ORIGINAL RESEARCH Open access

Regression-Based Multivariable Energy Signatures: A Comparative Study Using Real-World Building Data

Thomas Jensen[1*], Mikkel Hansen[1]

Abstract

Developing an ecosystem of building energy models is vital to accelerate retrofit initiatives aimed at boosting energy performance and fulfilling sustainability commitments. Such representations should illuminate a building's energy dynamics, the effect of external drivers, and the roots of operational waste. Energy Signatures can fulfill this need, notably regarding HVAC-related demand. Univariate formulations, which overlook the influence of solar irradiation, have been commonly applied to Energy Signature studies. This investigation puts forward Multivariable Energy Signatures incorporating ambient air temperature and incident solar radiation. Deploying multivariable non-parametric techniques on a real-world dataset differentiates this effort from earlier contributions within the ES domain. This yielded an average enhancement in the coefficients of determination from 0.768 to 0.804, evaluated over 103 real-world instances. Additionally, Neural Networks proved superior to multiple established algorithms in the literature concerning precision, stability, and capacity to handle growing data volumes. The study further examines challenges tied to input data granularity and puts forward dedicated visualization strategies to operationalize Multivariable Energy Signatures for diagnostic purposes.

Keywords Buildings, Energy audit, Energy signature, Multivariable regression analysis

*Correspondence:
Thomas Jensen
thje@dtu.dk
[1] Department of Intelligent Power Engineering, Faculty of Engineering, Technical University of Denmark, Lyngby, Denmark

Introduction

As curbing energy demand and greenhouse gas emissions has risen to prominence on the global agenda, building energy performance has drawn scrutiny from governing bodies and scientific circles alike [1]. This surge in focus originates from the enormous share of worldwide energy use attributable to buildings. According to a 2024 United Nations assessment, this sector was responsible for 34% of total planetary energy consumption. Concurrently, buildings account for roughly 10 gigatons of CO_2 emissions, which equate to 37% of all anthropogenic emissions [2]. To date, the measures implemented to mitigate this sector's climate impact have proven insufficient, necessitating extraordinary efforts over the next several years to bring it into line with climate targets. Indeed, despite gains in energy efficiency, only a modest contraction in the sector's environmental load was observed following the peak usage year of 2015 [3]. This stagnation stems partly from the relentless expansion of floor area globally, and the rising weight of particular end uses, such as comfort cooling. The latter made up approximately 9% of end-use energy in 2022 [4]. Despite this relatively contained fraction, it is worth emphasizing that cooling-related energy uptake has multiplied by a factor of 2 over the preceding two decades and continues to rise steeply. Even accounting for technological refinements in cooling apparatus and the concomitant performance gains, electricity consumption for cooling is forecast to double again before 2040. Owing to these dynamics, upgrading the existing built stock has been singled out in a United Nations publication [5] as a pivotal

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

intervention. Explicitly, the document advocates that refurbishment measures should reach 3% of the entire existing building inventory each year to honor sustainability pledges.

The foundational move to trigger such a broad-scale refurbishment wave involves supplying robust, readily implementable energy assessment tools, effectively forging an “ecosystem” of building energy representations [6]. These instruments ought to, among other capabilities, generate benchmark consumption profiles and aid in distilling a descriptive model of a building’s energy conduct. Both objectives can be met through regression-based techniques. These constitute data-centric procedures that deduce a mapping between one or several predictors, termed independent variables, and a response, or target variable. When the former aim is pursued, such algorithms operate in a predictive mode. Concretely, they serve to establish reference baselines or project novel data points, for example, prospective ones. These algorithms are usually judged on their fidelity in forecasting the target quantity, customarily energy usage. Alternatively, when regression routines serve the goal of model identification, they can yield profound understanding of characteristic building operation, facilitate the deciphering of usage signatures and inefficiency sources, and permit the extraction of defining coefficients and metrics. Within this scope, they can also be deployed to estimate upfront gains from implementing specific energy conservation interventions.

Over the years, scholars have introduced a variety of regression formalisms, which can be classified into two principal branches: parametric and nonparametric. The first category encompasses inferential structures that impose a predefined architecture on the mapping from predictors to the outcome. A case in point is multivariate linear regression, in which a rectilinear relationship is posited between independent variables and the dependent variable. Broadly, parametric formulations cover any inferential construct characterized by a bounded parameter set. By contrast, nonparametric formulations impose no rigid functional mold on the linkage bridging predictors and the target. Nonparametric constructs include, for example, deep Neural Networks.

Investigative work in the building science arena has bifurcated into two contrasting directions. One stream adheres to elementary parametric constructs, typified by single-variable linear regression [7]. Such a route is typically selected in deployments geared toward model extraction, given the interpretational clarity afforded by the limited parameter count that encodes the inferential tie. Nevertheless, owing to the stringent simplifications invoked to constrain model intricacy, this path delivers modest fidelity. The other stream apparent in academic output involves embracing highly elaborate nonparametric structures, such as deep Neural Networks, for prognostic applications. Here, superior reproduction of the reference output sequence may be attained. Yet, these models provide no principled interpretation of their internal coefficients, and explanatory frameworks have seldom been transferred to regression contexts [8]. Among the additional limitations of this paradigm, these regression constructs prove unusable for the essential task of pre-project energy-saving quantification.

Beyond this, there exists no unified stance on the temporal granularity best suited for regression modeling. Intervals spanning monthly, diurnal, hourly, and sub-minute acquisitions have all been exploited for regression undertakings in the building analysis field [7]. Densely sampled schemes reduce the time required to build an adequate corpus for fitting trustworthy regression models. Still, their efficacy may be compromised by complications arising from protracted thermal transients within structures, driven by heat storage effects [7], leading to misspecified representations. In contrast, sparsely sampled schemes risk demanding prohibitively long acquisition spans, on the order of years, to amass a satisfactory dataset for fitting serviceable models.

This work strengthens the research stream on regression-based building analytics by providing a systematic comparison across parametric and nonparametric formulations. Multiple regression techniques from the Machine Learning (ML) arena and classical statistical traditions have been benchmarked against a large-scale real-world case repository. This corpus covers 103 industrial sites from the Telecommunications (TLC) sector, a rapidly expanding, energy-intensive domain. Suitable working assumptions are adopted to match the experimental setting, accommodating multiple time granularities and multiple regressors. Beyond this, the reference baseline produced by the trained mappings is rendered and decoded as an MES to advance the comprehension of building energy dynamics and to extract meaningful insights into how

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

driving variables modulate building consumption. Within these MES, ambient climatic variables serve as predictors of building electrical load. The paper's contributions can be condensed as follows:

• Multivariable regression schemes have been formulated to lift model fidelity and yield insight into the contribution of multiple driving factors.

• A head-to-head comparison of six regression approaches—namely, Ordinary Least Squares (OLS), polynomial regression, quantile linear regression, local regression, kernel regression, and NN-based regression—has been performed.

• The regression approaches were assessed across the different temporal granularities in the Energy Signature (ES) scholarship, specifically hourly, daily, and monthly resolutions.

• Every model and time resolution was tested against an extensive real-world case collection comprising 2 years of observations from 103 industrial buildings in the TLC sector.

• The strongest models are then examined and translated as MES. Dedicated visualization tools enabled us to unpack the effect of explanatory variables across distinct portions of the domain.

The chosen multivariable strategy, combined with nonparametric regression methods and an experimental rollout on a real-world dataset, marks a departure from existing published work. Applying the introduced methodology aims to improve the precision of ES representations while maintaining the analytical lens focused on the variables most consequential for cooling-related demand in buildings.

Literature Review on Regression Models and Related Applications

Regression models are vehicles for formalizing a dependence to recover a target variable from one or more predictor variables. In other words, they can be understood as mappings that capture how the response variable responds to the explanatory variables. The generic formulation of a regression function is:

$$Y = f(X) + e, \quad (1)$$

Where X stands for the matrix of independent variables, Y designates the response vector, (X) represents the regression mapping, and e denotes an additive error contribution. This contribution may subsume statistical noise and the gap between actual output observations and the modeled output (X) attributable to unaccounted explanatory variables—the smaller the error contribution e, the more effective the regression model. A variety of routes can therefore be followed to compress the error term of regression mappings, among them, modeling a wider set of explanatory variables and adopting more capable regression functions f. Regression models have been widely used across the building analytics domain for multiple purposes, including projection, anomaly detection, and inverse characterization. Among the pivotal regression instruments in this space are ES. These models deduce the thermal end-use of buildings from outdoor climatic variables. They may rest on conventional statistical regression frameworks or on ML-based ones.

