AT-15-034

# An Ensemble Model for Predicting Energy Performance in Residential Buildings Using Data Mining Techniques

**Subramanian Manimaran** **Ibrahim AlBastaki** **J. Alamelu Mangai**


## ABSTRACT

*Concerns about waste of energy and its serious environmental impact have imposed a greater significance for studies and research in energy consumption. Since buildings contribute to 20% to 40% of the overall energy consumption in any country, there is a dire need for energy efficient building designs with more energy conservation properties. The success of designing an energy efficient building mainly depends on an accurate prediction of heating load (HL) and cooling load (CL) needed to maintain comfortable indoor air conditions. This paper proposes a data mining approach for predicting HL and CL for residential buildings. The proposed technique uses an ensemble method called* bagging*, which aggregates the predictions made by multiple REPTrees as base classifiers, instead of relying on a single prediction model. The technique uses information gain and reduced error pruning to build compact decision trees. The training data comprises details of 768 diverse residential buildings designed using Ecotect from the UCI machine learning repository (Bache and Lichman 2013). The proposed prediction model for HL has a correlation coefficient of 0.9985 and a mean absolute error MAE of 0.3811. The proposed prediction model for CL has a correlation coefficient of 0.983 and a mean absolute error MAE of 1.1065. The performance of the proposed ensemble method is also compared with neural networks as a base classifier.*


## INTRODUCTION

Energy consumption is usually split into three main sectors namely, industry, transport and "others." The third sector includes agriculture, service sector and residential. The rapid increase in living standards, desire for comfort levels, spending more time indoors, enhancement of building services, and use of computer and entertainment systems have increased building energy consumption on par with that of industry and transport. Building energy consumption accounts for 20% to 40% of the total energy consumption in the U.S. (Perez-Lombard et al. 2008). Buildings account for 40% of the total energy consumption in many countries and are also the major sources of carbon dioxide $CO_2$ emissions. With the increase in population, and hence the number of buildings, these $CO_2$ emissions have paved the way for more severe environmental impacts. The International Energy Agency (IEA) has identified building sectors as one of the most cost-effective sectors for reducing energy consumption. By reducing the overall energy demand, building energy efficient buildings can significantly reduce $CO_2$ emissions and promote nearly zero energy buildings NZEB. Potential energy savings can be achieved in residential buildings through proper heat, ventilation, and air-conditioning HVAC design, since this accounts for 40% of total building energy consumption. The HVAC systems help to maintain good indoor air quality. They also provide thermal comfort and are the largest energy consumers. Determining the heating load (HL) and cooling load (CL) is the most important step in deciding the size of the HVAC. Proper sizing of the HVAC system directly impacts the energy use, operational cost of the equipment, and the comfort level to the occupants. Several commercial buildings energy simulation software and tools exists in market for heating and cooling load calculations and estimating building design alternatives. However they are time consuming and need the end user to get them trained and later use the software. Also, the accuracy of the estimations made varies across different building management software (Tsanas and Xifara 2012).

**Subramanian Manimaran** and **Ibrahim AlBastaki** work for Emirates Global Aluminium, Dubai, United Arab Emirates. **J. Alamelu Mangai** works for BITS Pilani, Dubai Campus, Dubai, United Arab Emirates

Hence, this paper proposes a data mining approach for predicting the heating load and cooling load for residential buildings.

Data mining (DM) is an infant research field that has received significant interest in recent years. It is the process of extracting significant, non-trivial patterns from huge data repositories (Tan et al. 2006). It has its roots in statistics, artificial intelligence, machine learning, and data bases. It combines traditional statistical techniques with sophisticated algorithms to extract useful patterns from large data repositories. Ensemble mining methods combine the predictions made by multiple learners on a test data to increase the predictive accuracy. In this paper an ensemble method called *bagging* is proposed for accurate estimation of heating and cooling load for residential buildings.

The rest of the paper is organized as follows. Related Work is a survey of the various approaches proposed in research literature related to the present work, Proposed Prediction Model is a description of the proposed framework for prediction, Materials and Methods is a description of the data set used and the experimental results with discussion, and Conclusions and Future Work concludes with the highlights and findings of the present study.

## RELATED WORK

Simulation based modeling, statistical approaches, and artificial intelligence methods are used for energy prediction, as seen in research literature. Support vector machines (SVMs) (Dong et al. 2005) are used to forecast building energy consumption in the tropical region. The predictions made had a coefficient of variance less than 3% and error within 4%. A parallel implementation of support vector machine (Zhao and Magoules 2010) using Gaussian kernels is used to predict energy consumption in buildings. The need for longer training time, which is one of the major disadvantages of SVM, is overcome by parallelizing the kernel evaluations and kernel updates. This helped to accelerate the learning process. A third order polynomial support vector regression model is used for energy prediction and the proposed model is found to work well on small training sets (Mahdieh et al. 2013). Two forecasting models for building energy consumption are proposed (Noh and Rajagopal 2013); one of them is a time series model used for short term prediction and the second is a Gaussian model used for long term prediction. An artificial neural network (ANN) model for predicting chiller plant energy use in tropical climates is proposed in Yalcintas and Akkurt (2005). ANN is applied for forecasting building energy use based on building architectural and energy use parameters (Hawkins and Mumovic 2012). The load prediction potential for commercial buildings is improved using fuzzy logic (Jones 2012). The proposed algorithm uses modified learning from experience to generate membership functions using a training set. The membership functions are further improved using a recursive least squares method to generate a single power prediction output. Regression models are used (Zhang and Haghighat 2010) to predict the monthly heating demand for single-family residential sector in temperate climates. This model can be used by the architects or design engineers to find energy efficient solutions. A sky radiance distribution model is used (Hosobuchi et al. 2005) for calculating heating and cooling load in buildings.

