Applied Energy 341 (2023) 121074

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Modelling energy performance of residential dwellings by using the MARS technique, SVM-based approach, MLP neural network and M5 model tree

Paulino José García Nieto [a,*], Esperanza García–Gonzalo [a], Beatriz María Paredes–Sánchez [b], José Pablo Paredes–Sánchez [b]

[a] *Department of Mathematics, Faculty of Sciences, University of Oviedo, 33007 Oviedo, Spain*
[b] *Department of Energy, College of Mining, Energy and Materials Engineering, University of Oviedo, 33004 Oviedo, Spain*

HIGHLIGHTS

- The energy performance of residential buildings (EPB) is analysed in detail.
- The developed MARS model has predicted satisfactorily the HL and CL.
- The MARS model was compared with SVM, MLP and M5 tree techniques.
- The correlation coefficient of the MARS-relied model is about 0.99.
- The physico-chemical input variables are studied in depth.

ARTICLE INFO

Keywords:
Multivariate adaptive regression splines (MARS)
Support vector machines (SVMs)
Artificial neural networks (ANNs)
M5 model tree
Energy performance at residential dwellings
Regression analysis

ABSTRACT

Several previous studies indicate that the energy consumption of buildings has increased steadily during the last decades all over the world. Residential dwellings in European countries are lawfully required to meet the suitable minimum needs concerning energy efficiency according to the European Directives. Specifically heating, ventilation and air conditioning (HVAC) devices represent most of the energy use in dwellings as they have a principal purpose in controlling the inner climate. Hence, one manner to relieve the constantly growing request for supplementary energy supply is to carry more efficient dwelling designs from the energy point of view, which is to say, with superior energy conservation properties. In this sense, an accurate estimation of the *heating load* (HL) and *cooling load* (CL) is needed to calculate the detailed descriptions of the heating and cooling device needed to support comfortable indoor air conditions. The goal of this investigation was to acquire some foretold models to achieve a tangible calculation of the HL and CL (output variables) as a function of 8 specific input variables (concretely, relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and glazing area distribution) at residential dwellings. These eight input factors have been often employed in the literature about the energy performance of dwellings (EPB) to analyse energy-related themes in dwellings. Moreover, a support vector machines (SVM) approach with distinct kernels, an artificial neural network (ANN) of multilayer perceptron network (MLP) kind and M5 model tree were adjusted to the observed data for evaluation of differences. The outcomes of the current investigation are two-fold. First, the importance (or strength) of each input variable on the HL and CL (output variables) is presented through the MARS model. Secondly, the MARS-relied approximation was the most excellent predictor of the EPB. Indeed, a MARS regression was conducted and coefficients of determination equal to 0.9961 for the HL estimation and 0.9651 for the CL estimation were gotten when this approach was employed to 768 diverse residential buildings, respectively. The concordance between the observed data and those predicted with the MARS approximation verified the satisfactory performance of the latter. Finally, the conclusions of this original investigation are summarised.

[*] Corresponding author.
E-mail address: pjgarcia@uniovi.es (P.J. García Nieto).

https://doi.org/10.1016/j.apenergy.2023.121074
Received 27 September 2022; Received in revised form 7 March 2023; Accepted 1 April 2023
Available online 19 April 2023

1. Introduction

Research studies so far propose that the energy consumption of buildings has increased continuously in recent decades around the world [1,2]. Dwellings in the European Union (EU) are lawfully required to comply with the appropriate minimum energy performance requirements since the European Directive 2002/91/EC [3], which was modified by Directives 2010/31 and 2018/844 of the EU [4,5]. Furthermore, the price of electricity and gas is prohibitive for most families currently. Therefore, energy saving and suitable use of energy sources for more efficient buildings are driving forces in the energy transition period [6,7]. In addition, heating, ventilation and air conditioning (HVAC) systems play an important function in controlling the inside climate since they represent most energy employed in dwellings [8,9].

Hence, one form to relieve the going up request for energy is to obtain designs in buildings with better properties of energy saving. To determine an energy-efficient dwelling pattern, the calculation of the Heating load (HL) and the Cooling load (CL) is needed throughout the year to obtain the requirements and performance conditions of the HVAV equipment necessary to support convenient inside air states [10]. Clearly, the external side climate gives place to an effect on the temperatures that it is experienced inside. Thus, the HVAC systems must work strongly in the greatest extreme climates for the purpose of keeping an adequate inside environment. The HL and CL describe the amount of heating or cooling demanded, respectively, from an advisable internal home temperature [11,12].

Currently, simulation tools in buildings for the calculation of energy evaluation are employed to a large extent to analyse or foretell energy usage rates and obtain the best performance and energy management. Moreover, although there are building simulation programs based on energy balance, the precision of the estimated outcomes can change between distinct classical software packages. In this sense, Atam [13] shows that advanced modelling for energy-efficient dwellings is essential for the best possible energy savings, which fervently relies on the existence of suitable simulation tools based on real data. For all these reasons, it is used in this investigation *machine learning techniques* (MLT) to analyse the effect of diverse kinds of construction factors (e.g., relative compactness) on the two output factors of interest (e.g., HL and CL) due to the fact that this is most rapid and simpler if it has available a database of the physical factors involved [14,15]. Indeed, the use of statistical machine learning techniques in the *energy performance of buildings* (EPB) has provoked a large quantity of concern lately [16,17].

The relevance of this procedure that employs the multivariate adaptive regression splines (MARS) approximation [18–22] to identify HL and CL factors in residential dwellings, to the knowledge of the authors, has not been intended for coming before research at this time. Furthermore, SVM–relied models with different kernels [23], an artificial neural network (ANN) of multilayer perceptron (MLP) type [24] and M5 model tree [25] were also adjusted to the observed dataset for the purpose of estimating HL and CL output variables for the evaluation of differences purposes. MARS is a nonparametric approximation to solve regression problems and can be seen as a generalization of the linear model with ability to tackle nonlinearities and the presence of factor interactions [17–22].

Preceding investigations make visible that MARS is an acting instrument in a large number of areas, such as bioinformatics, biomedicine, geochemistry, bioenergy and engineering areas of expertise [21,22,26,27]. Certainly, some reasons behind the utility of the suggested MARS approximation are as follows [18–22]: (1) MARS approximations are more tractable than linear regression techniques; (2) The MARS approximation originates always the same ensemble of basis functions when it is applied to the identical initial dataset; (3) MARS approximations are relatively easy to comprehend and understand; (4) MARS can deal with both categorical and continuous data; and (5) MARS approximations supply an explicit mathematical expression of the output dependent factor as a function of the input factors through a summation of basis functions (either hinge functions or products of two or more hinge functions). This most recent characteristic is an essential disagreement in comparison with other choices because the majority acts properly like a black box. The main disadvantage of MARS models is that the resulting fitted function is not smooth (not differentiable along hinges), although this is not an obstacle so that they can make more precise and faster predictions than other methods based on statistical learning.

The principal objective of the current investigation is to evaluate the application of several machine learning techniques (specifically, an optimised MARS–based model as well as SVM–based models with different kernels, MLP-type approximation and M5 model tree) to foretell the Heating load (HL) and Cooling load (CL) from the physical input parameters of the residential buildings in the context of EPB. To this end, it has been investigated the action of eight input factors (concretely relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and glazing area distribution (GAD)) to evaluate the HL and CL dependent variables in residential dwelling with success (see Fig. 1).

2. Materials and methods

2.1. Observed dataset

The observed dataset is based on previous measurements from Tsanas and Xifara [15]. It has taken an elementary cube $(3.5 \times 3.5 \times 3.5 \text{ m}^3)$, generating 12 dwelling shapes where each dwelling shape is defined by 18 elementary cubes. All the dwellings have identical volume (771.75 m^3), but distinct surface areas and dimensions. The materials employed for all 12 dwelling shapes are identical [28]. The selection was based on the common and newest materials employed in the dwelling industry jointly with the lowest U-value (specifically, the associated U-values taken here are: walls (1.78), roofs (0.5) and windows (2.26)).

The modelling supposes that the residential dwellings are in the Mediterranean climate region [15,28]. It is used for 7 persons with little activity (~70 W). The inner design states were established in the following manner:

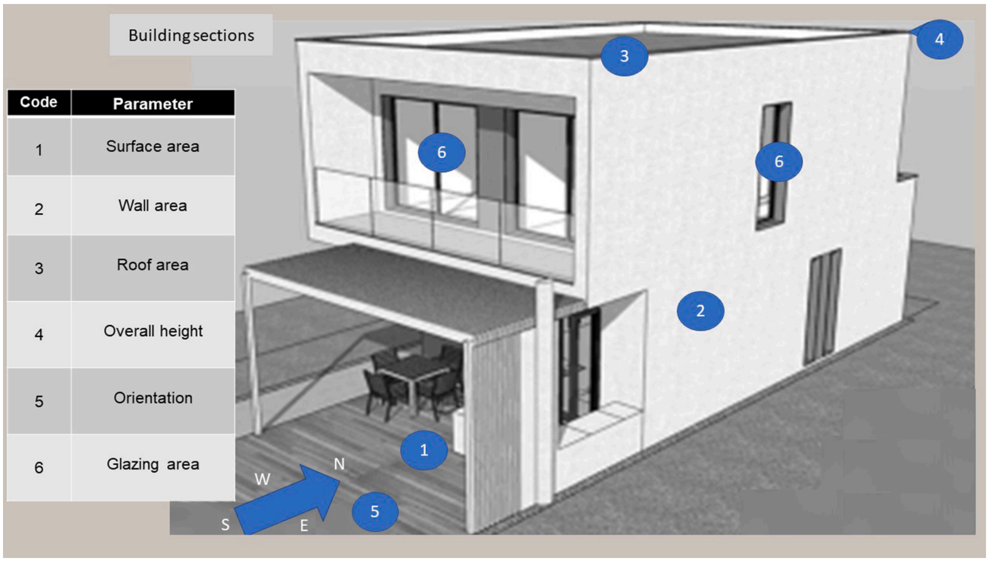
- level of light: 300 lx;
- speed of air: 0.30 m/s; and
- relative humidity: 60 %.

With respect to the thermal properties, it was employed the following combined manner: 95 % efficiency and thermostat temperatures ranging from 19 to 24 °C. The number of operation hours was ranging from 15 to 20 h on weekdays and from 10 to 20 h on weekends.

