**Selecting a statistical model to predict the energy efficiency for residential buildings**

***Walter CIOS (Ryerson Student ID: 500791945)***

***CKME136 Capstone Project: Fall 2017 / Final Results - Nov 20, 2017***

**Introduction**

The research on building simulation can be traced back to the 1960’s when the US government was involved in projects to evaluate the thermal environment in fallout shelters. Since its inception, building simulation has evolved and several building energy simulation tools are currently used in order to analyze or forecast the building energy consumption.

Building enclosures have become more energy efficient as energy codes have become more stringent in recent years, highlighting the importance of improvements such as better windows, orientation, or insulation. The heating and cooling load calculation is the first step of the HVAC design procedure.

Heating and cooling load are used to right-size the HVAC equipment of a building. The load calculations are determining the equipment selection and the duct design needed to deliver heating and conditioned air to the rooms of a house. In the HVAC industry there was and still is a resistance to perform accurate load calculation. The user of HVAC systems are complaining that multiple technicians are quoting loads needs different and most have as result and oversized equipment. This is no surprise since larger oversized equipment brings a higher profit to both suppliers and the installation technicians. It’s very simple to identify which stakeholders are in favor of a proper load calculation like users, architects, and building designers or against it like some suppliers and HVAC technicians. Oversizing the HVAC system is not only an issue of increased cost, is also detrimental to energy use, comfort, indoor air quality, building and equipment durability.

Statistical and machine learning tools can help create models that predict the heating and cooling load. Having a model that is properly trained, will help providing fast answers by varying some building design parameters.

The scope of this project is to build a statistical model to predict the cooling load and to select the best performing model that provides the highest accuracy. For this scope, RStudio will be used. The quantitative research will also investigate which variables are influencing the most the heating efficiency and will explore the trade-offs between accuracy and interpretability of the model.

**Research Questions**

* Can the cooling load be predicted accurately with a statistical model?
* Which variable are influencing the most the cooling efficiency?
* Explore trade-offs between accuracy and interpretability of the model.
* Investigate energy impact when changing some buildings parameters.

This model and analysis will help HVAC equipment selection and prevent purchasing of oversized equipment.

**Literature Review**

**Energy Performance Prediction**

Arlan Burdick - “Strategy Guidline: Accurate Heating and Cooling Load Calculations”

In order to predict the heating and cooling load calculation, the 2011 study from A. Burdick is answering questions on: Why we do a load calculation? What are the critical inputs? What are the risks of an incorrect calculation? The heating and cooling load calculation and prediction is the first step of a building HVAC design in order to select the equipment needed to maintain comfortable indoor air conditions. There is a tendency that HVAC contractors are trying to oversize the equipment what is detrimental from the point of view of comfort, energy use, air quality, and price. What I found important from this study is that orientation is a critical input and the degree of shading from the sun can have a great impact on a certain house heat gain.

Athanasios Tsanas, Angeliki Xifara – “Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools”

This study by two professors from the Oxford and Cardiff University in UK is the piece of literature that I studied in most detail since is based on the dataset that is being used in my current Capstone research project. Because I will be explaining the dataset in detail in another section, I want to comment here only on some methods used in the professor’s study. After assessing that the data is not normal, they used the Spearman rank correlation coefficient to review the association of each input variable with each of the output variables. The study then refers to a total new regression method for me, the iteratively reweighted least squares (IRLS) that adjusts weights in the coefficients of the classical regression scheme to reduce the effect of the outliers in producing the fitting curve and this way obtain an improved least square estimate. One of the interesting conclusions of the research was that the most important input variable is not the most correlated with either output variable. Altogether a very interesting study although in my opinion at a very high technical level.

“A Comparison of Ten Different Building Performance Simulation Tools” – S.Attia, L.Beltran, A.D.Herde. J. Hensen

The reason I researched this paper on different performance simulation tools was to find out more details on what are new simulation tools since Ecotect, the software used to produce the dataset I’m working on (see Autodesk Ecotect reference), was recently discontinued. The paper highlights that Ecotect was considered together with eQUEST and IES VE as tools used during the early design phases, and have been replaced by newer tools like DOE-2, EP, and EPSU. This review helps me understand what other building performance simulation tools are currently available.

“Modelling energy efficiency performance of residential buildings stocks based on Bayesian statistical model” – Marta Braulio-Gonzalo, Pablo Juan, Maria D. Bovea, Maria Jose Rua

This study presented at the University Jaume 1 from Castello, Spain is treating very similar subjects to my study. The dataset studied in this study is an energy simulation of 240 cases using an software called Energy Plus (my dataset had 768 simulation done with software Ecotect). The take–away are an INLA (Integrated Nested Laplace Approximation) package in R that was used and the approach to inference using differential equation that are considering the influence of a set of particular variables. What I also liked was in Fig. 9 the graphical presentation of correlation between the predicted and observed data for two of the models. Inspired by this study I will research the R-INLA package for my study.

“A methodology for predicting the energy performance and indoor thermal comfort of residential stocks on the neighbourhood and city scales – case study in Spain” Marta Braulio-Gonzalo, Maria D. Bovea, Maria Jose Rua, Pablo Juan

This study is by the same authors as above but this time they treat indoor thermal comfort what is exactly related to the heating and cooling that are the output variables in my study. The difference to my study is that here the prediction of energy performance is based on a complex prediction method that combines physics modelling, statistical inference in a GIS environment. Although I was not interested in the GIS and mapping part, this study is also using the INLA package in R and identifies key variables that affect the building thermal performance. A take-away was the way it was presented (in table 8) the response variable and the corresponding prediction equation. The study established that regarding orientation, the lowest energy demand for heating went to the north-oriented buildings. I will review to see if in my study I will come to similar orientation related conclusions.

“Analysis and diagnosis of the energy performance of buildings and districts: Methodology, validation and development of Urban Energy Maps – Fabrizio Ascione and group (see reference)

This study was done in Italy. Although energy maps was not my focus of interest, an interesting approach was looking at the energy performance for both winter and summer and concluding that the demand for cooling (in Benevento, Italy) in the summer is higher than the demand for heating in the winter. The article analyzing the shading effect is proposing percentages of reduction of cooling needs (see Table 11). The model is not presented in detail but the authors are mentioning that they are working on transferring the model to open source – e.g. Quantum GIS.

**Regression Models**

“Benchmarking the energy efficiency of commercial buildings” – William Chung, Y.V.Hiu, Y.Miu Lam

This study is looking at a benchmarking process by using multiple regression analysis. The energy-efficiency benchmarking is useful to monitor changes in energy efficiency. The authors selected backward elimination procedure and following this they determined a final regression model to be used for benchmarking. Four examples of transformation of primary indicators are also provided. The study realized an energy efficiency model using liner regression and the take away for me was the model selection. The shortcoming is that with a complex model, more data is required to make accurate predictions.

“Testing goodness of fit in regression: a general approach for specified alternatives”- Also Solari, Sakia le Cessie, and Jelle J. Goemans

The point of researching this study was to find out more about how to fit a model in regression and when a model can fail. I wanted to know what the cases are when models do not fit or how the authors are calling it “different types of lack of fit”. Although a difficult read, the take-away was that many existing test can be constructed as special cases to test lack of fit. One of the samples was related to detecting nonlinearity and a semiparametric alternative model was used. An interested approach was about the test of the model used for testing: if this test indicates that the model does not fit, than it can be helpful to suggest what should be modified in the model in order to fit. The study also mentions an R package “globaltest” that has specific functions for testing against several types of lack of fit.

**Decision Trees**

“Induction of Decision Trees” – J.R Quinlan

Although the author is focusing on specific ID3 knowledge-based system on inductive inference from examples, my focus was on study findings on decision trees classifications and training set potential issues. When talking about “noise”, this study explains that we are assuming that the information under the training set is accurate when in reality it may include attributes that might have errors due to measurements or judgements, what can result in misclassification. This is something to consider in order of selecting the best attribute-based test to form the root of the decision tree.

“Classification and regression trees” – Wei-Yin Loh

The author compares different classification trees methods and I was researching this study in order to find other than the Random Forest methods. On CART is explained how the method is to first grow an overly large tree and then prune it to a smaller size to minimize an estimate of the misclassification error. I found the table on “Comparison of Classification Tree Methods” as very useful to present per various features, methods like: C4.5, CART, CRUISE, GUIDE, and QUEST.

“Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms” – Halil Ibrahim Erdal, Onur Karakurt

This study on water resources planning and management in Turkey highlights the following ensemble techniques: bagging, boosting, and stacking that are used to deal with problems of having weak predictors. Even if this study refers primordially to water, streamflow, and hourly flood forecasting, all not related to my study, it has the benefit of explaining in detail bagging, gradient boosting and the differences between classification and regression trees (CART) and support vector regression (SVR). What I found as very useful is comparing CART with SVR, concluding that the results of CART in testing are not the best and enhancing CART with bagged regression trees, stochastic gradient boosted regression trees, to improve the results of the forecasting.

**Dataset**

The dataset analyzed is a public dataset, created by Angeliki Xifara & Athanasios Tsanas and available under: <https://archive.ics.uci.edu/ml/datasets/Energy+efficiency>.

In addition to the dataset the two authors made public in 2012 a study on this dataset called “Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools” (see reference for link). Athough this paper was a strong information source, this study will use the dataset for a different approach and analysis and occasionaly will reference to conclusion from this paper.

