A comparative analysis of regression algorithms and a real world application of multivariable energy signatures

Energy and AI 22 (2025) 100641

Contents lists available at ScienceDirect

Energy and AI

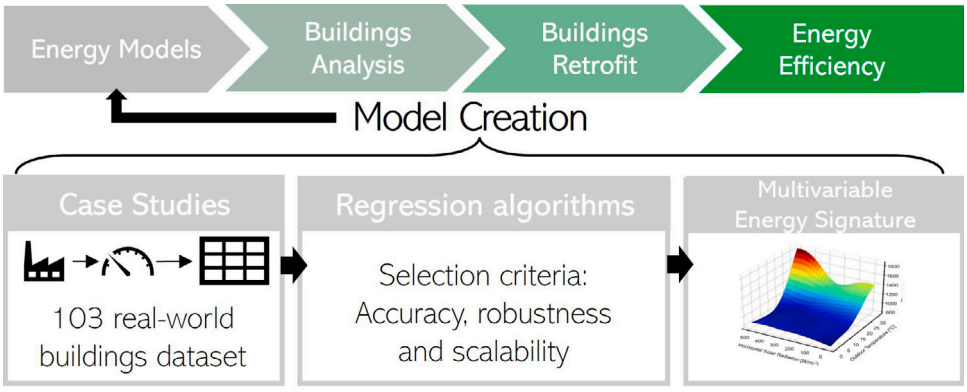
journal homepage: www.elsevier.com/locate/egyai

A comparative analysis of regression algorithms and a real world application of multivariable energy signatures

Simone Eiraudo [a,*], Daniele Salvatore Schiera [a], Luca Barbierato [a], Alena Trifirò [b], Lorenzo Bottaccioli [a], Andrea Lanzini [a]

[a] Energy Center Lab, Politecnico di Torino, Torino, 10138, Italy
[b] TIM S.p.A, Milan, 20123, Italy

GRAPHICAL ABSTRACT



HIGHLIGHTS

- An ecosystem of energy models is needed to improve energy efficiency.
- Multivariable Energy Signatures are employed to model buildings thermal behavior.
- A comparative analysis of regression algorithms and time resolutions is carried out.
- Experiments consider real-world data from 103 industrial buildings.
- Neural Networks are accurate, robust and scalable tools to estimate Multivariable Energy Signatures.

ARTICLE INFO

Keywords:
Buildings
Energy audit
Energy signature
Neural networks
Multivariable regression analysis

ABSTRACT

An ecosystem of energy models of buildings is needed to boost the retrofitting process to improve energy efficiency and meet sustainability goals. Such models should enhance the understanding of the energy behavior of a building, the impact of the external variables, and the causes of inefficiencies. Energy Signatures can fill this role, with particular regard to the consumption due to air conditioning. Univariate models, neglecting the impact of solar radiation, have been widely adopted for Energy Signature analysis. This paper presents Multivariable Energy Signatures considering outdoor temperature and solar radiation. The application on a real-world dataset of multivariable non-parametric approaches stands out from previous works in the ES sector.

* Corresponding author.
E-mail addresses: simone.eiraudo@polito.it (S. Eiraudo), danielesalvatore.schiera@polito.it (D.S. Schiera), luca.barbierato@polito.it (L. Barbierato), alena.trifiro@telecomitalia.it (A. Trifirò), lorenzo.bottaccioli@polito.it (L. Bottaccioli), andrea.lanzini@polito.it (A. Lanzini).

https://doi.org/10.1016/j.egyai.2025.100641
Received 10 April 2024; Received in revised form 31 July 2025; Accepted 27 October 2025
Available online 29 October 2025

This led to a mean improvement of 0.768 to 0.804 of the coefficients of determination calculated over 103 real-world case studies. Moreover, Neural Networks outperformed several literature algorithms regarding accuracy, robustness, and scalability. The paper also discusses issues regarding the time resolution of input data and introduces appropriate visualization tools to employ Multivariable Energy Signatures as diagnostic tools.

1. Introduction

Since the reductions of energy consumption and CO_2 emissions have become significant concerns of the international community, the energy efficiency of buildings gained the attention of both institutions and the research community [1]. This increased interest is due to the tremendous contribution of buildings to the global energy consumption. The United Nations estimated in 2024 that this contribution accounted for 34% of the total global energy consumption. Moreover, buildings are responsible for about 10 gigatons of CO_2 emission, corresponding to 37% of the total amount of human activities-related emissions [2]. To date, the actions undertaken to alleviate the carbon footprint of this sector revealed insufficient, and exceptional efforts need to be undertaken in the next years to align it with the climate policy goals. Indeed, although the improvements achieved in the field of energy efficiency, just a moderate reduction of the footprint of buildings was registered after the consumption peak in 2015 [3]. This is due, among other causes, to the ongoing growth of the building stock and the increasing impact of some specific final uses, such as space cooling. The latter represented about 9% of the final energy uses in 2022 [4]. Notwithstanding this modest contribution, it should be noted that the energy consumption from space cooling has doubled over the last 20 years and continues to grow consistently. Despite the technological advances in cooling systems and the corresponding efficiency improvements, the electrical demand for cooling is expected to double once more by 2040. For these reasons, the retrofit of existing buildings has been indicated in the United Nations report [5] as one of the key actions to be undertaken. Specifically, the report claims retrofit actions should affect 3% of the entire existing building stock per year to meet the sustainability goals.

The first step to start such an extensive retrofit process is to provide robust and easily deployable energy audit tools, creating an "ecosystem" of energy models of buildings [6]. These tools should, among others, provide reference consumption patterns, and enhance the identification of a model describing the energy behavior of a building. Both these tasks may be achieved using regression models. These are data-driven tools that infer some relationship between one or more inputs, or explanatory variables, and an output, or dependent variable. Whenever the former task is intended to be undertaken, such algorithms are employed in a forecast fashion. Specifically, they are used to provide reference patterns or to estimate new values, for instance, future ones. Such algorithms are generally evaluated considering the accuracy of prediction of the output variable, typically energy consumption. Instead, when regression algorithms are employed for model identification, they may provide valuable insights into the typical behavior of buildings, support the interpretation of the consumption patterns and the causes of inefficiencies, and allow characteristic values and parameters to be estimated. Furthermore, in this case, they may be employed for ex-ante evaluation of the benefits associated with the adoption of determined retrofit actions.

Over the years, researchers have developed several regression algorithms, which may be grouped into two main categories: parametric models and non-parametric ones. The first family of algorithms refers to those regression models assuming a predetermined shape for the function that links the input variables to the output one. This is the case, for instance, of multivariable linear regression, where a linear relationship is assumed to hold between the input and output variables. In general, parametric models comprehend any inferential model with a finite number of parameters. Instead, non-parametric models do not assume any particular form for the relationship linking explanatory variables and the output variable. Non-parametric models include, for instance, deep NNs.

As shown and discussed in Section 2, the research efforts in building analysis have tended to two opposite approaches. The first adopts trivial parametric models, such as univariate linear regression [7]. Such an approach is generally adopted in those implementations aimed at model identification due to the ease of interpretation derived from the minimal number of parameters used to describe the inferential relationship. Yet, because of the heavy assumptions made to reduce the model complexity, such an approach has a coarse accuracy. The second tendency in the literature is that of adopting highly complex non-parametric models, such as deep NNs, for forecasting purposes. In this case, a great accuracy of the reference output variable may be achieved. However, such models are not provided with some grounding of the parameters, and interpretative approaches have rarely been adopted for regression tasks [8]. Among the other shortcomings of this approach, these regression models cannot be employed for the fundamental task of the ex-ante evaluation of energy savings.

Moreover, no general agreement exists on the time granularity that should be employed for regression models. Monthly, daily, hourly, and minute-sampling measures have been used for regression purposes in the building analysis sector [7]. High-sampling approaches require less time to acquire sufficient data to train accurate regression models. Yet, their performance may be affected by issues related to the long dynamic phenomena occurring in buildings because of thermal inertia [7], resulting in inadequate models. On the contrary, low sampling rates may require too much time, for instance, years, to acquire enough data to train decent models.

