Energy Informatics

RESEARCH Open Access

On the surplus accuracy of data-driven energy quantification methods in the residential sector

Lars Wederhake[1], Simon Wenninger[2,3*], Christian Wiethe[2] and Gilbert Fridgen[4]

*Correspondence:
simon.wenninger@fim-rc.de

[1] credium GmbH, Katharinengasse 13, 86150 Augsburg, Germany
[2] Branch Business & Information Systems Engineering of the Fraunhofer FIT, Alter Postweg 101, 86159 Augsburg, Germany
[3] Research Center Finance & Information Management, University of Applied Sciences Augsburg, Alter Postweg 101, 86159 Augsburg, Germany
[4] SnT–Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg, 29 Avenue John F. Kennedy, 1855 Luxembourg, Luxembourg

Abstract

Increasing trust in energy performance certificates (EPCs) and drawing meaningful conclusions requires a robust and accurate determination of building energy performance (BEP). However, existing and by law prescribed engineering methods, relying on physical principles, are under debate for being error-prone in practice and ultimately inaccurate. Research has heralded data-driven methods, mostly machine learning algorithms, to be promising alternatives: various studies compare engineering and data-driven methods with a clear advantage for data-driven methods in terms of prediction accuracy for BEP. While previous studies only investigated the prediction accuracy for BEP, it yet remains unclear which reasons and cause–effect relationships lead to the surplus prediction accuracy of data-driven methods. In this study, we develop and discuss a theory on how data collection, the type of auditor, the energy quantification method, and its accuracy relate to one another. First, we introduce cause–effect relationships for quantifying BEP method-agnostically and investigate the influence of several design parameters, such as the expertise of the auditor issuing the EPC, to develop our theory. Second, we evaluate and discuss our theory with literature. We find that data-driven methods positively influence cause–effect relationships, compensating for deficits due to auditors' lack of expertise, leading to high prediction accuracy. We provide recommendations for future research and practice to enable the informed use of data-driven methods.

Keywords: Energy quantification methods, Data-driven methods, Building energy data, Data quality, Building energy performance, Prediction accuracy theory

Introduction

Efforts to mitigate climate change are timelier and more relevant than ever. The European Commission recently launched the "fit for 55" package to reduce net greenhouse gas emissions across sectors by at least 55% from 1990 levels by 2030 (European Commission 2021). With a share of 40% of energy consumption and 36% of energy-related greenhouse gas emissions, the building sector in the EU is a critical success factor to reach climate goals (European Commission 2020). In particular, residential buildings and their thermal energy for heating and hot water, with around 20% of the total energy use in Germany, offer great potential to reduce energy consumption and emissions (German

Federal Ministry for Economic Affairs and Energy 2018). Today's building stock comprises many older buildings adhering to less stringent construction codes. Correspondingly, these buildings have a high energy consumption. Low rates of new construction make this situation persist in the future, if there are not sufficient energetic retrofits per year (Deutsche Energie-Agentur GmbH 2016; Wenninger and Wiethe 2021). However, these energetic retrofits are too low to meet the EU's objectives (German Energy Agency 2018).

To improve the energy performance of buildings and increase the rate of energetic retrofits, the European Parliament and the Council passed a directive in 2002 already declaring the need for Energy Performance Certificates (EPC) (the European Parliament and the Council of the European Union 2002). EPCs are issued by qualified auditors and inform owners and occupants about the energetic condition of buildings, the associated operating costs, and recommendations for retrofitting (Arcipowska et al. 2014). Uncertainty is a major barrier to energetic retrofits, and thus the accurate assessment of building energy performance (BEP) and the expected energy savings from retrofits are crucial for EPCs (Amecke 2012; Walter et al. 2014; Ahlrichs et al. 2020). However, EPCs are yet under debate for their inadequate determination of BEP (Hardy and Glew 2019).

Today, auditors are bound to use legally prescribed engineering methods based on physical laws and anchored in norms and standards (e.g., DIN V 18599 in Germany) to quantify BEP (Zhao and Magoulès 2012a). These so-called engineering methods demand qualified auditors to conduct on-site visits to estimate physical building measures (e.g., thermal transmittance represented by the U-value of the building envelope's components) (Arcipowska et al. 2014). The apparent underlying assumption is that without these physical measures, BEP quantification is infeasible and that the average building occupant cannot assess these reliably.

