## A Data Mining Approach to Study Occupant Behavior Motivation

Xinyuyang Ren

Supervisors:

prof.ir. W. (Wim) Zeiler,

ir. G. (Gert) Boxem

dr. Y. Zhao

## ABSTRACT

This article introduced a data-based methodology to investigate the motivation behind people’s behavior, followed by a case study on people’s adjustment of ventilation system in a Dutch community. In the individual level, a machine learning based approach was applied to predict occupants’ behavior and reveal the motivation behind. In the community level, the results from different occupants were compared and discussed, aimed at seeking for possible behavior pattern among different people.

#### INTRODUCTION

Motivation

The real energy consumption of building depends 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. However, so far the occupant behavior has not been adequately modeled when calculating the expected performance of building. Consequently, field test studies all over Europe have shown discrepancies between real and expected performance of building [17] [18].

Also, in the frontier 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.

Therefore, an effective behavior model would contribute in more than one aspect to the advance of built environment research. Before the occupants’ behavior could be sufficiently modeled, it is critical to understand it from real records.

case description

*De Kroeven* in Roosendaal is a housing stock built around 1964. Between April 2010 and April 2011, it was completely renovated on the basis of passive house principles with comprehensive energy reduction measures, including a very good insulation shell, an effective sealing of cracks and a balanced ventilation system with heat recovery for each house. As the result, the energy consumption should decrease 60%-70% compared with before [4].



Figure De Kroeven Community

After the finish of the renovation work, in order to test if the presumed performance has been reached, a monitoring program was launched. Between the year 2013 and 2015, varies of sensors were installed in 10 experimental houses and recorded the information regarding the domestic energy consumption, indoor environment as well as system running parameters etc. This database is used to conduct the study introduced in this article.

Table De Kroeven Monitoring Program Database

|  |  |  |
| --- | --- | --- |
| **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** | Indoor Temperature [℃] | 3 min |
| Relative humidity [%] | 3 min |
| Concentration [ppm] | 3 min |
| Ventilation System Supply Air Temperature [℃] | 3 min |
| **Occupant Behavior** | Ventilation Panel Interaction | / |

Research question

The ventilation system installed in *De Kroeven* community has 3 different positions for ventilation flowrate, which is adjustable by occupants through a control panel. The occupants’ interaction with the ventilation control panel is selected as the case study. Then the main research questions could be listed as

* **Question 1** What is the motivation for an occupant to increase/decrease ventilation flowrate?
* **Question 2** For different occupants, whether do they behave in the same way and, how similar/different are they?

#### METHDOLOGY

To answer the research questions raised in the previous chapter, different techniques are introduced respectively.

The reason why people are adjusting the panel could be seen as a feature selection question in the perspective of data mining. Mathematically it’s possible to build a model to predict people’s behavior under a certain circumstance, then the features that are more informative could be evaluated quantitatively. In the machine learning domain, an L1-regularized logistic regression is a robust solution for feature selection.

In the community level, the task would be compare among different people and find similarities/differences: it’s called clustering in the data mining domain. This kind of algorithms, such as widely-used K-means, could group different samples into several clusters with the best optimized similarity inside each cluster and difference among different clusters.

In this section, the technique involved would be briefly introduced.

L1-regularized 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].

This is a standard linear regression formula

where x is a series of features, is a vector containing coefficients for each feature and 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

Finally, the equation of logistic regression becomes

The function is plotted in below Figure 2.1. It could be observed that the range of logistic regression output is between 0 and 1. A threshold, say 0.5 could be chose to divide two different categories (i.e. if output < 0.5, predict the case to be in category 0, else predict category 1).

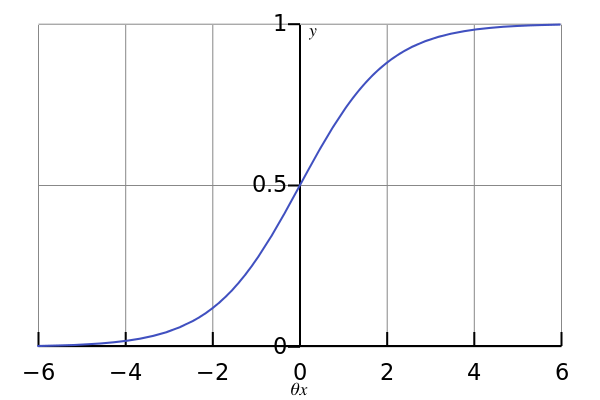


Figure Logistic Regression Output

After training with the dataset, which aimed at finding optimized to minimize the cost function

the model is adjusted to minimize the prediction error based on the training set and the coefficients of each feature, i.e. the could then be used to evaluate the relative importance of each feature in its classification process.