This paper contributes to the research efforts in the field of regression models for building analysis by providing a comprehensive comparative analysis of both parametric and non-parametric models. Indeed, several regression models belonging to the field of Machine Learning (ML) and traditional statistical models have been benchmarked, by testing their performance on an extensive real-world case dataset. This dataset comprehends 103 industrial buildings from the Telecommunication (TLC) sector, a fast-growing energy-intensive branch. Proper hypotheses are assumed to fit the experimental case study, considering different time granularities and multiple input variables. Moreover, the reference value provided by the trained models is visualized and interpreted as a MES to enhance understanding of the energy behavior of buildings and to provide helpful insights into the analysis of the impact of explanatory variables on the consumption of buildings. These MES consider outdoor weather variables as explanatory variables to infer the electrical energy consumption of buildings. This paper's contributions can be resumed as follows:

- Multivariable regression models have been designed to increase the accuracy of the models and to provide insights into the impact of multiple explanatory variables
- A comparative analysis of six regression models, that is, Ordinary Least Squares (OLS), polynomial regression, quantile linear regression, local regression, kernel regression, and NN-based regression, has been carried out
- The regression models were tested considering the different time granularities employed in the literature for Energy Signature (ES) analysis, namely, hourly, daily, and monthly time resolutions
- All the models and time resolutions were tested on an extensive real-world case study, including 2 years of measurements from 103 industrial buildings from the TLC sector

Acronyms

CP	Change Point
DC	Data Center
ES	Energy Signature
FFNN	Feed-Forward Neural Network
GBT	Gradient Boosting Tree
M&V	Measurement and Verification
MAPE	Mean Absolute Percentage Error
MARS	Multivariate Adaptive Regression Spline
MES	Multivariable Energy Signature
ML	Machine Learning
MLP	Multi-Layer Perceptron
NN	Neural Network
OLS	Ordinary Least Squares
PGNN	Physics-Guided Neural Network
PIML	Physics-Informed Machine Learning
PINN	Physics-Informed Neural Network
RMSE	Root Mean Squared Error
TLC	Telecommunication

Symbols and Variables

α	Envelope solar absorptivity, [−]
β	Vector of parameters
ϕ_{Cond}	Heat from the air conditioning system, [kW]
ϕ_{Sol}	Heat gain from solar radiation, [kW]
ϕ_{St}	Internal heat generation of buildings, [kW]
ϕ_T	Heat exchanged through the buildings' envelope, [kW]
χ	Activation function
b	Vector of bias
e	Error term
G	Horizontal solar irradiance, [W/m^2]
h_0	Convection heat transfer coefficient, [$\frac{W}{m^2*K}$]
I	Total solar irradiance, [W/m^2]
P_{Aux}	Electrical load from lights and auxiliaries systems, [kW]
P_{CLC}	Conditioning system electrical demand, [kW]
P_{DISS}	Electrical power conversion losses, [kW]
P_{TLC}	Electrical load of the telecommunication equipment, [kW]
P_{TOT}	Total electrical load, [kW]
t	Time
T_{ext}	Outdoor temperature, [°C]
T_{in}	Indoor temperature, [°C]
$T_{sol-air}$	Solar-air temperature, [°C]
w	Vector of weights
x	Vector of inputs
X	Matrix of inputs
Y	Vector of outputs

- The best models are then discussed and interpreted as MES. The interpretation of the impact of the explanatory variables on different intervals of the domain was achieved by using suitable visualization tools.

The adopted multivariable approach, along with the employment of non-parametric regression algorithms and the experimental application on a real-world dataset, constitute a novelty to the existing research works. The employment of the proposed methodology is intended to enhance ES models accuracy, while focusing the analysis on the most relevant variables affecting cooling consumption in buildings.

The remaining of the present work is structured as follows. Section 2 presents a collection of relevant works that employed regression models, with a particular focus on ES models. Section 3 introduces the formulation of ES models, describes the benchmarked regression algorithms, and discusses the issues regarding time granularity. Section 4 delves into the case study on which the methodology was tested and provides implementation details. Section 5 presents and discusses the results. Finally, Section 6 concludes this work and identifies future directions for further investigation.

2. Literature review on regression models and related applications

Regression models are tools for modeling a relationship to infer one output variable from one or more input variables. In other words, it can be seen as the function describing the response of the output variable to the explanatory variables. The general form of a regression functions is:

$$Y = f(X) + e, \tag{1}$$

where X is the matrix of the input variables, Y is the output vector, $f(X)$ is the regression function and e is an additive error term. This term may comprehend statistical noise and the discrepancy between real output values and the modeled output $f(X)$ due to the impact of un-modeled explanatory variables. The lower the error term e, the better the regression model. Many approaches may hence be adopted to minimize the error term of regression functions, including modeling multiple explanatory variables and employing more effective regression functions f. Regression models have been widely adopted in the building analysis sector for multiple purposes, including forecasting, anomaly detection, and inverse modeling. Among the most important regression tools in this field are ES. These models infer the thermal consumption in buildings based on outdoor weather variables. They may rely on traditional statistical regression approaches or ML-based ones.

Traditional approaches are parametric regression models, where the model can be described with a finite number of fixed parameters. This is, for instance, the case of linear regression, where the output variable is a linear combination of the explanatory variables weighted according to a vector of parameters. Such a regression model can be written as:

$$Y = \beta X + e, \tag{2}$$

where β is the vector of the parameters, with a length equal to $(p + 1)$, where p is the number of explanatory variables employed, and an additional parameter β_0 is included in the vector to represent an intercept term, while a column of ones is added to the input variables matrix X. For example, suppose a single input variable is employed in the regression model. This is the case of an univariate linear regression, where X is a 2xN matrix, being N is the number of elements considered in the problem, and β is a vector of length 2. It is worth noticing that, when working with time series, N represents the number of time steps considered. Many applications of this univariate linear regression model applied to ES analysis exist in the literature regarding both residential [9] and commercial buildings [10]. These approaches are generally implemented in a model parameter identification fashion. For instance, in [11], characteristic parameters of equivalent models of buildings are estimated. A slight modification of the linear regression approach is the Change Point (CP) model, established as reference statistical ES models for buildings thermal characterization [12]. CP models are univariate piece-wise regression models featuring 1 to 6

parameters describing the line offset, gradients, and breakpoints [13]. These models are generally estimated by using OLS. Yet, in [14], Meng et al. adopted quantile regression to fit a CP regression model. All these linear approaches rely on heavy and sometimes fragile assumptions. Besides, they all depend on the sole outdoor temperature as the explanatory variable. A few examples of multivariable linear regression models exist in the literature. Among the others, the authors in [15] employed a multivariable approach. Yet, the results were only validated against simulated data, and the regression algorithms were restricted to linear ones. Other authors took advantage of more complex parametric models. Nagaler et al. [16] applied a sigmoid ES approach on residential and commercial buildings, achieving good agreement between this data-driven model, physics-based simulations, and real data. Afshari et Liu employ multivariable regression approaches in their work [17] to model the energy consumption of Abu Dhabi. The consumption is considered at an aggregated level, hence no analysis is possible regarding single buildings. In this case, the impact of some input variables is modeled as linear, while others are assumed to depict a non-linear impact on the output variable. A multivariable approach is employed in [18] to predict the electrical consumption of the lighting system, considering, among the others, polynomial regression and using horizontal global radiation as an input. In general, parametric models are easy to train and enhance inferential statistics analysis. Yet, they lack accuracy and rely on the a priori assumption of the qualitative form of the regression function. With regard to the complex problem of modeling heating or cooling related consumption, researchers frequently relied on outdoor air temperature as the sole independent variable for statistical models [7].

Non-parametric models overcome such limitations of traditional models. These regression tools do not assume any shape of the distribution of the output variable with respect to inputs and can catch any particular relationship despite its complexity. A typical example of non-parametric models is Deep NNs. This family of algorithms has been widely adopted in the building sector, particularly for forecasting purposes. Yet, it has been rarely applied to estimate ES of buildings. One example of non-parametric ES was provided by Westermann et al. in [19] to investigate the thermal behavior of a set of buildings. On one hand, the authors correctly point out the highly informative contribution of ES to building analysis. On the other, the main drawback of non-parametric models is highlighted; that is, the models have no physical grounding. A comprehensive collection of non-parametric regression models, including NNs, Gradient Boosting algorithm, Support Vector Machine, and others, is presented by Sekeroglu et al. in [20]. Alizimir et al. [21] compared statistical approaches, such as Multivariate Adaptive Regression Spline (MARS), and ML-based algorithms, including Gradient Boosting Tree (GBT), NNs, and Adaptive Neuro-fuzzy systems, to forecast the daily solar radiation by employing temperature, wind speed, and humidity as inputs. GBT achieved superior performance compared to the other algorithms. The MARS algorithm was also capable of providing high forecasting accuracy. Another contribution to this category of algorithms regarding the field of ES is from Rouchier [10], which proposes a Hidden-Markov ES model and benchmarked it against the reference CP model. The non-parametric model proposed by the author enhances achieving higher regression accuracy and is finally employed to estimate the avoided energy consumption, that is, the energy savings, deriving from a retrofit action that has been undertaken. In the end, non-parametric models may enhance relevant accuracy improvements compared to traditional models. Yet, they are more data-intensive, as both the qualitative and quantitative aspects of the regression function have to be modeled and may not be easily employed for inferential statistics. Besides, concerning the task of Measurement and Verification (M&V), they cannot be used for an ex-ante estimate of savings deriving from a given retrofit intervention. On the contrary, in physics-grounded models, different retrofit scenarios can be simulated by varying physically significant model parameters [16], thus enhancing ex-ante energy savings assessment.