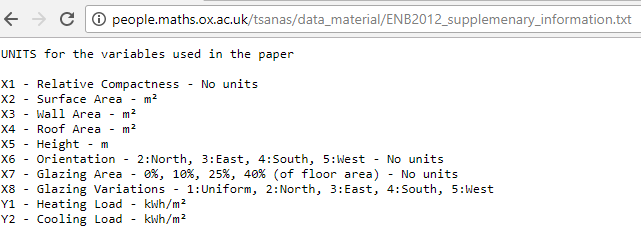
The data is a simulation of 768 residential seven person buildings, assumed to be in Athens, Greece and has eight input variables (surface area, wall area, roof area, orientation, glazing area…) and two output variables (heating and cooling load).

The scope of creating this simulated data was to support determining the heating and cooling load of energy efficient buildings based on the building parameters (input variables).

The dataset was created in 2012 using the software Ecotect. There were 12 building forms created, all having same volume of 771.75 m3, but having different surface areas and dimensions. The materials used for the building forms are the same.

The authors of the dataset are providing additional information of design conditions assumptions like humidity, air speed, lighting level, infiltration rate… that will be excluded from this analysis since are not input variables used in the dataset.

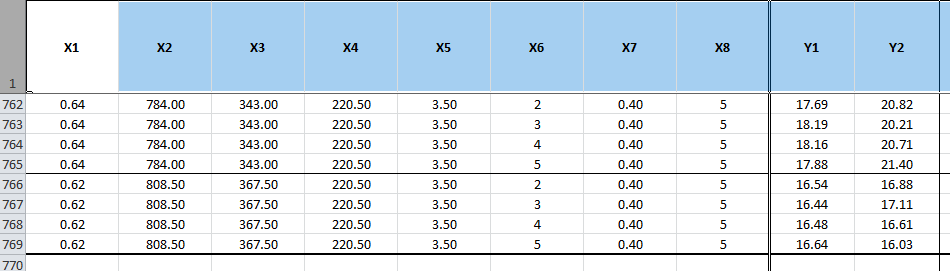
What this study will focus on is the variables included in the dataset and for this will include the below screen-shot from the supplementary dataset information:



It’s important to understand in detail how the 768 variation of buildings have been created and this is why the dataset author’s own words are quoted: “We used three types of glazing areas, which are expressed as percentages of the floor area: 10%, 25 %, and 40%. Furthermore, five different distribution scenarios for each glazing area were simulated: 1) uniform: with 25 % glazing on each side, 2) north: 55% on the north side and 15% on each of the other sides, 3) east: 55 % on the east side and 15 % on each of the other sides, 4) south: 55% on the south side and 15% on each of the other sides, and , 5) west: 55% on the south side and 15% on each of the other sides. In addition, we obtained samples with no glazing area. Finally, all shapes were rotated to face the four cardinal points. Thus, considering twelve building forms and three glazing area variations with five glazing area distribution each, for four orientations, we obtained 12 x 3 x 5 x 4 = 720 building samples. In addition, we considered twelve buildings forms for the orientation without glazing. Therefore, in total we studied 12 x 3 x 5x 4 + 12 \* 4= 768 buildings. Each of the 768 simulated buildings ca be characterized by eight building parameters (to conform to standard mathematical notation and facilitate the analysis in the work, henceforth these buildings parameter will be called input variables and will be represented with X) which we are interested in exploring further. Also, for each of the 768 buildings we recorded the Heating and Cooling Load (henceforth these parameters will be called output variables and will be represented with Y)”.

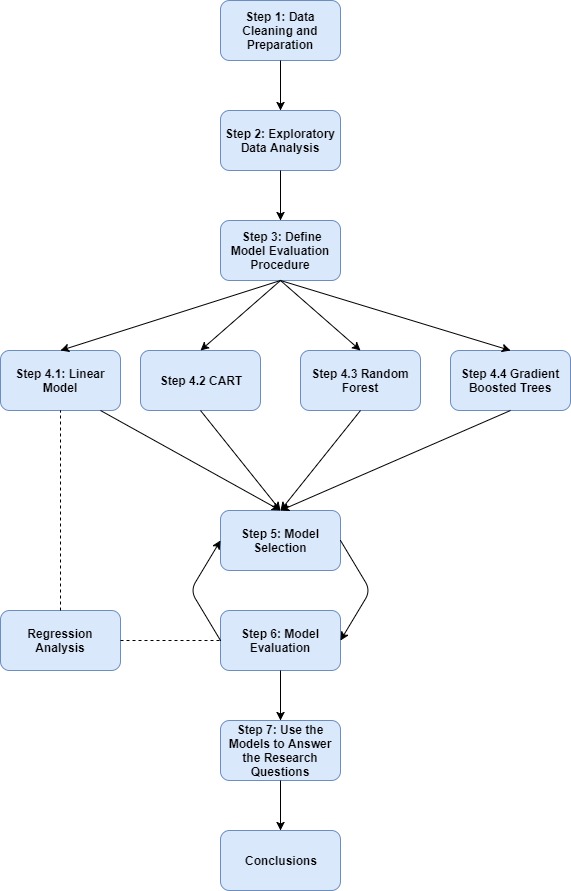
Descripting statistics of the attributes will be provided in the exploratory analysis section and to conclude the dataset description let’s summarize the following points: the dataset is public with data available in Excel format; is based on simulated data; has 768 observations, has eight input variables (X1…X8) and two output variables (Y1 and Y2).

The following screen shot sample from the dataset helps understanding how the raw data looks like:



**Approach**

The following Figure 1 block diagram shows the steps to analyze the data.



This block diagram includes the following steps that will be followed during the analysis:

* Step 1: Data Cleaning and Preparation
* Step 2: Exploratory Data Analysis
* Step 3: Define Model Evaluation Procedure
* Step 4: Models
  + Step 4.1: Linear Model (Regression Analysis)
  + Step 4.2: CART
  + Step 4.3:Random Forest
  + Step 4.4: Gradient Boosted Trees
* Step 5: Model Selection
* Step 6: Model Evaluation
* Step 7: Use the Model to Answer the Research Questions
* Step 8: Conclusions

# Step 1: Data Cleaning and Preparation

### Loading the data, review structure, and missing data

Loading the data: after examining the original Excel file with the dataset, it can be noticed that only the first sheet has data and only this sheet will be read from the file. The other two empty sheets (with greek names) will be ignored

The structure of “raw\_data” is reviewed to make sure that all 768 observations are loaded

str(raw\_data)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 768 obs. of 10 variables:  
## $ X1: num 0.98 0.98 0.98 0.98 0.9 0.9 0.9 0.9 0.86 0.86 ...  
## $ X2: num 514 514 514 514 564 ...  
## $ X3: num 294 294 294 294 318 ...  
## $ X4: num 110 110 110 110 122 ...  
## $ X5: num 7 7 7 7 7 7 7 7 7 7 ...  
## $ X6: num 2 3 4 5 2 3 4 5 2 3 ...  
## $ X7: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ X8: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Y1: num 15.6 15.6 15.6 15.6 20.8 ...  
## $ Y2: num 21.3 21.3 21.3 21.3 28.3 ...

Because readxl loaded all variables as numeric there cannot be blank spaces.

Review for missing data: the check for NA's is being done for the whole dataset and also as an alternative an NA check per columns (using “sapply”) is being reviewed. As expected there was found none NA.

### Change column names to real variable names

The column names have been changed to real variable names. The structure of “data” is being reviewed and the “raw\_data” will be kept in the workspace in case it's needed later in the analysis

data <- raw\_data  
names(data) <- c("Relative Compactness", "Surface Area", "Wall Area", "Roof Area", "Overall Height",  
 "Orientation", "Glazing Area", "Glazing Area Distribution", "Heating Load", "Cooling Load")  
  
str(data)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 768 obs. of 10 variables:  
## $ Relative Compactness : num 0.98 0.98 0.98 0.98 0.9 0.9 0.9 0.9 0.86 0.86 ...  
## $ Surface Area : num 514 514 514 514 564 ...  
## $ Wall Area : num 294 294 294 294 318 ...  
## $ Roof Area : num 110 110 110 110 122 ...  
## $ Overall Height : num 7 7 7 7 7 7 7 7 7 7 ...  
## $ Orientation : num 2 3 4 5 2 3 4 5 2 3 ...  
## $ Glazing Area : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Glazing Area Distribution: num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Heating Load : num 15.6 15.6 15.6 15.6 20.8 ...  
## $ Cooling Load : num 21.3 21.3 21.3 21.3 28.3 ...

Exclude from analysis 48 tuples where “Glazing Area Distribution” = “0”.

The original dataset has initially for "Glazing Area Distribution" six different values: "0" = No Glazing Area (no windows); 1 = Uniform (25 % windows on each side), 2 = North, 3 = East, 4 = South, 5 = West

Values like "0" (No Glazing Area) might never apply to residential buildings and could hinder in answering the research questions by influencing the model fit, since there is not possibility that a house has no windows. This is why it was decided to remove those "0" tuples from the Glazing Area Distribution.

By removing the 48 tuples with "0" Glazing Area Distribution, also all the "0" in "Glazing Area are being removed the same time.

Now, after removal, the Glazing Area Distribution has only 5 values and "0" has been removed.

The structure after the change displays the 720 remaining values.

