Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 121 (2015) 285 – 292

Procedia Engineering

www.elsevier.com/locate/procedia

9th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC) and the 3rd International Conference on Building Energy and Environment (COBEE)

Comparative Study on Machine Learning for Urban Building Energy Analysis

Lai Wei[a,b], Wei Tian[a,b]*, Elisabete A. Silva[c], Ruchi Choudhary[d], QingXin Meng[a], and Song Yang[a,b]

[a] *College of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin, China*
[b] *Tianjin Key Laboratory of Integrated Design and On-line Monitoring for Light Industry & Food Machinery and Equipment, Tianjin 300222, China*
[c] *Department of Land Economy, University of Cambridge, Cambridge, UK*
[d] *Department of Engineering, University of Cambridge, Cambridge, UK*

Abstract

There has been an increasing interest in applying machine learning methods in urban energy assessment. This research implemented six statistical learning methods in estimating domestic gas and electricity using both physical and socio-economic explanatory variables in London. The input variables include dwelling types, household tenure, household composition, council tax band, population age groups, etc. Six machine learning methods are two linear approaches (full linear and Lasso) and four non-parametric methods (MARS multivariate adaptive regression spline, SVM support vector machine, bagging MARS, and boosting). The results indicate that all the four non-parametric models outperform two linear models. The SVM models perform the best among these models for both gas and electricity. The bagging MARS performs only a little worse than the SVM for gas use prediction. The Lasso model has similar predictive capability to the full linear model in this case.

Peer-review under responsibility of the organizing committee of ISHVAC-COBEE 2015

Keywords: Urban buildings; Energy use; Machine learning; Comparative learning; Cross validation

* Corresponding author. Tel.: +086-22-60273495; fax: +086-22-60273495.
E-mail address: tjtianjin@126.com

Peer-review under responsibility of the organizing committee of ISHVAC-COBEE 2015
doi:10.1016/j.proeng.2015.08.1070

1. Introduction

The construction of reliable energy models is one of the most important tasks in building energy analysis [1]. This is because building energy models can be used to assess various energy saving strategies for both existing and new buildings in order to optimize building energy performance and reduce associated carbon emissions. As a result, there are various types of building energy models, including both engineering-based and statistical energy models [2]. Engineering models are based on the fundamentals of heat transfer required a number of detailed input parameters, whereas statistical energy models based on various machine learning methods need less parameters well suitable for building stock research, for instance, urban-scale energy assessment.

In recent years, there has been an increasing interest in applying machine learning methods to construct statistical energy models in urban environment [1, 3]. Tian and Choudhary [4] applied a MARS (multivariate adaptive regression spline) method to create non-parametric energy models for estimating energy use of secondary school buildings in London. Jain et al. [5] used support vector machine method to predict energy use of multi-family residential buildings in an urban area. Howard et al. [6] implemented a robust multivariate linear regression to analyze building energy use by end use in New York. Tian et al. [7] explored the spatial patterns of domestic energy in London using spatial error models. However, there are few studies that compare the predictive performance of machine learning methods in assessing energy performance of urban buildings.

Therefore, this research is focused on comparing predictive performance of six statistical machine-learning methods: full linear, Lasso (least absolute shrinkage and selection operator), MARS, SVM (support vector machine), bagging MARS, and boosting [8]. This study uses London as a case study for comparative analysis. The two output performance indicators are annual total electricity and gas consumption for domestic properties. The explanatory variables include dwelling types, household tenure, household composition, council tax band, population age groups, etc.

2. Data

Table 1. Explanatory variables used for estimating gas and electricity in London

Inputs	Number	Descriptions
Meters	3	Number of ordinary domestic electricity meter, number of Economy 7 electricity meter, number of domestic gas meter
Population	7	Number of total population, age 0-15, age 16-29, age 30-44, age 45-64, age of over 65, working age (16-64)
Household	1	All the households
Household composition	5	Couple household with/without dependent children, single parent household, one person household, other household types
Household tenure	4	Household with outright ownership, household with mortgage or loan, social rented household, private rented household
Dwelling type	6	Detached, semi-detached, terrace, flat, household with or without usual residents
Land area	1	Land area for domestic buildings (1000 m2)
Financial vulnerability	12	Households by financial capability segment. Bands range from 0 to 11, FV0 denotes the least vulnerable, FV11 denotes the most vulnerable
Council tax band	8	Dwellings in 8 council tax (CT) bands from A to H. The council tax A (CTA) is the lowest and the council tax G (CTG) is the highest.