As observed, literature efforts have been chiefly devoted to two approaches. The former is characterized by adopting heavy assumptions to reduce the complexity of the problem. This approach tends to neglect the impact of multiple input variables and adopt reference linear regression models. This results in easily interpretable models, which may be employed for parameter estimation, building characterization, and ex-ante energy retrofit savings estimation. Yet, it results in low accuracy and does not provide insights regarding most of the variables involved in the real building thermal problem. The second approach takes advantage of complex agnostic regression algorithms to model both quantitative and qualitative aspects of the regression functions. This results in high accuracy and may take multiple variables as input. Yet, it does not allow many applications except for forecasting. Indeed, without any physical grounding of the model or interpretative tools, few outcomes are possible regarding understanding the thermal behavior of buildings, potential savings assessment, and detection of causes of inefficiencies. Recently, a number of approaches, including Physics-Informed Neural Network (PINN) [22] and Physics-Guided Neural Network (PGNN) [23] have been introduced as part of the Physics-Informed Machine Learning (PIML) [24] research branch. Such algorithms aim at encoding and integrating physical laws into NN, enhancing training more general predictive models by using less data than ML algorithms. Yet, a few concerns of PIML algorithms, including optimization, convergence and computational time issues, must be address to develop a more stable and generalized framework [24]. For these reasons, further efforts are needed to establish an accurate, interpretable, and widely adoptable reference model for ES.

3. Methodology

Data-driven energy modeling approaches have proven to be effective in multiple applications, including energy modeling and operation. These approaches require lower computational and modeling efforts than physical models. Besides, they do not require extensive information about the thermal layout of buildings. Among data-driven approaches, one established approach is that of ES. These are highly informative data-driven regression models, embedding the physical laws derived from the thermal balance of buildings. They can therefore enhance understanding of the thermal behavior of buildings, the impact of the weather variables, and the possible causes of inefficiencies. A robust, proficient, and wide employment of such analysis tools may, in turn, support benchmarking of the building stock and the estimation of energy savings related to specific efficiency measures. In general, a building's thermal behavior may depend on *(i)* the internal temperature, *(ii)* the outdoor temperature, *(iii)* wind speed [25], *(iv)* solar radiation [17], *(v)* the presence of occupants and other contributions to internal heat generation, *(vi)* the buildings characteristic parameters, for instance, the total heat loss coefficient. The widely adopted approach of linear univariate ES operates strong assumptions regarding most of these aspects, reducing the problem to a mere investigation of the impact of the outdoor temperature to estimate the total heat loss coefficient. Yet, in many cases, one or more of the remaining aspects may heavily affect building consumption. To better understand the impact of these variables, it is useful to introduce the thermal balance equation of a generic building:

$$\frac{dT_{in}}{dt} = \phi_T + \phi_{Sol} + \phi_{Cond} + \phi_{St} \tag{3}$$

where $\frac{dT_{in}}{dt}$ is the variation of the indoor temperature over time; ϕ_T is the thermal power exchanged with the outdoor environment by conduction through the building envelope; ϕ_{Sol} is the heat gain determined by solar radiation; ϕ_{Cond} is the thermal power of the air conditioning system; and, finally, ϕ_{St} is the internal heat generation. Assuming an internal constant temperature is set, and the air conditioning system is capable of balancing the remaining contributions to maintain this setup point, we set $\frac{dT_{in}}{dt}$ equal to zero. Hence, the contribution of the

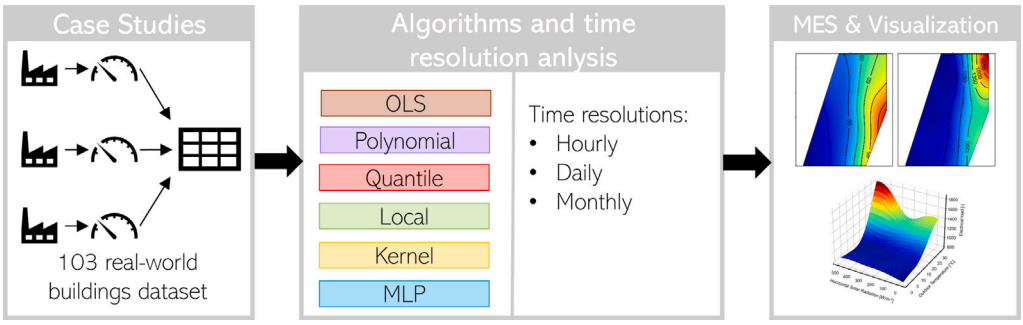


Fig. 1. Outlook of the methodology employed for the comparative analysis of regression algorithms.

conditioning system ϕ_{Cond} will equal the negative of the remaining contributions and assume a positive value if heating is needed to maintain the setup temperature, and vice versa, a negative value if cooling is required. The contribution of heat transmissions over the envelope is often assumed to depend only on outdoor temperature. Yet, considering the additional contribution of the solar radiation affecting the envelope, that is, using the concept of solar-air temperature [26], enhances a much more accurate description of the heat exchange through the envelope. The solar-air temperature is defined as:

$$T_{sol-air} = T_{ext} + \frac{\alpha * I}{h_0} \tag{4}$$

where T_{ext} is the outdoor temperature, α is envelope solar absorptivity, I is the total solar irradiance on the envelope, and h_0 is the convection heat transfer coefficient. This coefficient may, in turn, be affected by the wind speed and orientation. Indeed, the higher the wind speed, the higher the heat exchange within the air and the building envelope surface [25]. The second contribution in the thermal balance equation, ϕ_{Sol}, depends on the solar radiation, the sun's position, and some buildings' characteristic parameters, such as the Solar Heat Gain Coefficient. It may hence be noticed that solar radiation affects 2 out of three of the thermal contributions the conditioning system has to deal with. Finally, the internal heat gains contribution ϕ_{S_t} depends on the buildings' usage, the schedules, the presence of occupants, and other factors. Many of these factors are rarely available on real buildings' datasets. Besides, internal heat generation has a minor contribution to many case studies, such as the residential sector, or may be assumed as constant for others, such as the one we present in Section 4.1 (see Fig. 1).

For the above reasons, this paper proposes analyzing MES, considering outdoor temperature and solar radiation as input variables. This is expected to enhance the study of the thermal behavior of buildings by including the two most relevant variables affecting the thermal balance. Finally, the thermal demand is considered the regression models' output variable.

3.1. Benchmark of Multivariable Energy Signature models

To enhance a fair and comprehensive comparison of traditional and ML-based regression algorithms, six SoA regression architectures are considered. Each is trained on 103 case studies, that is, on 103 buildings, considering different time resolutions. By considering the SoA contributions presented in Section 2, the following algorithms, employed in the papers reported in parenthesis, are considered: *(i)* OLS linear regression, *(ii)* quantile linear regression [14], *(iii)* polynomial regression [18], *(iv)* local polynomial regression [27], *(v)* kernel regression [19], and *(vi)* NN-based regression [20,21]. Besides accuracy, the models should fit the criteria of scalability and interpretability [6]. The accuracy and scalability of the models are evaluated by considering respectively the coefficient of determination R^2, the MAPE and RMSE, and the computational time, calculated over the different time resolutions. Moreover, the data are split into train and test datasets, by means of time-series split, using a proportion of 2/3 and 1/3, respectively. The coefficient of determination is calculated separately for the two subsets of data to discuss the models' robustness and capability to generalize. Finally, the graphical tools introduced in 3.2 enhance the interpretation of the regression models.

3.1.1. Ordinary Least Squares Linear Regression

Multivariable Linear regression [28] foresees that the predicted variables are computed as a linear weighted combination of explanatory variables. In our case study, this can be written as:

$$\phi_{cond} = \beta_0 + \beta_T * T_{ext} + \beta_{Sol} * G + e, \tag{5}$$

where β_0 is the offset of the regression function, β_{Temp} is the temperature coefficient, β_{Sol} is the solar radiation coefficient, and G is the horizontal total radiation. The OLS method is the most commonly adopted approach to fit linear regression. In this case, the regression algorithm aims to minimize the residual sum of squares between the real values and the predicted output values. Notice that several hypotheses should be considered when adopting a linear regression model, including linearity, homoscedasticity, and lack of multicollinearity. In the literature, linear regression models have been frequently adopted without ensuring full compliance with the mentioned hypotheses. Homoscedasticity and linearity will be discussed in Section 5. Regarding multicollinearity, it is worth pointing out that the input variables considered for our multivariable regression task are expected to depict a strong correlation; that is, they are collinear. Notwithstanding the infraction of the hypotheses that should be adopted to use this model, this paper considers linear regression as a reference benchmark algorithm employed to enlighten and discuss some related issues. For the sake of brevity, we will hereafter refer to this approach simply as Linear Regression.

