015).

[17] Calì, Davide, et al. "Analysis of occupants' behavior related to the use of windows in German households." Building and Environment 103 (2016): 54-69.

[18] Andersen, Rune Vinther, Bjarne W. Olesen, and Jørn Toftum. "Modelling window opening behaviour in Danish dwellings." Proceedings of indoor air (2011).

[19] Shi, Shanshan, and Bin Zhao. "Occupants’ interactions with windows in 8 residential apartments in Beijing and Nanjing, China." Building Simulation. Vol. 9. No. 2. Tsinghua University Press, 2016.
[20] Andersen, Rune, et al. "Window opening behavior modelled from measurements in Danish dwellings." Building and Environment 69 (2013): 101-113.
[21] Wong, Johnny KW, Heng Li, and S. W. Wang. "Intelligent building research: a review." Automation in construction 14.1 (2005): 143-159.
[22] Carrie Armel, “Energy Disaggregation”, Precourt Energy Efficiency Center, Stanford December 2011.
[23] Bidgely Official Page http://www.bidgely.com/solutions/