The data used in this study are from London data store [9]. The spatial scale for both energy and explanatory data is based on LSOA since it is the smallest spatial scale available in London data store. LSOA stands for the lower super output area, designed by UK ONS (Office for National Statistics) for small area data analysis [10]. London has 4765 LSOAs, but the data from 4746 LSOAs are used in this analysis due to the availability of energy data. All the input data used here are summarized in Table 1. The input factors include type of gas and electricity meters (corresponding to fuel and tariff), population number at different age groups, household composition, household

tenure status, dwelling types, land area for domestic buildings, financial vulnerability, and council tax bands. These input variables can represent both household physical conditions and socio-economic influences. Figure 1 shows the spatial distributions for electricity and gas use at LSOA scale in London for the year of 2011.

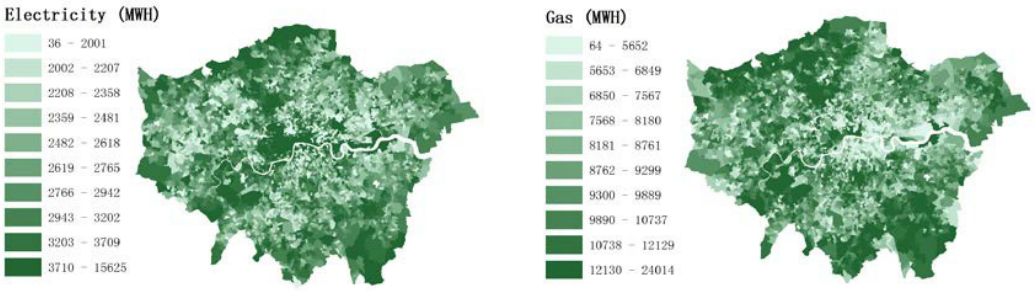

Fig.1. Electricity and gas use at LSOA (lower super output area) in London

3. Methods

The six machine learning methods will be used in this analysis to estimate both gas and electricity use in London. These methods are full linear, Lasso, MARS, SVM, Bagging MARS, and boosting. Except for the full linear model, the parameters in other five approaches need to be tuned to derive a model that best matches the data. In this study, the cross-validation resampling method is used to adjust/tune the parameters [11]. This means that a given dataset is randomly divided into k sets (also called k-fold) with almost the same size. A statistical model is obtained using all k-1 sets, except for one set (acted as a held-out sample). This held-out set is used to judge the predictive performance for a statistical model. In the next iteration, all the data except for the second set is used to fit a new model and the second set acts as a held-out sample to assess the performance of this new model. This procedure is carried out for all the k sets. It should be emphasized that the computational cost is very high for some machine learning methods. For example, if a method requires 2 parameters to be tuned, and one chooses to carry out 10-fold cross validation, then 120 statistical models need to be generated (each parameter can have 6 possible values). R caret package, developed by Max Kuhn, is used for all the statistical computation in this research [12].

These six machine learning methods are described briefly below and for more detailed information, please refer to two documents [8] and [13].

(1) The full linear model is the simplest in these methods using least square approach without reducing the number of predictors.

(2)The lasso method uses a shrinkage approach by retaining a subset of the most relevant predictors and deleting the remaining explanatory variables. In this method, the fraction of full solution is regarded as a tuning parameter. The smaller values of the fraction of full solution mean a greater shrinkage [14]. When the fraction of full solution is one, then the Lasso model becomes ordinary least square with no penalty (i.e. a full linear model). Although the training error from constructing models may be lower in comparison with a full linear model, the prediction error from this shrinkage model is not necessarily low due to a decrease in variance of predictions.

