Bus Inf Syst Eng 63(3):223–242 (2021)
https://doi.org/10.1007/s12599-021-00691-2

RESEARCH PAPER

Benchmarking Energy Quantification Methods to Predict Heating Energy Performance of Residential Buildings in Germany

Simon Wenninger · Christian Wiethe

Received: 1 July 2020 / Accepted: 25 January 2021 / Published online: 10 March 2021

Abstract To achieve ambitious climate goals, it is necessary to increase the rate of purposeful retrofit measures in the building sector. As a result, Energy Performance Certificates have been designed as important evaluation and rating criterion to increase the retrofit rate in the EU and Germany. Yet, today's most frequently used and legally required methods to quantify building energy performance show low prediction accuracy, as recent research reveals. To enhance prediction accuracy, the research community introduced data-driven methods which obtained promising results. However, there are no insights in how far Energy Quantification Methods are particularly suited for energy performance prediction. In this research article the data-driven methods Artificial Neural Network, D-vine copula quantile regression, Extreme Gradient Boosting, Random Forest, and Support Vector Regression are compared with and validated by real-world Energy Performance Certificates of German residential buildings issued by qualified auditors using the engineering method required by law. The results, tested for robustness and systematic bias, show that all data-driven methods exceed the engineering method by almost 50% in terms of prediction accuracy. In contrast to existing literature favoring Artificial Neural Networks and Support Vector Regression, all tested methods show similar prediction accuracy with marginal advantages for Extreme Gradient Boosting and Support Vector Regression in terms of prediction accuracy. Given the higher prediction accuracy of data-driven methods, it seems appropriate to revise the current legislation prescribing engineering methods. In addition, data-driven methods could support different organizations, e.g., asset management, in decision-making in order to reduce financial risk and to cut expenses.

Keywords Energy informatics · Energy quantification methods · Energy performance certificates · Benchmarking · Data-driven methods · Machine learning algorithms · Building energy · Data analytics

1 Introduction

Human-made climate change is in full swing and revealing first negative effects (Larsen et al. 2020). The United Nations' Paris Agreement declared ambitious climate goals and aims to decrease energy end-use below 1990 levels by 2030 (Boden et al. 2017). However, current efforts are not sufficient to achieve the intended goals and therefore additional steps are necessary (European Environment Agency 2019). One of the largest single energy consuming sectors in Germany are residential single- and two-family buildings, accounting for 11% of the overall final energy consumption, 84% of which relates to heating and hot

Accepted after two revisions by the editors of the special issue.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12599-021-00691-2.

S. Wenninger (✉)
FIM Research Center, University of Applied Sciences Augsburg,
Universitätsstraße 12, 86159 Augsburg, Germany
e-mail: simon.wenninger@fim-rc.de

S. Wenninger · C. Wiethe
Project Group Business & Information Systems Engineering of
the Fraunhofer FIT, Universitätsstraße 12, 86159 Augsburg,
Germany

C. Wiethe
FIM Research Center, University of Augsburg,
Universitätsstraße 12, 86159 Augsburg, Germany

water production, with similar figures for many other countries (Cao et al. 2016; Federal Ministry for Economic Affairs and Energy 2018). Moreover, 64% of the German residential buildings were erected before 1979, which were subject to less strict construction codes than today, thus offering considerable energy savings potential when conducting retrofits (Federal Statistical Office of Germany 2011). Nevertheless, retrofit measures on these buildings are carried out sparsely and the retrofit rate – the percentage of buildings that undergo retrofits in one year – is too low to reach the climate goals (Achtnicht and Madlener 2014).

In this vein, Energy Performance Certificates (EPC) have been designed to support achieving the climate goals in the EU and particularly in Germany (European Parliament and the Council 2002). EPCs are issued by qualified auditors and are intended to increase the retrofit rate by providing general information about buildings, their Final Energy Performance (FEP) – the annual amount of energy required for space and water heating, cooling, and ventilation per square meter effective building area – and possible retrofit measures (Arcipowska et al. 2014). To achieve its full effect, accurate prediction of the FEP is important to decide on purposeful retrofit measures, as uncertainty and incomplete information are substantial investment barriers (Amecke 2012). However, today's most frequently used and by law prescribed Energy Quantification Methods (EQM) are hotly debated in the research community, as they exhibit low prediction accuracy (Hardy and Glew 2019). The prescribed engineering EQM bases on physical laws to calculate thermal dynamics and energy behavior (Zhao and Magoulès 2012) and requires detailed information on building components, gathered by auditors during on-site inspections (Arcipowska et al. 2014). If the input data quality is low, e.g., because the insulation materials are not known and cannot be determined with reasonable effort, the result will also be erroneous.

To enhance the prediction accuracy, data-driven EQMs were introduced in research and obtained promising results in preliminary studies (Sutherland 2020). They learn underlying dependency structures from available data without relying on expert knowledge of building physics or precise information on building components (Amasyali and El-Gohary 2018). This allows data-driven EQMs to potentially overcome the shortcomings of engineering EQMs. However, there is a lack of studies on data-driven EQMs in residential buildings considering heating energy with a focus on long-term (annual) energy prediction, as required for EPCs (Amasyali and El-Gohary 2018). Furthermore, most studies are based on simulated building and energy data, which limits their practical applicability and the validity of the findings (Wei et al. 2018). It is therefore unclear whether data-driven methods can outperform the engineering EQM with respect to annual energy prediction of residential buildings necessary for EPCs, and, if so, which data-driven EQMs are particularly suited. Even though different EQMs have been applied in several case studies (Buratti et al. 2014; Tsanas and Xifara 2012), to the best of our knowledge no benchmarking of different EQMs on the same underlying real-world data has been performed, which is nonetheless essential for full comparability and transparency of the algorithms' performance in practice. Thus, we formulate our guiding research question as follows:

Which of the investigated energy quantification methods yields the highest accuracy for predicting final energy performance of real-world residential single- and two-family buildings in Germany?

In this sense, our goal on the methodological level is not to explain the underlying causality, but to predict energy consumption, allowing us to benchmark prediction accuracy (Shmueli and Koppius 2011). Since the computational performance of data-driven methods generally exceeds that of engineering methods after initial training, we focus on the prediction accuracy, i.e., effectiveness, and not efficiency.

We address the research question by implementing and tuning several machine learning algorithms – Artificial Neural Network (ANN), D-vine copula quantile regression, Extreme Gradient Boosting (XGB), Random Forest (RF), and Support Vector Regression (SVR) – on an extensive first dataset containing 25,000 real-world single and two-family buildings in Germany. We subsequently calculate the output accuracy (predictive power) by predicting the FEP of 345 additional buildings from a second dataset and comparing the prediction with the actual metered energy consumption. As the second dataset was gathered by qualified energy auditors and also encompasses the FEP stated in the EPCs based on the prescribed engineering EQM, we can further compare the data-driven EQMs to the engineering EQM. To ensure robust results and to comply with state-of-the-art machine learning practices, we benchmark the machine learning algorithms against each other in depth based on nested cross-validation on both building datasets, which is not possible for the engineering EQM due to data restrictions. By stratifying the Performance Evaluation Measures (PEM) based on a third dataset which contains information on the German building stock, we ensure representativeness.

Even though the applied solutions and the respective problem in this research are technically known, we argue that we contribute an improvement to existing solutions in terms of Gregor and Hevner (2013) for the following reasons: (1) We are among the first to compare existing solutions (i.e., different EQMs) in terms of solution maturity within a new application domain of the annual FEP prediction for residential buildings, filling the research gap of missing data at the residential building level and the application of data-driven EQMs. (2) Because data-driven EQMs must be designed for specific applications to unleash their full potential (Mosavi et al. 2019), existing knowledge about the performance of data-driven EQMs on non-residential buildings cannot be transferred to the residential building stock directly. Especially in countries like Germany, which have a very high percentage of single- and two-family buildings (Federal Statistical Office of Germany 2011), the improvement of the quantification of the energy efficiency of buildings is relevant to advance towards the set climate goals.

The remainder of this study is structured in seven sections: Sect. 2 summarizes the theoretical background of EPCs, previous research on EQMs, and PEMs to assess the EQMs' prediction accuracy. Section 3 presents the methodology and the study design for the benchmarking process. The datasets and pre-processing procedure are then introduced in Sect. 4. In Sect. 5 we display the model training as well as the model optimization and present the results in Sect. 6. We discuss the results and provide managerial and policy implications as well as limitations and prospects for further research in Sect. 7 before the final Sect. 8 concludes.

2 Problem Context and Theoretical Background

2.1 Energy Performance Certificates

The European parliament and council passed a directive in 2002 that declares the need for EPCs to improve the energy performance of buildings, aiming to inform owners, occupants, and property developers about the energetic building state and related operating costs (European Parliament and the Council 2002). EPCs are issued by qualified auditors and illustrate the energy performance of individual buildings as well as further information like building age, energy source of the heating system, recommendations for energetic retrofit measures, or the building's position in an energy efficiency ranking scheme which allows to compare different buildings (Poel et al. 2007).

Both literature and practice manifold discuss different aspects of EPCs (Li et al. 2019). Next to investigations about to which extent EPCs influence the real estate market as well as the impact and relevance of EPCs on retrofit and purchasing decisions, the energy performance gap is a major challenge of EPCs (Pasichnyi et al. 2019). The energy performance gap describes the phenomenon that the actually metered FEP differs significantly from the predicted FEP, with studies depicting deviations of up to 287% (Calì et al. 2016; Wilde 2014). Many studies are dedicated to the gap's existence, causes, and solutions to minimize it (Burman et al. 2014; Herrando et al. 2016; Menezes et al. 2012). One possible solution to address the energy performance gap are data-driven EQMs instead of engineering EQMs (Foucquier et al. 2013). Another option that can be used to minimize this gap is a demand-consumption comparison, which is regulated in addendum 1 of DIN V 18599 for retrofitting consulting, but not part of official EPCs (Beuth Verlag GmbH 2010). The norm defines key figures and correlations, in order to stepwise approximate the calculated demand to the measured consumption and thus to minimize the performance gap by improving retrofitting decisions (Bigalke and Marcinek 2016).

In Germany, the Energy Saving Ordinance forms the regulatory framework for EPCs with the FEP as target measure (Deutscher Bundestag 2013). EPCs for residential buildings concentrate mostly on space and water heating, as cooling and controlled ventilation systems are not common in Germany (Federal Ministry for Economic Affairs and Energy 2018). Thus, in our research, we focus on the FEP for space and water heating. Broadly speaking, an EPC is issued either by metering (measured EPC) or by calculations (calculated EPC). Measured EPCs reflect the actually metered annual consumption of all energy sources that have contributed to the heating, ventilation, and cooling of a house within the last three consecutive years, thus implicitly including occupant behavior. Calculated EPCs reflect the energy demand and determine the FEP by means of a technical analysis of a multitude of building parameters prescribed by the Energy Saving Ordinance. To collect the required information to carry out calculated EPCs, on-site inspections of qualified auditors are needed. The German engineering norm DIN V 18599 is the standard calculation scheme to determine the FEP of buildings (DIN e.V. and Beuth Verlag 2016). For residential buildings the norm DIN V 4108–6 can also be applied in combination with DIN V 4701–10 or DIN V 4701–12 (Deutscher Bundestag 2013). The current guidelines necessitate calculated EPCs for nearly two-thirds of all residential buildings in Germany. As a large part of these buildings was constructed before the heat insulation ordinance of 1977, thus offering great energy savings potential, we focus on calculated EPCs in the following (Federal Statistical Office of Germany 2011).

