Applied Energy 253 (2019) 113500

Contents lists available at ScienceDirect

Applied Energy

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

Building energy performance forecasting: A multiple linear regression approach

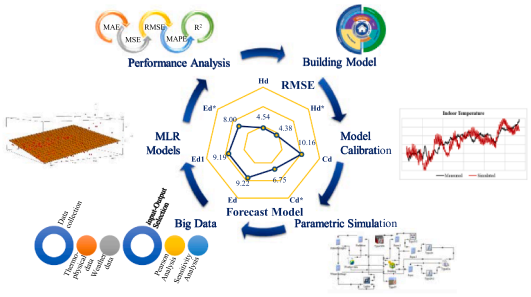
G. Ciulla*, A. D'Amico

University of Palermo, Department of Engineering (DI), Viale delle Scienze, Ed. 9, Italy

HIGHLIGHTS

- Building energy demand assessment designed with high-energy performance.
- Parametric simulation to develop an accurate energy database.
- Sensitivity analysis to identify the main parameters of the building energy balance.
- Forecasting of the building energy needs through the black box method.
- Multiple Linear Regression to identify simple correlations with high reliable degree.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Building energy demand
Sensitivity analysis
Forecast method
Dynamic simulation
Black box method
Multiple linear regression

ABSTRACT

Different ways to evaluate the building energy balance can be found in literature, including comprehensive techniques, statistical and machine-learning methods and hybrid approaches. The identification of the most suitable approach is important to accelerate the preliminary energy assessment. In the first category, several numerical methods have been developed and implemented in specialised software using different mathematical languages. However, these tools require an expert user and a model calibration. The authors, in order to overcome these limitations, have developed an alternative, reliable linear regression model to determine building energy needs. Starting from a detailed and calibrated dynamic model, it was possible to implement a parametric simulation that solves the energy performance of 195 scenarios. The lack of general results led the authors to investigate a statistical method also capable of supporting an unskilled user in the estimation of the building energy demand. To guarantee high reliability and ease of use, a selection of the most suitable variables was conducted by careful sensitivity analysis using the Pearson coefficient. The Multiple Linear Regression method allowed the development of some simple relationships to determine the thermal heating or cooling energy demand of a generic building as a function of only a few, well-known parameters. Deep statistical analysis of the main error indices underlined the high reliability of the results. This approach is not targeted at replacing a dynamic simulation model, but it represents a simple decision support tool for the preliminary assessment of the energy demand related to any building and any weather condition.

1. Introduction

In Europe, the building sector is considered to be the largest energy consumer being responsible for up to 40% of total energy use and 36% of total CO_2 emissions [1]; more specifically, the non-residential sector represents about 40% of total energy consumption in the building

* Corresponding author.
E-mail addresses: giuseppina.ciulla@unipa.it (G. Ciulla), antonio.damico@deim.unipa.it (A. D'Amico).

https://doi.org/10.1016/j.apenergy.2019.113500
Received 8 April 2019; Received in revised form 5 June 2019; Accepted 7 July 2019
Available online 10 July 2019
0306-2619/

Nomenclature

Acronyms

ANN	Artificial Neural Network
CFD	Computational Fluid Dynamics
FEMP	Measurement and Verification of Federal Energy Projects
GA	Genetic Algorithm
MLR	Multiple Linear Regression
SVM	Support Vector Machine
TMY	Typical Meteorological Year

MLR parameters

y_i	i-th independent variable
x_i	i-th explanatory variable
b_0	intercept of the linear regression
b_i	i-th regression coefficient
e	error of the linear regression

Error and performance parameters

CV-RMSE	Coefficient of Variation of the Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
NMBE	Normalize Mean Bias Error
RMSE	Root Mean Square Error
R^2	determination coefficient
StD	Standard Deviation

Other parameters

CDD	Cooling Degree Days [K day]
HDD	Heating Degree Days [K day]
Q_G	internal gains [kWh/year]
S_{op}	opaque surface [m^2]
S_w	surface of the glazed component [m^2]
S/V	shape factor [m^{-1}]
U	thermal transmittance [$W/(m^2 \cdot K)$]

Outputs of the models

C_d	cooling energy demand [$kWh/(m^2 \cdot year)$]
E_d	energy demand [$kWh/(m^2 \cdot year)$]
H_d	heating energy demand [$kWh/(m^2 \cdot year)$]

sector [2,3]. A knowledge of the energy performance of existing building stocks and the forecasting of the energy behaviour of newly designed buildings is fundamental to achieving the targets of the EPDB (Energy Performance of Building Directive) established by the European Union [1].

It is well-known that building energy assessment is quite complex owing to the influence of many factors, such as weather conditions, the building construction and its shape, the thermophysical properties of the materials used, the intended use, the occupancy and behaviour of the users, the lighting, the ventilation, and the heating/cooling systems along with their performance and operating schedules [4]. Furthermore, for new buildings, it is necessary for the choices to be based on high energy performance, securely guaranteeing the achievement of energy and environmental targets.

In general, the evaluation of the energy performance of an existing building and the design of new buildings integrating several energy-efficiency measures are solved via software programs with the aim of predicting improvements that could be made considering different design management. For careful energy planning, new methods have to be explored in order to support engineers and architects in their efforts to improve design, reduce computational time and increase energy performance.

Due to the complexity of the problem, the prediction of building energy consumption is quite difficult and has become one of the main objectives of several research studies. In recent years, a large number of both elaborate and simplified forecast approaches have been proposed and applied to several problems. Several of these cases are available, some based on knowledge of the building thermal balance and on the resolution of physical equations, and others on building data collection and on the implementation of forecast models developed by means of machine-learning techniques [5]. In the literature, it is possible to distinguish three main methods: "white box" or physical techniques, "black box" or statistical and/or learning approaches and "grey box" or hybrid approaches.

The "white box" approaches are used to model building thermal behaviour for several applications on different scales. These techniques, known also as engineering methods, are based on the use of physical principles to solve the equations describing the physical behaviour of heat transfer. In this category, it is possible to distinguish between simplified and detailed comprehensive methods. Among the simplified methods, the degree day method is one of the most used; several research studies affirm that meteorological data provide an effective tool for determining the energy demand and for calculating heating or cooling building requirements [6–8]. Another simplified method is based on the temperature frequency, which can be used to model large buildings where internal gains (Q_G) dominate [9]. For example, White et al. [10] attempted to use average monthly temperatures to predict monthly building energy consumption and Westphal et al. [11] forecasted the annual heating and cooling loads of non-residential buildings based on certain weather variables.

On the other hand, the detailed comprehensive methods use very elaborate physical functions to evaluate, step-by-step, the energy consumption of a building linked to its construction, operation of the systems, utility rate schedule of the equipment, external climate conditions and solar irradiance.

To solve such physical problems, a large number of numerical software programs are available and these have been compared [12,13]. Users can choose to select the mechanisms and the associated equations representing the system, but sometimes many software tools are badly adapted to taking into account moisture influences, and generally the effects of latent heat are neglected [13,14]. In the literature, three main thermal building models can currently be found: Computational Fluid Dynamics (CFD), zonal methods and the multi-zone technique. CFD is a branch of fluid mechanics that is based on numerical analysis to analyse and solve problems that involve flows. Nowadays, a huge number of CFD software programs are available, such as FLUENT [15], COMSOL Multiphysics [16], MIT-CFD, PHOENICS-CFD [17], and so on.

The zonal method is the first degree of simplification of the CFD technique; it involves dividing each building zone into several cells detailing the indoor environment and estimating a thermal comfort zone [18,19]. Specifically, this technique presents its efficiency in the description of the flow profiles within the building. The multi-zone technique, or nodal method, is based on the assumption that each building zone is a homogeneous volume characterised by uniform state variables. The solution is based on the application of two main methods: transfer functions or the finite difference method. In the field of energy efficiency and sustainability in buildings, and based on this last technique, several software tools have been developed, such as, Energy Plus [20], ESP-r [21], TRNSYS, IDA-ICE [22], Clim2000

[23,24], BSim [25,26] and BUILDOPT-VIE [27].

Although these simulation tools are effective and accurate, there are some practical difficulties in implementing a reliable model. Indeed, these tools require details of the building and environmental parameters which are not always simple to find and collect, and the lack of precise input can lead to a low-accuracy simulation; furthermore, to use these tools normally, an expert user is required, as is a careful calibration of the model.

The "black box" approaches are mainly used to deduce a prediction model from a relevant database (for example, to forecast energy consumption or heating/cooling load in a given building). These models do not require any information about physical phenomena but they are based on a function deduced only by means of sample data connected to each other and which describe the behaviour of a specific system. The black box methods mainly employed in the field of building energy forecasting are: Multiple Linear Regression (MLR) or statistical regression model, Genetic Algorithm (GA), Artificial Neural Network (ANN) the Support Vector Machine (SVM) [4,5]; an overview of these method is described in Li et al. [28].

MLR methods correlate the building energy consumption or energy indices with the influencing variables in a simple way. These empirical models are developed based on energy performance data collected previously. In certain simplified models, linear regression is used to correlate the energy consumption with climatic variables [29–31]; for example, Ansari et al. [32] calculated the total cooling load by adding up the cooling load of each building envelope component, while Dhar et al. [33,34] modelled heating and cooling loads using the outdoor dry bulb temperature as the only weather variable. Parti et al. [35] were the first to propose a new method using linear regression for the prediction of building energy consumption.

Kialashaki et al. [36] applied the regression and ANN models to evaluate the energy requirements of the residential sector.

The main advantage of this method is its ease of use; indeed, no specific expertise is required. As indicated in Aydinalp-Koksal et al. [37], regression models are easier to use, against the engineering methods. However, the MLR presents a major limitation in that it is unable to treat non-linear problems.

GA is a stochastic optimisation technique based on Darwin's theory of evolution. In building simulation, GA is used to find a prediction model deducing a simple equation which can fit the problem. An important advantage of GA is the fact that it deals with a powerful optimisation method which is able to solve every problem and give several final solutions to a complex problem [5].

Among artificial intelligence models, ANN's are the most widely used in the forecasting of building energy and are capable of solving both non-linear and complex problems [38,39]. The main advantage of ANN is its ability to determine non-linear relationships among different variables without any assumptions or any postulate of a model overcoming the discretisation problem. However, ANNs need to have a relevant database in order to obtain reliable solutions. In fact, it is really important to train an ANN with an exhaustive learning database with representative and complete samples [4].

Among artificial intelligence techniques, SVM, introduced by Vapnik et al. [40], is usually used to solve classification and regression problems. These are highly effective models even with small quantities of training data. Many studies [41,42] of these models were conducted on building energy analysis and demonstrate that SVMs can perform well in predicting hourly and monthly building energy consumption. When a problem cannot be completely solved by applying one of the methods previously described, it is possible to use a "grey box" method. These methods can overcome the limitations of each individual technique by coupling them so that the advantages of one method counteract the drawbacks of the other [5].

In the field of building energy planning, it would be more convenient to identify the best method that describes the investigated problem for the development of a generic decision support tool, characterised by low calculation time, a non-complex data collection phase, high reliability and a simple calculation language that can be used even by a non-expert user.

1.1. Contribution of the work

In this paper, the authors have tried to identify a simple method capable of solving the traditional building energy balance which will represent a decision support tool useful in the preliminary phases of an energy planning, when the user is not an expert or when it is necessary to speed up the decision-making phases.