3.1.2. Quantile Linear Regression

Another approach to Linear Regression is employing the method of Quantile Regression [29]. In this case, the algorithm assumes the conditional median as the reference output value instead of the conditional mean, which OLS regression employs. Quantile Linear Regression is more robust to outliers concerning OLS regression. In the present analysis, this algorithm is considered in particular to investigate whether it may provide more robust and generalizable regression models concerning the more widely adopted method of Ordinary Least Squares.

3.1.3. Polynomial Regression

Multivariable Polynomial Regression is a regression approach where explanatory variables comprehend both the original input variables and the higher-degree terms that can be derived from the same input variables. In this sense, it may expand the multivariable regression approach, considering additional input terms. This approach is implemented in this analysis considering terms up to the second degree. In

this case, the regression model can be described as follows:

$$\phi_{cond} = \beta_0 + \beta_{T^2} * T_{ext}^2 + \beta_{Sol^2} * G^2 + \\ + \beta_{T-Sol} * T_{ext} * G + \beta_T * T_{ext} + \beta_{Sol} * G + e \quad (6)$$

where β_{T^2}, β_{Sol^2}, and β_{T-Sol} refer to the regression coefficient of the second-degree terms. This formulation solves the regression problem as a standard linear regression problem.

3.1.4. Local Polynomial Regression

Local Polynomial Regression [30] is a classical method in the sense that it is based on the least squares estimation method. Yet, it provides a much more flexible approach than ordinary linear regression, as it is built on the intuition of fitting simple polynomial functions to localized data subsets [27]. This results in a regression function that can, in principle, fit any shape. For this reason, Local Polynomial Regression is considered a non-parametric method. This approach provides several advantages. In particular, it does not require a prior definition of the function qualitative form and is more flexible for standard least squares methods. Yet, it is computationally expensive, and a non-parametric model cannot be described using a finite parameters function. This paper implements this method by considering a second-order degree polynomial to fit data locally. This complies with the method's basic intuition, which is considering low-order polynomial for sub-domain estimation to avoid data over-fit [30]. For the sake of brevity, we will hereafter refer to this algorithm as Local Regression.

3.1.5. Kernel Regression

In the same way that Local Regression, Kernel Regression [31] is a statistical approach to regression problems. Besides, these two methods share a local adaption to data to provide a non-parametric regression model. Yet, while Local Regression fits a defined polynomial regression function to a limited local sub-domain, Kernel regression employs kernel functions, typically the Gaussian one, to account for the dataset elements' proximity to a specific point. Kernel regression may result in even higher computational complexity than Local Regression. Yet, its extreme flexibility enhances handling any complex relationship between the output and the explanatory variables.

3.1.6. Neural Network-based regression

NNs are a family of ML models characterized by a biology-inspired form, including several different architectures, which may be divided into two main sub-categories, namely Feed-Forward Neural Network (FFNN) and Recurrent Neural Networks. One of the simplest NN architecture is the Multi-Layer Perceptron (MLP), firstly described in [32]. This is a fully connected FFNN, including an input layer, an output one, and one or more hidden ones. Each layer comprehends one or more perceptrons. These are simple elements performing the following operation:

$$a = f(x) = \chi(\langle w, x \rangle + b) \quad (7)$$

where x is the input vector, w is the weighting trainable vector, b is a trainable bias vector, and χ is the activation function. By feeding the input layer with the explanatory variables and each successive layer with the output of the previous one, the network finally outputs the dependent variable. Assuming that the network features multiple hidden layers, comprehending several perceptrons for each one, it may be considered a non-parametric regression model. To the aims of our analysis, we employed a 2-hidden layer NN. The training procedure was realized by setting the batch size equal to one month of measurements. The remaining NN hyperparameters were determined by an automated procedure as described in Section 4.2.

3.1.7. Time resolution analysis

Different alternatives regarding the time resolution to be employed for ES have been proposed by researchers. For instance, the hourly time resolution is used in [10,16,19]. On the contrary, daily mean data were preferred in [25]. Finally, a few authors considered using monthly mean data [7]. The time resolution impact on building energy models is twofold. First, thermal load and weather time series generally show high autocorrelation values, which are linked to the building's thermal mass effects. The second aspect is time series periodicity, typically determined, regarding buildings, by operation schedules. According to Fu [7], using lower resolution data, for instance, daily, is an effective solution to this. Similar concerns are reported by other authors, such as Rasmussen in [25], which considers that the daily time resolution allows the analyst to neglect the impact of thermal inertia. Nevertheless, the decrease in time resolution would come with a cost. In real-world scenarios, too much time would be needed to collect enough data to train most regression models. Besides, energy models designed on coarse data would not be helpful for some tasks, such as Operation and Maintenance or anomaly detection. Considering the alternatives proposed in the literature regarding the time resolution to employ for ES purposes, this analysis comprehends hourly, daily, and monthly time granularities. The three time series resolutions are employed to train all six regression models separately and tested on the 103 real-world datasets. Finally, they are evaluated and commented on considering accuracy, scalability, and generalization capability.

3.2. Employment of Energy Signatures as inspection tools

Employing proper visualization tools may finally enhance the diagnostic task and the qualitative understanding of the thermal behavior of the considered buildings. To this aim, three visualization tools for MES are considered. The first option is to provide the most relevant input variable on the x-axis and the output variable on the y-axis. A reference line will describe the regression model in a univariate fashion, while the contribution of one or more additional variables determines the dispersion of points around this line. This visualization tool may eventually be employed for models considering any number of input variables, beholding a quick and trivial visualization of the impact of one fundamental explanatory variable [25]. On the contrary, the effect from only two input variables can be described for 3-D style MES and heat maps. Yet, in this case, a much more accurate description of the contribution of the second explanatory variable to the final output is achieved.

4. Case study and experimental set up

In addition to the methodological steps described in the previous sections, which represent the general framework and are intended to be applied to any building analysis, a few additional instances regarding the specific case study of application are presented here. In addition, for the sake of transparency and replicability of the results, we briefly comment on the experimental setup and the software details in Section 4.2.

4.1. Case study

The proposed methodology has been applied to a dataset of 103 buildings. The dataset contains two years of aggregated buildings' hourly electrical load values. These values hence include any contribution to electrical consumption. No information is available regarding single electrical contributions or final uses. Besides, the dataset includes the hourly measures for outdoor temperature and Horizontal Solar Radiation, expressed °C and W/m^2 respectively.

These measures were collected from 103 Data Centers (DCs) of an important TLC service provider in Italy. The TLC networks and their management buildings have massively increased their contribution to

final energy demand, reaching share of 2.8–3.8% of total electricity use in Europe, corresponding to a 70–95 TWh of electricity in 2022 [33]. DC are industrial buildings characterized by high internal loads, which determine high internal heat density generation. The buildings are located in different regions of Italy, covering a wide spectrum of climatic conditions. Indeed, the buildings stock investigated covers all the 7 climatic classes identified by the *Climatic Severity Index*. Most of the analyzed buildings are characterized by a sporadic presence of occupants, resulting in an irrelevant contribution to the final electrical demand by people, and by continuous operation. In such cases, the operative schedules are expected not to play an important role in the energy profiles of the buildings, mitigating the negative effect of periodicities while employing hourly resolution time series [34]. Yet, the dataset also includes DC buildings hosting some office areas. In this case, the time series will depict stronger periodicities, and using low-resolution data is expected to result in more reliable outcomes. In both cases, the two most important contributions to the buildings' final energy demand are the TLC equipment and the cooling systems employed to avoid overheating the equipment. The energy balance [34] of DCs is described as:

$$P_{TOT} = P_{TLC} + P_{DISS} + P_{CLC} + P_{Aux} \tag{8}$$

where P_{TOT} is the aggregated electrical load, P_{TLC} is the contribution from the TLC equipment, P_{CLC} is the conditioning system demand, P_{Aux} represents the load from the lighting system and the auxiliaries and P_{DISS} accounts for the energy conversion losses.

In order to provide an energy model in the form of a ES, the explained variable, in this case, the electrical consumption, shall be expressed as a function of the weather variables, T_{ext} and G. By considering that the contributions P_{TLC}, P_{Aux}, and P_{DISS} are not affected by the weather variables, it is derived that:

$$P_{TOT}(T, G) = P_{TLC} + P_{DISS} + P_{CLC}(T, G) + P_{Aux} \tag{9}$$

It is worth noticing that the conditioning system load P_{CLC} is the only load quota depending on the weather variables. P_{TLC} and P_{DISS} may generally assumed as constant over time for DC, hence will not affect the regression models, except for the determination of a constant offset of the ES models. Finally, P_{Aux}, which comprehends the auxiliary systems as well as lighting and other minor load quotas, is variable in time and may, in particular, depend on building schedules but may be considered independent of the weather variables. This may easily result in a non-causal correlation if hourly resolution is considered. For instance, in the case P_{Aux} features high values during the day, this load contribution would positively correlate with temperature and radiation, which generally depict higher values during the day. These considerations confirm what is anticipated in Section 3.1.7, that is, on the one hand, that the aggregated consumption values P_{TOT} may be safely employed for ES purposes whether daily or monthly resolution data is used. On the other hand, hourly values should be employed carefully, considering the issues related to non-causal correlation and thermal inertia.

