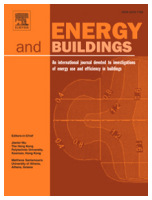
Energy & Buildings 349 (2025) 116495



Contents lists available at ScienceDirect

Energy & Buildings

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





Experimental application of a semi-parametric model for interpretable and accurate egression analysis of building energy consumption

Simone Eiraudo [a,b], Alfonso Gijón [b], Antonio Manjavacas [b], Daniele Salvatore Schiera [a,*], Luca Barbierato [a], Miguel Molina-Solana [b], Juan Gómez-Romero [b], Roberta Giannantonio [c], Lorenzo Bottaccioli [a], Andrea Lanzini [a]

[a] Energy Center Lab, Politecnico di Torino, Torino, 10138, Italy
[b] Department of Computer Science and Artificial Intelligence, Universidad de Granada, Granada, 18071, Spain
[c] TIM S.p.A, Via Gaetano Negri, 1, Milan, 20123, Italy

ARTICLE INFO

Keywords:
Semi-parametric models
Regression analysis
Hybrid models
Building consumption
Energy
Interpretable machine-learning

ABSTRACT

Regression analysis is a versatile tool with numerous applications across diverse domains. Its utility extends to several tasks, including forecasting, inverse modeling, anomaly detection, and pattern identification. Over the years, researchers have mainly focused on two regression categories: parametric and non-parametric analysis. In light of the benefits and drawbacks of both methods, this work introduces a semi-parametric approach, combining regression accuracy and interpretability. This is achieved by designing a hybrid model, that includes a physics-based sub-model and a neural network. The proposed data-driven pipeline is applied to a relevant case study from the energy sector, namely the analysis of building energy consumption, achieving high accuracy compared to the parametric approach. Results demonstrate an increase in the mean coefficient of determination, from 0.77 to 0.94, with a MAPE drop from 5.5 % to 2.2 %. Meanwhile, the semi-parametric model allows the assessment of the thermal behavior of the buildings, thereby offering an improvement over black-box approaches.

1. Introduction

Regression analysis involves estimating the relationship between one or more independent variables and one or more dependent variables. The former variables are named explanatory or input variables, while the latter represent the output of the regression models. Due to the widespread use of data-gathering in all scientific fields, regression analysis is currently used in sectors such as econometrics, engineering, sociology, and law [1]. The typical task of regression models is to examine how a variation in one observed variable affects, or appears to affect, the value of another [2]. This correlation between variables may be functional or not, and it can be described by means of either trivial or complex mathematical functions. In both cases, the use of regression analysis may be beneficial for analysts, as it can be employed for several tasks, such as forecasting, parameter identification, inverse modeling, measurement & verification [3], data screening, and others [4].

Regression models can be classified into two main categories: parametric and non-parametric models, depending on the mathematical function that describes the relationship between the input and output variables. The former are models that are defined by a finite number of parameters. Parameters are trainable coefficients whose values are determined on the basis of a fitting or training step. The simplest example of a parametric model is univariate linear regression. In such a case, a single output is calculated as a linear function of a unique explanatory variable. Similarly, multivariable linear models compute the dependent variable as a linear combination of the contribution of multiple input variables. However, more complex parametric regression models, such as polynomial or exponential ones, also exist. Parametric models are a computationally efficient solution that can be easily employed for physical parameter estimations. However, they typically exhibit low accuracy and are based on assumptions about the shape of the regression function.

Conversely, non-parametric regression models can capture complex relationships between outputs and explanatory variables. Their improved accuracy, combined with advances in computational capacity, has led to the widespread adoption of these regression algorithms. Indeed, Machine Learning (ML) algorithms — such as Neural Networks (NN), Gradient Boosting (GB), and Support Vector Machine (SVM) [5]— have garnered significant attention and become the focus of an

* Corresponding author.
E-mail addresses: simone.eiraudo@polito.it (S. Eiraudo), alfonso.gijon@ugr.es (A. Gijón), manjavacas@ugr.es (A. Manjavacas), daniele.schiera@polito.it (D.S. Schiera), luca.barbierato@polito.it (L. Barbierato), miguelmolina@ugr.es (M. Molina-Solana), jgomez@decsai.ugr.e (J. Gómez-Romero), roberta.giannantonio@telecomitalia.it (R. Giannantonio), lorenzo.bottaccioli@polito.it (L. Bottaccioli), andrea.lanzini@polito.it (A. Lanzini).

https://doi.org/10.1016/j.enbuild.2025.116495
Received 9 May 2025; Received in revised form 12 August 2025; Accepted 20 September 2025
Available online 25 September 2025
0378-7788/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

expanding body of literature, with applications spanning a wide range of research fields. A key aspect of non-parametric methods is that they do not rely on any prior assumptions about the data distribution, that is, they constitute an agnostic approach to regression. This represents a major advantage for its widespread application across different sectors. Indeed, these methods can easily be employed for those regression problems in which the functional laws underlying the distribution of the dependent variables with respect to the inputs are unknown, or are too complex to be modeled. Such tools make it possible, for instance, to analyze the impact of independent variables which, in a parametric approach, would have been neglected. However, non-parametric regression has a notable drawback: it lacks functional explainability, which means that it is not per se interpretable. Moreover, functional relationships may not be clear. This means that causality between the explanatory and dependent variables is not guaranteed, and issues of multicollinearity or non-causal correlation should be carefully considered when interpreting the results of non-parametric models [4].

Finally, semi-parametric models can be designed as hybrid approaches [6], combining a parametric function with a non-parametric term to address regression problems [7]. These models overcome most of the shortcomings of the previous approaches. Indeed, they can provide high regression accuracy [8], while preserving the interpretability of some modeled functional relationships. However, their widespread adoption is limited by a lack of extensive literature, which results in their employment being relegated to limited sectors, such as econometrics [9]. To the best of our knowledge, these models have rarely been applied in the energy sector field.