A comprehensive analysis of the energy performance of a specific building, although correctly interpreting the energy balance problem, requires an expert user with a knowledge of the physical problems associated and who is capable of constructing a model, collecting and implementing a large number of parameters, performing careful calibrations and explaining the results well. All of these steps require high computational time and do not always provide an immediately correct evaluation; with an incorrect assessment, the procedure must be restarted. Moreover, although a parametric simulation allows the simultaneous analysis of the energy needs of several case studies, the results cannot be generalised: a dynamic simulation of each individual case corresponds to a specific result. To try to overcome these limits and to accelerate the preliminary assessment phase, the authors investigated the reliability of an alternative method using the multiple linear regression to solve the building energy balance. With the implementation of a detailed, reliable energy database, representative of high energy performance non-residential building stocks, it was possible to apply the black box method and to compare the obtained results with the previous comprehensive analysis. Obviously, the method validity is linked to the reliability of the database used to identify the linear correlation. To ensure this high reliability, the authors based their work on a carefully calibrated TRNSYS model, implementing a parametric simulation which allowed the investigation of 195 scenarios representing different possible building combinations built with high energy performance and simulated in several climatic conditions for different thermophysical characteristics and different shape factors *(S/V)*. Furthermore, a careful sensitivity analysis through examination of the Pearson coefficient permitted the identification of the most suitable variables that influence the building energy balance during the heating/cooling period. The application of the MLR method to the energy database resulted in the definition of some simple correlations that identified heating (H_d), cooling (C_d) or comprehensive (E_d) energy demand with a high degree of reliability and these are valid for a representative building stock. These correlations, validated thanks to deep statistical error analysis, solve the building energy balance knowing only a few well-known parameters and without any computational time or physical knowledge. For these reasons, the solutions obtained from the application of the MLR method can be considered generic and applicable to any condition. The literature reports black box methods being applied to forecast the energy needs of a single building or a district level, yet in this work a methodology is proposed to allow the identification of a more flexible tool that can assess the energy requirements of an entire panorama of non-residential buildings. Indeed, once the correlations valid at a general level have been determined, it will be possible to provide an easy-to-use tool that can help identify the needs of a building without that the user knowing the physical problem or all the variables that come into play, simply by solving a linear equation. Furthermore, the added value of this paper lies in the generality of the results obtained thanks to the availability of an accurate database built on a high number of models that were simulated with parametric simulations. The high degree of reliability achieved from the results guarantees that this methodology can be replicated in any other climatic and building context, representing a forecasting tool to support decisions.

The simple form of the correlations could be used as a

supplementary evaluation criterion/tool to support standards and/or laws in the field of building energy performance. Although it is possible to develop AI-based models (SVM, ANN, GA, and so on), which in some cases, present more accurate solutions, these tools require a high knowledge of the physical, numerical and mathematical principles of the analysed sector. Moreover, as for the MLR application, such tools require for their implementation the use of an accurate database [5,43]. Another strength of the presented model in this work is that the application of the MLR method does not require, during the use phase, any calculation tool such as a personal computer or software program, but it is characterised by the resolution of a simple linear equation.

2. Method

The aim of this paper is to provide an improved method that allows the evaluation of building energy performance immediately and simply in any situation and boundary conditions. In this section, the main steps and the procedures used to achieve the objective of the work are illustrated. In the flow chart, displayed in Fig. 1, the entire procedure followed by the authors is represented.

As indicated in the flow chart, the idea is to develop a generic decision support tool that, without an expert user and with a high degree of reliability, immediately solves the traditional building energy balance in any case and in any conditions. In order to achieve this goal, it is necessary to develop a generic solution of a representative building stock which includes all possible building topologies and environmental conditions. For this reason, the authors decided to investigate, as representative, non-residential buildings designed with high energy performance according to the energy efficiency laws and standards in Italy (Section 3).

As explained in the introduction, to solve the building energy balance it is possible to choose several methods, the two methods applied in this work are reported in the flow chart. First of all, to determine the building thermal energy demand a comprehensive method (Section 4) was applied. A detailed TRNSYS model of a non-residential building was developed allowing, after a careful calibration with actual conditions, the determination of the heating and cooling energy requirements (Section 4.1). To collect the results that describe the energy balance of the representative building stock, a parametric simulation was developed. Based on the calibrated model and developing the parametric analysis, the authors simulated 195 scenarios of a non-residential building [44], which represent the possible combinations in 5 climatic zones and 15 cities, with 13 shape factors and different thermo-physical parameters, for a total of 1560 simulations (Section 4.2).

The identification of a series of restrictions such as data collection, knowledge of the physical problems, the tool language, the computational time and the lack of generality of results, because these are single answers to a specific condition (Section 4.3), prompted the authors to

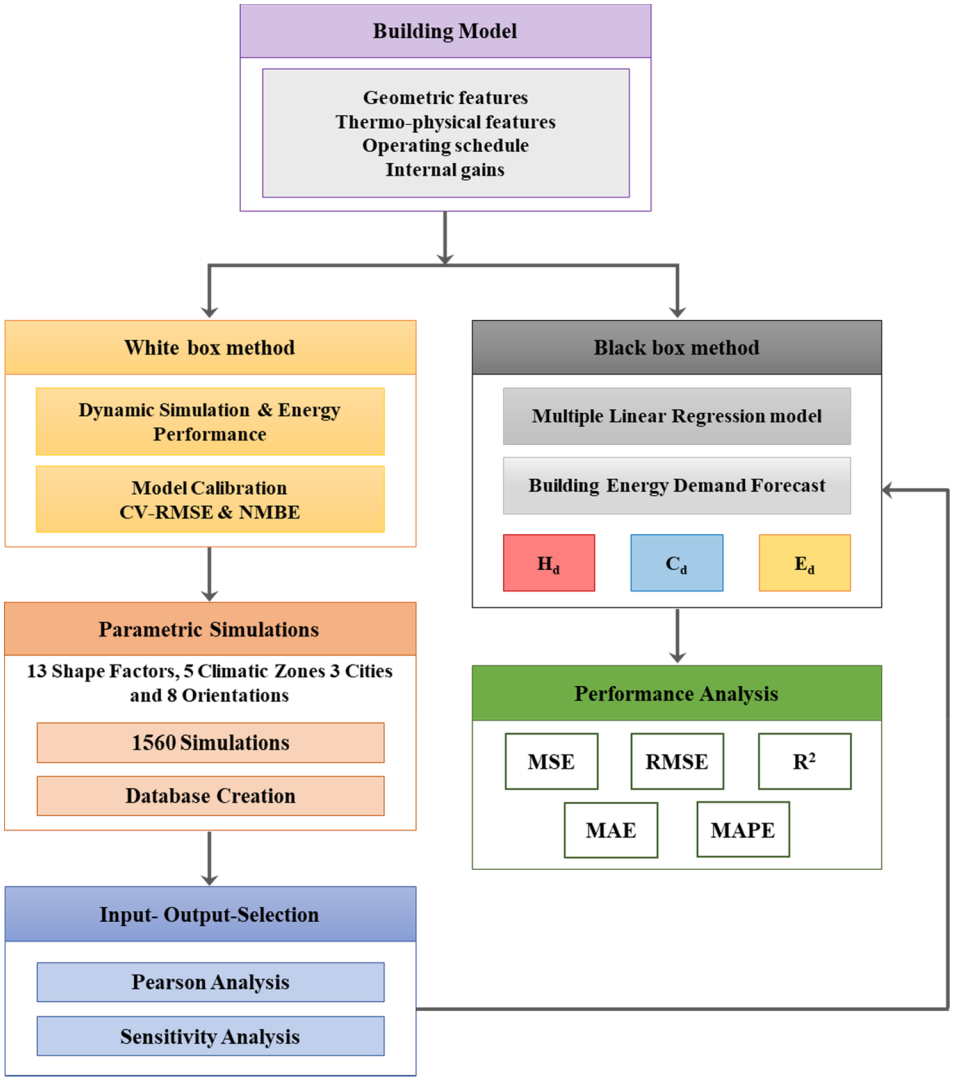


Fig. 1. Flow chart of the procedure method.

investigate other alternative methods that overcome these limits. As previously indicated, a good alternative for resolving this problem is represented by the black box methods (Section 5). Although they do not take into account the physics of the problem, they are able to identify a correlation or a dependence between the input and output data. The strong correlation or dependence between the data is guaranteed by the identification of the main parameters that characterise the building energy balance. In this case, for a generic solution, all main parameters that describe the building thermal energy balance and all thermal energy results obtained from the parametric simulations were collected in a matrix of 197 rows and 20 columns. This dataset was used to explore the MLR method (Section 6), which allows the modelling of a linear relationship between two or more explanatory variables (input of the model) and a response variable (building energy performance) through a fitting procedure.

The identification of the best solutions for calculating the building energy demand with high reliability is guaranteed thanks to preventive sensitivity analysis (Section 6.1), which allowed to identify the more correlate parameters with the heating and cooling demand, and so too the optimal input data for forecasting the building energy requirements. The performance analysis of this alternative method is illustrated by means of an error metric analysis (Section 6.2), which provides the most used error indices. This statistical analysis was applied for all correlation forms proposed: for the heating (Section 6.2.1), cooling (Section 6.2.2) and comprehensive energy demand assessment (Section 6.2.3). Owing to the reliability and flexibility of the energy database, this method was investigated with good results. The generic database, which identifies a solution that simultaneously responds to changes in climate and shape factor, gives generic solutions that can explain any possible building topology in any condition (Section 7).

3. Case study

As previously indicated, the authors proposed a method for the assessment of building energy needs that can be extended to any context, for any building and boundary conditions. In order to obtain a generic tool with these characteristics, it was necessary to investigate a representative building stock that includes all possible building types and environmental conditions. It is known in the energy efficiency field that every European country legislates autonomously and that the standards and laws require different transmittance limit values for the building envelope and different efficiency systems. In this case, the authors decided to analyse a representative building stock designed with high energy performance and non-residential use located in the Italian context, which had already been developed in a previous work [44].

Based on the Heating Degree Days (*HDD*) index, the Italian peninsula can be divided into six different climatic zones [45], where zone A represents the hottest and zone F represents the coolest. For each zone, the daily hours of heating system activity and the consequent yearly heating period are indicated (Table 1) and the transmittance limit values for the design of high-performance buildings are imposed (Table 2) [46].

As for Cooling Degree Days (*CDD*), the current Italian standards indicate values without changing the cooling periods for all Italian cities, and without making any distinction between the climate zones [47]. In order to represent the entire climate conditions, 3 cities, characterised by the maximum, minimum, and mean *HDD* value were selected for each zone [48]; the 15 selected cities, according to actual Italian laws and standards are collected in Table 3.

Furthermore, because in [47] the *CDD* values for all Italian cities are not indicated, among the cities chosen in this work, Cortina, Sestriere and Termoli have no indication of *CDD* values (bolded in Table 3). Based on the calculation procedure used in Italy [44] and the actual standard indications for the determination of the *CDDs*, the authors calculated the *CDD* values for these 3 cities.

From a geometrical point of view, the shape factor indicates the ratio between the surface exposed to the outside or to another ambient at different temperature, and the heated volume, representing thus, an energy loss index [4]. For this reason, to obtain generic results, it is necessary to investigate several geometrical configurations. In the following table (Table 4) all analysed geometrical configurations are listed.

On the basis of the real geometric constructions of a non-residential building with high energy performance, the authors have tried to investigate the greatest number of combinations, varying the S/V from 0.2 to 0.9 and respectively identifying the geometric dimensions.

A knowledge of the energy demand of each building, varying simultaneously the weather conditions, the shape factor and the thermal transmittance of the envelope allowed us to obtain an energy database of non-residential building stock designed with high performance representative of the Italian context.

4. White box methods

One of the most common white box methods for solving the building energy balance is the multi-zone technique for which several software programs have been developed. The application of these software provides optimal energy consumption estimations with some simplifications. Indeed, these tools require a detailed data collection phase, long computational times, calibration of the model and an expert user who knows how to use these and the physical phenomena of the problem. To achieve the aim, the authors first implemented and calibrated an "ideal building" and then, to generalise the results, developed a parametric simulation.