Conventional frameworks are parametric regression models, where the formulation is describable through a finite, fixed parameter set. This holds, for instance, for linear regression, where the target variable is a linear combination of the explanatory variables, governed by a parameter vector. Such a regression model can be written as:

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

$$Y = \beta X + e, \quad (2)$$

Where β denotes the parameter vector, its length corresponding to (p + 1), with p indicating the tally of explanatory variables in play, and an added parameter β_0 residing in the vector to capture an intercept term, while a column of ones is adjoined to the input variable matrix X. To illustrate, consider a single input variable employed within the regression setup. This describes a univariate linear regression case, where X is a 2xN matrix, N being the number of instances encompassed in the problem, and β a vector of length 2. Note that when manipulating temporal sequences, N corresponds to the count of time steps incorporated. Many deployments of this univariate linear regression model for ES analysis are reported in the literature for both residential [9] and commercial buildings [10]. These workflows are typically executed in parameter-extraction mode. For instance, in Baasch *et al*. [11], characteristic coefficients of simplified building analogs are determined. A modest twist on the linear regression theme is the Change Point (CP) model, which is entrenched as the benchmark statistical ES formulation for thermal characterization [12]. CP models are univariate segmented regression architectures with 1-6 parameters that capture the baseline offset, segment slopes, and transition thresholds [13]. These models are typically fitted via OLS. Yet, Meng *et al*. [14] resorted to quantile regression to calibrate a CP regression construct. All these linear paradigms rest on weighty, sometimes brittle, presuppositions.

Furthermore, they all rely solely on outdoor temperature as the explanatory variable. A sparse set of multivariable linear regression examples dots the literature. Among others, Tronchin *et al*. [15] embraced a multivariable tactic. However, the outcomes were substantiated only against synthetic data, and the regression tools were limited to linear varieties. Other investigators exploited richer parametric models. Nageler *et al*. [16] applied a sigmoid ES approach to residential and commercial stock, achieving strong concordance with this data-driven proxy, physics-grounded simulations, and field measurements. Afshari and Friedrich [17] pursue multivariable regression approaches to capture the energy profile of Abu Dhabi. Consumption is aggregated at the tier, so no building-level dissection is possible. In this setting, the effect of selected input variables is styled as linear, while others are posited to exert a non-linear influence on the target quantity. A multivariable approach is followed in Belany *et al.* [18] to project lighting system electrical demand, employing, among other techniques, polynomial regression and using horizontal global radiation as a predictor. Overall, parametric models are straightforward to fit and streamline inferential statistical evaluation. They nonetheless fall short on accuracy and lean on an a priori commitment to the qualitative shape of the regression mapping. When it comes to the intricate challenge of representing heating- or cooling-attributable consumption, researchers have largely defaulted to outdoor air temperature as the sole independent variable in statistical models [7].

Nonparametric formulations break free from the constraints that bind conventional approaches. These regression tools impose no prior assumptions about the distributional shape linking the output to the inputs and can faithfully reproduce any arbitrary relationship, no matter its complexity. A canonical example is Deep NNs. This algorithmic family has been extensively deployed across the building sector, overwhelmingly for forecasting applications. Its use in estimating building ES, however, remains rare. Westermann *et al*. [19] presented one example of a nonparametric ES, in which the thermal response of a cluster of buildings was examined. The authors rightly emphasize the rich diagnostic value ES brings to building assessment, yet they simultaneously foreground the central drawback of nonparametric formulations—their lack of physical interpretability.

Sekeroglu *et al*. [20] present a comprehensive compilation of nonparametric regression techniques spanning NNs, Gradient Boosting, Support Vector Machines, and more. Alizamir *et al.* [21] pitted statistical methods, such as Multivariate Adaptive Regression Spline (MARS), against ML-based algorithms, including Gradient Boosting Tree (GBT), NNs, and Adaptive Neuro-fuzzy systems, to predict daily solar radiation using temperature, wind speed, and humidity as predictors. GBT outperformed its counterparts. The MARS algorithm also demonstrated strong prediction capability. A further contribution to this category within the ES literature comes from Rouchier [10], who formulates a Hidden-Markov ES model and benchmarks it against the standard CP construct. The nonparametric model the author develops achieves higher regression fidelity and is ultimately deployed to gauge the avoided energy uptake—the energy savings—resulting from an executed retrofit measure. In sum, nonparametric models can deliver meaningful gains in accuracy compared to

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

traditional paradigms. At the same time, they exact a heavier data toll, since both qualitative structure and quantitative specifics of the regression mapping must be learned, and they may resist easy deployment for inferential statistics. Furthermore, regarding Measurement and Verification (M&V) tasks, they cannot support an ex-ante assessment of savings associated with a prospective retrofit measure. By contrast, within physics-anchored models, varied retrofit scenarios can be simulated by adjusting physically meaningful model coefficients [16], thereby enabling ex-ante evaluation of energy savings.

As the foregoing survey reveals, the literature has largely bifurcated into two camps. The first leans on imposing strict simplifications to tame problem complexity. This tendency often sidelines the role of multiple input variables and settles on off-the-shelf linear regression schemes. The payoff is models that are straightforward to interpret—well-suited for parameter extraction, building characterization, and ex-ante retrofit savings estimation—but that deliver modest accuracy and offer little visibility into the variables that shape real building thermal dynamics. The second camp capitalizes on complex, agnostic regression algorithms that learn both the quantitative and qualitative contours of the regression mapping. This yields high accuracy and can accommodate multiple inputs. However, it largely forecloses applications beyond forecasting. Without any physical mooring or interpretive tooling, little headway can be made in unpacking building thermal behavior, sizing up prospective savings, or tracing the causes of inefficiency. More recently, a stream of approaches—among them Physics-Informed Neural Networks (PINN) [22] and Physics-Guided Neural Networks (PGNN) [23]—has surfaced under the broader Physics-Informed Machine Learning (PIML) banner [24]. These algorithms aim to encode and weave physical laws into NNs, fostering the training of more generalizable predictive models that require sparser data than conventional ML pipelines. Yet, persistent concerns around PIML algorithms—including optimization, convergence, and computational cost—must be resolved to deliver a more stable and widely applicable framework [24]. For these reasons, further work is still needed to crystallize an accurate, interpretable, and broadly adoptable benchmark model for ES.

Materials and Methods

Data-driven energy modeling paradigms have proven their mettle across diverse applications, from energy characterization to operational management. They demand markedly less computation and modeling effort than physics-based models and dispense with the need for exhaustive information about a building's thermal fabric. Within the data-driven landscape, ES occupy a well-established niche. These are highly informative, data-driven regression models that embed the physical principles underlying the thermal equilibrium of buildings. They can therefore sharpen the understanding of building thermal behavior, the influence of weather variables, and the likely sources of inefficiency. A robust, capable, and widespread adoption of such analytical instruments may, in turn, underpin building stock benchmarking and the quantification of energy savings attributable to specific efficiency measures. In general terms, a building's thermal behavior can hinge on (i) indoor temperature, (ii) outdoor temperature, (iii) wind speed [25], (iv) solar radiation [17], (v) occupancy and other internal heat sources, and (vi) characteristic building parameters, such as the overall heat loss coefficient. The widely used linear univariate ES imposes severe assumptions across most of these dimensions, reducing the problem to a narrow examination of outdoor temperature's role in estimating the overall heat loss coefficient. However, in many settings, one or more of the remaining factors can markedly sway building consumption. To better appreciate the weight of these variables, it is helpful to set out the thermal balance equation for a generic building:

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

Where $\frac{dT_{in}}{dt}$ captures the indoor temperature drift over time;ϕ_T denotes the thermal power exchanged with the outside through the building envelope via conduction; ϕ_Sol is the solar-driven heat gain; ϕ_Cond is the thermal power supplied by the air conditioning system; and ϕ_St captures internal heat generation. Assuming a fixed indoor temperature is held,

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

and the air conditioning system can offset the residual contributions to uphold this setpoint, we set $\frac{dT_{in}}{dt}$ to zero. The conditioning system contribution, ϕ_Cond, then balances the negative of the remaining terms, registering a positive value when heating is needed to defend the setpoint and a negative value when cooling is required. Envelope heat transmission is frequently reduced to a function of outdoor temperature alone. Yet, folding in the additional influence of solar radiation incident on the envelope—via the solar-air temperature concept [26]—yields a far richer description of envelope heat exchange. The solar-air temperature is expressed as:

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

Where T_{ext} is the outdoor temperature, α is the envelope's solar absorptivity, I is the total solar irradiance on the envelope, and h_0 is the convective heat transfer coefficient. This coefficient itself can be shaped by wind speed and orientation; the brisker the wind, the more vigorous the heat exchange between air and the envelope surface [25]. The second term in the thermal balance, ϕ_Sol, is governed by solar radiation, solar position, and building-specific parameters such as the Solar Heat Gain Coefficient. Solar radiation can thus be seen to influence 2 out of the three thermal loads the conditioning system must handle. The internal gains term ϕ_St, finally, depends on usage patterns, schedules, occupancy, and other variables—many of which are rarely available in real building datasets. Moreover, internal heat generation plays a subordinate role in numerous case studies, such as the residential sector, or can be treated as steady in others.