1.1. Regression analysis in the energy sector

Regression algorithms are fundamental tools for the analysis of energy systems, with many applications in measurement & verification [10], inverse modeling and parameter identification [11,12], forecasting renewable energy production [13], and more. Two main trends can be identified in the literature on regression analysis within the energy sector. The first involves the use of simple models to characterize the parametric impact of some relevant explanatory variable. However, this approach often involves heavy assumptions to reduce the complexity of the problem, neglecting relevant observable variables. This is, for instance, the case of univariate linear Energy Signature (ES) models for buildings [3]. These models describe the relationship between a thermal force, which is generally modeled considering the outdoor temperature, and the thermal demand of a building. This approach provides interpretable models, which may be described by means of a finite number of physics-grounded parameters, which, in turn, may be used to evaluate buildings' characteristic parameters and efficiency [14]. Although they are easy to implement, as they only require one input variable to be monitored, the model may not capture some important information [15]. Other examples of parametrical regression models are multivariable ones [16], which may consider multiple input variables.

The second trend is that of the black-box forecasting approach. The use of multiple input variables, along with the introduction of a number of efficient advanced ML-based methods, enhances the possibility of achieving high forecasting accuracy. Olu-Ajayi et al. [17] found that Deep NNs have the highest forecasting accuracy in predicting the energy consumption of residential buildings. Other algorithms have been equally employed, such as SVM. Cai et al. [18] recently used this algorithm to forecast the heating and cooling demand in buildings, demonstrating its prediction accuracy and robustness.

Both the previous approaches suffer from specific drawbacks. On the one hand, parametric regression models generally fit poorly with real-world data, as the impact of most of the variables that concur to the thermal phenomena of the building are neglected. Besides, most applications consider a very limited number of input variables —even just one, in most cases—. On the other hand, non-parametric models can also entail certain issues related to computational complexity, data-intensity and lack of interpretability. Indeed, such models do not offer insights into the underlying issues of buildings' thermal behavior, the impact of some specific input variables, or the causes of inefficiencies.

Given the limitations of both parametric and non-parametric methods, the lack of applications of semi-parametric models in the energy sector, and the importance of regression analysis, this paper proposes a semi-parametric regression algorithm for analyzing building thermal behavior. This approach can be expected to guarantee a high regression fit with the data, while preserving the interpretability of the parametrically modeled functions [6]. The proposed model includes a univariate ES model, which describes the thermal response of a building to changes in the outdoor temperature, and an Neural Network (NN) regression model, which is employed to infer the deviation of the impact of temperature from the expected modeled relationship, as well as to describe the effects of additional explanatory variables.

The proposed methodology was applied to a real-world case study, which involved two years of hourly load measurements from 18 buildings. A semi-parametric model was trained for each building to predict the thermal load and to estimate the characteristic parameters of the ES models. The model was then compared with reference parametric and non-parametric models. The main novelties and contributions presented in this paper can be resumed as follows:

- A semi-parametric model has been proposed for a regression analysis to overcome the shortcomings of both parametric and non-parametric algorithms.
- This model is applied to a real-world dataset from the energy analysis sector, representing, to the best of our knowledge, one of the first semi-parametric approaches in this research field.
- The parametric branch of the model is based on a reference physics-based model. The typical physical parameters of the problem can be calculated, ensuring interpretability and providing valuable insights into the regression task.
- The non-parametric regression branch can be employed to infer the influence of any additional explanatory variables, including the effects of unmodeled physics, thereby enhancing regression accuracy.

The remaining of the paper is organized as follows. Section 2 presents the semi-parametric regression model and the data-driven pipeline for its application. Section 3 dives into the analyzed real-world case study and provides some practical details for implementation. Section 4 describes and discusses the experimental results, while final conclusions and future works are presented in Section 5.

2. Methodology

2.1. Semi-parametric regression model

Parametric models are regression algorithms that assume a certain distribution of the output variables with respect to the input ones. This distribution is described by means of a finite number of parameters, which are the coefficients of the regression function [20]. Regression functions can be expressed as:

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

where y is the real output variable, X is the matrix of the modeled inputs, f is a parametric regression function, and ϵ are the residuals between the output of the regression algorithm, $f(X)$, and the real output. The latter term can include noise, abnormal values, deviations from the expected behavior with respect to the inputs, or unmodeled inferences from other observable variables. The parametric function can feature a variety of characteristic distributions, from simple linear regression functions [11], to polynomial or sigmoidal [21], among others. In order to select a proper function, it is necessary to consider the expected relationship between the input and the output variables. In the engineering sector, a suitable choice could be the underlying physical law governing