(3) The MARS stands for multivariate adaptive regression splines, a generalization of stepwise linear regression. This method is well suited for high dimensional inputs and it is also fast to compute. The degree and the term numbers are two tuning parameters in MARS method. If the degree is 1, the regression model only considers linear effects. If the degree is 2, the interactions between two variables will be taken into account in the analysis. The higher the number of terms entered in the regression equation, the lower the training error is.

(4) The SVM method is a class of powerful flexible regression learning techniques, originally developed for classification models. SVM is closely related to robust regression by reducing the effects of outliers in regression process. The polynomial function is chosen as a kernel function in this research after preliminary analysis. Three main tuning parameters in this SVM regression are cost, degree, and scale. The cost parameter is to penalize the

large residuals of a model. The degree is associated with the polynomial kernel function used to describe linear or non-linear relationships in which a larger degree value tends to result in over-fitting. The scale factor can change predictor values to improve predictive performance for a SVM model.

(5) The bagging method means the bootstrap aggregation, a general approach to combine a number of models and obtain the averaged predictions for these models. In this analysis, the bagging method is used together with a MARS regression. The advantage of this bagging method is to reduce the variance of predictions through aggregation process. The MARS can sometimes yield unstable predictions and as a result, the combination of MARS and bagging methods may improve the prediction performance for a MARS model. The main tuning parameters for bagging MARS are the same as MARS method as described above.

(6) The boosting method combines the outputs from many "weak" regression models to obtain a better model, similar to the bagging method. The main difference is that the boosting method would use a weighted majority vote for different models for a final prediction, whereas the bagging method treats all the models equally. Stochastic gradient boosting is used in this analysis to obtain a boosting model by fitting a base model for a subsample of the training data randomly sampled without replacement. The shrinkage, iteration numbers, and max tree depth are the three main tuning parameters for this method. The shrinkage parameter is also called "learning rate" that can improve predictive capability using smaller values of shrinkage, but lead to an increase in computational cost. The optimal iteration number depends on the shrinkage values. Hence, analysts may firstly choose appropriate shrinkage and then select a proper iteration number based on cross-validation error. The tree depth can account for interaction terms between predictors in SVM regression.

Note that the data needs to be pre-processed in order to efficiently implement the machine learning methods described above. The three steps have been used in this analysis: (1) remove linear dependencies; (2) delete highly correlated predictors; (3) center and scale original data. Firstly, the QR decomposition is used to find the linear dependencies among input variables and remove them from the analysis. The benefit of this operation is to increase the stability of numerical computation. Indeed, some statistical methods cannot be properly used without removing these linear dependencies. The second step is to identify correlated predictors and delete these variables if the correlation coefficient is greater than 0.75 [8]. Reducing the number of predictors has some advantages. Fewer variables indicate a simpler model and also reduce computational cost, which often results in a parsimonious and more interpretable model. The third step is to center (zero mean) and scale (standard deviation of one) the predictors before training the statistical models [8]. These manipulations can increase the numerical stability for machine learning methods.

4. Results and discussion

4.1. Gas use

Firstly, a full linear model is obtained using least square method for estimating gas use in London. Then a Lasso model is created for the same purpose. It is found that the cross validation error is gradually decreased when the fraction of full solution increases. This means the models from the full linear and Lasso methods have the similar capability when estimating the gas consumption at LSOA in London. Figure 2 shows the resampling profile for tuning the parameters in MARS and bagging MARS models. The change of term number entered in the model has obvious effects on the cross validation error, especially at the starting stage. Then this influence becomes smaller for a large number of terms entered in the model. This means more terms added in the regression model would not increase the predictive capability of estimating gas consumption. Note that the bagging MARS becomes more unstable with an increase in term numbers although it can slightly increase the predictive performance in this case. Thus, analysts should be careful of choosing bagging MARS models in the analysis to avoid creating a more destructive model.

