WINDSOR CONFERENCE
*Rethinking comfort*
*12th - 15th April 2018*

# Using feature selection techniques to determine best feature subset in prediction of window behaviour

**Hailun Xie[1], Shen Wei[2], Li Zhang[1], Bobo Ng[1], Song Pan**[3]

[1] Faculty of Engineering and Environment, Northumbria University, Newcastle, UK, hailun.xie@northumbria.ac.uk;
[2] The Bartlett School of Construction and Project Management, University College London, London, UK, shen.wei@ucl.ac.uk;
[3] College of Architecture and Civil Engineering, Beijing University of Technology, Beijing, China, pansong@bjut.edu.cn;

**Abstract:** Previous studies have demonstrated diverse effects of different factors on occupant window behaviours. It is necessary to choose appropriate subsets of different behavioural window opening features, and to eliminate irrelevant and redundant features so as to avoid overfitting, noise and random fluctuations being learned by the model, and improve the accuracy of predictive models of window opening. The choice of protocols for the selection of features has been widely accepted as one of the most important steps in developing machine learning prediction algorithms. This study employed the use of both a recursive and a non-recursive feature selection method designed to consider all influencing factors simultaneously to explore the confounding effects inherent in various factors pertaining to the prediction of window opening behaviour. Two machine learning algorithms were applied as estimators in a recursive selection process, namely support vector classification (SVC), logistic regression (LR), and one in a non-recursive process, namely random forest (RF). Additionally, two processing schemes in the recursive method analysis were tried to determine the optimal feature subset based on corresponding algorithms, namely recursive feature elimination (RFE) and recursive feature elimination with cross validation (RFECV). Seven factors were considered in the feature selection process based on collected data, including: indoor temperature, outdoor temperature, relative humidity, concentrations of PM2.5, air quality index (AQI), wind speed and wind direction respectively. The results showed that different feature subsets can generate different prediction accuracy within the recursive method. RFECV can determine the most appropriate feature subset effectively with the consideration of the correlation among various factors. Both LR and SVC were proved to be effective as estimators embedded in RFECV, however SVC is more computationally expensive and LR shows a larger variance within the feature subset space. RF, as a non-recursive method, demonstrated real advantages in eliminating redundant features compared to the recursive feature selection process.

**Keywords**: Feature Selection; Recursive Feature Elimination; Cross Validation; Logistic Regression

## 1. Introduction

### 1.1. Study on window behaviour

Occupants in buildings can behave in a wide range of ways to maintain their comfort levels due to the many different adaptive opportunities they have to hand to adjust their thermal environment including the use of: thermostatic valves, HVAC system set points, window, blinds, shades operation, and plug loads. The ‘dark side’ of occupant behaviour in terms of building energy consumption can also result from inactivity and studies showed that in some commercial buildings 56% electricity was consumed during non-working hours due to leaving lights and equipment on at the end of day (Masoso and Grobler, 2010). Occupant behaviour can affect building energy consumption to an extent similar to that exerted by mechanical control, based on a study of experiment conducted in Switzerland (Filippín et al., 2005, Haas et al., 1998). Whether energy-saving strategies and technologies performs as expected has been found to largely depends on how occupants understand and interact while the building is in use (Yan et al., 2017). Therefore, to gain a better understanding the role played by occupant behaviour in the energy performance of a particular building is crucial for bridging gap between real and predicted energy performance.

The key intervention between perceived indoor environmental quality in buildings and the climate outdoors is the building envelop. Some parts of that envelope are immovable and some, the windows can open to allow the mixing of outdoor and indoor air. As a consequence, window operation is one of the most efficient strategies for producing a desired indoor micro-climate (D'Oca and Hong, 2014). Extensive studies focusing on various aspects of window opening behaviour have been conducted, including window behaviours under different heating or cooling mode, in different building types, and across different countries(Pan et al., 2018, Wei et al., 2013, Wei et al., 2014). According to previous studies, window behaviours can be influenced by a wide range of factors both "external" to occupant itself, (e.g. air temperature, air quality), and internal or "individual" (e.g. personal background, attitudes, preferences), and building properties (e.g. HVAC systems, ownership, building type). All drivers of window behaviour can be divided to five general categories: Physical (indoor and outdoor environment); Psychological (preferences, attitudes); Physiological (age, sex); Contextual (type of environment where the occupants are located); and social (income, lifestyle) (Fabi et al., 2012).

Statistical analysis has been extensively used to analyse associations and relationships among these various factors influencing building performance and occupant behaviour (D'Oca et al., 2014). Fritsch et al. developed a window opening angle predicting model related to outdoor temperatures in winter based on Markov chains in 1991 (Fritsch et al., 1990). Nicol was the first one using the method of probability distribution to predict window opening behaviour as logit functions of outdoor temperature (Nicol, 2001). Haldi and Robinson adopted three different methods, logistic regression, Markov chain, and random process, to make predictions on window behaviour respectively (Haldi and Robinson, 2009a). Based on previous extensive researches, drivers of window behaviour contain various factors both in numerical and categorical formats. In fact, different factors actually demonstrate different levels of importance in terms of the interrelationship with window behaviour. A predictive model with more variables doesn't necessarily represent a model with better predicting performance. On the contrary, the inclusion of more factors in predictive modelling probably leads to the increase of model dimensionality, which would lead to a higher risk of overfitting problems, especially with limited sample sizes (De Silva and Leong, 2015). Therefore, it is very necessary to adopt a feature selection process to select of the more or most relevant factors and remove irrelevant, redundant, or noisy information in order to avoid the overfitting problem, noise and random fluctuations being learned in the model, in the process of predicting window behaviour.