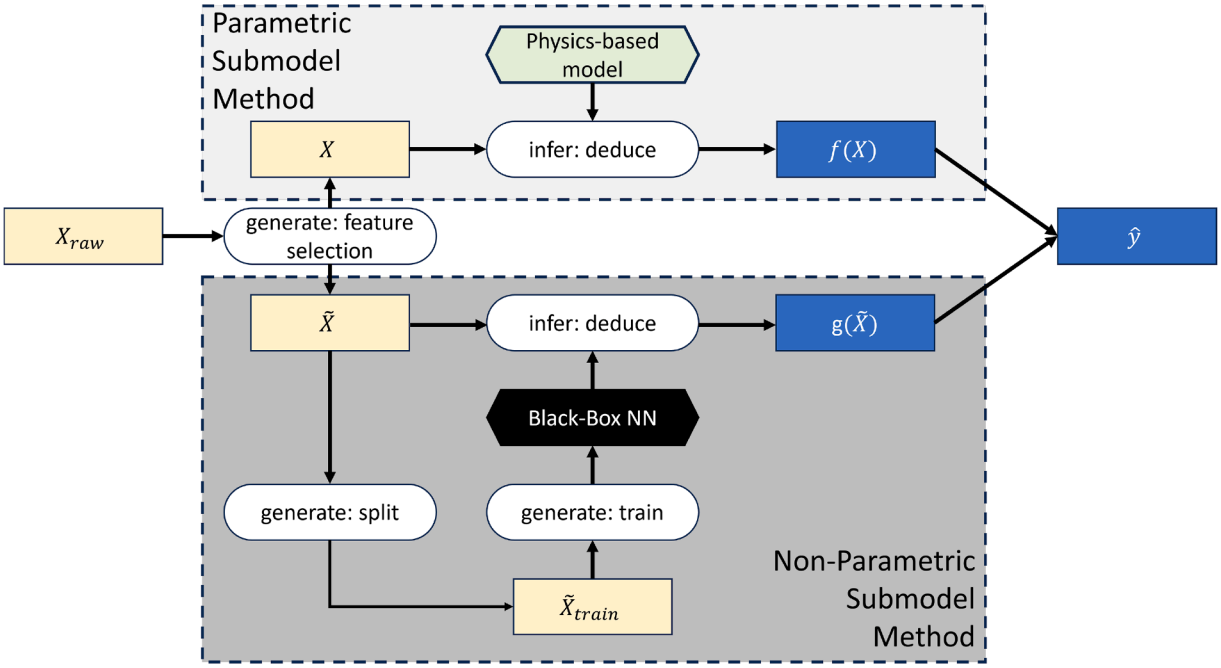


Fig. 1. Taxonomy of the proposed model. The rectangular, rounded and hexagonal blocks represent data, functions, and models, respectively, according to the method presented in van Bekkum et al. [19]. Yellow and blue blocks are used for inputs and outputs respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the problem. The typical model parameters can be trained by means of data-driven techniques or be predefined based on prior knowledge.

Non-parametric models, on the contrary, do not assume any specific shape for the function employed to calculate the dependent variable. They can be written in a similar manner to the previous one [22]:

$$y = g(\tilde{X}) + \epsilon \tag{2}$$

where g is the non-parametric regression function, and $\tilde{X}$ is the matrix of observed variables.

Non-parametric models, such as NNs, can capture complex relationships among the considered variables. Since such models do not require any domain-specific expertise to define the regression function, they are commonly employed. Moreover, additional input variables can be included in the regression task if necessary.

Semi-parametric models combine parametrical and non-parametrical sub-models one to overcome the limitations of the two previously mentioned approaches. In this paper, an additive semi-parametric model (see Fig. 1) is proposed, expressed as:

$$y = f(X) + g(\tilde{X}) + \epsilon \tag{3}$$

The first sub-model, $f(X)$, has to be selected according to the case study, as it will be later demonstrated for the building consumption analysis in the following subsection. The non-parametric sub-model, $g(\tilde{X})$ can be designed as a standard black-box regression approach. The aim of the resulting semi-parametric model is to minimize the error term by incorporating a larger number of input variables than the parametric model. Complex relationships are captured by using a NN as the non-parametric sub-model. At the same time, such a model is intended to preserve the benefits of the parametric regression task performed by means of a physics-based model, such as the identification of the characteristic coefficients of the problem, and to obtain a deep understanding of the relationship between the modeled variables.

2.2. Parametric model for the thermal characterization of buildings

Energy Signatures (ES) are energy models that are used to describe the responses of the thermal loads of buildings to variations in the outdoor variables. Univariate ES, which considers the outdoor temperature as an input, can be considered as a reference model for parametric regressions to describe the energy behavior of a building [3]. Such models prescribe that the power consumption from a building cooling system is a function of the outdoor temperature according to the following equation [23]:

$$P_{\mathrm{CLC}} = \begin{cases} 0 & \text{if } T_{\mathrm{ext}} <= T_{\mathrm{BP}} \\ \frac{k_{\mathrm{Tot}}(T_{\mathrm{ext}} - T_{\mathrm{BP}})}{\mathrm{COP}} & \text{if } T_{\mathrm{ext}} > T_{\mathrm{BP}} \end{cases} \tag{4}$$

where P_{CLC} is the electrical load of the cooling system, k_{Tot} is the total heat loss coefficient of the building, T_{ext} is the outdoor temperature, T_{BP} is the balance point temperature, and COP is the coefficient of performance of the cooling system. A detailed description of an univariate ES regression model is provided in Eiraudo et al. [23].

This approach, considering the relevant contribution of space cooling to the final electrical demand of many buildings, is indispensable for the energy characterization of buildings. Thus, this ES model is implemented as the parametric sub-model of the semi-parametric regression task.

2.3. Methodological framework

The regression tools presented in the previous sections were implemented in a data-driven pipeline, as shown in Fig. 2. This pipeline includes: *i)* a pre-processing step, the successive implementation of the *ii)* parametric sub-model and *iii)* non-parametric sub-model, and *iv)* a post-processing step.

The aim of the pre-processing step is to handle the raw historical series of observed variables, and to guarantee their exploitability for the regression task. To this aim, the format, resolution, and quality of the data are considered. First, a data-filtering step is performed. In this case, the input data are cleaned of any inconsistent measures. Elements containing *NaN* values are eliminated from the dataset. Moreover, the distributions of the most important explanatory variables, which, in this case, are the outdoor temperature, and the output variable, namely the electrical load, are considered to detect any possible measurement errors. For this purpose, a moving average of the value of the electrical load is calculated from the dataset elements, which are ordered according to

