Recursive Partitioning in Predicting Energy Consumption of Public Buildings

Marijana Zekić-Sušac
University of Josip Juraj Strossmayer in Osijek
Faculty of Economics in Osijek
Trg Ljudevita Gaja 7, 31000 Osijek, Croatia
marijana@efos.hr

Adela Has
University of Josip Juraj Strossmayer in Osijek
Faculty of Economics in Osijek
Trg Ljudevita Gaja 7, 31000 Osijek, Croatia
adela.has@efos.hr

Saša Mitrović
University of Zagreb
Faculty of Organization and Informatics Varaždin, PhD student
Pavlinska 2, 42000 Varaždin, Croatia
smitrovic@foi.hr

Abstract. *Recursive partitioning includes a number of algorithms that create a classification or a regression decision tree by splitting the values of independent variables. The aim of this paper is to compare the accuracy of four different recursive partitioning methods in predicting the electrical energy consumption of public buildings. The input space included 141 attributes of public buildings in Croatia describing their geospatial, construction, heating, cooling, meteorological and energy characteristics. Four methods that produce regression tree partitioning were trained and tested. The results show that the random forest (RF) has outperformed CART, conditional inference tree (CTREE), and gradient boosted tree (GBT). The selection of important predictors was also compared and discussed.*

Keywords. Recursive partitioning, energy consumption, public buildings

1 Introduction

Previous research has shown that buildings are the largest individual energy consumers. More precisely, the building sector itself contains 40% of total primary energy consumption (Tommerup et al., 2007). Efficient models for predicting energy consumption could be particularly useful in public sector, where the state institutions need to recognize large consumers and allocate resources for improving its energy efficiency. Most of public buildings in Croatia still uses non-renewable energy resources and greenhouse gases, and reduction of such energy consumption is in accordance with EU directives and national strategic and action plans. Zekić-Sušac (2017) has shown that several approaches were used by researchers in modelling energy consumption: (1) individual statistical methods such as linear regression, time series analysis, probability density functions, or similar methods, (2) comparison of statistical methods with machine learning methods, and (3) simulation modelling.

This paper focuses on machine learning approach, more precisely on recursive partitioning methods and investigates their potential in predicting energy consumption of public buildings. Four different methods were used: classification and regression trees (CART), conditional inference trees (CTREE), random forest (RF), and gradient boosted trees (GBT) on a real dataset of Croatian public buildings. The aim was to investigate which of the recursive partitioning methods best fits the data and has a potential to be used as a modelling approach in reducing the cost of energy in public sector.

2 Previous research

Energy consumption and management is a frequent topic in recent research, due to the global need of reducing pollution, usage of non-renewable natural resources and green gas emission. Zekić-Sušac (2016) brings and overview of methods used in modelling energy efficiency and consumption. There are efforts in building prediction models of energy consumption in different countries by using statistical methods, machine learning, simulation (Chou and Bui, 2014). Some authors build prediction models for households, such as Farzana et al. (2014) who predicted the energy demand in the urban residential buildings of Chongqing in south west China. They have compared the accuracy of artificial neural networks (ANN), Grey models, regression models, a polynomial model and a polynomial regression model. According to their research, the artificial neural networks outperformed other methods. The largest number of machine learning methods in this domain was used by Chou and Bui (2014) but they have used experimental datasets from the literature instead of real data. Their intensive methodological tests have shown that the ensemble of support vector regression (SVR) and ANN has outperformed CART method, chi-squared automatic interaction detector, general linear regression, and ensemble inference model. Regarding the input space, they have used only 8 input attributes and two output variables (cooling load (CL) and heating load (HL)). Mangold et al. (2015) used Swedish energy performance certificate data to describe energy usage in buildings. Chung and Park (2015) investigated energy consumption in buildings in South Korea.

The dataset from the public sector was used by Son et al. (2015) who predicted energy consumption of government-owned buildings based on an RreliefF variable selection algorithm and support vector machines method. Has and Zekić-Sušac (2017) has investigated the potential of artificial neural networks in predicting the energy efficiency level.

Yu et al. (2010) utilized decision tree method (C4.5 algorithm) with annual average air temperature, house type, construction type, floor area, heat loss coefficient, equivalent leakage area, number of occupants, space heating, hot water supply and kitchen as input variables on 80 residential buildings in Japan to estimate residential building energy performance. Research results of the aforementioned authors demonstrated that decision tree method can classify and predict building energy demand levels with a high accuracy (93% for training dataset and 92% for test dataset.