4.1. Base-case and calibration

The authors decided to consider a non-residential building located in the south of Italy the as "Base-Case". The building was constructed between 1962 and 1965 and it is used as the Department of Energy, Information Engineering and Mathematical Models (DEIM) at the University of Palermo. It has five elevations: the mezzanine floor and the third floor are intended for laboratory use, the first and second floors are mainly used as offices, and the basement floor is the location of the technical room. From a structural point of view, the building has a load-bearing system framed with pillars and beams in reinforced concrete with foundations made of reinforced concrete plinths connected with beams. Each floor is characterised by a surface of 1130 m^2 and the thermophysical main features are listed in Table 5.

The windows are made of aluminium and equipped with insulating thermoacoustic glass with plastic blinds. To solve the energy balance of this building, TRNSYS software (Fig. 2) was used and in order to simulate the thermal behaviour, the following were considered:

- detailed weekly and daily schedules regarding the utilisation of equipment, lighting systems, and presence of office users;
- actual monitored data recorded from 2000 to 2009 (TMY2) generated by Meteonorm software [49];

Table 1
Italian climatic zones.

Climatic Zone	*HDD*		Heating season		Daily hours
	From	To	From	To	
A	0	600	1st December	15th March	6
B	601	900	1st December	31st March	8
C	901	1400	15th November	31st March	10
D	1401	2100	1st November	15th April	12
E	2101	3000	15th October	15th April	14
F	3001	∞	No limit		

Table 2
Limit thermal transmittance values for each climatic zone.

Climatic zone	A-B	C	D	E	F
	U_{limit} [W/(m^2·K)]				
U_{wall}	0.45	0.38	0.34	0.30	0.28
U_{roof}	0.38	0.36	0.30	0.25	0.23
U_{floor}	0.46	0.40	0.32	0.30	0.28
U_{window}	3.20	2.40	2.00	1.80	1.50

Table 3
Selected Italian cities *HDD* and *CDD* values.

Climatic Zone	City	*HDD*	*CDD*
		[K day]	[K day]
B	Messina	707	260
	Palermo	751	309
	Crotone	899	255
C	Cagliari	990	222
	Bari	1185	314
	Termoli	1350	155
D	Genova	1435	115
	Firenze	1821	331
	Forlì	2087	108
E	Trieste	2102	125
	Torino	2617	166
	Bolzano	2791	135
F	Cuneo	3012	80
	Cortina	4433	0
	Sestriere	5165	0

Table 4
Geometric features of the investigated building models [44].

Case study	*S/V* [m^{-1}]	Width [m]	Depth [m]	Height [m]	Loss surface [m^2]	Heated surface [m^2]	Heated volume [m^3]
1	0.24	45	39	13.5	5797	7050	23,793
2	0.50	106	50	4.5	11,987	5293	23,793
3	0.90	118	8	3.16	2673	940	2970
4	0.35	15	30	13.5	2115	1800	6075
5	0.62	25	20	4.5	1405	500	2248
6	0.76	40	25	3.16	2411	1000	3160
7	0.4	25	15	10.5	1590	1125	3938
8	0.32	40	40	9	4640	3200	14,400
9	0.27	60	22	13.5	4854	5280	17,820
10	0.69	90	20	3.5	4370	1800	6300
11	0.70	45	60	3.2	6072	2700	8640
12	0.58	50	50	4	5800	2500	10,000
13	0.56	100	50	4	11,200	5000	20,000

- infiltration losses according to Appendix C of [50];
- a heat gain of 230 W per piece of equipment (one piece for each office worker and one piece per 50 meeting people); and
- the estimation of the presence of office workers with sedentary activity (1 met).

Furthermore, based on the heating and cooling periods, a heating period was set from 1st December to 31st March and a cooling period from 1st June to 30th September, eliminating weekdays and holidays, for 8 h per day based on the office occupancy rate.

The results obtained from the dynamic simulations were validated thanks to a model calibration. For the model validation, data recorded by two-channel Hobo-U10 Temp/RH temperature sensors positioned in some office rooms of the building, was used.

For a period from 25th February to 17th May 2006, the indoor air temperature and the indoor air average relative humidity trends were monitored. For example, the data relating to an area for office use located on the second floor is reported. This office was unoccupied for the entire period and, therefore, characterised by a low air turnover and negligible temperature changes.

In Fig. 3, the comparison between the hourly measured and simulated indoor air temperature is illustrated. As reported in Mustafaraj et al. [51] and Royapoor et al. [52], according to the main standards or guidelines (ASHRAE Guideline 14 [53], Measurement and Verification of Federal Energy Projects (FEMP) [54] and International Performance Measurement and Verification Protocol (IPMVP) [55]), the authors could validate the "Base-Case" model. In particular, two error indices were calculated: the Normalized Mean Bias Error (NMBE) and the Coefficient of Variation of the Root Mean Square Error (CV-RMSE):

- the NMBE (Eq. (1)) is a normalisation of the Mean Bias Error and provides the global bias between the expected and predicted data. Positive values of this index mean that the model provides an underestimated value with respect to the expected data. Negative values mean that the model provides an overestimated output data [56,57].

$$NMBE = 100 \cdot \frac{1}{N} \frac{\sum_{i=1}^{N} (x_i - y_i)}{\bar{x}} \quad (1)$$

- the CV-RMSE (Eq. (2)), providing a measure of the variability of the error between the expected and predicted data, is one of the most important measurements for evaluating the goodness-of-fit of the forecast model [58]. It provides a clear indication of the forecast ability of the model in the field of building energy evaluation [57,59]

$$CV - RMSE = 100 \cdot \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2}}{\bar{x}} \quad (2)$$

In Table 6, the limit values and ranges of applicability of all criteria for both indices for hourly calibration are reported; furthermore, in the last column the NMBE and the CV-RMSE calculated for the "Base-Case" model are indicated.

For all criteria the "Base-Case" model was calibrated; the NMBE is within the applicability ranges required and CV-RMSE is lower than the specified limit values.

4.2. Parametric simulation

As explained previously, to obtain a generic database useful for developing a reliable forecast model, it was necessary to perform a parametric simulation in a TRNSYS environment. Based on the calibrated model, it was possible to develop several scenarios for analysing the energy demand, varying different boundary conditions and several geometrical properties. From the "Base-Case", it was possible to construct an "ideal building" model that, by means of a parametric simulation, was simulated 1560 times, each time varying the shape factors, the heated volume, the building construction type, the thermo-physical features, the heating/cooling operational period, the climatic zones, the cities and the building orientation. It was, therefore, possible to generate a large building energy database representative of non-residential Italian building stocks designed with high energy performance [44]. An "ideal building" model for each climatic zone was developed. Varying the climatic zones change the limit transmittance values; the values of the thermal transmittance (U) used for the five climatic zones are collected in Table 7.

Each model was simulated for 13 geometrical configurations in 5 climatic zones represented by 3 different cities (Table 3). Moreover, because the building orientation and the wall azimuth influences the solar radiation received on the façade, each model was simulated eight

Table 5
Envelope thermal features of the ideal building model.

Components	Layers	Materials	Conductivity [W/mK]	Density [kg/m^3]	Thermal capacity [kJ/kg K]	Thickness [m]
External Wall	1	External coating	1.00	1800	0.84	0.02
	2	Lime cement	0.90	1800	0.96	0.015
	3	Tuff block	0.63	1500	0.70	0.30
	4	Internal Plaster	0.70	850	0.96	0.02
Floor	1	Cement Brick	2.00	2500	0.88	0.02
	2	Cement Screed	1.40	2000	1.20	0.06
	4	Concreate slab	1.91	1400	1.00	0.25
	5	Internal Plaster 2	0.70	800	0.837	0.02
Roof	1	External tiles	1.10	2100	0.84	0.01
	2	Bitumen	0.17	1200	1.40	0.02
	3	Lime cement	1.40	2000	1.20	0.06
	5	Concreate slab	1.91	1400	1.00	0.25
	6	Internal Plaster 2	0.70	800	0.84	0.02

times, varying the orientation by 45° each time, and averaging the results (for a total of 1560 simulations). In D'Amico et al. [44], the results obtained from the parametric dynamic simulation are collected.

4.3. White box method: strengths and weaknesses

The use of a white box method to solve the energy balance is a good solution but can be considered reliable only if the dynamic model is calibrated. As explained previously, the identification of the best software tool is not always simple and an expert user of the investigated problem and of the software language is necessary. Any simulation needs the collection of a multitude of parameters, which are not always easy to select or to implement. For careful building energy analysis, a preliminary collection and investigation phase is necessary. After calibrating the "Base-Case" model and implementing other reliable scenarios with a parametric simulation in order to extrapolate a generic relation that permits the identification of the energy demand of a generic building, all results must be analysed and elaborated because each single simulation is the answer to a specific condition. For this reason, the authors decided to explore an alternative method belonging to the black box category.

5. Black box methods

Although the application of comprehensive methods by means of a dynamic simulation software tool represents the optimal solution for evaluating building energy performance, the high number of difficulties encountered in the implementation of the model has led many researchers to study and develop alternative resolution techniques such as those represented by black box methods. Thanks to the availability of a large, generic database (Section 4.2), built through the application of a parametric simulation on 1560 models in the TRNSYS environment, this alternative approach could be applied with optimal results using the correlations between the expected and predicted data, as found in

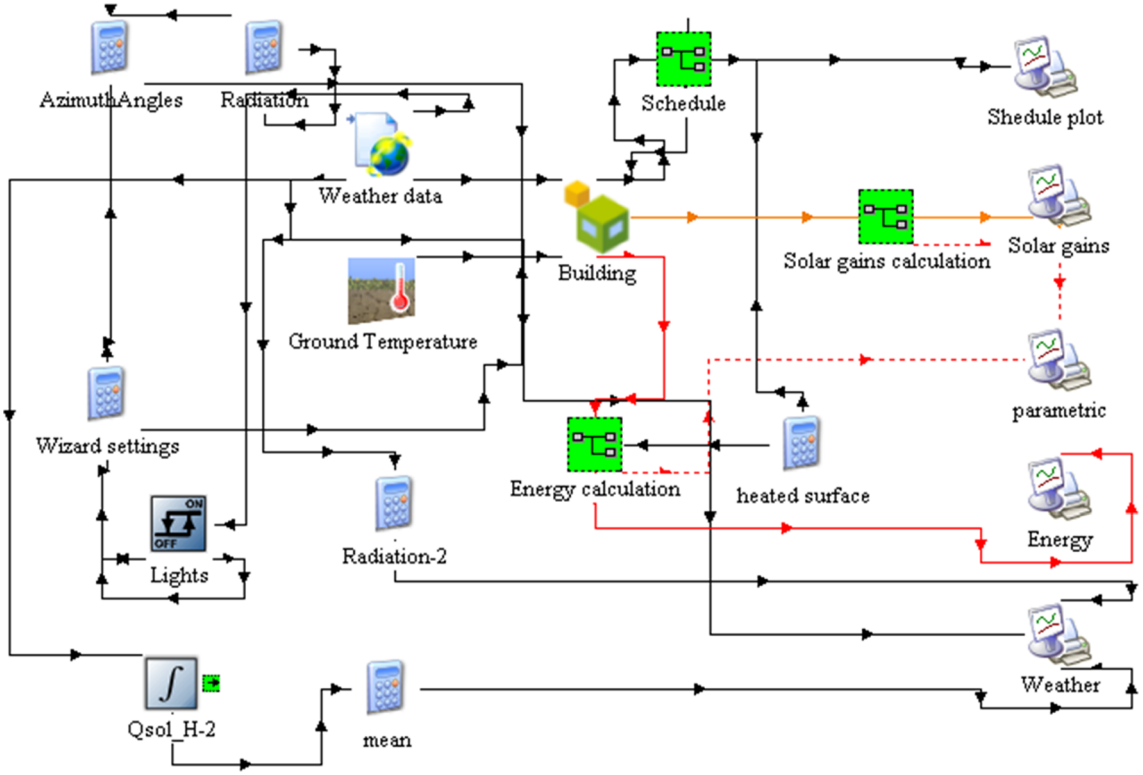


Fig. 2. TRNSYS schema.