2.2. Variables involved in the problem

The principal goal of this investigation was to obtain the HL and CL factors (output variables) relied on the eight physical input variables (see Table 1). It is commonly accepted that the method of EPB entails a large number of variables. The data knowledge employed for the distinct models (i.e., MARS–relied approximation, SVM–relied method with distinct kernels, MLP–relied method and M5 model tree) relies on these physical factors for modelling. Next, we are going to define the eight physical input variables employed in this investigation:

- Relative compactness (RC) (%): it is a numeric input variable. The relative compactness of the shape is obtained by comparing its ratio between volume and area with that of the most compact shape with the same volume [28]. RC relies exclusively on the shape, unlike the traditional compactness indications as the characteristic length. This



Code	Parameter
1	Surface area
2	Wall area
3	Roof area
4	Overall height
5	Orientation
6	Glazing area

Fig. 1. Picture of a residential dwelling and the factors analysed in it as an energy system.

Table 1
The physical input and output factors used in this investigation of buildings and their means and standard deviations.

Input variables	Name of the variable	Number of possible values	Mean	Standard deviation
Relative compactness (%)	RC	12	0.7642	0.1058
Surface area (m^2)	SA	12	671.7083	88.0861
Wall area (m^2)	WA	7	318.5	43.6265
Roof area (m^2)	RA	4	176.6042	45.166
Overall height (m)	OH	2	5.25	1.7511
Orientation	O	4	–	–
Glazing area (m^2)	GA	4	0.2344	0.1332
Glazing area distribution	GAD	6	2.8125	1.551
Output variables				
Heating load (kWh/m^2)	HL	768	22.3072	10.0902
Cooling load (kWh/m^2)	CL	768	24.5878	9.5133

last length relies on the form size calculated as the volume absolute value. Since its principal morphological sample entails forms of identical volume, it prefers here to use RC, because it describes better the individual perception of form compactness for the design professionals.

- Surface area (m^2): it is a numeric variable. The surface area of the solid object is a measure of the entire area that the surface of the building fills. Therefore, the surface area is a 2D-object, which is the base area of the building envelope.
- Wall area (m^2): it is a numeric parameter. It is the surface area of the inner walls in a rectangular space based on the length, width and height of the room in the considered building.
- Roof area (m^2): it is a numerical value. It is the approximate flat area of the built ceiling in a building.
- Overall height (m): it is a numeric variable. Generally, the overall height of a structure is measured to the peak of the building (i.e., the highest roof point). The higher the roof pitch, the taller the peak reaches.
- Orientation: it is a categorical datum. This input variable shows the direction of the building structure referred to the cardinal points.
- Glazing area (m^2): it is a numerical variable. The word *glazing* alludes to the glass part of a dwelling façade or inner surfaces. All glazed areas of a building include windows, sliding glass doors, glass doors, and skylights, among others.
- Glazing area distribution: This input variable refers to the five distinct scenarios of distribution for each glazing area considered in this research.

Hence, it has been employed three kinds of glazing areas (GAs) stated in form of percentages of the floor area. In this investigation, we have used 10 %, 25 %, and 40 %, respectively. Moreover, five distinct possible situations of distribution in each glazing area were simulated termed as:

(1) Uniform: each face has a 25 % glazing area;
(2) North: it indicates 55 % glazing area on the north face and 15 % on each of the other faces;
(3) East: it indicates 55 % glazing area on the east face and 15 % on each of the other faces;
(4) South: it indicates 55 % glazing area on the south face and 15 % on each of the other faces; and
(5) West: it indicates 55 % glazing area on the west face and 15 % on each of the other faces.

As well, samples without glazing areas are considered here. Conclusively, all forms were turned to look towards the four cardinal points.

Hence, taking into account 12 dwelling shapes and 3 variants of glazing area with 5 glazing area distributions each, for the four cardinal points (4 additional orientations), it obtains $12 \times 3 \times 5 \times 4 = 720$ dwelling samples. Moreover, it is considered 12 dwelling forms for the 4 cardinal points without glazing. To summarize, it has been studied in total $720 + 12 \times 4 = 768$ shapes of buildings (see Appendix A). Finally, for each of the 768 dwellings, it records HL and CL factors (see Table 1).

2.3. *Computational procedures*

2.3.1. *Multivariate adaptive regression spline (MARS)*

Multivariate adaptive regression splines (MARS) [18–20] is a flexible nonparametric methodology that permits to tackle of regression problems. This method is a generalization of both CART decision trees [16] and *SL* (*stepwise linear*) regression and, but able to overcome its limitations. Its principal main is the foretelling of the values of a continuous output (dependent) variable,$\boldsymbol{y}(n \times 1)$, from an ensemble of independent input variables,$\boldsymbol{X}(n \times p)$. The MARS technique is given by the following expression:

$$\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{e} \tag{1}$$

Here:

- f: it is a weighted sum of basis functions that depend on $\boldsymbol{X}$ and;
- $\boldsymbol{e}$: it is the error vector whose dimension is $(n \times 1)$.

MARS methodology does not demand a priori suppositions about the preexisting functional relationship among the dependent variable and independent variables. In this way, its mathematical expression is defined by means of a collection of piecewise polynomials of degree q (basis functions) with its coefficients derived completely from the entire regression dataset $(\boldsymbol{X}, \boldsymbol{y})$. The MARS approach is created by means of the fitting of basis functions to distinct intervals of the independent variables. Certainly, MARS employs as splines basis functions two-sided truncated power functions, whose equations are as follows [18–20]:

$$[-(x-t)]_+^q = \begin{cases} (t-x)^q & \textit{if } x < t \\ 0 & \textit{otherwise} \end{cases} \tag{2}$$

$$[+(x-t)]_+^q = \begin{cases} (t-x)^q & \textit{if } x \geqslant t \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

so that $q\,(\geqslant 0)$ is the power that defines the type of splines (linear, quadratic or cubic) and therefore the flatness degree of the consequential function estimate. On this matter, the case $q = 1$ corresponds to simple linear splines (the case of this investigation), $q = 2$ to quadratic splines, $q = 3$ cubic splines and so on. For instance, Fig. 2 shows a couple of splines for $q = 1$ at the knot $t = 3.5$.

As a result, the MARS approximation is a set of piecewise linear multivariate splines of a dependent variable $\boldsymbol{y}$ using M basis functions that respond to the following expression [21,29–32]:

$$\widehat{\boldsymbol{y}} = \widehat{f}_M(\boldsymbol{x}) = c_0 + \sum_{m=1}^{M} c_m B_m(\boldsymbol{x}) \tag{4}$$

In Eq. (4), we have that:

- $\widehat{\boldsymbol{y}}$: is the output-dependent variable prognosticated by using the MARS approximation;
- c_0: is a coefficient that remains constant termed *intercept*;
- $B_m(\boldsymbol{x})$: it is the m-th basis function; and
- c_m: it is the corresponding coefficient of this $B_m(\boldsymbol{x})$ basis function.

Moreover, MARS employs the generalised cross-validation (GCV) [29–32] to determine the basis functions that constitute the entire approximation. Certainly, the GCV is defined as the quotient between the mean squared residual error and a penalization element. This last factor is related to the model complexity. The mathematical expression of the GCV is as follows [18–21,29–32]:

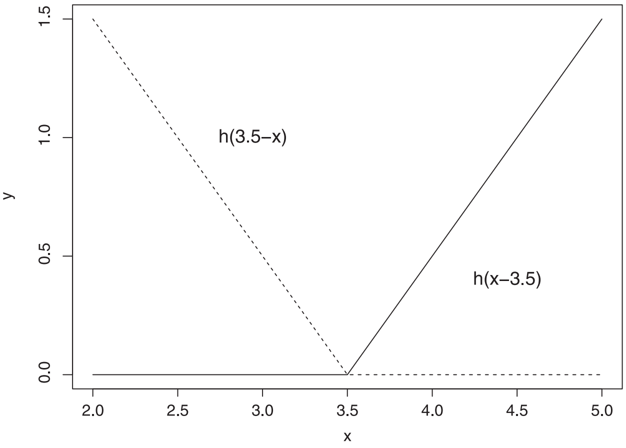


Fig. 2. Picture of a couple of linear hinge functions (spline basis functions). Dashed line represents the left spline $(x < t, -(x-t))$ while the solid line indicates the right spline $(x > t, +(x-t))$.

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}_M(\boldsymbol{x}_i)\right)^2}{\left(1 - C(M)/n\right)^2} \tag{5}$$

so that:

- $C(M)$ is a complexity penalization term that increases when the number of basis functions of the MARS model grows. This factor has the form [19,20,29]:

$$C(M) = (M+1) + dM \tag{6}$$

In such a way that:

- M is the number of basis functions in Eq. (5); and
- d is a penalization factor for each basis function defined by the model.

Likewise, it is obligatory the employ of the GCV factor described above as well as the factors N-subsets (this criterion is related to the number of subsets of the model in which each variable is integrated) and the residual sum of squares RSS [21,29–32] for the purpose of determining reliable outcomes.

2.3.2. Support vector regression (SVR) method

In this subsection, it will study the use of support vector machines (SVMs) to find a solution for regression problems. In these cases, they are usually called Support Vector Regression (SVR) [16,23,33,34]. Take into account a collection of training examples $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i \in \Re^d$ and $y_i \in \Re$, assuming that the y_i values of all examples of S can be fitted (or quasi-adjusted) by a hyperplane, the goal is to encounter the parameters $\boldsymbol{w} = (w_1, \ldots, w_d)$ that permit to describe mathematically the regression hyperplane $f(\boldsymbol{x}) = (w_1 x_1 + w_2 x_2 + \ldots + w_d x_d) + b = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$.

We define the random noise or disturbance $\varepsilon \sim N(\boldsymbol{0}, \sigma)$ as the measurement error of the value y, that is,$y = f(\boldsymbol{x}) + \varepsilon$. To allow for some noise in the training examples, it can relax the error condition between the value foretold by this function and the observed value. For this, the $\varepsilon -$ insensitive loss function is used,L_ε, given by [16,23]:

$$L_\varepsilon(\boldsymbol{x}) = \left\{ \begin{array}{cc} 0 & \text{if}|y - f(\boldsymbol{x})| \leqslant \varepsilon \\ |y - f(\boldsymbol{x})| - \varepsilon & \text{otherwise} \end{array} \right\} \tag{7}$$

It is a linear function with an insensitive zone of width 2ε, in which the loss function takes a null value. By choosing this function, it allows some flexibility in the solution function, so that all the examples that are confined to the tubular region will not be considered support vectors, since the cost associated with the loss function is 0. In practice, it is very difficult to achieve a linear regression model with zero prediction error, so it will resort to the concept of *soft margin.*

We define the slack variables as the distance to the sample measured from the tubular zone of the regression hyperplane. The slack variables ξ_i^+ and ξ_i^- will allow us to quantify the prediction error that it is willing to admit for each training example and, with the sum of all of them, the cost associated with the examples with a non-zero prediction error. It will take $\xi_i^+ > 0$ when the prediction of the example $f(\boldsymbol{x}_i)$ is greater than its actual value,y_i, in an amount greater than ε or equivalently $f(\boldsymbol{x}_i) - y_i > \varepsilon$. Similarly $\xi_i^- > 0$ when the actual value of the example is greater than its prediction by an amount greater than ε, that is,$y_i - f(\boldsymbol{x}_i) > \varepsilon$. In any other case, the slack variables take a value of zero. Note that both variables cannot simultaneously take a value other than zero, since it happens whenever $\xi_i^+ \cdot \xi_i^- = 0$(see Fig. 3).