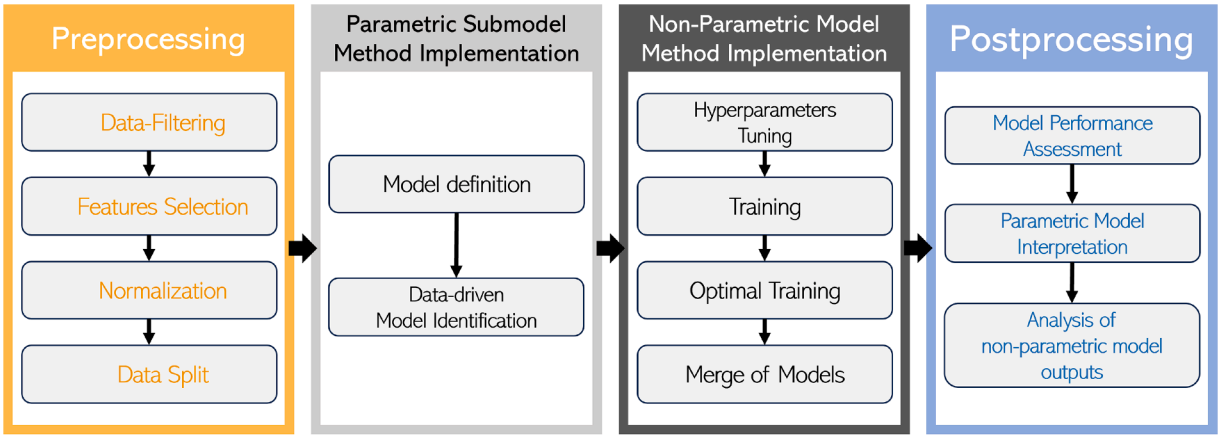


Fig. 2. Fundamental steps of the proposed data-driven pipeline.

the considered explanatory variables. Moreover, the standard deviation of the values is calculated over the domain by dividing the explanatory variable domain into bins. Values are filtered out of the dataset, if they fall outside the range of 3 standard deviations around the mean electrical load value from the bin to which they belong. From this first step, a filtered dataset, which we refer to as X_{raw}, of the inputs and outputs is obtained.

Secondly, the observed variables present in this dataset are considered. The variables that have to be modeled within the parametric branch are included in X. At the same time, all the variables that are expected to exert an impact on the output are included in the $\tilde{X}$ matrix, which will be the input of the non-parametric sub-model. It should be noted that the variables included in X could also be included in $\tilde{X}$ to investigate any potential deviations in their impact on the output variable, compared to the parametric sub-model output. Finally, the data are normalized to range $[0, 1]$ and split in both training and testing sets, with a percentage of 66 % and 33 % of the instances respectively.

The parametric sub-model implementation starts with the definition of the model to be employed. This can be done on the basis of physical laws, hypothesis, domain expertise or data exploration considerations. As explained in the previous section, the model considered for building analysis is the univariate ES model (see 4). Thus, when the coefficients of the parametric model are unknown, they can be estimated by either a training or a fitting step, as required by the non-parametric branch of the model.

The non-parametric sub-model is then implemented. In this case, this is done by considering the residuals from the predictions of the former branch, with respect to real values, as in following equation:

$$P_{\text{res}} = P - \hat{P}_{\text{phys}} \tag{5}$$

where P_{res} is the residual electrical load, P is the real load of the building, and $\hat{P}_{\text{phys}}$ is the output of the ES sub-model.

In the proposed model, a simple feed-forward NN was employed as the non-parametric regression algorithm. We employed a grid-search performed on a number of hyperparameters, including the activation function, learning rate, number of hidden layers and neurons per hidden layer. Training is then performed on the network that features the most adequate hyperparameters. Finally, the best achievable performance is obtained by considering the epoch at which the minimum loss function value was achieved in the previous step.

Hence, the outputs of the non-parametric branch are summed with those of the ES model to obtain the final model predictions. The performance of the model is compared with those of a parametric model and a non-parametric one, in terms of coefficient of determination and Mean Absolute Percentage Error (MAPE). The parametric sub-model is then employed to estimate the characteristic coefficients of the building studied. Finally, the impact of minor variables and the deviations from the modeled behavior can be discussed considering the outputs of the non-parametric branch.

3. Case study and experimental set up

3.1. Case study and variables

The model presented in the previous section was tested in a real-world case study, which involved 18 real-world buildings from an energy-intensive industrial sector. The analyzed buildings are all data

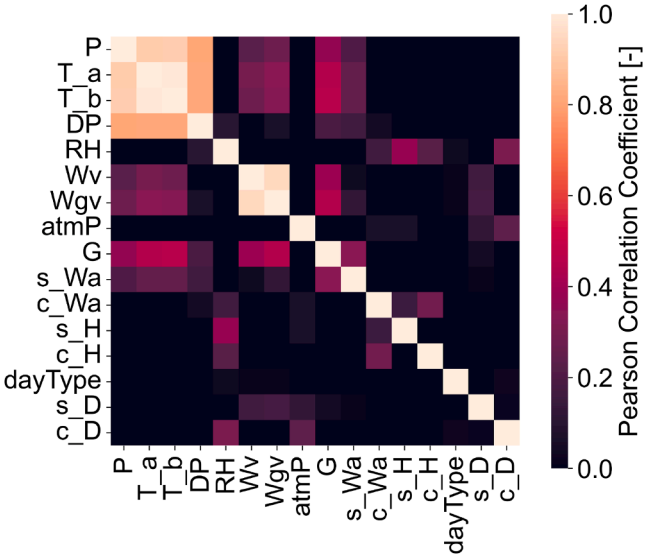


Fig. 3. Correlation matrix of the input and output variables for one of the studied real-world buildings. T_a and T_b both refer to outdoor temperature, but from two different meters.

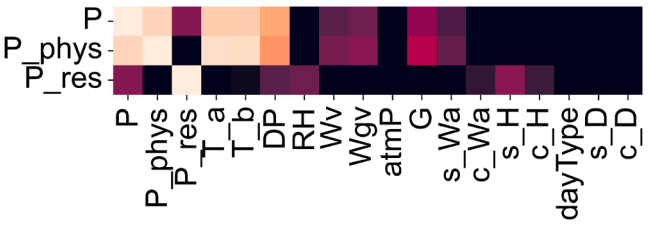


Fig. 4. Correlation matrix considering the output variable, whose predicted value is estimated by the physics-based parametric model and the residual term.