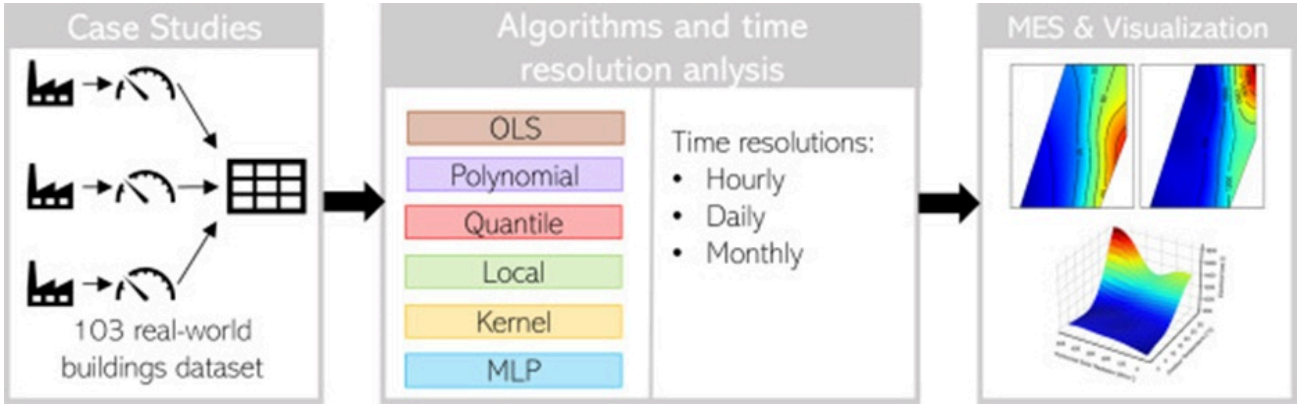

Figure 1. Schematic of the methodology employed for the comparative benchmarking of regression algorithms.

Given these considerations, this paper advocates examining MES that take outdoor temperature and solar radiation as inputs. This is expected to enrich the study of building thermal dynamics by introducing the two most influential variables that affect the thermal balance. Finally, thermal demand is treated as the regression model's output variable.

Benchmark of multivariable energy signature models

To ensure a balanced, wide-ranging comparison between classical and ML-based regression paradigms, six SoA regression configurations are evaluated. Each is trained on 103 case studies, corresponding to 103 individual buildings, across varying temporal granularities. Informed by the SoA contributions, the following algorithms—whose originating references appear in parentheses—are shortlisted: (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]. Beyond raw accuracy, the models must also satisfy the criteria of scalability and interpretability [6]. Accuracy and scalability are measured using the coefficient of determination (R^2), the Mean Absolute Percentage Error (MAPE), the Root Mean Squared Error (RMSE), and the computational time, all tracked across different time granularities. Furthermore, the dataset is divided into training and testing partitions via a time-series split, adopting proportions of and,

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

respectively. The coefficient of determination is computed independently on the two partitions to assess model robustness and generalization. Finally, the graphical aids help interpret the regression models.

Ordinary least squares linear regression

Multivariable Linear regression [28] assumes that the forecast quantities are derived as a linear weighted aggregate of the explanatory variables. For our case study, this translates to:

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

Where β_0 is the intercept term, β_Temp captures the temperature coefficient, β_Sol is the coefficient tied to solar radiation, and G denotes horizontal total radiation. The OLS method is the most widely used technique for fitting linear regression models. In this setting, the regression procedure aims to minimize the residual sum of squares between the observed values and the model outputs. Several assumptions—linearity, homoscedasticity, and the absence of multicollinearity—should be satisfied when employing a linear regression framework. In the literature, linear regression models have routinely been deployed without rigorous verification of these conditions. Regarding multicollinearity, it is worth noting that the input variables chosen for our multivariable regression task are expected to be strongly interdependent; in other words, they are collinear. Despite the violation of assumptions that ought to accompany the use of this model, this paper retains linear regression as a baseline benchmark to illuminate and problematize several related concerns. For brevity, we shall subsequently refer to this approach as Linear Regression.

Quantile linear regression

A different path to Linear Regression is to adopt the Quantile Regression method [29]. Here, the algorithm takes the conditional median as the reference output value rather than the conditional mean deployed by OLS regression. Quantile Linear Regression offers greater resistance to outliers than OLS. In the current investigation, this algorithm is included specifically to examine whether it can provide more robust and generalizable regression formulations than the more commonplace Ordinary Least Squares approach.

Polynomial regression

Multivariable Polynomial Regression is a regression framework in which the explanatory variable set includes both the original input features and higher-order terms derived from them. In this sense, it broadens the multivariable regression setup by admitting supplementary input terms. This approach is realized in the present analysis by incorporating terms up to the second degree. Under this scheme, the regression model can be expressed as:

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

Where β_{T^2}, β_{Sol^2}, β_{T-Sol} and signify the regression coefficients associated with the second-degree terms. This formulation reframes the regression problem as a standard linear regression task.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

Local polynomial regression

Local Polynomial Regression [30] qualifies as a classical method insofar as it builds upon least squares estimation. Yet, it delivers a far more flexible apparatus than ordinary linear regression, since it is based on the intuition of fitting simple polynomial functions to local neighborhoods of data [27]. The result is a regression mapping that, at least in theory, can conform to any shape. For this reason, Local Polynomial Regression is classified as a nonparametric method. This approach carries several strengths. In particular, it obviates the need to predefine the function's qualitative shape and affords greater flexibility than conventional least squares routines. That said, it is computationally costly, and a nonparametric model resists description through a finite-parameter function. This paper operationalizes the method by fitting a second-degree polynomial to local data windows. This choice is consonant with the method's guiding principle of using low-order polynomials for local estimation to guard against overfitting [30]. For brevity, we shall henceforth label this algorithm as Local Regression.

Kernel regression

In parallel with Local Regression, Kernel Regression [31] constitutes a statistical avenue to regression tasks. Moreover, both approaches share local data-adaptivity to furnish a nonparametric regression model. The difference, however, lies in the mechanism: whereas Local Regression fits a prescribed polynomial regression function to a bounded local sub-domain, Kernel Regression uses kernel functions—most typically the Gaussian kernel—to weight dataset elements based on their proximity to a designated target point. Kernel regression can incur even steeper computational complexity than Local Regression. Yet, its extreme pliancy enables it to capture any intricate linkage between the output and the explanatory variables.

Neural network-based regression

NNs form a family of ML models with a biologically inspired blueprint, spanning multiple distinct architectures that can be broadly divided into two subgroups: Feed-Forward Neural Networks (FFNN) and Recurrent Neural Networks. Among the simplest NN designs is the Multi-Layer Perceptron (MLP), first characterized in Rosenblatt [32]. This is a fully connected FFNN made up of an input layer, an output layer, and one or more hidden layers. Each layer houses one or more perceptrons—elementary units executing the following calculation:

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

Where x stands for the input vector, w is the trainable weight vector, b is a trainable bias vector, and χ denotes the activation function. By supplying the input layer with the explanatory variables and each successive layer with the output of its predecessor, the network eventually returns the dependent variable. Provided the network possesses multiple hidden layers, each containing several perceptrons, it may be regarded as a nonparametric regression model. To achieve the objectives of our study, we used a 2-hidden-layer neural network NN). The training regime was set by fixing the batch size to one month of measurements. The remaining NN hyperparameters were determined through an automated procedure.