### 1.2. Feature selection in window behaviour modelling

When it comes to the issue of the feature selection process in the prediction of window opening behaviour, it appears that strategies for choosing key features have been largely undiscussed in previous studies. Based on feature selection issues as raised in current studies three general categories stand out, which also reflect three problems of feature selection modelling of window behaviour.

Problem 1: The criterion and details for selecting a suitable feature subset in the prediction of window behaviours are not clear or thoroughly illustrated in some studies, which would make the selection results unsolid and increase the uncertainty about achieving an optimal prediction performance or being able to confidently compare results between parallel studies. One of the typical example is as follows: D'Oca and Hong (D'Oca and Hong, 2014) employed logistic regression to identify factors influencing window behaviour with monitored data from 16 private offices. Coefficients of all applied variables in logistic regression were calculated for each office. Somehow, they provided no further explanation about how to decide which feature was chosen based on these coefficients in logistic regression. However, several conclusions were made without specific description and analysis, for example, indoor air temperature, arrival time, occupant presence, time of day and outdoor temperature are some of the main factors influencing window opening behaviours. However, it still remains unclear and dubious about whether the results of selected feature subset in this study provide the optimal solution or not, because no criterion of selection was demonstrated in this study.

Problem 2: Feature selection processes in previous studies generally failed to take into account the collective effects of various factors on window behaviour simultaneously. Features in the prediction model were selected mainly by analysing and measuring the statistical correlations for every factor separately with window behaviour. By evaluating different factors separately, this correlation analysis cannot measure the confounding effects on window behaviour inflicted by the collective interaction of all factors.

In another typical example Herkel et al. (Herkel et al., 2008) carried out a study of window opening behaviour in 21 south-facing offices in Germany, in which seasonal effects, outdoor temperature, indoor temperature, time of the day and building occupancy were considered. Each factor was analysed and discussed separately to evaluate its significance to window behaviour. Then outdoor temperature and user occupancy depending on the time of the day were selected to construct a user model. Despite the elaboration in Herkel's study, it failed to take into account collective and confounding effects of various factors due to the separation of different variables, which made this study unable to determine the best feature subset capable of  achieving the highest prediction accuracy.

Problem 3: Few studies on window behaviour modelling formulate a search strategy which can be effectively applied to deal with a large number of factors in the process of feature selection. For very limited number of studies which include feature selection processes before establishing their prediction model, each one only conducted several tests of intentional combinations of two or three factors, rather than proposing a complete and viable method to execute the possible feature combinations in the whole feature space. In 2009, Haldi and Robinson (Haldi and Robinson, 2009b) conducted a comprehensive study of interactions with window opening behaviours by office occupants based on seven years of continuous measurements and three modelling approaches. When dealing with feature selection, Haldi and Robinson adopted a 'wrapper method' using different attempts at including univariate, multivariate and polynomial logistic models to determine the better feature subset. Their feature selection process was wrapped inside the process of model training so that selected features can maintain its conformity to the calculation of the prediction algorithm to achieve a better performance. This is so far the most complete study on feature selection in prediction of window behaviour based on logistic regression. However, the researchers only made several trials of combining some factors with best relevance rather than provided a search strategy for feature combination in multivariate regression process, which makes this research unable to examine the confounding effects among factors attached with different importance levels. The criterion for determining the best model is dependent on parameters of goodness-of-fit, which can show a good performance in the model training stage, but it may not work effectively on a new dataset.

### 1.3. Aim of the study

In general, most feature selection processes in previous studies only considered each feature separately, thereby feature dependencies and redundancies could not be analysed, which may reduce their classification performance when compared to other types of feature selection techniques. Therefore, the study of window opening behaviour prediction currently lacks systematic feature selection techniques and protocols. In order to make the prediction models more accurate and lay a solid foundation for the application of far more complicated prediction algorithms in future, this study aims to make practitioners of window behaviour prediction aware of the necessity of feature selection and demonstrate both a recursive and non-recursive feature selection method, which can consider all influence factors simultaneously so as to take into account confounding effects among various factors. Two algorithms were used as estimators in recursive selection process, namely support vector machine (SVM), logistic regression (LR), and one in non-recursive process, namely random forest (RF). Cross validation and non-cross validation methods, recursive and non-recursive methods are discussed and compared based on the training results of the real-life data.