Researchers study and propose data-driven methods to address issues with the accuracy of engineering methods, mostly relying on Machine Learning Algorithms (MLA) (Sutherland 2020; Bourdeau et al. 2019). In contrast to methods utilizing relationships from physical laws, MLA can learn from input data, which may also represent non-physical measures such as building age (Amasyali and El-Gohary 2018). Various studies compare engineering and data-driven methods with a clear advantage for data-driven methods regarding prediction accuracy for BEP. To that end, Wenninger and Wiethe (2021) found that data-driven methods exceed the engineering method used in Germany by almost 50% in terms of prediction accuracy for residential buildings.

So far, research has only investigated whether and how much data-driven methods perform better than engineering methods in terms of prediction accuracy of BEP (Wenninger and Wiethe 2021; Tsanas and Xifara 2012). However, it is unclear what reasons and cause–effect relationships lead to the often observed and reported superior prediction accuracy of data-driven methods. To further identify potentials for improvement in either class of methods, it is relevant to capture their underlying mechanics' embedded in the real process of issuing EPCs (as opposed to lab environments). That is what this piece of research targets. To do so, we develop and discuss a testable theory to identify reasons and causes why data-driven methods exhibit superior prediction accuracy.

For our contribution, we first provide theoretical and practical foundations about EPCs as well as about engineering and data-driven methods for BEP quantification in

"Literature" section. "Research method and study design" section then presents the study's design, before we derive method-agnostically cause–effect relationships for model quality, data quality, and output accuracy influencing the quantification of BEP in "Conceptual analytic model and research hypotheses" section. We further derive several design candidates based on data quality and model quality to account for the application of either engineering or data-driven methods and the conduct of on-site visits. The differentiation whether on-site visits are conducted or not allows us to investigate and consider the degree of expertise of the auditor issuing an EPC. We then qualitatively model the expected accuracy of the BEP prediction by the different design candidates before we evaluate our theory in "Evaluation" section and discuss it in the context of existing literature and case studies in "Discussion and implications" section. "Conclusion" section concludes the study.

With our work, we contribute to the scientific body of knowledge by giving insights on cause–effect relationships of data-driven methods that can compensate, e.g., for the lack of auditor expertise and lead to high prediction accuracy. We further provide recommendations to develop data-driven methods, enable the informed use of data-driven methods, and propose perspectives for further research.

Literature

Energy performance certificates

The introduction of EPCs dates to 2002 when the European parliament and council passed a directive that declared the need for EPCs to improve the BEP of the building stock (The European Parliament and the Council of the European Union 2002). EPCs are issued by qualified auditors and aim to inform owners, occupants, and property developers about the BEP (typically the annual final and primary energy consumption as well as associated carbon dioxide emissions), related operating costs, and recommendations for energetic retrofitting (Droutsa et al. 2016). EPCs further allow comparing different buildings' BEP in an energy efficiency ranking scheme independent of their location, year of measurement, and climate effects (Poel et al. 2007). EPCs for residential buildings thereby focus on space heating, water heating, and cooling (Arcipowska et al. 2014), which constitutes the biggest share of residential households' energy consumption [e.g., 85% of the final energy consumption of German residential households (Energieeffizienz in Zahlen 2018)]. Other energy flows like electricity consumption are not considered in EPCs as these are highly occupant dependent (Gram-Hanssen 2013).

Two options are available to derive the BEP of a building: one refers to the final energy demand (demand-oriented), while the other is based on the final energy consumption (consumption-oriented) (Arcipowska et al. 2014). The demand-oriented EPC reveals the energetic quality of a building with the above-mentioned engineering methods based on a technical analysis of manifold building parameters. Building parameters comprise building geometry, building type, condition of the heating system, and material of building components like walls or thermal insulation. The consumption-oriented EPC reports on the metered or data-related final energy consumption of a building, therefore implicitly including occupant behavior (Semple and Jenkins 2020). According to Arcipowska et al. (2014), no EU country relied exclusively on consumption-oriented EPCs, but either

on demand-oriented EPCs or both options. Thus, we also focus on demand-oriented EPCs.

