A Data Mining Approach to Study Occupant Behavior Motivation

Xinyuyang Ren, Yang Zhao

Keywords: *Data Mining, Occupant Behavior, Motivation Pattern*

ABSTRACT

This study proposed a data-based method to investigate the occupants' behavior, supported by a case study of analyzing people's adjustment on ventilation system. In the individual level, a logistic regression based approach was applied to classify occupants' behavior of increasing/decreasing the ventilation flowrate and reveal the motivation behind. In the community level, the behavior motivation derived from different occupants were compared and three motivational behavior patterns were summarized.

INTRODUCTION

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 occupants' behavior. However, so far the occupant behavior has not been adequately modeled when calculating the expected performance of a building. Consequently, field test studies have shown discrepancies between real and expected performance of building [1] [2].

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 [3]. With a better understanding of people's behavioral pattern, the building control system could generate tailored strategies for its occupants.

Therefore, building an effective behavior model could contribute in more than one aspect to the advance of built environment. Before which, it is critical to understand occupants' behavior and their motivation from real records.

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. As the result, the energy consumption should decrease 60%-70% compared with before [4].

After the renovation, to test if the presumed performance has been reached, a monitoring program was launched. Between the year 2013 and 2015, sensors were installed in 10 experimental houses and recorded varies of information including the domestic energy consumption, indoor environment as well as people's operation on light/ventilation etc. A part of this database, introduced in *Table 1*, is used to conduct the study introduced in this article.

Table 1 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	Increase/decrease ventilation flow on control panel	/

The occupants' interaction with the ventilation control panel is chose for the case study. The two research questions listed would be answered

- **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?

METHODS

To find the reason why people adjust the ventilation 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 and then quantitatively evaluate the importance of each feature. L1-regularized logistic regression is a robust solution for this purpose by practice.

Up to the community level, comparing among different samples and grouping ones with similarities is 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 in-cluster similarity and inter-cluster difference.

In the following of this chapter the technique mentioned above will be briefly introduced.

Logistic regression [5], 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.

This is a standard linear regression formula

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

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

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

Finally, the equation of logistic regression becomes

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

The function is plotted in *Figure 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).

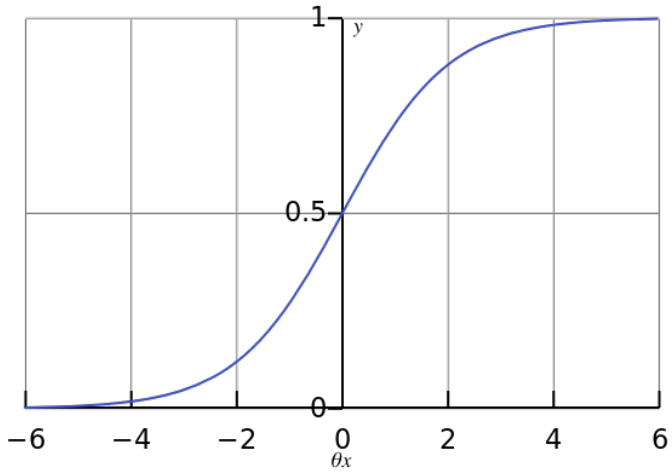

Figure 1 Logistic Regression Output

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

$$\mathrm{J}(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log\left(h_\theta x^i\right) + \left(1 - y^i\right)\log\left(1 - h_\theta x^i\right)$$

the model is adjusted to minimize the prediction error based on the training set and the coefficients of each feature.

Based on its linear nature, the coefficient of each feature in a trained logistic regression model is widely used to evaluate the importance of this feature. The effectiveness, interpretability and robustness of this approach have been validated by many peer researchers [1] [2] [6] [7] [8].

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. The model runs repeatedly with different λ to make a *grid search*. Finally stopes at the parameter combination that gives the best cross validation accuracy,

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log\left(h_\theta x^i\right) + \left(1 - y^i\right)\log\left(1 - h_{\square} x^i\right) + \lambda \sum_{i=1}^{n} |\vartheta_i|$$

As linear model penalized with L1 norm tends to give sparse solutions i.e. many of its estimated coefficients would be zero, thus it will make the feature selection more significant [9].