Tsanas & Xifara (2012) compared a classical linear regression approach, Reweighted Least Squares (IRLS), against non-parametric random forests (RF) method for predicting heating load and cooling load of 768 diverse residential buildings by using relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution as input variables. Their research has shown that RF outperformed IRLS. Wang et al. (2018) used random forest (RF) for predicting energy consumption of two educational buildings in Florida state in the United States. Their dataset consisted of 11 input variables (meteorological, occupancy and time related data) while the methodology included RF, and regression tree (RT), and Support Vector Regression (SVR). The comparison has revealed that RF was more accurate then RT and SVR.

Papadopoulos et al. (2018) evaluated performances of random forests, extremely randomized trees (extratrees), and gradient boosted regression trees on Tsanas & Xifara (2012) dataset of 768 diverse residential buildings with 8 input variables. The results showed that tested tree partitioning methods outperformed other methods in recently published works of other researchers. Following the experience and guidelines from previous research which did not exploit recursive partitioning enough in this area, it was our challenge to put more focus on this type of machine learning methods. In addition, previous authors emphasized the importance of using occupational data (Liang et al., 2016) in addition to building characteristics. Our dataset is among the most extensive ones including a large number of building attributes as well as geospatial, heating, cooling, occupational, and meteorological data.

3 Data and model evaluation

The dataset was extracted from the database maintained by the Agency for Legal Trade and Real Estate Brokerage (APN) in Croatia. The initial data consisted of 2048 public buildings from Croatia such as kindergartens, schools, medical buildings, sport objects, health institutions, military, and all other types of buildings that are owned or occupied by the public sector. They were described by 141 attributes that can be grouped into geospatial, construction, heating, cooling, meteorological and energy data. The variable names are given in Table 1. The output variable was the total electricity consumption and the total natural gas consumption of each building in 2016. In the pre-processing stage, the outliers (i.e. the cases above the upper quartile) were detected and removed from the dataset variable, thus the final sample consisted of 1858 cases. For the purpose of this research the variable reduction was not used in the pre-modeling stage since all the tested methods incorporate algorithms for retaining the most important variables in the training process. Therefore, the variable reduction was conducted in the post-modeling phase.

Table 1. Variables describing public buildings

No.	Group of variables	Variable name
1.	Geospatial data	county, object region, type of object, object geo type, cultural heritage building
2.	Construction data	share of use of total building area, year of completion of construction, year of last restoration, flat gross floor area of building, useful area surface of building, object dim cooled area, object dim cooled surface area, object dim cooled volume area, number of floors, internal project temperature, share of windows surface
3.	Heating data	heated surface of the building, heated volume area of the building,

		installed power el. motor for pumps heat, type of heat pump, energy generating product, heating pump, total heat capacity of heat pump, total body heat radiator, total power body heat radiator, total body heat function oil, total power body heat function oil, total body heat other, total power body other, thermal power of heaters, primary heat sys using electrical heaters, installed capacity of electrical heaters, primary heating sys using split sys, installed electrical power of split sys heat, installed heat power of split sys heat, total heating power, factor of building shape f0, h1max. allowed coefficient of transmission heat loss per surface, transmission coefficient of heat loss, annual thermal energy needed 4heat, number of interior light luminaries
4.	Cooling data	object dimension of cooled area object dimension of cooled surface area object dimension of cooled volume area
5.	Meteorological data	air temperature
6.	Occupational data	number of employees, number of users, number of working days per week, number of working days per year, no of working hours per workday
7.	Energy coefficients of 9 specific parts of buildings d1,…,d9	object construction coeff. Transmission, object construction iso. thickness, object construction surface, object construction thickness (d1=roof, d2=floor, d3=windows, d4=shades, d5=heated ceiling, d6=unheated ceiling, d7=external wall, d8=doors, d9=unheated wall)
8.	**Output variable**	**Yearly electricity consumption (kWh)**

For the purpose of training and testing recursive partitioning methods, the total sample was randomly divided into the train and the test data, such that 70% of data (1486 cases) was used for training and 30% (372 cases) for testing. Data were normalized before training. The mean square error (MSE) is used as a common measure of performance in regression trees to determine final splitting. However, for final comparison of accuracy, we have followed the suggestion of Tofallis (Tofallis, 2015) to use the symmetric mean average percentage error (SMAPE) since it more fairly treats positive and negative residuals. The measure was computed according to:

$$SMAPE = 100\frac{1}{n}\sum_{i=1}^{n}\frac{|y_t - y_c|}{|y_t| + |y_c|} \qquad (1)$$

where y_t the real is output value, y_c is the predicted value, and *n* is the number of cases in the test sample.