## 2. Research Methods

### 2.1. The data set

Data on window behaviour was collected based on an office building in Beijing University of Technology (BJUT). The field monitoring was conducted during two transitional seasons in 2014, from 16th March to 30th April, so that data of how occupants operate windows can be obtained without the interference of air conditioning systems. Five offices, each with two south-facing gliding windows as shown in Figure 1, on the first floor were chosen to monitor for occupancy (1min interval), window state (1min interval) and indoor temperature ($T_i$, 5min interval). Simultaneously, outdoor parameters, including outdoor temperature ($T_o$), PM2.5, air quality index (AQI), relative humidity (RH), wind direction (WD), and wind speed (WS), were also monitored by a weather station installed locally on the roof of case study building (Shen et al., 2015). All monitored factors are shown in Table 1 as followed.

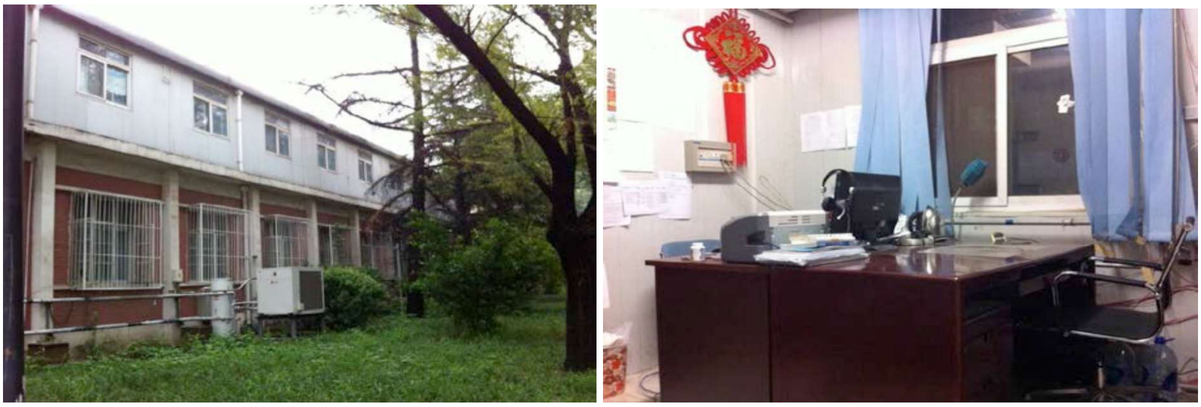

Figure 1. The case study building and the outlook of monitored office

Table 1. Monitored factors in this study

| | Monitored Factors |
|---|---|
| **Outdoor Parameters** | outdoor temperature, relative humidity, AQI, PM2.5, wind direction, wind speed |
| **Indoor Parameters** | Indoor temperature |

### 2.2. Estimators in feature selection process

Many machine learning models can generate feature rankings inherently from their internal structures, or can be constructed for feature selection. This applies to regression models, random forest, SVM, etc. In this paper, different machine learning methods and processing approaches will be studied on the selection of relevant features to window behaviour.

(1) Logistic Regression (LR):
Logistic regression is a sigmoidal classification able to predict the probability of an event having binary outcome (0-1) occurrences, which has been extensively applied in prediction of window behaviour in previous studies. Logistic regression allows to express the magnitude of coefficients of each related variable as a function of the binary outcome.

$$Log\left(\frac{P}{1-P}\right) = a + b_1 \cdot X_1 + \cdots + b_n \cdot X_n + \cdots \quad (1)$$

where:

- P is the probability
- a is intercept
- $b_{1-n}$ are coefficients
- $x_{1-n}$ are variables

(2) Support Vector Classification (SVC):
Support vector machine can construct a hyperplane, which can be used to make classifications. In SVC, a hyperplane is selected to best separate the points in the input variable space by their classes, which is to maximize the margin between the two classes. SVC can not only solve the problem of linear classification, but also the problem of non-linear classification by applying for kernel function.

$$\min \frac{1}{2} ||w||^2 \;\; s.t. \; , y_i(w^T x_i + b) \geq 1, i = 1, \cdots, n \quad (2)$$

where:

- $w$ is the vector of the coefficients

(3) Random Forest (RF)
In many practical applications, it is almost impossible to generate a specific functional relationship between inputs and output. The decision tree method is conceptually simple, yet powerful nonlinear method that often provides excellent results (Tsanas and Xifara, 2012). Random forest applied in this study is an ensemble learning method by constructing a group of decision trees during training stage. The input features are successively split into different branches with smaller sub-regions so that similar response can end up in the same set. The tree stops growing until it is impossible to split anymore or a certain criterion has been met. Besides, tree models can be directly used for feature selection by the measure of

impurity. Based on averaged impurity decrease values from each feature, features can be decided whether to be chosen or not. For classification, this measurement is typically called Gini impurity and information gain/ entropy, as followed.

$$H(T) = Entropy = -p * \log(p) - (1-p) * \log(1-p) \tag{3}$$

$$IG(T,a) = H(T) - H(T|a) \tag{4}$$

where:

- P is the percentage of positive samples
- a is corresponding attribute
- $H(T)$ is information entropy
- $IG(T,a)$ is information gain

### 2.3. Recursive feature elimination and cross validation

(1) Recursive Feature Elimination

Given the chosen estimator or classifier, different weights can be assigned to features, for example, the coefficients in generalized linear model. Recursive feature elimination (RFE) is based on the idea of selecting features by recursively considering smaller and smaller sets of features. Firstly, the estimator was trained on the initial set with all features, and importance of features can be obtained through training process by the attribute of estimator. Then, feature with least importance are pruned from current set of features. This procedure is recursively repeated until the desired number of features to select is eventually reached as shown in Figure 2. RFE is an effective method to get rid of some unimportant features preliminarily so as to reduce dimension of feature space when there are too many factors in training data.