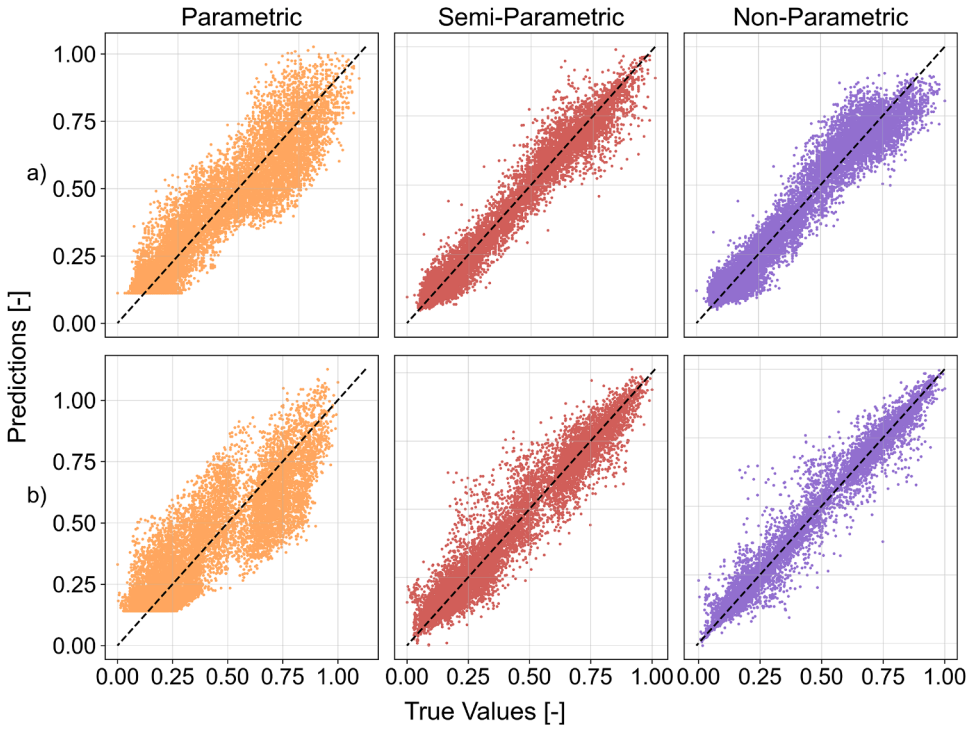


Fig. 5. Predicted and real output variable for two case studies (*Building A* and *Building B*), estimated by the parametric, hybrid and black-box approaches.

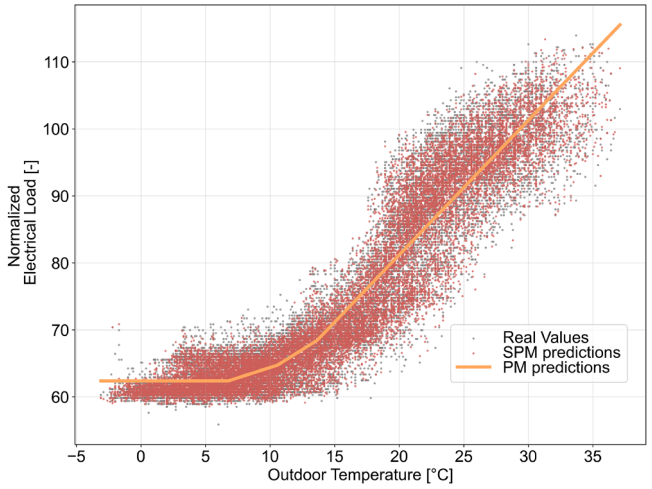


Fig. 6. Reference Energy Signature provided by the parametric model (PM) and predictions from the semi-parametric model (SPM) for one case study (*Building A*).

centers located in different regions throughout Italy. Although these facilities share some features with pure data centers, they are in fact telecommunication (TLC) central offices, which host network equipment mostly, like switching racks, servers and routers. Unlike high-variability data centers, TLC central offices typically operate under more stable and predictable IT loads, with less pronounced short-term variability. This characteristic increases the relevance of weather-dependent models, such as the ES approach, while still leaving room for improvement through the modelling of secondary influences. Indeed, cooling makes a fundamental contribution to the final energy demand of such buildings, because of the dense internal heat generated by the computing equipment. As discussed in Eiraudo et al. [23], the ES is particularly relevant for this case study, as air conditioning is the primary factor contributing to variations in the final electrical demand.

The dataset considered for this study contains two years of measurements with an hourly resolution. The variable included in the modeled matrix, X, is the outdoor temperature, while the input matrix of the NN, $\tilde{X}$, is obtained as follows:

$$\tilde{X} = \{T, DP, RH, W_v, W_{gv}, P_{\text{atm}}, G, s_{Wa}, c_{Wa}, s_H, c_H, day, s_D, c_D\} \quad (6)$$

where T is the outdoor temperature, DP is the dew point, RH is the relative humidity, W_v is the wind velocity, W_{gv} is the wind gust speed, P_{atm} is the atmospheric pressure, G is the solar radiation, s_{Wa} and c_{Wa} are the cyclic encodings of the wind angle (sine and cosine), s_H and c_H are the cyclic encodings of the hour of the day, *day* is a binary value that represents whether a day is a working day or a holiday, and s_D and c_D are the encodings of the day of the year. All these variables are expected to exert an effect on the energy demand of the building, but a parametric model that could clearly describe them would be too complex to design.

3.2. Software settings

All the algorithms were implemented and trained using Python 3.9. The experiments were run using a laptop with an 11th Gen Intel Core i7-1165G7 processor, 16 GB RAM and an Intel® Iris® Xe graphic card. NNs models were implemented and trained using the Keras library [24], while the grid-search procedure was conducted using the HyperBand tuner [25].