The linear regression model is compared with random forests (Tsanas and Xifara 2012) for predicting heating load and cooling load in residential buildings. The results of this study have proved that machine learning tools can be used conveniently and accurately to estimate building parameters provided the test data resembles the training data used to build the model *a priori*. Regression models (Catalina et al. 2008) and a decision tree based model (Yu 2010) is used for predicting energy use intensity in buildings. Some of the machine learning algorithms like neural networks, decision trees, and support vector machines have been successfully employed for energy prediction as seen in research literature. However, the support vector machine based learners have some disadvantages, such as they need longer training time, more memory, selection of appropriate parameters, and operate better with small feature space than with large feature space. The neural network classifiers are unstable in nature. A small change to the training set will result in large variations in their predictions.

Ensemble methods are intensively studied in the fields of machine learning and pattern recognition. These methods use multiple base classifiers/regression models instead of one to predict the test data. The prediction made by each base classifier/regression model is then aggregated to make the final prediction. This resembles the process of building multiple experts and letting them vote (Tan et al. 2006). The probability of wrong predictions made is lower in the ensemble methods than using a single classifier/prediction model. Besides, they have the advantage of dealing with small sample size, high dimensionality, and complex data structures. Ensemble methods manipulate the training set in different ways to generate different subsets of the training set. One base classifier/regression model is modeled using each training set. The predictions made by all base classifiers/regression models on a test data are combined to get the final ensemble. The three common ensemble methods are bagging, boosting, and random forests. A detailed description of these methods are given in Upitz and Maclin (1999). Recently, their use in computational biology (Yang et al. 2010) weather forecasting is increasing (Gneiting and Raftery 2005).

## PROPOSED PREDICTION MODEL

### Bagging

Bootstrap aggregation also called *bagging* is a technique (Breiman 1996) used with any classification or regression methods to reduce the variance in prediction and hence improves the prediction process. Figure 1 shows the architecture of the bagging ensemble (Tan et al. 2006).

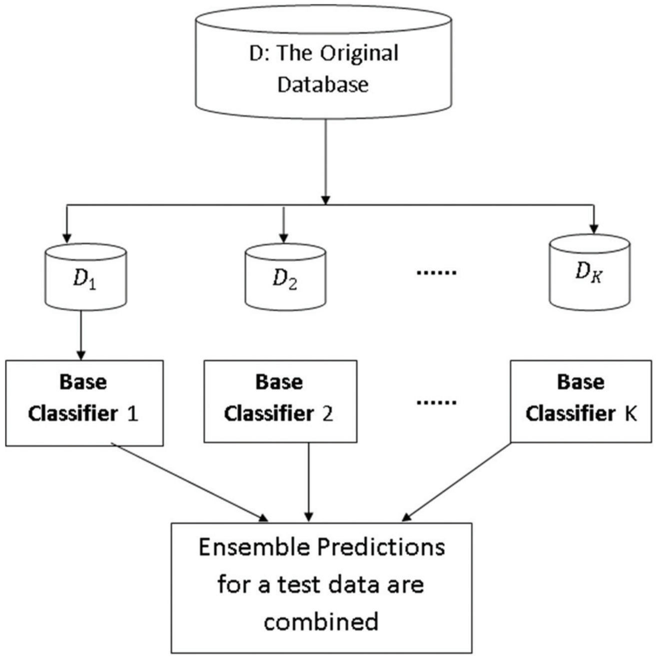


***Figure 1*** *An illustration of bagging ensemble.*

Many bootstrap samples are created from the training set, some prediction method is applied to each bootstrap sample, and the predictions made for a given test data are then combined using averaging for regression and majority voting for classification. These predictions are then aggregated to get the overall prediction, which will have a low variance from the actual data due to averaging. The details of this algorithm (Tan et al. 2006) are illustrated below.

***Algorithm Bagging.***

*Input:* The training data set $D$ of $N$ examples, No. of Base Classifiers/Regression Models $K$, the Test Data Set $T$

*Output:* The predicted value for each test instance in $T$. The output is a nominal value in case of classification and numeric in case of regression.