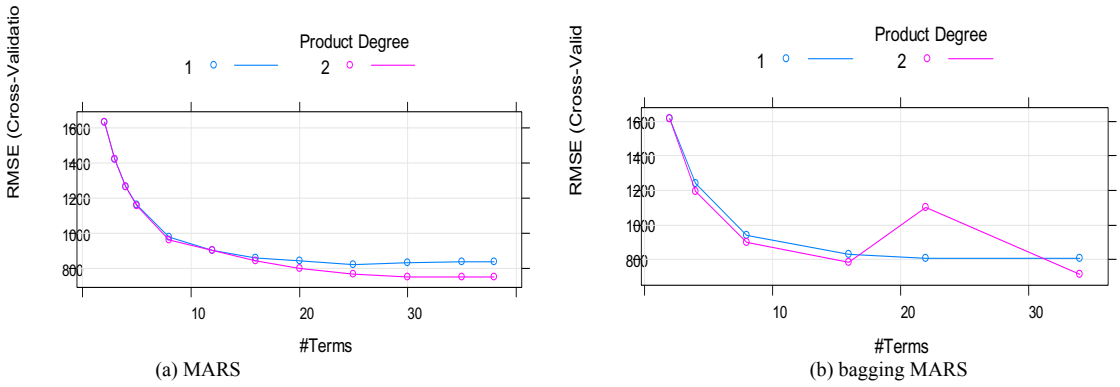

Fig. 2. Parameter tuning profile of MARS and bagging MARS for gas use in London

Figure 3 reveals that the effects of cost and scale are dependent on the degree number for SVM regression. When the degree number is taken as 1, the cross validation error becomes stable after a scale of 0.01. When the degree number is 2, the predictive errors show more irregularities. The final optimal model is quadratic (i.e. a degree of 2) with a scale factor of 0.01 and a cost value of 1. Figure 4 indicates that the RMSE decreases with an increase of both tree depth and boosting interactions for the boosting method. The small shrinkage values can lead to an increase in RMSE for this case. The final boosting model has a depth of 8 with a shrinkage of 0.04 and a tree number of 1500.

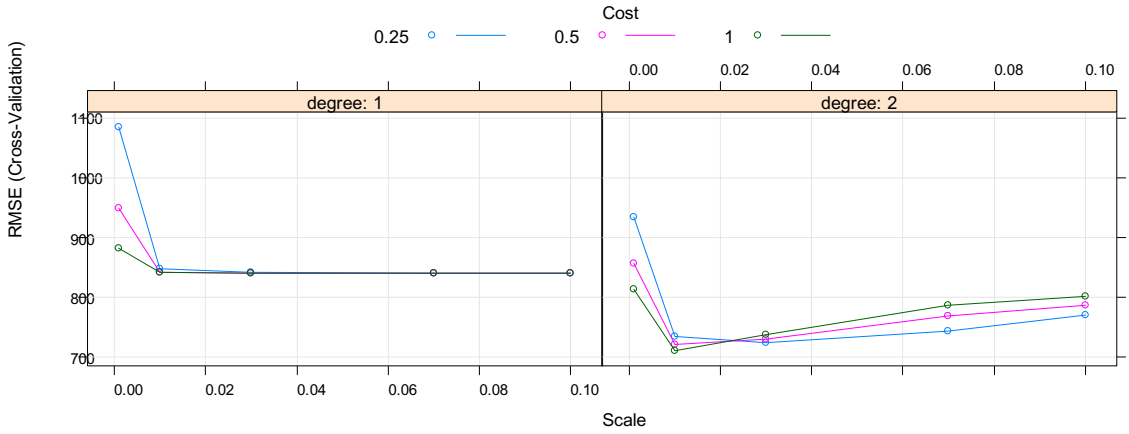

Fig. 3. Parameter tuning profile of support vector machine methods for gas use in London

Figure 5 compares the model accuracy based on the resampling results with box plots for these six models. The performance indicators are RMSE (root mean square error) and R2 (coefficient and determination). The RMSE can be regarded as the average model error and the R2 can be interpreted as a proportion of the information explained by a regression model. The lower the RMSE, the better a model is. In contrast, a higher R2 indicates a better model. The SVM model emerges as the winner in this case in terms of both RMSE and R2. The model created using bagging MARS has very similar performance to the SVM model. Two linear models (full linear and Lasso) perform worse than all the non-parametric models in this case.