For a better understanding of the following sections, we describe necessary calculation rules of EPCs. The FEP is related to the effective building area A_e [m^2], which does not correspond to the more common living space A_l [m^2] (Deutscher Bundestag 2013). The effective building area includes areas that are heated indirectly like corridors or stairways, and thus turns out to be larger than the living space. According to national legislation, the effective building area depends on the heated building volume and the story height, but can also be approximated with the living space and the factor f_c using Eq. (1).[1] The factor f_c is used for approximating the effective area A_e with the living space A_l, which is more commonly available for tenants or homeowners. The conversion factor f_c is 1.35 for buildings which contain no more than two apartment units with a heated basement, and 1.2 for all other buildings (Deutscher Bundestag 2013).

$$A_e = f_c \cdot A_l. \tag{1}$$

To meaningfully compare buildings from different locations, i.e., with different climatic conditions, the FEP is weather-rectified by referring to the climate of the reference location of Potsdam in a test reference year (Deutscher Bundestag 2013). To extract weather effects, the broadly accepted and normative formalized method of climate factors (CF) based on heating degree days (HDD) is established in research (You et al. 2014). A degree day is defined as the difference between an indoor comfort temperature (τ_I) and the average daily outdoor temperature (τ_i). The HDD equal the sum of degree days over a certain period of N days, where τ_i is below the heating limit (τ_L) (e.g., 15 °C in Germany for existing buildings (Olonscheck et al. 2011)), as depicted in Eq. (2) (Baltuttis et al. 2019):

$$HDD(\tau_L, \tau_I) = \sum_{i=1}^{N} 1_{\tau_L \geq \tau_i}(\tau_I - \tau_i). \tag{2}$$

The indicator function $1_{\tau_L \geq \tau_i}$ takes the value 1 if the average outdoor temperature τ_i is below or equal to the heating limit τ_L and is 0 for all other cases. By calculating the HDD for two locations (X, Y) the climate factor CF can be derived according to Eq. 3:

$$CF = \frac{HDD(X)}{HDD(Y)}. \tag{3}$$

Based on the climate factor CF, the measured consumption of location Y can be adjusted to the weather conditions of location X. With the help of the climate factor and the effective building area, we can calculate the FEP of a building from any location the same way it is given in EPCs by rectifying the final energy demand or measured consumption C using Eq. (4). For EPCs the HDDs of location X refer to the climatic conditions of the reference location of Potsdam and the corresponding test reference year (Deutscher Bundestag 2013). This enables us to compare buildings' energy performance independently of their location, size, and weather-related temperature effects.

$$FEP = C \cdot \frac{CF}{A_e}. \tag{4}$$

2.2 Energy Quantification Methods

Quantifying buildings' energy performance is a challenging task with multiple influencing factors like building geometry, occupant behavior, thermal properties, or weather (Wei et al. 2018). Accordingly, the field of EQM research is diverse and methods differ significantly regarding their level of detail and purpose (Wang et al. 2012). Common dimensions to distinguish the scope of EQM studies are building types, prediction time horizon, and the scope of energy performance (Amasyali and El-Gohary 2018). Thereby, most studies currently focus on the prediction of overall energy performance for commercial and/or educational buildings with an hourly time horizon (Wei et al. 2018). In their extensive literature reviews, where they examined collectively over 200 articles, Amasyali and El-Gohary (2018), Bourdeau et al. (2019), and Wei et al. (2018) independently conclude that there is a lack of research for residential buildings and specifically for long-term annual energy prediction. Especially, the combination necessary for EPCs in residential buildings has not been sufficiently analyzed by means of data-driven EQMs. This also holds true for 2019 onwards, as indicated in Table 1. Real-world applications and data are necessary to obtain reliable results, because synthetic data from simulation models use simplifications and required input parameters are often not available (Wei et al. 2018). Nonetheless most studies currently use synthetic data instead. There are many reasons for the lack of large and reliable real-world datasets for residential buildings, as collecting data for residential buildings is a difficult and time-consuming task. The building stock is extremely diverse (Bourdeau et al. 2019), and the data sources are not standardized, which requires extensive questionnaires and tools for data collection. In addition, parameters and terms are often interpreted differently, making it difficult to align datasets (Carpino et al. 2019). With our study we directly address this research gap, focusing on residential buildings, using real-world data, and predicting annual heating energy performance.

In general, EQMs are categorized into engineering methods, data-driven methods, and hybrid methods combining the former (Foucquier et al. 2013). In literature there is no consistent terminology for EQMs. As a generic term, methods or approaches are often used, both for engineering and for data-driven methods (Bourdeau et al. 2019). For data-driven methods, depending on the research domain,

[1] For the sake of completeness, we refer to the Energy Saving Ordinance for further details on the determination of the effective area for calculated EPCs. In our study we make use of the simplification for measured EPCs, as our datasets do not contain any information about the heated building volume and story height necessary for the calculation scheme of calculated EPCs (cf. Sect. 4).

Table 1 Recent studies (2019–2021) of data-driven energy quantification methods and energy prediction (list not conclusive)

Source	Building type	Time horizon	Type of energy performance	Type of datasets
Ciulla and D'Amico (2019)	Non-residential	Annually	Heating/cooling	Simulated
Ali et al. (2020)	Residential	Annually	Overall	Simulated/real-world
Gao et al. (2020)	Non-residential	Daily	Electricity	Real-world
Sendra-Arranz and Gutiérrez (2020)	Non-residential	Hourly	Ventilation	Real-world
Pan and Zhang (2020)	Non-residential	Annually	Overall	Real-world
Seyedzadeh et al. (2020)	Non-residential	Annually	Overall/Emissions	Simulated
Thrampoulidis et al. (2021)	Residential	Annually	Overall/Emissions	Simulated
This work's focus	Residential	Annually	Heating	Real-world

the term (machine learning) algorithm is widely established (Amasyali and El-Gohary 2018). In this study we use the terminology "methods" when referring to data-driven, engineering, or hybrid EQMs in general and "machine learning algorithms" when referring to individual instances of data-driven EQMs, e.g., RF or ANN. Even though hybrid methods try to exploit the advantages of engineering as well as data-driven methods while simultaneously minimizing their disadvantages, the necessary knowledge about both EQMs as well as computational inefficiencies poses a great challenge, which makes the hybrid methods less attractive (Wei et al. 2018). Thus, in our study we focus on engineering and data-driven EQMs. Engineering EQMs model the thermal behavior of heat flows in buildings based on physical laws (Amasyali and El-Gohary 2018). Figure 1 displays exemplarily the heat flows considered in engineering EQMs. These include, for example, transmission heat losses H_T through the building shell (e.g., walls, windows, roof, etc.), ventilation heat losses H_V, caused by airing or leakages in the building shell, solar heat gains Q_S, and internal heat gains Q_i (e.g., electrical consumers or heat radiated by occupants). The heating energy demand Q_h provided with a heating system is consequently calculated from the heat losses, to ensure a constant room temperature. In addition, the demand for hot water heating Q_{tw} must be calculated and the heating system's efficiency considered (Ettrich 2008).

Over the past 50 years, different types of engineering EQMs varying in model complexity and prediction accuracy were developed (Zhao and Magoulès 2012). For the case of calculated EPCs from Germany, quasi-steady-state methods are prescribed by the Energy Savings Ordinance (Eicker et al. 2018). Generally, engineering EQMs require detailed information about all building components and its environment, like external climate conditions, geometrical data, building construction, material properties, or operation (Zhao and Magoulès 2012). Especially for existing buildings the required information and parameters are hardly accessible, thus costly and time consuming to collect (Wang et al. 2012). Furthermore, engineering EQMs are widely discussed for their prediction accuracy, revealing high energy performance gaps, as highlighted in Sect. 2.1.

In contrast to engineering EQMs, data-driven EQMs do not require detailed knowledge about building physics and technical aspects, but use machine learning algorithms to predict building energy performance by learning from available data (Amasyali and El-Gohary 2018). Data-driven EQMs require algorithm training, testing, and validation (Bourdeau et al. 2019). In addition, previous work has to be put in data collection and pre-processing (Kaymakci et al. 2021). Data-driven EQMs have shown convincing results in research regarding prediction accuracy and have surpassed engineering EQMs in several studies (Wei et al. 2018). Researchers agree that data-driven EQMs designed for a particular application achieve the highest degree of accuracy (Mosavi et al. 2019). Yet, a major limitation of data-driven EQMs is the data availability and data quality (Foucquier et al. 2013).

ANN, SVR, and decision trees (or RF and XGB as decision tree ensembles) are the three most used machine learning algorithms for predicting building energy performance (Amasyali and El-Gohary 2018). Even though Bourdeau et al. (2019) and Amasyali and El-Gohary (2018) indicate that SVM and ANN may be the best performing data-driven EQMs to predict building energy performance, there is no consistent picture in the literature yet as to which EQM performs best in terms of prediction accuracy (Ahmad et al. 2018; Aydinalp et al. 2004; Wei et al. 2018). Different advantages and disadvantages of data-driven EQMs like dealing with incomplete data, complexity of the models' training process, or computation speed are discussed. Particularly interesting is the novel D-vine copula quantile regression. Copulas are essentially d-dimensional distribution functions, which can also be used for energy quantification or prediction. They are especially suited for complex prediction tasks, as copulas are able to capture complex dependence patterns even in the tails of the

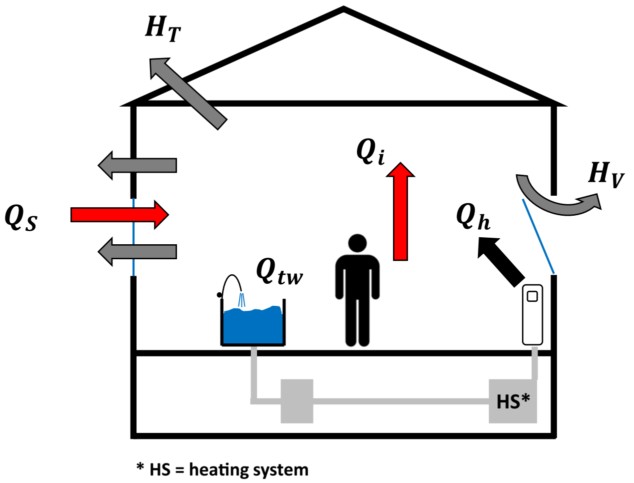

Fig. 1 Generic illustration of heat flows considered in engineering EQMs to calculate the heating energy demand (own illustration based on Ettrich (2008))

distributions (Czado 2019; Nelsen 2010). So far, copulas have been applied to various fields of study and have convinced with promising results (Kraus and Czado 2017; Schallhorn et al. 2017; Töppel et al. 2019).

2.3 Performance Evaluation Measures

Predictive analytics requires empirical predictive models and methods for evaluating their predictive power – PEMs (Shmueli and Koppius 2011). In literature several PEMs are broadly discussed. Amasyali and El-Gohary (2018) provide an overview of the most commonly-used PEMs for predicting building energy consumption. As the most widely used PEMs they mention the Coefficient of Variation (CV), the Mean Absolute Percentage Error (MAPE), the Root-Mean-Square Error (RMSE), and the Mean Absolute Error (MAE). Table 2 gives an overview of the respective PEMs, including their formal definitions, units, value ranges, and optima.