K-means clustering [10] is one of the simplest unsupervised learning algorithms that solve the 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.

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

where $\left\| x_i^{(j)} - c_j \right\|$ 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 [11].

Below Figure 2 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.

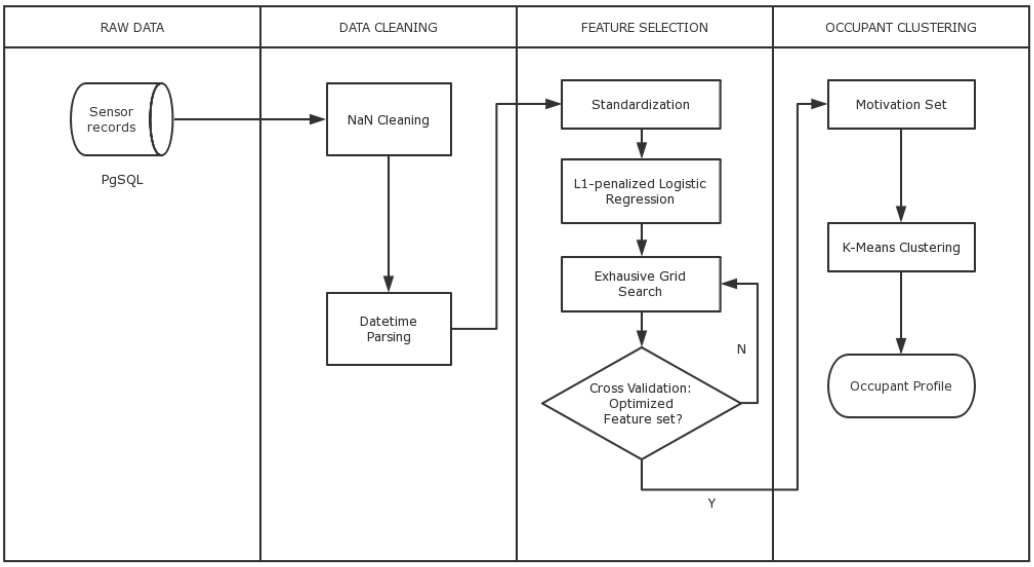

Figure 1 the Data Pipeline

RESULT AND DISCUSSION

The model is designed to predict whether an increase/decrease adjustment on the ventilation system will happen according to information including time and indoor environment etc.

Firstly, the training set was standardized, which means all the features are rescaled into zero-mean and unit-variance distributions. Then the dataset is fed into a L1-penalized logistic regression classifier, which will optimize the cost function to predict occupants' reaction in a certain circumstance. As the feature scale is standardized, the coefficient of the linear model trained could indicate the relative importance of the feature it corresponds to. For example, *Figure 3* shows the motivation factor importance for occupant no.1, with the model cross-validated precision reached 86%.