In addition, in this project the logistic regression kernel used 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.

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

The logistic regression runs repeatedly with different to make a *grid search*. Finally, from the model with best cross validation accuracy, the most informative feature combination could be determined, which implies the motivation of occupants’ behavior.

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.

Specifically, the algorithm follows a simple 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.

where 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.

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].

**WORKFLOW**

Below Figure 3 shows the overall logic design, or the *data pipeline* of this study. It describes generally how will the data stream ‘flow’ throughout the whole process and defines the basic blocks and their own functionalities.

Firstly, the related dataset stated in table 1, including weather data, indoor environment data and occupant behavior records, was extracted from the monitoring program database. After essential data cleaning and mapping, the logistic regression model was then trained to find the motivation combination. Finally, the motivation sets from different people were compared and grouped into several occupant profiles.

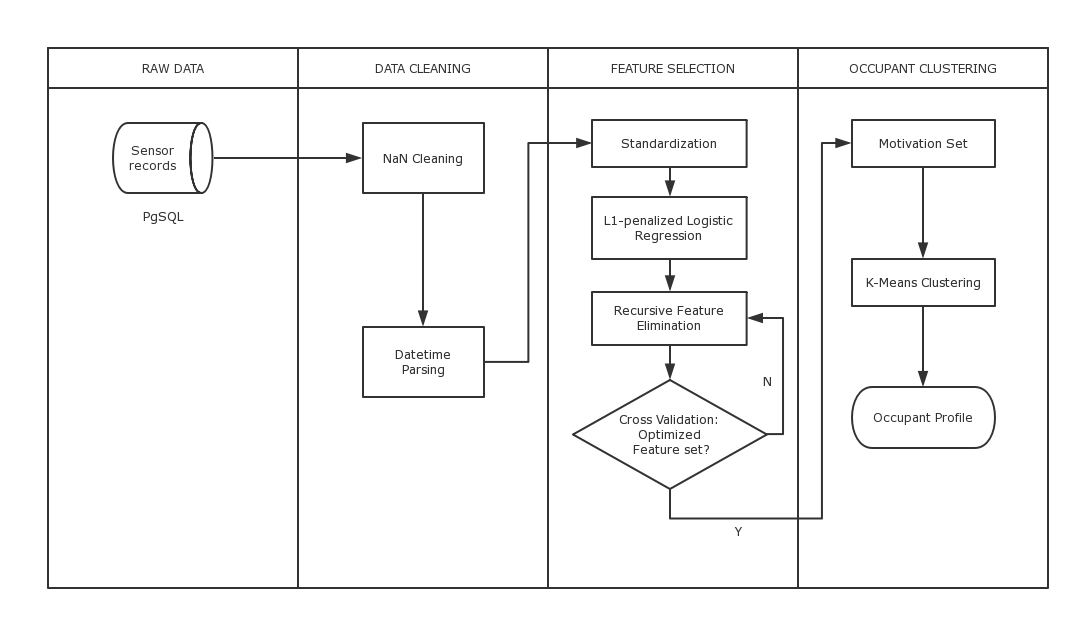
****

Figure the Data Pipeline

##### MOTIVATION INVESTIGATION AND OCCUPANT PROFILING

This chapter discuss the motivation analysis result in the individual level.

**MOTICATION INVESTIGATION**

Firstly, the training set was standardized*,* which means all the features are rescaled into zero-mean and unit-variance distributions, so that the coefficient of the linear model used could indicate the relative importance of the feature it corresponds.

Then the normalized dataset is fed into a L1-penalized logistic regression classifier, which will try to predict the occupants’ reaction in a certain circumstance. 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, Figure 10 below shows an example of the coefficients output for the classier.