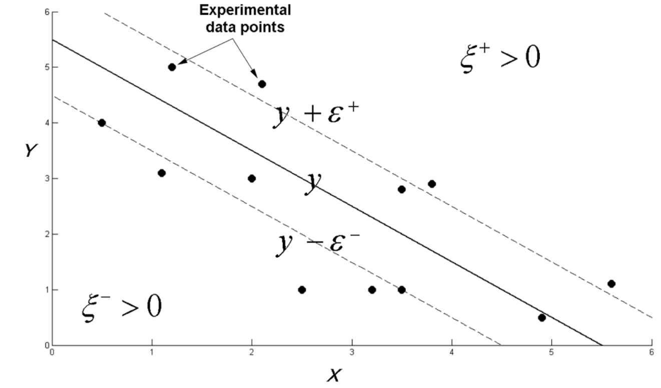


Fig. 3. An illustration of the $\varepsilon -$ insensitive tube in the event of regression.

With all this, it can now pose the problem to be optimised. The goal is to minimise the sum of the associated loss functions, each one to an example of the training set $\sum_{i=1}^{n} L_\varepsilon(y_i, f(x_i)) = \sum_{i \in \text{non - tubular zone}} |y_i - f(x_i)| - \varepsilon$. This is equivalent to maximizing the tubular zone defined by the loss function, in which it takes a null value. Therefore maximizing ε is equivalent to minimizing $\|\boldsymbol{w}\|$. All this together with the penalty imposed by the slack variables define the following optimization problem with soft margin, so that C is called the *regularisation constant* [16,23]:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}^+,\boldsymbol{\xi}^-} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\left(\xi_i^+ + \xi_i^-\right)$$

subject to

$$\left\{ \begin{array}{ll} (\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - y_i - \varepsilon - \xi_i^+ \leqslant 0 & i = 1, \ldots, n \\ y_i - (\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - \varepsilon - \xi_i^- \leqslant 0 & i = 1, \ldots, n \\ \xi_i^+, \xi_i^- \geqslant 0 & i = 1, \ldots, n \end{array} \right\} \tag{8}$$

After, the transformation to the dual problem with four families of Lagrange multipliers $(\alpha_i^+, \alpha_i^-, \beta_i^+, \beta_i^-)$ is conducted [33,34]:

$$\max_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-} \sum_{i=1}^{n}\left(\alpha_i^- - \alpha_i^+\right)y_i - \varepsilon\sum_{i=1}^{n}\left(\alpha_i^- + \alpha_i^+\right) - \frac{1}{2}\sum_{i,j=1}^{n}\left(\alpha_i^- - \alpha_i^+\right)\left(\alpha_j^- - \alpha_j^+\right)\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

subject to

$$\left\{ \begin{array}{ll} \sum_{i=1}^{n}\left(\alpha_i^+ - \alpha_i^-\right) = 0, & \text{with} \\ 0 \leqslant \alpha_i^+, \alpha_i^- \leqslant C & i = 1, \ldots, n \end{array} \right\} \tag{9}$$

The obtained regressor is [16,23,33,34]:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}\left(\alpha_i^- - \alpha_i^+\right)\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + b^* \tag{10}$$

The optimal value b^* is obtained from the restrictions resulting from the application of the second *Karush-Kuhn-Tucker* (KKT) condition and the restrictions on the dual problem, so that [16,23]:

$$\left\{ \begin{array}{ll} b^* = y_i - \langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + \varepsilon & \text{if} \quad 0 < \alpha_i^+ < C \\ b^* = y_i - \langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle - \varepsilon & \text{if} \quad 0 < \alpha_i^- < C \end{array} \right\} \tag{11}$$

Note that to define the regression hyperplane it considers the examples with a non-zero loss function, that is, those that are outside the tubular region. Viewed in terms of the parameters introduced above, for the support vectors it gathers from the Karush-Kuhn-Tucker (KKT) conditions that $\alpha_i^+ \cdot \alpha_i^- = 0$, so [16,23]:

- for the examples that are outside the tubular zone it will be fulfilled $\xi_i^+ \cdot \xi_i^- = 0$, if $\xi_i^- = 0$ and $\xi_i^+ > 0$, then $\alpha_i^+ = C$ and $\alpha_i^- = 0$; and if $\xi_i^- > 0$ and $\xi_i^+ = 0$, then $\alpha_i^- = C$ and $\alpha_i^+ = 0$;
- the support vectors that are just into the border of the sensitivity zone verify that if $0 < \alpha_i^+ < C$, then $\alpha_i^- = 0$. In that case, it must be $\xi_i^+ = 0$ and $\xi_i^- = 0$. Similarly for the other case.

The examples which $\alpha_i^+ = \alpha_i^- = 0$(are not considered support vectors) are found within the tubular region.

When the examples cannot be fitted by a linear function (nonlinear problems), it resorts to using *kernel functions*. Through a suitable kernel, a Hilbert space is induced, also called a feature space, in this it is possible to adjust the transformed examples using a linear regressor, which has the following expression [33,34]:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}\left(\alpha_i^- - \alpha_i^+\right)K(\boldsymbol{x}, \boldsymbol{x}_i) \tag{12}$$

Now the coefficients are obtained solving the dual problem that results from Eq. (9) with dot products substituted for kernel functions given by [16,23]:

$$\max_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-} \sum_{i=1}^{n}\left(\alpha_i^- - \alpha_i^+\right)y_i - \varepsilon\sum_{i=1}^{n}\left(\alpha_i^- + \alpha_i^+\right) - \frac{1}{2}\sum_{i,j=1}^{n}\left(\alpha_i^- - \alpha_i^+\right)\left(\alpha_j^- - \alpha_j^+\right)K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

subject to

$$\left\{ \begin{array}{ll} \sum_{i=1}^{n}\left(\alpha_i^+ - \alpha_i^-\right) = 0, & \text{with} \\ 0 \leqslant \alpha_i^+, \alpha_i^- \leqslant C & i = 1, \ldots, n \end{array} \right\} \tag{13}$$

To solve regression problems using SVRs, it must choose a suitable kernel and a C parameter as well as the selection of a suitable ε. The value of the parameter C expresses the balance between the flatness of the objective function and the decrease of the model complexity [16,23]. In the case of noisy regression problems, the parameter ε should be selected to express the variance of the noise in the data, since in most practical cases it is possible to obtain an approximate measure of the noise variance from the training data. The methodology employed to choose the optimal values of C and the rest of the kernel parameters is normally based on cross-validation techniques.

Several frequent functions used as kernels in the research publications are given by [16,23,33,34]:

- Polynomial kernel:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sigma \boldsymbol{x}_i \cdot \boldsymbol{x}_j + a\right)^b \tag{14}$$

- RBF (radial basis function) kernel:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\sigma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2} \tag{15}$$

where a, b and σ are the kernel hyperparameters.

Hence, to find the solution of a complicated regression problem like this, it has used the SVM technique here with data that is not linearly separable. To this end, it is mandatory to choice a kernel type along with its optimal parameters so that these data become linearly separable in a space of higher dimension known as *feature space*.

2.3.3. Neural network: Multilayer perceptron

Minsky and Papert showed in 1969 [24] that the simple perceptron and ADALINE (adaptative linear element) cannot solve nonlinear problems (for example, XOR). The combination of several simple perceptrons could solve certain nonlinear problems, but there was no automatic mechanism to adapt the weights of the hidden layer. Rumelhart and other authors presented in 1986 [35] the *Generalised*

Delta Rule (GDL) to adapt the weights by propagating the errors backwards, that is, propagating the errors towards the lower hidden layers. In this way, it is possible to work with multiple layers and with nonlinear activation functions. It can be shown that this multilayer perceptron (MLP) is a universal approximator [35,36]. A multilayer perceptron can approximate nonlinear relationships present between input and output data. This ANN has become one of the most common architectures.

The MLP is kind of artificial neural network (ANN) composed of multiple layers, in such a way that it can encounter solutions to nonlinearly separable problems [35]. This matter is the foremost limitation of the simple perceptron. However, MLP can be fully or locally connected. To be fully connected all the neurons of a layer must be attached with all the neurons of the next layer while this condition is not present in a locally connected MPL.

The layers of an MLP can be classified into three types (see Fig. 4) [35,36]:

- Input layer: the information of the independent variables enters through this layer and there is no process here.
- Output layer: the connection with the dependent variables is made here.
- Hidden layers: are layers located between the input and output layers that pass and process the information from the input to the output layers.

Backpropagation (also known as error backpropagation or *generalised delta rule*) is the mathematical rule to train this type of neural networks [36]. In this sense, MLP is also termed as a backpropagation artificial neural network (BP-ANN). Additionally, the main quality of this kind of networks is that the transfer functions of the processing elements (neurons) must be *derivable*.