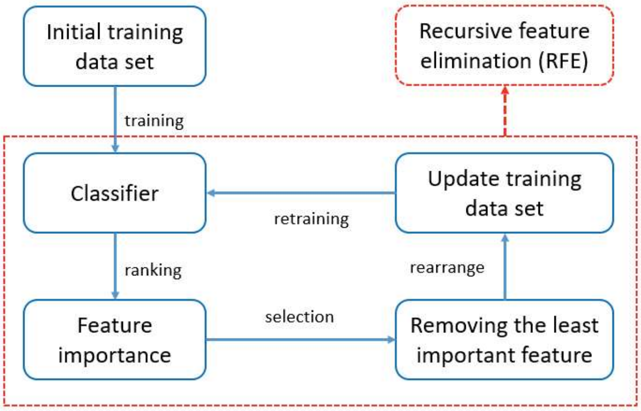
Initial training data set
Recursive feature elimination (RFE)
training
Classifier
Update training data set
retraining
ranking
rearrange
selection
Feature importance
Removing the least important feature

Figure 2. Diagram of recursive feature elimination (RFE)

(2) Cross Validation

When employing RFE, the number of remained features needs to be defined by practitioner rather than determined by some objective standards, which would bring in uncertainty of the final results. Cross validation can solve this problem by holding out part of data in training set as test data, then use trained model to predict on them. The best feature subset is the one with smallest error on the hold out test data. The prediction accuracy of test data in cross validation can provide criterion for RFE to determine the best feature subset. Therefore, recursive feature elimination with cross validation (RFECV) was applied to select features as shown in Figure 3.

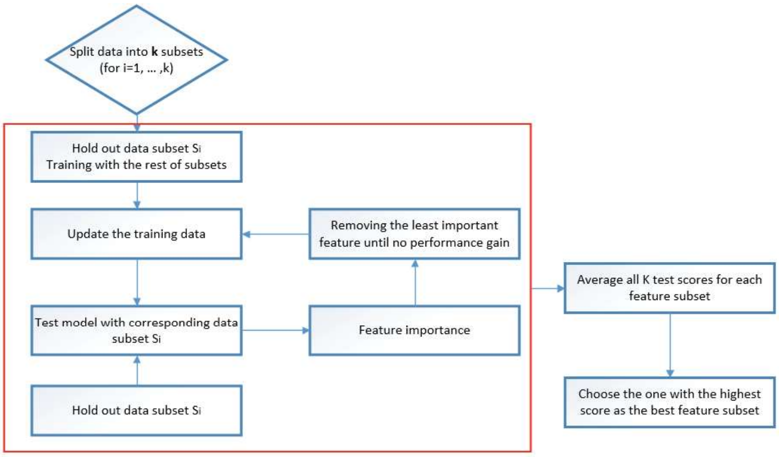
Split data into k subsets (for i=1, … ,k)
Hold out data subset Si Training with the rest of subsets
Update the training data
Removing the least important feature until no performance gain
Average all K test scores for each feature subset
Test model with corresponding data subset Si
Feature importance
Choose the one with the highest score as the best feature subset
Hold out data subset Si

Figure 3. Diagram of recursive feature elimination with cross validation (RFECV)

# 3. Results and Discussion

## 3.1. Analysis of correlation among features

Perfectly corrected features are truly redundant in the sense that no additional information is gained by including all of them in the model. If highly correlated features are present, individual features would exhibit similar performance to the collective feature subset and computational price would accordingly increase (De Silva and Leong, 2015). Besides, including redundant features in predicting model may also mislead certain modelling algorithms and reduce prediction accuracy (Liu and Motoda, 1998). Therefore, eliminating redundant features before establishing predicting model is necessary in order to improve the model performance.