Steps:

1. Partition $D$ into $K$ subsets of size $N$ called *bootstrap samples*, using sampling with replacement. Let them be $\{D_1, D_2, D_3, \ldots\ldots\ D_K\}$.
2. Learn $K$ prediction methods $\{P_1, P_2, P_3, \ldots\ldots\ P_K\}$ each using the corresponding $\{D_1, D_2, D_3, \ldots\ldots\ D_K\}$ as training samples.
3. For each $t \in T$ do the following:
   a. Get the predicted output of each $P_i$ for $i = 1$ to $K$ for this test data.
   b. Combine the predictions made by all $P_i$s by averaging to get the overall prediction.

## Tree Based Classification and Regression

Tree-structured classification and regression methods are nonparametric and are the best options for a data analyst to get accurate and quick results (Sutton 2005). They require less domain knowledge and are better suited for data sets of larger dimensions and larger training set and are resistant to outliers. In case of classification, the target variable to be predicted is of nominal type and it is numeric in case of regression. Among the different algorithms available in literature, the classification and regression trees CART (Breiman et al. 1984) is a sophisticated program for fitting trees into data. The section "Classification Trees" is a short description about the use of CART for classification, the section "Regression Trees" is a short description about the use of CART for regression, and the section "REPTrees" is a short description of the REPTrees used in the proposed model for predicting heating load and cooling load in residential buildings.

**Classification Trees.** A decision tree model used for classification has one or more internal nodes and one external node. All external nodes are class labels. The internal nodes represent the attribute test condition. One or more branches descend from each internal node depending on the number of possible values of the splitting attribute. The subset of training examples whose attribute value matches with a certain branch are further split into smaller subsets if it is impure. A subset of training examples are said to be pure if all belong to the same class. In each iteration the tree growing process tries to find an optimal splitting attribute that results in pure partitions with the highest gain. Different tree growing algorithms use different measures to determine the purity of a set and also different splitting criteria. A measure used by the tree growing algorithm, namely entropy and information gain measure, for two class training examples are stated in Equation 1 and Equation 2.

$$\text{Entropy } S = \sum_{i=1}^{c} -P_i \log_2 P_i \tag{1}$$

where $S$ is the given set of training examples, $c$ is the number of classes, and $P_i$ is the probability of class $i$ in $S$.

$$\begin{gathered}\text{Information Gain } = \\ \text{Entropy before splitting } S \\ -\sum_{v \in \text{value}(A)} \frac{|D_v|}{|D|}\text{Entropy}(D_v)\end{gathered} \tag{2}$$

where $A$ is the candidate splitting attribute and $|D_v|$ is the number of training examples where the attribute $A$ has the value $v$.

**Regression Trees.** CART creates a regression tree for the given data set if the target variable is numeric. It uses the same procedure for classification, except that the leaf nodes are now a numeric value, which is the average of the value of the target variable of all training examples that reach that node. The best splitting attribute that is chosen at each stage of the tree growing process is the one that results in minimum sum of the squared differences between the target values of training examples corresponding to a node and their mean. If a set of


Published in ASHRAE Tranactions, Volume 121, Part 2

training examples $S$ can be divided into two subsets, $S_1$ and $S_2$, using a numeric attribute $x_i$, where $i$ = 1 to total number of training examples, then the squared difference criteria to be minimized is given in Equation 3

$$\sum_{x_i \in S_1} (t_i - \bar{t}_{S_1})^2 + \sum_{x_i \in S_2} (t_i - \bar{t}_{S_2})^2 \tag{3}$$

where $\bar{t}_{S_1}$ is the mean of the target values of the training examples in the subset $S_1$ and $\bar{t}_{S_2}$ is the mean of the target values of the training examples in the subset $S_2$. The trees generated are also pruned in order to avoid overfitting the training data. Several algorithms for pruning decision trees are proposed in the literature.

**REPTrees.** The proposed model for predicting heating load and cooling load in residential building uses bootstrap aggregation with REPTrees as base classifiers (Witten and Frank 2005). It is a fast decision tree learner and it builds a regression tree using variance. The resulting tree is pruned using reduced error pruning. The numeric values of the attribute are sorted only once.

The performance of the model is also compared with two other existing models, namely a neural network based model called MLP (Walter et al. 2005) and a support vector machine based regression model (Shevade et al. 2000) called SMO-reg. These are the two commonly used models as in research literature for energy performance in residential buildings.

**Performance Evaluation Metrics.** The performance of the predictive model is evaluated using two metrics, namely correlation coefficient and root mean square error RMSE. Correlation coefficient between actual and predicted values of an attribute is defined as in Equation 4:

$$\text{Correlation coefficient} = \frac{S_{pa}}{\sqrt{S_p S_a}} \tag{4}$$

where $S_{pa} = \sum_i^N (p_i - \bar{p})(a_i - \bar{a}) / (N-1)$ where $\bar{p}$, $\bar{a}$ are the averages, respectively, and

$$S_p = \sum_i^N (p_i - \bar{p})^2 / (N-1) \quad \text{and}$$

$$S_a = \sum_i^N (a_i - \bar{a})^2 / (N-1) \tag{5}$$

A correlation coefficient of 1 indicates that the values are perfectly correlated, while a correlation coefficient of 0 implies no correlation exists between them.