This kind of learning happens in the multilayer perceptron (MLP) by altering the weights of the connections considering the disagreement between the expected and the obtained output values. This change is performed using backpropagation which is a generalization of the lowest mean square (LMS) used on the linear perceptron. For data point n the error at node j is $e_j(n) = d_j(n) - y_j(n)$, being d the observed value and y the value predicted by the multilayer perceptron. The total error to correct is [35,36]:

$$\varepsilon(n) = \frac{1}{2}\sum_j e_j^2(n) \tag{17}$$

Employing the *gradient descent method*, we find that the change of the weights is given by [35,36]:

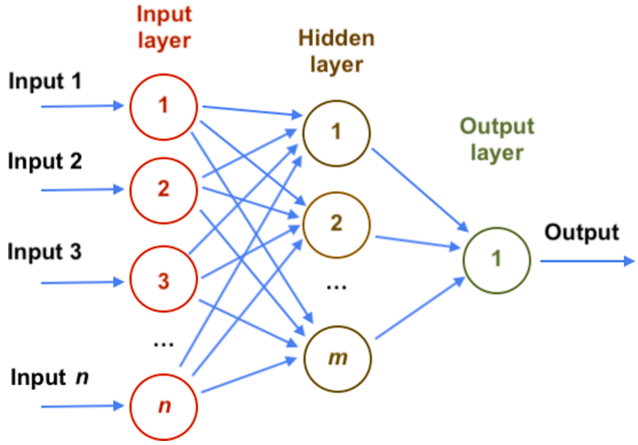


Fig. 4. Picture of an MLP-type artificial neural network (ANN) (in this case, picture shows m neurons in the hidden layer, n neurons in the input layer and a single neuron in the output layer).

$$\Delta w_{ji}(n) = -\eta \frac{\partial\varepsilon(n)}{\partial v_j(n)} y_i(n) \tag{18}$$

where:

- η is the *learning rate.* It must be chosen carefully: a small value produces a slow convergence while a big value can hamper the convergence of the optimization. Adequate values range from 0.1 to 0.8; and
- y_i is the output obtained from the neuron in the previous layer.
- v_j is the local induced field. It can be proved that for a given output node:

$$-\frac{\partial\varepsilon(n)}{\partial v_j(n)} = e_j(n)\cdot\phi'\left(v_j(n)\right) \tag{19}$$

being ϕ' the derivative of the activation function. The variation of the weights for the nodes of the hidden layer is given by:

$$-\frac{\partial\varepsilon(n)}{\partial v_j(n)} = \phi'\left(v_j(n)\right)\sum_k -\frac{\partial\varepsilon(n)}{\partial v_k(n)} w_{kj}(n) \tag{20}$$

k is the subscript of the nodes from the output layer and these nodes affect the change of the weights of the hidden layer. We start changing the weights of the output layer taking into account the derivative of the activation function and then this process backpropagates modifying the weights of the previous layers.

2.3.4. M5 Model tree

This approximation is also relied on the machine learning and it employed the following an inspired thought [25,37,38]: the parameter space can be divided into subspaces and after a linear regression approach is constructed in each of them. The consequential approximation would be considered a modular method, so that the linear fits specialise in the specific subsets of the input space.

The mathematical technique termed algorithm M5 is employed to force a model tree [37–40]. Indeed, a group of T training data is considered here. Each instance is depicted by the values of a not variable collection of input attributes as well as a related goal output value. The principal goal is to build a method that connects an objective value of the training data with their input attribute values. The model excellence will usually be assessed if it foretells the objective values of the unknown cases accurately.

The method used to build tree-based machine learning models is *divide-and-conquer* [41–43]. The set T is connected to a leaf or several tests are selected to divide T into subsets. This splitting algorithm is applied recursively. The division criterion used by the M5 model tree technique makes use of the standard deviation of the class values arriving at a node for the purpose of measuring the error at that node and then the calculation of the anticipated reduction of this error to check every attribute in that node. Certainly, the lowering of the standard deviation (SDR) can be determined by using the following expression [25,37–43]:

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \tag{21}$$

where T indicates the number of examples arriving at the node, T_i signifies the subset of cases that have on the ith outcome of the potential collection, and sd is the standard deviation [25,37–43].

Next a thorough examination of all possible divisions, the M5 model tree choices the element that fully improves the anticipated error lowering [40–43]. This M5 model tree splitting mechanism ends when the class values of all instances arriving at a node disagree by only a very small tolerance (stopping criterion), or else when only a few instances remain. This persistent splitting process often gives place to much elaborated frameworks that must be pruned, i.e. replacing a subtree by a leaf. With time, it is necessary to carry out a smoothing process to counterbalance for the abrupt discontinuities that will predictably happen among contiguous linear models at the leaves of the pruned tree, in particular for several models built from a lower number of training data. During this procedure, the contiguous linear equations are upgraded so that the foretold outputs for the contiguous input vectors related to the distinct equations are transformed very close in their expressions.

2.4. The goodness–of–fit of this approach

The main goodness–of–fit statistics for the regression problem posed in this paper is the coefficient of determination R^2 [44,45]. If the observed and predicted values are t_i and y_i, respectively, it considers the following expressions [44,45]:

- $SS_{reg} = \sum_{i=1}^{n} (y_i - \bar{t})^2$: is the explained sum of squares.
- $SS_{tot} = \sum_{i=1}^{n} (t_i - \bar{t})^2$: this addition is directly related to the variance of the sample.
- $SS_{err} = \sum_{i=1}^{n} (t_i - y_i)^2$: is the residual sum of squares.

so that $\bar{t}$ is the average value of the experimental data given by:

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{22}$$

The coefficient of determination is then defined by the expression [44,45]:

$$R^2 \equiv 1 - \frac{SS_{err}}{SS_{tot}} \tag{23}$$

The closer the R^2 statistic is to the value 1.0, the smaller the disagreement between the experimental and foretold data. Similarly, the mathematical expressions for the other two statistics used in this study (*RMSE* and *MAE*) are as follows [44,45]:

$$RMSE \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n} (t_i - y_i)^2} \tag{29}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |t_i - y_i| \tag{30}$$

Higher values of R^2 are preferred, i.e. closer to 1 means better model performance and regression line fits the data well. Conversely, the lower the *RMSE* and *MAE* values are, the better the model performs.

As well, the MARS approximation relies fervently on the following hyperparameters [18–22]:

- Maximum number of basis functions (Maxfuncs): it corresponds to the maximum number of terms when the forward phase finishes.
- Penalty parameter (d): it corresponds to the penalization associated to the complexity of the MARS approximation in Eq. (5) of the Generalised Cross Validation (*GCV*) penalty per knot. Observe that if $d = 0$ penalizes terms, but not knots. Certainly, if $d = -1$, there is no penalisation;
- Interactions: it corresponds to the greatest degree of interaction among factors.

Right now, it has been built distinct approximations (concretely in this investigation, the MARS–relied approximation, SVM–relied approach with distinct kernels, MLP-type ANN approximation and M5 model tree) taking as dependent factors the HL and CL from the 8 remaining factors (input variables) in residential buildings [15], analysing their efficacy for the purpose of optimizing its estimation employing of the coefficient of determination R^2 successfully.

Moreover, as earlier referred to the MARS approximation, is dependent very much on the MARS hyperparameters: greatest number of hinge functions (Maxfuncs); penalization parameter (d); and in the end the interaction degree among variables. The customary manner of enacting hyperparameter optimization is known as *grid search*, or *parameter sweeping*, which is plainly a thorough seeking by hand through a stated subset of the parameter space employing that statistical machine learning algorithm. In this investigation, it has employed the grid search method [46,47] with success.

Indeed, the dataset is divided into two sets: 80 % is used in the training set and the rest of the data, 20 %, is for the testing set. In this sense, it employs the training collection to construct the MARS model. For this purpose, it calibrates the parameters of the MARS model employing the grid search algorithm using a 10-fold cross-validation method. When the optimum parameters have been found, it built the model with the whole training dataset. Then, it proceeds to get predictions with this model for the elements of the testing set. These predictions are compared with the actual values and the goodness-of-fit of the model evaluated. At this point, Fig. 5 displays the process diagram of this MARS–relied approximation implemented in this investigation.

Furthermore, cross-validation is also the common method utilised here for encountering the authentic coefficient of determination (R^2) [44,45,48]. Certainly, for the purpose of assessing the predictive capacity of the MARS–relied approximation, a total 10-fold cross-validation method was implemented in this investigation [48]. To do this, the regression modelling has been enacted with MARS approximation, employing the ARESlab package [49,50]. The variation intervals of the solution space employed in this investigation are displayed in Table 2.

For the purpose of optimizing the MARS parameters, the grid search technique was employed. In this way, the grid search seeks the most excellent Maxfuncs, Penalty, and Interactions parameters by means of the evaluation of differences of the cross-validation mistake in every iteration. The variation space is three-dimensional (one dimension per each parameter). Thus, the goal function or principal fitness factor is the coefficient of determination (R^2) in this investigation.

3. Results and discussion

Tables 3 and 4 point out the best possible hyperparameters of the most excellent adjusted MARS–relied approximation encountered with the grid search technique for the HL and CL, respectively.

Tables 5 and 6 indicate a list of principal basis functions for the fitted two MARS–relied approximations with their coefficients c_i for the HL and CL, respectively. Observe that a hinge function is described by the expression:

$$h(x) = \left\{ \begin{array}{lll} x & \text{if} & x > 0 \\ 0 & \text{if} & x \leqslant 0 \end{array} \right\} \tag{24}$$

All in all, the MARS approximation is a kind of nonparametric regression approach and can be taken as a generalization of linear methods that models in an automatic way the presence of nonlinearities as well as interactions between input variables employing a weighted summation of *hinge functions* defined above [18–22,29–32].

As well and by comparison, a SVR–relied approach with distinct kernels together with an artificial neural network of MLP-type and M5 model tree have been adjusted to the observed dataset related to the HL and CL output factors too.

Tables 7 and 8 display the determination and correlation coefficients as well as the root mean square error (RMSE) and mean absolute error (MAE) for the most excellent MARS–relied approximation, SVM–relied models for different types of kernel, MLP-type ANN and M5 model tree for the HL and CL output factors for the testing data, respectively.