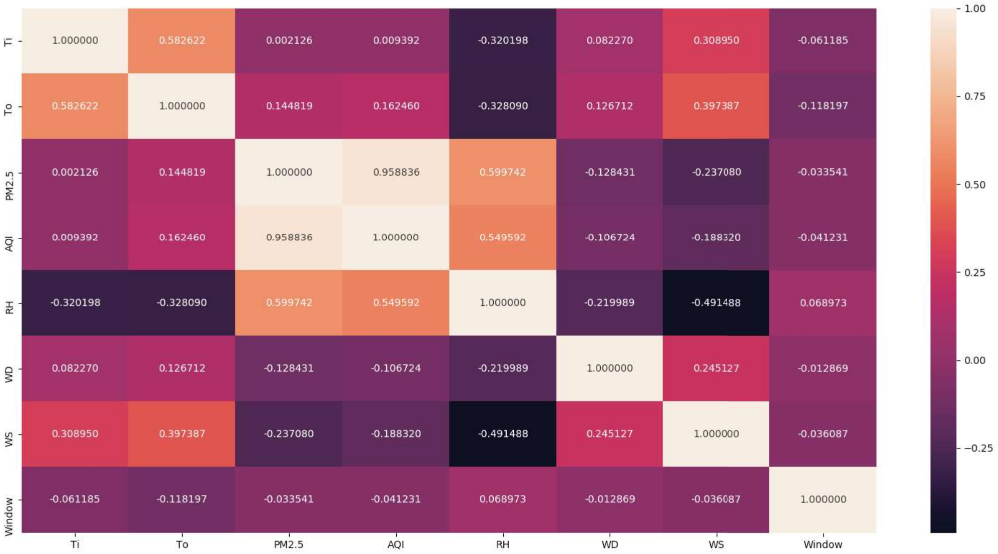

Figure 4. Correlation coefficients between different factors

The correlation coefficients among seven features have been calculated as shown in Figure 4. The highest correlation coefficient of 0.96 occurs between PM2.5 and AQI, which reasonable because PM2.5 is a sub-index in the evaluation of AQI. Besides, strong correlation can also be observed between indoor and outdoor temperature (0.58), relative humidity (RH) and PM2.5 (0.60), RH and AQI (0.55), wind speed (WS) and RH (-0.49). Such various correlations among all factors may change the prediction performance of each factor on window behaviour to some degree by imposing complex effects between each other. In order to eliminate the effects of redundant features and improve prediction accuracy, appropriate methods were applied by considering feature subsets rather than individual feature relevance assessment as followed.

### 3.2. Recursive feature elimination (RFE)

Recursive feature elimination has a great advantage in the elimination of unimportant features when the feature dimension of the model is relatively large. Although only seven features were considered which probably makes it not particularly necessary to apply RFE process, RFE was still employed in this study in order to provide insights in dealing with large feature dimension in window prediction and improve the universality of this study. Therefore, RFE with an estimator of logistic regression (LR) was applied to demonstrate a complete process for dealing with feature selection. Two of least important features among all seven are ruled out, which means five features are remained in process of RFE as shown in Table 2.

Table 2. The coefficients of remained features

| **Name of feature** | $T_i$ | $T_o$ | PM2.5 | RH | WS |
|---|---|---|---|---|---|
| **Coefficient** | 0.3926 | -0.1855 | -0.0006 | -0.0475 | 0.0127 |

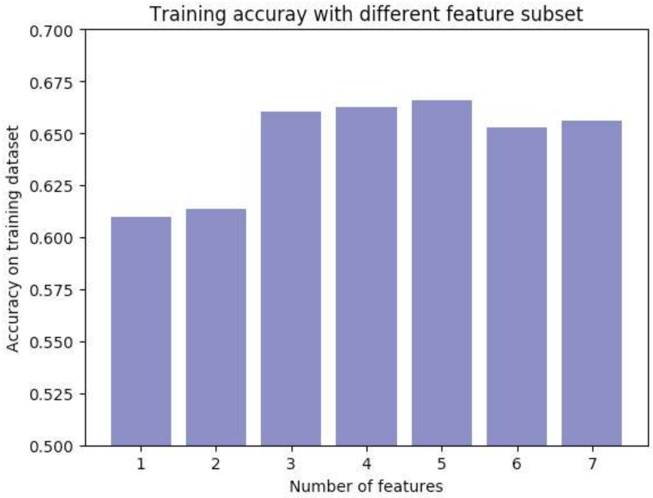

Figure 5. Training accuracy of RFE on different feature subset

Based on the results, AQI and wind direction are superseded among seven features, which means these two don’t have a contribution as significant as other features to the prediction accuracy of window behaviour. Previous researches have pointed out that wind speed is not particularly correlated with window operation (Haldi and Robinson, 2009b), which is identical with the results in this study. As for AQI, it refers to the severity that air has been polluted, and is considered as an important and essential index for air quality. The reason that AQI is ruled out in RFE process is mainly because the data collection in this study was conducted during transitional seasons in Beijing, when central heating system in a city scale had been turn off during that period, so did all coal boilers used for central heating. Hence, the pollutants concentration in air was not as high as it was in winter, which can adequately explain why AQI is considered as an irrelevant feature by the RFE process. Additionally, the coefficient for PM2.5 is quite low compared with other coefficients in table 2, which indicates that PM2.5 has little correlation with window operation. The reason behind this low relevance of PM2.5 is quite similar as that of AQI, because PM2.5 is actually a sub-index in the evaluation of AQI in China.

When the raw data include too many features, it is very effective to apply RFE preliminarily removing part of irrelevant features. However, there is one drawback for RFE, which is the best feature subset cannot be decided by RFE process. Although the model accuracy based on training data can be calculated, there is no validation process to measure the predicting performance for various feature subsets combined based on training data. As shown in Figure 5, just because the subset with 5 features generates the best accuracy based on training data among all 7 feature subsets in RFE, it doesn't mean that this subset with 5 features would be exactly the best choice for the model because there is no prediction process on a new group of data to validate this idea.  In order to obtain the best feature subset, the recursive feature elimination with cross validation (RFECV) can be applied based on the new feature dimension selected by RFE as a complimentary process.