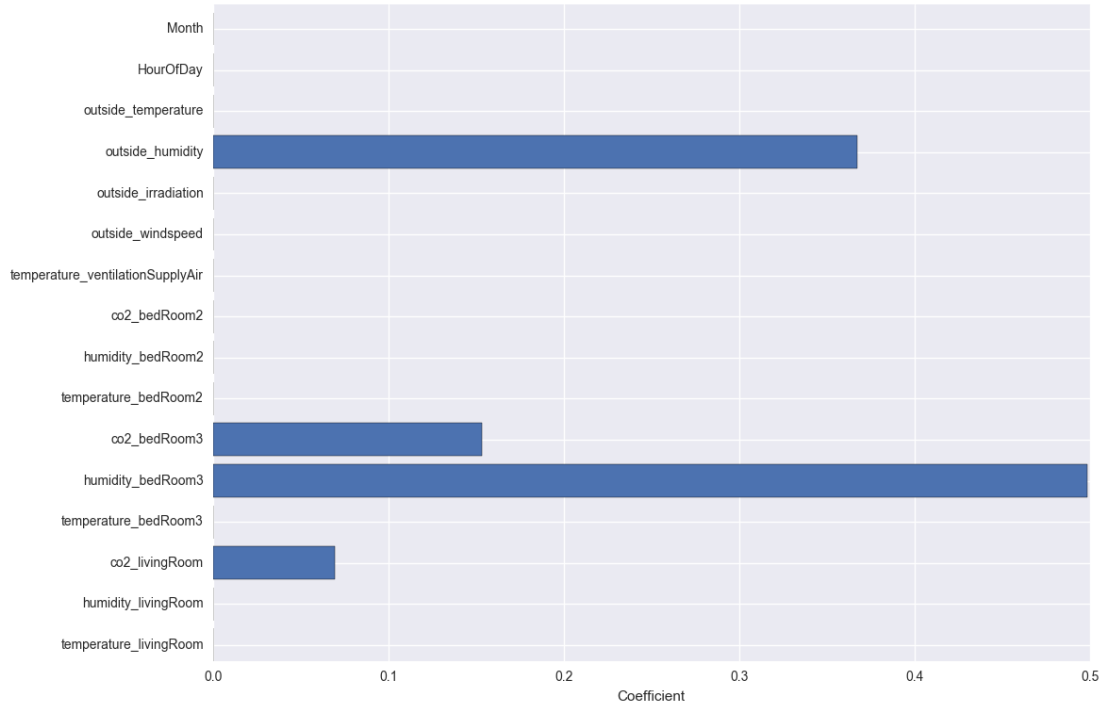

Figure 2 Feature importance output

It could be observed the less informative features for this occupant were filtered out with zero coefficients, while the remaining indicates the indoor CO2 concentration and humidity are the most important motivational drivers for this occupant to adjust the ventilation flow rate.

Train similar models for every occupant could reveal the main motivational driven factors for their 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, expanded to the community level, a clustering analysis could group occupants 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. All the main driven factors fall into two categories: *time-related factors* including month, weekday/weekend, hour of day info and *environment-related factors*, including indoor temperature, relative humidity, CO2 concentration and outside weather info. According to those two dimensions and with essential re-scaling, the 10 occupants took part in the experiment could be represented in *Figure 4*. The horizontal axis represents the importance of *indoor environment factors* in determining occupants' behavior, while the vertical axis represents the importance of *time-related factors*.