As stated in these last statistical estimates, the MARS approximation is the most excellent model for estimating both dependent variables (HL and CL factors) in EPB, because the adjusted MARS approximations have coefficients of determination R^2 equal to 0.9961 and 0.9651, and cor-

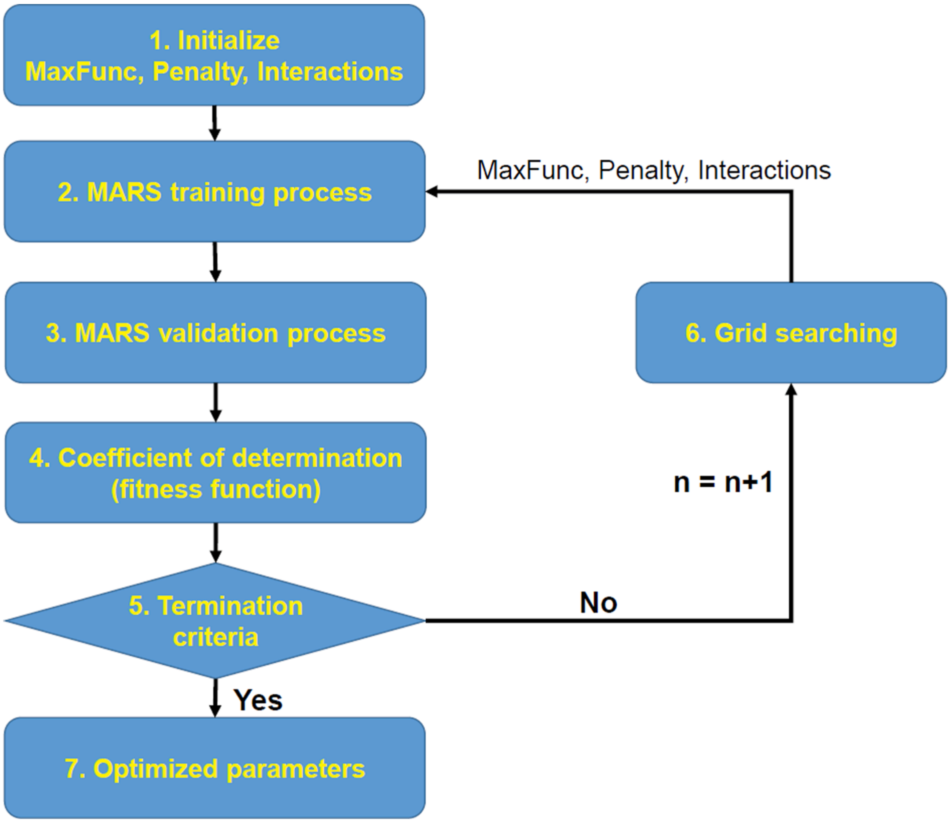


Fig. 5. Process diagram of the MARS–relied approximation.

Table 2
Variation ranges of the three hyperparameters of the MARS–relied approximation adjusted in this research.

MARS hyperparameters	Lower limit	Upper limit
Maximum number of basis functions (MaxFuncs)	3	200
Penalty parameter (*d*)	−1	4
Interactions	1	6

Table 3
Best possible hyperparameters of the most excellent fitted MARS approximation for the Heating load (HL).

Hyperparameters	Optimal values
MaxFuncs	17
Penalty (*d*)	1
Interactions	2

Table 4
Best possible hyperparameters of the most excellent fitted MARS approximation for the Cooling load (CHL).

Hyperparameters	Optimal values
MaxFuncs	19
Penalty (*d*)	1
Interactions	2

Table 5
List of hinge functions of the most excellent fitted MARS–relied approximation for the Heating load (HL) and their coefficients c_i.

B_i	Definition	c_i
B_1	1	52.9469
B_2	$h(\mathrm{RA} - 122.5)$	−0.9162
B_3	$h(122.5 - \mathrm{RA})$	0.9415
B_4	$h(GA - 0.1)$	20.3601
B_5	$h(0.1 - \mathrm{GA})$	−82.6879
B_6	$h(SA - 637)$	−0.8231
B_7	$h(637 - SA)$	−0.3278
B_8	$h(\mathrm{RA} - 122.5) \times h(SA - 612.5)$	0.0101
B_9	$h(\mathrm{RA} - 122.5) \times h(612.5 - SA)$	0.0169
B_{10}	$h(\mathrm{RA} - 122.5) \times h(WA - 343)$	−0.0047
B_{11}	$h(\mathrm{RA} - 122.5) \times h(343 - WA)$	0.0016
B_{12}	$h(\mathrm{RA} - 122.5) \times h(\mathrm{GA} - 0.1)$	−0.0853
B_{13}	$h(\mathrm{RA} - 122.5) \times h(0.1 - \mathrm{GA})$	0.4013
B_{14}	$h(SA - 637) \times h(0.66 - \mathrm{RC})$	1.2203
B_{15}	$h(GA - 0.1) \times h(147 - RA)$	0.1899
B_{16}	$h(GA - 0.1) \times h(GAD - 2)$	−0.7392
B_{17}	$h(GA - 0.1) \times h(2 - GAD)$	0.7392

relation coefficients equal to 0.9981 and 0.9824 for the HL and CL factors, respectively. Hence these outcomes display a reliable goodness of fit, that is to say, an adequate concordance is acquired between MARS approximations and the experimental data. Additionally, a computer with a CPU Intel Core i7-4770 @ 3.40 GHz with eight cores and 15.5 GB RAM memory was used, taking 540 s (approximately 9 min) to obtain the Heating load (HL) factor and 600 s (approximately 10 min) for the Cooling load (CL) factor.

As a supplementary outcome of these estimations, the importance order for the eight input factors foretelling the HL and CL (output dependent factors) in this complex investigation is displayed in Tables 9 and 10, and Figs. 6 and 7, respectively. Therefore, according to MARS models, the most significant variable in the foretelling of HL and CL output variables is the input variable Roof area (RA), followed by surface area (SA), glazing area (GA) and glazing area distribution (GAD) for the HL; and wall area (WA), glazing area (GA) and glazing area distribution (GAD) for CL, respectively.

Therefore, the most important independent variable in the prediction of HL (see Fig. 6) in the adjusted MARS–relied approximation is roof

Table 6
List of hinge functions of the most excellent fitted MARS–relied approximation for the Cooling load (CL) and their coefficients c_i.

B_i	Definition	c_i
B_1	1	34.9228
B_2	$h(\mathrm{RA}-122.5)$	−0.2427
B_3	$h(122.5-\mathrm{RA})$	−0.4235
B_4	$h(GA-0.1)$	19.6777
B_5	$h(0.1-\mathrm{GA})$	−41.0694
B_6	$h(WA-343)$	−0.3735
B_7	$h(343-WA)$	−0.1625
B_8	$h(318.5-WA)$	0.1842
B_9	$h(WA-318.5)\times h(RC-0.64)$	2.2239
B_{10}	$h(WA-318.5)\times h(0.64-RC)$	3.4734
B_{11}	$h(\mathrm{RA}-122.5)\times h(\mathrm{RC}-0.82)$	2.3886
B_{12}	$h(\mathrm{RA}-122.5)\times h(0.82-RC)$	0.3528
B_{13}	$h(\mathrm{RA}-122.5)\times h(GA-0.25)$	−0.0807
B_{14}	$h(RA-122.5)\times h(0.25-\mathrm{GA})$	0.1275
B_{15}	$h(O-4)$	2.0515
B_{16}	$h(4-O)$	0.3419
B_{17}	$h(O-4)\times h(RC-0.9)$	−21.4757
B_{18}	$h(O-4)\times h(0.9-RC)$	−7.8057
B_{19}	$h(O-4)\times h(GAD-3)$	−0.3443

Table 7
Coefficients of determination (R^2), correlation coefficients (r), root mean square error (RMSE) and mean absolute error (MAE) for the MARS–relied model, SVM–relied approaches with linear, quadratic, cubic and RBF kernels, MLP and M5 model tree adjusted in this investtigation for the Heating load (HL) factor and for testing data.

Model	R^2	*R*	*RMSE*	*MAE*
MARS	**0.9961**	**0.9981**	**0.6899**	**0.5709**
SVM–linear kernel	0.9380	0.9685	2.7327	2.0413
SVM-quadratic kernel	0.9490	0.9742	2.5774	1.9418
SVM-cubic kernel	0.9680	0.9839	1.9942	1.3948
SVM–RBF	0.9361	0.9675	2.7324	1.8485
MLP	0.9893	0.9946	1.5663	1.3396
M5 model tree	0.9348	0.9668	0.7639	0.5573

Table 8
Coefficients of determination (R^2), correlation coefficients (r), root mean square error (RMSE) and mean absolute error (MAE) for the MARS–relied model, SVM–relied approaches with linear, quadratic, cubic and RBF kernels, MLP and M5 model tree adjusted in this investigation for the Cooling load (CL) factor and for testing data.

Model	R^2	*R*	*RMSE*	*MAE*
MARS	**0.9651**	**0.9824**	**1.8111**	**1.2884**
SVM–linear kernel	0.8968	0.9470	3.2106	2.2672
SVM-quadratic kernel	0.9155	0.9568	2.8798	1.9877
SVM-cubic kernel	0.9388	0.9689	2.3667	1.5505
SVM–RBF	0.9011	0.9493	3.0718	2.1803
MLP	0.9374	0.9682	1.8553	1.3774
M5 model tree	0.9599	0.9797	1.8230	1.3370

Table 9
Importance order for the variables entailed in the most excellent adjusted MARS–relied approximation for the Heating load (HL) as stated criteria Nsubsets, *GCV* and *RSS*.

Input variable	Nsubsets	*GCV*	*RSS*
Roof area	15	100	100
Surface area	14	52.80	22.92
Glazing area	13	49.9	16.0
Glazing area distribution	1	0.32	0.012

Table 10
Importance order for the variables entailed in the most excellent adjusted MARS–relied approximation for the Cooling load (CL) as stated criteria Nsubsets, *GCV* and *RSS*.

Input variable	Nsubsets	*GCV*	*RSS*
Roof area	17	100	100
Wall area	9	55.90	2.50
Glazing area	15	28.13	16.0
Glazing area distribution	3	0.56	0.02

area, followed by surface area, glazing area and glazing area distribution. In this sense, several works have studied the mechanisms of heat transfer in buildings, Babiarz and Szymański [51] show that in changing performance conditions, energy phenomena occurring in buildings are influenced by different aspects, such as: building shapes or environmental conditions. In this sense, the heat balance in buildings results from the analysis of heat losses and gains through the building structure. Additionally, Moss [52] analysed that heat transfer is particularly important for convection mechanisms in the dwelling structure. Natural convection is consequential from pressure differences among distinct sections of the dwelling structure because of the difference of temperatures due to air environment conditions both outside and within the household. Essentially, when a fluid is near to a heat source and it is hotter than ambient temperature, fluid pressure is lower, which produces air movement. The main reason of this phenomenon, referred to a building structure, is the location of the heat source for heating and cooling performance in the building throughout the year related to the outside environment. It is inside for heating action and outside for cooling process for a cold and hot weather, respectively, in the outside environment.

In this sense, the HL produces an internal air movement by convection with the main influence in the roof area for the airlift, an additional influence is the surface area and glazing area based on the convection movement. Moreover, the heat transfer by conduction in the glazing area is an enabler for heat exchange, glazing area causes heat gains and losses in cooling and heating condition, respectively, by associated U-value [53].

On the other hand, the most important independent variable in the prediction of CL (see Fig. 7) in the fitted MARS–based model is roof area, followed by wall area, glazing area and glazing area distribution. In a similar way, the effect of the air movement over the external surface in CL, based on the sun as heat source of the weather, and the heat transfer in glazing area are drive forces. These circumstances justify the influence of the defined input variables based on thermal energy balance and heat transfer process in buildings. Additionally, the environmental parameters condition the thermal energy production and potential implementation of renewable energies in energy systems based on thermal energy balance [54].