F_i and A_i are the predicted and actual values for the FEP for an instance i, N is the sample size, and $\overline{A}$ is the mean of all actual values A_i. Each PEM exhibits different characteristics, leading to different outcomes of prediction accuracy. Outlier sensitivity is an important characteristic, as high deviations between predicted and actual values are not beneficial for EQMs. Furthermore, a unitless measure provides intuitive interpretation and understanding of the PEMs for readers not familiar with this subject. Both characteristics support the fact that the CV is the most commonly-used PEM, as well as its recommendation for energy consumption prediction models by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (American Society of Heating, Refrigerating and Air-Conditioning Engineers 2002). As the selection of the best suited PEM is not trivial, comparing several PEMs is preferable (Botchkarev 2019). Therefore, in this study, despite focusing primarily on the CV, we additionally provide information on the other three PEMs as well.

3 Methodology and Study Design

To address the research question and benchmark different EQMs, a suitable methodology and study design are necessary. Benchmarking is a well-known and often used term recognized as an essential instrument for improving product and organizational performance, even if benchmarking activities may vary strongly today (Ketter et al. 2015). To meaningfully structure the benchmarking of different EQMs, we derived a seven-step process illustrated in Fig. 2, which is based on the Cross Industry Standard Process for Data Mining (CRISP DM) and the guidelines by Müller et al. (2016) for conducting big data analysis.

Generally, the CRISP DM provides a standardized process to increase business understanding by applying data mining methods in six steps: "Business

Table 2 Overview of the most common Performance Evaluation Measures in analogy to Amasyali and El-Gohary (2018) and the Mean-Squared Error used for model learning

Equation number	Performance evaluation measure	Equation	Unit, value range	Best value
(5)	Coefficient of variation (CV)	$\mathrm{CV} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - A_i)^2}}{\overline{A}}$	$-, [0, \infty)$	0
(6)	Mean absolute percentage error (MAPE)	$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left\lvert\frac{F_i - A_i}{A_i}\right\rvert \cdot 100$	$\%, [0, \infty)$	0
(7)	Root-mean-square error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(F_i - A_i)^2}{N}}$	$\frac{kWh}{m^2a}, [0, \infty)$	0
(8)	Mean absolute error (MAE)	$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert F_i - A_i\rvert$	$\frac{kWh}{m^2a}, [0, \infty)$	0
(9)	Mean-squared error (MSE)	$\mathrm{MSE} = \frac{\sum_{i=1}^{N}(F_i - A_i)^2}{N}$	$\left(\frac{kWh}{m^2a}\right)^2, [0, \infty)$	0

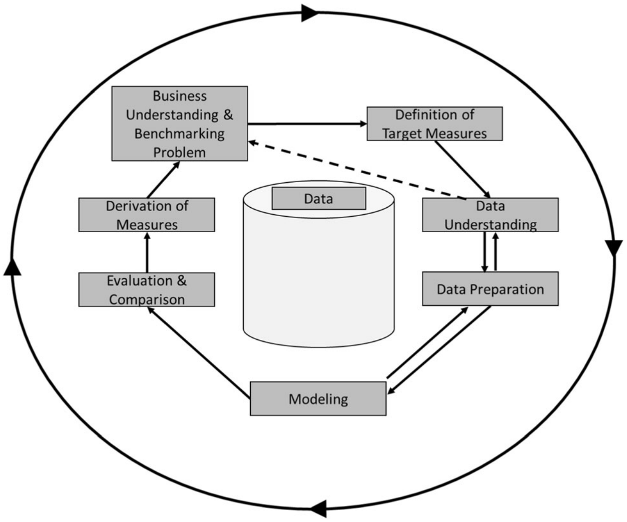

Fig. 2 Derived process to benchmark energy quantification methods for predicting building energy performance (own illustration based on Wirth and Hipp (2000))

Understanding", "Data Understanding", "Data Preparation", "Modeling", "Evaluation", and "Deployment" (Wirth and Hipp 2000). We explain our derived process steps in the following:

"**Business Understanding and Benchmarking Problem**": We extend the initial first stage of "Business Understanding" with our main objective of solving the benchmarking problem of different EQMs. In addition, we modify the intention of the business understanding to collect domain specific knowledge about building energy performance and EQMs, which is necessary for the benchmarking problem, as we do not intend to get deeper business insights by applying data mining methods. We presented domain specific knowledge in Sect. 2 providing the theoretical background for EPCs and EQMs. As benchmarking candidates, we choose the legally required standard engineering EQM and some well-selected data-driven EQMs. We use the three most commonly used machine learning algorithms in literature for predicting the energy performance of buildings, namely ANN, SVR, and RF (Amasyali and El-Gohary 2018). In addition, we consider the ensemble learning algorithm XGB and D-vine copula quantile regression that showed promising results in recent case studies (Schallhorn et al. 2017; Touzani et al. 2018). With this selection, we can investigate a wide range of models from simpler models like RF to more complex models like SVR. After selecting our benchmarking candidates, we modify the CRISP DM again by introducing our target measure the FEP, before proceeding with data understanding.[2]

"**Data Understanding**": This step was not modified. In our study we dispose of a training and a separately collected validation dataset, which will be explained in Sect. 4.

"**Data Preparation**": This step was not modified either. We prepare the data, such that they are available in high quality and can be further used appropriately. For this purpose, we apply the two-stage LANG approach to check for semantic and syntactic data constraints in Sect. 4 (Zhang et al. 2019).

"**Modeling and Evaluation**": In these steps we implement, train, and tune our EQMs (c.f. Sect. 5). With the trained models we predict the FEP for each building in the validation dataset, which allows us to meaningfully compare the different EQMs based on the PEMs in Sect. 6. Thereby, we conduct two benchmarking analyses. (1) We train the data-driven EQMs on the first dataset and benchmark them against the engineering EQM on the out-of-sample second dataset encompassing the FEP calculated by the energy auditors according to the normative framework. (2) We subsequently benchmark only the data-driven EQMs based on nested cross-validation on all available data to get the most robust results while complying with state-of-the-art machine learning techniques. Since the calculated FEP is not given for the first dataset, we do not include the engineering EQM in the second benchmarking analysis.

"**Deployment**": This step largely coincides with the step of deployment in the original CRISP DM. We discuss our results and present derived implications for policy, research, and commercial application in Sect. 7.

Last, our results contribute to solve the defined benchmarking problem with our research question in the first step and close the process cycle. Further iterative rounds of the process could be used to adapt single process steps for further insights.

4 Data and Pre-processing

In this study, we used three real-world datasets to derive the target measure FEP for the benchmarking of the EQMs. The first dataset comprises 25,000 single and two-family buildings from Germany with 74 attributes containing information on the building characteristics, e.g., physical

[2] Note, that our datasets were already at our disposal when we started with our study. Hence, we did not include "Data Collection" to our derived process. Nevertheless, this step could be set in parallel with the definition of the target measure to allow a general process application and to enable further studies starting without available datasets.

building attributes and geometry, the installed heating system, the location, and the annual metered thermal energy consumption.[3] Information about the occupants is not available. This dataset serves as training and test data for the data-driven EQMs. The second dataset originates from two German energy consulting companies that employ qualified energy auditors and includes 345 additional single and two-family buildings with 35 attributes each, which were collected during on-site inspections by the employed auditors in the period between 2016 and 2018. Next to the metered annual thermal energy consumption, the dataset also contains the calculated annual energy demand from EPCs, which represents the engineering EQM. We therefore use this second dataset as validation data for the benchmarking against the engineering EQM. The calculation rules and specifics for the creation of EPCs are updated frequently (Platten et al. 2019). To compare EPCs correctly requires that they follow the same calculation rules. The calculated EPCs in this dataset were each created according to the standard DIN V 4108–6 in combination with DIN V 4701–10. As there were no normative changes concerning the FEP during the period of the survey, the dataset does not need to be adjusted (Beuth Verlag GmbH 2004, 2016). The third dataset is a statistical survey from the German micro census 2011, which represents the household and building stock of Germany (Federal Statistical Office of Germany 2011). This dataset will later be used for stratification purposes to ensure representative results.

To calculate the target measure FEP we had to make some assumptions. Following Eq. 4, the FEP is calculated from the consumption, the climate factor, and the effective building area. Since the latter two were not directly included in the datasets, we assumed that each building contains a heated basement and applied Eq. (1) with $f_c = 1.35$ to derive the effective building area. We further retrieved the mean climate factor over the period the datasets were gathered from historical data by mapping the buildings to the nearest weather station based on the zip code. Finally, we inserted these values in Eq. 4 to calculate the FEP.

To ensure high data quality we cleansed the training and validation datasets. First, we reduced the attributes to the intersection between the two datasets. This is necessary, because otherwise we would train the EQMs on data we cannot provide for validation. Nonetheless, the datasets shared a large intersection in the most important attributes, containing identical or similar attributes that could be easily converted. Second, we excluded attributes lacking explanatory power for the FEP, like identification numbers, as well as attributes with few entries. Also, we deleted faulty or contradicting data entries, e.g., when the age of the roof is older than the building age itself. Third, we eliminated outliers in the attributes living space and final energy consumption, using the thresholds of Metzger et al. (2019). The resulting datasets contained 20,348 and 330 data entries, respectively, with a total of 15 attributes, illustrated in Table 3.

Some data-driven EQMs require further processing steps to increase their prediction accuracy. Because these processing steps are not identical for all EQMs, we further processed the data algorithm-specifically. For the ANN, this involved normalizing all numerical attributes to [0,1] and one-hot encoding all non-numerical attributes, i.e., introducing a binary dummy variable for $n-1$ instantiations (Jovanović et al. 2015). For SVR, we only performed one-hot encoding, while no further pre-processing is required for the RF and XGB. For the copula, we applied continuous convolution to each attribute (Nagler 2018a, b).

To ensure representativeness of our study, we post-stratified our results with regard to building age based on the third dataset according to the German building stock.[4] Stratification describes a sampling procedure, in which representativeness with regard to a desired attribute is ensured by sampling in the respective relation from the different subpopulations (Bowley 1925). Post-stratification takes place after data collection. We post-stratify our results by adjusting the PEM to the German building stock. First, we calculate the PEM for each subpopulation – in our case the building age class –, then calculate a weighted average according to the building age class distribution in the German building stock. This method is used with great success in various fields of study (Bowley 1925); Heinisch 1965; Miratrix et al. 2012). Table 4 shows the percentages of the overall German building stock and our datasets, illustrating why post-stratification is necessary. Henceforth, when we refer to our PEMs, we use the stratified PEMs when applicable.

Table 5 summarizes the individual pre-processing steps.

[3] The data originate from the nationwide “Modernisierungs-Kompass” (Modernization Compass) offered by the EN-OP Institute (enop.de). It comprises free written modernization consultation for owner-occupied single- and two-family buildings, which was used more than 300,000 times between 1983 and 2014. The data used in this study were transmitted between 2007 and 2014. The service was discontinued in 2014 and the company ISO GmbH became a legal successor of the EN-OP Institute.

[4] Note, that the census only provides aggregated information to mitigate the risk of information leakage. Further, the census distinguishes between a total of eight classes. Due to the low retrofit potential, the newest two classes were considered jointly as class 7.