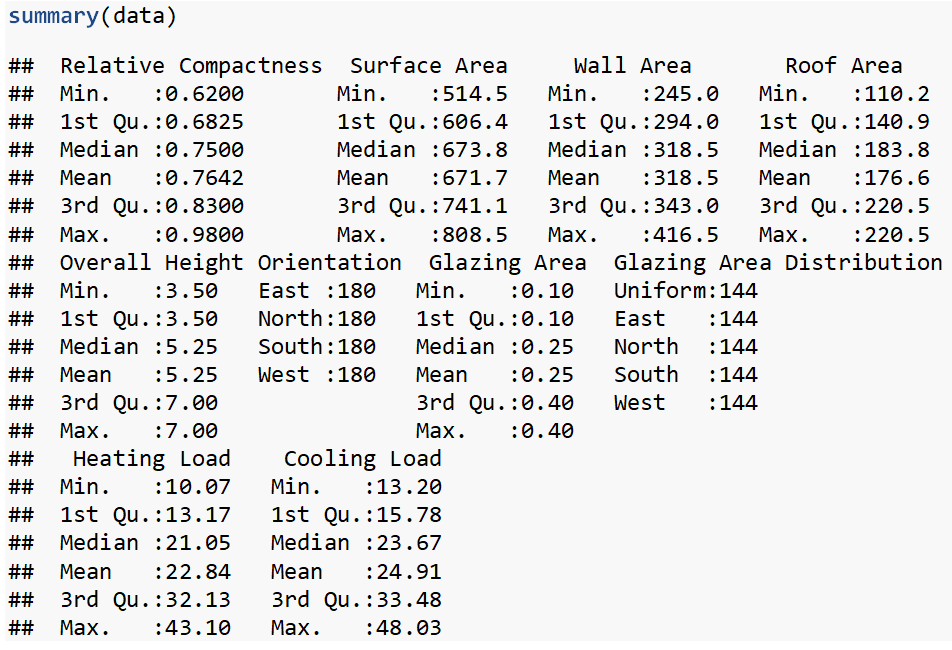
### 

### Converting categorical variables to factors

As per the dataset description, it was established that there are two categorical variables: "Orientation" and "Glazing Area Distribution". The variable "Orientation" has only four values, the cardinal directions: N,W,S,E. The variable "Glazing Area Distribution" has all the above four and in addition "1" for "Uniform".

Using the dataset dictionary, the integers from the raw data have been mapped to categorical values.

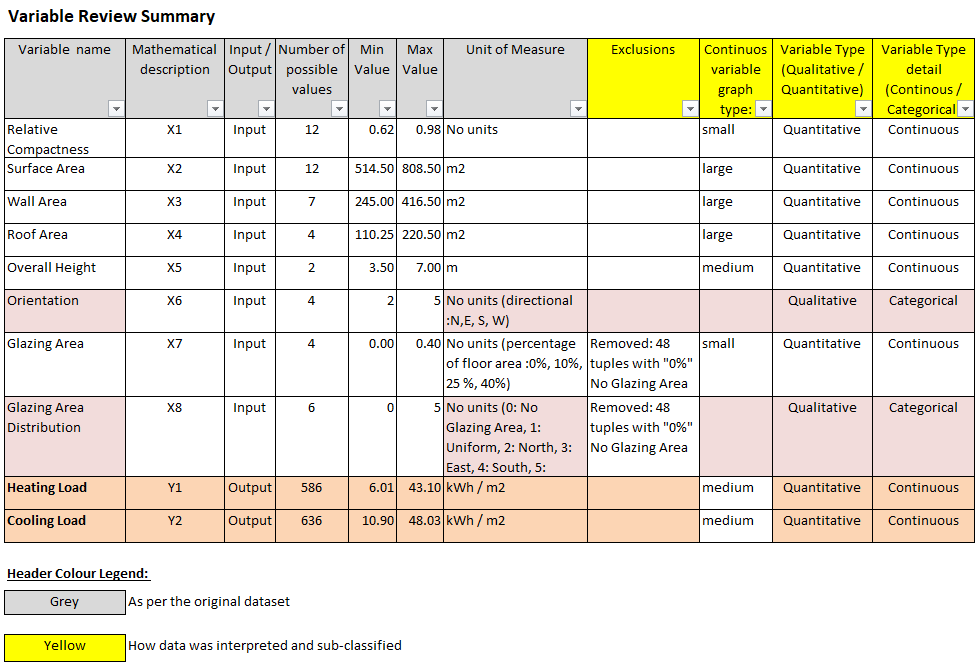
### Review summary for "data"



From the summary, can be concluded that Orientation and Glazing Area Distribution, are qualitative-categorical variables, the other variables are quantitative-continuous. Those conclusions and the Min and Max will be used in an Excel table (Fig. 2) that explains the dataset in more detail.

Using the count of the unique values per each variable, the following Excel table was created to provide a better understanding of the variables included in the dataset.

Fig.2

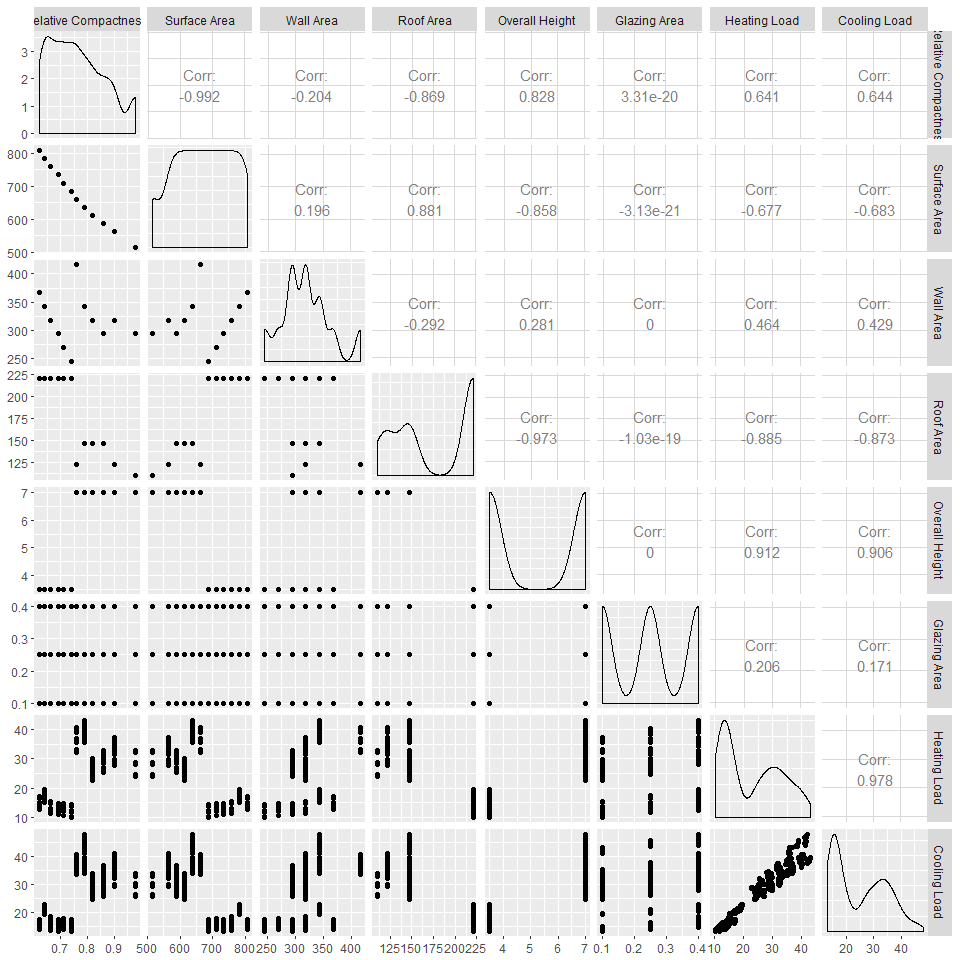


# Step 2: Exploratory Data Analysis

The categorical variables are split from the continuous variables, into a new dataframe called "data\_categorical" and the continuous variables are kept in "data\_numeric". This is done to make it clear which methods apply to the continuous variables (correlations) and which apply to the categorical variables (counting with the table function).

Next the analysis visually checked for correlated variables in the numerical dataset using an optimal display of plot and selected the package "GGally" where the function ggpairs is being used for creating scatterplots between all pairs of continuous variables from "data\_numeric".

ggpairs(data\_numeric)



The following are the interpretations of the major correlations:

Relative Compacteness & Surface Area / strong almost perfect negative correlation at -0.992

Relative Compacteness & Wall Area / low negative correlation

Relative Compacteness & Roof Area / strong negative correlation

Surface Area & Overall Height / strong negative correlation

Surface Area & Roof Area / strong positive correlation

Relative Compactness & Glazing Area / "0" correlation

Relative Compactness & Heating Load /medium strong positive correlation

Wall Area & Heating Load / low positive correlation

Relative Compacteness & Cooling Load / positive high correlation

Surface Area & Cooling Load / negative high correlation

Wall Area & Cooling Load / low positive correlation

Roof Area & Cooling Load / strong negative correlation

Overall Height & Cooling Load / strong positive correlation

Glazing Area & Cooling Load / low positive correlation

Heating Load & Cooling Load / strong almost perfect positive correlation

### Outlier detection

Review if there are outliers by using boxplots. Going forward this analysis will try to use only ggplot for all graphs to create a consistent view.

Looking at the Excel table "Variable Review Summary" that has the Min and Max, it was decided to split the boxplots of data-numeric into three categories by cluster of data: small (values 0-1); medium (values >1-50); large (>50)

#### Boxplots are created for the variables that are grouped by value (see Excel Table Fig2) into small, medium and large.

A couple of features have been used to make the boxplots better visible on paper:

* The boxplots were plotted using ggplot2;
* In the geom\_boxplot, the inside was filled "green";
* For the median, it was used "lwd" = 1.2, slightly more than normal, ("lwd" stands for line width) to display a thicker line and this way to be better visible;
* Used geom\_jitter to add a small amount of variation and make the dots more visible by changing: the shape to be "diamond" = shape # 18, colour "blue" to show better on green, and increase slightly the width of the blue diamond dots

The numeric data will be split in three data frames because the variables have different scales.