The code used for data preprocessing, building, training, and evaluating the models is available at this Github repository: https://github.com/alfonsogijon/Buildings_hybrid.

4. Results and discussion

The data-driven pipeline detailed in Section 2 was applied to the case study presented in Section 3. The pre-processing step led to a reduction of the dataset of about 9.9 %. The filtered out rows were predominantly samples containing NaN values and, albeit less frequently, some values that were detected as being abnormal measurements by the pre-processing algorithm.

Table 1
Coefficient of determination and MAPE calculated using the different regression models. The numbers report the mean value and standard deviation for the 18 buildings considered.

	R^2		MAPE	
	Train	Test	Train	Test
Parametric	0.772 ± 0.091	0.771 ± 0.093	$5.53\,\% \pm 1.48\,\%$	$5.54\,\% \pm 1.50\,\%$
Semi-Parametric	0.938 ± 0.044	0.905 ± 0.059	$2.22\,\% \pm 1.11\,\%$	$3.10\,\% \pm 1.03\,\%$
Non-Parametric	0.934 ± 0.048	0.905 ± 0.057	$2.32\,\% \pm 1.06\,\%$	$3.06\,\% \pm 0.95\,\%$

Before conducting the model fit process, a cross correlation analysis of the data was performed to observe the impact of the input variables on the output. These steps are performed to ensure the correct adoption of the parametric model, in this case, the ES model. Notice how, as shown in Fig. 3, the electrical load is significantly more correlated with the outdoor temperature than any other observed variable. This provides a data-based endorsement, in addition to the physical one, of the suitability of adopting the ES model.

A total of 18 parametric models were then trained, one for each studied building, according to the ES model described in Eiraudo et al. [23] and to the Eq. (4). The resulting trained models were employed to calculate the errors with respect to the predictions, which, in turn, were used to train the non-parametric sub-model, so as to finally provide the semi-parametric model predictions. Table 1 summarizes the performance results achieved by the proposed semi-parametric model and the considered benchmark models for the 18 case studies. As can be observed, the semi-parametric model outperforms the reference ES model, with an increase of 0.771 to 0.905 of the mean coefficients of determination, R^2. This results in the regression error being almost halved. The resulting model performance essentially matches that of the reference NN-based black-box approach. These results prove that semi-parametric models can combine the interpretability of parametric models with the regression accuracy of complex ML methods.

It is worth noting that a further confirmation of the reliability of the physics-based model is obtained from a correlation analysis of the electrical load predicted by the parametric sub-models and the residuals. Indeed, as can be observed from the correlation heat map in Fig. 4, the ES model is capable of effectively capturing all the linear correlation with the outdoor temperature and, in turn, overriding any minor relationships detected from variables cross-correlated with the temperature, like the radiation. This is easily observed from a quick look at the almost uncontaminated correlation vector of the residual electrical load, P_{res}. Notwithstanding this, the successive implementation of a non-parametric model led to the aforementioned boost in accuracy. This confirms the appropriateness of adopting a NN to model the complex inference of the minor explanatory variables. It is worth noting that, due to the relatively stable internal loads of TLC central offices, the most significant NN contributions arose from systematic corrections to the ES baseline rather than from tracking frequent, large, abrupt IT-driven changes. Nevertheless, the NN branch successfully learned and reproduced certain recurring operational patterns, such as weekday versus weekend cooling demand differences, which are not explicitly captured by the parametric model.

Fig. 5 depicts the results for 2 of the 18 considered case studies. The two rows show the prediction accuracy of the 3 considered regression approaches for *Building A* and *Building B*. The first case reports MAPE values of 4.70 %, 2.08 %, and 3.09 % for the parametric, semi-parametric, and non-parametric models, respectively. In this case, the semi-parametric model not only preserves the advantages of a physic-grounded regression model, compared to the agnostic black-box approach, it also provides a consistent improvement in performance. On the other hand, the non-parametric model shows the best performance in the second case study, with MAPE values of 5.88 %, 2.62 %, and 1.51 %, respectively. In this case, the lower performance of the semi-parametric model may be linked to the marked deviations of the real values from the modeled ES, as shown by the dispersion of the points in the plot in the bottom left. In the case of TLC facilities, sudden large load spikes are rare compared to pure data centers. However, the hybrid model still demonstrated its ability to anticipate smaller but recurring load variations linked to operational routines. For example, in Building A, regular weekday morning increases in load-likely associated with scheduled network operations, were well captured by the hybrid approach, reducing the MAPE by more than 50 % compared to the parametric baseline. Conversely, rare, non-repeating anomalies (e.g., maintenance shutdowns) remained difficult to predict without retraining.

The integration of the non-parametric sub-model with the physics-based one may be more intuitively understood by analyzing Fig. 6, where the ES and the predictions of the semi-parametric model from *Building A* are shown. It can be observed how the physics-based model shows a good fit to the data. Such a model enhances the identification of the important characteristics parameters of the studied building, such as the balance point temperature and cooling system COP, which depict values of 6.9 °C and 3.72, respectively.

On the other hand, the hybrid approach leads to an effective performance boost, as the predictions of the model move from the reference ES line to a more precise fit of the output variable values. The displacement of the prediction points is due to the complex inference estimated by the NN model from the additional explanatory variables contained in $\tilde{X}$.