4 Recursive partitioning methods

The four recursive partitioning methods were used, namely CART, CTREE, random forest and gradient boosted tree. All the computations were conducted using R software.

4.1 Classification and regression tree (CART)

The classification and regression tree (CART) suggested by Breiman et al. (1984) is the basic and most commonly used recursive partitioning method. In this research the regression variant of the CART was used such that the output produces a real number instead of a class probability. In its standard form, it builds a binary tree by splitting the input vectors at each node according to a function of a single input. For each input variable, the parent node is divided into child nodes by separating the objects with values lower and higher than the split point with the highest reduction of impurity. After repeating the splitting process for all input variables using each node as a new parent node until the tree reaches its maximum size, the stage of pruning begins and the algorithm prunes the tree back using the cross-validation procedure to select the right-sized tree. The algorithm considers all possible tree splits in order to find the most successful one by Gini index defined as (Apté et al., 1997):

$$Gini(t) = 1 - \sum_i p_i^2 \qquad (2)$$

where *t* is a current node and p_i is the probability of class *i* in *t*. Prune of misclassification error was used as the stopping rule, with minimum n=5. In the regression tree, the response for any observation is computed by following the path from the root node down to the appropriate terminal node of the tree, where the values for the splitting variables are observed, and the predicted response value is calculated by averaging response in that terminal node [10]. The limitation of CART trees is in their biasness regarding the variable selection, since it does not treat fair the variables of different types, categories, or missing values (Grömping, 2009).

4.2 Conditional inference tree (CTREE)

The conditional inference tree (CTREE) was proposed by Hothorn et al. (2006) as a tree partitioning method which does not use pruning, and is therefore faster than the CART and also overcomes the variable selection bias which exists in CART. The CTREE uses multiplicity-adjusted conditional tests to determine the predictors of an output and to generate a tree structure. It conducts the test of the null hypothesis of no association between an input variable and the output, and it calculates it both globally for each node and separately for each individual variable in non-terminal nodes. The smallest p-value is used to determine the variable which becomes a split variable. The tree grows until there is no further statistically significant split. This method is also robust since it can work with different types of variables and missing data.

4.3 Random forest (RF)

The random forest (RF), as the name of the method says, is a tree partitioning method that generates a collection of decision trees based on a random subset of the data, and each split within each tree is created based on a random subset of candidate variables (Hartshorn, 2016). The final response is obtained by averaging responses of the individual trees. According to the ambiguity decomposition (Krogh et al., 1995, in Louppe, 2014), the generalization error of the ensemble is to be lower than the average generalization error of its constituents. That is also the main advantage of the random forest method, since ensemble overcomes instability of single-tree techniques and improves its performance. The main shortcomings are in its complexity and computing time (Fan and Gray, 2005), (Grömping, 2009).]. The RF-CART algorithm was used in our experiments a random forest based on CART. In order to save computing time in pruning trees, we have used the complexity parameter cp=0.01 with ANOVA, such that the overall R-squared must increase by cp at each step. Splits that do not improve the fit by cp will likely be pruned off by cross-validation. The maximum depth parameter of any node was set to 30.

4.4 Gradient boosted trees (GBT)

This method uses boosting process in generating trees, meaning that trees are grown sequentially such that each successive tree uses information from previously grown trees. The aim is to minimize the error of the previous trees (Garreth et al., 2014). In this research the gradient boosting implemented in the R package xgboost was used. It involves resampling of observations and columns in each round of 10 cross-validation steps. The fitting of a decision tree is obtained by using the model residual errors as the outcome variable of the new model. The new decision tree adjusted by a shrinkage parameter *lambda* is added into the fitted function and the residuals are updated. *Lambda* of 0.01 was used in our research. Since it builds trees sequentially, this method is time-consuming comparing to other recursive partitioning methods. However, previous research has shown that it often outperformed other methods (Touzani et al., 2018).