In short, this research permits to estimate the HL and CL output factors in concordance with the real observed values employing the MARS–relied approximation with excellent precision successfully. Certainly, Fig. 8 displays the evaluation of differences among the HL values experimental and foretold employing the M5 model tree (see Fig. 8(a)), MLP approach (see Fig. 8(b)), SVM approach with linear kernel (see Fig. 8(c)), SVM approach with quadratic kernel (see Fig. 8 (d)), SVM approach with cubic kernel (see Fig. 8(e)), SVM approach with RBF kernel (see Fig. 8(f)) and MARS–relied approximation (see Fig. 8(g)). Similarly, Fig. 9 indicates the comparison among the HL values experimental and foretold employing the M5 model tree (see Fig. 9(a)), MLP approach (see Fig. 9(b)), SVM approach with linear kernel (see Fig. 9(c)), SVM approach with quadratic kernel (see Fig. 9 (d)), SVM approach with cubic kernel (see Fig. 9(e)), SVM approach with RBF kernel (see Fig. 9(f)) and MARS–relied approximation (see Fig. 9(g)). Hence, it is mandatory the employ of a MARS approximation for the purpose of achieving the most excellent approach for this regression problem. Clearly, these outcomes concur again with the

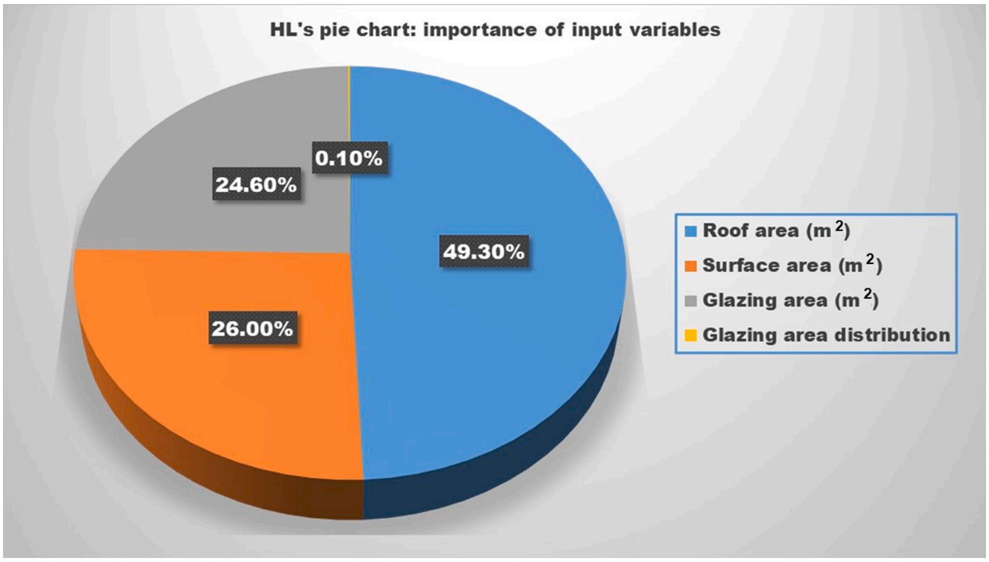


Fig. 6. Comparative importance of the input operation variables to foretell the Heating load (HL) in the adjusted MARS–relied approximation.

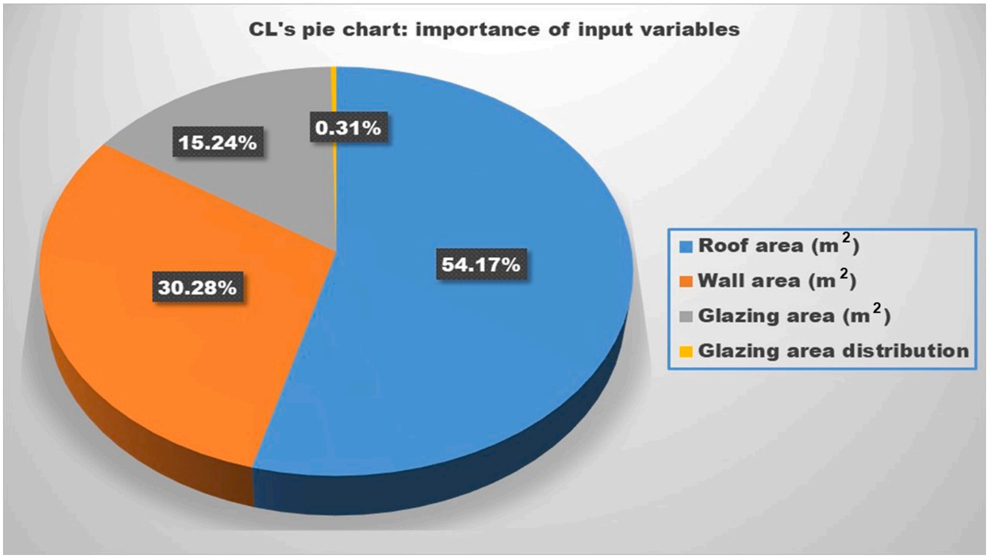


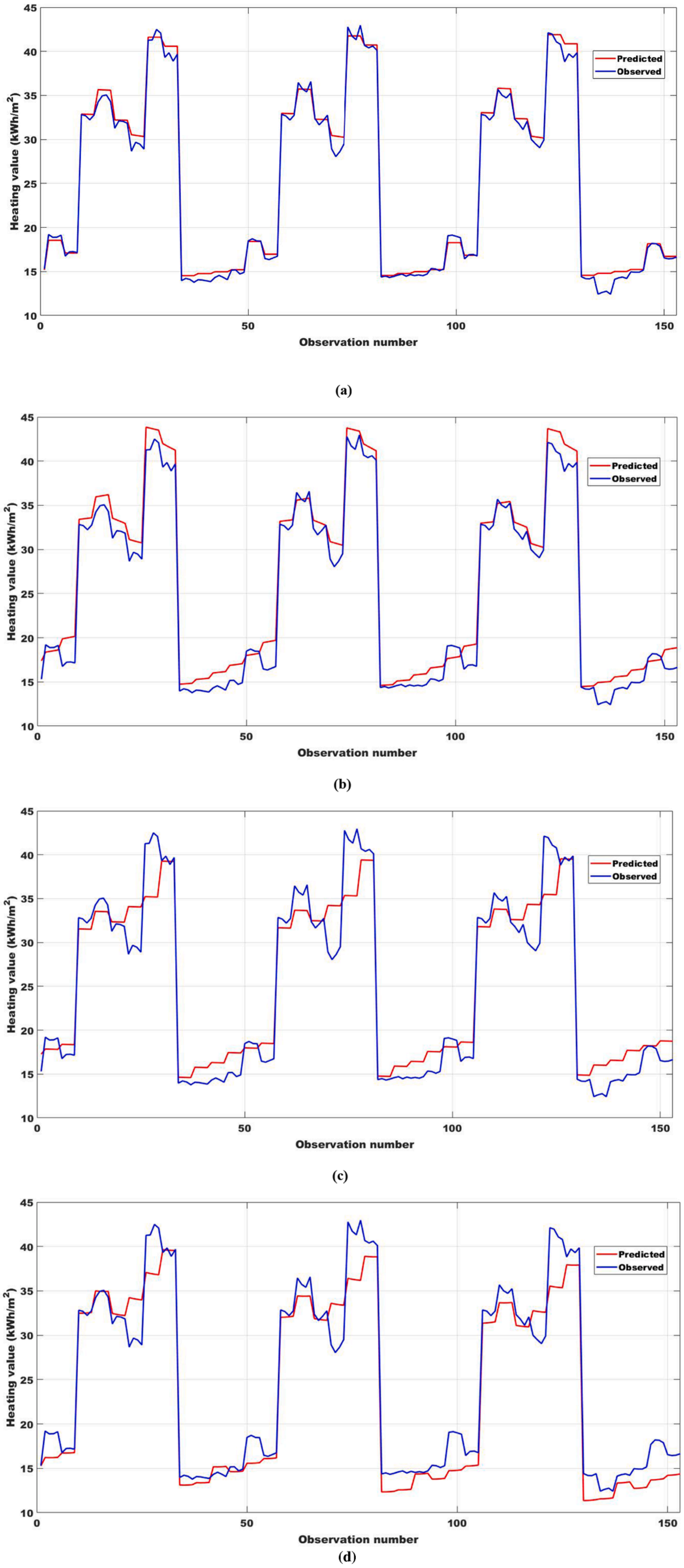
Fig. 7. Comparative importance of the input operation variables to foretell the Cooling load (CL) in the adjusted MARS–relied approximation.

important statistical criterion of ‘goodness of fit’ (R^2) in such a way the MARS–relied approximation has been the most excellent fitting.

4. Conclusions

Comparing the numerical and experimental outcomes, the principal discoveries of this investigation can be synthesised as follows:

- Firstly, the determination of the HL and CL factors requires addressing the solution of a very complex heat transfer problem, taking into account the three forms of transmission: conduction, convection and radiation. The consequential complete model implies the solution of partial differential equations (EDPs). In practice, the solution of this EDP requires numerical methods (for example, the finite element method, finite differences method, etc.) and some additional heuristic approximations cause that the solutions to differ considerably; consequently, the obtaining of additional analisys methods relied on MLT is extremely noteworthy. In a specific way, the MARS–relied approximation employed in this investigation is an adequate choice to assess HL and CL variables in residential dwellings.
- Secondly, the hypothesis that Heating and Cooling loads can be precisely calculated employing a MARS–relied approximation in the construction industry was verified.
- Thirdly, coefficients of determination with values of 0.9961 and 0.9651 are gotten when the MARS–relied approximation is applied to testing data (20 % experimental data not used for training) in the calculation of the dependent variables Heating load (HL) and Cooling load (CL), respectively.
- Fourthly, since MARS approximations generate an explicit mathematical expression of the output variables (HL and CL in this case) from the input variables as a summation of basis functions (known as hinge functions and its products of two or more functions of this type), the MARS approximation may therefore be set up on a cheap-price microcontroller-relied system to accomplish the operation of the EPB foretelling (home automation applications).
- Fifthly, it is feasible to create a significance order of the input variables entailed in the foretelling of the Heating and Cooling loads.