4.2. Experimental set up and software details

All the algorithms were designed and trained in the Python environment and conducted using a Python 3.9 interpreter. All the experiments were carried out with a laptop with an 11th Gen Intel Core i7-1165G7 processor and 16 GB RAM.

The linear and polynomial regression models were designed using the Scikit-Learn library [35]. The *QuantileRegressor* class from the same library was considered to perform quantile regression. In our implementation, *highs* was employed as a solver. Local Regression was performed using the *localreg* library [36]. The Epanechnikov kernel function is employed and considers a domain fraction of 0.66. A grid-search procedure was undertaken to set the value of the hyperparameters, namely the regression degree, the kernel function and its width. Kernel Regression was implemented in the *statsmodels* library [37]. The hyperparameters of the algorithm were optimized by means of the cross-validation least squares method implemented by *statsmodels*. The NN was designed using the Scikit-Learn library. The *MLPRegressor* class and the library's model selection tools were employed. This allows an automatic grid-search-based selection of the NN hyperparameters. In this analysis, the hyperparameters considered in the procedure were the number of neurons, the activation functions, the learning rate α, and whether it was a constant or adaptive rate.

5. Experimental results and discussion

The methodology presented in Section 3 was tested considering the case study presented in Section 4. As mentioned, the electrical load data from 103 buildings were considered along with the historical weather variables series. The raw data first underwent a coarse preprocessing step to clear the dataset from the possible presence of measure errors. First, the data, proceeding from different meters and provided with varying resolutions of time, were re-sampled considering hourly mean values. A unique dataset, indexed by the data and time of the measures, was obtained. Then, a domain was considered for both input and output variables. In particular, the solar radiation values were imposed to be non-negative and lower than 1376 W/m^2, and the electrical load was set to be positive, thus excluding NaN and zero values. The measures presenting one or more variables outside the domain were excluded from the dataset. Hence, the electrical load data were altered and anonymized to guarantee privacy and data protection to the provider. The electrical load data are hereafter handled as anonymous ones. For this reason, they rely on an arbitrary unit measure, and the electrical load labels on the plots do not report any unit of measure.

A second re-sampling step was undertaken to produce the daily and monthly resolution datasets. Finally, the datasets from each single case study, that is, from each single building, were divided into explanatory and dependent variables and train and test subsets. In particular, weather variables were included in the input variables dataset, while the electrical load was set as the output variable. The test dataset was built on a randomly selected 33% portion of the measures from each case study.

5.1. Benchmark of regression models

Each of the six regression models presented in Section 3.1 was fed with the data provided with each of the three-time resolutions and trained for every one of the 103 case studies, resulting in a total of 1854 trained models. An overview of the results from the different regression models is presented in Table 1, where mean values of the coefficient of determination R^2 from the 103 case studies are reported. The first column reports the reference univariate regression model results, implemented according to the methodology detailed in [38]. The bold numbers indicate the highest score for each single regression model for train and test subsets. It may be easily seen that the use of the data with monthly resolution resulted in the highest model train fit for all the considered regression models. Yet, it is evident that the models are frail and have a low generalization capability. Indeed, the R^2 score values reported for the test set were consistently lower for all the regression models considered. The low generalization ability of the models is even more evident for more complex models, in particular the Kernel, Local Quadratic, and NN-based regression models, while the benchmark univariate model stood up to be the most reliable model to employ monthly resolution data thanks to a mean R^2 of 0.74. The poor generalization capability achieved by using this time resolution may depend, to some extent, by the reduced number of points (24 in this case, since the dataset contained 2 years of measurements). This limitation would remain for many practical cases, as the acquisition of a sufficient amount of measurement points would require too much time.

Table 1
Mean coefficient of determination retrieved for the 103 considered case studies. Bold numbers indicate the best result over different time granularities, the cell background indicates the best algorithm performing on the train set, while a green background highlights the best algorithm regarding the test set.

	Univariate ES		Linear		Polynomial		Quantile		Kernel		Basic MLP		Local Quadratic	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Monthly	**0.832**	0.740	**0.819**	0.677	**0.887**	0.687	**0.791**	0.646	**0.887**	0.637	**0.953**	0.612	**0.941**	0.290
Daily	0.766	**0.768**	0.726	**0.718**	0.783	**0.780**	0.716	**0.709**	0.831	**0.803**	0.825	**0.804**	0.794	**0.787**
Hourly	0.660	0.660	0.647	0.648	0.692	0.694	0.639	0.641	0.726	0.722	0.720	0.721	0.702	0.703

Table 2
MAPE retrieved for the 103 considered case studies. Bold text and cell background colors are employed in the same way as in Table 1.

	Univariate ES		Linear		Polynomial		Quantile		Kernel		Basic MLP		Local Quadratic	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Monthly	**3.88**	**4.23**	**4.22**	**5.11**	**3.03**	**4.58**	**3.95**	**5.29**	**2.94**	4.58	**1.61**	5.13	**2.06**	6.53
Daily	5.47	5.69	5.79	5.69	4.92	4.72	5.70	5.61	4.26	**4.33**	4.36	**4.34**	4.81	**4.63**
Hourly	6.64	6.72	6.99	7.04	6.23	6.30	6.88	6.94	5.78	5.89	5.85	5.91	6.10	6.17

Table 3
RMSE retrieved for the 103 considered case studies. Bold text and cell background colors are employed in the same way as in Table 1.

	Univariate ES		Linear		Polynomial		Quantile		Kernel		Basic MLP		Local Quadratic	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Monthly	**5.88**	**7.09**	**6.36**	**7.88**	**4.92**	**7.43**	**6.78**	**8.39**	**5.05**	7.43	**3.23**	8.35	**3.51**	9.95
Daily	7.87	7.86	8.72	8.73	7.61	7.68	8.92	8.89	6.79	**7.19**	6.87	**7.14**	7.45	**7.55**
Hourly	11.45	10.97	11.19	10.53	10.43	9.73	11.33	10.67	9.91	9.29	10.02	9.32	10.32	9.62

Table 4
Confidence intervals estimated for the different performance metrics (R^2, MAPE, and RMSE), and regression algorithms, reported for the results from daily data resolution.

	Univariate ES	Linear	Polynomial	Quantile	Kernel	Basic MLP	Local quadratic
R2	0.748–0.786	0.700–0.743	0.759–0.803	0.680–0.732	0.784–0.823	0.782–0.820	0.770–0.804
MAPE	4.59–7.44	5.22–6.20	4.38–5.19	5.17–6.11	4.02–4.68	4.04–4.69	4.31–4.99
RMSE	5.88–15.36	6.31–12.67	5.33–10.93	6.36–12.74	5.11–9.84	4.97–9.93	5.28–9.78

The use of hourly resolution was found to lead to low regression accuracy, as it can be seen by considering the results reported for the R^2, MAPE and RMSE metric in Tables 1, 2, and 3 respectively. Contrary to what resulted from the monthly resolution, in this case, the models featured a good generalization ability, with R^2 score values aligned with the model's performance on the train set. The more complex regression models achieved the best performance. Specifically, the Kernel regression algorithms slightly outperformed the NN-based regression model, featuring a coefficient of determination for the test set of 0.722 against 0.721, and a MAPE of 5.89% against 5.92%. Yet, the overall performance was often below the one reported using the monthly resolution datasets and always below the one from daily data. Indeed, daily data sampling resulted in the best performance for all the regression models, in terms of R^2 score. In the case of the kernel, local and NN-based regression algorithms, such time resolution also achieved the best results in terms of MAPE and RMSE. Besides, the values reported from the train and test sets were aligned. All the models, except the two linear models (OLS and Quantile), could outperform the benchmark univariate model. The NN-based regression model slightly overcomes the performance from the Kernel regression, resulting in a 0.804 mean R^2 score. A more extensive representation of the performance of the models is reported in Fig. 2. It is worth remembering that the Kernel and NN-based regression models achieved coefficient of determinations over 0.6 for 100 and 101 over 103 buildings, respectively. Besides, their R^2 values overcame 0.8 for 58 and 59 buildings, respectively. The superior performance of the MLP and Kernel-based methods with respect to the others can be demonstrated by considering the 95% confidence intervals of the performance metrics calculated by means of the Bootstrap method, as reported in Table 4 for daily time resolution.