Time resolution analysis

Scholars have proposed differing stances on the optimal temporal granularity for ES modeling. For example, hourly resolution appears in Rouchier [10], Nageler *et al*. [16], and Westermann *et al.* [19]. In contrast, daily mean values were chosen in Rasmussen *et al*. [25]. A limited set of authors also considered monthly averaged observations [7]. The effect of time resolution on building energy models is twofold. On one front, thermal demand and meteorological sequences tend

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

to exhibit strong serial correlation, a phenomenon linked to the heat-storage characteristics of the building mass. The second front concerns periodic patterns embedded in the time series, which, for buildings, are largely governed by usage schedules. Fu *et al*. [7] suggest that using coarser-grained data, such as daily aggregates, is a practical way to address this. Similar considerations are raised by other investigators, including Rasmussen *et al.* [25], who maintain that daily resolution allows thermal inertia effects to be safely sidelined. Yet, degrading the temporal detail incurs a trade-off. In real-world deployments, excessively long monitoring periods would be required to amass a dataset large enough to fit most regression algorithms. Moreover, energy models trained on coarse data would prove ill-suited to certain applications, such as Operations and Maintenance or anomaly flagging. Given the range of options for ES temporal resolution in the literature, the present investigation encompasses hourly, daily, and monthly granularities. The three temporal scales are used to independently train all six regression models, which are then tested on 103 real-world building records. The outcomes are subsequently evaluated and discussed with respect to accuracy, scalability, and generalization power.

Employment of energy signatures as inspection tools

Deploying appropriate visualization tools can further enhance the diagnostic value and qualitative insight into the thermal behavior of the studied buildings. With this objective, three visualization schemes for MES are examined. The first approach places the most influential input variable on the x-axis and the response variable on the y-axis. A reference trace then captures the regression model in an effectively univariate fashion, while the influence of one or more additional variables registers as point scatter around this trace. This visualization tool can, in principle, accommodate models with any number of input variables, providing a quick and intuitive depiction of the weight assigned to a single principal explanatory variable [25]. By contrast, when the scope is confined to just two input variables, 3-D style MES plots and heat maps become viable. In this latter mode, a far more faithful rendering of how the second explanatory variable shapes the final output is achieved.

Case Study and Experimental Set Up

In addition to the methodological sequence detailed in earlier sections—which delineates the overarching framework intended for application to any building study—several specifics tied to the concrete application case are set out below. Furthermore, to promote openness and the reproducibility of the findings, we briefly discuss the experimental setup and software.

Case study

The proposed methodology was field-tested on a dataset covering 103 buildings. The dataset holds two years of aggregated hourly electrical load readings for the buildings. These readings consequently bundle all electrical end uses; no disaggregation by individual load component or final use is available. The dataset further includes hourly readings of outdoor temperature and Horizontal Solar Radiation, given in °C and W/m^2, respectively.

These measurements were sourced from 103 Data Centers (DCs) belonging to a major Italian TLC service provider. TLC infrastructure and the buildings that house them have sharply expanded their footprint in final energy demand, reaching a 2.8%–3.8% slice of total European electricity consumption, corresponding to 70–95 TWh of electrical energy in 2022 [33]. DCs are industrial facilities characterized by high internal loads, resulting in dense internal heat generation. The buildings are scattered across different Italian regions, thereby sampling a wide climatic spectrum. In fact, the building portfolio under examination covers all seven climatic zones identified by the Climatic Severity Index. Most of the studied buildings are characterized by infrequent human presence—making the occupancy contribution to final electrical draw negligible—and by round-the-clock operation. In such settings, operational schedules are unlikely to figure prominently in building energy signatures, thereby dampening the detrimental effect of periodicities when hourly-resolution series are deployed

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

[34]. Nonetheless, the dataset also includes DC facilities with office areas. In these cases, the time series will display sharper cyclic patterns, and turning to coarser-resolution data is expected to produce more trustworthy outputs. In both scenarios, the two overriding factors in the buildings' final energy draw are the TLC equipment and the cooling plant used to prevent equipment overheating. The energy balance [34] of DCs is expressed as:

$$P_{TOT} = P_{TLC} + P_{DISS} + P_{CLC} + P_{Aux} \quad (8)$$

Where P_{TOT} designates the total electrical load, P_{TLC} corresponds to the TLC equipment draw, P_{CLC} is the conditioning system load, P_{Aux} bundles lighting and auxiliary consumption, and T_{ext} accounts for conversion losses.

To cast an energy model in the form of an ES, the target variable—here, the electrical consumption—must be rendered as a function of the climatic variables T_{ext} and G. Observing that the terms P_{TLC}, P_{Aux}, and P_{DISS} are not modulated by weather variables, it follows that:

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

It is important to underline that the conditioning load P_{CLC} represents the sole load fraction responsive to weather conditions. P_{TLC} and P_{DISS} can typically be treated as steady over time in the DC domain and will therefore only imprint the regression models through a constant intercept in the ES formulations. Finally, P_{Aux}, which covers auxiliary systems, lighting and other marginal loads, varies temporally and may be especially shaped by building schedules, yet can be regarded as weather-independent. This state of affairs can readily induce a non-causal association if hourly resolution is adopted. For instance, were

P_{Aux} to exhibit elevated values during the daytime, this load fraction would spuriously covary with temperature and radiation, which also tend to peak during the day. These observations bear out what was anticipated namely, that aggregated consumption figures P_{TLC} and P_{DISS} can be soundly leveraged for ES purposes when daily or monthly resolution data are used. In contrast, hourly values require careful handling, with due regard for the complications posed by spurious correlation and thermal inertia effects.

Experimental setup and software details

The full suite of algorithms was developed and fitted within a Python 3.9 runtime. All experimental runs were performed on a laptop equipped with an 11th Gen Intel Core i7-1165G7 CPU and 16 GB of system memory.

Linear and polynomial regression specifications were implemented using the Scikit-Learn package [35]. Quantile regression made use of the library's QuantileRegressor class, with highs serving as the numerical solver. For Local Regression, the localreg package [36] furnished the implementation; the Epanechnikov kernel was paired with a domain fraction parameter of 0.66. A systematic grid search was performed to determine the hyperparameter configuration—covering the polynomial degree, kernel function, and associated bandwidth. Kernel Regression leveraged the statsmodels toolkit [37], where hyperparameter selection was guided by the library's built-in cross-validated least squares optimization. The neural network architecture was specified using Scikit-Learn's MLPRegressor class and its model selection module, enabling automated grid-based exploration of the hyperparameter space. The search spanned neuron counts, activation function types, learning rate values, and the decision between constant and adaptive rate scheduling.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

Results and Discussion

The framework elaborated was tested on the case study described before. As previously noted, electrical load streams for 103 buildings were analyzed alongside their corresponding historical weather records. The unprocessed data passed through an initial coarse filtering stage to weed out probable sensor errors. Since the readings originated from various meters with non-uniform timestamping, the first operation resampled everything to hourly averages, yielding a single dataset aligned by date-time index. A permitted range was then stipulated for the predictors and the response alike. Solar radiation entries were bounded between zero and 1376 W/m², electrical load values were required to be strictly positive, and NaN as well as zero records were eliminated. Any row where at least one variable fell outside its designated range was dropped. After this cleaning phase, the load data underwent transformation and anonymization to comply with the data provider's confidentiality and privacy stipulations. From this point forward, the load figures are treated as anonymized readings, expressed in an arbitrary unit of measurement; accordingly, the load axes in all plots omit unit labels.

A second resampling pass generated the daily-averaged and monthly-averaged versions of the dataset. The records for each building—each constituting an independent case study—were subsequently separated into feature matrices and target vectors, and then split into training and holdout folds. Weather attributes formed the predictor set, while the electrical load served as the response. The holdout fold comprised a randomly chosen 33% fraction of each building's observations.

Benchmark of regression models

Each of the six model classes was trained on the three datasets with different resolutions, producing a total of 1854 trained instances. A condensed view of the outcomes across model types is presented in **Table 1**, which reports the mean coefficient of determination (R^2) computed across the 103 case studies. The leftmost column provides the figures for the baseline univariate specification, constructed according to the procedure outlined in Eiraudo *et al*. [38]. Boldface entries indicate the top score attained by a given model form on the training and testing splits. One immediate observation is that the monthly sampling frequency yielded the strongest training-set fit across all model varieties. At the same time, these models clearly suffer from brittleness and limited generalization capacity. R^2 values on the test partition were uniformly lower than their training counterparts across all model types. The deficit in generalization is thrown into sharper relief for the more expressive algorithms—Kernel, Local Quadratic, and NN-based regression stand out in this regard.