### 3.3. Recursive feature elimination with cross validation (RFECV)

In the cross validation process, the whole training data will be divided into 10 folds, 9 of them used for training and one hold-out fold used for validation. The difference between RFECV and RFE is that in each feature subset the estimator will be examined in terms of making predictions on the data of hold-out fold in RFECV, hence the best feature subset can be determined by the rankings of CV scores, which is actually the prediction accuracy obtained by using number of correct predictions divided by the number of hold-out samples in cross validation.

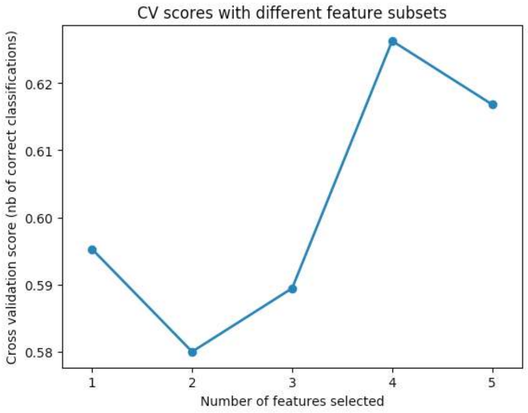

Figure 6. CV scores with remained five features

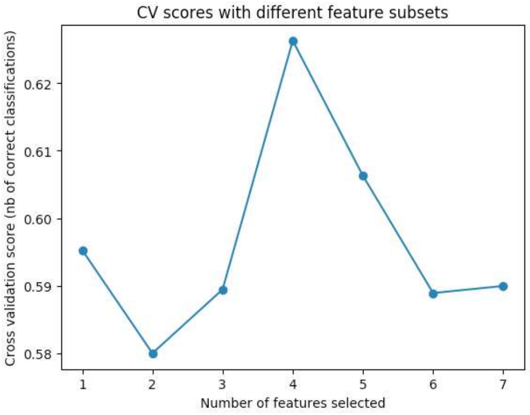

Figure 7. CV scores with original seven features

The RFECV has been applied firstly on the data set with remaining five features from the RFE process as shown in Figure 6. Based on the results, different feature subsets demonstrate different CV scores. The cross validation score reached up to a maximum of 0.626 when four features are retained in the model including $T_i$ (indoor temperature), $T_o$ (outdoor temperature), RH (relative humidity), and WS (wind speed), which means one more feature, PM2.5, can be further got rid of.

The strong relativity between indoor temperature, outdoor temperature, and window behaviour has been proved by many researches before (Parys et al., 2011), however RH and WS have been uniformly ignored because of their separate insignificant statistical correlation with window behaviour (Haldi and Robinson, 2009b). The problem is that the statistical significance analysis in previous researches was conducted without considering the confounding effects among features combinations on the prediction of window behaviour (Herkel et al., 2008, Shen et al., 2015), hence features with smaller relevance to window behaviour were ruled out at the first step. Based on the results in Figure 6, however $T_i$, $T_o$, combined together with RH and WS generates the best CV scores rather than merely temperature parameters, which in turn indicates that just because features don't have strong correlation with window behaviour, doesn't mean the combination of them cannot reach a better prediction accuracy. On the contrary, when features closely correlate with each other, they're likely to become redundant features which couldn't provide more useful information for the prediction in the model (Guyon and Elisseeff, 2003).

To validate the result, the original data set with all seven features were used again to apply RFECV as shown in Figure 7, the same result was obtained. Besides, the results also demonstrate that a model with more features doesn't necessarily lead to a better prediction accuracy, on the contrary, sometimes it would be totally counterproductive. When selected feature number is less than four, low CV scores indicate that the prediction model is likely to result in underfitting, which means the model cannot capture characteristics of the problem very well. Similarly, when more than four features are selected, the model is probably overfitting. Actually, when number of features is not very high in the model, RFECV can be directly used for determining the best subset of features rather than established on RFE. However, RFECV is more computationally expensive than RFE, in that case using RFE to deal with high dimensionality data firstly is very helpful. Therefore, according to results of RFECV the best feature subset on this training data includes four features, which are indoor temperature, outdoor temperature, relative humidity and wind speed respectively. It should be noted that this conclusion is only suitable in this dataset and estimator, rather than a universal conclusion.

### 3.4. Comparison between LR and SVC on results of RFECV

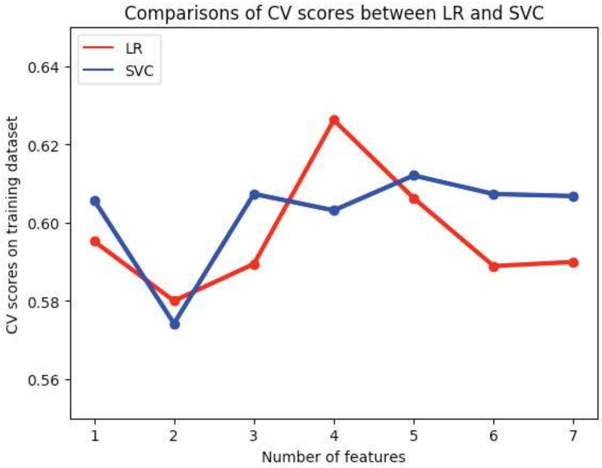

Figure 8. Comparison of CV scores between LR and SVC on various feature subset

The estimators mentioned above in RFE and RFECV are all based on logistic regression (LR). In this section, support vector classification (SVC) is applied for the estimator as a substitute of logistic regression. Based on Figure 8, CV scores for SVC in the whole range of feature subset demonstrate some variances to a degree, which has been previously proved in LR model that different feature subset can lead to different prediction accuracy. The highest CV score reaches up to 0.612 and occurs when five features are chosen, which are indoor temperature, outdoor temperature, PM2.5, relative humidity, and wind speed respectively.