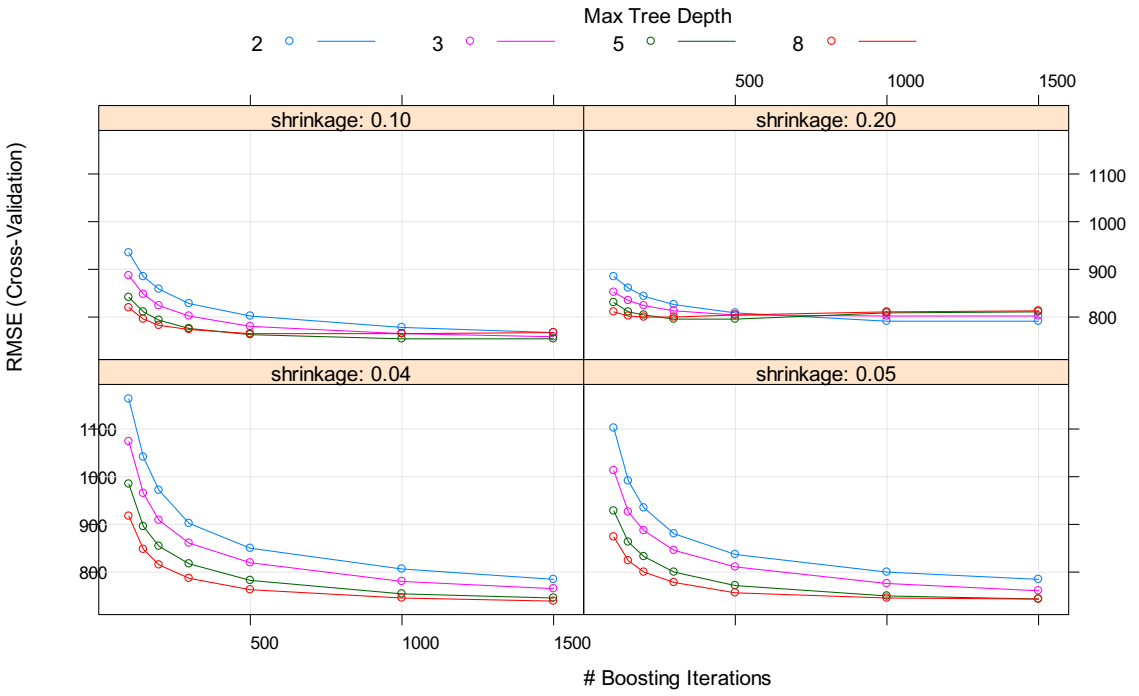

Fig. 4. Parameter tuning profile of generalized boosted regression for gas use in London

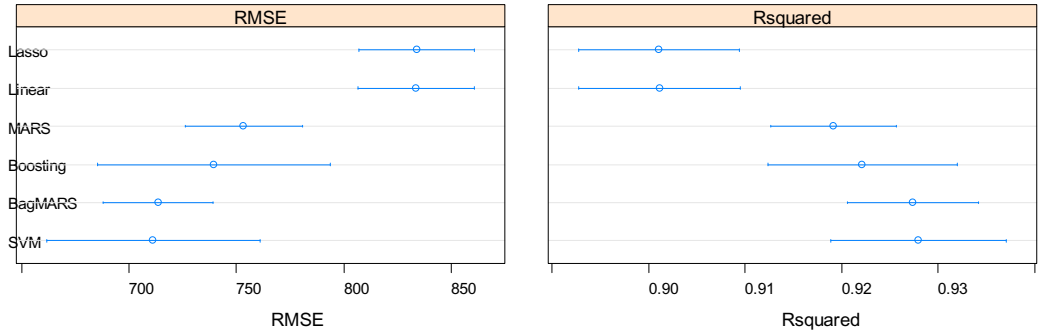

Fig. 5. Comparison of model performance for gas use in London based on six machine learning methods in terms of RMSE (root mean square error) and Rsquared (coefficient of determination, R2) (boxplot with confidence level 0.95).