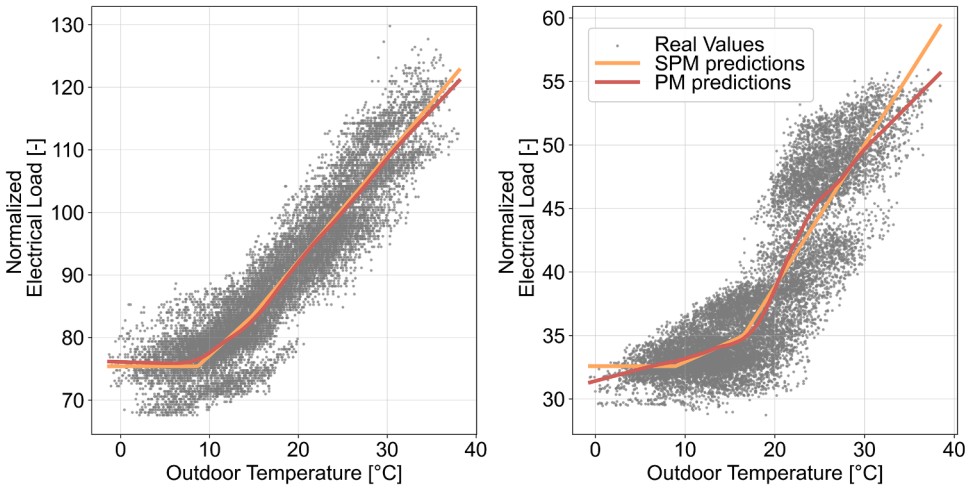


Fig. 7. Deviations from the modeled impact of the outdoor temperature on the energy consumption for *Building C* and *Building B*.

Finally, the outputs of the semi-parametric model can be investigated in more detail to provide further insights of the regression problem. In particular, it may be interesting to discuss the reliability of the hypothesis adopted when assuming a parametric model. For instance, in this case, linear behavior was assumed, since it is the shape that ES models fit. The predictions of the hybrid model can be projected with respect to the modeled explanatory variable, as shown in Fig. 7. The resulting reference line can then be compared with the one obtained when using the ES model. In the case shown in the subplot on the left, which refers to *Building C*, the correctness of the assumed impact of the outdoor temperature on the electrical load can be confirmed by considering the irrelevant deviations of the projected line obtained by the hybrid model, with respect to the original ES. On the contrary, if the reference outputs for *Building B* in the subplot on the right are considered, three important deviations from the assumed behavior can be observed at lower temperatures than 5 °C, from about 20 °C to 25 °C, and over 3 °C, respectively. This suggests that this particular case study needs to be studied in more detail to understand the possible causes of the deviation from the expected response to changes in the outdoor temperature.

5. Conclusions

This paper introduces a hybrid model approach for semi-parametric regressions, combining a parametric physics-based sub-model and a black-box non-parametric one. The former one is modeled according to the physics of the problem, while the latter comprises a feed-forward NN. The data-driven pipeline, which includes pre-processing, model fitting, hyperparameter tuning, and post-processing, is described.

The proposed model has been applied in a real-world case study related to the energy sector and, specifically, to the analysis of energy consumption in buildings. The model was employed to predict the energy demand of 18 buildings, and its performance was evaluated against reference parametric and non-parametric models. The proposed model was able to reduce the mean regression error by almost half, compared to the parametric model, and to achieve a mean absolute percentage error of 3.10 % and a determination coefficient of 0.905. These performance metrics match those of non-parametric models. However, compared to the latter, the semi-parametric model was able to preserve the physics-grounding, and thus the interpretability of the trained model. Because of that, it is better suited to estimate the typical parameters of the analyzed buildings and to provide relevant insights into the problem dynamics. While the present application focuses on TLC central offices, whose internal loads are generally more stable than those of high-variability data centers, the proposed methodology remains applicable to a wide range of building types. In scenarios with greater operational volatility, the non-parametric branch could be expected to play an even more prominent role in capturing rapid load changes, provided such patterns are sufficiently represented in the training dataset.

Considering the importance of the interpretation of regression models, as well as the scarce literature discussing this topic, an explainability analysis step regarding the non-parametric sub-model will be undertaken as the first future development of the present study. This step can be used to assess the impact of the additional input variables and to shed light on the unknown physics underlying the regression task.

The performance of the models in terms of computational complexity and data intensity should also be analyzed in depth to assess the potential for Big Data applications. Furthermore, the modular design of the proposed semi-parametric model is intended to support applications across different research sectors. Therefore, future work will focus on analyzing datasets from diverse case studies and implementing specific parametric models for each application.

CRediT authorship contribution statement

Simone Eiraudo: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation; **Alfonso Gijón:** Writing – review & editing, Software, Methodology, Investigation, Conceptualization; **Antonio Manjavacas:** Writing – review & editing, Software, Methodology, Investigation, Conceptualization; **Daniele Salvatore Schiera:** Writing – review & editing, Supervision, Formal analysis, Data curation; **Luca Barbierato:** Writing – review & editing, Supervision; **Miguel Molina-Solana:** Writing – review & editing, Supervision; **Juan Gómez-Romero:** Writing – review & editing, Supervision; **Roberta Giannantonio:** Supervision, Resources; **Lorenzo Bottaccioli:** Writing – review & editing, Supervision; **Andrea Lanzini:** Writing – review & editing, Supervision, Project administration, Formal analysis.

Data availability

The data that has been used is confidential.

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 his Ph.D. scholarship. This work was partially funded by the European Union NextGenerationEU/PRTR and the Spanish Ministry of Economic Affairs and Digital Transformation (IA4TES project, MIA.2021.M04.0008). AG, AM, MMS and JGR were also funded by FEDER/Junta de Andalucía (D3S project, P21.00247; and SE2021 UGR IFMIF-DONES) and by MICIU/AEI/10.13039/501100011033 (SINERGY project, PID2021.125537NA.I00).