5 Results

The graphical presentation of the tree partitioning for the first two methods that produce a single final tree is shown in Figure 1 (for CART) and in Figure 2 (for CTREE). It can be seen from Figure 1 that the CART method has extracted a smaller tree, with only 6 final splits and using only four features. The selected features are V9 (heated surface of the building) in the root node of the tree, V40 (total building power of cooling in kW) in the left branch, and variables V17 (number of working hours per workday) and V5 (number of users) in the right branch. The accuracy of the CART decision tree model in the sense of MAPE was 33.7324%. Parameters for CART were: Gini index, cp = 0.01, minsplit = 20, cp = 0.01, maxcompete = 4, maxsurrogate = 5, usesurrogate = 2, no. of cross-validation = 10, max.depth = 30.

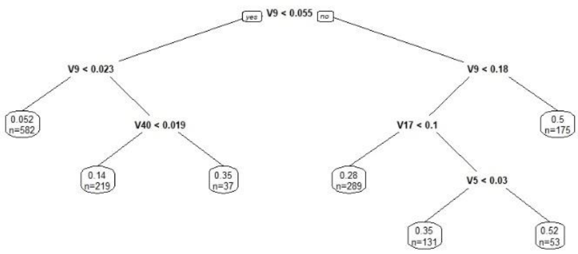

Figure 1. CART decision tree

The CTREE has produced a larger number of splits as shown in Figure 2. It has extracted 16 predictors as important, with XXX splits. Parameters for CTREE were: 19 terminal nodes, teststat = c("quad", "max"), testtype=c("Bonferroni", "MonteCarlo", "Univariate", "Teststatistic"), mincriterion=0.95, minsplit=20, minbucket=7, stump=FALSE, nresample=9999, maxsurrogate = 0, mtry = 10, savesplitstats = TRUE, maxdepth = 0. The accuracy of CTREE was slightly higher than the accuracy of CART, since it has produced MAPE of 31.21%.

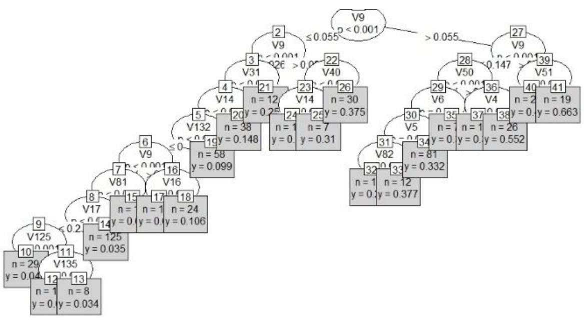

Figure 2. Conditional inference tree

The random forest method has created an ensemble of maximum 500 trees, while the number of variables tried at each split was 26. Since it does not produce a single tree as the output, it is possible to graphically observe the error conversion according to the number of generated trees, which is shown in Figure 3.

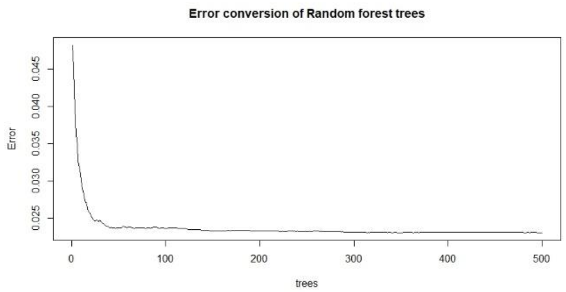

Figure 3. Error conversion of random forest

It can be seen in Figure 3 that the error converges as the number of trees increases and is relatively stable after 300 trees. The number of predictors extracted by random forest was 15, while its accuracy (MAPE) is 18.80%, which is higher than the previously tested CART and CTREE. The gradient boosted tree has produced the MAPE of 26.75%. The summary of the results of all four models is presented in Table 2.

Table 2. Accuracy of tested models

Recursive partitioning method	SMAPE (%)	No. of selected predictors
CART	33.73	4
CTREE	31.21	16
Random forest	18.80	15
Gradient boosted tree	26.75	20

It can be seen from Table 2 that the most accurate model was produced by the random forest method of recursive partitioning. In order to test the significance of the differences in results, the t-test of difference in proportions was conducted. The test has shown that the RF model is significantly different (p=0.000) from the CART model, as also significantly different from the CTREE model (p=0.000) and GBT model (p=0.049). Thus, it can be concluded that random forest performs significantly better than the other tested methods in modelling electricity consumption of public buildings.

Regarding the variable importance, it is interesting to analyse if the methods differ among themselves in selecting the important predictors. The selection of variables in all decision trees is obtained by using the information gain for regression trees calculated as (Hartshorn, 2016):

$$\boldsymbol{Information\ gain = n(MSE_{before} - MSE_{after})}$$

where $\boldsymbol{MSE_{before}}$ is the MSE before the split, $\boldsymbol{MSE_{after}}$ is the MSE after the split, and n is the number of data points that split operated on. The variable importance is obtained by calculating information gain across all splits for a certain variable. Due to a lack of space, only first five highly ranked predictors extracted by each method were presented in Table 3.