4.2. Electricity use

The general patterns from these machine learning methods for electricity use are similar to the gas use. Thus, only brief discuss on performance differences is presented here and the tuning profiles for SVM and Boosting are not shown in this paper. For the bagging MARS model, the tuning process for electricity (Figure 6) is more unstable in comparison with gas consumption (Figure 2). This suggests that a higher number of terms in bagging MARS may have worse performance although an appropriate number of terms may have better prediction power. For the MARS method, the trend between gas and electricity use is different after the number of terms reaches 16. The cross validation error for electricity use is higher for the degree of 2 than that for the degree of 1, whereas the opposite is the case for gas use. This suggests that the interactions between input variables may be more significant for gas use compared to electricity use. For the final models from MARS and bagging MARS, the number of terms entered for

electricity estimation is less than gas use. This means the statistical energy models for electricity are more parsimonious than those for gas prediction. Note that the quality of energy data itself may have important implication for this analysis.

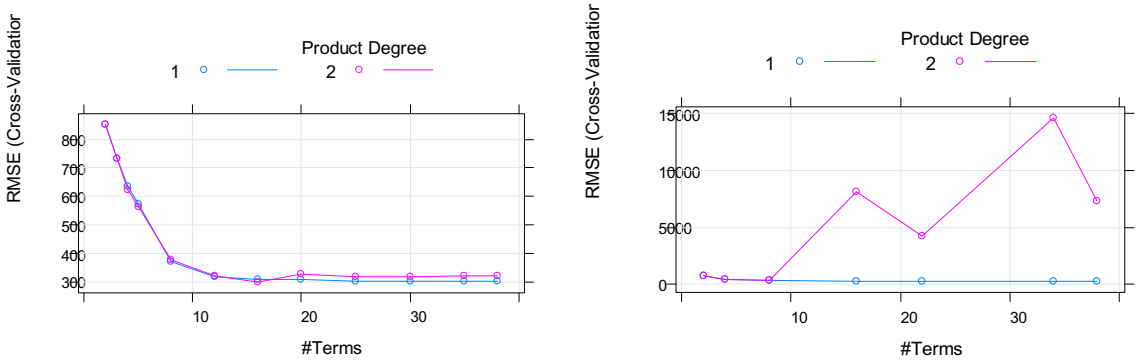

Fig. 6. Parameter tuning profile of MARS and bagging MARS for electricity use in London

For SVM models, the final tuning parameters for electricity are the same as the gas, except for the cost of 0.5 for the electricity model (1 for the gas model). The cost value in SVM is a main parameter to adjust the complexity of the models. The smaller the cost in SVM, the simpler the final model is. This conclusion from the SVM method confirms the statement from comparing the MARS model from gas and electricity as discussed in the last paragraph. The statistical energy models for electricity estimation are more parsimonious than those for predicting gas use. This indicates that less explanatory variables listed in Table 1 are required to predict the electricity use, whereas more predictors are needed for gas calculation. The optimal parameters of boosting models are the same for both electricity and gas prediction in this case.

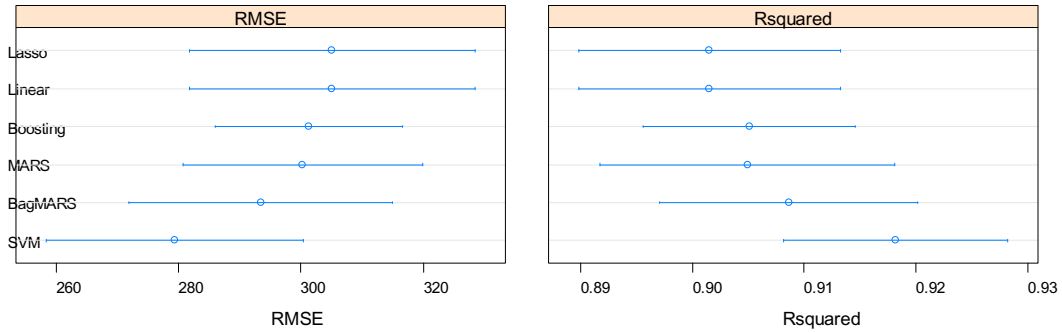

Fig. 7. Comparison of model performance for electricity use in London based on six machine learning methods in terms of RMSE (root mean square error) and Rsquared (coefficient of determination, R2) (boxplot with confidence level 0.95).