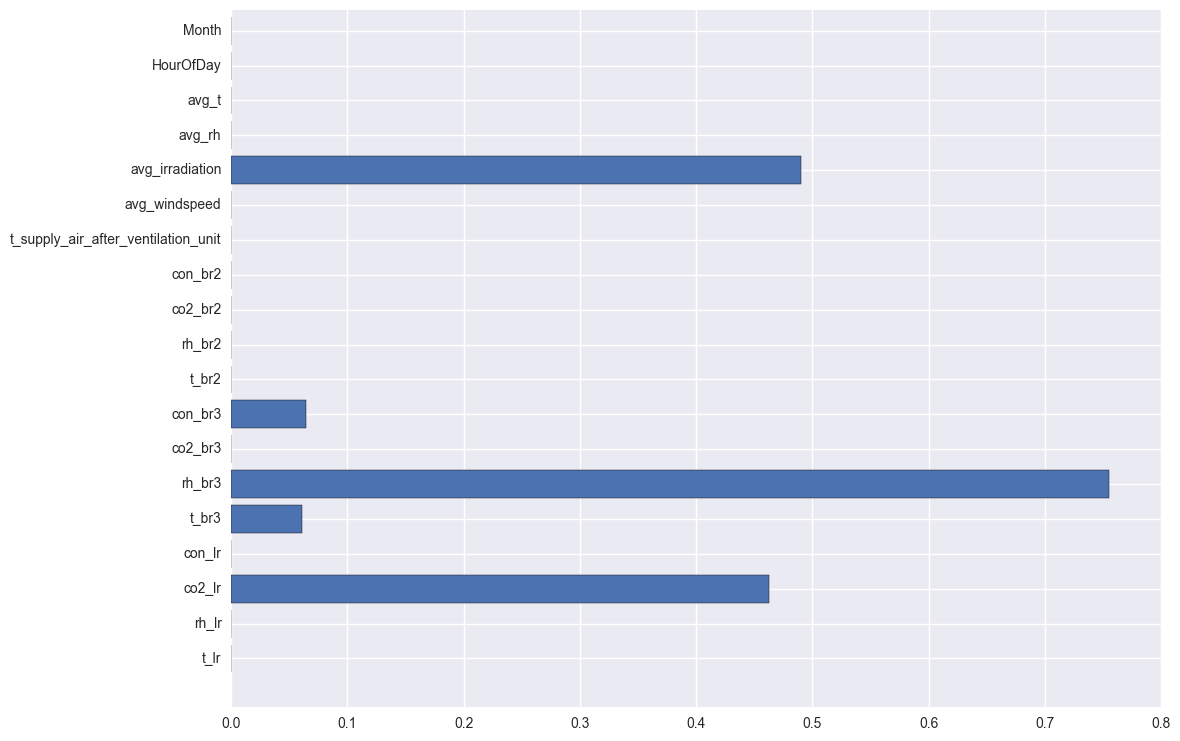


Figure 3.5 Feature importance output of increase operation

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 2

Table Abbr. Explanation

|  |  |
| --- | --- |
| **Abbreviation** | **It stands for** |
| openwin | No. of opened windows in the house |
| 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 flow rate.

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.

#### MOTIVATION PATTERN

The motivation investigation discussed in the previous section is to reveal the main driven factors of occupants’ behavior in the individual level. It could be expected different people should hold different preferences and not likely to behave in the same way. Thus in this section the analysis is expanded to the community level. By a clustering process, occupants were grouped into several motivational behavior patterns.

The most informative feature set for each occupant, with its coefficients, is extracted from the output of logistic regression model. 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. With essential re-scaling, the 10 occupants took part in the experiment could be plotted into 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.



Figure 4.3 Cause pattern of ventilation system operation

Indicated by K-Means algorithm, 3 different types of occupants could be observed:

* Indoor environment sensitive occupants: 2, 4, 6, 8
* Time sensitive occupants: 7, 9
* Mixed type occupants: 1, 3, 5, 10

The complexity of occupants’ motivational behavior pattern could be seen from the data analysis result. The Indoor environment sensitive occupants are more likely to interact with their ventilation control panel when they feel slightly 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.

## Conclusion and Discussion

In the previous chapters, details of the data mining framework developed in this research was elaborated with a case study *De Kroeven*. 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 *1st 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

1. 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.
2. 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.
3. 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.

## 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/