In contrast, the benchmark univariate model proved the steadiest performer at the monthly scale, with a mean R^2 of 0.74. The weak out-of-sample performance associated with this temporal granularity can, at least in part, be chalked up to the scarcity of observations (24 data points, reflecting the 2-year measurement window). This constraint would carry over to many real-world deployments, where the timeline for building an adequate sample size would be prohibitively long.

Table 1. Mean coefficient of determination across the 103 case studies. Bold numerals pinpoint the best outcome across the three time granularities; cell shading marks the strongest performer on the training fold, while a green-filled cell identifies the top algorithm on the test fold.

	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

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

When shifting to hourly-resolution data, regression accuracy was comparatively weak, as indicated by the R^2, MAPE, and RMSE values in **Tables 1-3**. Unlike the monthly-resolution picture, the models here displayed healthy generalization, with test-set R^2 closely mirroring training-set performance. The richer model classes claimed the highest scores. Kernel regression held a razor-thin advantage over the NN-based specification, posting a test-set R^2 of 0.722 versus 0.721 and a MAPE of 5.89% against 5.92%. Still, overall performance at this frequency tended to trail both the monthly-resolution figures and, without exception, the daily-resolution results. Daily averages, in fact, yielded the peak R^2 reading for every model type. For Kernel, Local, and NN-based regression, daily data also yielded the lowest MAPE and RMSE. Train and test values at this granularity remained tightly coupled. Except for the two linear formulations (OLS and Quantile), all models eclipsed the univariate baseline. The NN-based variant held a slender edge over Kernel regression, attaining a mean R^2 of 0.804. A richer statistical portrait of model behavior is presented in **Figure 2**. It is worth underscoring that Kernel and NN-based regression achieved R^2 values of 0.6 on 100 and 101 buildings, respectively, across the full set of 103 buildings, and exceeded 0.8 on 58 and 59 buildings, respectively. The advantage of the MLP and Kernel-based approaches relative to the remaining contenders is supported by the 95% bootstrap confidence intervals tabulated in **Table 4** for the daily resolution setting.

Table 2. MAPE across the 103 case studies. Bold text and cell shading follow the same logic 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 across the 103 case studies. Bold text and cell shading follow the same logic 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

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

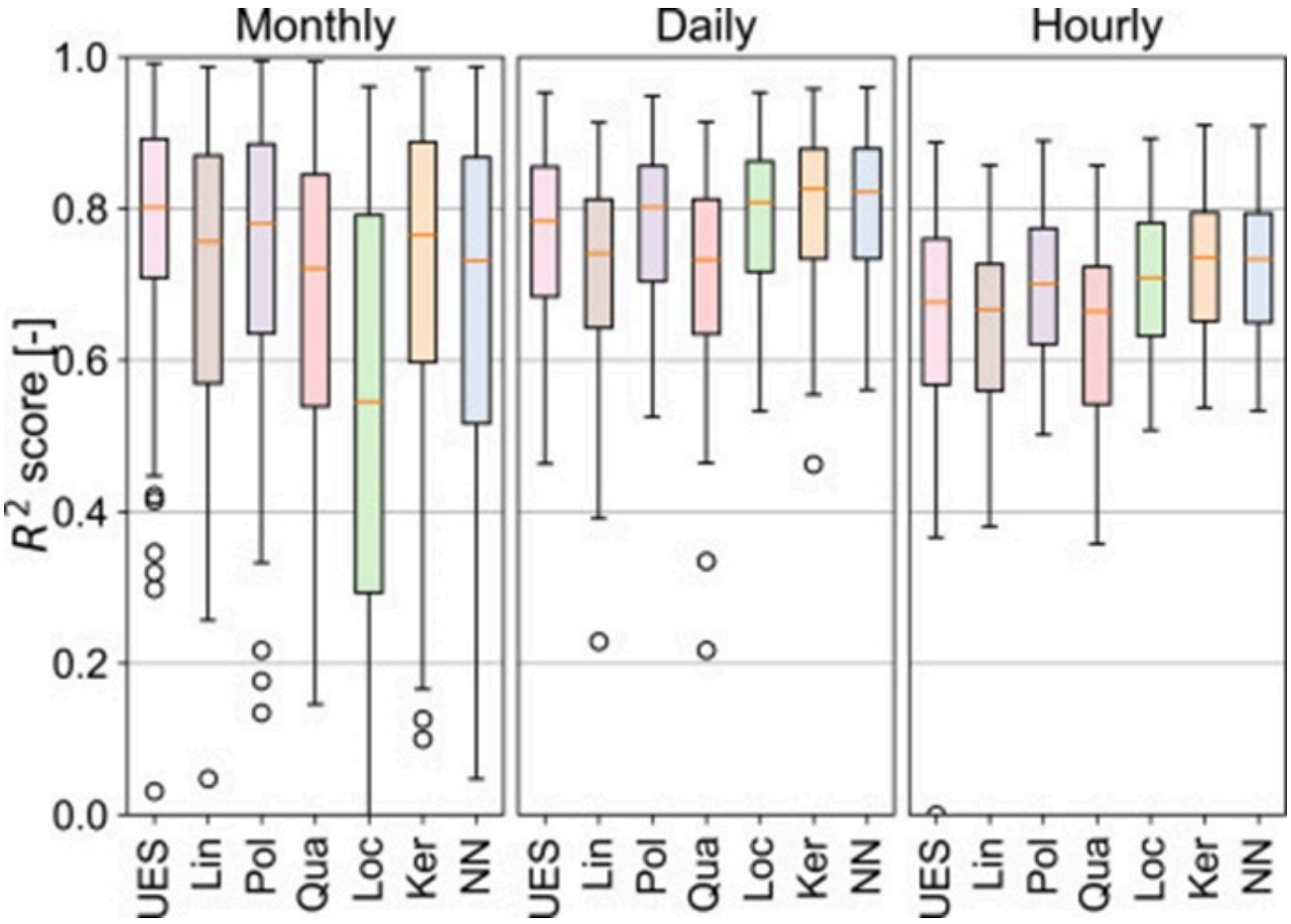

Figure 2. Statistical summary of the coefficient of determination R² achieved by the regression algorithms under study over the 103 real-world case studies.

Table 4. Bootstrap-derived 95% confidence intervals for the performance metrics (R², MAPE, and RMSE) and regression algorithms, computed from the daily-resolution results.

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

Discussion on model performance and computational time

What follows is an examination of the findings yielded by the different model classes and temporal resolutions, set against the backdrop of the literature gaps catalogued in Fu *et al*. [7]. The discussion weighs the relative merits and drawbacks of the conventional approaches [28, 29] alongside the ML-driven alternatives [30-32] , and homes in on the sampling frequency best suited to ES work. The fitted models are then unpacked to provide further insight into the energy profiles of the buildings under study.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

A battery of residual diagnostics—spanning normality, independence, and homoscedasticity—was first conducted to test whether the foundational regression assumptions were met. Residual independence held across every algorithm and the overwhelming majority of case studies. By contrast, only the local, kernel, and NN-based schemes delivered residuals that simultaneously satisfied normality and constant variance for most of the buildings examined.

The source of the performance edge enjoyed by NN and Kernel regression becomes more intelligible upon examining the fitted models shown in **Figure 3**. These panels, taken from a representative building at daily resolution, map the primary explanatory variable onto the x-axis and the response variable onto the y-axis. The information extracted from the secondary predictor—solar radiation, in this instance—results in a vertical shift of the predicted values (the coloured markers) relative to the reference univariate regression curve. Two observations leap out from the figure. The first is that the linear variants (marked Lin and Qua) fit the empirical distribution poorly, particularly at the low and high extremes of the temperature range. This suggests that the linearity assumption is tenable only over a restricted segment of the data, with sizeable departures elsewhere. The polynomial specification (Pol) labours under a comparable limitation, as does the local regression method (Loc), albeit to a lesser extent. The Kernel and NN-based versions (NN and Ker), on the other hand, closely conform to the data, with their gradients shifting across three distinct temperature zones. At the cool end, electrical load plateaus at a constant value, independent of outdoor temperature—the signature of an unconditioned regime. Across the intermediate zone, load increases roughly linearly with temperature. These two zones mirror the classic building thermal response that informs the benchmark univariate model and is further discussed in Eiraudo *et al*. [38].