5.2. Discussion on models performance and computational time

The results retrieved for the different models and time resolutions are here discussed, considering the gaps existing in the literature, as reported in [7], describing the advantages and limitations of the traditional [28,29] and ML-based algorithms [30–32] presented in Section 3.1, and determining the most adequate time sampling to be employed for ES. Then, the models are interpreted to better draw insight into the energy behavior of the investigated buildings.

First, normality, independence, and homoscedasticity of residuals tests were undertaken to assess the validity of the regression assumptions. While independence of residuals was proved for any regression algorithm and the vast majority of case studies, only the local, kernel, and NN-based regression algorithms provided also normal and homoscedastic residuals for most of the analyzed buildings.

The reasons for the superior performance of NN and Kernel regression may be better understood by considering the fitted regression models shown in Fig. 3. These plots, derived from a case study and considering daily resolution data, show the most relevant explanatory variable on the x-axis and feature the output on the y-axis. The inference drawn from the additional explanatory variable, in this case, the solar radiation, is responsible for the displacement of predictions (i.e. colored points) from the reference univariate regression line. Two clues are particularly evident in this Figure. First, the linear regression models (*Lin* and *Qua* in the figure) poorly fit the data distribution, particularly regarding the regions close to the domain boundaries at low and high temperatures. This would confirm the hypothesis of linearity of the models only for a limited part of the data, while relevant deviations from this behavior happen in the remaining parts. A similar

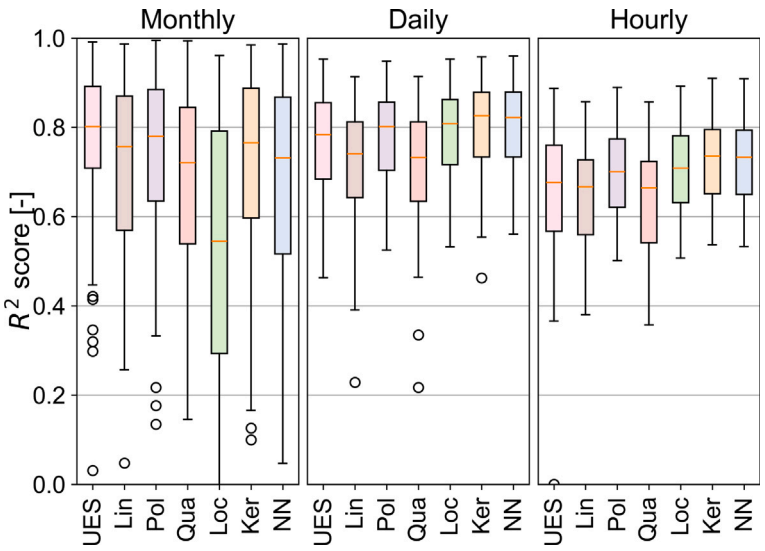


Fig. 2. Statistical representation of the coefficient of determination R^2 retrieved by the considered regression algorithms over the 103 real-world case studies.

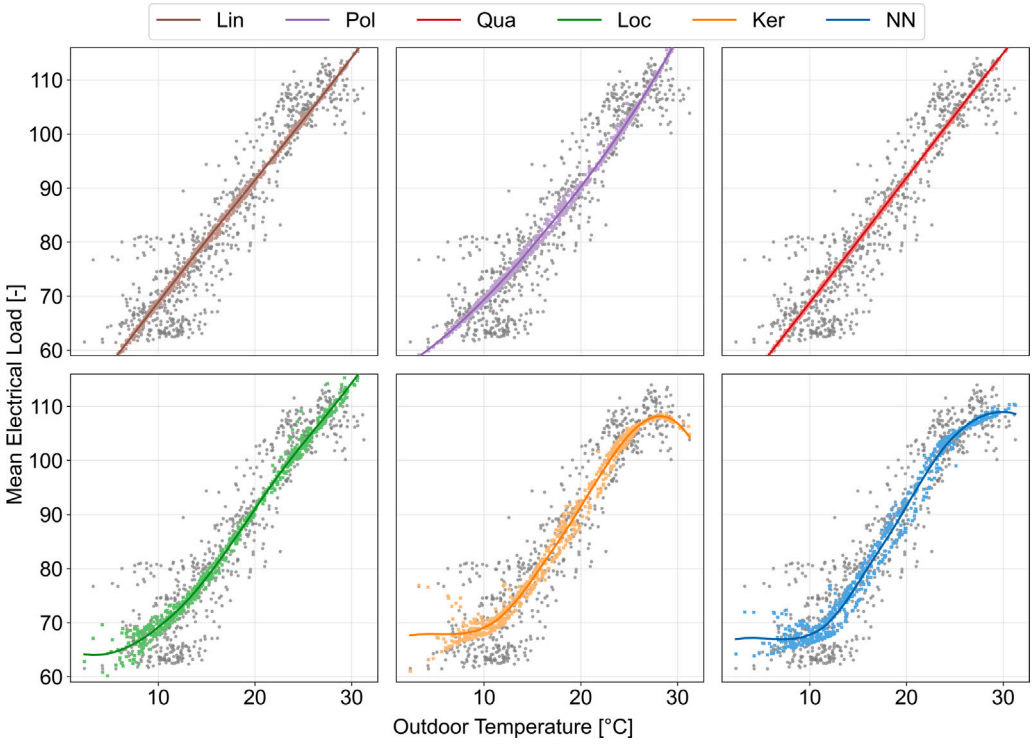


Fig. 3. Reference regression lines and predictions from the different regression algorithms for a case study (building *G*). The predictions are displayed by means of colored points, while the reference lines represent the mean expected output, at a given outdoor temperature. The gray points represent real values.

limitation affects the polynomial regression model (*Pol* in the figure) and, to a minor extent, the local regression algorithm (*Loc* in the figure). On the contrary, the Kernel and the NN-based models (*NN* and *Ker* in figure) properly fit the data, featuring varying slopes on low, intermediate, and high temperatures regions. The first region is characterized by a constant electrical load, whose value is hence independent of the outdoor temperature. This behavior is typical of an unconditioned region. The load increases in the intermediate region, showing a quasi-linear relationship with the outdoor temperature. These two behaviors comply with the typical thermal behavior of buildings, which is assumed to design the benchmark univariate regression model and described more in-depth in [38].

Finally, a deviation from the linear relationship may be observed at high temperatures. The second clue from this figure concerns the dispersion of points around the reference regression line, which, as mentioned, depends on the inference drawn by the model from the additional explanatory variable. The OLS, the quantile, and the polynomial models feature the lowest dispersion of predictions. It is saying that these models could not capture and model solar radiation's impact on electrical load. On the contrary, the other models all show a relevant dispersion of points. It is worth noticing that the distance of the points from the line is not homogeneous over the domain. This implies that the impact of solar radiation is modeled differently in different temperature ranges, as will be later discussed.

Table 5
Mean computational times, reported in seconds, needed to train a regression algorithm for one case study from the different MES models, according to the different time resolutions.

	Lin	Pol	Qua	Loc	Ker	NN
Monthly	$8.9 \cdot 10^{-4}$	$2.4 \cdot 10^{-3}$	$2.7 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$	$4.2 \cdot 10^{-1}$	$1.4 \cdot 10^{0}$
Daily	$1.1 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$	$1.1 \cdot 10^{-1}$	$8.3 \cdot 10^{0}$	$1.4 \cdot 10^{0}$
Hourly	$1.6 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$	$2.5 \cdot 10^{0}$	$2.3 \cdot 10^{1}$	$1.1 \cdot 10^{3}$	$6.8 \cdot 10^{0}$

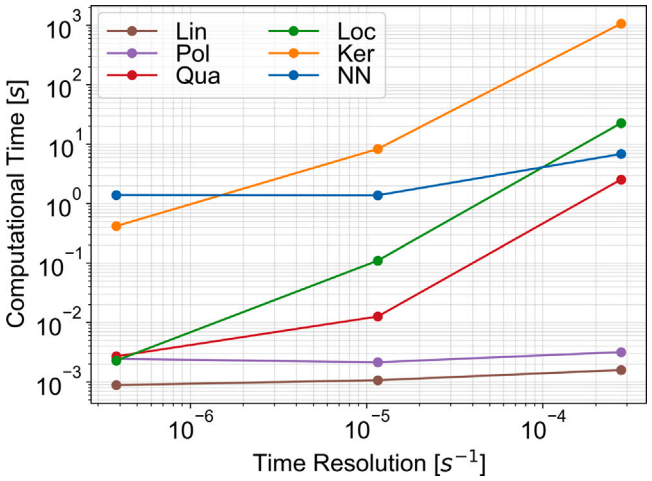


Fig. 4. Logarithmic representation of the mean computational time needed to train the different regression algorithms for one case study.