Table 3. First five predictors extracted by each method

Rank	CART	CTREE	Random forest	Gradient boosted tree
1	V9 (heated surface of the building)	V9 (heated surface of the building)	V9 (heated surface of the building)	V9 (heated surface of the building)
2	V40 (total building power of cooling in kW)	V31 (installed electric power of split system for heating in kW)	V4 (number of employees)	V4 (number of employees)
3	V17 (number of working hours per workday)	V40 (total building power of cooling in kW)	V5 (number of users)	V5 (number of users)
4	V5 (number of users)	V14 (cooled volume area of the building in m^2)	V29 (total installed thermal power of heaters in kW	V6 (number of working days per week)
5	V4 (number of employees)	V132 (cool energy generating product code)	V70 (number of interior light luminaries)	V14 (cooled volume area of the building in m^2)

Table 3 reveals that the tested methods show certain similarities in selecting important predictors. The variable *V9 (heated surface of the building)* was ranked as the most important by all four methods. That variable belongs to the group of heating data (see Table 1) as well as the variable *V31 (installed electric power of split system for heating in kW)* extracted as the second most important by CTREE, *V29 (total installed thermal power of heaters in kW* and *V70 (number of interior light luminaries)* extracted as important variables by RF. The two variables from the group of occupational data: *V5 (number of users)* and *V4 (number of employees)* were selected by CART, RF, and GBT among the five most important ones but not by CTREE. The CART method has extracted an additional occupational variable: *V17 (number of working hours per workday)* while GBT extracted *V6 (number of working days per week)*. The three variables from the group of cooling data were extracted in Table 3. The CART and CTREE have extracted *V40 (total building power of cooling in kW)*, while CTREE and GBT additionally selected *V14 (cooled volume area of the building in m2)* and *V132 (cool energy generating product code).*
The choice of important predictors generally shows that variables related to heating have the highest impact to electricity consumption, followed by the cooling-related variables, and occupational data.

6 Potentials for model implementation in reducing energy consumption in public sector

The experiments conducted in this research are a part of the research project “Methodological Framework for Efficient Energy Management by Intelligent Data Analytics” that aims to contribute the reduction of energy consumption of non-renewable natural resources by machine learning methods, such as artificial neural networks, recursive partitioning, support vector machines, and other methods. Due to the fact that buildings are the largest energy consumers, and that the state is in position to directly influence the energy consumption of public sector by allocating resources into measures to improve its energy efficiency, creating models that will support decisions on resource allocation is highly desirable.

Croatia has made significant steps by establishing the central Information system of energy management (ISGE) managed by the Agency for Legal Trade and Real Estate Brokerage (APN). However, the system still does not use machine learning to create prediction models or to extract important predictors of energy consumption. The models created in this research, especially the one based on random forest method can be implemented into ISGE as an intelligent module, a part of the web-based and mobile Internet of Things (IoT) applications that will automatically collect data, create models, and enable decision makers in determining actions that will lead to decreased energy consumption. Business analytics tools such as Alteryx, IBM Watson Analytics, Microsoft Azure Machine Learning, Amazon Web Services, and others enable to import data from ISGE system, R or Python scripts that define the algorithms, and create predictive models based on the methodological framework which selects the machine learning method that best fits the data and produce a prediction model.

7 Discussion and conclusion

The paper compares the accuracy and variable selection across four different recursive partitioning methods: classification and regression tree (CART), conditional inference tree (CTREE), random forest (RF), and gradient boosted tree (GBT) in modelling energy consumption of buildings in public sector. After data pre-processing, each method is trained and tested by using random subsampling procedure. The results have shown that the most accurate model was the one produced by the RF method which yielded the symmetric mean average percentage error (SMAPE) of 18.80%. The RF method significantly outperformed other tested recursive partitioning methods. The reason could be in the fact that RF uses ensemble of decision trees which improves the error in case of high-dimensional data that were present in this research. Although previous research showed that GBT usually produces more accurate results than other tree-partitioning methods, that was not the case in this research possibly due to a large number of input variables with a very similar effect on the error.

The most accurate RF model extracted 15 out of 141 predictors which belonged to the group of heating, cooling and occupational data. All four methods have extracted the variable *heated surface of the building* as the most important one.