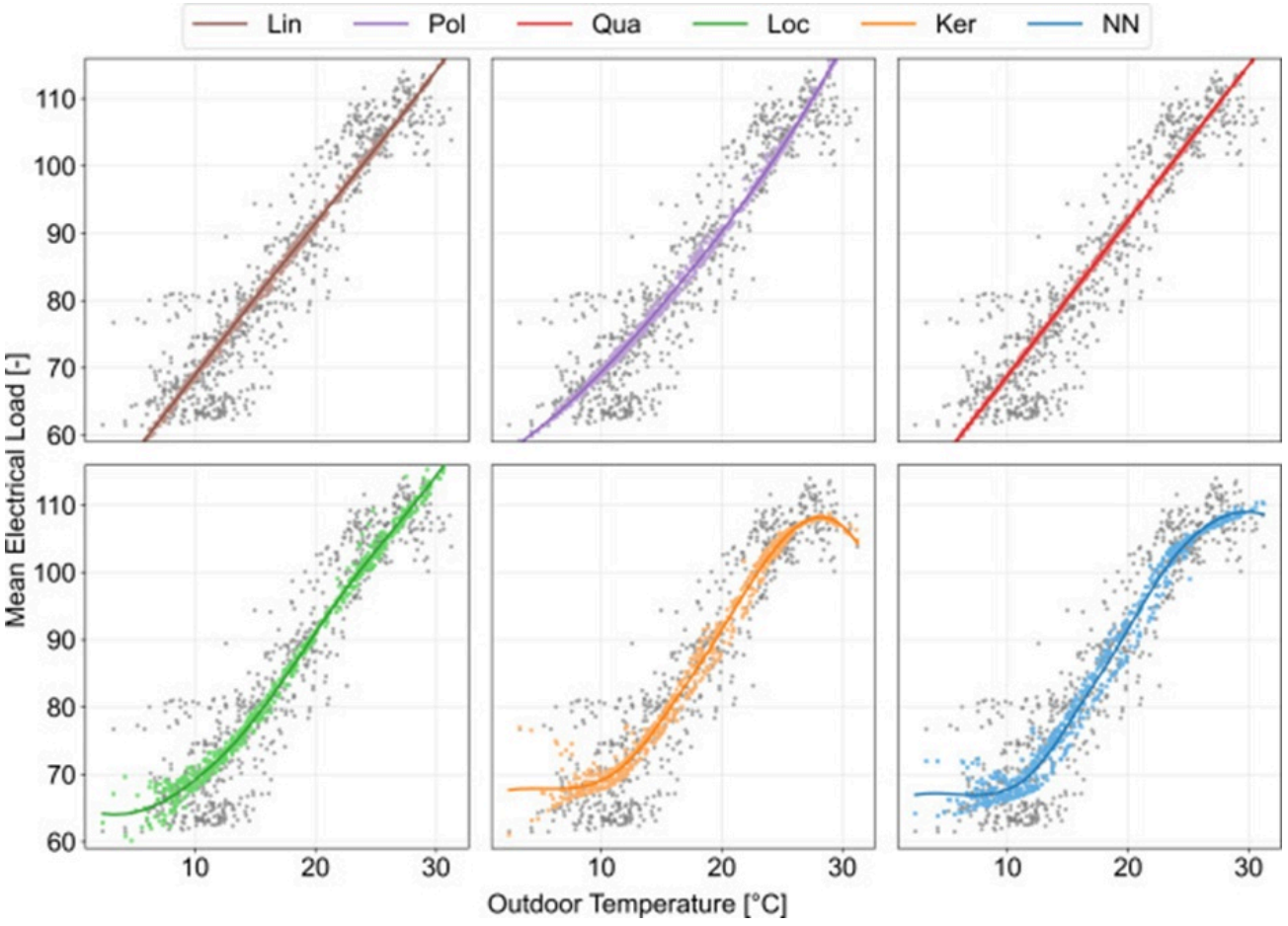

Figure 3. Reference regression curves and predictions produced by the different regression algorithms for a single case study (building G). Predictions are rendered as coloured markers, while the reference curves trace the expected mean output at a given outdoor temperature. Grey markers denote actual measurements.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

At high temperatures, a break from linearity becomes apparent. The second instructive feature in **Figure 3** is the degree of scatter surrounding the reference regression curve, which, as noted earlier, reflects the model's success in harnessing the additional explanatory variable. The OLS, quantile, and polynomial models produce the tightest clustering of predictions—a tell-tale sign that these models are unable to register and represent solar radiation's bearing on electrical load. The remaining models, meanwhile, all generate appreciable spread. Importantly, the vertical displacement from the curve is not uniform across the domain, suggesting that solar radiation's modelled influence varies with temperature band, a matter returned to below.

The question of scalability is then taken up through the lens of training runtimes. **Table 5** collates the mean training duration per case study for each regression approach, and **Figure 4** plots these same quantities on a logarithmic scale for easier visual digestion. The plot makes plain that parametric models (OLS, quantile, polynomial) operate orders of magnitude faster than their nonparametric cousins. That said, quantile regression incurs a steep increase in training costs as the dataset grows. The Local Quadratic model follows a broadly similar trajectory. The NN regressor is the slowest option when fed sparsely sampled data, with mean training times exceeding 1 second per case study. Yet its scaling profile is favourable: training time grows by only about 4 as resolution is sharpened 720-fold, from monthly to hourly records. Kernel regression tells a starkly different story—it proves fundamentally unscalable, with training complexity exploding exponentially across granularities, pushing mean training times to roughly 8.3 seconds for daily data and around 1100 seconds per case study for hourly data. Balancing predictive accuracy, generalization capacity, and scaling behaviour, the NN-based model trained on hourly observations stands out as the most judicious regression choice. This model reduced ES accuracy to a MAPE of 5.2%, an improvement over the benchmark's 5.7%.

Table 5. Mean computational durations, in seconds, for training a single regression algorithm on one case study from the various MES models, broken down by time resolution.

	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	$1.6 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$	$2.5 \cdot 10^{0}$	$2.3 \cdot 10^{1}$	$1.1 \cdot 10^{3}$	$6.8 \cdot 10^{0}$