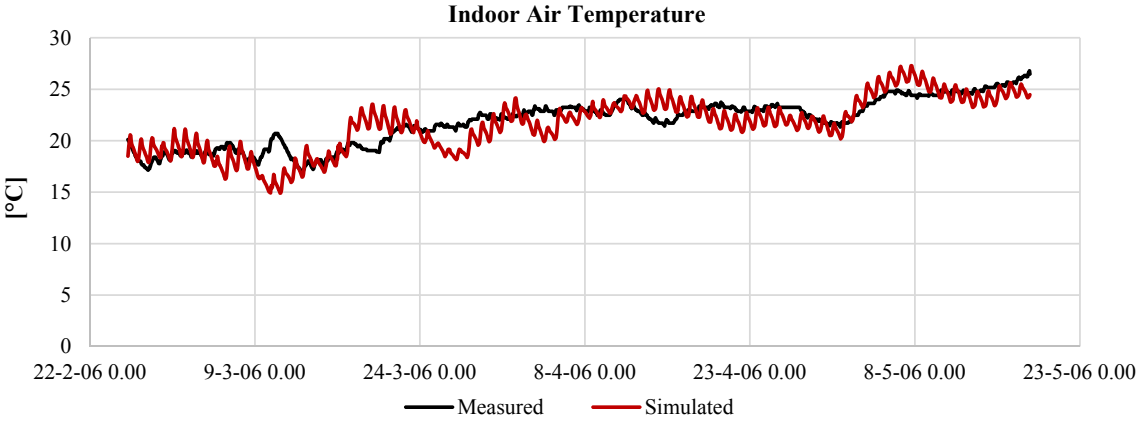


Fig. 3. Comparison between the measured and simulated indoor air temperature trend.

Table 6
Criteria and error indices for the model calibration.

Calibration Criteria	Index	FEMP	ASHRAE	IPMVP	Base Case
Hourly criteria %	NMBE	± 10	± 10	± 5	+1.33
	CV-RMSE	30	30	20	8.13

Table 7
Thermal transmittance values used in the TRNSYS models.

Climatic zone	A-B	C	D	E	F
	U_{model} [W/(m^2·K)]				
U_{wall}	0.444	0.379	0.336	0.297	0.276
U_{roof}	0.377	0.353	0.303	0.249	0.234
U_{floor}	0.445	0.385	0.307	0.287	0.268
U_{window}	2.760	2.260	1.760	1.760	1.40

the literature [60,61]. For this reason, the authors explored the applicability of the black box method developing a linear regression model. In order to validate this method, the authors used 85% of the available data for the determination of the MLR model equations, while the remaining 15% was used to evaluate and test the reliability achieved by each relationship. To provide some information on the reliability of each model, a first analysis on the distribution of residuals (differences between expected and predicted values by the models) through their representation in scatter plots was conducted. Despite being a simplistic analysis, this provides the first feedback on the goodness of fit of the built model; a distribution of the residuals around zero is indicative of model accuracy in forecasting building energy needs. However, deep statistical analysis on forecasting model errors should satisfy five evaluation criteria:

1. measurement validity;
2. reliability;
3. ease of interpretation;
4. clarity of presentation; and
5. support of statistical evaluation.

Hence, the authors provided the following statistical errors [62]:

- the Mean Absolute Error (MAE) represents the direct deviation between expected and predicted output values (Eq. (3)) [59]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{3}$$

- the Mean Square Error (MSE) calculates the variance between the target of a model and what is going to be forecasted (Eq. (4)) [63]:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{4}$$

- the Root Mean Square Error (RMSE) represents the square root of the quadratic mean of the differences between predicted and expected values (Eq. (5)).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{5}$$

- the Mean Absolute Percentage Error (MAPE) evaluates the absolute percentage deviation between the predicted and expected values. It indicate the percentage error size that could be used as a measure of the quality of a model's output (Eq. (6)) [64]:

$$MAPE = 100\cdot\frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - y_i|}{x_i} \tag{6}$$

- the determination coefficient (R^2) evaluates the manner in which a model approximates the real data points, which is a measure of the predictability degree of the model [65]; the higher R^2, the more efficient the developed model (Eq. (7)) [66]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{7}$$

where

x_i is the i-th expected output;
y_i is the i-th predicted output;
$\bar{x}$ is the average of the whole desired output; and
N is the number of the identification set samples.

The MAE, MSE, RMSE and MAPE allow a comparison of the deviation between the predicted and expected values of the building energy demand [65,67]. However, because the first three are based on absolute errors, it is not possible to identify a specific criterion to find an optimal value for each of them, but smaller values correspond to more precise models. Instead, the MAPE, being independent of the scale, is more significant [67].

6. MLR model

The multiple linear regression model allows an immediate assessment of building energy requirements. As discussed before, an MLR

model is one of the black box categories and one of the easiest and most intuitive approaches of prediction. This method, excluding a knowledge of the physical phenomena, still allows the prefixed objective to be reached without excessive computational cost. Nonetheless, a knowledge of a large survey database on which the model can be constructed is necessary. Therefore, if compared to the physical model, MLR models have the advantage of minimising the amount of input data, avoiding tedious work and the necessity for powerful informatics equipment [62]. The aim of this method is to explain the relationship between the dependent variable (annual heating, cooling or comprehensive energy demand) and two or more explanatory variables or regressors (climate and thermophysical parameters) using linear combinations of the latter [68]. The MLR models were developed according to the most general equation form (Eq. (8)) [69]:

$$y_i = b_0 + b_1x_1 + b_2x_2 + ...+b_px_p + e \quad (8)$$

where

y_i represent the i-th independent variable (output);
x_i represent the i-th explanatory variable (input);
b_0 is the intercept of the relationship;
b_i is the i-th regression coefficient that determines the used weight by the equation on the i-th explanatory variable to provide the estimate output; and
e is the error related to the i-th observation.

The objective function for constructing the MLR model is the least square method, with the goal of minimising the sum of the least square errors between the expected and predicted outputs as illustrated in the following equation (Eq. (9)) [68]:

$$Min \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}b_j - b_0\right)^2 \quad (9)$$

6.1. Sensitivity analysis and input variable selection

Owing to the complexity of the building energy balance resolution, selection of the phenomenon explanatory variables is a crucial step in the modelling of forecasting methods, because the input data determines both the equation form and the partial regression coefficient values that affect the results [70]. This is widely recognised by the scientific community and input data selection is applied in many works: in Lahouar et al. [71], an autocorrelation plot was used to identify the input that most influences the output variables; in Gunay et al. [72] and Kapetanakis et al. [73], the Pearson and Spearman correlation coefficients were applied to identify the strongest correlation between the building load and weather parameters. Other input selection methods are represented by the clustering methods; for example, Yan Ding et al. [74] applied the K-means and hierarchical clustering methods to study the accuracy of cooling load prediction models in office buildings influenced by the input data. In the same manner, David Hsu [75] used the K-means and clusterwise regression methods for an energy needs prediction model. To identify the mean parameters that mostly influence the heating and cooling energy demands of the building stock studied, the authors applied the Pearson correlation coefficient (r) analysis. This method, deducing simple correlations between the explanatory variables and the dependent variable, is one of the simplest and fastest methods for selecting and identifying the most influential input variables useful for forecast models [70]. Given two statistical variables, the Pearson correlation r coefficient is defined as the ratio between the covariance of the two variables and the standard deviation of each as indicated in the following (Eq. (10)):

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad (10)$$

where σ_{xy} is the covariance between the x and y variables and is calculated as:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y}) \quad (11)$$

and σ_x and σ_y are the standard deviation of each variable and are calculated as:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \quad (12)$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} \quad (13)$$

The r coefficient measures the linear correlation between the two analysed variables and it may assume a value between $-1 < r < 1$; the value 1 represents a total positive linear correlation, the value -1 indicates a total negative linear correlation and 0 means that there is not a linear correlation.

The authors calculated the r coefficient for each parameter representative of the energy database constructed in Section 4.2 and then applied a sensitivity analysis to identify those parameters that affect the building thermal balance more and that can be used in the MLR model. In the following graphs (Figs. 4–9), the linear regression of the main variables affecting the dynamic behaviour of the "ideal building" model both for heating and cooling energy demand are illustrated: *HDD*, *CDD*, external temperature (T), S/V, glazed surface (S_w), opaque surface (S_{op}) and internal gains (Q_G). For each trend, the determination coefficient (R^2) and the r coefficients are also displayed.

A first criterion for the identification of the significant variables for the studied phenomenon could be that of sorting the variables in descending order of the absolute value of the coefficient r and selecting those that have a value of r significantly different from zero. Another identification criterion is represented by an empirical rule that (for high value of n) selects those variables in which the value of r is greater than

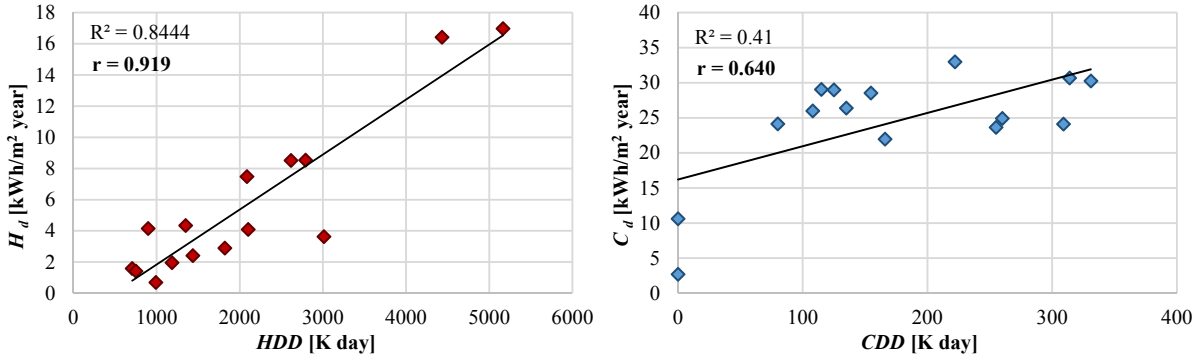


Fig. 4. Linear regression analysis between the H_d and *HDD* (a) and between the C_d and *CDD* (b).

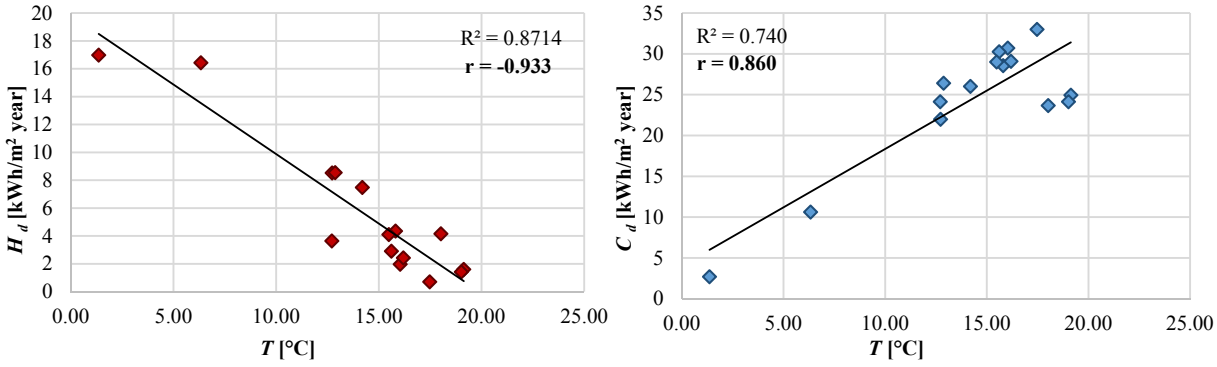


Fig. 5. Linear regression analysis between the H_d and T (a) and between the C_d and T (b).