References

[1] A.O. Sykes, An introduction to regression analysis (1993).
[2] N.R. Draper, Applied Regression Analysis, McGraw-Hill. Inc, 1998. https://doi.org/10.1002/9781118625590
[3] evo, 2020. Efficiency Valutation Organization (EVO). IPMVP's Snapshot on advanced measurement & verification. 2020.
[4] H. Fu, J.-C. Baltazar, D.E. Claridge, Review of developments in whole-building statistical energy consumption models for commercial buildings, Renew. Sustain. Energy Rev. 147 (2021) 111248. https://doi.org/10.1016/j.rser.2021.111248
[5] B. Sekeroglu, Y.K. Ever, K. Dimililer, F. Al-Turjman, Comparative evaluation and comprehensive analysis of machine learning models for regression problems, Data Intell. 4 (3) (2022) 620–652. https://doi.org/10.1162/dint_a_00155
[6] M. Manfren, M. Sibilla, L. Tronchin, Energy modelling and analytics in the built environment—A review of their role for energy transitions in the construction sector, Energies 14 (3) (2021) 679. https://doi.org/10.3390/en14030679
[7] A. Gijón, S. Eiraudo, A. Manjavacas, L. Bottaccioli, A. Lanzini, M. Molina-Solana, J. Gómez-Romero, Explainable hybrid semi-parametric model for prediction of power generated by wind turbines, in: L. Franco, C. de Mulatier, M. Paszynski, V.V. Krzhizhanovskaya, J.J. Dongarra, P.M.A. Sloot (Eds.), Computational Science – ICCS 2024, Springer Nature Switzerland, Cham, 2024, pp. 299–306. https://doi.org/10.1007/978-3-031-63775-9_21
[8] P. Stoffel, C. Löffler, S. Eser, A. Kümpel, D. Müller, Combining data-driven and physics-based process models for hybrid model predictive control of building energy systems, in: 2022 30th Mediterranean Conference on Control and Automation (MED), IEEE, 2022, pp. 121–126. https://doi.org/10.1109/MED54222.2022.9837277
[9] B. Xu, Y. Luo, R. Xu, J. Chen, Exploring the driving forces of distributed energy resources in China: using a semiparametric regression model, Energy 236 (2021) 121452. https://doi.org/10.1016/j.energy.2021.121452
[10] S. Rouchier, Bayesian workflow and hidden Markov energy-signature model for measurement and tion, Energies 15 (10) (2022) 3534. https://doi.org/10.3390/en15103534
[11] C. Rasmussen, P. Bacher, D. Calì, H.A. Nielsen, H. Madsen, Method for scalable and automatised thermal building performance documentation and screening, Energies 13 (15) (2020) 3866. https://doi.org/10.3390/en13153866
[12] F. Fernández de la Mata, A. Gijón, M. Molina-Solana, J. Gómez-Romero, Physics-informed neural networks for data-driven simulation: advantages, limitations, and opportunities, Physica A: Statistical Mechanics and its Applications 610 (2023) 128415. https://doi.org/10.1016/j.physa.2022.128415
[13] A. Gijón, A. Pujana-Goitia, E. Perea, M. Molina-Solana, J. Gómez-Romero, Prediction of wind turbines power with physics-informed neural networks and evidential uncertainty quantification, arXiv preprint arXiv:2307.14675 (2023). https://doi.org/10.48550/arXiv.2307.14675

[14] A. Acquaviva, D. Apiletti, A. Attanasio, E. Baralis, L. Bottaccioli, F.B. Castagnetti, T. Cerquitelli, S. Chiusano, E. Macii, D. Martellacci, et al., Energy signature analysis: knowledge at your fingertips, in: 2015 IEEE International Congress on Big Data, IEEE, 2015, pp. 543–550. https://doi.org/10.1109/BigDataCongress.2015.85

[15] G. Baasch, A. Wicikowski, G. Faure, R. Evins, 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, pp. 223–232. https://doi.org/10.1145/3360322.336083

[16] L. Tronchin, M. Manfren, B. Nastasi, Energy analytics for supporting built environment decarbonisation, Energy Procedia 157 (2019) 1486–1493. https://doi.org/10.1016/j.egypro.2018.11.313

[17] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, S. Ajayi, Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques, J. Build. Eng. 45 (2022) 103406. https://doi.org/10.1016/j.jobe.2021.103406

[18] W. Cai, X. Wen, C. Li, J. Shao, J. Xu, Predicting the energy consumption in buildings using the optimized support vector regression model, Energy 273 (2023) 127188. https://doi.org/10.1016/j.energy.2023.127188

[19] M. van Bekkum, M. de Boer, F. van Harmelen, A. Meyer-Vitali, A.t. Teije, Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases, Appl. Intell. 51 (9) (2021) 6528–6546. https://doi.org/10.1007/s10489-021-02394-3

[20] H.F.F. Mahmoud, Parametric versus semi and nonparametric regression models, arXiv preprint arXiv:1906.10221 (2019). https://doi.org/10.48550/arXiv.1906.10221

[21] P. Nageler, A. Koch, F. Mauthner, I. Leusbrock, T. Mach, C. Hochenauer, R. Heimrath, Comparison of dynamic urban building energy models (UBEM): sigmoid energy signature and physical modelling approach, Energy Build. 179 (2018) 333–343. https://doi.org/10.1016/j.enbuild.2018.09.034

[22] J. Fox, Nonparametric regression, Appendix to: An R and S-PLUS Companion to Applied Regression (2002) 1–7. https://doi.org/10.1002/0470013192.bsa446

[23] S. Eiraudo, D.S. Schiera, L. Mascali, L. Barbierato, R. Giannantonio, E. Patti, L. Bottaccioli, A. Lanzini, Neural network-based energy signatures for non-intrusive energy audit of buildings: methodological approach and a real-world application, Sustain. Energy Grids Netw. 36 (2023) 101203. https://doi.org/10.1016/j.segan.2023.101203

[24] F. Chollet, et al., Keras, 2015, (https://keras.io).

[25] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar, Hyperband: a novel bandit-based approach to hyperparameter optimization, J. Mach. Learn. Res. 18 (185) (2018) 1–52. https://doi.org/10.48550/arXiv.1603.06560