Table 3 Input parameters for data-driven Energy Quantification Methods

Category	Attributes	Values
Miscellaneous	Basement available	Yes, no
	Building construction year	Year
	Building type	Detached, attached
	Living space	m^2
Wall insulation	Double skin construction insulation	cm
	Outer wall construction year	Year
	Outer wall insulation thickness	cm
	Presence of double skin construction insulation	Yes, no
Heating system	Type of energy source	Oil, gas, district heat, etc
	Boiler construction year	Year
Roof	Presence of roof insulation	Unknown/none, partial, full
	Year of the last roof covering	Year
Windows	Material of window frame	Wood, plastic, aluminum
	Type of window-glazing	Single, double old, double modern, triple-glazing thermal insulated
	Window construction year	Year
Final energy consumption	(Target measure)	$\frac{\text{kWh}}{\text{m}^2\text{a}}$
	Weather effects adjusted final energy consumption	
Final energy demand	(Representing the engineering EQM)	$\frac{\text{kWh}}{\text{m}^2\text{a}}$
	Calculated final energy demand by energy auditors from the EPCs	

Table 4 Distribution of building age classes in Germany (census) and in our datasets

Classes	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Building age	– 1918	1919–1948	1949–1978	1979–1986	1987–1990	1991–2000	2001-
Percentage in census	14%	13%	46%	10%	3%	8%	6%
Percentage in training dataset	6%	10%	55%	14%	5%	8%	2%
Percentage in validation dataset	6%	11%	43%	15%	9%	14%	3%

5 Model Fitting and Tuning

As mentioned, the results for the engineering EQM are only available for the validation dataset. This in turn means, that benchmarking the engineering EQM is also only possible on this dataset and, consequently, benchmarking against the engineering EQM applying nested cross-validation on the larger training dataset is not possible. Thus, to receive the most reliable results and make the best use of our available data, we conducted two benchmarking analyses. In the first analysis, we applied cross-validation on the training dataset and evaluated the prediction accuracy on the validation dataset for all EQMs including the engineering EQM, while in the second analysis we further benchmarked only the data-driven EQMs against each other based on nested cross-validation on all data. We implemented and tuned each algorithm in the statistical programming language R.

For the first analysis, we applied cross-validation on the training dataset with hyperparameter tuning based on genetic algorithms (Friedrichs and Igel 2005; Goldberg 2012). In this vein, we defined areas for all relevant hyperparameters and randomly initialized a population. For each hyperparameter specification we trained a model and evaluated its fit based on the CV, handing the best performing specifications over to the next generation. Additionally, new hyperparameter specifications were added by crossbreeding and mutating the more successful

Table 5 Methods used for data pre-processing

Step	Method/procedure	Applied to
1	Pre-screening of variables and underlying EPC calculation rules	Train and validation data
2	Identification of variable intersection	Train and validation data
3	Conversion of data to same scales and ordinal specifications	Train and validation data
4	Calculation of the FEP (including effective building area and climate factors)	Train and validation data
5	Deletion of variables lacking relevance and completeness (explanatory power for the FEP, missing information, etc.)	Train and validation data
6	Deletion of data entries lacking correctness (exceeding definition ranges, contradictory, etc.)	Train and validation data
7	Outlier deletion according to Metzger (living space and FEP)	Train and validation data
8	Algorithm-specific pre-processing and feature engineering	Train and validation data
9	Post-stratification on the predictions with respect to the building age	Results

specifications while ensuring parameter constraints (e.g., integer values for the hidden layers or non-negativity constraints). This procedure was applied over 200 generations, or until an early callback indicated no further improvement in CV. Once we identified the best performing hyperparameter specification, we trained a model with the tuned hyperparameters on the entire training dataset to not lose any information before evaluating their prediction accuracy on the validation dataset.

For the second analysis, i.e., to benchmark only the data-driven EQMs in-depth, we proceeded mostly in the same way with the sole difference, that we applied nested cross-validation on all data instead, which had not been possible before due to missing results for the engineering EQM in the larger training dataset. This two-stage approach allowed us to compare the data-driven methods against the engineering EQM while still receiving robust results for the benchmarking of the data-driven methods.

In what follows we cover method-specific details on the model tuning process. However, because a holistic introduction to all relevant hyperparameters for the different algorithms is neither content wise nor in terms of space within the scope of this manuscript, we refer to the literature for thorough explanations and only provide the information necessary to reproduce this study. To lever comparability, we used MSE where applicable for model training and CV for (outer-)fold performance evaluation. The respective tables in the Appendix A1 (available online via http://link.springer.com) show the final set of hyperparameters and their value ranges during the tuning process.

Random Forest: For the RF we used the R package “randomForest” (Breiman et al. 2018). Because we apply regression, we fitted each individual tree minimizing the MSE as error metric instead of the information gain used for classification.

Extreme Gradient Boosting: For the XGB we used the R package “xgboost” (Chen et al. 2020) and proceeded similar to the RF. We again used regression minimizing the MSE.

ANN: For the ANN we used the R packages “keras” and “tensorflow” (Falbel et al. 2020a, b). We fitted the individual models using Adam as optimizer based on rectified linear units as activation functions for the hidden layers and a linear output function. The model was trained minimizing the MSE on 500 epochs, however using early callback if there was no significant improvement in the test data to avoid overfitting.

SVR: For the SVR we used the R package “e1071” (Meyer et al. 2019). We used radial basis kernel functions and applied Epsilon-regression. This procedure prioritizes good model fit over simple solutions, which is in line with the overall goal of this study. We then fitted the individual models according to the underlying optimization function.

Copula: For the copula we used the R packages “vinereg” and “VineCopula” (Nagler et al. 2019; Nagler 2019). For the copula there are no hyperparameters to be tuned in the classical sense. Instead, we applied a parsimonious forward selection algorithm by Kraus and Czado (2017), which sequentially builds up the model using the Akaike Information Criterion based on conditional log-likelihood as termination criterion. The algorithm thereby automatically fixes the tree sequences in the vine copula structure. Once the termination criterion threshold is no longer breached when adding variables, the algorithm stops. The resulting variable selection and tree structure can be found in the Appendix A2.

6 Results

6.1 Benchmarking against the Engineering Energy Quantification Method

The prediction accuracy of the different EQMs measured by the PEMs is presented in Fig. 3. Here, the EQMs are depicted on the x-axis, while the y-axis indicates the magnitude of the PEMs. Focusing first on the CV, we notice that the engineering EQM lags significantly behind with a CV of 0.614, while the data-driven EQMs provide results in approximately the same range between 0.33 and 0.35. This means that the prediction of the engineering EQM deviates roughly 60% on average from the mean actual FEP. The XGB shows the highest prediction accuracy with a CV of 0.329 which equals a decrease in error of almost 50%. To ensure robustness, we validated these results by means of further PEMs.[5] Thereby, the general tendency remains the same with only minor variations in the exact outcomes. The only notable difference occurs for the MAPE, where the ANN shows the highest prediction accuracy, reducing error by more than 50% compared to the engineering EQM. However, the difference is slight and minor variations in the order of EQMs were expected. Table 6 provides detailed numeric values.

Next, we have a closer look at the individual predictions of the EQMs. Figure 4 shows scatterplots in which we compare the predicted and the metered (weather rectified) FEP for each EQM. The *x*-axes show the predicted values, while the y-axes show the metered values. The blue circles represent the buildings in the validation dataset. For easier interpretation, we provide an angle bisector and a regression line. Ideally, we want all observations to lie on the angle bisector.

The engineering EQM exhibits the highest standard deviation in the errors with $\sigma^{Eng} = 55.45\,\mathrm{kWh/(m^2a)}$, for a mean metered FEP of 126.44 $\mathrm{kWh/(m^2a)}$ over the whole dataset. The engineering EQM is followed by the SVR with $\sigma^{SVR} = 43.99\,\mathrm{kWh/(m^2a)}$ and ANN with $\sigma^{ANN} = 43.28\,\mathrm{kWh/(m^2a)}$. The copula and the RF on the other hand exhibit slightly less standard deviation in the errors with $\sigma^{Cop} = 42.84\,\mathrm{kWh/(m^2a)}$ and $\sigma^{RF} = 42.72\,\mathrm{kWh/(m^2a)}$. The XGB has the smallest standard deviation of $\sigma^{XGB} = 42.07\,\mathrm{kWh/(m^2a)}$. At the same time, the engineering EQM and the RF both overestimate the FEP on average by 50 and 4 $\mathrm{kWh/(m^2a)}$, respectively, while the ANN, SVR, and XGB underestimate the FEP on average by 15, 6, and 5 $\mathrm{kWh/(m^2a)}$, respectively. The copula underestimates the FEP on average very slightly by 0.05 $\mathrm{kWh/(m^2a)}$. Again, we notice that there is high unexplained variance which could stem from different factors like occupant behavior and cannot be explained by the EQMs based on the building characteristics alone.

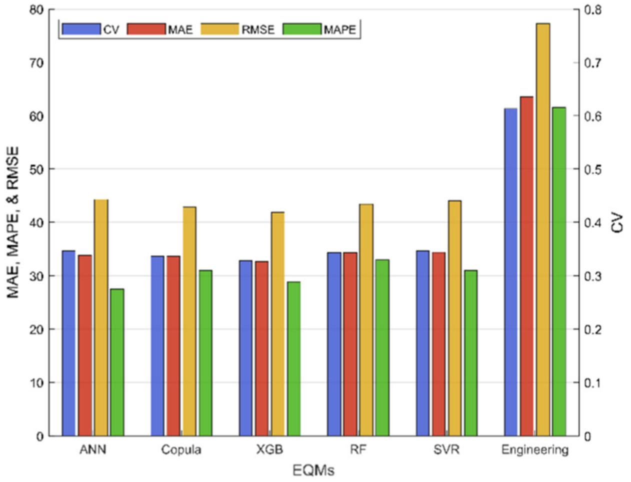

Fig. 3 Performance evaluation measures for the different energy quantification methods

To obtain a more complete picture, we disaggregated the predictions for different instantiations of the variables building age and living space, and analyzed whether there are significant differences. The idea behind this is that systematic errors might have been made when one of the variables takes extreme values, e.g., a very poor prediction accuracy for old buildings. For better readability we aggregated the variables into building age classes and living space bins. For the building age we chose the building age classes from the census to obtain comparability with other studies. For the living space bins, we took the different deciles as separators for a total of ten living space bins.[6] Figure 5 shows the results for the building age classes on the left-hand side and for the living space bins on the right-hand side. The figures are structured analogously, with the x-axes indicating the instantiations of the variables and the y-axes indicating the CV.

For the building age classes, the XGB, copula, and RF show slightly higher prediction accuracy, as expected

[5] Note that hyperparameter tuning was applied based on the CV, thus the remaining PEMs probably leave room for further optimization. The appendix (Table A 5) provides further information on the performance metrics and on overfitting assessments.

[6] We also checked other categorizations for building age classes and living space bins, which yielded similar results.