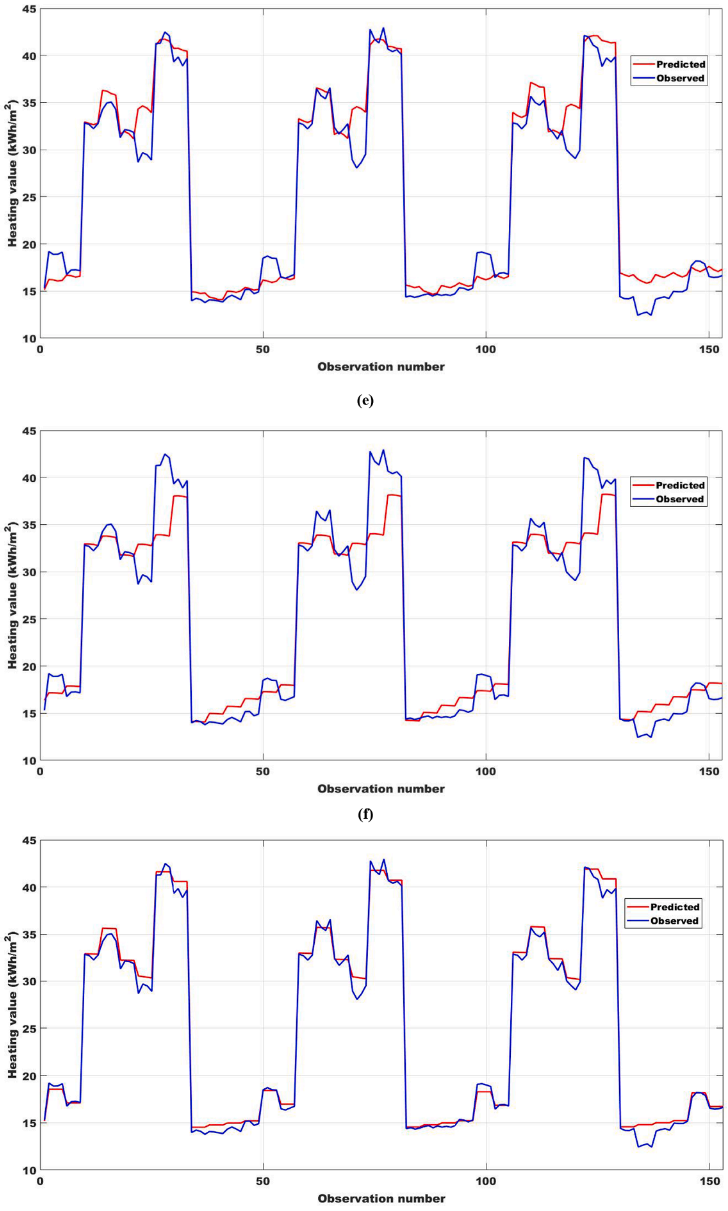
(a)

(b)

(c)

(d)

Fig. 8. Evaluation of differences between the Heating Load values experimental and foretold employing the following models: (a) M5 tree model ($R^2 = 0.9348$); (b) MLP network ($R^2 = 0.9893$); (c) SVM model with linear kernel ($R^2 = 0.9380$); (d) SVM model with quadratic kernel ($R^2 = 0.9490$); (e) SVM model with cubic kernel ($R^2 = 0.9680$); (f) SVM model with RBF kernel ($R^2 = 0.9361$); and (g) MARS model ($R^2 = 0.9961$).



(e)

(f)

(g)

Fig. 8. (*continued*).

They are validated by the performance of heat transfer in buildings and the thermal energy systems for cooling and heating condition. This is one of the principal discoveries of this investigation. Exactly, the independent variable Roof area (RA) could be considered the most important factor in the foretelling of HL and CL output variables, followed by SA, GA and GAD for the HL; and WA, GA and GAD for CL, respectively.

- Sixthly, the principles described here are extremely widespread and could, theoretically, be spread out to cover supplementary independent variables (for instance, some of the factors that are supposed to be constant in this investigation, as occupancy or climate, could be considered as new input variables) in future research.
- Finally, the outcomes of this investigation vigorously warn against the blind use of obtainable mathematical methods to large extent that are in many cases relied on the normal behavior of the data. Afterward, an efficient MARS–relied approximation is a functional answer to the question of the EP in residential dwellings corresponding to the construction industry.

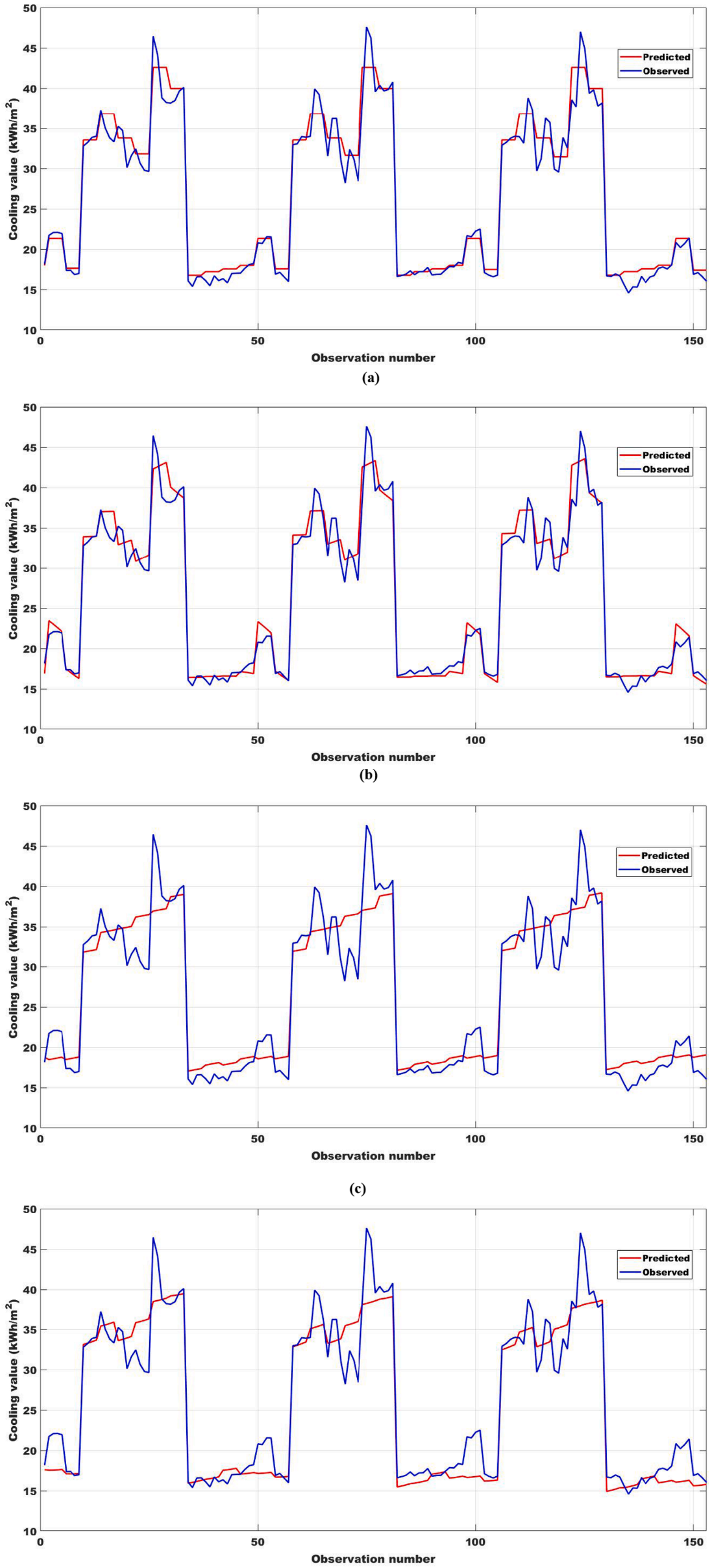
To conclude, this MARS approach could be employed in other types of residential buildings with like or distinct geometries and materials successfully, but it is required to consider the peculiarities of each construction every time.

Data availability

Dataset is made available on Appendix A.

CRediT authorship contribution statement

Paulino José García Nieto: Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Formal analysis, Visualization, Investigation, Supervision, Writing – review & editing. **Esperanza García–Gonzalo:** Conceptualization, Methodology, Software, Validation, Data curation, Writing – original draft, Formal analysis, Visualization, Investigation, Writing – review & editing. **Beatriz María Paredes–Sánchez:** Conceptualization, Methodology, Writing – original draft, Formal analysis, Visualization, Investigation,



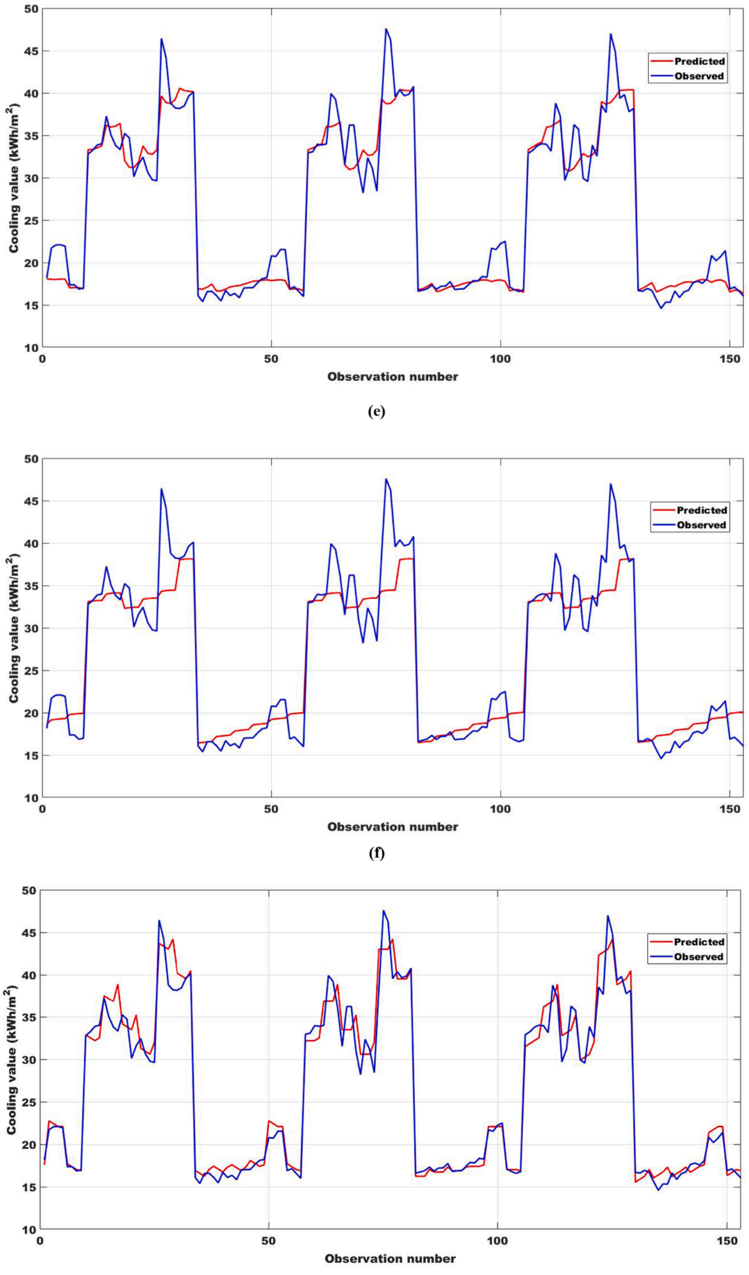
(a)

(b)

(c)

(d)

Fig. 9. Evaluation of differences between the Cooling Load values experimental and foretold employing the following models: (a) M5 tree model ($R^2 = 0.9599$); (b) MLP network ($R^2 = 0.9374$); (c) SVM model with linear kernel ($R^2 = 0.8968$); (d) SVM model with quadratic kernel ($R^2 = 0.9155$); (e) SVM model with cubic kernel ($R^2 = 0.9388$); (f) SVM model with RBF kernel ($R^2 = 0.9011$); and (g) MARS model ($R^2 = 0.9651$).