Recent research discovered deviations between both options summarized under the phenomenon of the energy performance gap, which describes the phenomenon that the metered BEP differs significantly from the calculated BEP (Wilde 2014; Hertle et al. 2005). Studies on the energy performance gap depict deviations of up to 287% across Europe (Wilde 2014; Calì et al. 2016). Given that EPCs are supposed to provide a reliable basis for decision-makers (Arcipowska et al. 2014), these strong deviations are unacceptable, as uncertainty and incomplete information are substantial investment barriers in energy efficiency (Amecke 2012). Therefore, many studies try to find causes and solutions to minimize the energy performance gap (Burman et al. 2014; Herrando et al. 2016; Menezes et al. 2012). In terms of the energy quantification methods used to determine the BEP, the current methods can either be gradually improved (Zhao and Magoulès 2012a; Bigalke and Marcinek 2016) or completely replaced by more accurate methods, e.g., data-driven methods (Wenninger and Wiethe 2021; Foucquier et al. 2013; Deutscher Bundestag 2013), to minimize the energy performance gap. The potentials of data-driven methods are expected to originate especially from the first two steps of creating EPCs. The creation of EPCs typically follows three generic process steps (Hardy and Glew 2019; Li et al. 2019). First, the auditor collects necessary input data (Hardy and Glew 2019) during on-site visits to ensure a high level of data quality (Arcipowska et al. 2014). The collected data may stem from, e.g., photos, plans, or sketches and are then converted and pre-processed for the following calculations. Second, the auditor calculates the EPC's target value, the BEP, and identifies possible retrofitting recommendations with the prescribed engineering methods, mostly implemented in software tools (Hardy and Glew 2019; Li et al. 2019). Third, the auditor prepares the EPC document and presents the results (Pasichnyi et al. 2019).

Energy quantification methods

Quantifying BEP is challenging since multiple influencing factors like building geometry, occupancy behavior, thermal properties, or weather must be considered to achieve accurate results (Wei et al. 2018). Today's energy quantification methods (EQMs) can be categorized, as mentioned before, into engineering and data-driven EQMs. A third type is hybrid EQMs, which combine the former two types (Amasyali and El-Gohary 2018; Foucquier et al. 2013; Borgstein et al. 2016). Hybrid EQMs require a deep knowledge of both types of EQMs and are computationally inefficient, posing a major challenge and making them less attractive (Wei et al. 2018; Coakley et al. 2014). Thus, not being in that niche, we focus on engineering and data-driven EQMs in our study. The purpose and level of detail of EQMs can vary widely (Wang et al. 2012). For instance, different prediction periods, building types, or the type of predicted energy consumption can be distinguished (Amasyali and El-Gohary 2018). Besides quantifying annual BEP for EPCs, the predicting granular future energy consumption for the coming (quarter)hours or days is a relevant field of research and in practice allowing energy management systems to optimize operation (Qiao et al. 2021). Engineering EQMs differ strongly in complexity and accuracy from detailed computational fluid dynamic models to simplified lumped parameter models (Foucquier et al. 2013; Andrade-Cabrera et al. 2018). The thermal

behavior of heat flows in buildings sets the basis for these physical models (Foucquier et al. 2013). To calculate the BEP for heating (or cooling), transmission heat losses through the building shell, ventilation heat losses, solar heat gains, and internal heat gains are part of the physical models (Ettrich 2008). Thus, to depict the BEP, all these heat flows must be considered and derived. In the EU, quasi-steady-state EQMs are typically used (Semple and Jenkins 2020; Eicker et al. 2018). For an accurate calculation of the heat flows, the correlations of all input parameters must be considered. Input parameters include detailed information about building location, building geometry, materials used (e.g., insulation or masonry material), and the heating system(s). As a result of the high demand for information and the sophistication of the computation, software is commonly used to carry out demand-oriented EPCs (Foucquier et al. 2013). Since documentation is rare (in existing older buildings), and test drillings are cost-intensive, collecting the necessary information (such as materials' heat transmission coefficients) by identifying the materials used as well as the isolation thickness is challenging, time-consuming, and costly. Thus, for engineering EQMs, the data quality necessary for accurate calculations is a centrally limiting factor. Therefore, the research concludes that engineering EQMs may be more appropriate in the design phase of buildings than assessing the BEP of existing buildings (Qiao et al. 2021).