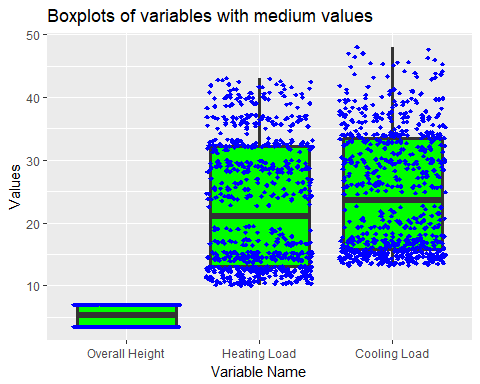
"Melt"" from the reshape2 package was used because the boxplots have the variable name on the x-axis and the long data format (one column with the variable name and another one with the value). "Melt" can be easily used by ggplot2 to create the side-by-side boxplots.

To show a sample of the boxplots, in the following only one of the three boxplots created will be displayed, the “Boxplot of variables with medium values”. The other two boxplots, can be reviewed in the RMarkdown document attached to this study.

The boxplots of variables with “small” values (Relative Compactness, Glazing Area) are showing that we have no outliers and that there are several data points that have the same values what is the result of the simulated values in the dataset. "Relative Compacteness" takes 12 values only and "Glazing Area" only four values.

The boxplots of variables with “medium” values (Overall Height, Heating Load, and Cooling Load) are showing that we have no outliers and that there are several data points clustered in the bottom part of the boxplot of the Heating Load and Cooling Load (between 10 and 15). For the Overall Height, this has only two values 3, 5 and 7. For Heating Load we see that the data is mostly clustered in the first quartile of the boxplot. For Cooling Load we see more variability in the data but most data is grouped together in the first quartile.

Below is a the boxplot based on the “medium” variables (sample graph since “small” and “large” are not displayed in this paper but are available in the RMarkdown attachment)



The boxplots of variables with high values (Surface Area, Roof Area, and Wall Area) are showing that there are no outliers and that Surface Area has only 12 possible values, Wall Area has only 7 values, and Roof Area has only 4 values.

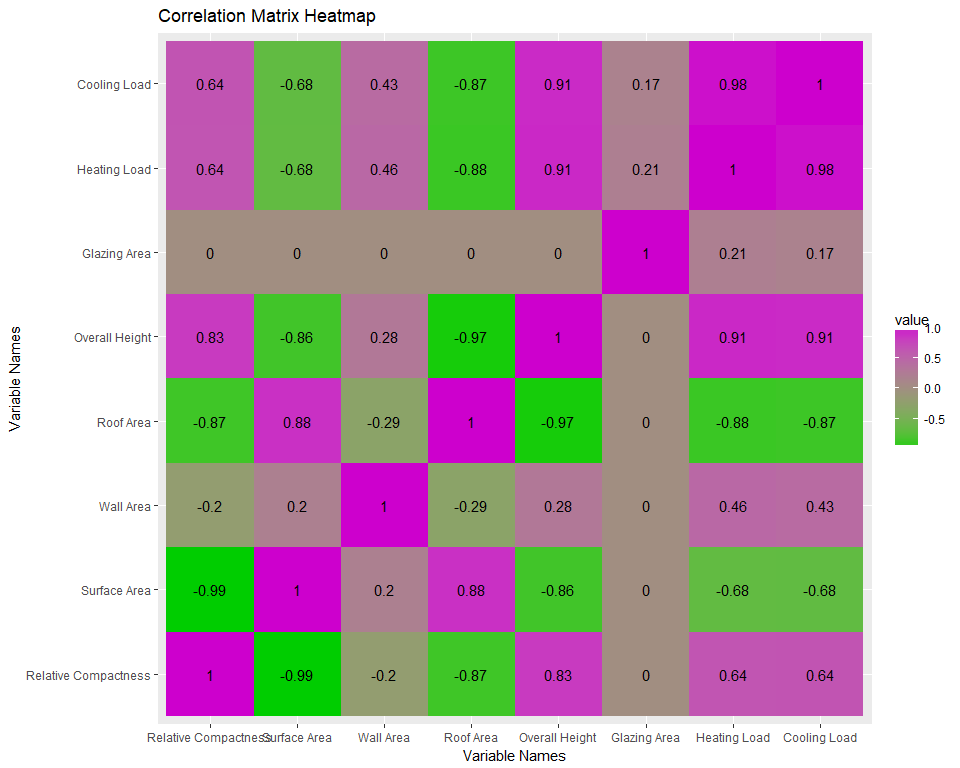
### Creating a Correlation Matrix Heatmap

A Correlation Matrix Heatmap was created (using “ggplot”) because colours are a better way to quickly visualize any large or small correlations.

A couple of features have been used to make the Correlation Matrix Heatmap a better visualization:

* The cor method "Pearson"
* geom\_text to write the cor values as text in the square
* Selected 2 colours "green3" and "magenta 3" From: <http://sape.inf.usi.ch/quick-reference/ggplot2/colour>
* A scale gradient of colours with values between -1 and 1
* The green colour is showing the negative correlation and the magenta colour the positive correlation
* The geom\_tile is displaying the rectangles
* Set the figure height at 8 and width at 10 to better show up when printed.

Having the below Correlation Matrix Heatmap it is easy to visualize the lighter and darker colours that stand for low and high correlation.



### Investigating the Categorical Variables

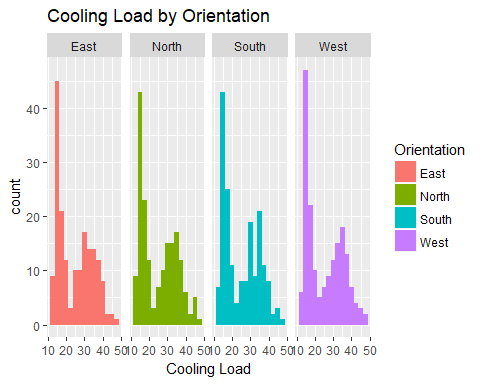
The two categorical variables are "Orientation” and “Glazing Area Distribution”. "Orientation" has only the four values of the cardinal directions: East, North, South, and West, each with 180 observations

The tables are displayed for both categorical variable "Orientation" and "Glazing Are Distribution" to show that the "Glazing Area Distribution" has the four values of the cardinal directions and includes also "Uniform". Each “Glazing Area Distribution” value appears 144 times.

Next, facet\_grid will be used to plot the histograms of the Cooling Load for each value of the categorical variables.

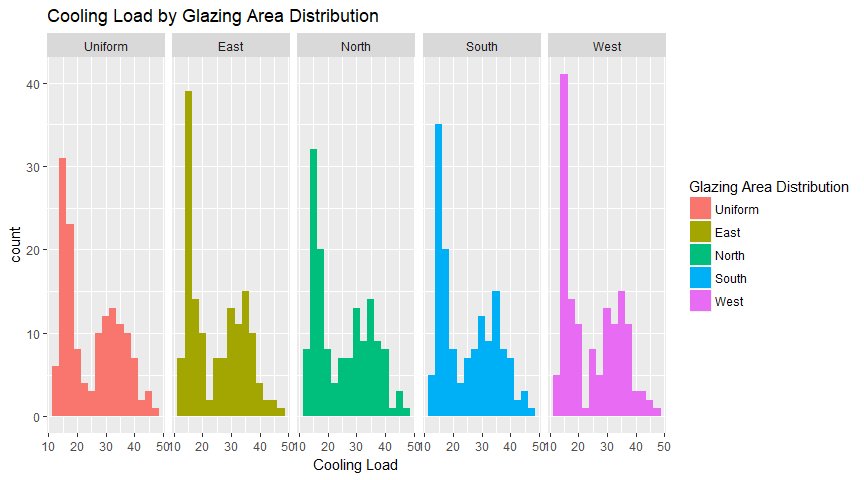
The two graphs of the “Cooling Load by Orientation” and “Cooling Load by Glazing Area Distribution” are being created (using ggplot, geom\_histogram).

The plot of the categorical variable "Orientation" against the predicted variable "Cooling Load" is displayed below



For all orientations the Cooling Load distribution is roughly bimodal, with one peak around 15 and the other around 35.

Similar results will be obtained for Glazing Area Distribution.



The initial intent was to use One-way ANOVA to check if these categorical variables would have an impact on the predicted variable, but because the distributions within the groups are clearly not normal, this can’t be used.

# 

# Step 3 Define Model Evaluation Procedure

### Methodology

Regression will be the main focus of this study because the dataset contains one numeric variable (Cooling Load) that needs to be predicted, and a set of independent variables which can be used to make that prediction.

In general, the objectives of regressions are:

* To review if there is a relationship between two variables and especially if there is a statistically significant relationship between those two
* To predict new observations.

The objectives of this study are to determine which predictors influence the response variable and the nature of the relationships between the predictors and the response

The models that will be used are:

* Random forest with all the variables because it is a very powerful model capable of capturing non-linear relationships between the predictors and the response variable and this was the best performing model in the original study
* CART because it is an interpretable model which could help in answering some of the research questions and because the scatterplots in the previous step had some groups which could be handled by the splits
* Multiple linear regression because it can represent very simple linear relationships between the predictors and the response variable and if it performs well, it could answer more questions than the other two models

The models will use the following variables:

* Random forest: all variables
* CART: all variables
* Linear model: multiple models created with backward elimination, starting from the model with all variables and eliminating the variables one by one

For all the models, prediction and residual analyses will be performed, but this is especially important in the case of the linear model, which makes the assumption that the errors are normally distributed.

To avoid overfitting, all the measures will be computed on different datasets than the ones that were used to train the models. After the best performing models are selected, they will be trained on the complete dataset.