(e)

(f)

(g)

Fig. 9. (*continued*).

Writing – review & editing. **José Pablo Paredes–Sánchez:** Conceptualization, Methodology, Writing – original draft, Formal analysis, Visualization, Investigation, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Authors desire to thank for the computational support given by the Department of Mathematics at University of Oviedo. Moreover, we would desire to express gratitude to Anthony Ashworth for his correction of English grammar and spelling of this investigation.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apenergy.2023.121074.

References

[1] Perez-Lombard L, Ortiz J. A review on buildings energy consumption information. Energy Build 2008;40(3):394–8.

[2] Cai WG, Wu Y, Zhong Y, Ren H. China building energy consumption: situation, challenges and corresponding measures. Energy Policy 2009;37(6):2054–9.

[3] European Commission, Directive 2002/91/EC of the European Parliament and of the council of 16th December 2002 on the energy performance of buildings, Official journal of the European Communities, L1/65–L1/71, 04/01/2003.

[4] Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings. L153/13 18/06/2010.

[5] European Parliament and of the Council. Directive (EU) 2018/844 of the European Parliament and of the Council, of May 30, 2018, amending Directive 2010/31/EU on the energy efficiency of buildings and Directive 2012/27/EU on to energy efficiency. L156/75, 19/06/2018.

[6] García-Nieto PJ, García-Gonzalo E, Paredes-Sánchez BM, Paredes-Sánchez JP. Forecast of the higher heating value based on proximate analysis by using support vector machines and multilayer perceptron in bioenergy resources. Fuel 2022;317:122824.

[7] Paredes-Sánchez BM, Paredes-Sánchez JP, García-Nieto PJ. Evaluation of Implementation of Biomass and Solar Resources by Energy Systems in the Coal-Mining Areas of Spain. Energies 2021;15(1):232.

[8] Platt G, Li J, Li R, Poulton G, James G, Wall J. Adaptive HVAC zone modelling for sustainable buildings. Energy Build 2010;42:412–21.

[9] Afram A, Janabi-Sharifi F. Review of modeling methods for HVAC systems. Appl Therm Eng 2014;67(1–2):507–19.

[10] Kreider JF, Curtiss PS, Rabl A. Heating and cooling of buildings: design for efficiency. Boca Ratón, Florida, USA: CRC Press; 2009.

[11] Hu M. Net zero energy building: predicted and unintended consequences. London: Routledge; 2019.

[12] Corner DB, Fillinger JC, Kwok AG. Passive house details: solutions for high-performance design. London: Routledge; 2017.

[13] Atam E. Current software barriers to advanced model-based control design for energy-efficient buildings. Renew Sust Energ Rev 2017;73:1031–40.

[14] Wan KKW, Li DHW, Liu D, Lam JC. Future trends of building heating and cooling loads and energy consumption in different climates. Build Environ 2011;46:223–34.

[15] Tsanas A, Xifara A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tool. Energy Build 2012;49:560–7.

[16] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning. New York: Springer-Verlag; 2003.

[17] García-Nieto PJ, García-Gonzalo E, Sánchez Lasheras F, Paredes-Sánchez JP, Riesgo FP. Forecast of the higher heating value in biomass torrefaction by means of machine learning techniques. J Comput Appl Math 2019;357:284–301.

[18] Friedman JH. Multivariate adaptive regression splines. Ann Stat 1991;19:1–141.

[19] Sekulic SS, Kowalski BR MARS:. A tutorial. J Chemometr 1992;6:199–216.

[20] Friedman JH, Roosen CB. An introduction to multivariate adaptive regression splines. Stat Methods Med Res 1995;4:197–217.

[21] Xu QS, Daszykowski M, Walczak B, Daeyaert F, De Jonge MR, Heeres J, et al. Multivariate adaptive regression splines—studies of HIV reverse transcriptase inhibitors. Chemometr Intell Lab 2004;72(1):27–34.

[22] Vidoli F. Evaluating the water sector in Italy through a two stage method using the conditional robust nonparametric frontier and multivariate adaptive regression splines. Eur J Oper Res 2011;212(13):583–95.

[23] Steinwart I, Christmann A. Support vector machines. Cambridge, Massachusetts (USA): Springer; 2008.

[24] Bishop CM. Neural networks for pattern recognition. Oxford (UK): Oxford University Press; 1995.

[25] Quinlan JR. Learning with continuous classes. In: Proceedings of Australian Joint Conference on Artificial Intelligence, Singapore: World Scientific Press; 1992, p. 343–48.

[26] Majeed F, Ziggah YY, Kusi-Manu C, Ibrahim B, Ahenkorah I. A novel artificial intelligence approach for regolith geochemical grade prediction using multivariate adaptive regression splines. Geosyst Geoenviron 2022;1(2):100038.

[27] Chen W–H, Lo H–J, Aniza R, Lin B–J, Park Y–K, Kwon EE, Sheen H–K, Grafilo LADR. Forecast of glucose production from biomass wet torrefaction using statistical approach along with multivariate adaptive regression splines, neural network and decision tree. Appl Energ 2022;324:119775.

[28] Pessenlehner W, Mahdavi A. A building morphology, transparency, and energy performance. In: Eighth International IBPSA Conference Proceedings, Eindhoven, Netherlands; 2003; p. 1025–32.

[29] Álvarez Antón JC, García Nieto PJ, de Cos Juez FJ, Sánchez Lasheras F, Blanco Viejo C, Roqueñí GN. Battery state-of-charge estimator using the MARS technique. IEEE Trans Power Electron 2013;28:3798–805.

[30] Chen M-Y, Cao M–T. Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. ApplSoft Comput 2014;22:178–88.

[31] García-Nieto PJ, García-Gonzalo E, Bové J, Arbat G, Duran-Ros M, Puig-Bargués J. Modeling pressure drop produced by different filtering media in microirrigation sand filters using the hybrid ABC–MARS–based approach, MLP neural network and M5 model tree. Comput Electron Agric 2017;139:65–74.

[32] Kisi O. Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 2015;528:312–20.

[33] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. New York (USA): Cambridge University Press; 2000.

[34] Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, Massachusetts (USA): The MIT Press; 2001.

[35] Hassoun M. Fundamentals of artificial neural networks. Boston (USA): MIT Press; 2003.

[36] Ripley BD. Pattern recognition and neural networks. Cambridge (UK): Cambridge University Press; 2014.

[37] Pal M. M5 model tree for land cover classification. Int J Remote Sens 2006;27(4):825–31.

[38] Pal M, Deswal S. M5 model tree based modelling of reference evapotranspiration. Hydrol Process 2009;23(10):1437–43.

[39] Solomatine DP, Xue YP. M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Hual River in China. J Hydrol Eng 2004;9(6):491–501.

[40] Rahimikhoob A, Asadi M, Mashal M. A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. Water Resour Manage 2013;27(14):4815–26.

[41] Behnood A, Olek J, Glinicki MA. Predicting modulus elasticity of recycled aggregate concrete using M5′ model tree algorithm. Constr Build Mater 2015;94:137–47.

[42] Khorrami R, Derakhshani A, Moayedi H. New explicit formulation for ultimate bearing capacity of shallow foundations on granular soil using M5’ model tree. Measurement 2020;163:108032.

[43] Seghier M, Keshtegar B, Correia J, de Jesús A, Lesiuk G. Structural Reliability Analysis of Corroded Pipeline made in X60 Steel Based on M5 Model Tree Algorithm and Monte Carlo Simulation. Procedia Struct Integr 2018;23:1670–5.

[44] Freedman D, Pisani R, Purves R. Statistics. New York: W.W. Norton & Company; 2007.

[45] Agresti A, Kateri M. Foundations of statistics for data scientists: with R and Python. Boca Ratón, Florida (USA): Chapman and Hall/CRC; 2021.

[46] Aggarwal CC. Linear algebra and optimization for machine learning. Springer, New York, USA: Springer; 2020.

[47] Theodoridis S. Machine learning: A bayesian and optimization perspective. London (UK): Academic Press; 2020.

[48] Picard R, Cook D. Cross-validation of regression models. J Am Stat Assoc 1984;79(387):575–83.

[49] Milborrow S. Earth: multivariate adaptive regression spline models. R Package, version 2014;3:2–7.

[50] Jekabsons G. ARESlab User’s manual: adaptive regression splines toolbox for Matlab/Octave, version 1.13.0; 2016.

[51] Babiarz B, Szymański W. Introduction to the dynamics of heat transfer in buildings. Energies 2020;13(23):6469.

[52] Moss KJ. Heat and mass transfer in buildings. London (UK): Routledge; 2015.

[53] Ficco G, Iannetta F, Ianniello E, Alfano FRDA, Dell’Isola M. U-value in situ measurement for energy diagnosis of existing buildings. Energy Build 2015;104:108–21.

[54] Paredes-Sánchez JP, Las-Heras-Casas J, Paredes-Sánchez BM. Solar energy, the future ahead. In: Vasel A, Ting DK, editors. Advances in sustainable energy. Cham: Springer; 2019. p. 113–32.