If $p_i$ is the predicted value for $i$th instance, $a_i$ is the actual value for $i$th instance, and $N$ is the total number of instances in the given data set, the root mean square error (RMSE) is given by Equation 6.

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(p_i - a_i)^2}{N}} \tag{6}$$

The smaller the RMSE, the better is the performance of the model.

## MATERIALS AND METHODS

### Data Set Description

The details of the attributes in the training data set (Bache and Lichman 2013) used to build the proposed prediction model are shown in Table 1. This data set is collected after performing energy analysis on 12 different building shapes simulated in Ecotect. It has 768 instances, 8 input features, and 2 output features.

**Table 1. Data Set Description**

| Feature | Input/Output | Min | Max | Mean | S.D |
|---|---|---|---|---|---|
| Relative Compactness | Input | 0.62 | 0.98 | 0.764 | 0.106 |
| Surface Area | Input | 514.5 | 808.5 | 676.708 | 88.086 |
| Wall Area | Input | 245 | 416.5 | 318.5 | 43.626 |
| Roof Area | Input | 110.25 | 220.5 | 176.604 | 45.166 |
| Overall Height | Input | 3.5 | 7 | 5.25 | 1.751 |
| Orientation | Input | 2 | 5 | 3.5 | 1.119 |
| Glazing Area | Input | 0 | 0.4 | 0.234 | 0.133 |
| Glazing Area Distribution | Input | 0 | 5 | 2.813 | 1.551 |
| Heating Load | Output | 6.01 | 43.1 | 22.307 | 10.09 |
| Cooling Load | Output | 10.9 | 48.03 | 24.588 | 9.513 |

The proposed model is implemented using the WEKA (Witten and Frank 2005) data mining tool. The prediction model is induced using ten-fold cross validation.

## Results and Discussion

The results of the prediction model for heating load and cooling load are shown in Table 2.

The model has generated ten REPTrees, one in each fold for predicting each target feature. The results of the tenth fold for heating load prediction are given in the Appendix A. The results of the ten fold for cooling load prediction are given in the Appendix B. For simplicity in interpreting the model, the attributes in the training data set are transformed to simple variable names from X1 to X2, as illustrated in Table 3. The internal nodes of the tree are test conditions on these variable names. The leaf nodes of a branch are the regression models that fit the subset of training examples that reach this node.

The performance of the proposed model is also compared with the most commonly used artificial neural network model MLP and regression based support vector machine SMO-reg for predicting the target variables. The parameters used for modeling MLP are hidden layers = 4, learning rate = 0.3, momentum = 0.2, and training time = 500. The SMO-reg was modeled using Polykernel function with an exponent 1.0 and a regularization constant 1.0.

The Gaussian RBF kernel based SVM was also modeled using gamma value of 0.01. Table 4 shows a comparative analysis of these models with the proposed model for heating load prediction. As seen from the results, the proposed model has a higher correlation coefficient than the other two models. It also has a lower value of MAE and RMSE than the other models, which is a desirable characteristic of any prediction model. The difference in the performance of the SMO-reg with both kernels is statistically insignificant. Table 5 shows a comparative analysis of these models with the proposed model for cooling load prediction.

As seen from the results, the proposed model performs better than the other two models for the same data set.

## CONCLUSION AND FUTURE WORK

Due to the increase in awareness of environmental safety, optimizing energy usage has become the dire need. Buildings account for 40% of the total energy consumption in many countries and are the major sources of $CO_2$ emissions. This

**Table 2. Results of the Proposed Prediction Model**

| Target Feature | Correlation Coefficient | MAE | RMSE |
|---|---|---|---|
| Heating Load | 0.9985 | 0.3811 | 0.5521 |
| Cooling Load | 0.983 | 1.1065 | 1.7433 |

**Table 3. Description of the Internal Nodes of the Proposed Model**

| Attribute | Variable Name | Attribute | Variable Name |
|---|---|---|---|
| Relative Compactness | X1 | Overall height | X5 |
| Surface Area | X2 | Orientation | X6 |
| Wall Area | X3 | Glazing Area | X7 |
| Roof Area | X4 | Glazing Area Distribution | X8 |

**Table 4. Comparative Analysis of Various Heating Load Prediction Models**

| Model | Correlation Coefficient | MAE | RMSE |
|---|---|---|---|
| MLP | 0.9941 | 0.8401 | 1.1111 |
| SMO-reg - Polykernel | 0.9546 | 2.0402 | 3.0051 |
| SMO-reg - Gaussian kernel | 0.9529 | 2.0314 | 3.0761 |
| **Proposed** | **0.9985** | **0.3811** | **0.5521** |

**Table 5. Comparative Analysis of Various Cooling Load Prediction Models**

| Model | Correlation Coefficient | MAE | RMSE |
|---|---|---|---|
| MLP | 0.9771 | 1.5178 | 2.0825 |
| SMO-reg - Polykernel | 0.9363 | 2.2057 | 3.3605 |
| SMO-reg - Gaussian kernel | 0.9362 | 2.2772 | 3.4255 |
| **Proposed** | **0.983** | **1.1065** | **1.7433** |

paper has implemented a data mining model that will help in an efficient HVAC design for residential buildings and hence optimize the energy usage. The predictive model automatically predicts the HL and CL for a proposed residential building. Experiments are done with a benchmarking data set and the model proposed has a correlation coefficient of 0.9985 for HL and 0.983 for CL. Our future work is to identify the relevant significant parameters that can be used to train the prediction model.

