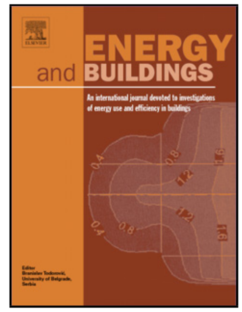
Accepted Manuscript



Title: Estimation models of heating energy consumption in schools for local authorities planning

Author: Alfonso Capozzoli Daniele Grassi Francesco Causone

PII: S0378-7788(15)30136-5
DOI: http://dx.doi.org/doi:10.1016/j.enbuild.2015.07.024
Reference: ENB 6017

To appear in: *ENB*

Received date: 3-3-2015
Revised date: 2-7-2015
Accepted date: 11-7-2015

Please cite this article as: A. Capozzoli, D. Grassi, F. Causone, Estimation models of heating energy consumption in schools for local authorities planning, *Energy and Buildings* (2015), http://dx.doi.org/10.1016/j.enbuild.2015.07.024

- The annual heating energy consumptions of eighty school buildings are analysed
- Two energy estimation models were developed to support public authorities planning
- A multiple regression model was built using nine different influencing variables
- CART enables also non-expert users to extract information for decision making
- MAE, RMSE and MAPE were calculated to compare the performance of estimation models

Estimation models of heating energy consumption in schools for local authorities planning

Alfonso Capozzoli[1*], Daniele Grassi[1], Francesco Causone[2]

1. TEBE Research group, Department of Energy, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.

2. Department of Energy, Politecnico di Milano, Via Lambruschini 4, 20156 Milano, Italy

* Corresponding author: Alfonso Capozzoli

Tel.: +39 011 0904413; fax: +39 011 0904499

E-mail addresses: alfonso.capozzoli@polito.it (A. Capozzoli), grassi.daniele88@gmail.com (D. Grassi), francesco.causone@polimi.it (F. Causone)

Keywords: School buildings; Multiple Linear Regression; Heating energy estimation; Decision Tree; Public authorities planning

Abstract

Large building stocks should be well managed, in terms of ordinary activities and formulating strategic plans, to achieve energy savings through increased efficiency. It is becoming extremely important to have the capability to quickly and reliably estimate buildings' energy consumption, especially for public authorities and institutions that own and manage large building stocks. This paper analyses the heating energy consumption of eighty school buildings located in the north of Italy. Two estimation models are developed and compared to assess energy consumption: a Multiple Linear Regression (MLR) model and a Classification and Regression Tree (CART). The CART includes interpretable decision rules that enable non-expert users to quickly extract useful information to benefit their decision making. The output of MLR model is an equation that accounts for all of the major variables affecting heating energy consumption. Both models were compared in terms of Mean Absolute Error (MAE), Root Mean Square error (RMSE), and Mean Absolute Percentage error (MAPE). The analysis determined that the heating energy consumption of the considered school buildings was mostly influenced by the gross heated volume, heat transfer surfaces, boiler size, and thermal transmittance of windows.

Nomenclature

β	Estimated Coefficient of Multiple Linear Regression Model
CART	Classification and Regression Tree
D-W	Durbin-Watson test
E	Error associated to the tree
EUI	Energy Use Intensity
EUI_{st}	Standard Energy Use Intensity

$EUI_{st,s}$	Standard and Specific Energy Use Intensity
F	Fisher-Snedecor test
HDD_{con}	Conventional Heating Degree Days
HDD_{real}	Real Heating Degree Days
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLR	Multiple Linear Regression
POW	Boiler Size (Heat Input)
R^2	Coefficient of Determination
R^2_{adj}	Adjusted Coefficient of Determination
RMSE	Root Mean Square Error
SUR	Heat Transfer Surface
$U_{windows}$	Thermal Transmittance of Windows
VIF	Variance Inflation Factors
VOL	Heated Gross Volume

1. Introduction

1.1 Energy consumption analysis in school buildings

Buildings are responsible for about 40 % of the total energy consumption in developed countries [1]. In countries like Italy, about 60 % of the existing building stock is more than 40-years-old [2]. A rapid and substantial energy retrofit program is therefore required for these existing buildings. There are about two million households in Italy that live in buildings requiring either demolition and rebuilding or refurbishment. Directive 2010/31/EU (EPBD recast) requires buildings, or parts of buildings, to meet a minimum energy performance or be subject to a retrofit or refurbishment. These requirements may be met by renovating a building's envelope and systems, but an effective management can also significantly impact a building's energy consumption. The EU Directive [3, 4, 5] requires public buildings to play an exemplary role in terms of energy savings. Public utilities and buildings that are typically owned and managed by municipalities include: street lighting, schools, administrative buildings, public transport, and sport centres, such as swimming pools and gymnasiums [6]. Local governments would clearly benefit from having access to energy consumption data. Further, being able to understand the savings potential of these assets would help to prioritise energy and environmental projects and better illuminate their financial aspects [7]. According to the US Department of Energy, school buildings constitute a major part of the public building stock. Around 25 % of the energy expenses in schools could be saved through better building designs and more energy-efficient technologies, combined with improvements in operation and maintenance [8].

De Santoli at al. [9] evaluated the energy performance of public schools in Rome. They defined intervention strategies to reduce energy consumption and identified action priorities by means of a simple payback time analysis (PBT). Dimoudi et al. [10] conducted an energy simulation to study the energy savings potential of school buildings in Greece. Kim Tae Woo et al. [11] analysed the energy consumption of some elementary schools in South Korea by utilising monitoring data from January 2006 to December 2010. They determined that electrical energy was consumed the most, followed by gas and oil. During the monitoring period, electrical energy continued to increase its relevance on the energy breakdown because of cooling/heating system replacements. These and other studies were carried out in recent years to estimate the energy consumption of school buildings. The literature shows that there are two main approaches for estimating a building's energy consumption: the direct approach or the inverse approach. The first approach calculates the energy demand by running an energy simulation under a steady state or dynamic conditions. The second approach uses historical data to produce data driven models that estimate the energy consumption.

A large part of the current literature focuses on the inverse approach. Analysts and decision makers have access to several applications of new or recast versions of existing models. Corrado et al. [12] defined a simplified method for predicting future consumption based on climatic and real use data on a stock of 120 school buildings. Corgnati et al. [13] then validated this method, using another stock of 118 schools, as did Ariaudio et al. [14]. Amber et. al. [15] gathered daily values of a school building's electrical consumption on the Southwark campus of the London South Bank University from 2007 to 2013 and then developed a multiple regression model to estimate future daily electrical consumption. Beusker et al. [16] evaluated the energy consumption of schools and sports facilities in Germany using different linear and nonlinear regression models. Thewes et al. [17] presented a regression model with categorical variables to predict the electrical and heating energy consumption of school buildings in Luxembourg.

Innovative techniques, including machine learning, data mining, and knowledge discovery in databases, have also been successfully applied to building energy consumption data in recent years [18]. In particular, a classification tree which consists of a multi-stage decision-making process that is useful to categorise observations in a finite number of classes, can be a powerful estimation tool. This method has not yet been applied in other studies to estimate the energy consumption of school buildings.

In this paper, the heating energy consumption of a school building stock located in the north of Italy is analysed using a Multiple Linear Regression (MLR) model and a Classification and Regression Tree (CART). Both MLR model and CART are data driven models that have been successfully applied to estimate a building's energy demand. Nevertheless, the outcome of MLR is an equation, while the output of CART are decision rules that allow users to quickly extract relevant information [19]. This characteristic substantially changes the practical applicability of the two models.

1.2 Implementation of multiple regression analysis and classification tree for buildings' energy use estimations

In recent years, numerous researchers successfully employed multiple regression model as a tool for energy consumption estimations. Al-Garni et al. [20] correlated electrical energy consumption with relevant climatic variables (air temperature, relative humidity, solar radiation), and variable occupant populations through statistical methods (regression model) to forecast the overall electrical energy consumption in Eastern Saudi Arabia. Aranda et al. [21] developed three regression models to predict the Spanish banking sector's annual energy consumption. The first model can be used to estimate the energy consumption of the whole banking sector, while the second estimates the energy consumption for branches under conditions of a low severity winter climate and the third under conditions of a high severity winter climate. The variance reported for the three models is 58 %, and 68 %, respectively. Korolija et al. [22] developed regression models to predict the annual heating, cooling, and electrical auxiliary energy consumption of five different types of HVAC systems (variable air volume – VAV, constant air volume – CAV, fan-coil system with dedicated air (FC), and two chilled ceiling systems with dedicated air, radiator heating, and either embedded pipes – EMB - or exposed aluminium panels – ALU) for office buildings in the UK. Freire et al. [23] used independent variables like energy consumption, ventilation and air conditioning power, outdoor temperature, relativity humidity, and total solar radiation to develop a regression equation to predict the indoor air temperature and relative humidity for two buildings with low and high thermal mass. The literature demonstrates therefore that regression models offer a robust methodology for estimating a building's energy consumption (e.g., heating, cooling, lighting, etc.).

Decision trees belong to the *machine learning algorithms* family. This method is recognised as an emerging analysis tool and is currently receiving plenty of attention from applied research. Yu et al. [18] used the decision tree to classify and predict building energy consumption. This method was applied to Japanese residential buildings for predicting and classifying building Energy Use Intensity (EUI) levels based on training data. This tool was then evaluated on a sample test.

Zhao et al. [24] used a C4.5 decision tree algorithm, locally weighted naïve Bayes and support vector machine, to classify occupant behaviour and to create schedule models for building energy simulation. The results show that the C4.5 algorithm correctly classified 90 % of individual behaviour and this allowed getting closer to the real group schedule. Mikučionienė et al. [25] used a decision tree to increase the sustainability and improve the criteria for evaluating energy efficiency measures in a public building renovation in Lithuania. By analysing and weighting each variable (related to insulation of external walls, roof insulation, heating substation renovation, reconstruction of the entire heating system, and installation of a ventilation system with exhaust air heat recovery), the researchers created a decision tree to evaluate the influence of each variable on energy consumption. The results show that this algorithm reduces the amount of data that must be understood by transforming it into a more compact form while still preserving the basic substance. The researchers determine whether the data are characterised by well-separated object classes and finally, this algorithm determines the precise relationship between attributes and their class.