The limitations of the recursive partitioning lie in the fact that in order to create efficient models for energy consumption of public buildings, other types of energy should be also considered, such as natural gas and water. In order to create a complete methodological framework, more machine learning methods should be compared and used in integrative manner. The created models have shown a potential of recursive partitioning methods in managing energy consumption in public buildings, and if implemented, could directly decrease energy consumption and expenditures in public sector, and significantly impact the state budget.

Acknowledgments

This work has been fully supported by Croatian Science Foundation under Grant No. IP-2016-06-8350 "Methodological Framework for Efficient Energy Management by Intelligent Data Analytics" (MERIDA).

References

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2-3), 197-210.

Breiman, L., Friedman, J., Ohlsen, R., & Stone, C. (1984). *Classification and regression trees*, Belmont, CA: Wadsworth International Group.

Chou, J. S., & Bui, D. K. (2014). Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Buildings*, 82, 437-446.

Chung, M., & Park, H. C. (2015). Comparison of building energy demand for hotels, hospitals, and offices in Korea. *Energy*, 92, 383-393. DOI: 10.1016/j.energy.2015.04.016

Fan, G., & Gray, J. B. (2005). Regression tree analysis using TARGET. *Journal of Computational and Graphical Statistics*, 14(1), 206-218.

Farzana, S., Liu, M., Baldwin, A., & Hossain, M. U. (2014). Multi-model prediction and simulation of residential building energy in urban areas of Chongqing, South West China. *Energy and Buildings*, 81, 161-169.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, *63*(4), 308-319.

Has, A., Zekić-Sušac, M., Modelling energy efficiency of public buildings by neural networks and its economic implications, in Zadnik-Stirn, L., Drobne, S. (Eds.), *Proceedings of the 14th International Symposium on Operations Research in Slovenia*, 27-29th September 2017, Bled, Slovenia. pp. 461-466.

Hartshorn, S. (2016). *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners*. Amazon Digital Services LLC: Seattle, Washington, USA.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3), 651-674.

Garreth, J., Witten, D, Hastie, T. & Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company.

Liang, X., Hong, T., & Shen, G. Q. (2016). Improving the accuracy of energy baseline models for commercial buildings with occupancy data. *Applied energy*, 179, 247-260. doi: 10.1016/j.apenergy.2016.06.141.

Louppe, G. (2014). Understanding random forests, From theory to practice, PhD dissertation, University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science. https://arxiv.org/pdf/1407.7502.pdf.

Mangold, M., Österbring, M., & Wallbaum, H. (2015). Handling data uncertainties when using Swedish energy performance certificate data to describe energy usage in the building stock. *Energy and Buildings*, 102, 328-336. doi: 10.1016/j.enbuild.2015.05.045.

Son, H., Kim, C., Kim, C., & Kang, Y. (2015). Prediction of government-owned building energy consumption based on an RReliefF and support vector machine model. *Journal of Civil Engineering and Management*, 21(6), 748-760.

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8), 1352-1362.

Tommerup, H., Rose, J., Svendsen, S. (2007). Energy-efficient houses built according to the energy performance requirements introduced in Denmark in 2006. *Energy and Buildings*, 39(10), 1123-1130.

Touzani, S., Granderson, J., Fernandes, S. (2018). Gradient boosting machine for modelling the energy consumption of commercial buildings, *Energy and Buildings*, 158, pp. 1533-1543.

Zekić-Sušac, M. (2017). Overview of prediction models for buildings energy efficiency. In Mašek Tonković Anka (Ed.), *Proceedings of the 6th International Scientific Symposium Economy of Eastern Croatia – Vision and Growth*. (pp. 697 - 706). Faculty of Economics in Osijek, 25.05. - 27.05.2017.

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. doi:10.1016/j.enbuild.2012.03.003

Yu, Z., Haghighat, F., Fung, B. C. M., & Yoshino, H. (2010). A decision tree method for building energy demand modeling. Energy and Buildings, 42(10), 1637-1646. doi:10.1016/j.enbuild.2010.04.006

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., & Ahrentzen, S. (2018). Random Forest based hourly building energy prediction. Energy and Buildings, 171, 11-25. doi:10.1016/j.enbuild.2018.04.008

Papadopoulos, S., Azar, E., Woon, W. L., & Kontokosta, C. E. (2018). Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. Journal of Building Performance Simulation, 11(3), 322-332. doi:10.1080/19401493.2017.1354919