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

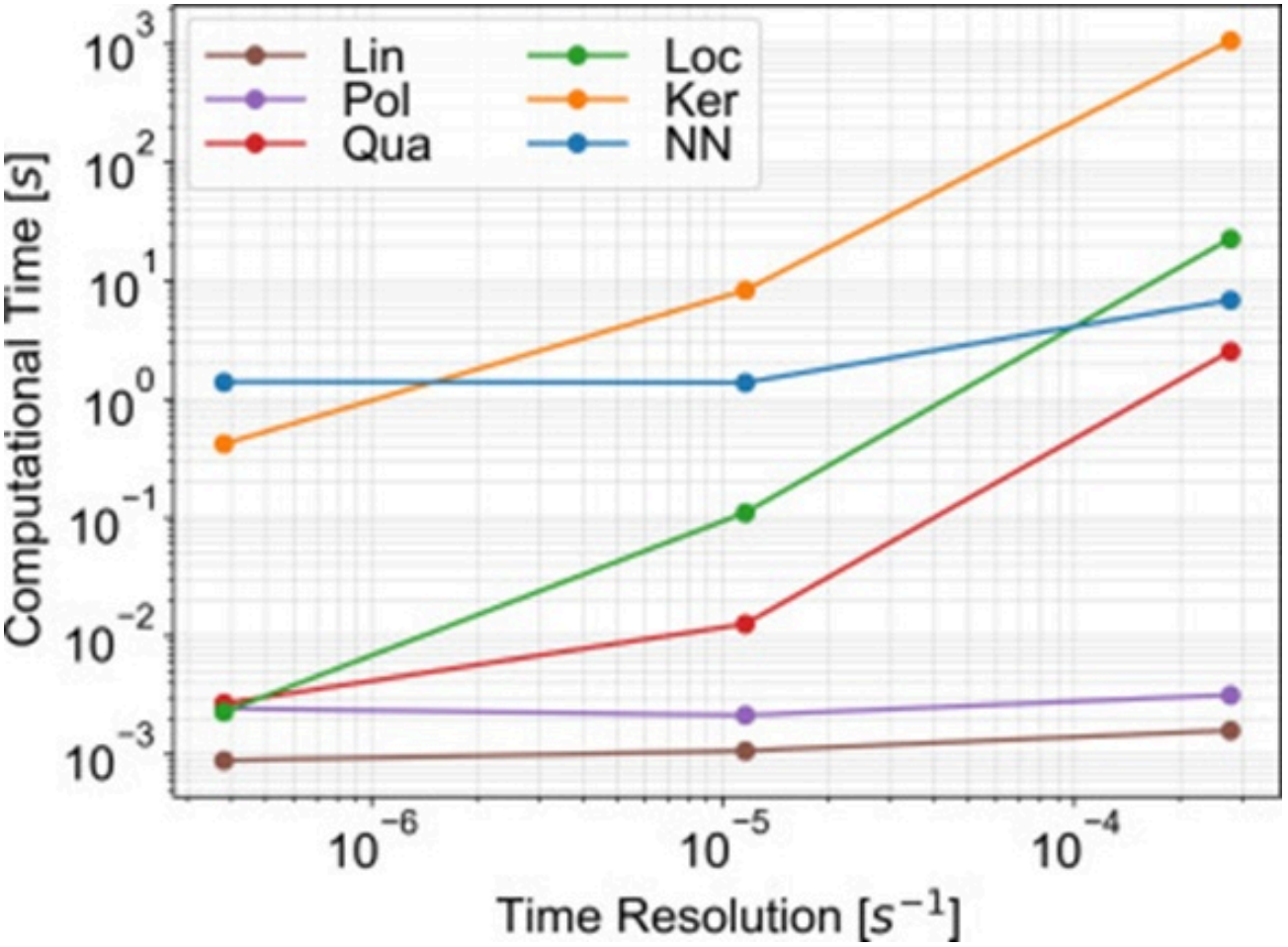

Figure 4. Logarithmic plot of the mean computational duration required to train the different regression algorithms on a single case study.

Interpretation of the multivariate energy signatures

In the final analysis, MES can function as a diagnostic instrument for decoding building thermal behaviour. To more fully apprehend the joint influence of both explanatory variables, the fitted regression surfaces need to render the predictors and the response together in explicit form—a task well served by a 3-D surface plot **(Figure 5)** or, in a more compressed format, a heatmap **(Figure 6)**. The former depicts the nonparametric regression surface learned for one of the buildings in the case-study pool. The MES depicted in **Figure 5** highlights the archetypal thermal response identified in the previous section. Most notably, it reveals a band of flat, subdued consumption at low outdoor temperatures—the unconditioned zone, within which the cooling machinery need not activate to hold DC equipment within thermal limits. Across this band, neither outdoor temperature nor solar radiation exerts any meaningful pull on electrical load. Next, an intermediate, conditioned zone is captured, marked by a quasi-linear increase in consumption as outdoor temperatures rise. In this region, solar radiation ordinarily leaves only a faint imprint on electrical load. Finally, a high-temperature band comes into view. Across the 103 buildings studied, solar radiation was observed to make the greatest contribution to electrical load in this band. However, the form and magnitude of that contribution fluctuate markedly from building to building.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

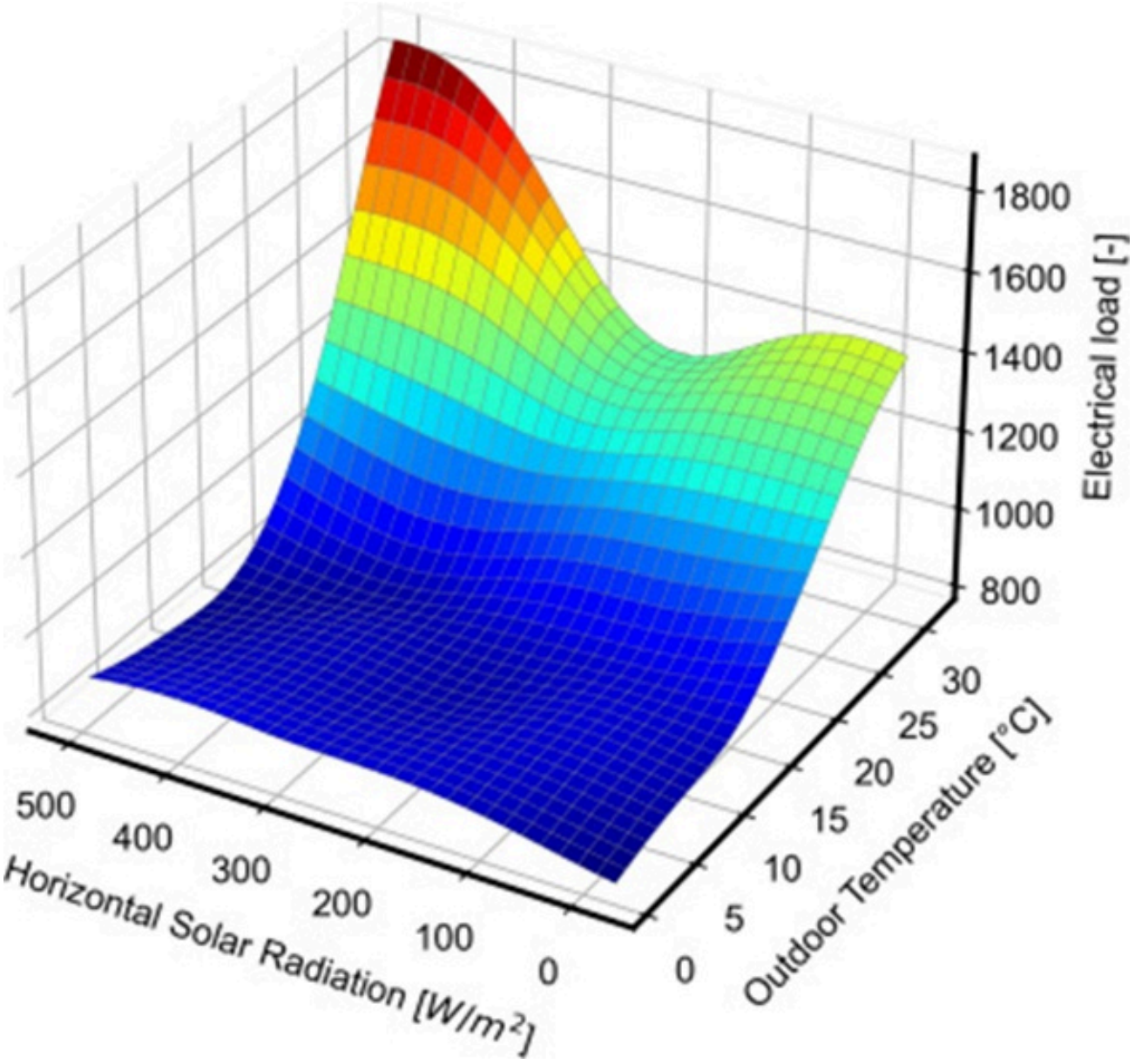

Figure 5. 3-D style MES for building B.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