In this paper, two different estimation models are developed using a database consisting of 80 school buildings located in the province of Turin. The estimation models include climatic, envelope and heating system variables, and annual metered heating energy consumption. They are:

- a Multiple Linear Regression (MLR) model that estimates the energy use for heating based on geometrical, climate, and thermo-physical characteristics. This model creates an equation that relates n independent variables to the dependent variable;
- a Classification and Regression Tree (CART) which consists of a multi-stage decision-making process to classify observations in a finite number of classes. The model's output is a flowchart constructed by subdividing the observations into homogeneous subsets with respect to the dependent variable or response (represented in our model by heating energy consumption).

The two estimation models are compared to determine which one is more accurate in terms of a residuals analysis and errors (MAE, RMSE, MAPE). The possibilities and limitations of the two models are ultimately contrasted, highlighting advantages and disadvantages for their use by a final operator, such as a consultant or a decision maker. Moreover, this paper discusses the practical application and robustness of the constructed estimation models.

2 Methodology

A wide range of theoretical and practical factors that are relevant to each building should be considered to create estimation models that analyse building energy consumption. The methodology followed in this study is schematised as shown in the flowchart in Figure 1.

An existing database was initially analysed to evaluate the consistency of the school building stock. The available variables are associated with a building's envelope, heating/cooling systems, and location. This step is useful to understand the limits of applicability of the models, which may be applied to other building stocks with similar features, once they are validated. In the second step of the analysis, two estimation models were implemented (MLR model and CART). Finally, in the third phase, the two models were developed and compared, highlighting their usefulness for public school managers.

2.1 Pre-processing analysis

The database contains information from 80 school buildings without sport facilities (sport halls), situated in the Province of Turin (Italy). The initial dataset was composed of 120 school buildings located in the same area, but the sample was reduced to 80 schools due to missing heating energy consumption data from 40 school buildings. The analysed influencing variables are related to the opaque and transparent building envelope, heating systems, building geometry features, and climatic data.

From a climatic point of view, the Province of Turin is located in the Italian climate zones E and F. The analysed buildings are located in a climate with Conventional Heating Degree Days (HDD_{conv}) ranging from 2517 to 3197 DD. Figure 2 shows the frequency distribution for the gross heated volume and heat transfer surface of the sampled buildings. The majority of schools have a gross heated volume lower than 35000 m^3 (about 60 %). Schools with a higher gross heated volume are composed of two or more buildings. In addition, about 60 % of the sampled schools have values of heat transfer surface lower than 10000 m^2. For this reason, most of the sample is composed of buildings with an aspect ratio (ratio of heat transfer surface on gross heated volume) range from 0.25 to 0.40 m^{-1}. The heat losses mainly depend on the quality of the building envelope and not from the building shape.

Figure 3 shows the frequency distribution of the sampled buildings for the thermal transmittance of walls and windows. As can be seen, most of the buildings are characterised by a thermal transmittance of windows higher than 4 $W/(m^2 K)$ (about 65 % of the sample is composed of single glazing) and by opaque walls without thermal insulation (80 % of the sample is characterised by values higher than 0.40 W/m^2K).

Figure 4 shows the frequency distribution of the sampled buildings for boiler size (heat input) and average system efficiency. The boiler size (heat input) ranges from values lower than 500 kW to values higher than 8000 kW. 12 % of the schools are equipped with a boiler size lower than 500 kW, 43 % from 500 to 1500 kW, 17 % from 1500 to 2000 kW, and only 28 % by a boiler size higher than 2000 kW. Analysing the frequency distribution of the average seasonal system efficiency (the ratio between building energy need and primary energy) reveals that 83 % of the sample have values lower than 0.70. This figure denotes the presence of high thermal losses in subsystems. Moreover, the schools are equipped with old emission subsystems (cast iron radiators), old distribution subsystems (non-insulated pipes), and old control subsystems, i.e. centralised control that is only installed at the generation system level (e.g. climatic control).

Several variables related to school buildings should be considered in a comprehensive analysis of energy consumption. Ventilation rates, hours of use, set points and time clock settings, infiltration rates, internal heat gains, solar gains, geometrical building characteristics, building envelope physical variables, heating system features, outdoor temperature, and number of pupils and classes are all considered important variables in characterising a school's energy use. Moreover, occupant behaviour can significantly impact energy consumption, particularly the opening and closing of windows. However, some of these variables (e.g. infiltration and ventilation rates or variables related to occupant behaviour) are very difficult to obtain.

In [26], it was claimed that the floor surface and/or the volume (mostly the volume) primarily influenced the heating energy consumption and the electrical energy in school buildings in their analysed sample [27]. In some cases, it was also verified that the data related to the transmittance of opaque components of the façades, the boiler size, and the daily period of use significantly influenced the heating energy consumption [27].

In our work the available data collected for the analysed sample to characterise the heating energy consumption for each school building are: *real heating degree day, gross heated volume, heat transfer surface, aspect ratio, floor heated* *area, building height, numbers of floors, thermal transmittance of walls, thermal transmittance of windows, boiler size (heat input), number of classrooms, number of pupils, annual operating time, average seasonal system efficiency.*

Table 1 provides a list of the variables with the definition of the data location, central tendency, and dispersion for each of them. From the literature [26, 27], we know that the selected variables can be considered as the most influential factors. No major effect of controls may be reported in the database, since no local control was installed in the school. Occupant behaviour definitely affected the final energy performance by opening and closing windows. However, information about occupant behaviour were not available and are very difficult to get.

In order to standardise the impact of climate on heating energy consumptions, the degree-days method was applied [28]. For this purpose, standard heating energy consumptions (EUI_{st} - Standard Energy Use Intensity) for each school buildings was defined as:

$$EUI_{st} = EUI \cdot \left(HDD_{conv} / HDD_{real} \right) \tag{1}$$

where EUI is the heating energy consumption [kWh], HDD_{con} are the Heating Conventional Degree Days, and HDD_{real} are the Heating Real Degree Days (related to the year 2012). The average heating energy consumption of the school building stock EUI_{st} is equal to 830 MWh/year (Figure 5). In order to compare buildings of different sizes, the EUI_{st} was normalised by the gross heated volume ($EUI_{st,s}$– Standard and Specific Energy Use Intensity). The $EUI_{st,s}$ ranges from 17.74 to 61.12 kWh/m^3/year (Figure 5).

2.2 Outliers detection

A pre-processing analysis [29] is required to identify outliers before creating an estimation model. An observation is an outlier when it departs from other members of the sample and appears to be inconsistent with the remaining dataset. The presence of one or more outliers could reduce the capacity of the models to estimate the heating energy consumption. Outliers should be eliminated from the dataset [30, 31], however, their treatment is not simple. Several indexes were evaluated to identify outliers in the dataset. In fact, these indexes can be used together to perform an accurate screening of the database:

- z-score;
- Mahalanobis Distance;
- Index of Mardia;
- Distances Cook;
- Leverage Value.

These techniques made it possible to detect outliers at both multivariate and univariate levels. The first index (z-score) detected outliers at the univariate level, i.e. for each variable. The other four indexes detected multivariate outliers by considering a combination of different variables. These outlier detection categories are complementary and should be used together. Indeed, a case cannot be considered an outlier if it only has one single distorted value. At the same time, multivariate outliers represent a pattern of responses that are unlikely to be comparable to the rest of the sample.

The definitions of the analysed indexes are briefly explained in the following.

The z-score is used to measure when the observed value deviates from the mean value. It is expressed by the mean of the following equation.

$$z\text{-}score = (x\text{-}\bar{x})/DS \tag{2}$$

where x is the observed value, $\bar{x}$ is the average value, and DS is the standard deviation. On the basis of Chebyshev's theorem, if z-score ≥ the values are potential outliers.

The Mahalanobis distance (D_{hk}) is a statistical measure of the distance between the units. It is calculated by taking into account the correlation between variables:

$$D_{hk} = \sqrt{(x_h - x_k)^T \cdot W^{-1} \cdot (x_h - x_k)} \text{ with } h \neq k = 1, \ldots\ldots, n \tag{3}$$

where x_h and x_k are the vectors with the observations on the samples h and k, and W is the variance-covariance matrix between the observed variables. As a rule of thumb, values of D_{hk} higher than the chi-square critical (α = 0,001, degree of freedom = predictors) are considered abnormal points.

The Index of Mardia checks if the relationship between variables can be considered linear. The multivariate normality is met if the Index of Mardia is less than a critical value:

$$Mah_{critical} = k \cdot (k + 2) \tag{4}$$

where k is the number of predictors.

The Distances Cook (D_i) is the distance between the regression line that includes all observations and the regression line that does not include the i-th observation:

$$D_i = |\hat{y} - \hat{y}_i| / p \cdot DS \tag{5}$$

where ŷ is the expected value, $\hat{y}_i$ is the expected value without the use of the i-th case, p is the order of the multiple regression analysis, and DS is the standard deviation. Generally, values higher than 1 are considered abnormal points.

The Leverage Value ($L_{average}$) is a measure of how much the specified value of the independent variable deviates from its mean. The values vary between zero (no influence) and (n/(n-1)) (greatest influence). The average value corresponds to:

$$L_{average} = (k+1)/n \tag{6}$$

where k is the number of predictors and n is the number of cases analysed. As a rule of thumb, values higher than two or three times the average values are considered abnormal points.

2.3 Multiple Linear Regression model

The multivariate statistical analysis [32, 33] can estimate the value of some variables, if the parameters included in the model are actually relevant for the building's final energy consumption. The MLR model (classical model for parameter estimation) is expressed as follows:

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n + \varepsilon \tag{7}$$

where Y is the dependent variable, β_0 is the intercepts, $\beta_{1,\ldots p}$ are the estimated coefficient of MLR model, and ε is the statistical error. The regression model's coefficients $\beta_{1,\ldots p}$ are estimated by using the ordinary least square or linear least square method. This method tries to minimise the sum of the squares of the error terms.