Except for the subset with first two features, all other subsets display a relatively stable CV score, which is in the range of 0.60-0.62 with little fluctuations.

When compared with SVC, LR shows a larger variance through the whole range of feature numbers, and a higher maximum of 0.626 at the subset with four features. The general trend of the variety in CV scores is quite similar between SVC and LR, but prediction accuracy of SVC seems more robust on different the feature number compared to LR. However, this stability needs to be investigated further by testing on new dataset. Generally, the results show the validity of both methods in feature selection based on similar CV scores of both, however, the better one of them can only be determined by using new data to make predictions and comparing the accuracy of predicting results among these two methods in terms of bias and variance.

### 3.5. Random forest on feature selection

Unlike RFE or RFECV, tree models perform feature selection process by the measure of impurity embedded inside the algorithm rather than by iterations, which makes tree models much more computationally efficient than RFE methods. In this study, random forest (RF) has been employed as a non-recursive method to complete feature selection process and constructed by 10 decision trees.  For classification, feature importance can be evaluated by the reduction of Gini impurity.

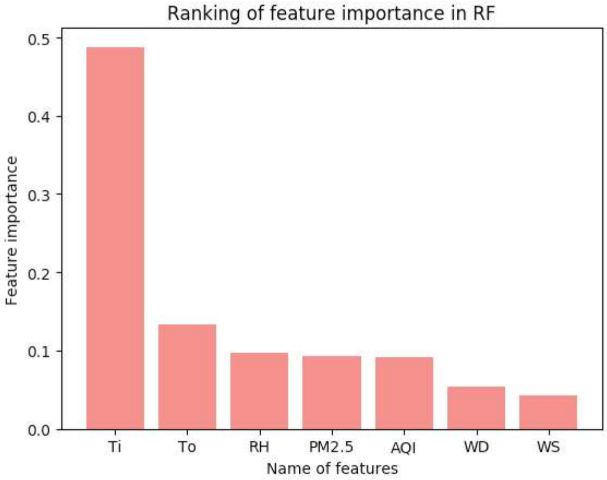

Figure 9. Feature rankings based on RF

As shown in Figure 9, indoor temperature is most important among all seven features and has a great advance in terms of the normalized values of feature importance. Then outdoor temperature comes second, followed by relative humidity, PM2.5 and AQI with little variance in feature importance. The similar importance level among relative humidity, PM2.5 and AQI based on tree model conforms the results of high correlation among these three features in 3.1. However, it should be noted that when there are highly correlated features present in the tree model, any of these correlated features can be chosen as predictor with no preference one over others. Once one of them is determined, the importance of others would be significantly reduced since the reduction of impurity have been mainly executed by the first chosen feature, which can lead to a lower reported importance for other features. It would benefit the feature selection process and reduce overfitting for the tree model, however it does not indicate that the feature with lower importance in tree models is also insignificant in statistics. On the contrary, the feature with lower importance may turn out to be equally important as the chosen feature, which can be proven by the obvious difference of feature importance between indoor and outdoor temperature in Figure 9.

By comparison with recursive method RFECV, both RF and RFECV consider and evaluate features in the model altogether and deemed indoor temperature, outdoor temperature, and relative humidity as important features, which verified the validity of feature selection results with each other between these two methods. For Random Forest, the calculated feature importance of each factor actually indicate the reduction of impurity in random forest, which make the feature selection results more explainable compared to RFECV. Besides, RF is much more computationally cheaper because it can avoid recursive process.

## 4. Conclusions

This study demonstrated both recursive and a non-recursive feature selection methods, which are each capable of considering all influencing factors simultaneously so as to take into account confounding effects among various factors in the prediction of window opening behaviours. Two machine learning algorithms were applied as estimators in the recursive selection process, namely support vector classification (SVC), logistic regression (LR), and one in non-recursive process, a random forest method (RF). A complete feature selection scheme has been demonstrated by the combination of recursive feature elimination (RFE) and recursive feature elimination with cross validation (RFECV). In general, several main conclusions about the feature selection in window behaviour can be made as followed:

(1) Based on the review of current study on window behaviour prediction, three problems exist related to the influencing factors or features in such prediction models: no clear criterion exist to guide the feature selection process; features are separated from each other without a consideration of their confounding effects among features; no comprehensive feature subset search strategy is involved to deal with problems associated with large feature numbers. The status quo of the study of feature selection in window behaviours also manifests the importance and necessity of re-examining the criterion and approach when making decisions on feature selection.

(2) Factors correlate with each other to different degrees, which can make some of factors with high correlations become redundant features in the prediction of window behaviour. The individual feature relevance assessment, which has been applied in previous researches, has limited effects in the elimination of redundant features.