$2/\sqrt{n}$ [76]. In Fig. 10, sensitivity analysis among the input variables and the heating/cooling energy demand based on the r coefficient is displayed.

To calculate the r correlation coefficient, the values assumed in each selected city (for the climatic parameters) and the values assumed in each "ideal building" model (for the thermophysical and geometric parameters) were considered. In this manner, applying the empirical selection criterion previously described, only the values with an r correlation coefficient greater than 0.55 can be considered for the implementation of the regression model. Based on these considerations and on the sensitivity analysis emphasised in Fig. 10, the *HDD*, *T*, *S/V* and S_w for the heating energy demand forecast and *CDD*, *T* and S_{op} for the cooling energy demand evaluation should be selected. However, based on the previous results obtained in Ciulla et al. [48,77] and in D'Amico et al. [44], the authors, also for the building cooling load evaluation, considered the *S/V* parameter indispensable. Instead, because the two climatic indices are a function of the external temperatures, it was decided to exclude the temperature from input variables in the linear regression model because of its redundancy. Further, the determination of *HDD* and *CDD* data, often tabulated in laws and standards, is easier than determining the average monthly temperatures Regarding the high values of linear correlation assumed by the glazed and opaque surfaces for the heating and cooling energy demand respectively, it is possible to affirm that, fixing all other conditions, with increases of the glazed surface the solar gain increases and obviously H_d decreases and C_d increases.

6.2. MLR evaluation

The investigation of the MLR method allowed the identification of the best correlation form for determining H_d, C_d and E_d. For the heating and cooling demand two correlations were identified for each of them; the first is a function of the weather index and the shape factor, and the second is also a function of S_w for H_d and S_{op} for C_d. Regarding E_d evaluation, the authors proposed two equation forms that considered *HDD*, *CDD* and *S/V* simultaneously, and another correlation in which the dependence from S_w and S_{op} are also indicated.

6.2.1. MLR and heating energy demand evaluation

The first form of the heating energy demand as a function of *HDD* and *S/V* is represented by Eq. (14):

$$H_d = k + \alpha_1 \cdot HDD + \alpha_2 \cdot \frac{S}{V} \tag{14}$$

where

α_1 is the first regression coefficient [kWh/(m^2 year K day)];
α_2 is the second regression coefficient [kWh/m year]; and
k is the intercept [kWh/m^2 year].

The graphical representation of Eq. (14) is plotted in Fig. 11.

While the second form of H_d as a function of *HDD*, *S/V* and S_w is represented by Eq. (15):

$$H_d^* = k + \alpha_1^* \cdot HDD + \alpha_2^* \cdot \frac{S}{V} + \alpha_3^* \cdot S_w \tag{15}$$

where

α_1^* is the first regression coefficient [kWh/(m^2 year K day)];
α_2^* is the second regression coefficient [kWh/m year];
α_3^* is the third regression coefficient [kWh/m^5 year]; and
k is the intercept [kWh/m^2 year].

The values of all regression coefficients, intercepts and R^2 for both equations, obtained from the application of the least square method between the expected and predicted outputs for 85% of the database values are collected in Table 8; in both cases R^2 is close to 0.9.

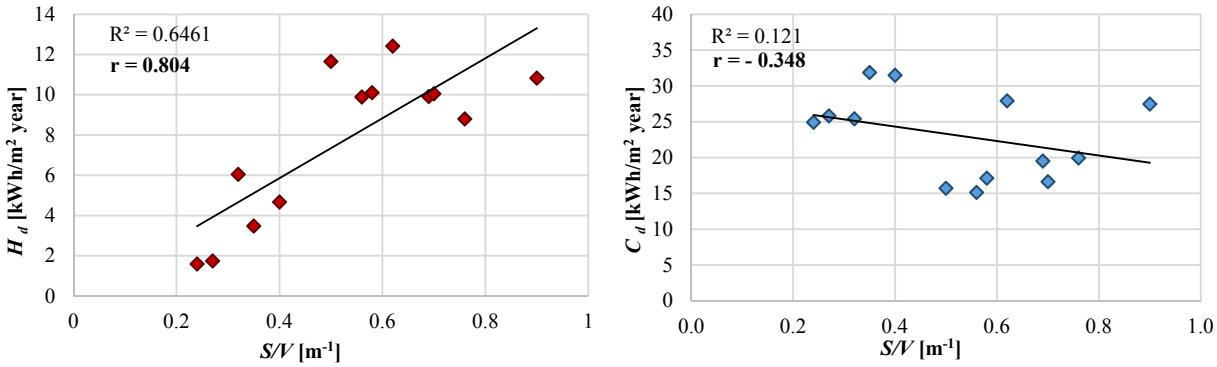


Fig. 6. Linear regression analysis between the H_d and S/V (a) and between the C_d and S/V (b).

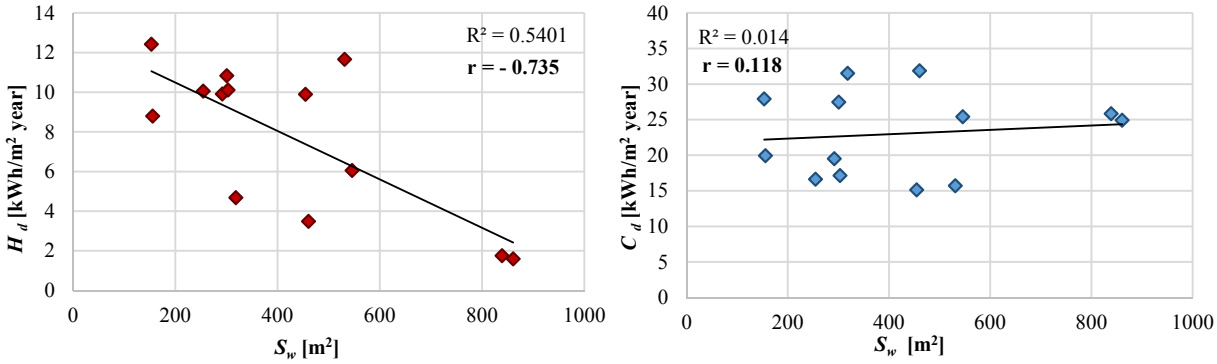


Fig. 7. Linear regression analysis between the H_d and S_w (a) and between the C_d and S_w (b).

6.2.2. MLR and cooling energy demand evaluation

In the same way, the first form of the cooling energy demand as a function of *CDD* and *S/V* is represented by Eq. (16) and is plotted in Fig. 12:

$$C_d = k + \beta_1 \cdot CDD + \beta_2 \cdot \frac{S}{V} \quad (16)$$

where

β_1 is the first regression coefficient [kWh/(m^2 year K day)];
β_2 is the second regression coefficient [kWh/m year]; and
k is the intercept [kWh/m^2 year].

The second form is represented by Eq. (17):

$$C_d^* = k + \beta_1^* \cdot CDD + \beta_2^* \cdot \frac{S}{V} + \beta_3^* \cdot S_{op} \quad (17)$$

where

β_1^* is the first regression coefficient [kWh/(m^2 year K day)];
β_2^* is the second regression coefficient [kWh/m year];
β_3^* is the third regression coefficient [kWh/m^5 year]; and
k is the intercept [kWh/m^2 year].

Also in this case, the values of all regression coefficients, intercepts and R^2 for both equations were obtained from the application of the least square method for 85% of the data.

It should be noted that some of the data from the sample was purged due to an inconsistency between the value of *CDD* and the demand value for cooling calculated with the TRNSYS models. More specifically, for the cities where the *CDD* value was not initially provided, the parameter was calculated by the authors, and in particular for Cortina and Sestriere, the value of *CDD* was assessed as zero (Section 3). These values could imply the non-ignition of the cooling system, but since the current standard establishes a standard cooling period valid for all Italian cities without distinction of area, the simulation in TRNSYS has provided an unjustified cooling requirement. Therefore, for the determination of the cooling energy demand, it was agreed to eliminate the values linked to the models of the cities of Cortina and Sestriere (26 fewer scenarios). In Table 9, all parameters of Eqs. (16) and (17) are collected and, in general, the R^2 values are higher than 0.9.

6.2.3. MLR and comprehensive energy demand evaluation

To determine the comprehensive energy demand, two different forms of correlation were investigated. As indicated in Eq. (18), the first form considers, as a first explanatory variable, the sum of the *HDD* and *CDD* indices and the regression plan is plotted in Fig. 13:

$$E_d = k + \gamma_1 \cdot (HDD + CDD) + \gamma_3 \cdot \frac{S}{V} \quad (18)$$

where

γ_1, is the first regression coefficient [kWh/(m^2 year K day)];
γ_2 is the second regression coefficient [kWh/m year]; and
k is the intercept [kWh/m^2 year].

The second correlation form, instead, considers the two weather indices in two different explanatory variables (Eq. (19)):

$$E_{d1} = k + \gamma_1 \cdot HDD + \gamma_2 \cdot CDD + \gamma_3 \cdot \frac{S}{V} \quad (19)$$

where

γ_1, γ_2 are the first and second regression coefficients [kWh/(m^2 year K day)];
γ_3 is the third regression coefficient [kWh/m year]; and

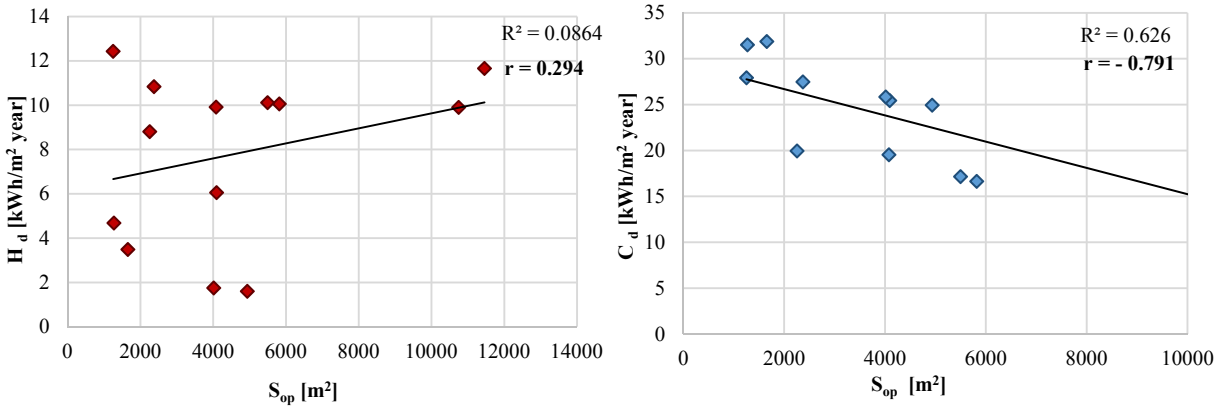


Fig. 8. Linear regression analysis between the H_d and S_{op} (a) and between the C_d and S_{op} (b).