MLR model can be evaluated using statistical tools. The adjusted coefficients of determination (R^2_{adj}), is a statistical index that provides information about the goodness of fit of a model. It represents the proportion of the variation in the dependent variable that is attributable to the explanatory variables:

$$R_{adj}^2 = 1 - \left[(1 - R^2) \cdot (n-1) / (n-p-1) \right] \tag{8}$$

where R^2 is the coefficient of determination, n is the number of observations, and p is the number of variables included in the model.

The coefficient t student is used to test the null hypothesis, i.e. when the values of the estimated coefficients of MLR model are not significant:

$$t = \beta_{1,\dots,p} / SE_{\beta_{1,\dots,p}} \tag{9}$$

where $\beta_{1,\dots p}$ are the estimated coefficients of MLR model and $SE_{\beta 1,\dots p}$ is the standard error of each the estimated coefficient. Generally, if, $t \leq |2|$, $\beta_{1,\dots p}$ is less significant.

A method for testing the significance of the MLR model is the Fisher-Snedecor test (F). It is conducted on the entire model and is based on the decomposition of deviance:

$$F = MS / MS_R \tag{10}$$

where MS is the Means Square of the model and MSr is the Residual Mean Square. If the value of F does not exceeds the critical value (default value for a given probability), the correlation between the variables is not linear. Therefore, there could be a different correlation.

Durbin-Watson $(D-W)$ is a statistic test used to detect the presence of autocorrelation in the residuals (estimation errors) of a MLR model.

$$D - W = \sum_{i=2}^{n} (e_i - e_{i-1})^2 \Big/ \sum_{i=1}^{n} e_i^2 \tag{11}$$

where $e_i = y_i - \hat{y}_i$ and y_i and $\hat{y}_i$ are respectively the observed values and the expected value of the response variable for individual *i*. The value of D-W always lies between 0 and 4. As a rule of thumb, D-W = 2 indicates no autocorrelation, D-W < 2 indicates negative autocorrelation, and D-W > 2 indicates positive autocorrelation.

2.4 Classification and Regression Tree method

The CART [34, 35] is a binary decision tree that is constructed by splitting a parent node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample. The CART can easily handle both numerical and categorical variables.

A decision tree generation consists of a two-step process: learning and classification. In the first step, the dataset is divided into a training set and a testing set. The creation of these two subsets is the most delicate part of the technique. It is important that the training set and the testing set come from the same population and that they are disjointed. In the classification process, the results obtained from the training set are the input to test the decision tree. The accuracy of the model is measured by comparing the estimated values of each "leaf node" with the real values contained in the test sample. If the estimation is acceptable, the decision tree can be applied to new datasets for classification and estimation. Initially, all records in the training data are grouped together into a single unit. At each iteration, the algorithm chooses a predictor attribute that can "best" separate the target class values. Measures of impurity are used to estimate the ability of a predictor to separate the target class values. The CART consists of three parts:

1. Construction of maximum tree: the classification tree is built in accordance with the splitting rule. Each time data must be divided into two parts with the highest homogeneity. The Least Square Deviation measures the impurity of a node *t* and is defined as:

$$i(T) = \frac{1}{N_w(t)} \cdot \sum_{i=t} w_i \cdot f_i \cdot (y_i - \bar{y}(t))^2 \quad (12)$$

where N_w is the weighted number of cases in node *t*, w_i is the weight of the variable in the case, f_i is the frequency value of the variable, y_i is the value of the response variable, and $\bar{y}(t)$ is the weighted average value of the variable at node *t*. The best split s^* of a generic node t is what determines the greater decrease of the i(t). For each split s of node *t* into t_r and t_l the following algorithm is valid:

$$\Delta i(s^*, t) = i(T) - (i(t_l) + i(t_r)) \quad (13)$$

where $\Delta i(s^*,t)$ is the decrease of impurities in a generic node, i(T) is the Least Square Deviation of the given node, $i(t_{l,r})$ are the Least Square Deviation of the two nodes split.

2. Choice of the right tree size. Two rules for the stop can be used in practice: optimisation by number of points in each node (minimum number of cases in the parent's node and the child's node) and the error E associated with the tree. The E parameter allows the tree to be properly built:

$$E = \sum_{i=1}^{n} \left(n_i / n_i^* \cdot \varepsilon_i \right) \quad (14)$$

where n_i is the total number of records in the training set which terminate in the leaf, and n_i^* is the number of records classify bin the leaf i.

The optimal condition is obtained by setting the error E=1. In fact, in this case the tree correctly classifies all of the records in the training set. The optimisation of the tree size is important, because the maximum trees may turn out to be very complex and may consist of hundreds of levels.

3. Classification of new data: each of the new observations will be set to one of the terminal nodes of the tree by means of a set of questions. A new observation is assigned with the dominating class/response value of the terminal node, where this observation belongs.

3. Results: development of models

3.1 Outliers detection analysis

In order to find potential outliers, a pre-processing analysis was carried out prior to creating the estimation model. All of the variables should show a sufficient range of variability and have skewness and kurtosis values of less than |1.00|. Indeed, including variables whose distribution is too different from the normal value into the MLR model can lead to the violation of the assumptions of linearity and homoskedasticity of the residual anomalies. The variable *floor heated area* was excluded, because its values were missing for 18 schools.

The pre-processing analysis identified 14 potential outliers. After conducting an accurate frequency distribution analysis of the sample, it was observed that the detected outliers belonged to the tails of distribution for each variable. In particular, it was verified that these outliers influence the mean and standard deviation for each variable, causing a non-normal distribution for all of them. Even if the detected outliers can be considered reliable from an energy measurement point of view, it was verified that they decrease the performance of the estimation models. A detailed analysis on these buildings showed that they are characterised by high thermal transmittance values, low system energy efficiencies, low number of pupils, very low or very high volume. This is the reason why the outliers belong to the tails of distribution for each variable and therefore determine anomalous values of heating energy consumption. Table 2 shows the indexes evaluated for the sample without outliers.

3.2 Multiple Linear Regression model

In order to develop the MLR model, all of the reported anomalies were deleted by excluding the identified outliers from the database. The assumptions relating to the specification of the model (do not omit relevant predictors and do not include irrelevant predictors) were verified by evaluating the bivariate correlations between the independent variables and the dependent variable constituted by heating energy consumption (Figure 6 and Table 3).

The variables *building height* and *number of floors* have a correlation coefficient of less than 0.20 (Figure 6). For this reason, they were not included in the model.

The values of the parameter Variance Inflation Factors (VIF) are reported in Table 3. VIF allows detecting the presence of multicollinearity between the explanatory variables. In general, a multicollinearity occurs if the value of VIF exceeds 10.

The results analysis found a strong correlation between:

- number of pupils and number of classrooms;

- aspect ratio and gross heated volume;

- aspect ratio and heat transfer surface.

Given these findings, the variables included in the MLR model were reduced to: *real heating degree days, gross heated volume, heat transfer surface, thermal transmittance of walls and windows, boiler size, number of pupils, annual operating time,* and *average seasonal system efficiency*, as summarised in Table 4.

The sample was randomly split into the training dataset (39 records were selected from the database, i.e. 70 % of the sample) and testing dataset (the remaining 27 records, i.e. 30 % of the sample). The estimation model was therefore developed on the basis of a training sample. The training set does not include the outliers previously identified. Each variable was standardised by the z-score method (Eq.1) to compare variables between them by assuming the same distribution ($\mu = 0$; $\sigma = 1$). The most accurate estimation for heating energy consumptions (measured in kWh) is calculated by means of the following equation:

$$EUI_{st} = 662765 + \beta_1 \cdot X_1^* + \beta_2 \cdot X_2^* + \beta_3 \cdot X_3^* - \beta_4 \cdot X_4^* - \beta_5 \cdot X_5^* - \beta_6 \cdot X_6^* + \beta_7 \cdot X_7^* + \beta_8 \cdot X_8^* - \beta_9 \cdot X_9^* \quad (15)$$

where the variables of the model are shown in Table 5, including detailed information about the estimated coefficients (β) of MLR model, partial standardised regression coefficients (b), and the t-values.

All of the examined variables within this study can theoretically impact heating demand. The *gross heated volume, boiler size,* and *thermal transmittance of windows* exhibit the greatest impact in the model, with partial standardised regression coefficients of 0.86, 0.64, and 0.61, respectively. The t-test identifies the inference on individual coefficients β. In particular, it verifies whether every single variable X^* influences the response variable. The variable *annual operating time* and *average seasonal system efficiency* are the only two variables with a t-value of less than |2|. For this reason, both of their estimated coefficients of MLR model are less significant. The variance showed by the model compared to the total variance of the sample is 86 % (R^2_{adj}), therefore, 86 % of the heating energy consumption variance can be explained by the nine variables used in the model. Moreover, there is an absence of auto-correlations among residuals (D-W = 2.05 ≈ 2) and the value of the Fisher-Snedecor test (F = 27) is greater than the critical value (F_{crit} = 4). As such, the MLR model can be considered robust. To assess the quality of the estimation model, Figure 7 shows the distribution plot between the estimated EUI_{st} and the monitored EUI_{st} using the testing dataset.

The best fit is affected by an error of 1 %, while the bad fit of the model is affected by an error of 40 %. The average error is equal to 15 % and for testing dataset the value of R^2 is 86 %. The model tends to underrate the energy consumption; in fact, 16 cases of 27 show an estimated heating energy consumption that is lower than the actual value. The validation test demonstrates that the model has an adequate estimation ability.

3.3 Classification and Regression Tree

In order to develop the CART, the training dataset and the test dataset used are the same as the ones used in the regression model. The CART algorithm selected five parameters from the database to model input variables (Table 6). The decision tree was constructed to estimate the heating energy consumptions. The rules set for the arrest of the tree are as follows:

- minimum number of cases (*parents node)*: 2
- minimum number of cases (*children node):* 2
- E = 1.