Data-driven EQMs, in contrast, exclusively rely on data, requiring no expertise of physical phenomena to describe thermal behavior (Foucquier et al. 2013). Instead, by learning from correlations between input and output parameters, the underlying model builds the knowledge to predict BEP for unknown buildings with new data. Regarding the historical data, these may only comprise easily collectible building and energy consumption or demand data, as well as information on implemented energy retrofit measures. Several EQMs exist for data-driven prediction of BEP. The following data-driven EQMs for the prediction and classification of BEP are the most popular (Wei et al. 2018): *Artificial neural networks (ANN), Support vector machines (SVM), statistical regressions*, and *decision tree genetic algorithms.* Thereby, ANNs and SVMs are particularly well suited to predict BEP, although requiring a large amount of high-quality data to make accurate predictions (Zhao and Magoulès 2012b; Kaymakci et al. 2021). Thereby ANNs are computationally less intensive at runtime than SVMs (Wei et al. 2018). On the other hand, Wenninger and Wiethe (2021) found that after reasonably hyperparameter optimization and thorough training, ANNs, SVMs, Random Forests, Extreme Gradient Boosting, and D-Vine Copulas perform comparably well. In addition, emerging methods from the data science disciplines gain momentum in energy research. Researchers seek improvements in prediction accuracy using deep neural networks or ensemble learning approaches combining several data-driven EQMs (Qiao et al. 2021). Thereby, ensemble learning assumes that more diverse compositions of models lead to more diverse errors that may cancel each other out, leading to higher prediction accuracy (Polikar 2006). Next to enhancing prediction accuracy, a new stream of explainable artificial intelligence analysis aims to shed light on the black box phenomenon of data-driven EQMs and derive insights into buildings' energy behavior (Wenninger et al. 2022a).

Literature reports clear advantages in prediction accuracy for data-driven EQMs when comparing engineering and data-driven EQMs. Deb and Schlueter (2021) state that the accuracy surplus for data-driven EQMs results from training on data with no prior

defined model structure. Wenninger and Wiethe (2021) benchmarked several data-driven EQMs against the status-quo engineering EQM for EPCs in Germany and found that all data-driven EQMs tested exceeded the engineering EQM by almost 50% in prediction accuracy. Further, various case studies on different BEP prediction tasks reveal high prediction accuracy for data-driven EQMs (Tsanas and Xifara 2012; Li et al. 2010; Fernandez et al. 2011). Remarkably though, literature focuses only on the prediction performance, i.e., whether and how much data-driven EQMs perform better than engineering EQMs (Wenninger et al. 2022b). However, it is not clear why data-driven EQMs often depict higher prediction accuracy in BEP studies. To identify further potential improvements for both engineering and data-driven EQMs, it is important to address this research gap and capture the underlying mechanics embedded in the real-world process of predicting BEP and issuing EPCs outside the laboratory.

Research method and study design

The process of issuing EPCs involves both a technological (the software for BEP quantification) and a social system (auditor collecting data and working with software), which interact with one another. According to Lee (2001), socio-technical systems are quite uniquely the unit of analysis of the information systems (IS) discipline. As a subdiscipline of IS, based on Watson et al. (2010), Energy Informatics is concerned with these systems in the energy domain, including BEP quantification for EPCs. Epistemologically, we thus follow and refer to research from these fields to develop a theory on the design of EQMs for the BEP quantification for EPCs. With that as one goal of this study, we refer to Doty and Glick (1994) for the epistemology, stating that there is apparent consensus among the theory-building experts that the minimal definition of a theory must comply with three primary criteria:

1. Constructs must be identified,
2. Relationships among these constructs must be specified; and
3. These relationships must be falsifiable.

In this sequence, a theory can be viewed as a set of "statements of relationships among constructs that can be tested" (Gregor 2006). As a prominent example, Davis (1985) follows this sequence of steps unfolding his theory on technology acceptance. Constructs in this field, for example, are perceived usefulness and perceived ease-of-use representing user beliefs, as well as technology acceptance behavior. Davis stipulates that there is a specific relationship among these constructs so that both former constructs predict higher rates of technology acceptance (latter construct). These relationships are formulated so that they can always be tested empirically (e.g., Davis 1989) and thus potentially be falsified.

Turning to the study design for the theory on BEP prediction as the main contribution of this research, we similarly follow Doty and Glick (1994), who prescribed the aforementioned primary criteria to establish a theory. Additionally, this study strives to go beyond mere prediction by characterizing causal explanations in line with, e.g., Bhattacherjee and Premkumar (2004).

To do so, we follow Niehaves and Ortbach (2016) and gradually develop our predictive-explanatory theory. First, we form an inner structural model, and second amend this inner model by outer constructs and relationships, i.e., the outer model. Introducing the process of issuing EPCs, we characterize the inner model, which is concerned with data quality, model quality, and (output) accuracy. These serve as constructs while the three essential cause–effect relationships (CER1, CER2, CER3) between those constructs define the relationships. We describe all CERs in detail in "Cause–effect relationships between model quality, data quality, and output accuracy" section.

In "Design candidates based on data quality and model quality" section, we extend the inner model by the constructs of the outer model. The outer model falls into a design and a measurement model. The design model describes the design options of energy performance assessments (EPA), in particular the EQM. Design options influence the constructs of the inner model. The combinations of choices for the design options make up concrete design candidates. The measurement serves to evaluate the different design candidates on output accuracy as the outcome of the cause–effect relationships. Thereby, the measurement model allows for the requirement of a theory to be falsifiable.