## APPENDIX A—REPTREE FOR PREDICTING HEATING LOAD

```
REPTree
============



X4 < 183.75
| X1 < 0.81
| | X7 < 0.18
| | | X7 < 0.05 : 27.02 (7/5.07) [0/0]
| | | X7 >= 0.05
| | | | X2 < 649.25 : 36.36 (15/0.43) [11/0.52]
| | | | X2 >= 649.25 : 32.59 (10/0.16) [7/0.16]
| | X7 >= 0.18
| | | X7 < 0.33
| | | | X3 < 379.75
| | | | | X6 < 3.5
| | | | | | X8 < 3.5 : 39.3 (4/0.66) [2/0.66]
| | | | | | X8 >= 3.5 : 39.92 (6/0.01) [0/0]
| | | | | X6 >= 3.5 : 39.19 (8/0.35) [4/0.2]
| | | | X3 >= 379.75 : 36.38 (14/0.22) [6/0.24]
| | | X7 >= 0.33
| | | | X3 < 379.75 : 41.85 (17/0.44) [3/0.62]
| | | | X3 >= 379.75
| | | | | X8 < 4.5
| | | | | | X8 < 3.5
| | | | | | | X8 < 2.5 : 40.46 (4/0.05) [3/0.15]
| | | | | | | X8 >= 2.5 : 39.41 (3/0.15) [4/0.14]
| | | | | | X8 >= 3.5 : 40.49 (4/0.06) [1/0.01]
| | | | | X8 >= 4.5 : 39.26 (4/0.14) [2/0.17]
| X1 >= 0.81
| | X7 < 0.33
| | | X7 < 0.05 : 18.19 (6/5.16) [6/3.25]
| | | X7 >= 0.05
| | | | X1 < 0.84
| | | | | X7 < 0.18 : 23.73 (11/0.14) [4/0.34]
| | | | | X7 >= 0.18 : 25.5 (15/0.29) [8/0.6]
| | | | X1 >= 0.84
| | | | | X7 < 0.18
| | | | | | X3 < 306.25
| | | | | | | X4 < 128.63 : 24.29 (13/0.02) [6/0.04]
| | | | | | | X4 >= 128.63
| | | | | | | | X8 < 1.5 : 25.74 (2/0.3) [2/0.26]
| | | | | | | | X8 >= 1.5 : 26.48 (12/0.21) [5/0.66]
| | | | | | X3 >= 306.25
| | | | | | | X8 < 4.5
| | | | | | | | X8 < 2.5 : 28.78 (5/0.32) [4/0.25]
| | | | | | | | X8 >= 2.5
| | | | | | | | | X6 < 3.5 : 28.38 (6/0.21) [0/0]
| | | | | | | | | X6 >= 3.5 : 29.36 (2/0.01) [2/0.57]
| | | | | | | X8 >= 4.5 : 29.44 (4/0.15) [0/0]
| | | | | X7 >= 0.18
| | | | | | X3 < 306.25
| | | | | | | X4 < 128.63 : 28.46 (12/0.05) [9/0.04]
| | | | | | | X4 >= 128.63 : 29.18 (17/0.21) [4/0.37]
| | | | | | X3 >= 306.25 : 32.4 (10/0.27) [8/0.34]
| | X7 >= 0.33
| | | X2 < 600.25
| | | | X3 < 306.25
| | | | | X1 < 0.92
| | | | | | X8 < 2.5
| | | | | | | X6 < 3.5 : 32.6 (2/0.02) [2/0.03]
| | | | | | | X6 >= 3.5 : 31.99 (4/0.04) [0/0]
| | | | | | X8 >= 2.5 : 31.98 (8/0.18) [2/0.02]
| | | | | X1 >= 0.92 : 32.51 (10/0.06) [10/0.05]
| | | | X3 >= 306.25
| | | | | X8 < 1.5 : 36.71 (5/0.24) [0/0]
| | | | | X8 >= 1.5 : 35.49 (7/0.25) [9/0.49]
| | | X2 >= 600.25 : 29.13 (12/0.31) [9/0.35]
X4 >= 183.75
| X7 < 0.05
| | X1 < 0.65 : 9.21 (6/0.91) [0/0]
| | X1 >= 0.65
| | | X1 < 0.7 : 6.98 (11/0.03) [1/0.05]
```

```
| | | X1 >= 0.7 : 6.25 (8/0.03) [2/0.02]
| X7 >= 0.05
| | X2 < 771.75
| | | X7 < 0.33
| | | | X7 < 0.18
| | | | | X2 < 698.25 : 10.3 (12/0.02) [7/0.01]
| | | | | X2 >= 698.25
| | | | | | X1 < 0.68 : 11.51 (14/0.02) [5/0.02]
| | | | | | X1 >= 0.68
| | | | | | | X6 < 4.5 : 10.93 (18/0.07) [8/0.06]
| | | | | | | X6 >= 4.5
| | | | | | | | X8 < 4.5 : 10.95 (3/0.14) [4/0.04]
| | | | | | | | X8 >= 4.5 : 12.42 (2/1.6) [0/0]
| | | | X7 >= 0.18 : 12.46 (38/0.1) [34/0.19]
| | | X7 >= 0.33
| | | | X1 < 0.68 : 15.12 (12/0.03) [6/0.03]
| | | | X1 >= 0.68
| | | | | X8 < 4.5 : 14.45 (38/0.05) [11/0.03]
| | | | | X8 >= 4.5
| | | | | | X2 < 722.75
| | | | | | | X1 < 0.73 : 12.54 (3/0.02) [0/0]
| | | | | | | X1 >= 0.73 : 14.35 (2/0.01) [2/0.01]
| | | | | | X2 >= 722.75 : 14.25 (4/0.01) [0/0]
| | X2 >= 771.75
| | | X7 < 0.33
| | | | X2 < 796.25
| | | | | X7 < 0.18 : 15.26 (12/0.01) [6/0.02]
| | | | | X7 >= 0.18 : 16.93 (13/0.1) [9/0.03]
| | | | X2 >= 796.25
| | | | | X7 < 0.18 : 12.85 (15/0.02) [5/0.02]
| | | | | X7 >= 0.18
| | | | | | X8 < 1.5 : 15.09 (6/0) [1/0]
| | | | | | X8 >= 1.5 : 14.28 (9/0.09) [6/0.11]
| | | X7 >= 0.33
| | | | X1 < 0.63 : 16.66 (12/0.55) [9/0.07]
| | | | X1 >= 0.63 : 18.78 (15/0.31) [7/0.07]
```