The tree includes a total of eight leaf nodes that represent the final classes. The estimation of the heating energy consumptions of each leaf node corresponds with the average of cases included in it. The algorithm can be translated into a set of decision rules that take the following form: *if* antecedent conditions, *then* consequent conditions. Table 7 presents the results of the CART in terms of the decision rules for the training dataset, starting from the root node and following all the way to each leaf node.

The decision rules can be used to estimate the EUI_{st} target level of a new school building having similar features. For example, looking at the first rules (Table 7), the EUI_{st} level can be estimated as follows:
Step 1: The root node is the starting point for the estimation. Table 7 shows that the value of the VOL variable should be examined first. If the VOL is higher or equal to 33195 m^3, then it is possible to go to the next step.
Step 2: examine the value of the SUR variable; if SUR is lower or equal to 12818 m^2, the EUI_{st} level of the school building is 968 MWh.
The CART carries out a sensitivity analysis before creating the decision rules in order to select the variables more correlated with the heating energy consumptions. In fact, the variables selected by the algorithm are characterised by high correlation coefficients (see Figure 6). The other factors (heating degrees days, thermal transmittance of walls, number of pupils and classrooms, average seasonal system efficiency, and annual operating time) do not appear in the decision tree, because they were excluded during the pruning process.
As previously mentioned, the accuracy of the decision tree must be evaluated before it is applied to a new dataset. Since the estimated values correspond to the mean value of the data included in the node, the estimation will always be affected by an error. For this reason, it is appropriate to associate a confidence interval for each estimated value. The confidence intervals with a 95 % probability of containing the true parameters were calculated and the results are shown in Table 8.

The decision tree was applied to the testing dataset and the results are reported in Figure 8. The best fit is affected by an error of 2 %, while the model's bad fit is affected by an error of 33 %. The average error is equal to 13 %, and the value of R^2 for the testing dataset is 86 %.

3.4 Models comparison

Two estimation models were evaluated based on their ability to estimate heating energy consumption in school buildings. Although they share the same goal, these models are based on different methodologies. To understand the possibilities and limitations of both, a residuals analysis is needed. The residuals are often used as an indicator to validate a model. The basic hypothesis is that the residuals, i.e. the errors, are randomly distributed. Moreover, they should not be correlated with the dependent or independent variables and the average value of the residuals should be equal to zero. The last hypothesis was verified for both of the models. The value is less uncertain for the residuals of CART than for the MLR model. No significant correlation was identified between the variables in both of the models. Comparing the goodness of fit of the two models, in percentage terms, MLR model commits an average error of 15 %, while the CART commits an average error equal to 13 %. In particular, the variance explained by the two methods is equal to 86 %. The residuals analysis and the coefficient of determination are not enough to evaluate the performance of the estimation models. Three other criteria have been used to test the performance of both models, the Mean Absolute Error (MAE) the Root Mean Square Error (RMSE) and the Mean Absolute Percentage error (MAPE).These parameters are shown in the following equations:

$$MAE(N) = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{16}$$

$$RMSE(N) = \sqrt[2]{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \Big/ N} \tag{17}$$

$$MAPE(N) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100 \tag{18}$$

where y_i is the monitored heating energy consumptions and $\hat{y}_i$ is the estimated heating energy consumptions and N is the sample size.

Table 9 gathers the results of the error analysis. The three indexes for the CART are always lower than the values estimated for the MLR model. The MAPE values are very similar for both models, but the MAE and the RMSE, considering the same testing dataset, are lower for the CART. A low RMSE value means that the error is characterised by a low dispersion. This is more clear in Figure 9, where a comparison between measured and estimated EUI_{st} is reported with a box-plots representation.

The range of the measured EUI_{st} is comparable with the estimated outputs of both models. In fact, the median values for the three bars is equal. The lower and upper quartile values of the measured data and the data estimated with CART are quite close. This confirms that the MLR model tends to underrate the heating energy consumption (minimum, median, and lower quartile are closer compared to the monitored data), while the CART's output is unbiased. The results show a strong relationship between the dependent variables and the heating energy consumptions. In particular, the most influential variables in the MLR model (gross heated volume, heat transfer surface, boiler size, and thermal transmittance of windows) are exactly the same factors that CART selected to create the decision rules. It can be concluded that both of the estimation models were correctly developed. CART's performance is slightly better than that of MLR model, as demonstrated by the results presented earlier (residual analysis, R^2, MAE, RMSE, and MAPE).

4. Discussion

The analysis of CRESME [2] shows that the Italian school building stock could achieve energy savings of about 48.3 % and shift from a current energy consumption rate of 9.6 TWh/yr to a target value of 5.0 TWh/yr. Similar results may be obtained for public directional buildings and dwellings. However, a more detailed analysis may be conducted at the local level, enabling a better definition of actual planning actions and economic assessments.

The actions that a local authority may adopt to challenge energy savings in the construction sector include:

- defining a public building portfolio (building stocks) and reference performance benchmarks;
- setting simple thresholds for energy performance, using existing energy data;
- setting a priority action list for the energy management and renovation of the building portfolio;
- adopting economic policies to promote the most relevant actions.

These actions may be part of an effective energy plan, but they require a set of technical steps: all of the considered asset's fundamental data must be collected, gathered, and assembled in an appropriate database, eventual outliers must be processed, and an analysis then ultimately produces effective decision rules. These include both the planning of ordinary management activities and strategic planning targeting energy efficiency improvements.

The choice of the most adequate and accurate estimation model to perform the required analysis, knowing its possibilities and limitations, is crucial in order to correctly inform the following local authority actions.

The estimation accuracy of the two models analysed in the present paper showed to be influenced by the nature of the dataset, in terms of its density and how uniform the frequency distribution is. The CART is based on binary splitting criteria of the response variable as a function of the influencing variables. It performs well when the leaf nodes are characterised by values that are close to the mean (low confidence intervals). In this case also numerical variables can be used as target attribute. However, generally the CART algorithm is used to classify categorical attribute. On the other hand, a non-uniform dataset could make the MLR model incapable of estimating unbiased regression coefficients.

The MLR model requires knowing the exact values of all input variables. This is a weakness, in fact, as the precise value of some of the variables, such as the thermal transmittance of walls/windows or heat transfer surface, is not easily obtained or readily available for existing buildings. Moreover, the model requires the input parameters to be standardised. For this reason, it is not easy for inexperienced users to interpret and use. On the other hand, the MLR model can be used to create benchmarks [36]. A benchmark value may be used as a target to be reached or exceeded and may prove quite useful to guide designers towards the optimal technical and economical solution. This is not possible with the CART.

Despite being a particular data mining technique, the CART's output consists of a set of decision rules that even non-experts can easily understand and use. Useful information can be obtained from this model, for example, it helps to understand a building's energy consumptions pattern and how to optimise a building's design. The algorithm automatically selects the different parameters as predictors. These are used to split the nodes of the decision tree, and their proximity to the root node indicates the strength of the influence and the number of records impacted. By examining the decision rules (see table 7), one can identify what primary factors account for the energy demand profiles of the schools. Among the considered factors, the root node, i.e. gross heated volume, indicates that the size of the schools is the most important element in determining energy demand. The heating energy consumption of large school buildings (Rules 1 – 2) is only influenced by the gross heated volume and heat transfer surface. Instead, in medium size school buildings (Rules 6 – 7 – 8), the heating energy consumption is influenced by four significant factors (gross heated volume, heat transfer surface, boiler size, and thermal transmittance of windows). Finally, the heating energy consumption of small size school buildings (Rules 3 – 4 – 5) is a function of two geometric factors (gross heated volume and heat transfer surface) and one construction feature (thermal transmittance of windows).

The accuracy of MLR and CART models results quite good also compared with values obtained by other researches. For example in [16], eight different regression models (Linear – Logarithmic – Quadratic – Cubic – Inverse – Linear and Inverse – Power – S – Exponential) were developed to estimate the energy consumptions of 105 schools located in Germany. The linear and inverse regression model showed the best fit with a MAPE of 17 %. Thewes et al. [17] developed a regression model able to explain 53 % (R^2 = 53 %) of electrical and heating energy consumption variance for 68 school buildings situated in Luxembourg. In [15], a multiple regression model was developed to estimate daily electricity consumption of an administration building located at the Southwark campus of London South Bank University in London. The final model has an adjusted R^2 value of 88 %.

No example of the CART model used for the estimation of heating energy consumption in schools is available. Therefore, only results obtained for other building types may be reported. Yu et al. [18], used a decision tree and decision rules to classify the EUI level of a new residential building in Japan. The C4.5 algorithm was used with a percentage error between 0.2 and 141.9 % (average error equal to 25 %), on the basis of 55 records for the training dataset and 12 records for the test dataset.

5. Conclusion

The present work studied two estimation models, based on different modelling methodologies, and applied them to estimate the heating consumption of school buildings in the north of Italy. The methods compared are: a multiple linear regression model and a classification and regression tree. While MLR model have been successfully applied in former works, data mining techniques, such as the decision tree, are a newly emerging analysis tool. The application of the decision tree to school buildings was demonstrably reliable in terms of the heating energy consumption estimation. The variance explained by both models is 86 %, but the decision tree shows lower errors, evaluated by means of the MAE, RMSE, and MAPE. Moreover, the gross heated volume, heat transfer surface, boiler size, and thermal transmittance of windows, were the parameters identified as primarily influencing the heating energy consumption among the considered school building stock.

The two methods are complimentary, not antagonistic, and show different strengths and weaknesses, as discussed in this paper. The greatest advantage of the CART is that the output consists of a set of practical decision rules that decision makers can quickly use. It also provides useful information on the influencing variables for each leaf node representing a sub-dataset, i.e. a homogenous class of school buildings.

The MLR model output consists of an equation including all of the major variables affecting heating energy consumption. Moreover, the partial standardised regression coefficients provide information on the most influencing input variables, making it possible to carry out a sensitivity analysis. Finally, the MLR model can be used to perform benchmark analyses.

Since the variability of the analysed sample is large enough to represent all school buildings in the north of Italy, the developed models can be used to estimate the heating energy requirements of new structures whose characteristics are within the ranges (for each variable) reported for the training dataset of the models. Similar models may moreover be developed, on the basis of other database, to estimate the heating energy requirements of different building types, since the model showed to be reliable and robust.