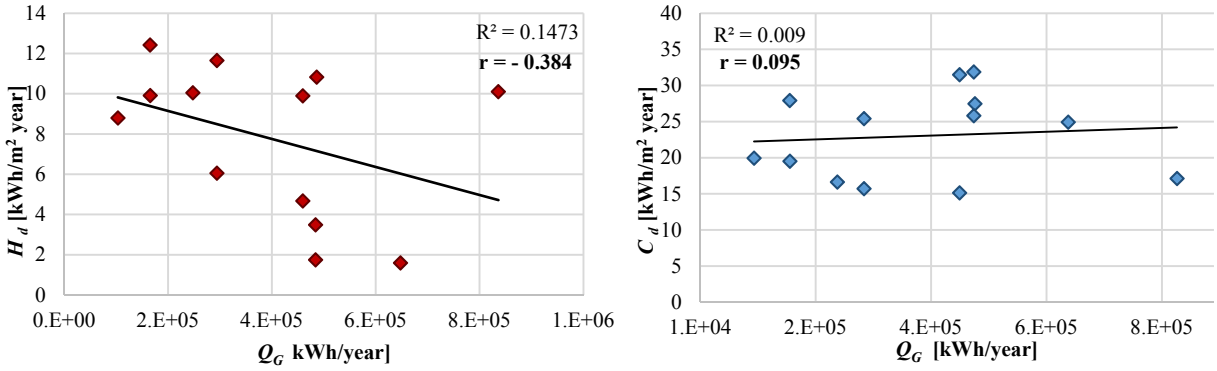


Fig. 9. Linear regression analysis between the H_d and Q_G (a) and between the C_d and Q_G (b).

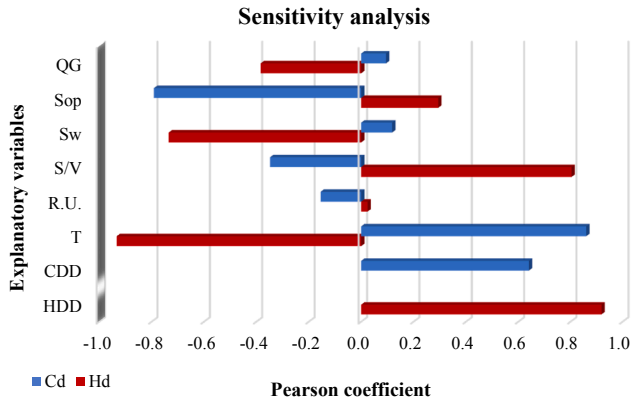


Fig. 10. Pearson correlation coefficients of input variables for H_d and C_d.

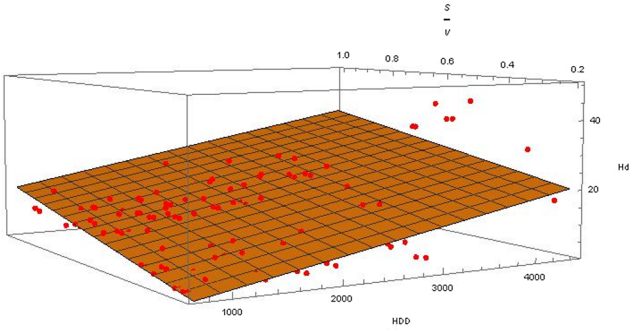


Fig. 11. Scatter plot and regression plan for the H_d.

Table 8
Partial regression coefficient and R^2 for the H_d.

	k	α_1	α_2	α_3	R^2
H_d	−7.3203	0.0053781	19.4008	–	0.898
H_d^*	−2.3015	0.0053839	14.4288	−0.0056909	0.900

k is the intercept [kWh/m^2 year].

Finally, to consider the strong correlation among the energy demand and the S_w and S_{op} parameters, a more complicated correlation is proposed in which the value of E_d is a function of five parameters (Eq. (20)):

$$E_d^* = k + \gamma_1^* \cdot HDD + \gamma_2^* \cdot CDD + \gamma_3^* \cdot \frac{S}{V} + \gamma_4^* \cdot S_w + \gamma_5^* \cdot S_{op} \quad (20)$$

where

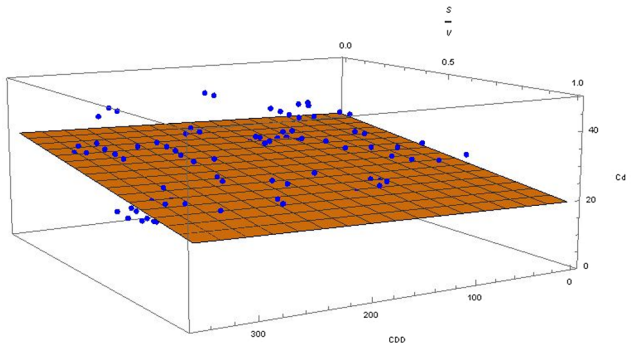


Fig. 12. Scatter plot and regression plan for the C_d.

Table 9
Partial regression coefficient and R^2 for the C_d.

	k	β_1	β_2	β_3	R^2
C_d	30.5767	0.0064923	−11.0297	–	0.906
C_d^*	41.4031	0.0041604	−13.0856	−0.0020440	0.962

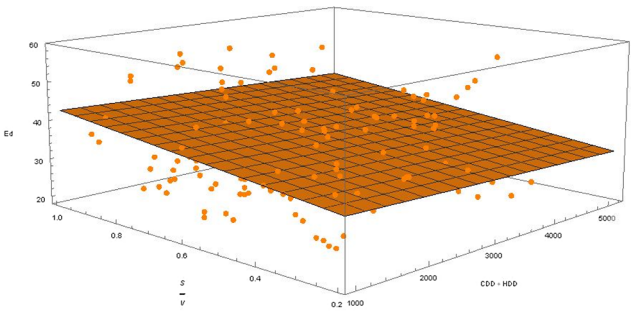


Fig. 13. Scatter plot and regression plan for the E_d.

γ_1^*, γ_2^* are the first and second regression coefficients [kWh/(m^2 year K day)];
γ_3^* is the third regression coefficient [kWh/m year];
γ_4^*, γ_5^* are the fourth and fifth regression coefficients [kWh/m^5 year]; and
k is the intercept [kWh/m^2 year].

The collection of the regression coefficients and the intercept values for each correlation, and the comparison of the determination coefficients is reported in Table 10.

The results confirm that the use of *HDD* and *CDD* as a unique explanatory variable or two distinct variables is indifferent, so much so that the determination coefficient is the same; in all cases higher than 0.95.

Table 10
Partial regression coefficient and R^2 for the E_d.

	k	γ_1	γ_2	γ_3	γ_4	γ_5	R^2
E_d	32.5597	−0.0006188	10.7855	–	–	–	0.950
E_{d1}	33.6326	−0.0008445	−0.0041735	10.8133	–	–	0.950
E_d^*	49.342	−0.0008874	−0.0058240	−1.35286	−0.0131923	−0.0007279	0.959

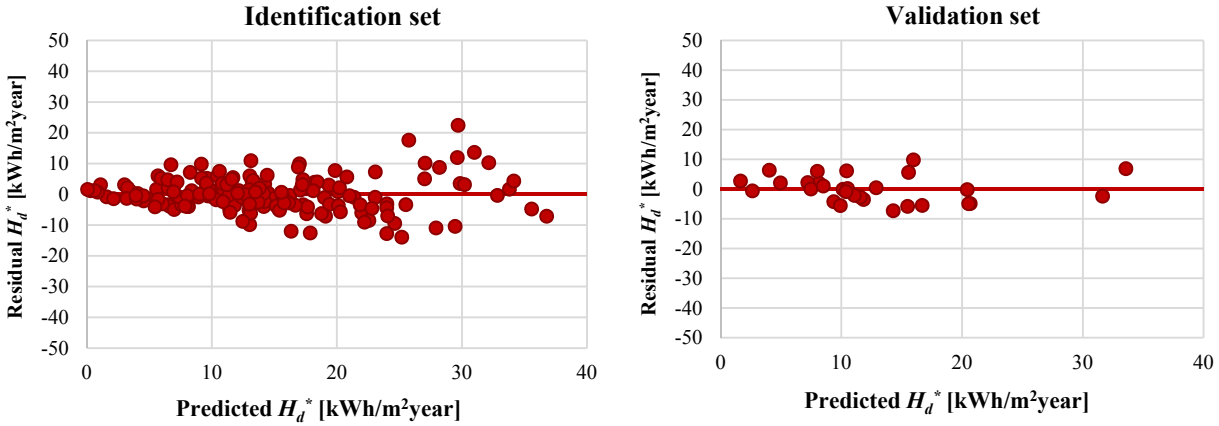


Fig. 14. Residual trend of H_d^* correlation for identification and validation set.

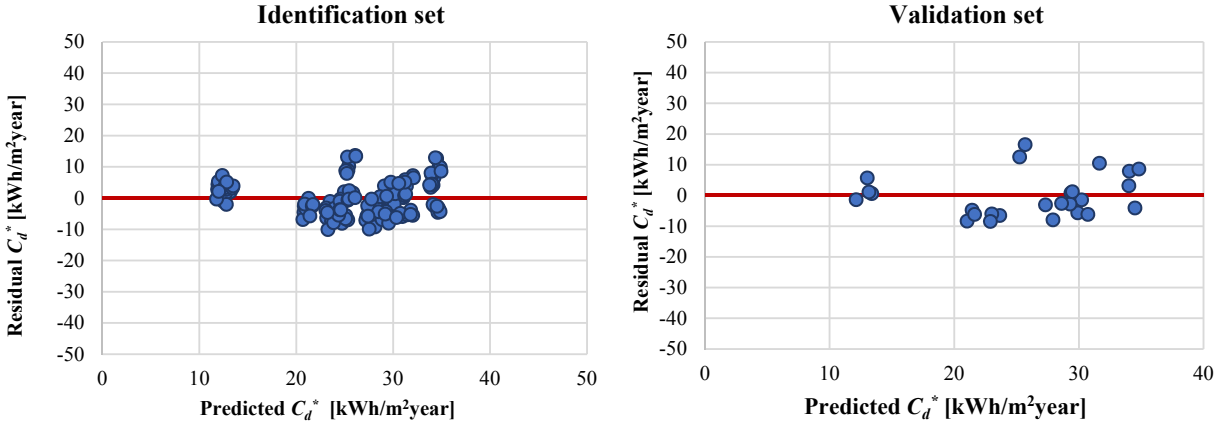


Fig. 15. Residual trend of C_d^* correlation for identification and validation set.

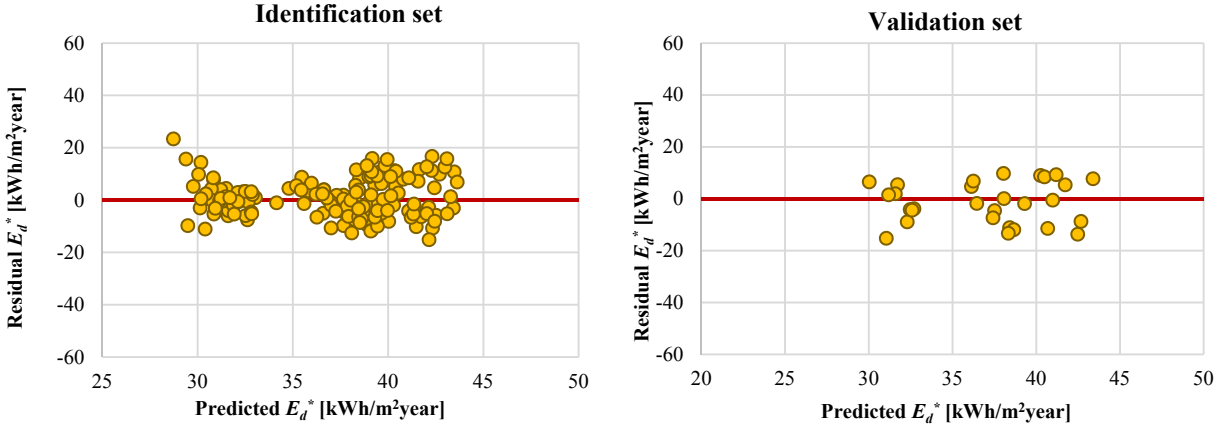


Fig. 16. Residual trend of E_d^* correlation for identification and validation set.