## APPENDIX B—REPTREE FOR PREDICTING COOLING LOAD

```
REPTree
============

X5 < 5.25
| X7 < 0.18
| | X1 < 0.65
| | | X1 < 0.63
| | | | X7 < 0.05 : 12.11 (4/0) [0/0]
| | | | X7 >= 0.05 : 14.22 (15/0.05) [5/0.02]
| | | X1 >= 0.63
| | | | X7 < 0.05 : 16.74 (2/0) [0/0]
| | | | X7 >= 0.05 : 19.28 (12/0.01) [6/0.01]
| | X1 >= 0.65
| | | X8 < 0.5
| | | | X3 < 281.75 : 11.27 (8/0.08) [2/0]
| | | | X3 >= 281.75 : 12.13 (11/0.04) [1/0.19]
| | | X8 >= 0.5
| | | | X1 < 0.73
| | | | | X8 < 3.5
| | | | | | X6 < 2.5 : 13.78 (4/0.12) [5/0.08]
| | | | | | X6 >= 2.5 : 14.14 (16/0.2) [9/0.2]
| | | | | X8 >= 3.5
| | | | | | X2 < 722.75 : 14.07 (6/0.03) [0/0]
| | | | | | X2 >= 722.75 : 14.64 (11/0.04) [3/0.08]
| | | | X1 >= 0.73 : 13.47 (12/0.03) [7/0.02]
| X7 >= 0.18
| | X3 < 330.75
| | | X7 < 0.33
| | | | X3 < 306.25
| | | | | X8 < 2.5 : 15.4 (12/0.06) [11/0.15]
| | | | | X8 >= 2.5 : 15.01 (19/0.17) [17/0.25]
| | | | X3 >= 306.25 : 16.03 (7/0.03) [6/0.08]
| | | X7 >= 0.33
| | | | X1 < 0.68 : 17.87 (12/0.04) [6/0.44]
| | | | X1 >= 0.68
| | | | | X6 < 3.5
| | | | | | X8 < 2.5
| | | | | | | X6 < 2.5 : 16.59 (7/0.04) [3/0.1]
| | | | | | | X6 >= 2.5 : 17.14 (3/0.04) [1/0.02]
| | | | | | X8 >= 2.5
| | | | | | | X8 < 3.5 : 15.87 (6/0.12) [0/0]
| | | | | | | X8 >= 3.5 : 16.44 (9/0.3) [1/0.08]
| | | | | X6 >= 3.5
| | | | | | X8 < 4.5 : 17.18 (17/0.08) [7/0.04]
| | | | | | X8 >= 4.5
| | | | | | | X1 < 0.73 : 16.2 (3/0.39) [0/0]
| | | | | | | X1 >= 0.73 : 16.78 (2/0.01) [1/0.01]
| | X3 >= 330.75
| | | X2 < 796.25
| | | | X7 < 0.33
| | | | | X8 < 1.5 : 21.14 (3/0) [0/0]
| | | | | X8 >= 1.5
| | | | | | X8 < 3.5 : 20.43 (3/0) [6/0.06]
| | | | | | X8 >= 3.5 : 19.95 (7/0.07) [3/0.01]
| | | | X7 >= 0.33
| | | | | X8 < 2.5 : 22.16 (8/0.16) [1/0.01]
| | | | | X8 >= 2.5
| | | | | | X8 < 4.5 : 21.44 (3/0.16) [6/0.36]
| | | | | | X8 >= 4.5 : 20.79 (4/0.18) [0/0]
| | | X2 >= 796.25
| | | | X7 < 0.33
| | | | | X8 < 1.5 : 15.8 (6/0) [1/0.03]
| | | | | X8 >= 1.5 : 15.11 (9/0.05) [6/0.09]
| | | | X7 >= 0.33
| | | | | X6 < 3.5 : 17.14 (7/0.01) [4/0.06]
| | | | | X6 >= 3.5 : 16.59 (5/0.15) [5/0.23]
X5 >= 5.25
| X2 < 624.75
| | X7 < 0.33
| | | X8 < 0.5 : 23.93 (6/5.69) [6/8.51]
| | | X8 >= 0.5
```

```
| | | | X2 < 600.25
| | | | | X3 < 306.25
| | | | | | X7 < 0.18
| | | | | | | X2 < 551.25 : 26 (13/0.04) [6/0.08]