Future research should concern the possibility to couple the two models, in order to increase further the estimation capability. When the data of each final class of the CART are characterised by a large confidence interval, the performance of the model decreases. For each of these nodes is therefore possible to develop a MLR model, since each node is made by a sub-dataset of buildings with homogenous features. This combination may exploit the best characteristic of each of the two models; however, it requires new and lager training and testing database.

REFERENCES

[1]L. Perez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy and Building 40 (2008) 394-398.

[2] CRESME, RIUSO03: Ristrutturazione edilizia riqualificazione energetica rigenerazione urbana – estratto della ricerca CRESME, 2014 (Italian text).

[3]EU Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings.

[4]EU Directive 2010/31/EU of the European Parliament and of the Council of 19May 2010 on the energy performance of buildings.

[5]EU Directive 2012/27/EU of the European Parliament and of the Council of25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC.

[6]D. Fiaschi, R. Bandinelli, S. Conti, A case study for energy issues of public buildings and utilities in a small municipality: Investigation of possible improvements and integration with renewables, Appl. Energy. 97 (2012) 101–114.

[7]D. Radulovic, S. Skok, V. Kirincic, Energy efficiency public lighting management in the cities, Energy. 36 (2011) 1908–1915.

[8]US DOE-Energy Smart Schools, US Department of Energy, Energy Efficiency and Renewable Energy, Building Technologies Program. Available from: <http://www.eere.energy.gov/buildings/energysmartschools/resources.html>.

[9]L. De Santoli, F. Fraticelli, F. Fornari, C. Calice, Energy performance assessment and a retrofit strategies in public school buildings in Rome, Energy Build. 68 (2014) 196–202.

[10]A. Dimoudi, P. Kostarela, Energy monitoring and conservation potential in school buildings in the C' climatic zone of Greece, Renew. Energy. 34 (2009) 289–296.

[11]T.W. Kim, K.G. Lee, W.H. Hong, Energy consumption characteristics of the elementary schools in South Korea, Energy Build. 54 (2012) 480–489.

[12]S.P. Corgnati, V. Corrado, M. Filippi, A method for heating consumption assessment in existing buildings: A field survey concerning 120 Italian schools, Energy Build. 40 (2008) 801–809.

[13]S.P. Corgnati, F. Ariaudo, L. Rollino, Definizione di un indice semplificato per la previsione dei consumi per il riscaldamento di un patrimonio edilizio esistente a destinazione d'uso prevalentemente scolastica, III Congresso Nazionale AIGE, Parma 4-5 June 2009.

[14]F. Ariaudio, S.P. Corgnati, M. Filippi, Heating consumption assessment and forecast of existing buildings: investigation on Italian school Buildings, Proceedings of the 5th IBPC, KYOTO (2012) May 28-31.

[15]K.P. Amber, M.W. Aslam, S.K. Hussain, Electricity consumption forecasting models for administration buildings of the UK higher education sector, Energy Build. 90 (2015) 127–136.

[16]E. Beusker, C. Stoy, S.N. Pollalis, Estimation model and benchmarks for heating energy consumption of schools and sport facilities in Germany, Build. Environ. 49 (2012) 324–335.

[17]A. Thewes, S. Maas, F. Scholzen, D. Waldmann, A. Zürbes, Field study on the energy consumption of school buildings in Luxembourg, Energy Build. 68 (2014) 460–470.

[18]Z. Yu, F. Haghighat, B.C.M. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy Build. 42 (2010) 1637–1646.

[19]G.K.F. Tso, K.K.W. Yau, Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy. 32 (2007) 1761–1768.

[202]A.Z. Al-Garni, S.M. Zubair, J.S. Nizami, A regression model for electric-energy consumption forecasting in Eastern Saudi Arabia. Energy1 19 (1999) 1043-1049.

[21]A. Aranda, G. Ferreira, M.D. Mainar-Toledo, S. Scarpellini, E. Llera Sastresa, Multiple regression models to predict the annual energy consumption in the Spanish banking sector, Energy Build. 49 (2012) 380–387.

[22]I. Korolija, Y. Zhang, L. Marjanovic-Halburd, V.I. Hanby, Regression models for predicting UK office building energy consumption from heating and cooling demands, Energy Build. 59 (2013) 214–227.

[23]R.Z. Freire, G.H.C. Oliveira, N. Mendes, Development of regression equations for predicting energy and hygrothermal performance of buildings, Energy Build. 40 (2008) 810–820.

[24]J. Zhao, B. Lasternas, K.P. Lam, R. Yun, V. Loftness, Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining, Energy Build. 82 (2014) 341–355.

[25]R. Mikučionienė, V. Martinaitis, E. Keras, Evaluation of energy efficiency measures sustainability by decision tree method, Energy Build. 76 (2014) 64–71.

[26]L. Dias Pereira, D. Raimondo, S.P. Corgnati, M. Gameiro da Silva, Energy consumption in schools – A review paper, Renew. Sustain. Energy Rev. 40 (2014) 911–922.

[27]S.P. Corgnati,E. Fabrizio, F. Ariaudo, L. Rollino: Edifici tipo, indici di benchmark di consumo per tipologie di edificio, ad uso scolastico (medie superiori e istituti tecnici) applicabilità di tecnologie innovative nei diversi climi italiani, Report RSE/2010/190.

[28]N. Gaitani, C. Lehmann, M. Santamouris, G. Mihalakakou, P. Patargias, Using principal component and cluster analysis in the heating evaluation of the school building sector, Applied Energy. 87 (2010) 2079–2086.

[29]R.E. Walpole, R.H. Myers, S.L. Myers, K. Ye., Probability & Statistics for Engineers& Scientists, nineth ed., Academic Press, New York, 2011.

[30]M. Heidarinejad, M. Dahlhausen, S. Mcmahon, C. Pyke, J. Srebric, Cluster analysis of simulated energy use for LEED certified U . S. office buildings, Energy Build. 85 (2014) 86–97.

[31]F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy Build. 75 (2014) 109–118.

[32]J. Wakefield, Bayesian and Frequentist Regression Model, Spinger-Verlag, New York, 2012.

[33]S.J. Sheather, A Modern Approach to Regression with Regression Model, Spinger-Verlag, New York, 2009.

[34]C. Charu Aggarwal, Data Classification Algorithm and Application, Chapman & Hall/CRC, New York, 2014.

[35]L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, Chapman & Hall/CRC, New York, 1984.

[36]W. Chung, Y. V. Hui, Y.M. Lam, Benchmarking the energy efficiency of commercial buildings, Applied Energy. 83 (2006) 1–14.

Variables	Minimum	Maximum	Mean	Standard Deviation
Real Heating Degree Days [DD]	2537	3197	2701	161
Gross Heated Volume [m^3]	2900	86830	31464	20010
Heat Transfer Surface [m^2]	1905	25946	9866	5404
Aspect Ratio [m^{-1}]	0.20	0.80	0.35	0.10
Floor Heated Area [m^2]	1578	18254	6791	4073
Building Height [m]	5	28	14.50	4.50
Numbers of Floors [n°]	2	7	3	1.5
Thermal Trasmittance of Walls [$W/m^2 K$]	0.40	2.39	1.05	0.50
Thermal Trasmittance of Windows [$W/m^2 K$]	2.90	6.50	4.30	0.90
Boiler Size (Heat input) [kW]	106	8000	1755	1345
Number of Classrooms [n°]	6	65	27	14
Number of Pupils [n°]	75	1340	583	320
Annual Operating Time [h]	889	1888	1444	2450
Average Seasonal System Efficiency [%]	0.45	0.86	0.64	0.07

Tab 1 Statistical description of the variables influencing the heating energy consumption

Parameters	Limit Values	Calculated Values
z-score	< 3	No Values Higher
Mahalanobis Distance	27.87	No Values Higher
Index of Mardia	99	94.85
Cook Distance	1	0.29
Leverage Values	0.60	No Values Higher

Tab. 2 Indexes for the outliers detection analysis

Variables	VIF
Real Heating Degree Days [DD]	1.53
Gross Heated Volume [m^3]	24.28
Heat Transfer Surface [m^2]	19.25
Aspect Ratio [m^{-1}]	4.96
Building Height [m]	2.55
Numbers of Floors [n°]	2.46
Thermal Trasmittance of Walls [W/m^2 K]	1.54
Thermal Trasmittance of Windows [W/m^2 K]	1.13
Boiler Size (Heat input) [kW]	3.10
Number of Classrooms [n°]	35.90
Number of Pupils [n°]	36.75
Annual Operating Time [h]	2.02
Average Seasonal System Efficiency [%]	1.347

Tab. 3 Variance Inflation Factors (VIF)

Variables	Minimum	Maximum	Mean	Standard Deviation
Real Heating Degree Days [DD]	2537	3113	2696	159
Gross Heated Volume [m^3]	5065	78532	28745	17497
Heat Transfer Surface [m^2]	1905	24206	9354	5122
Thermal Trasmittance of Walls [$W/m^2 K$]	0.40	2.39	1.05	0.40
Thermal Trasmittance of Windows [$W/m^2 K$]	2.90	6.50	4.25	0.95
Boiler Size (Heat input) [kW]	141	3807	1424	845
Number of Pupils [n°]	115	1194	530	266
Annual Operating Time [n°]	889	1848	1426	250
Average Seasonal System Efficiency [%]	0.45	0.77	0.64	0.06

Tab 4 Statistical description of the variables included into MLR model

	Variables	β	b	t-value
	Intercepts	662757	-	3.82
X_1^*	Z: Real Heating Degree Days	16954	0.36	2.11
X_2^*	Z: Gross Heated Volume	203734	0.86	6.29
X_3^*	Z: Heat Transfer Surface	22609	-0.44	2.49
X_4^*	Z: Thermal Trasmittance of Walls	-12285	-0.39	-2.44
X_5^*	Z: Thermal Trasmittance of Windows	-24142	-0.61	-2.89
X_6^*	Z: Boiler Size (Heat input)	103167	-0.64	3.40
X_7^*	Z: Number of Pupils	55185	0.51	2.39
X_8^*	Z: Annual Operating Time	9940	0.03	0.35
X_9^*	Z: Average Seasonal System Efficiency	-1131	-0.04	-0.55