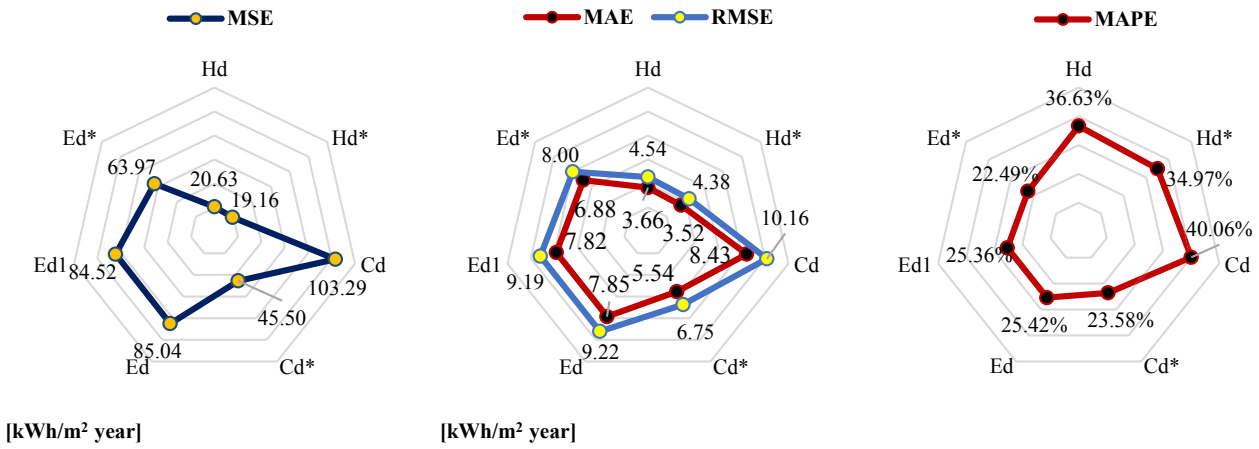


Fig. 17. Statistical analysis of the Validation set for the MLR models.

Table 11
MLR correlations and respectively statistical errors.

Correlations	R^2	MAE	MSE	RMSE	MAPE
$H_d = -7.3203 + 0.0053781 \cdot HDD + 19.4008 \cdot \frac{S}{V}$	0.90	3.66	20.63	4.54	37%
$H_d^* = -2.3015 + 0.0053839 \cdot HDD + 14.4288 \cdot \frac{S}{V} - 0.0056909 \cdot S_w$	0.90	3.52	19.16	4.38	35%
$C_d = 30.5767 + 0.0064923 \cdot CDD - 11.0297 \cdot \frac{S}{V}$	0.91	8.43	103.3	10.16	40%
$C_d^* = 41.4031 + 0.004604 \cdot CDD - 13.0856 \cdot \frac{S}{V} - 0.002044 \cdot S_{op}$	0.96	5.54	45.50	6.75	24%
$E_d = 32.5597 - 0.0006188 \cdot (HDD + CDD) + 10.7855 \cdot \frac{S}{V}$	0.95	7.85	85.04	9.22	25%
$E_{d1} = 33.6326 - 0.0008445 \cdot HDD - 0.0041735 \cdot CDD + 10.8133 \cdot \frac{S}{V}$	0.95	7.82	84.52	9.19	25%
$E_d^* = 49.342 - 0.0009 \cdot HDD + -0.0058 \cdot CDD - 1.3527 \cdot \frac{S}{V} - 0.0132 \cdot S_w - 0.0007 \cdot S_{op}$	0.96	6.88	63.97	8.00	22%

7. Results and discussion

The analysis of the results obtained from the application of the MLR model to the evaluation of building energy performance confirms that this procedure is a valid alternative to a more complex method. All correlations identified for the heating, cooling and comprehensive energy demand are characterised by optimal determination coefficients higher than 0.9. In all cases, the more complex correlations (H_d^*, C_d^* and E_d^*) are the best. For these correlations, in the following graphs (from Figs. 14–16), the residual values calculated for the identification and validation set are displayed. As previously explained, only 85% of the total data was used to determine the correlations, while 15% was used to validate these.

Put simply, a residual is the error in a result and in these cases the value is between ± 20%, both in the identification and validation sets.

In Fig. 14, the residual trends of H_d^* correlation for all data from the identification and validation set are plotted, whereas in Fig. 15, there are the residual trends of C_d^* correlation. In this second case, as explained in Section 6.2.2, the data sample is represented by a lower number of cases because there are no model results related to the cities of Cortina and Sestriere. In Fig. 16, the residual trends of E_d^* correlation are plotted.

In addition to the calculation of R^2 values, to validate the reliability of the MLR models, four other statistical errors were calculated: the MSE, MAE, RMSE and MAPE. In Fig. 17, the statistical analysis of the error based on the validation dataset is represented.

The MSE distribution highlighted as the best performance is related to the heating energy demand correlations (Eqs. (14) and (15)), while the worst is the cooling energy demand C_d (Eq. (16)). Among the energy comprehensive correlations, the best is E_d^* (Eq. (20)). The same considerations are valid for MAE and RMSE. As for MAPE, the best results are indicated by E_d^*, while the heating energy demand correlations are less efficient. Generally, in all cases, the solution E_d^*, H_d^* and C_d^* are the best correlations for solving the thermal energy balance of a building; these results are also confirmed by the high R^2 values determined in Section 6.2. In the following (Table 11), all correlations and respective statistical errors are collected.

As explained previously, the more complicated correlations are characterised by better quality and reliability; in general, the high value of R^2 and the low values of MAE and RMSE justify the use of the MLR methodology as a good alternative for determining the building energy performance. The MLR method represents a simple and immediate tool which can solve a complex problem, such as the building energy balance, and can accelerate and help some aspects of energy planning.

8. Conclusion

In this work, the authors explain that the selection of the most suitable method for solving a determinate problem is important because it allows to overcome certain limits, in order to identify a generic solution able to interpret any condition and to accelerate the resolution with high reliability. After a review of the main types of methods for solving the building energy balance widespread in literature, the authors investigated two of these: a comprehensive analysis with TRNSYS software and the Multiple Linear Regression method.

As explained in the paper, the first method, belonging to the white box category, allows the determination of the building energy performance with a high degree of reliability if the model is correctly developed and calibrated. Indeed, high reliability is a function of a detailed data collection phase (representative of the model), careful calibration, and the presence of an expert user who knows the software tool language and the studied physical phenomena. These conditions permit the development an accurate model which represents the actual conditions well. Based on this first result, in order to obtain a generic

solution, a parametric simulation was developed that solves the building energy balance, simultaneously changing the weather conditions, the shape factor and the thermophysical characteristics of the building. In this way, 1560 simulations of a representative building stock were obtained for non-residential buildings designed with high energy performance located in the Italian peninsula.

However, although the parametric simulation solves several scenarios, simultaneously obtaining 1560 results, it is not able to give a generic indication because each single simulation gives a single specific response for a model under certain boundary conditions and characterised by specific thermophysical choices. Indeed, to generalise the results, it is necessary to analyse all of the thermal energy results obtained from the parametric simulation. Careful sensitivity analysis on the 1560 simulation results, based on the identification of the Pearson coefficient, allowed the identification of the main parameters that influence the building thermal balance during the heating, cooling and entire climatisation periods. Thanks to this analysis and the use of all simulation data, the authors decided to explore the Multiple Linear Regression technique belonging to the black box methods. This method allowed a linear relationship to be modelled between two or more explanatory variables, which represent the inputs of the model and a response variable through a fitting procedure. As a result, some simple correlations were developed knowing only a few groups of well-known parameters, and identifying the heating, cooling and comprehensive energy needs of a building with a high degree of reliability. Indeed, these correlations are characterised by optimal statistical error values; for example, the determination coefficients are higher than 0.9 and the Mean Absolute Error and Root Mean Square error are lower than 10 kWh/m^2 year. The reliability and flexibility of the energy database allowed the identification of solutions that simultaneously respond to changes in climate and building shape factor, obtaining generic solutions which can explain any possible building topology in any conditions.

The promising results justify the use of Multiple Linear Regression as an alternative method, issuing a simple and immediate tool that can solve a complex problem like building energy balance, thereby accelerating and helping some evaluation phases in energy planning, presenting a valid criteria that could be indicated in standards and laws in the field of the building energy performance.

References

[1] European Parliament and Council. Directive 2010/31/EU on the energy performance of buildings. Off J Eur Union 2010. https://doi.org/10.3000/17252555.L_2010.153.eng.

[2] Poel B, van Cruchten G, Balaras CA. Energy performance assessment of existing dwellings. Energy Build 2007;39:393–407. https://doi.org/10.1016/j.enbuild.2006.08.008.

[3] Balaras CA, Gaglia AG, Georgopoulou E, Sarafidis Y, Lalas D, Mirasgedis S. European residential buildings and empirical assessment of the Hellenic building stock, energy consumption, emissions and potential energy savings. Build Environ 2007;42:1298–314. https://doi.org/10.1016/j.buildenv.2005.11.001.

[4] Zhao H-X, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16:3586–92. https://doi.org/10.1016/j.rser.2012.02.049.

[5] Foucquier A, Robert S, Suard F, Stéphan L, Jay A. State of the art in building modelling and energy performances prediction: a review. Renew Sustain Energy Rev 2013;23:272–88. https://doi.org/10.1016/j.rser.2013.03.004.

[6] Scafetta N, Fortelli A, Mazzarella A. Meteo-climatic characterization of Naples and its heating-cooling degree day areal distribution. Int J Heat Technol 2017;35. https://doi.org/10.18280/ijht.35sp0119.

[7] Atalla T, Gualdi S, Lanza A. A global degree days database for energy-related applications. Energy 2018;143:1048–55. https://doi.org/10.1016/j.energy.2017.10.134.

[8] Gi K, Sano F, Hayashi A, Tomoda T, Akimoto K. A global analysis of residential heating and cooling service demand and cost-effective energy consumption under different climate change scenarios up to 2050. Mitig Adapt Strateg Glob Chang 2018;23:51–79. https://doi.org/10.1007/s11027-016-9728-6.

[9] Al-Homoud MS. Computer-aided building energy analysis techniques. Build Environ 2001;36:421–33. https://doi.org/10.1016/S0360-1323(00)00026-3.

[10] White JA, Reichmuth R. Simplified method for predicting building energy consumption using average monthly temperatures. IECEC 96. Proc 31st intersoc energy convers eng conf IEEE; 1996. p. 1834–9. https://doi.org/10.1109/iecec.1996.553381.

[11] Westphal FS, Lamberts R. The use of simplified weather data to estimate thermal loads of non-residential buildings. Energy Build 2004;36:847–54. https://doi.org/10.1016/j.enbuild.2004.01.007.

[12] Crawley DB, Hand JW, Kummert M, Griffith BT. Contrasting the capabilities of building energy performance simulation programs. Build Environ 2008;43:661–73. https://doi.org/10.1016/j.buildenv.2006.10.027.

[13] Brun A, Spitz C, Wurtz E, Mora L. Behavioural comparison of some predictive tools used in a low-energy building. 11th Int IBPSA Conf Build Simul 2009. 2009. p. 1185–90.