| | | | | | | X2 >= 551.25 : 28.53 (14/6.11) [7/8.09]
| | | | | | X7 >= 0.18
| | | | | | | X1 < 0.92 : 31.42 (17/5.76) [4/7.69]
| | | | | | | X1 >= 0.92 : 29.78 (12/0.09) [9/0.1]
| | | | | X3 >= 306.25
| | | | | | X7 < 0.18 : 31.48 (17/4.32) [6/3.25]
| | | | | | X7 >= 0.18 : 34.55 (10/5.54) [8/3.07]
| | | | X2 >= 600.25
| | | | | X7 < 0.18
| | | | | | X6 < 3.5 : 26.53 (6/1.95) [1/2.76]
| | | | | | X6 >= 3.5
| | | | | | | X8 < 3.5 : 24.84 (3/0.02) [2/0.18]
| | | | | | | X8 >= 3.5 : 27.2 (2/1.45) [1/6.33]
| | | | | X7 >= 0.18 : 27.47 (15/1.35) [8/1.81]
| | X7 >= 0.33
| | | X1 < 0.88
| | | | X1 < 0.84 : 30.39 (12/3.27) [9/2.4]
| | | | X1 >= 0.84
| | | | | X6 < 2.5 : 34.08 (3/3.25) [0/0]
| | | | | X6 >= 2.5
| | | | | | X6 < 4.5 : 31.96 (7/3.47) [4/2.96]
| | | | | | X6 >= 4.5 : 33.79 (4/6.42) [0/0]
| | | X1 >= 0.88
| | | | X1 < 0.94
| | | | | X8 < 2.5
| | | | | | X6 < 4.5
| | | | | | | X6 < 2.5 : 38.82 (2/0.22) [0/0]
| | | | | | | X6 >= 2.5 : 36.02 (3/0.05) [1/0.7]
| | | | | | X6 >= 4.5 : 40.83 (2/0.03) [0/0]
| | | | | X8 >= 2.5 : 35.99 (5/2.85) [8/5.88]
| | | | X1 >= 0.94
| | | | | X6 < 3.5 : 33.17 (4/0.01) [6/0.09]
| | | | | X6 >= 3.5
| | | | | | X8 < 1.5 : 32.9 (2/0) [1/0.04]
| | | | | | X8 >= 1.5 : 33.87 (4/0.01) [3/0.01]
| X2 >= 624.75
| | X2 < 649.25
| | | X7 < 0.33
| | | | X6 < 3.5
| | | | | X8 < 3.5
| | | | | | X7 < 0.05 : 35.58 (3/9.27) [0/0]
| | | | | | X7 >= 0.05
| | | | | | | X6 < 2.5 : 42.45 (6/1.07) [2/2.15]
| | | | | | | X6 >= 2.5
| | | | | | | | X8 < 2.5 : 36.39 (3/1.23) [2/2.62]
| | | | | | | | X8 >= 2.5 : 42.03 (3/1.32) [0/0]
| | | | | X8 >= 3.5 : 36.54 (8/0.64) [4/14.94]
| | | | X6 >= 3.5
| | | | | X7 < 0.18 : 38.66 (6/14.38) [5/15.15]
| | | | | X7 >= 0.18
| | | | | | X8 < 2.5 : 43.97 (4/14.62) [1/6.28]
| | | | | | X8 >= 2.5 : 41.08 (4/11.88) [3/13.58]
| | | X7 >= 0.33 : 42.61 (17/13.12) [3/21.26]
| | X2 >= 649.25
| | | X7 < 0.18
| | | | X8 < 0.5 : 29.66 (3/0.03) [0/0]
| | | | X8 >= 0.5 : 33.98 (10/0.06) [7/0.1]
| | | X7 >= 0.18
| | | | X7 < 0.33 : 36.88 (14/0.34) [6/0.31]
| | | | X7 >= 0.33
| | | | | X8 < 2.5 : 40.2 (4/0.17) [3/0.38]
| | | | | X8 >= 2.5
| | | | | | X8 < 3.5 : 39.17 (3/0.41) [4/0.75]
| | | | | | X8 >= 3.5
| | | | | | | X8 < 4.5 : 40.2 (4/0.11) [1/0.44]
| | | | | | | X8 >= 4.5 : 39 (4/0.68) [2/1]
```