Tab. 5 Estimated coefficients（β） partial standardized regression coefficients (b) and t-values

N°	VARIABLE	NAME	TYPE	Unit of measure
1	Standard Energy Use Intensity	EUI_{ST}	Numerical	MWh
2	Heated Gross Volume	VOL	Numerical	m^3
3	Heat Transfer Surface	SUR	Numerical	m^2
4	Boiler Size (Heat input)	POW	Numerical	kW
5	Thermal Transmittance of Windows	$U_{windows}$	Numerical	W/m^2K

Tab 6 Variables selected in the CART

N° RULES	DECISION RULES
1	If VOL ≥ 33195 m^3 and SUR is < 12818 m^2 then EUI_{st} is 968 MWh
2	If VOL < 33195 m^3 and SUR is ≥ 12818 m^2 then EUI_{st} is 1183 MWh
3	If VOL < 33195 m^3 and SUR < 2460 m^2 then EUI_{st} is 140 MWh
4	If VOL < 33195 m^3 and SUR < 6203 m^2 and SUR ≥ 2460 m^2 and $U_{windows}$ is < 4.65 W/m^2K then EUI_{st} is 303 MWh
5	If VOL < 33195 m^3 and SUR < 6203 m^2 and SUR ≥ 2460 m^2 and $U_{windows}$ is ≥ 4.65 W/m^2K then EUI_{st} is 421 MWh
6	If VOL < 33195 m^3 and SUR is ≥ 6203 m^2 and $U_{windows}$ is < 4.54 W/m^2K then EUI_{st} is 521 MWh
7	If VOL < 33195 m^3 and SUR is ≥ 6203 m^2 and $U_{windows}$ is ≥ 4.54 W/m^2K and POW is < 1336 kW then EUI_{st} is 708 MWh
8	If VOL < 33195 m^3 and SUR is ≥ 6203 m^2 and $U_{windows}$ is ≥ 4.54 W/m^2K and POW is ≥1336 kW then EUIst is 816 MWh

Tab 7 Decision rules

RULES	ESTIMATED EUI_{st} [MWh]	UPPER CONFIDENCE LEVEL [MWh]	LOWER CONFIDENCE LEVEL [MWh]
1	968	1004	931
2	1183	1304	1061
3	140	214	56
4	303	328	279
5	421	472	369
6	521	604	439
7	708	774	641
8	816	832	801

Tab 8 Confidence interval of estimated EUI_{st} (CART)

INDEX	MLR MODEL	CART
MAE (MWh)	108	102
RMSE (MWh)	145	142
MAPE (%)	15	14

Tab 9 Error comparison for MLR model and CART

Fig.1 Framework of the research

Fig.2 Sample description: gross heated volume and external walls surface

Fig.3 Sample description: thermal transmittance of walls and windows

Fig.4 Sample description: boiler size (heat input) and average seasonal system efficiency

Fig.5 Sample description: heating energy consumption

Fig.6 Correlation coefficients

Fig. 7 Distribution plot between the monitored and the estimated EUI_{st} (MLR Model testing dataset)

Fig. 8. Distribution plot between the monitored and the estimated EUI_{st} (CART - testing dataset)