Table 6 Performance evaluation measures for the different energy quantification methods

	CV	Mean absolute error	Root-mean-square error	Mean absolute percentage error
ANN	0.347	33.829	44.314	**27.541**
Copula	0.337	33.673	42.819	30.982
XGB	**0.329**	**32.724**	**41.921**	28.917
RF	0.344	34.359	43.441	32.983
SVR	0.347	34.446	44.021	31.073
Engineering (Benchmark)	0.614	63.577	77.273	61.549

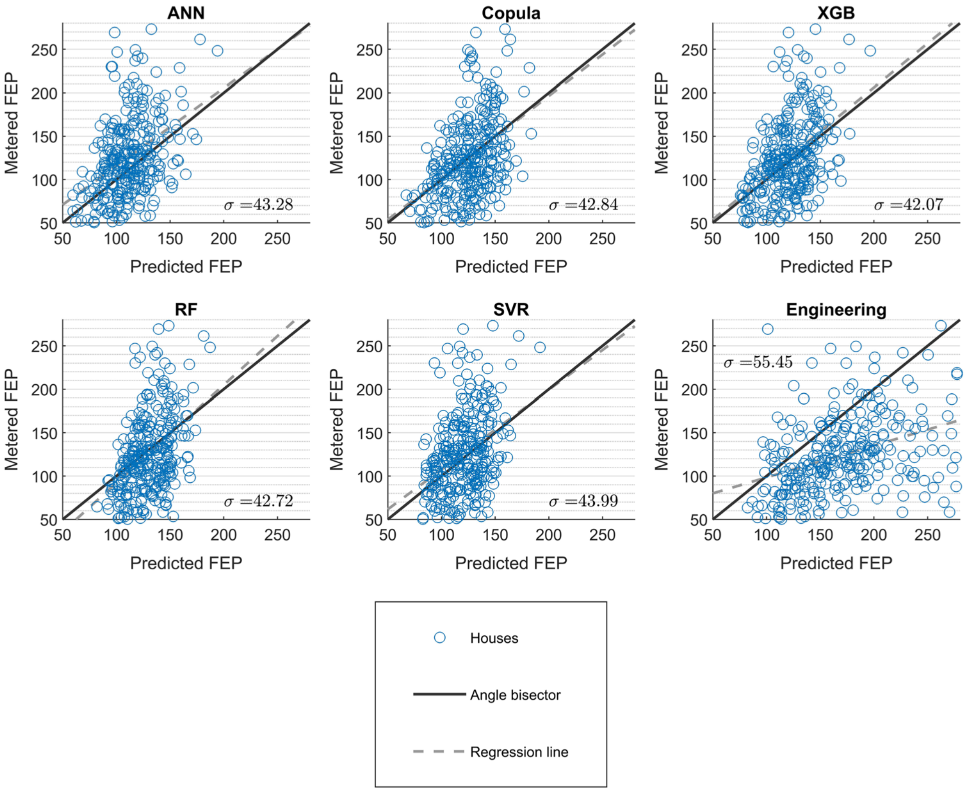

Fig. 4 Scatterplots of predicted and metered final energy performance for the different energy quantification methods

based on the aggregated results. While the data-driven EQMs produce similar results throughout all building age classes, the engineering EQM increases in prediction accuracy towards newer buildings until 1990. This is in line with findings in literature of higher measurement errors for buildings with lower energetic efficiency in England and Wales (Crawley et al. 2019), thus older buildings with less strict regulations, which can be explained by the underlying data quality. As mentioned, engineering EQMs require exact inputs and expert knowledge to produce viable outcomes. For older buildings this is often not the case, especially when construction methods or building materials used are unknown. The final increase in the last two building age classes is partially explained by the CV being a relative PEM. Stricter building construction regulations came into place in Germany from 1977, followed by further aggravations leading to lower overall FEP, which in turn yields higher CVs for the same absolute error (Deutsche Energie-Agentur GmbH 2016). For the living space, we notice an overall trend towards more accurate predictions for larger buildings. However, this trend is less pronounced when compared to the building age classes and therefore does not allow for conclusions. Again, the XGB and RF show superior prediction accuracy for most living space bins.

Last, we evaluated the individual over- and underestimations for different building age classes, as literature describes a general overestimation bias of FEP for older

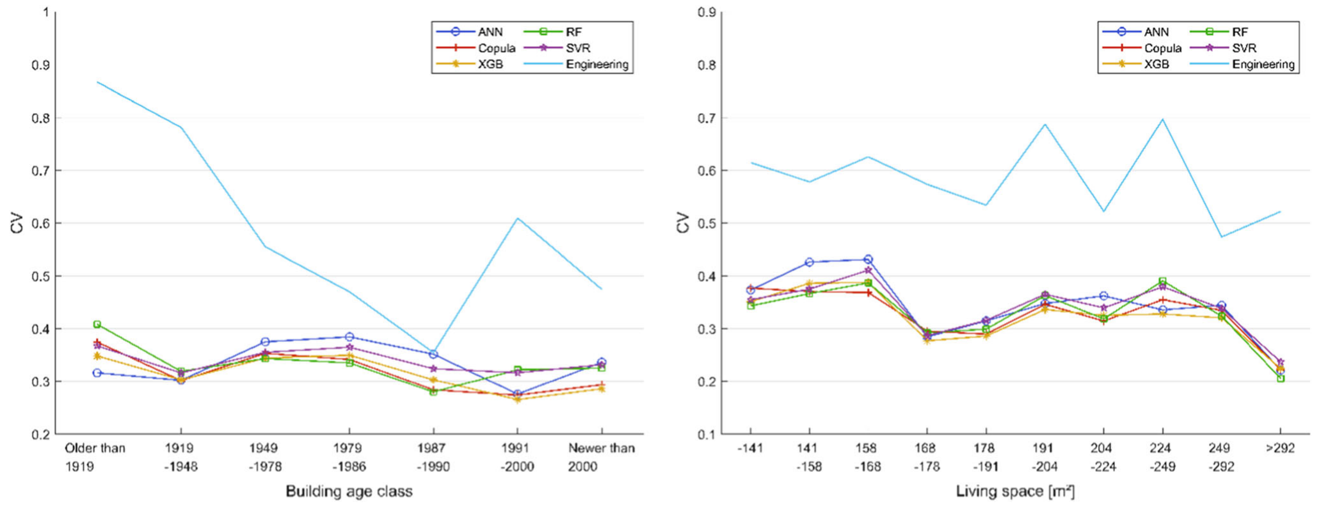

Fig. 5 Coefficient of Variation for the different Energy Quantification Methods for instantiations of the variables building age on the left-hand side, aggregated into building age classes, and living space on the right-hand side, aggregated into living space bins

and an underestimation for newer buildings (Greller et al. 2010). Figure 6 reports on the results, whereby the *x*-axis again depicts the building age classes and the *y*-axis depicts the mean prediction error as the difference between predicted/calculated and the real measured values. We note that the EQMs indeed overestimate the FEP for older buildings, however, cannot validate an underestimation of newer buildings. Analogously, in Fig. 4 the slope of the regression line from the engineering EQM is lower than the bisector. This supports the findings of Cozza et al. (2020), who found a lower actual consumption for energy inefficient residential buildings and a higher actual consumption for efficient residential buildings in the Swiss building stock. The fact that the overestimation is greater in older buildings with poorer energy efficiency must be urgently improved, so that the negative publicity and unreliable statements for buildings in need of renovation do not prevent investments in retrofitting measures. The lower estimation error for data-driven EQMs may result from the fact that the training dataset contains measured energy consumption and thus implicitly considers occupant behavior. Following Greller et al. (2010), the higher deviations in older buildings could be due to a more savings-conscious user behavior on average for less energetic efficient buildings, which is associated with a higher rejection of temperature comfort than assumed in the standards for calculation.

6.2 Benchmarking the Data-driven Methods

In this subsection we present the results for the in-depth benchmarking of the data-driven EQMs only. Because we applied nested cross-validation with five outer folds and ten inner folds, we obtain as a result not one but five tuned models per algorithm which performed best for their respective outer folds. To still present the results in a clear and understandable way, we aggregated the prediction accuracies by calculating the mean PEMs. Figure 7 presents an overview over the prediction accuracies of the different EQMs measured by the PEMs.

We notice that the differences in prediction accuracy almost completely vanish when we use nested cross-validation for performance evaluation instead of the validation set. When aggregated, the accuracies differ by less than 1% regarding CV. We further notice that the overall prediction accuracy increases slightly for most PEMs. Both effects are to be expected, as the repeated evaluation procedure yields more robust results and allows for in-sample training. The XGB and SVR slightly outperform their competitors in most cases. Table 7 further reports on the exact values and the standard deviations given in brackets. The standard deviations in the results reveal that the ANN mostly exhibits the highest standard deviation in prediction accuracy, thus its results should be treated with more caution. RF on the other hand scores very consistently.

Last, we provide some insights into variable importance to increase the explainability of the models. However, Shmueli and Koppius (2011) state that explanation and prediction should be best thought of as separate modeling goals. Consequently, any model trying to encompass both will have to compromise. This means that the following analyses should be interpreted with caution, as our goal was prediction and not explanation. To derive the variable importance for each of the models, we used the method initially proposed by Breiman (2001). The importance is derived by permuting the predictor variables and measuring the decrease in accuracy. Figure 8 shows the results for the five most important variables of the data-driven EQMs,

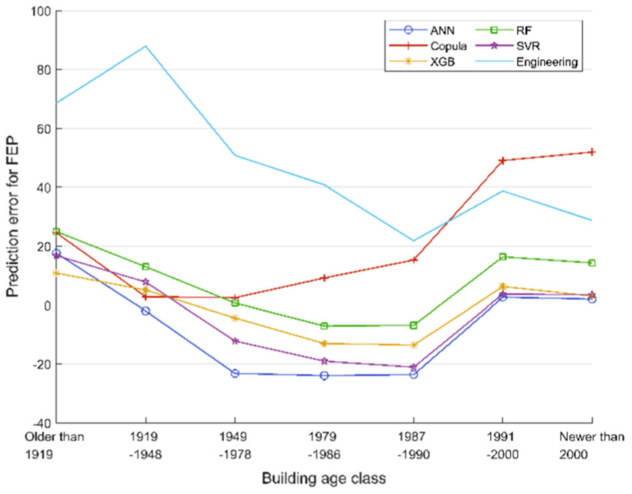

Fig. 6 Mean prediction error of the Final Energy Performance for the building age classes

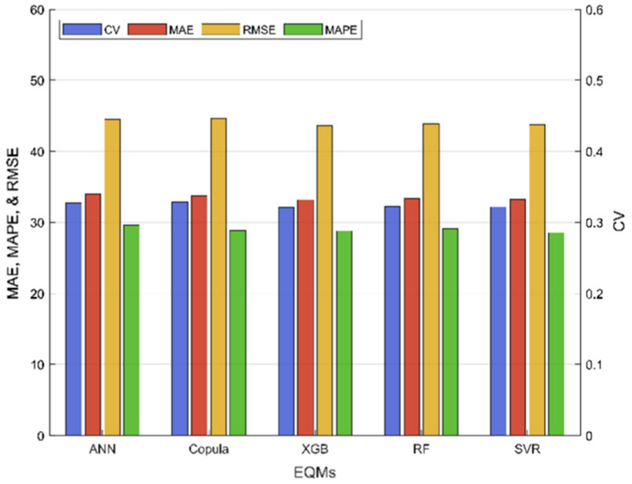

Fig. 7 Mean performance evaluation measures for the different data-driven energy quantification Methods over the five outer folds

with higher values corresponding to higher importance. A complete enumeration of all variables and their respective importance for each algorithm can be found in the appendix (Figure A 2).