## REFERENCES


Dong, B., C. Cao, and S.E. Lee. 2005 Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings* 37:545–53.

Bache, K., and M. Lichman. 2013. UCI machine learning repository University of California, School of Information and Computer Science: Irvine, CA. http://archive.ics.uci.edu/ml.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24:123–40.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and regression trees*. Pacific Grove, CA: Wadsworth.

Catalina, T., J. Virgone, and E. Blanco. 2008. Development and validation of regression models to predict monthly heating demand for residential buildings. *Energy and Buildings* 40:1825–32.

Delashmit, Walter H., and Michael T. Manry. 2005. Recent developments in multilayer perceptron neural networks. *Proceedings of the 7th Annual Memphis Area Engineering and Science Conference*, MAESC.

Gneiting, Tilmann, and Adrian E. Raftery. 2005. Weather forecasting with ensemble methods. *Science* 310(14):248–49.

Hawkins, David, and Dejan Mumovic. 2012. Applying artificial neural networks to estimate building energy use in the early stages of architectural design. CIBSE ASHRAE Technical Symposium, Imperial College, London UK. 18th and 19th April 2012.

Hosobuchi, Hayato, Harunori Yoshida, and Yoshiaki Uetani, 2005. Calculation Of The heating and cooling load of buildings using a sky radiance distribution model. Ninth International IBPSA Conference, Canada, August 15–18, 42734.

IEA. Available at www.iea.org. Paris: International Energy Agency.

Jones, Tyler C., David M. Auslander, Jay Taneja, Jason Trager, Michael Sankur, and Therese Peffer. 2012. Improved methods to load prediction in commercial buildings. ACEEE Summer Study on Energy Efficient

Buildings, August. http://www.cs.berkeley.edu/~taneja/publications/jones12baseline.pdf.

Mahdieh, Mahdis, Milad Mohammadiy, and Pooya Zhsaniz. 2013. A prediction reference model for air conditioning systems in commercial buildings. http://cs229.stanford.edu/proj2013/MahdiehMohammadiEhsani-Y2E2BuildingEnergyStudy.pdf.

Melek Yalcintas, Sedat Akkurt, Artificial neural networks applications in building energy predictions and a case study for tropical climates", Int. J. Energy Res. Vol. 29, pp. 891–901, 2005.

Noh, Hae Young, and Ram Rajagopal. 2013. Data-driven forecasting algorithms for building energy consumption. Proceedings of the SPIE 8692, Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems.

Perez-Lombard, L., J. Ortiz, and C. Pout. 2008. A review on buildings energy consumption information. *Energy and Buildings* 40(3):394–98.

Shevade, S.K., S.S. Keerthi, C. Bhattacharyya, and K.R.K. Murthy. 2000. Improvements of the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks* 11(5).

Sutton, C.D. 2005. Classification and regression trees, bagging and boosting. *Handbook of Statistics* 24:303–29.

Tan, P., M. Steinbach, and V. Kumar. 2006. *Introduction to data mining*. New York: Pearson Education.

Tsanas, Athanasios, and Angeliki Xifara. 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings* 49:560–567.

Upitzm, David, and Richard Maclin. 1999. Popular ensemble methods: An empirical study. *Journal Artificial Intelligence Research* 11:109–98.

Witten, I.H., and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques with java implementations*. Morgan Kaufmann: San Francisco.

Yang, Pengyi, Yee Hwa Yang, Bing B. Zhou, and Albert Y. Zomaya. 2010. A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics* 4:296–308.

Yu, Z., F. Haghigrat, B.C.M. Fung, and H. Yoshimo. 2010. A decision tree method for building energy demand modeling. *Energy and Buildings* 42:1637–46.

Zhang, J., and F. Haghighat. 2010. Development of artificial neural network based heat convection for thermal simulation of large rectangular cross-sectional area earth-to-earth heat exchanges. *Energy and Buildings* 42(4):435–40.

Zhao, Hai Xiang and Frederic Magoules. 2010. Parallel Support Vector Machines Applied to the Prediction of Multiple Buildings Energy. *Journal of Algorithms & Computational Technology* 4(2): 231–50.



Published in ASHRAE Tranactions, Volume 121, Part 2