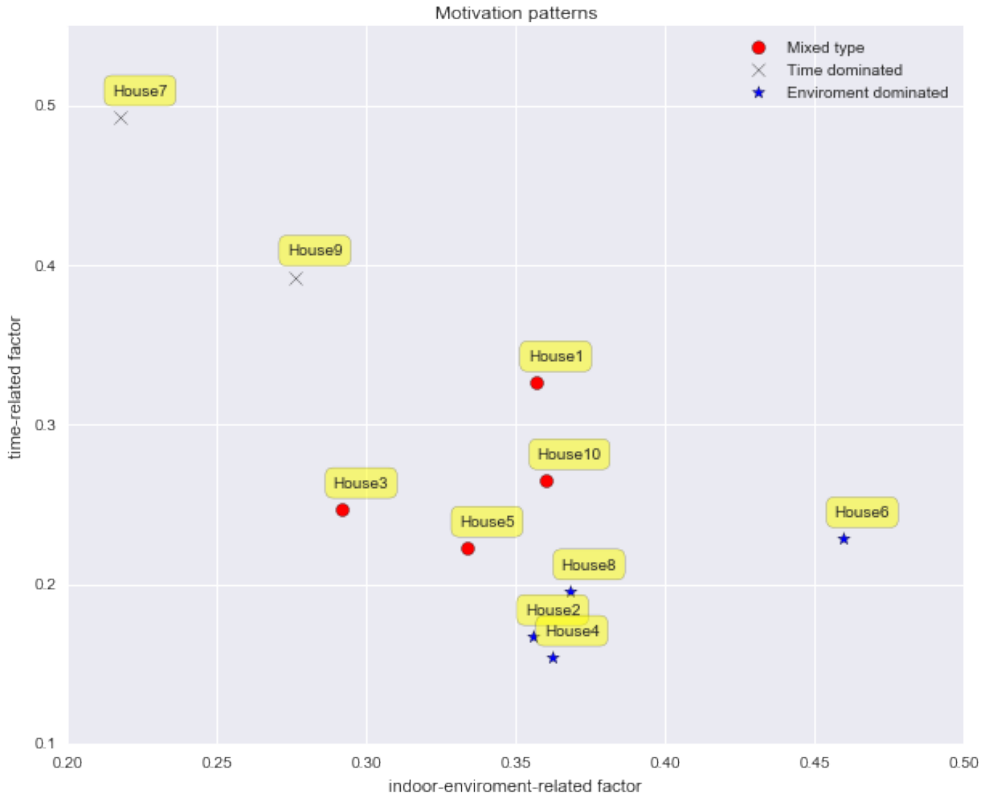

Figure 3 Motivation pattern of ventilation system operation

K-Means algorithm indicates 3 different types of occupants:

- Indoor environment sensitive occupants (plotted in star): 2, 4, 6, 8
- Time sensitive occupants (plotted in cross): 7, 9
- Mixed type occupants (plotted in dots): 1, 3, 5, 10

The complexity of occupants' motivational behavior pattern could be observed from the data mining 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 behave with fixed timetables (e.g., as soon as they wake up or come back from work etc. they adjust the ventilation). Of course, there are also some people in between, as mixed-type occupants their behaviors are effected considerably by both factors in the same time.

CONCLUSION

In this study, a data mining framework was implied to study the occupant behavior of adjusting the ventilation flow in a recently-renovated community in the Netherlands. The objective is to reveal the hidden motivation behind occupants' behavior and seek for possible behavior patterns among different people. A *L1-regularized 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 a bigger picture, the comparison among different occupants indicated 3 unique motivational patterns. Namely the *environment-driven* type, corresponds the occupants who are more sensitive to the environmental factors. *Time-driven* type, corresponds to the occupants

who hold relative fixed temporal habits. As well as *mixed-type* occupants, whose behavior is more randomized with no single preference pattern which is clear enough on environment and temporal factors.

The data-based method to investigate occupants' behavior introduced in this study enabled new possibility to leverage the BMS data. The learning drawn from the study could be used either to model people's behavior more precisely in the building simulation program as well as to contribute to the improvement of intelligent building. Also, besides the traditional approaches to investigate people's behavior by conducting a survey or interview, the algorithmic method is more robust with less man-made disturbances.

REFERENCE

[1] Calì, Davide, et al. "Analysis of occupants' behavior related to the use of windows in German households." Building and Environment 103 (2016): 54-69.
[2] Andersen, Rune Vinther, Bjarne W. Olesen, and Jørn Toftum. "Modelling window opening behavior in Danish dwellings." Proceedings of indoor air (2011).
[3] Wong, Johnny KW, Heng Li, and S. W. Wang. "Intelligent building research: a review." Automation in construction 14.1 (2005): 143-159.
[4] The report of monitoring program Kroeven 2013
[5] Hosmer Jr, David W., and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.
[6] 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.
[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] Andersen, Rune, et al. "Window opening behavior modelled from measurements in Danish dwellings." Building and Environment 69 (2013): 101-113.
[9] Dodge, Yadolah, ed. Statistical data analysis based on the L1-norm and related methods. Birkhäuser, 2012.
[10] Andrew Moore: "K-means and Hierarchical Clustering - Tutorial Slides"
http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html
[11] 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.
[12] Wei, Shen, et al. "Analysis of factors influencing the modelling of occupant window opening behavior in an office building in Beijing, China." (2015).
[13] 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.
[14] Khan, Imran, et al. "Fault detection analysis of building energy consumption using data mining techniques." Energy Procedia 42 (2013): 557-566.
[15] Ren, Xiaoxin, Da Yan, and Tianzhen Hong. "Data mining of space heating system performance in affordable housing." Building and Environment 89 (2015): 1-13.
[16] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research 12.Oct (2011): 2825-2