We notice that the living space is highly important for all data-driven methods. This is explained by changing heating behavior and usage patterns of rooms depending on the available living space. Because the number of residents in single and two-family houses does not generally increase with the living space, the utilization of all rooms decreases. For example, rooms are used as storage rooms, for sports or as repair shops and are not necessarily heated. Since the data-driven EQMs were trained on measured consumption, they could learn this correlation. Next to the living space, the energy source is consistently important. This is also not surprising, as the heating system is of central importance for the overall energy efficiency. The remaining variables are less consistent in their importance. Nonetheless, we notice similarities between the two tree-based algorithms XGB and RF, as well as between the ANN and SVR which both use one-hot-encoding. For the copula, the importance can be inferred to a certain degree from the tree structure for the bivariate copula building blocks. Moreover, due to the parsimonious forward selection algorithm applied for model fitting, the copula disposes of less variables (c.f. Figure A 1 and Figure A 2).

7 Discussion

Our results show that the energy performance gap generally holds true for single- and two-family buildings in Germany. The engineering EQM produces approximately the values for the energy performance gap as expected in literature. The data-driven EQMs are also in the expected range but exhibit a considerably lower error. The lack of literature for our specific benchmarking problem of predicting annual heating energy performance in residential buildings does not allow a holistic discussion of the accuracy gap between data-driven and engineering EQMs. Nevertheless, compared to the results of Neto and Fiorelli (2008), who compared an engineering EQM with an ANN for time series prediction of energy consumption of buildings, the data-driven EQMs in our study show an even

Table 7 Mean performance evaluation measures for the different data-driven energy quantification methods over the five outer folds (standard deviation given in brackets)

	CV	Mean absolute error	Root-mean-square error	Mean absolute percentage error
ANN	0.328 (0.0081)	34.003 (0.7381)	44.505 (0.9654)	29.577 (1.0671)
Copula	0.329 (0.0057)	33.736 (0.4565)	44.689 (0.8473)	28.805 (**0.2716**)
XGB	**0.321** (0.0060)	**33.173** (0.4735)	**43.618** (0.7638)	28.784 (0.5329)
RF	0.323 (**0.0052**)	33.384 (**0.4559**)	43.865 (**0.7199**)	29.114 (0.4757)
SVR	0.322 (0.0064)	33.212 (0.4566)	43.773 (0.8201)	**28.544** (0.4189)

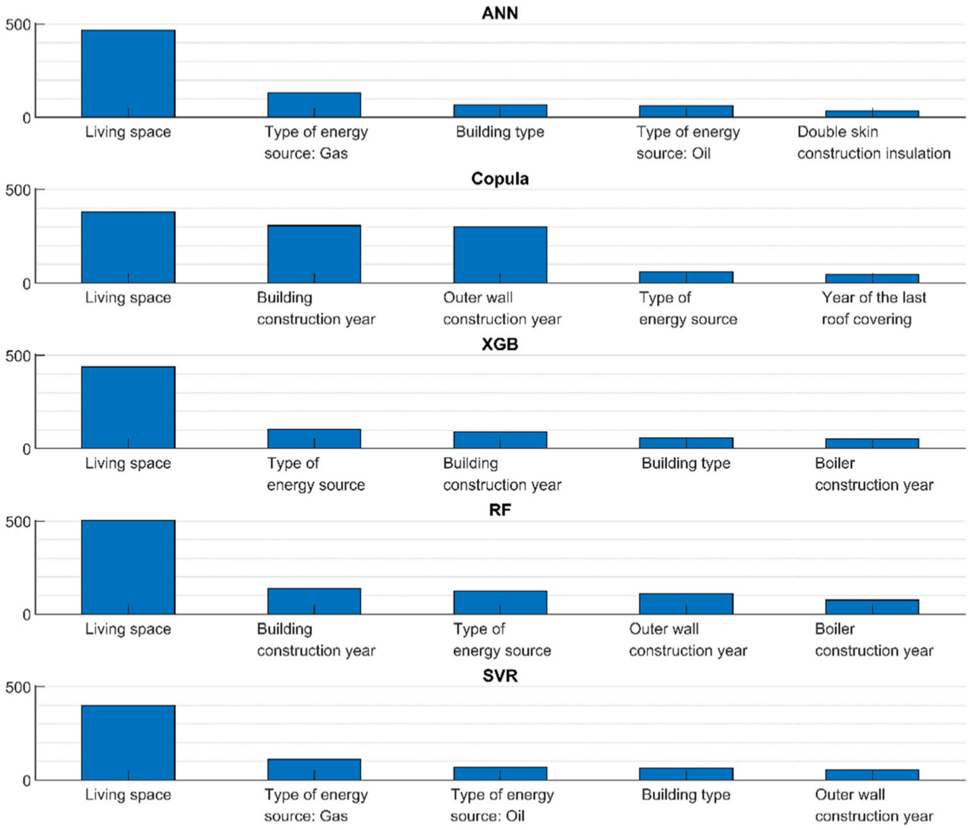

Fig. 8 Variable importance plots for the data-driven energy quantification methods

greater advantage in terms of prediction accuracy. In their study the ANN achieved a 3-percentage point advantage, whereas our data-driven EQMs achieve almost a 30-percentage point advantage over the engineering EQM. However, our analyses do not confirm previous findings in literature that ANN and SVR possess generally better prediction accuracy for building energy performance than less complex machine learning algorithms like RF (Amasyali and El-Gohary 2018). Rather, XGB exhibited the highest prediction accuracy for most analyses conducted, closely followed by SVR and RF. ANN, on the other hand, performed worst to second worst among the tested data-driven EQMs. However, the differences in prediction accuracy were slight and the standard deviations indicate that these results should be treated with caution. Consequently, we refrain from stating that one data-driven EQM is particularly suited for this task. Nonetheless, this supports that each application requires a specifically designed EQM to reach its highest accuracy, and that there is no strictly dominant EQM (Mosavi et al. 2019). Because our data-driven EQMs rely solely on few attributes which are relatively easy to grasp compared to the engineering EQMs, we argue that data-driven EQMs exhibit further advantages regarding their handling and applicability. Thus, using data-driven EQMs instead of engineering EQMs saves money and time while simultaneously increasing prediction accuracy.

Our results have several managerial and policy implications. First, they provide clear guidelines for policymakers. The current state of the low-carbon transition paths requires higher retrofitting rates for residential buildings to still reach the climate goals. Therefore, we advocate to revise the current legislation to allow for data-driven EQMs instead of the prescribed engineering EQM with significantly worse prediction accuracy. This potentially raises the residential building retrofitting rate by decreasing the uncertainty of energy efficiency measures, thereby removing investment barriers and contributing to achieving the climate goals. Two different applications are conceivable at present, either the direct replacement of the engineering EQM, or the complementary application used for transitional quality assurance of the engineering EQM to check for outliers or incorrect data. The verification could be automated and thus be realized cost-efficiently and without human involvement. The quality assurance can be rolled out nationwide and increase confidence in the EPC, thus offering a more reliable foundation for decision-making. Potential challenges are the acceptance and ensured quality of the underlying models. Homeowners may perceive unfair treatment if EPCs depicting low energy efficiency are issued based on calculation methods that are not or hardly comprehensible such as black-box approaches, as this reduces the resale value of houses. Using more explainable methods, like RF or XGB might mitigate this challenge. However, there is a whole field of Explainable Artificial Intelligence discussed controversially in literature (Rudin 2019). In addition, inexplicable miscalculations can arise for the data-driven methods, resulting in highly distorted results. We argue, however, that the currently prescribed methods are also highly error-prone if not performed correctly, therefore data-driven methods are to be preferred, due to the generally significantly higher prediction accuracy. When putting data-driven EQMs into a use case perspective, a distinction must be made between EPCs for existing and new buildings. Data-driven EQMs learn from available data, limiting their suitability for creating EPCs for new buildings. Since the construction rate in Germany is comparatively low and the energy saving potentials in existing buildings are much greater, as well as the determination of consumption being more costly and error-prone, the focus should be placed on this use case (Deutsche Energie-Agentur GmbH 2016). Second, we suggest the usage of data-driven EQMs for other applications as well, such as asset management, city planning, insurance, etc., to enhance their business models with more economic decision-making, minimization of risk, and higher profits. The energy efficiency evaluation of buildings is a central element in many areas and can be decisive for the economic success of companies (Bozorgi 2015). To collect cost-efficient information is particularly relevant for the initial energy evaluation of real estate if EPCs are not yet at hand, as energy-efficient buildings yield higher returns and higher rents than energy-inefficient buildings (Cajias and Piazolo 2013). Insurance companies could enhance claim prediction models, or asset management companies could optimize their portfolios with data-driven investment strategies. However, both should be extremely careful with the implementation since miscalculations in investment portfolios are comparatively worse than miscalculations in EPCs. Third, our results imply that more focus should be put onto the benchmarking of different machine learning algorithms, as for our specific use case XGB almost consistently yielded better results than the algorithms ANN and SVR which are favored in literature. Most literature investigated focused on one machine learning algorithm only and disregarded comparisons and benchmarks. This, however, results in a limited generalizability of their results.

Naturally, this research is beset with some limitations. First, we focused on annual heating FEP of German residential buildings. Other results might hold true for, e.g., commercial, or industrial buildings, as well as for other geographical regions or time horizons. Second, because the validation dataset was gathered by qualified energy auditors, there might be a systematic selection bias in the individual data points. However, the fact that we validated out of sample, i.e., that the data-driven EQMs could not learn this potential systematic bias, suggests that the relative improvement over the engineering EQM is presumably even more substantial than this study predicts. Third, several important building characteristics were missing in the dataset, e.g., upper floor insulation and basement insulation. More importantly, we also have no information on socio-economic factors or occupant behavior. This leaves a large margin of variance in the data unexplained. Fourth, for the calculation of the target measure in accordance with the current norms, some assumptions were made regarding basement availability and heating. We approximated the effective building area for all buildings where only the living space was given, but did not find any signs in our analysis that this approximation would lead to higher errors. In contrast, in the case of the buildings that are approximated by the living space, the errors in the building energy performance are consistently smaller. Nevertheless, future research could start here by training and analyzing on a complete dataset also including this information. Moreover, for the rectification of weather effects, we used the mean of the climate factor for each weather station over the period the datasets were gathered, because the datasets did not contain the exact year of data collection, but a span of seven years. These assumptions and simplifications could possibly lead to minor deviations in the final outcomes. In addition, the measured consumption could have been further rectified with regard to room and heating threshold temperatures that deviate from the standard assumption, vacancies, or measurement inaccuracies for non-network-bound energy sources (e.g., wood pellets or heating oil) if corresponding data were at hand (Bigalke and Marcinek 2016). Fifth, there exist further EQMs that were not considered in this study, which does not allow to state a final recommendation. Nevertheless, these EQMs can also be benchmarked by applying our methodology adapted from the CRIPS DM cycle. We are convinced that our derived process is generally applicable in the context of benchmarking and can be used in the future for comparison and benchmarking in various situations. Also, even though we tried to provide a comparable basis for all EQMs, by changing individual steps and spending more time in the optimization procedures improvements in prediction accuracy could have been achieved.