The most important criterion for choosing the best model will be R2. This study will not use ANOVA to compare the stepwise models, instead it will use information criteria (AIC and/or BIC) at the end of the procedure to select the model with the fewest variables that still has a good R2. ANOVA will not be used to stop eliminating variables, the backward elimination procedure will be stopped when R2 becomes too small. For non-linear models like CART and Random Forest, k-fold cross-validation (k = 10) will be used to pick the best performing model according to R2.

Going forward MSE and MAE will be used as evaluation measures.

The interpretability of the models and how can they help us to answer the original research questions, will also be reviewed.

From the two output variables the prediction of the "Cooling Load" only will be analyzed (that was marked in the original study with "y2"), and this is why the "Heating Load" will be excluded.

The set.seed will be used to ensure the reproducibility of the results. The data is shuffled because it is generated and the order in which it was generated would create biased folds in cross-validation.

The analysis will start with random forest and 10-fold cross-validation. In other models, a classical train-test split (70-30) will be used.

# Step 4 Models

The regression variant of Random Forest and CART models will be used to predict the cooling load because this is a numeric output variable.

The analysis will start with Random Forest with 10-fold cross-validation.

Step 5 "Model Selection" and step 6 "Model Evaluation" will be done individually for each model.

# Step 4.3 Random Forest

Random Forest is defined as per Wikipedia as "a way of averaging multiple decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally boost the performance of the final model".

### Random Forest for Regression

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set with responses bagging repeatedly (B times) selects a "random sample with replacement" of the training set and fits trees to these samples:

For b = 1,...,B there are samples with replacements taken and "n" training examples from X, Y, and call these , .

In addition to bagging, to further decorrelate the trees there is an additional technique used by random forest, specifically only using a limited subset of variables when trying to make a split. Usually these subsets are of a fixed size and that size is chosen before training and used as a hyperparameter called "mtry".

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x'.

Here "X" = the entire dataset of numerical and categorical variables.

X = Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, Glazing Area Distribution

Y = Cooling Load

Prediction formulas on an unseen data point x':

It can be clearly seen from the formula that the prediction is a simple average on the predictions from all the trees.

### Training with cross-validation

The cross-validation results will be used to pick the best performing model. A simple grid search will be used for the hyperparameter selection. In the grid values have been selected such that the best performing model is not at the extremes. The intervals of the hyperparameters (mtry, nodesize, ntree, and nodedepth) were picked after several trial runs.

The following will explain the grid search method. The optimization of the model hyperparmeters can be done using either a "randomized search" or a "grid search". In this study the method of full grid search over all parameters was selected in order to improve the quality of the results. Other factors that could have been investigated for improvements of results could have been: better feature engineering, using different regularization methods for regression or different regression methods.

Grid search is done by picking a set of values for each hyperparameter, generating all possible combinations, then training and evaluating all the models and picking the one which scored best according to the evaluation criteria.

#### Defining the evaluation function and the hyperparameter grid

For random forest the hyperparameter grid will be generated based on the following values:

* mtry(number of possible choices at each split): 2, 3, 4
* nodesize(average number of data points in a leaf): 5, 10, 15
* ntree(number of trees): 50, 100, 150
* nodedepth(maximum depth): 2, 3, 4, 5, 6, 7

For evaluation, several measures will be computed (R2, MSE, MAE) and extract the mean over all folds. A total of 162 different random forest models will be fit, but only display the best 10 and the worst 5, sorted by R2 in descending order(after training). There are 162 models because it is the cardinality of the cartesian product between all the hyperparameters (3 from mtry x 3 from nodesize x 3 from ntree x 6 from nodedepth = 162 models).

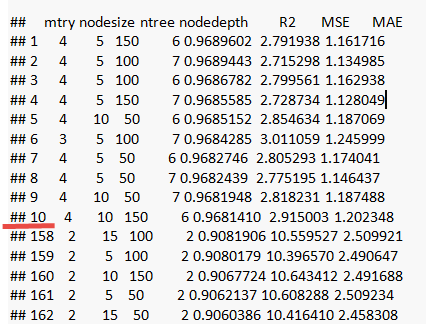
str(RFparams)

## 'data.frame': 162 obs. of 4 variables:  
## $ mtry : int 2 3 4 2 3 4 2 3 4 2 ...  
## $ nodesize : num 5 5 5 10 10 10 15 15 15 5 ...  
## $ ntree : num 50 50 50 50 50 50 50 50 50 100 ...  
## $ nodedepth: int 2 2 2 2 2 2 2 2 2 2 ...

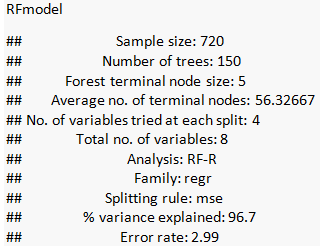
#### Model Training

For each model, the cross-validation results are computed as the mean over all folds, and then arranged in descending order by R2.

Results are presented below in Fig .3 with top 10 and bottom 5:

Fig 3 

Next, will fit the model that performed best in cross-validation on the whole dataset because the models were trained on the cross-validation folds. Next, the model will also be displayed in Fig 4.

Fig 4 

The model comparison will be done after all the models are trained.

# Step 4.2 CART

CART (that stands for Classification and Regression Trees) is a recursive partitioning method to build regression trees to predict the value of a continuous variable from one or more predictor variables that can be continuous and/or categorical.

The CART model will be used since it is a more interpretable model compared to random forest.

When creating the CART process of computing the regression trees, these four steps should be followed:

* Specify the criteria for predictive accuracy
* Select the splits
* Determining when to stop splitting
* Select the right-size tree

For CART a train-test split will be used, but would still use the grid search method for hyperparameter selection. In the grid values have been picked such that the best performing model is not at the extremes. The intervals were picked after several trial runs.

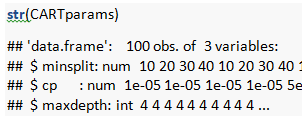
One of the reasons why the CART model was used here, originated in the fact that it was also used in the following study included in the "Literature Review" section: "Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms" - by Halil Ibrahim Erdal, Onur Karakurt.

#### Defining the evaluation function and the hyperparameter grid for CART

The hyperparameter grid will be generated based on the following values:

* minsplit(minimum number of points in node before split): 10, 20, 30, 40
* cp(cost complexity factor): 0.00001, 0.0005, 0.0001, 0.005, 0.001
* maxdepth(maximum tree depth): 4, 5, 6, 7, 8

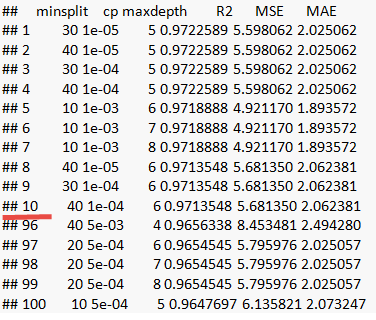
100 CART models will be evaluated on the training set. See Fig 5

Fig 5. 

The CART model performance will be displayed similar to how it was presented for Random Forest.

#### CART Model training

Results are presented below in Fig .6 with top 10 and bottom 5:

Fig 6. 

Now the CART model will be fit on the whole dataset because the previous models were fit on the training set.

The prediction analysis and model comparison will be done after training the linear model.

# Step 4.1 Linear Model

The linear model assumes that the dependent variable can be accurately predicted using a linear combination of the dependent variables by using the coefficients vector . is the error term, which represents what cannot be learned by the model.

To make a prediction for an instance of the variable Y, if we have the outcomes for the variables and we know the coefficients , we only need to compute the linear combination:

The R's lm function will be used to estimate the coefficient vector .

### Steps 5 and 6 - Model Selection and Evaluation for the Linear Model

The most important measure used to evaluate the model is the R2, but this study will also use Akaike Information Criterion (AIC), a measure that penalizes the objective function for linear regression by the number of variables included in the model. Penalizing complex models is used to avoid overfitting.

AIC is an estimator of the relative quality of a statistical model for a given data set. When given several models, AIC estimates the quality of each model, relative to each of the other models. AIC is used to make "apples to apples" comparison of different models over the same dataset.

In AIC, the model with the lowest AIC score will have the smallest divergence and will be the better, preferred model.

First the evaluation method will be defined, which includes the model AIC, along with the same train-test (70/30) split that was used to evaluate the CART model.

**Explaining the usage of backward elimination**

The model selection methods which will be used is backward elimination, starting from the complete model with all variables and eliminating variables one by one. Some variables might have multiple coefficients in the model and they will be removed together.

"Backward" is one four most used variable selection methods: forward, backwards, stepwise, and all subsets. Selecting the "backward" method in this was a result of a study analyzed in the "Literature Review" section, called "Benchmarking the energy efficiency of commercial buildings" by William Chung, Y.V. Hui, and Y. Miu Lam and where the "backward" elimination method is referred to as: "applied to select the regression models where insignificant explanatory variables are eliminated.

### Training the model using backward elimination

**Model "completeLinearModel" with all variables**

Next a linear model will be trained on all the variables in the train dataset.

The results of the evaluation will be loaded into a new blank LMeval dataframe where all the models will be analysed

The Roof Area variable will be removed because its coefficient is NA, which means it are linearly dependent to the other variables in the dataset.

**Model "model7Var" with all variables except Roof Area**

Based on the p-values and for the Glazing Area Distribution no coefficients seem relevant, so will be excluded the first because the p-values for all the coefficients are large.

**Model "model6Var" with all variables except: Roof Area, Glazing Area Distribution**

Now the coefficient p-values are analyzed for Orientation.

Even though OrientationWest seems a little relevant, it will be excluded from the next model and the evaluation table will be used at the end to see if it was the right choice.