Fig. 9. Box-plots of monitored and estimated heating energy consumption

Figure 1
Click here to download high resolution image

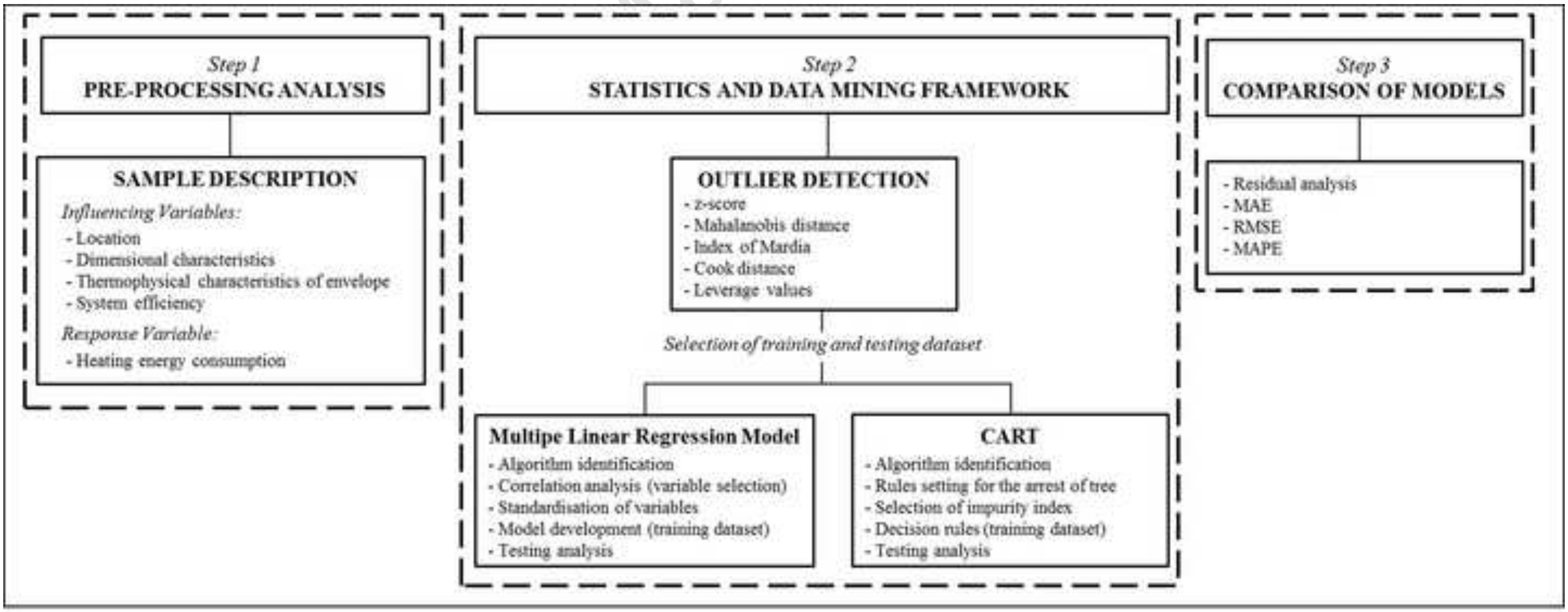


Figure 2
Click here to download high resolution image

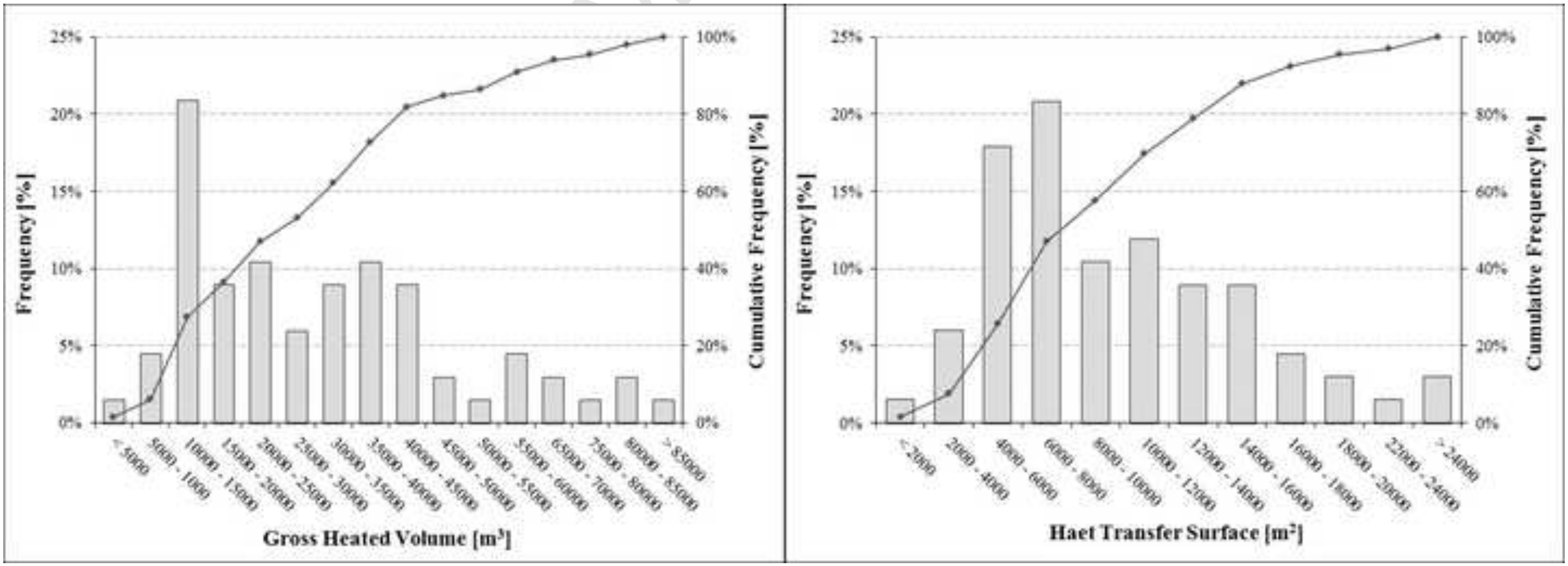


Figure 3
Click here to download high resolution image

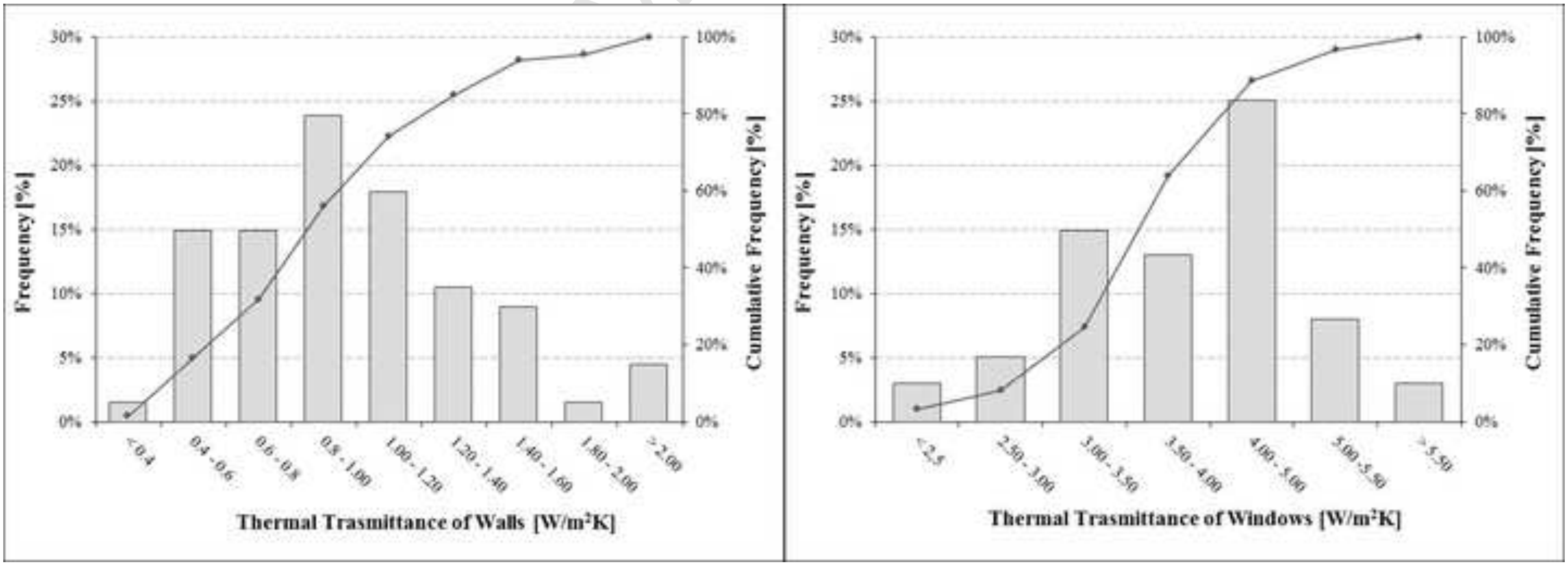


Figure 4
Click here to download high resolution image

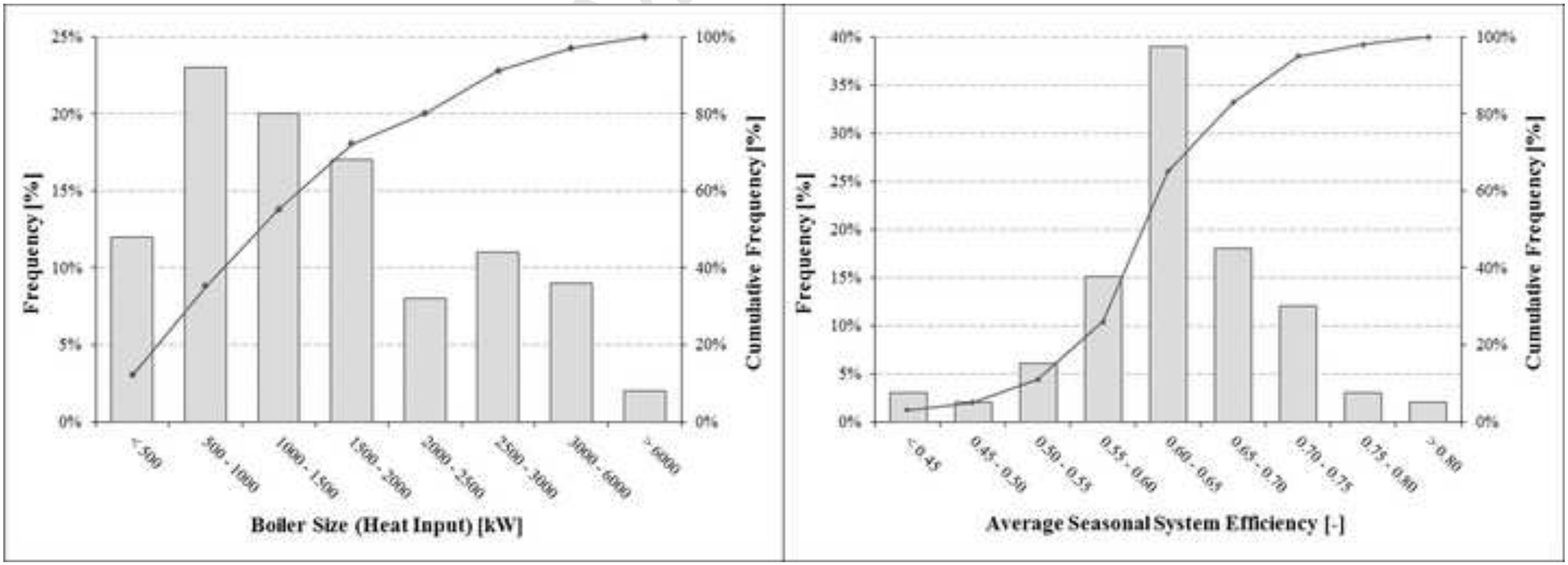