However, these limitations give rise to new research potential. One natural direction includes gathering additional high-quality data points, which include all necessary building characteristics as well as occupant behavior. However, this procedure might prove cumbersome. Another direction includes examining further EQMs as well as tuning them to a higher extent. In particular for the copula, we expect the more general R-vines to perform significantly better. To the best of our knowledge, no implementations of R-vine quantile regression exist in any statistical programming language, but promising theoretical advances have been recently made. Also, the focus on only one country may be relaxed, incorporating other geographical areas with different characteristics of buildings, climate conditions, and other normative frameworks for EPC calculation to assess whether our findings are generalizable for these areas and circumstances. This could also be an interesting task for transfer or federated machine learning to take advantage of decentralized datasets for large scale machine learning. All in all, further research is necessary in this field, as current research is scarce. This is most likely due to scarce publicly available and processable data as highlighted in literature (Carpino et al. 2019). Since most institutions with the necessary database are state-regulated, we suggest that policymakers enter into cooperation with scientific institutions, since a sufficiently large and high-quality database is essential to obtain reliable and more generally valid results from which to derive meaningful long-term political incentive mechanisms to curb climate change. In the same course of the structured recording of large quantities of quality-assured data, data on occupant behavior should be recorded. This would make it possible to analyze the causes of the significant differences between measured and calculated EPCs as well as between the different EQMs. Based on the obtained knowledge, more precise statements can be made about energy consumption and savings after potential retrofit measures. This in turn enables investment decisions to be taken on a sound basis, while at the same time reducing barriers to energy efficiency investments by minimizing the investment risk (Ahlrichs et al. 2020). In addition, a large high-quality database might allow to reproduce our results and benchmark further EQMs more systematically over all regions in Germany, to essentially mitigate the major drawbacks of our study. Our research also contributes to the theoretical body of knowledge by identifying potential for improvement in the currently established methods and benchmarking multiple EQMs in terms of predictability. Regarding the classification of Shmueli and Koppius (2011), this corresponds to role six (assessing predictability of empirical phenomena) and peripherally touches role four (comparing existing methods).

8 Conclusion

In this study, we benchmarked different Energy Quantification Methods (EQM) for residential buildings, applying a derived process based on the CRISP DM. In doing so, we are among the first to focus on the interface of predicting heating Final Energy Performance for residential buildings, based on real-world data with annual energy predictions.

More precisely, we compared Artificial Neural Networks, D-vine copula quantile regression, Extreme Gradient Boosting, Random Forest, and Support Vector Regression with the engineering EQM currently established by German law. We used an extensive real-world dataset of 25,000 German single- and two-family buildings for model training and testing and another out of sample dataset of 345 additional buildings for validation, also containing Energy Performance Certificates issued by qualified auditors, which represent the engineering EQM. Our results provide strong evidence that the data-driven EQMs outperform the engineering EQM by a large margin, reducing the prediction error by almost 50%. We additionally benchmarked only the data-driven EQMs against each other based on nested cross-validation. In contrast to existing literature, Extreme Gradient Boosting exhibits the highest prediction accuracy for most cases, closely followed by Support Vector Regression, which is favored in literature, and Random Forest. To ensure robustness of our results, we examined several Performance Evaluation Measures and analyzed two variables – the building age and the living space – in more detail to account for potential systematic biases. Despite minor variations, the general tendency holds, indicating robust results. We conclude that data-driven EQMs in general are more suitable for residential building energy quantification. Therefore, we advocate to revise the current legislation to allow for the use of data-driven EQMs in Energy Performance Certificates for existing buildings.

Acknowledgements The authors gratefully acknowledge the unknown reviewers' valuable and profound comments, the support of the Bavarian state ministry of economics and media, energy, and technology (Grant Number IUK-1606-0002//IUK491/001), the Ministry of the Environment, Climate Protection and the Energy Sector Baden-Wuerttemberg (Grant Number BWT17004), the Federal Ministry of Education and Research (Grant Number 03SBE116C) as well as of the Ministry of Economic Affairs, Innovation, Digitalization and Energy of the State of North Rhine-Westphalia (Grant Number PRO 0079E) for making this paper possible. The authors are responsible for the contents of this publication.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Achtnicht M, Madlener R (2014) Factors influencing German house owners' preferences on energy retrofits. Energy Policy 68:254–263. https://doi.org/10.1016/j.enpol.2014.01.006

Ahlrichs J, Rockstuhl S, Tränkler T, Wenninger S (2020) The impact of political instruments on building energy retrofits: a risk-integrated thermal energy hub approach. Energy Policy 147:111851. https://doi.org/10.1016/j.enpol.2020.111851

Ahmad T, Chen H, Guo Y, Wang J (2018) A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: a review. Energy Build 165:301–320. https://doi.org/10.1016/j.enbuild.2018.01.017

Ali U, Shamsi MH, Bohacek M, Hoare C, Purcell K, Mangina E, O'Donnell J (2020) A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings. Appl Energy 267:114861. https://doi.org/10.1016/j.apenergy.2020.114861

Amasyali K, El-Gohary NM (2018) A review of data-driven building energy consumption prediction studies. Renew Sustain Energy Rev 81:1192–1205. https://doi.org/10.1016/j.rser.2017.04.095

Amecke H (2012) The impact of energy performance certificates: a survey of German home owners. Energy Policy 46:4–14

American Society of Heating, Refrigerating and Air-Conditioning Engineers (2002) ASHRAE guideline: measurement of energy and demand savings. Atlanta

Arcipowska A, Anagnostopoulos F, Mariottini F, Kunkel S (2014) Energy performance certificates across the EU. A mapping of national approaches. http://bpie.eu/wp-content/uploads/2015/10/Energy-Performance-Certificates-EPC-across-the-EU.-A-mapping-of-national-approaches-2014.pdf. Accessed 12 Dec 2020

Aydinalp M, IsmetUgursal V, Fung AS (2004) Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks. Appl Energy 79:159–178. https://doi.org/10.1016/j.apenergy.2003.12.006

Baltuttis D, Töppel J, Tränkler T, Wiethe C (2019) Managing the risks of energy efficiency insurances in a portfolio context: an actuarial diversification approach. Int Rev Financial Anal. https://doi.org/10.1016/j.irfa.2019.01.007

Beuth Verlag GmbH (2004) DIN V 4108-6 Berichtigung 1:2004-03, Berichtigungen zu DIN V 4108-6:2003-06. Beuth, Berlin

Beuth Verlag GmbH (2010) DIN V 18599 Beiblatt 1:2010 01, Energetische Bewertung von Gebäuden – Berechnung des Nutz-, End- und Primärenergiebedarfs für Heizung, Kühlung, Lüftung, Trinkwarmwasser und Beleuchtung – Beiblatt 1: Bedarfs-/Verbrauchsabgleich. Beuth, Berlin

Beuth Verlag GmbH (2016) DIN SPEC 4701-10/A1:2016-05, Energetische Bewertung heiz- und raumlufttechnischer Anlagen – Teil 10: Heizung, Trinkwassererwärmung, Lüftung; Änderung A1. Beuth, Berlin

Bigalke U, Marcinek H (2016) Auswertung von Verbrauchskennwerten energieeffizienter Wohngebäude. dena-Studie, Berlin

Botchkarev A (2019) A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdiscip J Inf Knowl Manag 14:45–76. https://doi.org/https://doi.org/10.28945/4184

Bourdeau M, Xq Z, Nefzaoui E, Guo X, Chatellier P (2019) Modeling and forecasting building energy consumption: A review of data-driven techniques. Sustain Cities Soc 48:101533. https://doi.org/10.1016/j.scs.2019.101533

Bowley AL (1925) Measurement of the precision attained in sampling. Institut international de statistique, Rome (1925) Rapport de la Commision sur l'application des méthodes représentatives dans les diverses statistiques. Cambridge University Press, Cambridge, Annexe

Bozorgi A (2015) Integrating value and uncertainty in the energy retrofit analysis in real estate investment – next generation of energy efficiency assessment tools. Energy Effic 8:1015–1034. https://doi.org/10.1007/s12053-015-9331-9

Breiman L (2001) Random forests. Kluwer, Boston

Breiman L, Cutler A, Liaw A, Wiener M (2018) Random forest: Breiman and Cutler's Random Forest for classification and regression. https://cran.r-project.org/web/packages/randomForest/index.html. Accessed 12 Dec 2020

Buratti C, Barbanera M, Palladino D (2014) An original tool for checking energy performance and certification of buildings by means of artificial neural networks. Appl Energy 120:125–132

Burman E, Mumovic D, Kimpian J (2014) Towards measurement and verification of energy performance under the framework of the European directive for energy performance of buildings. Energy 77:153–163. https://doi.org/10.1016/j.energy.2014.05.102

Cajias M, Piazolo D (2013) Green performs better: energy efficiency and financial return on buildings. J Corp Real Estate 15:53–72. https://doi.org/10.1108/JCRE-12-2012-0031

Calì D, Osterhage T, Streblow R, Müller D (2016) Energy performance gap in refurbished German dwellings: lesson learned from a field test. Energy Build 127:1146–1158. https://doi.org/10.1016/j.enbuild.2016.05.020

Cao X, Dai X, Liu J (2016) Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. Energy Build 128:198–213. https://doi.org/10.1016/j.enbuild.2016.06.089

Carpino C, Mora D, de Simone M (2019) On the use of questionnaire in residential buildings. A review of collected data, methodologies and objectives. Energy Build 186:297–318. https://doi.org/10.1016/j.enbuild.2018.12.021

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y (2020) xgboost: extreme gradient boosting. https://cran.r-project.org/web/packages/xgboost/index.html. Accessed 12 Dec 2020

Ciulla G, D'Amico A (2019) Building energy performance forecasting: a multiple linear regression approach. Appl Energy 253:113500. https://doi.org/10.1016/j.apenergy.2019.113500

Cozza S, Chambers J, Patel MK (2020) Measuring the thermal energy performance gap of labelled residential buildings in Switzerland. Energy Policy 137:111085. https://doi.org/10.1016/j.enpol.2019.111085

Crawley J, Biddulph P, Northrop PJ, Wingfield J, Oreszczyn T, Elwell C (2019) Quantifying the measurement error on England and Wales EPC ratings. Energies 12:3523. https://doi.org/10.3390/en12183523

Czado C (2019) Analyzing dependent data with vine copulas: a practical guide with R. Lecture Notes in Statistics, vol 222. Springer, Cham

de Wilde P (2014) The gap between predicted and measured energy performance of buildings: a framework for investigation. AutomConstr 41:40–49. https://doi.org/10.1016/j.autcon.2014.02.009

Deutsche Energie-Agentur GmbH (2016) dena-Gebäudereport: Statistiken und Analysen zur Energieeffizienz im Gebäudebestand. https://www.dena.de/fileadmin/dena/Publikationen/PDFs/2019/dena-GEBAEUDEREPORT_KOMPAKT_2019.pdf. Accessed 12 Dec 2020

Deutscher Bundestag (2013) Novelle der Energieeinsparverordnung und des Energieeinsparungsgesetzes. https://energie-m.de/images/energie/EnEV-2013_lesefassung_2015-10-24.pdf. Accessed 09 Nov 2020

DIN e.V., BeuthVerlag (2016) DIN V 18599 - Energetische Bewertung von Gebäuden: Berechnung des Nutz-, End- und Primärenergiebedarfs für Heizung, Kühlung, Lüftung, Trinkwarmwasser und Beleuchtung, Ausgabe 2016, 6th edn. Beuth, Berlin

Eicker U, Zirak M, Bartke N, Romero Rodríguez L, Coors V (2018) New 3D model based urban energy simulation for climate protection concepts. Energy Build 163:79–91. https://doi.org/10.1016/j.enbuild.2017.12.019