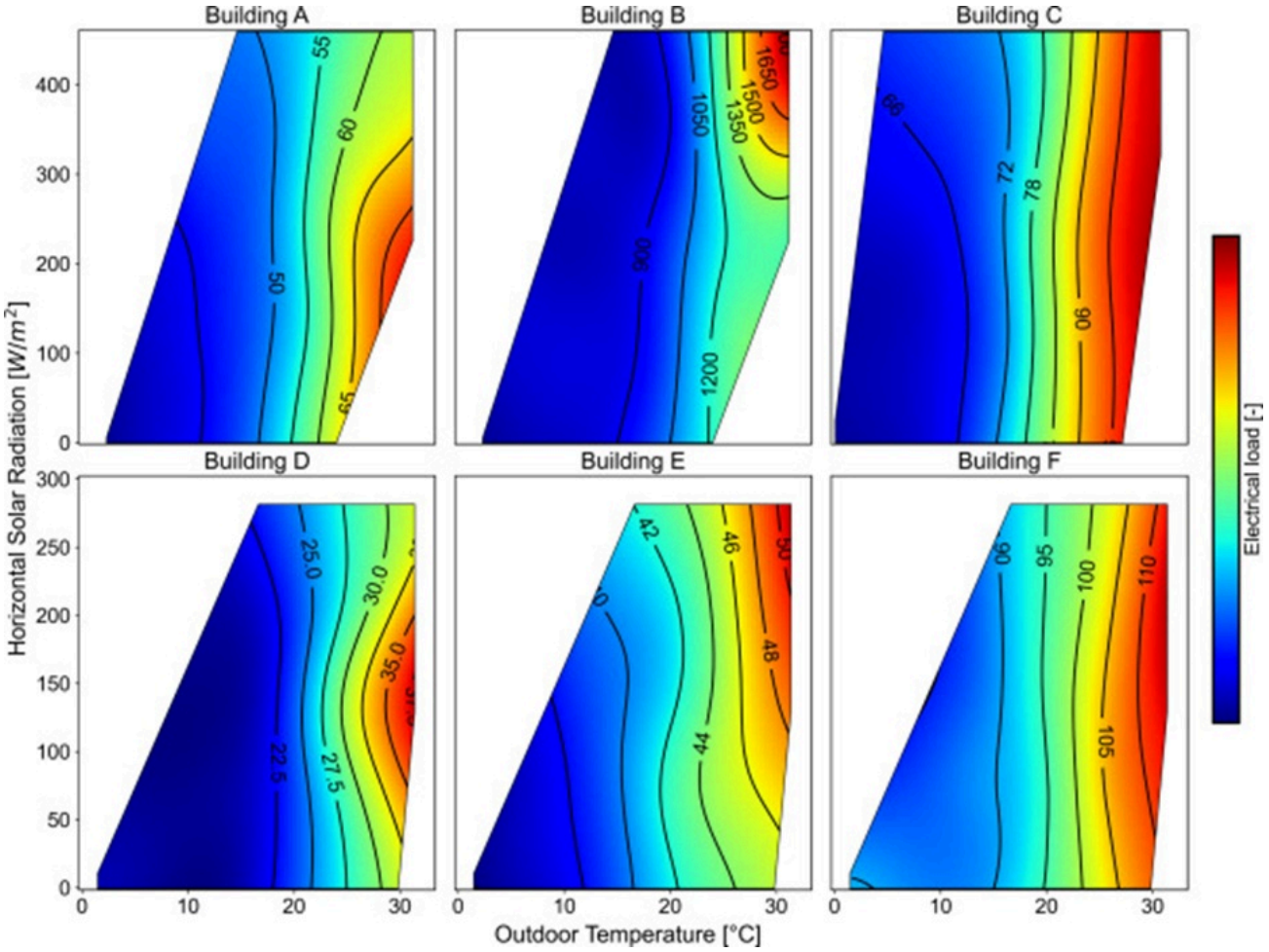

Figure 6. MES for six buildings, produced by the NN-based regression model using daily input data. The MES surfaces are cropped to an extended envelope computed from each building's local weather records.

In the specific building shown in **Figure 5**, a dramatic spike in electrical load materialises at the upper end of the temperature range as solar radiation reaches peak values. At the same time, a local dip is observed at moderate radiation levels, while the regression surface appears to project a local peak under high temperatures and minimal solar radiation. The convoluted signature of solar radiation on electrical load arises partly from its simultaneous grip on multiple load fractions—chief among them the cooling plant, with a secondary pathway through lighting, as treated in Eiraudo *et al*. [39]. Even so, a glance at the underlying input data suggests this particular region may lie beyond the empirical support of the observations. As underscored earlier, the MES surfaces are ultimately constructed from daily averaged quantities. For the location tied to building B, ambient temperatures above 30 °C can only materialise when mean solar radiation exceeds roughly 200 W/m². Each building's data envelope must therefore bound any faithful reading of MES. This discipline is enforced in **Figure 6**, where each surface is confined to the convex hull of the training data, augmented by a small margin around each predictor, with surface regions that spill beyond this expanded envelope trimmed off. The figure provides a comparative panorama across six buildings, revealing the heterogeneous imprint of solar radiation. For buildings C and F, shifts in solar radiation barely register on electrical demand. In the cases of buildings B (discussed above) and E, however, solar radiation supplies a sizeable additional load contribution within the high-temperature band. Two more idiosyncratic motifs stand out in buildings A and D: here, heightened solar radiation coincides with a drop in electrical load in the high-temperature region. This seemingly anomalous signature may hint at the presence of an absorption chilling plant or another peculiar building attribute. The thermal behaviour of building D presents an even more vexing interpretive challenge, as its global load peak occurs not at the highest temperature but at a mid-range value of solar radiation. Further probing of these buildings would be prudent to disentangle their thermal signatures and, as

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

appropriate, flag anomalies or inefficiency culprits. The majority of the remaining case studies broadly conformed to the canonical patterns illustrated by buildings B, C, E, and F.

Conclusion

A collection of six regression techniques from the published literature, spanning both conventional approaches and those rooted in Machine Learning, was applied to characterize building thermal dynamics through a Multivariable Energy Signature lens. The regression objective was to establish a link between electrical load and climatic drivers. In a break from the habitual practice of univariate Energy Signature formulations, solar radiation was admitted as a second explanatory variable. Beyond this, three levels of temporal aggregation—hourly, daily, and monthly—were put to the test. This benchmarking exercise drew upon a substantial real-world dataset comprising 2 years of time-series data from 103 industrial sites within the TLC sector. The headline takeaways are as follows:

• Daily-averaged records emerged as the most effective temporal granularity across all regression schemes, consistently outperforming monthly or hourly observations.

• Both Kernel regression and Neural Networks outperformed Ordinary Least Squares, quantile regression, polynomial regression, and local quadratic regression on measures of predictive fidelity.

• Neural Networks exhibited attractive scaling properties as the volume of training data expanded. Kernel regression, in sharp contrast, experienced an exponential increase in computational demand as the number of data points multiplied. On these grounds, Neural Networks were ultimately singled out as the most apt algorithmic choice for Multivariable Energy Signatures.

• The addition of solar radiation as a supplementary predictor yielded a meaningful uplift in model performance; a mean MAPE of 5.2% was achieved, compared with 5.7% for the benchmark univariate Energy Signature model.

• Lastly, the Neural Network-driven Multivariable Energy Signature achieved robust predictive performance, with a mean coefficient of determination of 0.804 across the 103 buildings examined.

Multivariable Energy Signatures were ultimately operationalized as diagnostic tools to probe the thermal fingerprint of buildings. Their deployment in energy auditing may be particularly pertinent in sectors where cooling loads dominate final building energy use.

Looking ahead, future research should expand the empirical scope to include case studies from other building typologies —the residential sector being a prime candidate—to validate the conclusions drawn here and broaden the evidence base for generalizability. Furthermore, subsequent studies ought to pursue a more granular formulation of the thermal load–weather linkage, guided by the solar-air temperature construct. This would, in due course, pave the way for treating wind as yet another independent variable bearing on the thermal equilibrium of buildings and, hence, on their energy draw. A deeper dive into the contributions of weather variables, including a formal feature importance assessment, would also represent a valuable direction for further work.

Acronyms

CP Change Point

DC Data Center

ES Energy Signature

FFNN Feed-Forward Neural Network

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

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, [$\frac{\mathrm{W}}{\mathrm{m}^{2*}\mathrm{K}}$]

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

I Total solar irradiance, [W/m^2]

$\mathrm{P}_{\mathrm{Aux}}$ Electrical load from lights and auxiliary systems, [kW]

$\mathrm{P}_{\mathrm{CLC}}$ Conditioning system electrical demand, [kW]

$\mathrm{P}_{\mathrm{DISS}}$ Electrical power conversion losses, [kW]

$\mathrm{P}_{\mathrm{T\ LC}}$ Electrical load of the telecommunication equipment, [kW]

$\mathrm{P}_{\mathrm{TOT}}$ Total electrical load, [kW]

t Time

$\mathrm{T}_{\mathrm{ext}}$ 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

Acknowledgements

Simone Eiraudo acknowledges support from TIM S.p.A. through the Ph.D. scholarship. The work of D. S. Schiera was supported by the “Network 4 Energy Sustainable Transition–NEST,” through the Project code PE0000021, Concession Decree No. 1561 adopted by the Ministero dell’Università e della Ricerca (MUR), under Grant CUP E13C22001890001; and in part by the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3–Call for tender No. 341 of MUR; funded by the European Union–NextGenerationEU .

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 25 Nov 2021 Revised: 30 Jan 2022 Accepted: 16 Feb 2022

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

Published online: 15 November 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. European Commission. Energy efficiency directive. 2024 [cited 2024 Mar 29]. Available from: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-directive_en

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. https://doi.org/10.59117/20.500.11822/45095.

3. IEA. World Energy Outlook 2024. Int Energy Agency. 2024. Available from: 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: IEA; 2023. Available from: 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. Available from: 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 JC, 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, et al. 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, et al. 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, et al. 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): Los Alamos National Lab. (LANL); 2017.

28. Yan X, Su X. Linear regression analysis: theory and computing. Singapore: World Scientific; 2009.

29. Koenker R. Quantile regression. Cambridge: 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.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

33. Kamiya G, Bertoldi P. 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, et al. 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, et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12:2825-30.

36. Marholm S. localreg 0.5.0 [Internet]. Available from: 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, et al. 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.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press