Figure 5
Click here to download high resolution image

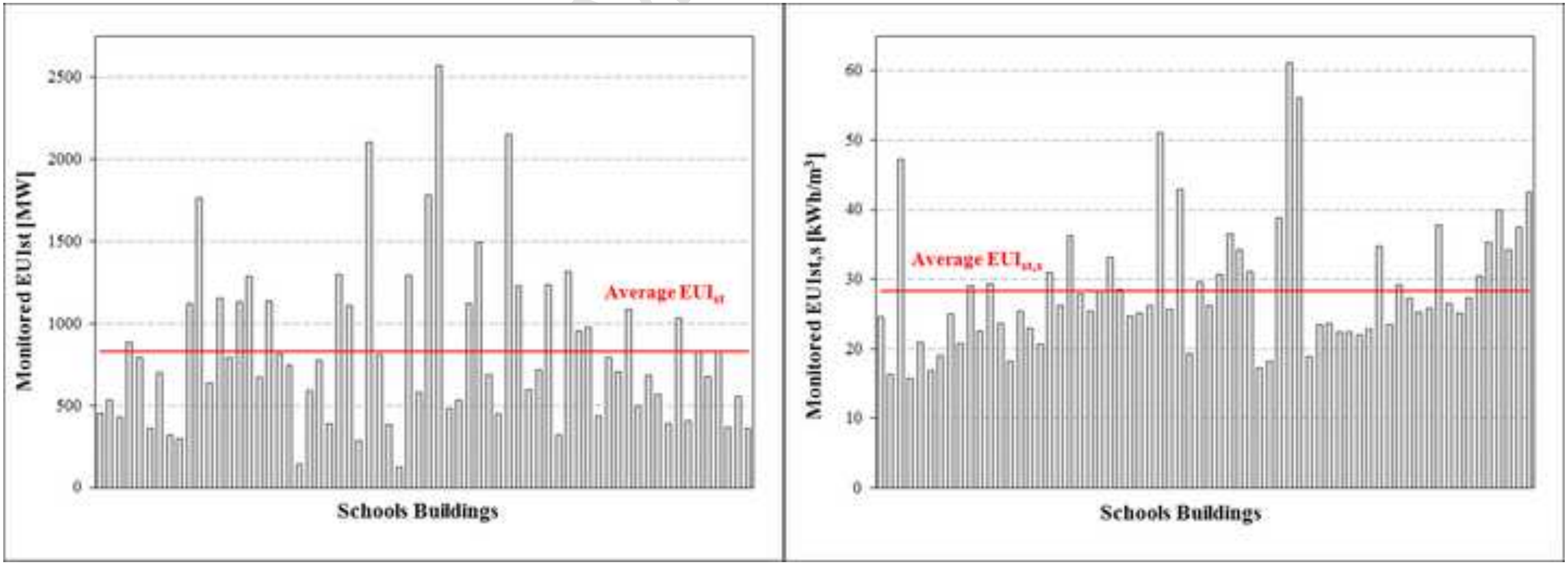


Figure 6
Click here to download high resolution image

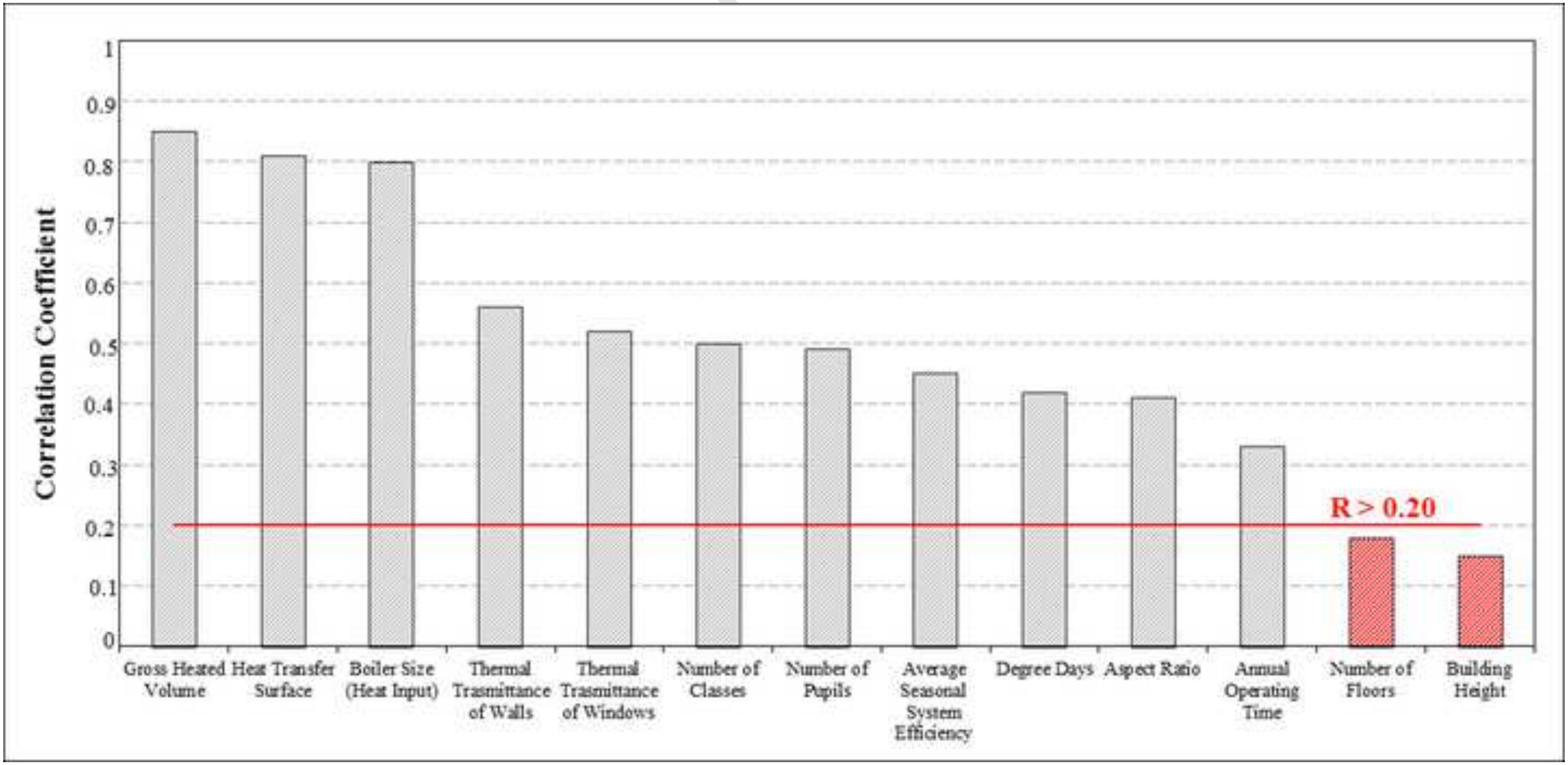


Figure 7
Click here to download high resolution image

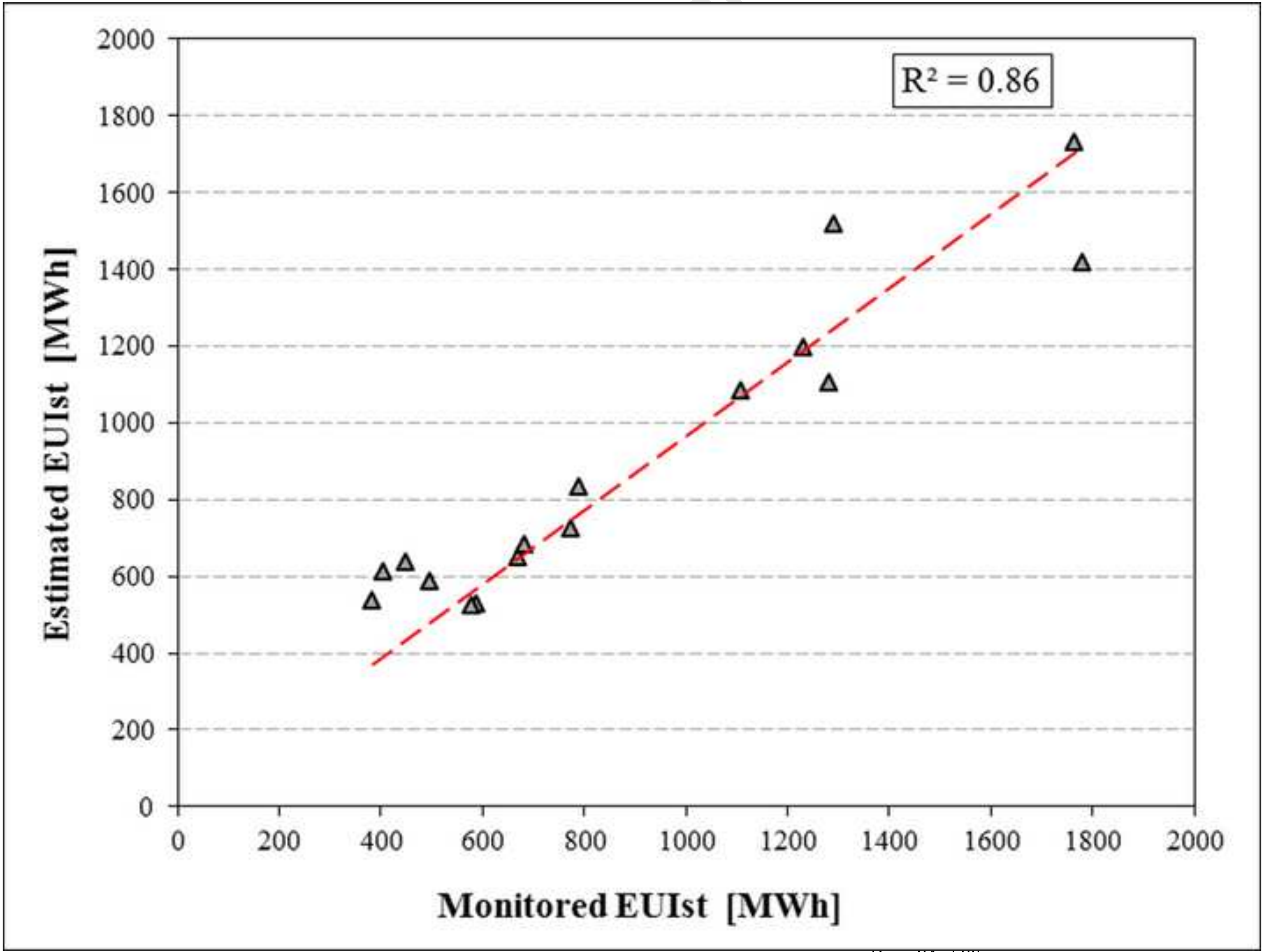


Figure 8
Click here to download high resolution image

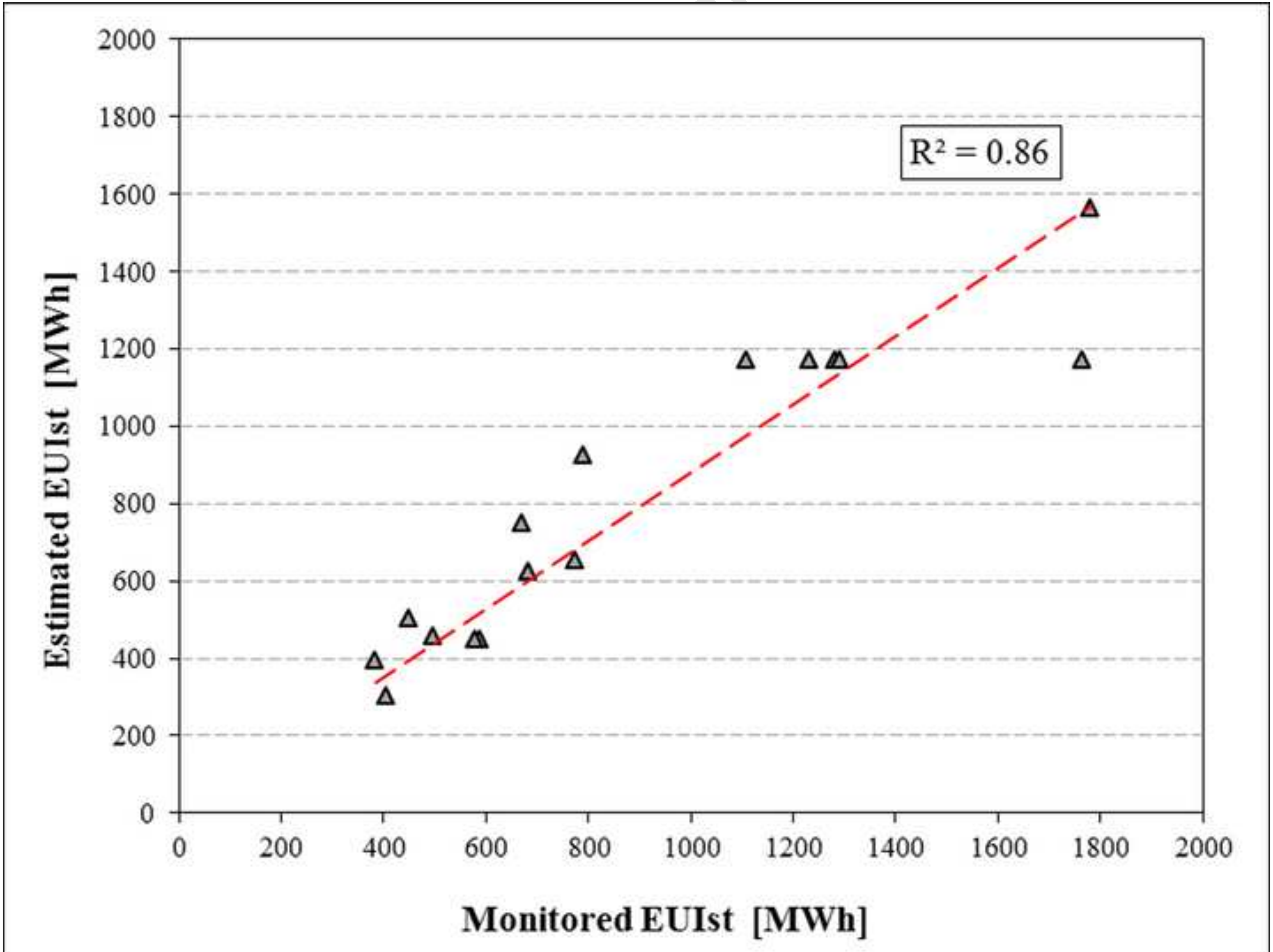


Figure 9
Click here to download high resolution image