Ettrich M (2008) Rechenverfahren im Wohnungsbau. https://www.regierung.oberbayern.bayern.de/imperia/md/content/regob/internet/dokumente/bereich3/energieeffizientesbauen/veranstaltungen/ettrich_rechenverfahren_wohnungsbau_18_07_2008.pdf. Accessed 26 Aug 2019

European Environment Agency (2019) Europe's state of the environment 2020: change of direction urgently needed to face climate change challenges, reverse degradation and ensure future prosperity. https://www.eea.europa.eu/highlights/soer2020-europes-environment-state-and-outlook-report. Accessed 7 June 2020

European Parliament and the Council (2002) Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings, vol 2002

Falbel D, Allaire JJ, Chollet F, RStudio, Google, Tang Y, van der Bijl W, Studer M, Keydana S (2020a) keras: R interface to 'Keras'. https://cran.r-project.org/web/packages/keras/index.html. Accessed 12 Dec 2020

Falbel D, Allaire JJ, RStudio, Tang Y, Eddelbuettel D, Golding N, Kalinowski T, Google (2020b) tensorflow: R interface to 'TensorFlow'. https://cran.r-project.org/web/packages/tensorflow/index.html. Accessed 12 Dec 2020

Federal Ministry for Economic Affairs and Energy (BMWi) (2018) Energieeffizienz in Zahlen: Entwicklungen und Trends in Deutschland 2018, Berlin

Federal Statistical Office of Germany (2011) Ergebnisse des Zensus 2011: Gebäude und Wohnungen sowie Wohnverhältnisse der Haushalte. https://ergebnisse.zensus2011.de/auswertungsdb/download?pdf=00&tableId=1&locale=DE&gmdblt=1. Accessed 5 Sep 2019

Foucquier A, Robert S, Suard F, Stéphan L, Jay A (2013) State of the art in building modelling and energy performances prediction: a review. Renew Sustain Energy Rev 23:272–288

Friedrichs F, Igel C (2005) Evolutionary tuning of multiple SVM parameters. Neurocomputing 64:107–117. https://doi.org/10.1016/j.neucom.2004.11.022

Gao Y, Ruan Y, Fang C, Yin S (2020) Deep learning and transfer learning models of energy consumption forecasting for a building with poor information data. Energy Build 223:110156. https://doi.org/10.1016/j.enbuild.2020.110156

Goldberg DE (2012) Genetic algorithms in search, optimization, and machine learning, 30th edn. Addison-Wesley, Boston

Gregor S, Hevner AR (2013) Positioning and presenting design science research for maximum impact. MIS Q 37(2):337–355

Greller M, Schröder F, Hundt V, Mundry B, Papert O (2010) Universelle Energiekennzahlen für Deutschland – Teil 2: Verbrauchskennzahlentwicklung nach Baualtersklassen. Bauphysik 32:1–6. https://doi.org/10.1002/bapi.201010001

Hardy A, Glew D (2019) An analysis of errors in the Energy Performance Certificate database. Energy Policy 129:1168–1178. https://doi.org/10.1016/j.enpol.2019.03.022

Heinisch O (1965) Cochran, W. G.: Sampling techniques. Biom J 7:203. https://doi.org/https://doi.org/10.1002/bimj.19650070312

Herrando M, Cambra D, Navarro M, La Cruz L, de, Millán G, Zabalza I, (2016) Energy performance certification of faculty buildings in Spain: the gap between estimated and real energy consumption. Energy Convers Manag 125:141–153. https://doi.org/10.1016/j.enconman.2016.04.037

Jovanović RŽ, Sretenović AA, Živković BD (2015) Ensemble of various neural networks for prediction of heating energy consumption. Energy Build 94:189–199. https://doi.org/10.1016/j.enbuild.2015.02.052

Kaymakci C, Wenninger S, Sauer A (2021) A holistic framework for AI systems in industrial applications. In: 16. Internationale Tagung Wirtschaftsinformatik

Ketter W, Peters M, Collins J, Gupta A (2015) Competitive benchmarking: an IS research approach to address wicked problems with big data and analytics. SSRN J. https://doi.org/10.2139/ssrn.2700333

Kraus D, Czado C (2017) D-vine copula based quantile regression. Comput Stat Data Anal 110:1–18. https://doi.org/10.1016/j.csda.2016.12.009

Kühl N, Hirt R, Baier L, Schmitz B, Satzger G (2020) How to conduct rigorous supervised machine learning in information systems research: the supervised machine learning reportcard [in press]. https://doi.org/https://doi.org/10.5445/IR/1000124438

Larsen M, Petrović S, Radoszynski AM, McKenna R, Balyk O (2020) Climate change impacts on trends and extremes in future heating and cooling demands over Europe. Energy Build 226:110397. https://doi.org/10.1016/j.enbuild.2020.110397

Li Y, Kubicki S, Guerriero A, Rezgui Y (2019) Review of building energy performance certification schemes towards future improvement. Renew Sustain Energy Rev 113:109244. https://doi.org/10.1016/j.rser.2019.109244

Menezes AC, Cripps A, Bouchlaghem D, Buswell R (2012) Predicted vs. actual energy performance of non-domestic buildings: using post-occupancy evaluation data to reduce the performance gap. Appl Energy 97:355–364. https://doi.org/10.1016/j.apenergy.2011.11.075

Metzger S, Jahnke, Katy, Walikewitz, Nadine, Otto M, Grondey A, Fritz S (2019) Wohnen und Sanieren: Empirische Wohngebäudedaten seit 2002 - Hintergrundbericht. https://www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2019-05-23_cc_22-2019_wohnenundsanieren_hintergrundbericht.pdf. Accessed 20 Oct 2019

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C-C, Lin C-C (2019) e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. https://cran.r-project.org/web/packages/e1071/index.html. Accessed 12 Dec 2020

Miratrix LW, Sekhon JS, Yu B (2012) Adjusting treatment effect estimates by post-stratification in randomized experiments. Wiley-Blackwell

Mosavi A, Salimi M, Faizollahzadeh Ardabili S, Rabczuk T, Shamshirband S, Varkonyi-Koczy A (2019) State of the art of machine learning models in energy systems, a systematic review. Energies 12:1301. https://doi.org/10.3390/en12071301

Müller O, Junglas I, Vom Brocke J, Debortoli S (2016) Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur J InfSyst 25:289–302. https://doi.org/10.1057/ejis.2016.2

Nagler T (2018) A generic approach to nonparametric function estimation with mixed data. Stat ProbabilLett 137:326–330. https://doi.org/10.1016/j.spl.2018.02.040

Nagler T (2018) Asymptotic analysis of the jittering kernel density estimator. Math Meth Stat 27:32–46. https://doi.org/10.3103/S1066530718010027

Nagler T (2019) vinereg: D-vine quantile regression. https://cran.r-project.org/web/packages/vinereg/index.html. Accessed 12 Dec 2020

Nagler T, Schepsmeier U, Stoeber J, Brechmann EC, Graeler B, Erhardt T, Almeida C, Min A, Czado C, Hofmann M, Killiches M, Joe H, Vatter T (2019) VineCopula: statistical inference of vine copulas. https://cran.r-project.org/web/packages/VineCopula/index.html. Accessed 12 Dec 2020

Nelsen RB (2010) An introduction to copulas, 2nd edn. Springer, New York

Neto AH, Fiorelli FAS (2008) Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. Energy Build 40:2169–2176. https://doi.org/10.1016/j.enbuild.2008.06.013

Olonscheck M, Holsten A, Kropp JP (2011) Heating and cooling energy demand and related emissions of the German residential building stock under climate change. Energy Policy 39:4795–4806. https://doi.org/10.1016/j.enpol.2011.06.041

Pan Y, Zhang L (2020) Data-driven estimation of building energy consumption with multi-source heterogeneous data. Appl Energy 268:114965. https://doi.org/10.1016/j.apenergy.2020.114965

Pasichnyi O, Wallin J, Levihn F, Shahrokni H, Kordas O (2019) Energy performance certificates – New opportunities for data-enabled urban energy policy instruments? Energy Policy 127:486–499. https://doi.org/10.1016/j.enpol.2018.11.051

Poel B, van Cruchten G, Balaras CA (2007) Energy performance assessment of existing dwellings. Energy Build 39:393–403. https://doi.org/10.1016/j.enbuild.2006.08.008

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x

Sendra-Arranz R, Gutiérrez A (2020) A long short-term memory artificial neural network to predict daily HVAC consumption in buildings. Energy Build 216:109952. https://doi.org/10.1016/j.enbuild.2020.109952

Seyedzadeh S, Pour Rahimian F, Oliver S, Rodriguez S, Glesk I (2020) Machine learning modelling for predicting non-domestic buildings energy performance: a model to support deep energy retrofit decision-making. Appl Energy 279:115908. https://doi.org/10.1016/j.apenergy.2020.115908

Shmueli K (2011) Predictive analytics in information systems research. MIS Q 35:553. https://doi.org/10.2307/23042796

Sutherland BR (2020) Driving data into energy-efficient buildings. Joule 4:2256–2258. https://doi.org/10.1016/j.joule.2020.10.017

Thrampoulidis E, Mavromatidis G, Lucchi A, Orehounig K (2021) A machine learning-based surrogate model to approximate optimal building retrofit solutions. Appl Energy 281:116024. https://doi.org/10.1016/j.apenergy.2020.116024

Töppel J, Tränkler T, Wiethe C (2019) The impact of energy-economical behavior on long-term energetic retrofitting road-maps: a vine copula quantile regression approach. In: Proceedings of 11th International Conference on Applied Energy, Part 1, Västerås

Touzani S, Granderson J, Fernandes S (2018) Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy Build 158:1533–1543. https://doi.org/10.1016/j.enbuild.2017.11.039

von Platten J, Holmberg C, Mangold M, Johansson T, Mjörnell K (2019) The renewing of Energy Performance Certificates – reaching comparability between decade-apart energy records. Appl Energy 255:113902. https://doi.org/10.1016/j.apenergy.2019.113902

Tsanas A, Xifara A (2012) Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Build 49:560–567. https://doi.org/10.1016/j.enbuild.2012.03.003

Wang S, Yan C, Xiao F (2012) Quantitative energy performance assessment methods for existing buildings. Energy Build 55:873–888. https://doi.org/10.1016/j.enbuild.2012.08.037

Wei Y, Zhang X, Shi Y, Xia L, Pan S, Wu J, Han M, Zhao X (2018) A review of data-driven approaches for prediction and classification of building energy consumption. Renew Sustain Energy Rev 82:1027–1047. https://doi.org/10.1016/j.rser.2017.09.108

Wirth R, Hipp J (2000) CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, London

You Q, Fraedrich K, Sielmann F, Min J, Kang S, Ji Z, Zhu X, Ren G (2014) Present and projected degree days in China from observation, reanalysis and simulations. ClimDyn 43:1449–1462. https://doi.org/10.1007/s00382-013-1960-0

Zhang R, Indulska M, Sadiq S (2019) Discovering data quality problems. Bus Inf Syst Eng 61:575–593. https://doi.org/10.1007/s12599-019-00608-0

Zhao H, Magoulès F (2012) A review on the prediction of building energy consumption. Renew Sustain Energy Rev 16:3586–3592. https://doi.org/10.1016/j.rser.2012.02.049