(3) In recursive methods, RFE and RFECV can be combined to solve the window behaviour prediction problems related to abundances of influencing factors. Recursive feature elimination (RFE) can be applied at the preliminary stage to get rid of some less relevant features and reduce the dimensionality of the model when the number of features is relatively high in the collected data. Recursive feature elimination with cross validation (RFECV) can be further employed to search for the most appropriate feature subset by eliminating less relevant features recursively when considering all features together.

(4) The algorithm in the prediction model can demonstrate different prediction accuracies with different feature subsets. Different algorithms can also perform

diversely with same feature subset, although the obtained feature subset results obtain using LR, SVC, and RF are not totally identical, all algorithms have successfully identified indoor temperature, outdoor temperature, and relative humidity as most important features when predicting window behaviour.

(5) Logistic regression (LR) and support vector classification (SVC) demonstrated different traits in recursive feature elimination. LR shows a higher maximum in CV scores while SVC shows a stronger stability during the selecting process. Both of them can be considered efficient in this study due to their similar CV scores. It should be noted that chosen algorithms should remain constant through the feature selection process and model establishing stage. Random forest analysis performs well in eliminating redundant features, for example the low feature importance of outdoor temperature in the results. It is also computationally cheaper compared to recursive methods.

## References

D'OCA, S. & HONG, T. 2014. A data-mining approach to discover patterns of window opening and closing behavior in offices. *Building and Environment,* 82, 726-739.

D'OCA, S., FABI, V., CORGNATI, S. P. & ANDERSEN, R. K. 2014. Effect of thermostat and window opening occupant behavior models on energy use in homes. *Building Simulation,* 7, 683-694.

DE SILVA, A. M. & LEONG, P. H. W. 2015. Feature Selection. *Grammar-Based Feature Generation for Time-Series Prediction.*

FABI, V., ANDERSEN, R. V., CORGNATI, S. & OLESEN, B. W. 2012. Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models. *Building and Environment,* 58, 188-198.

FILIPPÍN, C., FLORES LARSEN, S., BEASCOCHEA, A. & LESINO, G. 2005. Response of conventional and energy-saving buildings to design and human dependent factors. *Solar Energy,* 78, 455-470.

FRITSCH, R., KOHLER, A., NYGÅRD-FERGUSON, M. & SCARTEZZINI, J. L. 1990. A stochastic model of user behaviour regarding ventilation. *Building and Environment,* 25(2), 173-181.

GUYON, I. & ELISSEEFF, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*.

HAAS, R., AUER, H. & BIERMAYR, P. 1998. The impact of consumer behavior on residential energy demand for space heating. *Energy and Buildings,* 27, 195-205.

HALDI, F. & ROBINSON, D. 2009a. Interactions with window openings by office occupants. *Building and Environment,* 44(12), 2378-2395.

HALDI, F. & ROBINSON, D. 2009b. Interactions with window openings by office occupants. *Building and Environment,* 44, 2378-2395.

HERKEL, S., KNAPP, U. & PFAFFEROTT, J. 2008. Towards a model of user behaviour regarding the manual control of windows in office buildings. *Building and Environment,* 43, 588-600.

LIU, H. & MOTODA, H. 1998. *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers.

MASOSO, O. T. & GROBLER, L. J. 2010. The dark side of occupants' behaviour on building energy use. *Energy and Buildings,* 42, 173-177.

NICOL, J. F. 2001. Characterising occupant behaviour in buildings: towards a stochastic model of occupant use of windows, lights, blinds, heaters and fans. *Seventh International IBPSA Conference, Rio De Janeiro*, 1073-1078.

PAN, S., XIONG, Y., HAN, Y., ZHANG, X., XIA, L., WEI, S., WU, J. & HAN, M. 2018. A study on influential factors of occupant window-opening behavior in an office building in China. *Building and Environment,* 133, 41-50.

PARYS, W., SAELENS, D. & HENS, H. 2011. Coupling of dynamic building simulation with stochastic modelling of occupant behaviour in offices – a review-based integrated methodology. *Journal of Building Performance Simulation,* 4, 339-358.

SHEN, W., CHUANQI, X., SONG, P. J. S., YUNMO WANG, XIAOYAN LUO, TAREK M HASSAN, & STEVEN FIRTH, F. F., RORY JONES, PIETER DE WILDE 2015. Analysis of factors influencing the modelling of occupant window opening behaviour in an office building in Beijing, China. *Proceedings of BS2015: 14th International Conference of the International Building Performance Simulation Association*.

TSANAS, A. & XIFARA, A. 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings,* 49, 560-567.

WEI, S., BUSWELL, R. & LOVEDAY, D. 2013. Factors affecting 'end-of-day' window position in a non-air-conditioned office building. *Energy and Buildings,* 62, 87-96.

WEI, S., JONES, R. & DE WILDE, P. 2014. Driving factors for occupant-controlled space heating in residential buildings. *Energy and Buildings,* 70, 36-44.

YAN, D., HONG, T., DONG, B., MAHDAVI, A., D'OCA, S., GAETANI, I. & FENG, X. 2017. IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings,* 156, 258-270.