**Model "model5Var" with all variables except: Roof Area, Glazing Area Distribution, and Orientation**

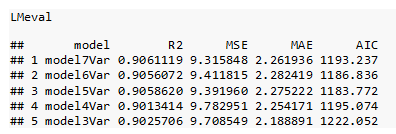
Next, Surface Area will be removed from the model, because its coefficient has the largest p-value.

**Model "model4Var" with all variables except: Roof Area, Glazing Area Distribution, Orientation, and Surface Area**

The next variable to be removed is Wall Area, because its coefficient estimate has the highest p-value.

**Model "model3Var" with only: Wall Area, Overall Height, and Glazing Area (here Relative Compactness was removed)**

This will be the last linear model included in the evaluation because the R2 is already becoming significantly worse compared to the other models. In Fig 7 are the complete results of the linear model:

Fig 7 

To summarize, the variables included in each linear model are:

* model7Var:Glazing Area, Glazing Area Distribution, Orientation, Overall Height, Relative Compactness, Surface Area, Wall Area
* model6Var:Glazing Area, Orientation, Overall Height, Relative Compactness, Surface Area, Wall Area
* model5Var:Glazing Area, Overall Height, Relative Compactness, Surface Area, Wall Area
* model4Var:Glazing Area, Overall Height, Relative Compactness, Wall Area
* model3Var:Glazing Area, Overall Height, Wall Area

The names of the models are based on how many variables they use.

Based on the above results the best selected model is "model5Var" (without Roof Area, Orientation and Glazing Area Distribution) because it has the second best R2 and MSE, but a lower number of variables. This is also indicated by AIC, which has the lowest value for this model.

Now model5Var will be fit on the whole dataset and analyzed further in the next step.

### Step 4.4 Gradient Boosted Trees

In general, boosting is used for weak learners and our decision tree was already a very good fit on the data with a R2 of over 0.95 in Random Forest and CART and this is why this model of gradient boosted trees will not be pursued this further.

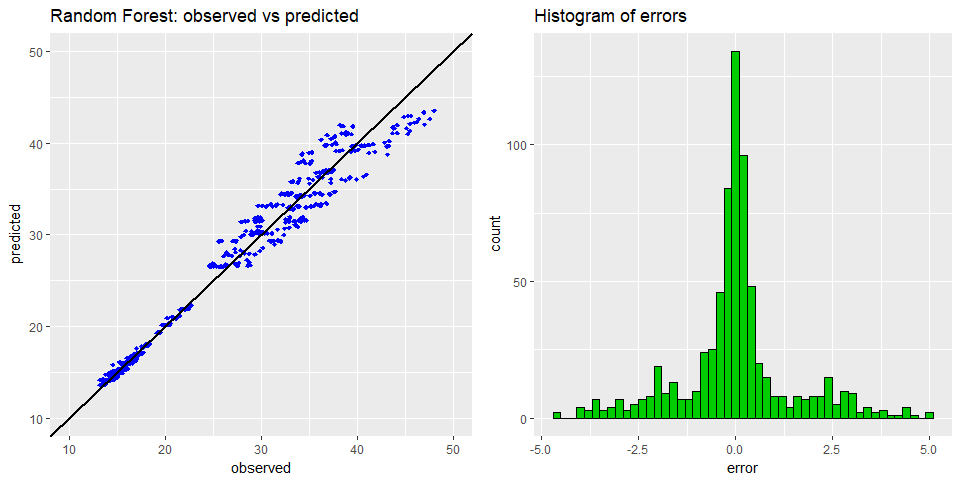
# Step 6: Model Evaluation / Model Comparison

### A more in-depth look

Although the CART MSE is almost twice the random forest MSE, the CART R2 is surprisingly larger, so this is why a more in-depth look at the predictions is needed.

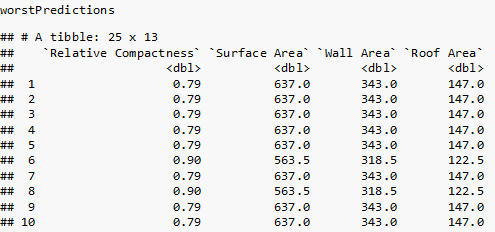
Further the errors will be ranked and the next step will be to determine the conditions that lead to large errors in the predictions.

#### Random Forest Prediction Analysis



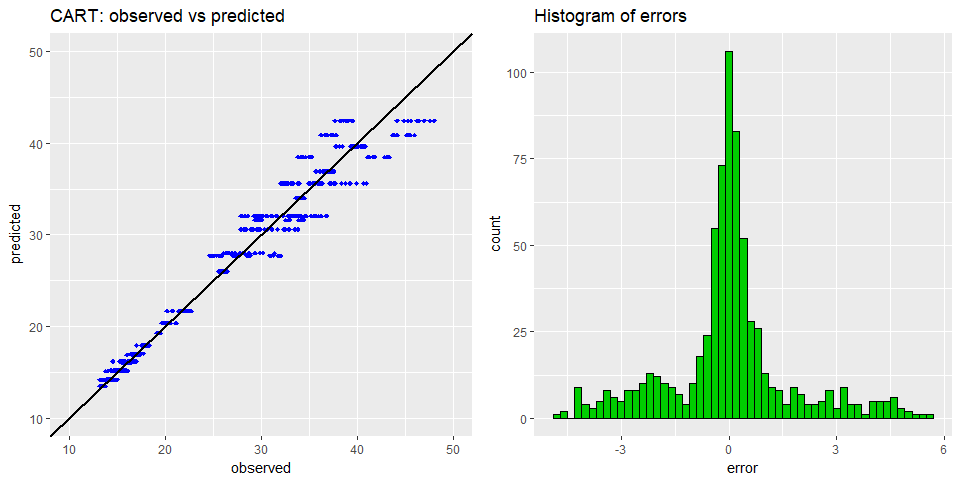
The error distribution is centered at 0 and it seems to be unimodal.

Now the worst predictions will be examined in order to determine if there is a pattern in the data Fig 8

Fig 8 

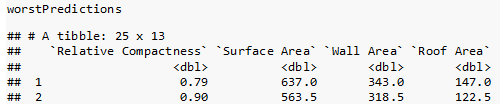
It can be seen that most of the worst predictions are made for one type of building. Because there are very few data points it is difficult to generalize this result and find all the building types for which the model might not work.

#### CART Prediction Analysis



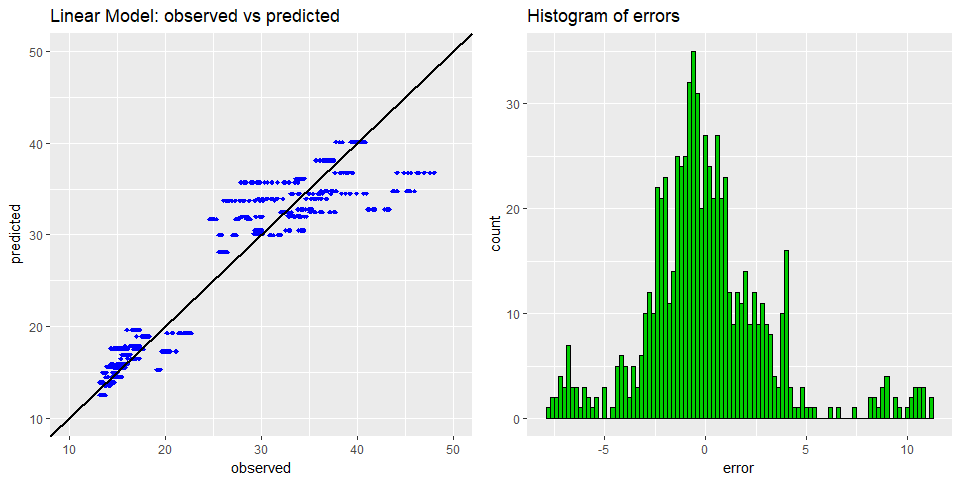
The error distribution is centered at 0 and it seems to be unimodal, although there seem to be more negative errors than positive ones.

Now the worst predictions will be examined in order to determine if there is a pattern in the data points that the model gets wrong. We will only look at the top 25 errors in Fig 9.

Fig 9. 

We can see that most of the worst predictions are made for one type of building. Because there are very few data points it is difficult to generalize this result and find all the building types for which the model might not work.

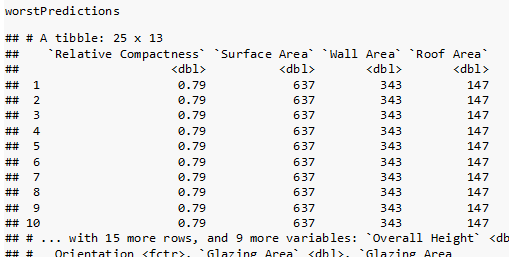
#### Linear Model Prediction Analysis



The assumption that the linear model makes is that the errors are independent and normally distributed. By doing a histogram of the residuals and a QQ plot the feasibility of this assumption will be investigated.

The residual distribution is not quite centered at 0, more like at -1, there are some data points with large errors that biased the whole model.

Now the worst predictions will be examined in order to determine if there is a pattern in the data points that the model gets wrong. Only the top 25 errors will be analyzed in Fig 10

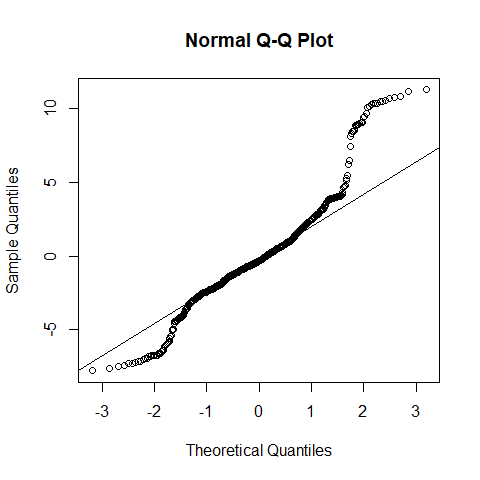
Fig 10. 