Figure 7 compares the resampling results from six models for electricity use in London. The SVM comes out the winner. The two linear models (Lasso and full linear) have similar predictive capability, which are both worse than all the four non-parametric models. By comparing Figure 5 and Figure 7, the predictive performance for final models is better for gas estimation in comparison with electricity calculation since the coefficients of determination (R2) are higher for gas than those for electricity.

5. Conclusions

This paper investigates the predictive performance of six machine learning methods in order to create reliable energy model for estimating both gas and electricity use in London. The explanatory variables include the population age, household number, household tenure status, dwelling type, council tax band, financial segment, etc. Six machine learning methods contain two linear methods (full linear and Lasso) and four non-parametric methods (MARS, SVM, bagging MARS, boosting). The results indicate that the SVM models perform the best among the six methods for both electricity and gas prediction in this research. The bagging MARS model nearly matches the SVM model. Further analysis will compare these machine learning methods for electricity and gas use per household in order to understand the patterns of domestic energy intensity, not only total gas and electricity in London. The methods used in this analysis can be also applied by implementing various types of machine learning approaches to other cities, such as Tianjin and Beijing in China, depending on data availability. Based on the results from this study, the SVM and bagging MARS models are recommended to be used together to provide reliable energy estimation in a city-scale research. The full linear model can be used as a reference case to compare predictive capabilities with non-parametric methods.

Acknowledgements

This research is supported by the Tianjin Research Program of Application Foundation and Advanced Technology (No. 14JCYBJC42600) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry of China.

References

[1] H.X. Zhao, F. Magoulès, A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews.16 (2012) 35 86-92.
[2] N. Fumo, A review on the basics of building energy estimation. Renewable and Sustainable Energy Reviews.31(2014)53-60.
[3] N. Wu, E.A. Silva. Artificial intelligence solutions for urban land dynamics: a review. Journal of Planning Literature. 24(2010)246-65.
[4] W. Tian, R. Choudhary, A probabilistic energy model for non-domestic building sectors applied to analysis of school buildings in greater London. Energy and Buildings.54(2012)1-11.
[5] R.K. Jain, K.M. Smith, P.J Culligan, J.E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Applied Energy.123(2014) 68-78.
[6] B. Howard, L. Parshall, J. Thompson, S. Hammer, J. Dickinson, V. Modi, Spatial distribution of urban building energy consumption by end use. Energy and Buildings.45(2012)41-51.
[7] W. Tian, J. Song, Z. Li, Spatial regression analysis of domestic energy in urban areas. Energy.76(2014) 29-40.
[8] M. Kuhn, K. Johnson. Applied predictive modeling: Springer, 2013.
[9] London Datastore. LSOA atlas, Greater London Authority, UK, Accessed 03-02-2015, http://data.london.gov.uk/dataset/lsoa-atlas. 2015.
[10] ONS. Digital boundaries for 2001 output areas (OAs) and super output areas (LSOAs and MSOAs) for England and Wales, ONS (Office for National Statistics), UK, Accessed 02-03-2015, http://www.ons.gov.uk/ons/guide-method/geography. 2015.
[11] E.A. Silva, K.C. Clarke. Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal. Computers, Environment and Urban Systems.26(2002)525-52.
[12] M. Kuhn, Building predictive models in R using the caret package. Journal of Statistical Software.28(2008)1-26.
[13] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction: Springer; 2009.
[14] W. Tian, R. Choudhary, G. Augenbroe, S.H. Lee, Importance analysis and meta-model construction with correlated variables in evaluation of thermal performance of campus buildings. Building and Environment.92(2015)61-74.