Then, the scalability of the models is discussed, considering the computational times reported for the models' training. The mean computational time needed to train each regression model for one single case study is reported in Table 5. Besides, these results are represented in the logarithmic plot in Fig. 4 to enhance a quicker and more effective analysis. In this graph, it may easily be observed that parametrical models (the OLS, the quantile, and the polynomial) are computationally much less expensive than non-parametrical ones. Yet, the quantile regressions suffer a relevant increase in computational time for training as the considered datasets include more data points. The Local Quadratic model features a similar behavior. The NN regression model is the most time-expensive as low-sampling data are used, with a mean computational time over 1 s for each case study. Yet, the model features good scalability, with computational time increasing by just about 4 times as time resolution is augmented by 720 times, moving from monthly to hourly data. On the contrary, the kernel regression model was revealed as not scalable, as the complexity of training exponentially increases over different time granularities, leading to mean training times of about 8.3 and 1100 s per case study for the daily and hourly data sampling, respectively. Hence, considering the accuracy of predictions, the generalization capabilities, and the good scalability, the NN-based model considering hourly measures is selected as the most proper regression tool. This model could enhance the ES accuracy to a MAPE of 5.2%, concerning the 5.7% reported by the benchmark model.

5.3. Interpretation of the Multivariate Energy Signatures

Finally, MES may be employed as diagnostic tools to interpret the thermal behavior of buildings. To better understand the impact of both the considered explanatory variables, the fitted regression models should represent both the input and output variables explicitly, for instance, by employing a 3-D surface plot (Fig. 5) or simply a heatmap (Fig. 6). The former represents the non-parametric regression model trained for one of the considered case studies. The MES represented in

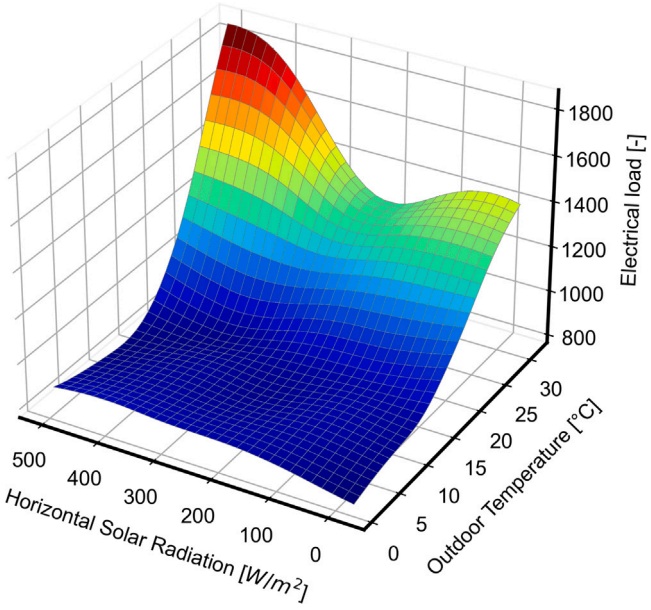


Fig. 5. 3-d style MES from building *B*.

Fig. 5 highlights the typical thermal behavior mentioned in the previous section. In particular, it features a typical constant and low consumption region corresponding to low outdoor temperatures, representing the un-conditioned region. In this area, activating the cooling machines is not necessary to cool down the DC equipment and guarantee the respect of the prescribed temperature set points. In this region, neither the outdoor temperature nor the solar radiation typically affect electrical load. Second, a conditioned intermediate region is modeled. A quasi-linear increase in consumption characterizes this area as the outdoor temperature grows. Generally, solar radiation may only produce a minimal effect on the electrical load. Finally, a high-temperature region is detected. In the 103 buildings analyzed in this study, it was seen that solar radiation produced the most relevant contribution to the electrical load in this region. Yet, such a contribution's shape and magnitude may vary consistently from building to building.

In the case study in Fig. 5, a dramatic boost of the electrical load may be noticed in the high temperatures regions as the solar radiation reaches its highest values. On the contrary, a local minimum is detected at moderate solar radiation values, while the regression model seems to foresee a local maximum for high temperatures and minimum solar radiation values. The complex contribution of solar radiation on electrical load is due, among the other factors, to the impact of this variable on multiple load quotas, which include, in particular, a predominant one, m regarding the cooling system and a minor one regarding lighting, as discussed in [39]. Yet, considering the input data, it might be found that the considered region falls outside the domain. Indeed, as commented before, the final MES models are fed with daily mean values. For the location considered for building B, temperatures over 30 °C may only happen if mean solar radiation over about 200 W/m^2 is registered. The final MES should consider each case study data domain to fairly interpret the analysis outcomes. This is done in Fig. 6 by considering the domain from the train set, with an additional tolerance which is set for each explanatory variable, and cutting off the MES regions falling outside this extended domain. This figure provides a comparative overview of six buildings. The variegate impact of solar radiation may be observed in the different case studies. For instance, according to the MES derived in the cases of buildings *C* and *F*, a variation of the value of solar radiation would not affect the electrical demand. Yet, it results in an important additional contribution in the high-temperature region for buildings *B* (commented before) and *E*.

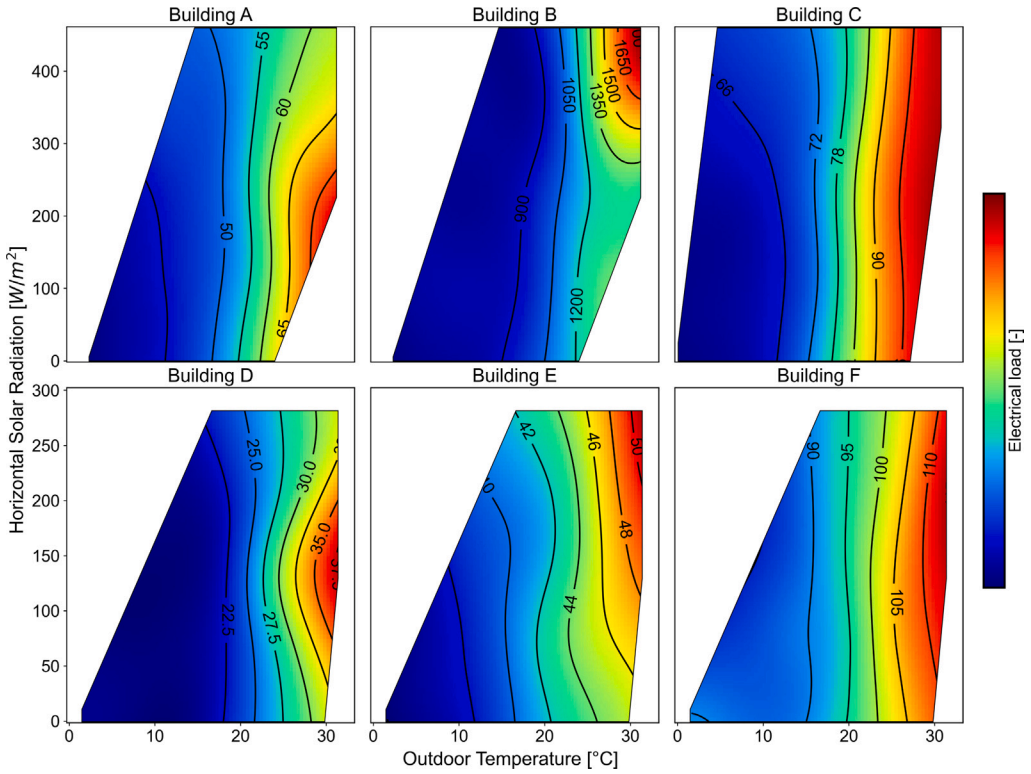


Fig. 6. MES from six buildings, retrieved by the NN-based regression model considering daily input data. The MES are limited to an extended domain calculated considering the weather data from each building's location.

Two more typical patterns exist regarding buildings *A* and *D*. Solar radiation determines a drop in electrical load consumption in the high temperatures region. This anomalous behavior might depend on the presence of some absorption chilling machine or other characteristic feature of this building. The possible reasons to explain the thermal behavior of building *D* are even more challenging to find, as the global maximum electrical load is not detected at the maximum temperature but around intermediate solar radiation values. Further investigation into these buildings is advisable to understand their typical thermal behavior and eventually consider some anomalies or causes of inefficiency. The remaining case studies mostly depicted the typical behavior described for buildings *B, C, E* and *F*.