It is the same building like for the other models, but the errors are even bigger and more domain knowledge is necessary to decide if removing that building from the training set makes sense.

#### Residual Analysis for the Linear Model

The residuals should form a normal distribution centered at 0, but the plot shows that this is not the case.

qqnorm(predictionData$error)  
qqline(predictionData$error)

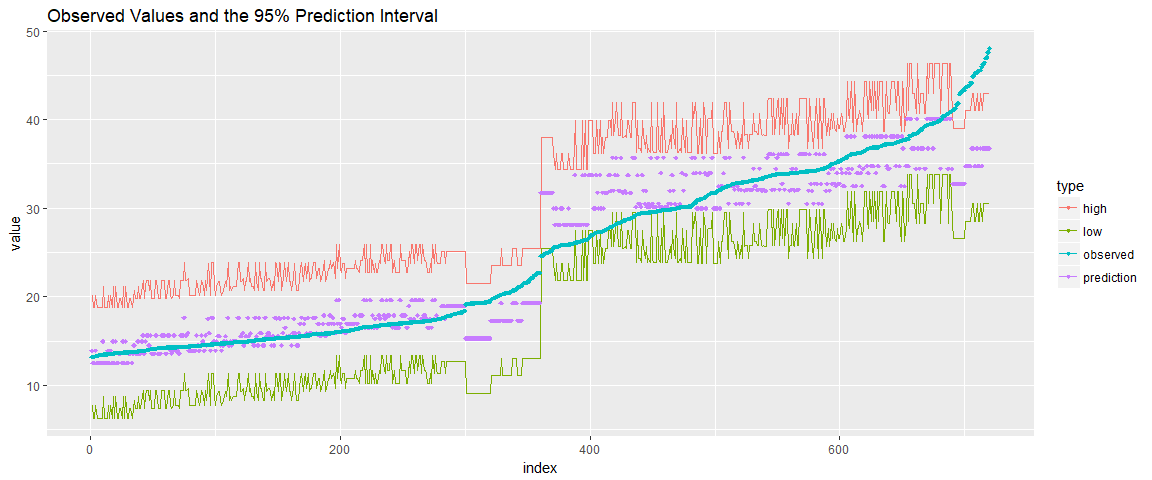


The QQ plot shows that the distribution of the residuals is quite far from normal. Theoretical Quantiles are from the Normal distribution and Sample Quantiles are from the residuals. We can see from the plot that there are more values at the extremes than in a normal distribution with the same parameters.

#### Prediction interval (95%) for the Linear Model

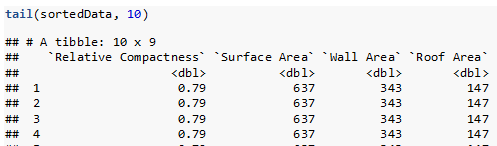
For both models Random Forest and CART this study selected not to investigate confidence intervals because there is no standard theoretically sound way to compute them like there is for linear models.

First the data is sorted by the observed values(Cooling Load), then the 95% prediction intervals are computed, then everything is plotted to see in how many instances the actual values(blue dots) are outside the prediction interval(upper - red, lower - green). The x-axis contains the indices, and the y-axis is used for the values, which are between 13 and 48 for Cooling Load. See Fig 11 below

Fig 11 

The observed values are outside the interval for buildings with high cooling load values, which means that those types of buildings don't fit the pattern learned by the model. A solution would be to gather more data about buildings with high cooling load.

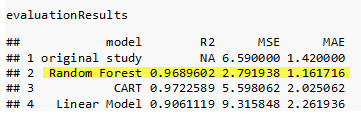
After a more in detail look, it can be observed that all the observations with a high cooling load belong to that building which all the models had trouble predicting, so this might be a data quality issue. Either that building is an outlier, or more data is needed for buildings with high cooling loads. See Fig 12



# Results

### Model Comparison

At this point we will compare the results of all the models that have been trained until now.



As it can be seen from the table, the random forest with our chosen parameters seems to outperform the model from the original study, however there are some other facts that contribute to this difference:

* 48 building with "No Glazing Area" have been removed; we experimented with models that included them and the errors on these buildings were the largest and our concern was that they biased the models and we couldn't answer the research questions effectively
* The "original study" repeated the cross-validation many times and the standard error for their MSE was 1.6, which puts us roughly at 2 standard errors from their average MSE
* A newer random forest package was used that didn't exist when the authors of the original study did their study in 2011.

### The best performing model

Both models: Random Forest and CART have very similar R-squared (Random Forest has 0.9689 and CART has 0.9722)

The best performing Model from the Initial Results seems to be Random Forest since it has a lower MSE at 2.79 compared to CART that has 5.59

In case of the best linear model, the R-squared is 0.906 what brings it on the last place.

### Step 7: Answering the Research Questions

**Can the cooling load be predicted accurately with a statistical model?**

As a result of the initial analysis, yes this can be done by using the Random Forest statistical model. The error is small relative to the values of the predicted variable and the R2 looks very good (over 0.968).

**Explore trade-offs between accuracy and interpretability of the model**

It can be seen that while CART is more interpretable, random forest gives a much lower error. The easy way to see how the prediction is calculated has to be given up for a more accurate prediction.

For the linear model, splitting variables into groups produces a more accurate prediction, at the expense of interpretability. Figuring out what happens when those variables change is harder to do when there are two coefficients associated with each variable. Also, the AIC recommended to pick a model with fewer variables instead of the model with all the variables and a slightly bigger R2.

Although from an engineering perspective, the most important variable should be the Glazing Area, in the random forest it is almost insignificant, while in CART and the linear model it is considered important. This means that these models might still be useful, despite the lower accuracy because they capture the real relationships better.

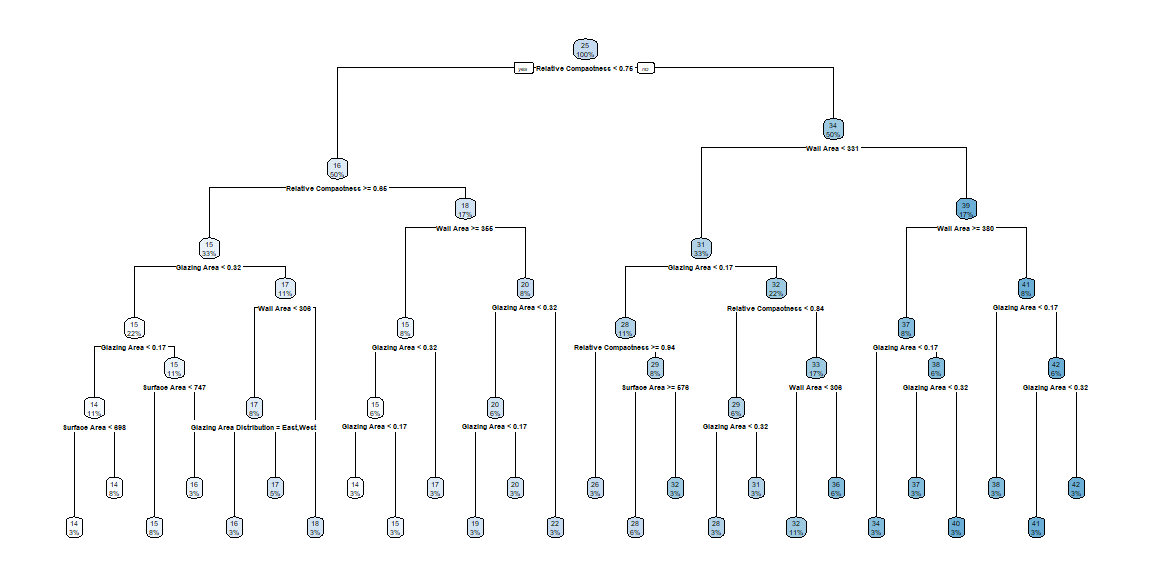
**Which are the most influential variables when predicting the cooling efficiency?**

All the models will be reviewed to answer this question and then compare the results:

*CART interpretation*

As was mentioned in the beginning, CART is a very interpretable model and even a non-technical person can look at the tree and follow the branches to see how a prediction was made and which variables influenced it.

rpart.plot(CARTmodel, uniform = T)



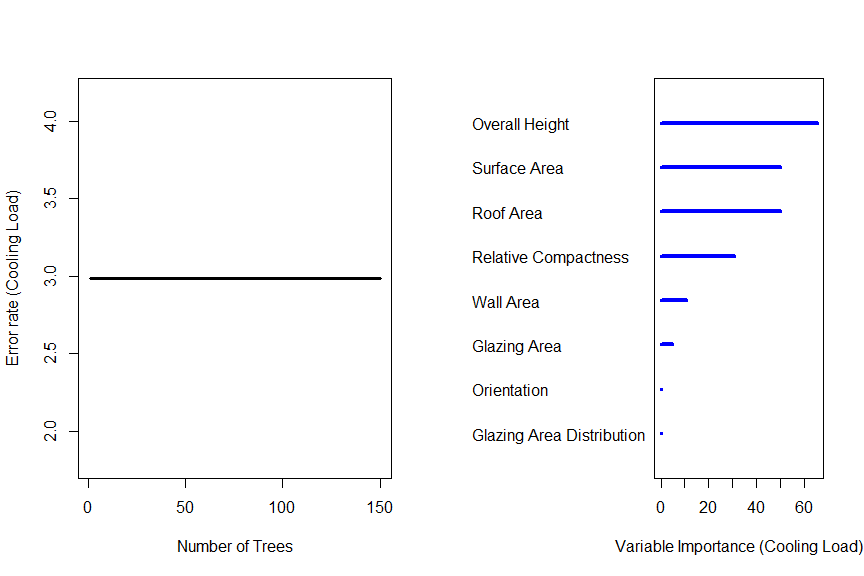
When we look at the tree it seems that Relative Compactness, Wall Area and Glazing Area are the most important variables, as they make up the first three levels. We can see the Surface Area and Glazing Area Distribution show up a few times on the last 2 levels. We don't see at all the Roof Area, Orientation and Overall Height.