[14] Woloszyn M, Rode C. Tools for performance simulation of heat, air and moisture conditions of whole buildings. Build Simul 2008;1:5–24. https://doi.org/10.1007/s12273-008-8106-z.

[15] ANSYS. ANSYS Fluent Software | CFD Simulation; 2012.

[16] COMSOL. Simulation Software COMSOL Multiphysics®; 1998 [n.d].

[17] CHAM. CHAM | PHOENICS; 2005.

[18] Wurtz E, Mora L, Inard C. An equation-based simulation environment to investigate fast building simulation. Build Environ 2006;41:1571–83. https://doi.org/10.1016/j.buildenv.2005.06.027.

[19] Haghighat F, Li Y, Megri AC. Development and validation of a zonal model - POMA. Build Environ 2001;36:1039–47. https://doi.org/10.1016/S0360-1323(00)00073-1.

[20] EnergyPlus™; n.d.

[21] ESP-r; n.d.

[22] Building Performance - Simulation Software | EQUA; n.d.

[23] Bonneau D, Rongere FX, Covalet D, Gautier B. Clim2000: Modular software for energy simulation in buildings. Proc IBPSA. 1993. p. 93.

[24] Woloszyn M, Rusaouen G, Covalet D. Whole building simulation tools: Clim 2000. IEA Annex 2004;41.

[25] Rode C, Grau K. Whole building hygrothermal simulation model. ASHRAE Trans 2003:572–82.

[26] Rode C, Grau K. Integrated calculation of hygrothermal conditions of buildings. Proc 6th Symp Build Phys Nord Ctries, vol. 1. 2002. p. 23–30.

[27] BuildOpt-VIE; n.d.

[28] Li Z, Han Y, Xu P. Methods for benchmarking building energy consumption against its past or intended performance: an overview. Appl Energy 2014;124:325–34.

[29] Bauer M, Scartezzini J-L. A simplified correlation method accounting for heating and cooling loads in energy-efficient buildings. Energy Build 1998;27:147–54. https://doi.org/10.1016/S0378-7788(97)00035-2.

[30] Westergren K-E, Högberg H, Norlén U. Monitoring energy consumption in single-family houses. Energy Build 1999;29:247–57. https://doi.org/10.1016/S0378-7788(98)00065-6.

[31] Pfafferott J, Herkel S, Wapler J. Thermal building behaviour in summer: long-term data evaluation using simplified models. Energy Build 2005;37:844–52. https://doi.org/10.1016/J.ENBUILD.2004.11.007.

[32] Ansari FA, Mokhtar AS, Abbas KA, Adam NM. A simple approach for building cooling load estimation. Am J Environ Sci 2005;1:209–12.

[33] Dhar A, Reddy TA, Claridge DE. A Fourier series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable. J Sol Energy Eng 1999;121:47–53.

[34] Dhar A, Reddy TA, Claridge DE. Modeling hourly energy use in commercial buildings with Fourier series functional forms. J Sol Energy Eng 1998;120:217–23.

[35] Parti M, Parti C. The total and appliance-specific conditional demand for electricity in the household sector. Bell J Econ 1980:309–21.

[36] Kialashaki A, Reisel JR. Modeling of the energy demand of the residential sector in the United States using regression models and artificial neural networks. Appl Energy 2013;108:271–80.

[37] Aydinalp-Koksal M, Ugursal VI. Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector. Appl Energy 2008;85:271–96. https://doi.org/10.1016/j.apenergy.2006.09.012.

[38] Olofsson T, Andersson S, Östin R. A method for predicting the annual building heating demand based on limited performance data. Energy Build 1998;28:101–8. https://doi.org/10.1016/S0378-7788(98)00004-8.

[39] Ekici BB, Aksoy UT. Prediction of building energy consumption by using artificial neural networks. Adv Eng Softw 2009;40:356–62. https://doi.org/10.1016/j.advengsoft.2008.05.003.

[40] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. https://doi.org/10.1007/BF00994018.

[41] Dong B, Cao C, Lee SE. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005;37:545–53. https://doi.org/10.1016/J.ENBUILD.2004.09.009.

[42] Lai F, Magoulès F, Lherminier F. Vapnik's learning theory applied to energy consumption forecasts in residential buildings. Int J Comput Math 2008;85:1563–88. https://doi.org/10.1080/00207160802033582.

[43] Ciulla G, D'Amico A, Lo Brano V, Traverso M. Application of optimized artificial intelligence algorithm to evaluate the heating energy demand of non-residential buildings at European level. Energy 2019. https://doi.org/10.1016/J.ENERGY.2019.03.168.

[44] D'Amico A, Ciulla G, Panno D, Ferrari S. Building energy demand assessment through heating degree days: the importance of a climatic dataset. Appl Energy 2019;242:1285–306. https://doi.org/10.1016/J.APENERGY.2019.03.167.

[45] Il Presidente della Repubblica. Regolamento recante norme per la progettazione, l'installazione, l'esercizio e la manutenzione degli impianti termici degli edifici ai fini del contenimento dei consumi di energia, in attuazione dell'art. 4, comma 4, della legge 9 gennaio 1991, n. 10. Gazz Uff Della Repubb Ital SO; 1993.

[46] Decreto 26 giugno 2015. Applicazione delle metodologie di calcolo delle prestazioni energetiche e definizione delle prescrizioni e dei requisiti minimi degli edifici; Adeguamento del decreto del Ministro dello sviluppo economico, 26 giugno 2009 - Linee guida nazionali per la cer; 2015.
[47] Ente Nazionale Italiano di Normazione. UNI 10349:2016 "Riscaldamento e raffrescamento degli edifici - Dati climatici. Ente Naz Ital Di Normaz; 2016.
[48] Ciulla G, Lo Brano V, D'Amico A. Modelling relationship among energy demand, climate and office building features: a cluster analysis at European level. Appl Energy 2016. https://doi.org/10.1016/j.apenergy.2016.09.046.
[49] Meteonorm- Global Meteorological Database- Version7. Software and data for engineers, planners and education; n.d.
[50] ISO, EN. EN ISO 13790: 2008. Energy performance of buildings-Calculation of energy use for space heating and cooling. Brussels: Eur Comm Stand (CEN); 2008.
[51] Mustafaraj G, Marini D, Costa A, Keane M. Model calibration for building energy efficiency simulation. Appl Energy 2014;130:72–85. https://doi.org/10.1016/j.apenergy.2014.05.019.
[52] Royapoor M, Roskilly T. Building model calibration using energy and environmental data. Energy Build 2015;94:109–20. https://doi.org/10.1016/j.enbuild.2015.02.050.
[53] ANSI/ASHRAE. ASHRAE Guideline 14: measurement of energy and demand savings; 2014.
[54] DOE US. M&V guidelines: measurement and verification for performance-based contracts - version 4.0. Fed Energy Manag Progr; 2015https://doi.org/10.1039/c8ew00545a.
[55] Efficiency Valuation Organization. International performance measurement and verification protocol: concepts and options for determining energy and water savings, vol. I. Energy Proj Financ Resour …; 2012. doi:DOE/GO-102002-1554.
[56] Yun K, Luck R, Mago PJ, Cho H. Building hourly thermal load prediction using an indexed ARX model. Energy Build 2012;54:225–33. https://doi.org/10.1016/j.enbuild.2012.08.007.
[57] Ruiz GR, Bandera CF. Validation of calibrated energy models: common errors. Energies 2017. https://doi.org/10.3390/en10101587.
[58] Chae YT, Horesh R, Hwang Y, Lee YM. Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. Energy Build 2016;111:184–94. https://doi.org/10.1016/j.enbuild.2015.11.045.
[59] Yezioro A, Dong B, Leite F. An applied artificial intelligence approach towards assessing building performance simulation tools. Energy Build 2008;40:612–20. https://doi.org/10.1016/j.enbuild.2007.04.014.
[60] Catalina T, Virgone J, Blanco E. Development and validation of regression models to predict monthly heating demand for residential buildings; n.d. https://doi.org/10.1016/j.enbuild.2008.04.001.
[61] Ciulla G, D'Amico A, Lo Brano V, Beccali M. ANN decision support tool for the prediction of the thermal energy performance of European top rated energy efficient non-residential buildings. Conf Proc 12th SDEWES Held Dubrovnk, 4 to 8 Oct 2017. 2017.
[62] Catalina T, Iordache V, Caracaleanu B. Multiple regression model for fast prediction of the heating energy demand. Energy Build 2013. https://doi.org/10.1016/j.enbuild.2012.11.010.
[63] Ahmad T, Chen H. Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches. Energy Build 2018;166:460–76. https://doi.org/10.1016/j.enbuild.2018.01.066.
[64] Xuan Z, Xuehui Z, Liequan L, Zubing F, Junwei Y, Dongmei P. Forecasting performance comparison of two hybrid machine learning models for cooling load of a large-scale commercial building. J Build Eng 2019;21:64–73. https://doi.org/10.1016/j.jobe.2018.10.006.
[65] Fud G. Deep belief network based ensemble approach for cooling load forecasting of air-conditioning system. Energy 2018;148:269–82. https://doi.org/10.1016/j.energy.2018.01.180.
[66] Elhami B, Khanali M, Akram A. Combined application of Artificial Neural Networks and life cycle assessment in lentil farming in Iran. Inf Process Agric 2017;4:18–32. https://doi.org/10.1016/j.inpa.2016.10.004.
[67] Son H, Kim C. Short-term forecasting of electricity demand for the residential sector using weather and social variables. Resour Conserv Recycl 2017;123:200–7. https://doi.org/10.1016/j.resconrec.2016.01.016.
[68] Deng H, Fannon D, Eckelman MJ. Predictive modeling for US commercial building energy use: a comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy Build 2018. https://doi.org/10.1016/j.enbuild.2017.12.031.
[69] Darlington RB, Hayes AF. Regression analysis and linear models: concepts, application and implementation; 2016. https://doi.org/10.1016/0141-1187(83)90072-X.
[70] Abdipour M, Younessi-Hmazekhanlu M, Ramazani SHR, Hassan Omidi A. Artificial neural networks and multiple linear regression as potential methods for modeling seed yield of safflower (*Carthamus tinctorius* L.). Ind Crops Prod 2019;127:185–94. https://doi.org/10.1016/j.indcrop.2018.10.050.
[71] Lahouar A, Ben Hadj Slama J. Day-ahead load forecast using random forest and expert input selection. Energy Convers Manage 2015. https://doi.org/10.1016/j.enconman.2015.07.041.
[72] Gunay B, Shen W, Newsham G. Inverse blackbox modeling of the heating and cooling load in office buildings. Energy Build 2017. https://doi.org/10.1016/j.enbuild.2017.02.064.
[73] Kapetanakis DS, Mangina E, Finn DP. Input variable selection for thermal load predictive models of commercial buildings. Energy Build 2017. https://doi.org/10.1016/j.enbuild.2016.12.016.
[74] Ding Y, Zhang Q, Yuan T, Yang F. Effect of input variables on cooling load prediction accuracy of an office building. Appl Therm Eng 2018. https://doi.org/10.1016/j.applthermaleng.2017.09.007.
[75] Hsu D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Appl Energy 2015. https://doi.org/10.1016/j.apenergy.2015.08.126.
[76] May R, Dandy G, Maier H. Review of input variable selection methods for artificial neural networks. Artif neural networks-methodological adv biomed appl. InTech; 2011.
[77] Ciulla G, D'Amico A, Lo Brano V. Evaluation of building heating loads with dimensional analysis: application of the Buckingham π theorem. Energy Build 2017;154:479–90. https://doi.org/10.1016/J.ENBUILD.2017.08.043.