6. Conclusion and future works

Six literature regression algorithms, comprehending both traditional and Machine Learning-based ones, have been considered to analyze the thermal behavior of buildings using a Multivariable Energy Signature approach. The regression was aimed at inferring between the electrical load and weather variables. Unlike what is usually done in reference univariate Energy Signature models, this analysis included solar radiation as an additional explanatory variable. Moreover, 3 time resolutions were considered, namely hourly, daily, and monthly. This comparative analysis was tested on an extensive real-world dataset including 2 years data series from 103 industrial buildings from the TLC sector. The most important outcomes of this analysis may be summarized as follows:

- Mean daily data resulted to be the most appropriate time resolution for all the considered regression algorithms, leading to a consistently higher accuracy than when monthly or hourly data were employed.
- The Kernel regression and Neural Networks outperformed the Ordinary Least Squares, quantile, polynomial, and local quadratic regression algorithms in terms of accuracy.
- Neural Networks showed good scalability for the increased dataset sizes. The Kernel Regression instead suffered from an exponential boost of computational time as the data points augmented. Hence, Neural Networks was finally selected as the most proper algorithm for Multivariable Energy Signatures.
- The inclusion of solar radiation as an additional explanatory variable led to an important increase in the performance of the models, and a mean MAPE equal to 5.2% was achieved, compared to the 5.7% of the reference univariate Energy Signature regression algorithm.
- Finally, the Neural Network-based Multivariable Energy Signature achieved good prediction performances, with a mean coefficient of determination equal to 0.804 for the 103 buildings considered in the analysis.

Multivariable Energy Signatures were finally employed as diagnostic tools for the thermal behavior of buildings. Their us for energy analysis may be of special interest for all those sectors where air conditioning represents a major contribution to the final consumption of buildings.

Future works will extend the analysis to other case studies from different sectors, such as the residential one, to confirm the outcomes and guarantee a wider applicability. Moreover, future implementations should involve a more detailed design of the relationship between the thermal load and weather variables, as described by the solar-air temperature concept. Eventually, this will lead to analyzing wind as an additional independent variable that could affect the thermal balance of buildings, and hence their energy consumption. In addition, an even more extensive analysis regarding weather variables shall be undertaken, such as a feature importance analysis.

CRediT authorship contribution statement

Simone Eiraudo: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daniele Salvatore Schiera:** Writing – review & editing,

Supervision, Software, Data curation, Conceptualization. **Luca Barbierato:** Writing – review & editing, Supervision. **Alena Trifirò:** Supervision, Resources. **Lorenzo Bottaccioli:** Writing – review & editing, Supervision. **Andrea Lanzini:** Writing – review & editing, Supervision, Project administration, Formal analysis, Conceptualization.

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.

Acknowledgments

Simone Eiraudo acknowledges support from TIM S.p.A. through the Ph.D. scholarship.

Data availability

The data that has been used is confidential.

References

[1] Energy efficiency directive. 2024, URL https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-directive_en. (Accessed 29 March 2024).
[2] United Nations Environment Programme. Global status report for buildings and construction-beyond foundations: Mainstreaming sustainable solutions to cut emissions from the buildings sector. 2024, http://dx.doi.org/10.59117/20.500.11822/45095.
[3] IEA. World Energy Outlook 2024. Int Energy Agency 2024. doi:https://www.iea.org/reports/world-energy-outlook-2024.
[4] International Energy Agency. Net zero roadmap: A global pathway to keep the 1.5 °C goal in reach. Paris, France: IEA; 2023, URL https://www.iea.org/reports/net-zero-roadmap-a-global-pathway-to-keep-the-15-0c-goal-in-reach.
[5] United Nations Environment Programme. Global status report for buildings and construction 2021. U. N Environ Program 2021. URL https://globalabc.org/resources/publications/2021-global-status-report-buildings-and-construction.
[6] Manfren M, Sibilla M, Tronchin L. Energy modelling and analytics in the built environment—A review of their role for energy transitions in the construction sector. Energies 2021;14(3):679.
[7] Fu H, Baltazar J-C, Claridge DE. Review of developments in whole-building statistical energy consumption models for commercial buildings. Renew Sustain Energy Rev 2021;147:111248.
[8] Letzgus S, Wagner P, Lederer J, Samek W, Müller K, Montavon G. Toward explainable artificial intelligence for regression models. 2022.
[9] Rose J, Kragh J, Nielsen KF. Passive house renovation of a block of flats–Measured performance and energy signature analysis. Energy Build 2022;256:111679.
[10] Rouchier S. Bayesian workflow and hidden Markov energy-signature model for measurement and verification. Energies 2022;15(10):3534.
[11] Baasch G, Wicikowski A, Faure G, Evins R. Comparing gray box methods to derive building properties from smart thermostat data. In: Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation. 2019, p. 223–32.
[12] Efficiency Valutation Organization (EVO). Ipmvp's snapshot on advanced measurement & verification. 2020.
[13] Kissock JK, Haberl JS, Claridge DE. Inverse modeling toolkit: Numerical algorithms. ASHRAE Trans 2003;109:425.
[14] Meng Q, Xiong C, Mourshed M, Wu M, Ren X, Wang W, Li Y, Song H. Change-point multivariable quantile regression to explore effect of weather variables on building energy consumption and estimate base temperature range. Sustain Cities Soc 2020;53:101900.
[15] Tronchin L, Manfren M, Nastasi B. Energy analytics for supporting built environment decarbonisation. Energy Procedia 2019;157:1486–93.
[16] Nageler P, Koch A, Mauthner F, Leusbrock I, Mach T, Hochenauer C, Heimrath R. Comparison of dynamic urban building energy models (UBEM): Sigmoid energy signature and physical modelling approach. Energy Build 2018;179:333–43.
[17] Afshari A, Friedrich LA. Inverse modeling of the urban energy system using hourly electricity demand and weather measurements, part 1: Black-box model. Energy Build 2017;157:126–38.
[18] Belany P, Hrabovsky P, Sedivy S, Cajova Kantova N, Florkova Z. A comparative analysis of polynomial regression and artificial neural networks for prediction of lighting consumption. Buildings 2024;14(6):1712.
[19] Westermann P, Deb C, Schlueter A, Evins R. Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. Appl Energy 2020;264:114715.
[20] Sekeroglu B, Ever YK, Dimililer K, Al-Turjman F. Comparative evaluation and comprehensive analysis of machine learning models for regression problems. Data Intell 2022;4(3):620–52.
[21] Alizamir M, Kim S, Kisi O, Zounemat-Kermani M. A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions. Energy 2020;197:117239.
[22] Muther T, Dahaghi AK, Syed FI, Van Pham V. Physical laws meet machine intelligence: current developments and future directions. Artif Intell Rev 2023;56(7):6947–7013.
[23] Faroughi SA, Pawar NM, Fernandes C, Raissi M, Das S, Kalantari NK, Kourosh Mahjour S. Physics-guided, physics-informed, and physics-encoded neural networks and operators in scientific computing: Fluid and solid mechanics. J Comput Inf Sci Eng 2024;24(4):040802.
[24] Cuomo S, Di Cola VS, Giampaolo F, Rozza G, Raissi M, Piccialli F. Scientific machine learning through physics–informed neural networks: Where we are and what's next. J Sci Comput 2022;92(3):88.
[25] Rasmussen C, Bacher P, Calì D, Nielsen HA, Madsen H. Method for scalable and automatised thermal building performance documentation and screening. Energies 2020;13(15):3866.
[26] Forouzandeh A. Comparative analysis of sol-air temperature in typical open and semi-closed courtyard spaces. In: Building simulation. Springer; 2022, p. 1–17.
[27] Garimella RV. A simple introduction to moving least squares and local regression estimation. Tech. rep., Los Alamos, NM (United States): Los Alamos National Lab. (LANL); 2017.
[28] Yan X, Su X. Linear regression analysis: theory and computing. world scientific; 2009.
[29] Koenker R. Quantile regression, vol. 38, Cambridge University Press; 2005.
[30] Frank EH. Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis. 2015.
[31] Bierens HJ. The Nadaraya–Watson kernel regression function estimator. 1988.
[32] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 1958;65(6):386.
[33] Kamiya G, Bertoldi P, et al. Energy consumption in data centres and broadband communication networks in the EU. Eur Comm Jt Res Cent 2024.
[34] Eiraudo S, Barbierato L, Giannantonio R, Porta A, Lanzini A, Borchiellini R, Macii E, Patti E, Bottaccioli L. A machine learning based methodology for load profiles clustering and non-residential buildings benchmarking. IEEE Trans Ind Appl 2023.
[35] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in python. J Mach Learn Res 2011;12:2825–30.
[36] Marholm S. localreg 0.5.0. URL https://pypi.org/project/localreg/.
[37] Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: 9th python in science conference. 2010.
[38] Eiraudo S, Schiera DS, Mascali L, Barbierato L, Giannantonio R, Patti E, Bottaccioli L, Lanzini A. Neural network-based energy signatures for non-intrusive energy audit of buildings: Methodological approach and a real-world application. Sustain Energy Grids Netw 2023;36:101203.
[39] Eiraudo S, Barbierato L, Giannantonio R, Patti E, Lanzini A, Bottaccioli L. Non-intrusive load disaggregation of industrial cooling demand with LSTM neural network. In: 2022 IEEE international conference on environment and electrical engineering and 2022 IEEE industrial and commercial power systems Europe. EEEIC/I&CPS Europe, IEEE; 2022, p. 1–6.