The fact that the Orientation is less important resulted also in another study that was referenced in the Literature Review section: "Modelling energy efficiency performance of residential buildings stocks based on Bayesian statistical model" - by Marta Braulio-Gonzalo, Pablo Juan, Maria D. Bovea, Maria Jose Rua. This referenced study was working with aother parameters (like shape factor of the building and year of construction) where it was specified that "The least significant (parameter) is the solar orientation of the main facade of the building"

*Random Forest interpretation*

Because aside from making accurate predictions, this study investigates which predictors have the biggest influence on the response variable, a more simple way will be used for this, to avoid plotting and examining all the trees individually.

vimp is a function from the randomForestSRC package that computes the variable importance, similar to the coefficient p-values in linear models.

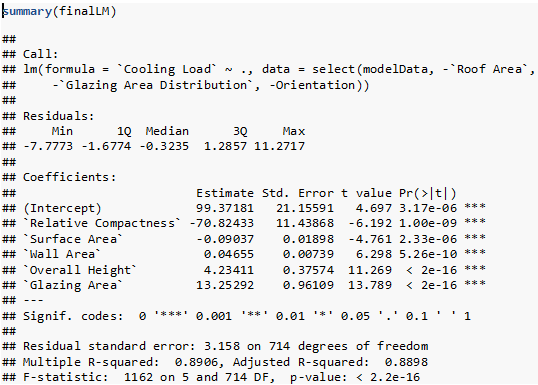


As we can see from the variable importance plot, the Random Forest model gives different results. The most important variable is the Roof Area, followed closely by the Surface Area, while the Overall Height and Relative Compactness have an average importance. Wall Area and Glazing area have minor influence on the predictions, while Orientation and Glazing Area Distribution don't seem to matter for this model.

*Linear Model interpretation*

The variables used by the model and their units of measure:

* Relative Compactness / no units
* Surface Area / m2
* Wall Area / m2
* Overall Height / m
* Glazing Area / no unit



The formula used to make a prediction will be:

The formula was obtained by putting the numeric values of the coefficients in the regression equation.

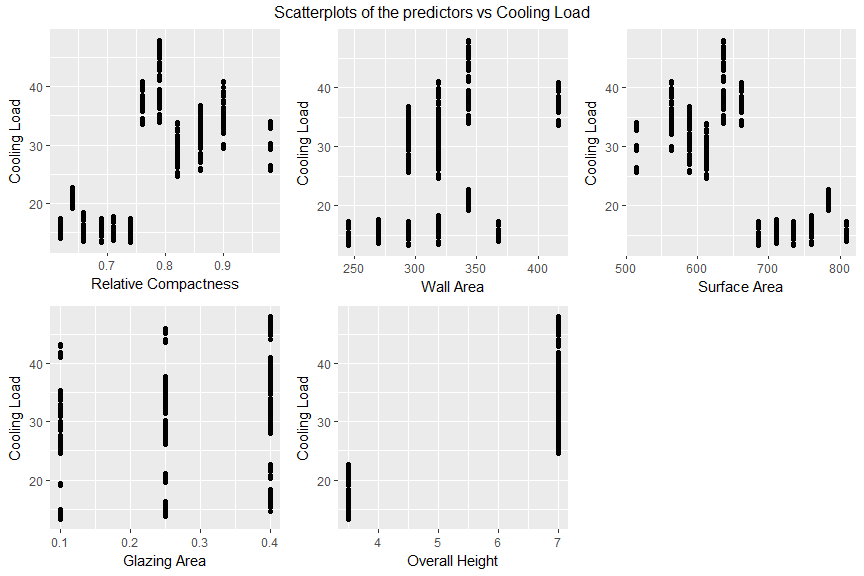
p-values were considered when doing variable selection but they will not be analysed further. This study recommends a more in depth analysis that takes the effect size into account.

*The most important variables*

The most important variables will be mostly determined by the best performing model, the random forest and these are Surface Area, Overall Height, Relative Compactness and Roof Area. It can be seen that some of these variables are important in the other models as well.

**Investigate energy impact when changing some building parameters**

The following scatterplots will be used to determine for which variables the change impact on the Cooling Load can be analyzed.



It is clear from the scatterplots that Glazing Area, Overall Height have very few values and Wall Area has such an irregular pattern that extrapolating outside those values is infeasible. It cannot be assumed that only the Surface Area can change, because other variables like Relative Compactness cannot stay constant when it changes because Relative Compactness is inversely proportional to the Surface Area. The nature of the dataset makes any independent numeric variable change analysis infeasible.

# Conclusions

The cooling load can be predicted accurately using a statistical model, but the best performing model does not capture the real relationships between the variables adequately (Glazing Area is insignificant in random forest for example).

An attempt was made to study the impact of changing predictors on the cooling load, but the nature of the dataset made this infeasible. A future study might be able to do this with more real(non-simulated) data.

This study concludes that the random forest should be used in together with one of the more interpretable models (CART or linear) by using random forest to make predictions and the other model to figure out what variables influenced the prediction.

In conclusion, these results will help building owners, architects, engineers, and other stakeholders to optimize the energy efficiency of the HVAC equipment in their buildings. At the same time those models can help urban planners, developers, and local authorities to identify residential areas that require a higher energy demands and where residential buildings require more urgent energy efficiency refurbishments.

**References**

1. Arlan Burdick, Strategy Guidline, Accurate Heating and Cooling Load Calculations, US Department of Energy, Energy Efficiency & Renewable Energy, Building Technologies Program, June 2011
2. Athanasios Tsanas, Angeliki Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, March 2012, <https://archive.ics.uci.edu/ml/datasets/Energy+efficiency>.
3. S. Attia L.Beltran, A.D.Herde, J. Hensen “A Comparison of Ten Different Building Performance Simulation Tools” - Eleventh International IBPSA Conference Glasgow, Scotland Jul 27-30, 2009

<http://www.ibpsa.org/proceedings/bs2009/bs09_0204_211.pdf>

1. Marta Braulio-Gonzalo, Pablo Juan, Maria D. Bovea, Maria Jose Rua, Modelling energy efficiency performance of residential building stocks based on the Bayesian statistical model , 20 May 2016 <http://www.sciencedirect.com.ezproxy.lib.ryerson.ca/science/article/pii/S1364815216301542?via%3Dihub>
2. Marta Braulio-Gonzalo, Pablo Juan, Maria D. Bovea, Maria Jose Rua, A methodology for predicting the energy performance of indoor thermal comfort of residential stocks on the neighbourhood and city – A case study in Spain, Dec 2016 <http://www.sciencedirect.com.ezproxy.lib.ryerson.ca/science/article/pii/S0959652616312070>
3. Fabrizio Ascione, Rosa Francesca De Masi, Filippo de Rossi, Romano Fistola, M.Sasso, Giuseppe Peter Vanoli, Analysis and diagnostic of the energy performance of buildings and districts: methodology, validation and development of Urban Energy Maps, Apr 2013 http://www.sciencedirect.com.ezproxy.lib.ryerson.ca/science/article/pii/S026427511300053X
4. William Chung, Y.V.Hiu, Y.Miu Lam, Benchmarking, the energy efficiency of commercial buildings, 20 Nov 2004, <https://www.researchgate.net/publication/222515455_Benchmarking_the_energy_efficiency_of_commercial_buildings>
5. Also Solari, Saskia le Cessie, Jelle J.Goemans, Testing the goodness of fit in regression: a general approach for specified alternatives, Dec 31, 2012, <https://journals-scholarsportal-info.ezproxy.lib.ryerson.ca/details/02776715/v31i0028/3656_tgofiragafsa.xml>
6. J.R.Quinlan, “Induction of Decision trees” , 1986 Kluwer Academics Publisher, Boston <http://storm.cis.fordham.edu/~gweiss/selected-papers/quinlan86.pdf>
7. Wei-Yin Loh, “Classification and regression trees” – 2011 John Wiley & Sons, Inc <http://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf>
8. Halil Ibrahim Erdal, Onur Karakurt, Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms, 6 Nov 2012, <https://ac-els-cdn-com.ezproxy.lib.ryerson.ca/S0022169412009791/1-s2.0-S0022169412009791-main.pdf?_tid=0b44cc36-a96b-11e7-ae93-00000aab0f01&acdnat=1507166467_e83e9466b94c4ee9c5dc4ac243af51a6>
9. Athanasios Tsanas, Publications <http://people.maths.ox.ac.uk/tsanas/publications.html>
10. StatSoft Electronic Cersion Textbook, Boosting Trees for Regression and Classification, StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK, <http://www.statsoft.com/Textbook/Boosting-Trees-Regression-Classification>
11. Heating Load vs Cooling Load – <http://hvac-talk.com/vbb/showthread.php?314442-Heating-Load-VS-Cooling-Load>
12. DnI Institute, “CART Algorithm for Decision Tree” - July 15,2017,

<http://dni-institute.in/blogs/cart-algorithm-for-decision-tree/>

1. Leo Breiman, Jerome Friedman, Richard A.Olshen, Charles J. Stone, Classification and regression trees, 1998 CRC Press, ISBN 0-412-04841-8
2. Leo Breiman, “Random Forest” – Jan 2001, Berkely, CA <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>