In "Characterizing output accuracy of the design candidates" section, we then systematically characterize the functional relationship between these types of auditors, their expertise (as one design option), importantly the EQM (as the other design option), and the (output) accuracy. Thus, we analyze the design candidates backed by the CERs.

Subsequently, in "Conceptual analysis of design candidates" section, we present design candidates and derive their suggested preferability over others regarding (output) accuracy based on a conceptual analytic model, i.e., the design candidates' relative position to one another in terms of (output) accuracy. The conceptual analytic model follows from the constructs and relationships described in "Cause–effect relationships between model quality, data quality, and output accuracy", "Design candidates based on data quality and model quality" and "Characterizing output accuracy of the design candidates" sections.

Finally, in "Summary of the analysis of the design candidates" section, we summarize and illustrate the inner and outer model according to Niehaves and Ortbach (2016) so that each relationship can be tested individually and the proposed predictive-explanatory theory as a whole. In this study, we refer and cite existing research supporting and validating the concepts as part of our evaluation in "Evaluation" section following evaluation guidance by Sonnenberg and Vom Brocke (2011) and on the evaluation criteria specified by March and Smith (1995).

Conceptual analytic model and research hypotheses

As mentioned in "Energy performance certificates" section, creating EPCs covers three high-level generic process steps: data gathering, BEP quantification, and assessment presentation. In theory, quantifying BEP (for EPCs) is an information-only process that would not necessarily involve physical actions such as on-site visits. However, on-site visits are still highly encouraged in practice, even if they are not compulsory (Arcipowska et al. 2014). If conducted, on-site visits occur during the first step of this process, while the other two steps are performed remotely. Since on-site visits add cost, regulators who enforce them and auditors who willingly perform them expect larger benefits than the thereby incurred costs. Regulators otherwise fear future costs of low-accuracy

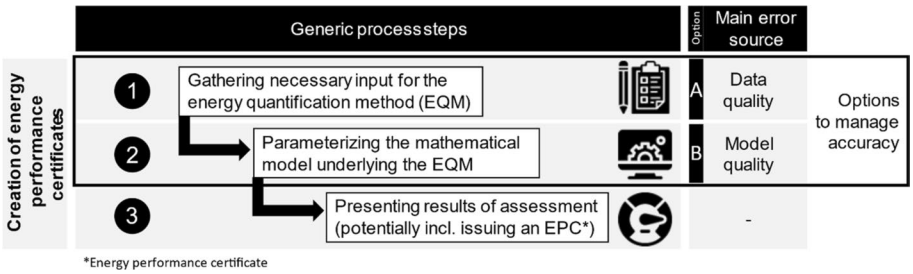

Fig. 1 Process of EPC creation and options to manage accuracy

assessments, e.g., economically inefficient retrofit decisions (Heo et al. 2012). Auditors, similarly, fear future costs from low-accuracy assessments because of revisions and penalties (Arcipowska et al. 2014). Irrespective of local policies, we find two options to manage accuracy:

(A) *Data quality*, e.g., by changing what and how to gather input data, and
(B) *Model quality*, e.g., by choosing an EQM and configuring its underlying model for BEP quantification.

Figure 1 illustrates the generic process steps and the two options to manage accuracy (highlighted by the frame).

Cause–effect relationships between model quality, data quality, and output accuracy

Since option (A) and option (B) are subtly linked to output accuracy, we derive three theoretical cause–effect relationships (CERs):

- CER1: Higher levels of *model quality* lead to higher levels of *output accuracy*, i.e., BEP prediction accuracy
- CER2: Higher levels of *data quality* lead to higher levels of *output accuracy*
- CER3: Higher levels of *data quality* during model design lead to higher levels of *model quality*

CER1: Model quality refers to the degree of how well a model captures real-world relationships, i.e., theoretical accuracy under perfect information. For example, computational fluid dynamic simulations are known for their high theoretical accuracy (Zhao and Magoulès 2012a). In contrast, existent EQMs based on linear (regression) models, e.g., the US ENERGY STAR rating, are despised for capturing relationships insufficiently (Papadopoulos and Kontokosta 2019).

CER2: Data quality describes the fit of data for its purpose (Klobas 1995), making it context-specific (Strong et al. 1997). In the context of EPCs, the purpose of the input data is to quantify BEP as accurately as possible. To assess the fit of data, quality is considered multi-dimensional (Pipino et al. 2002), with *completeness* and *accuracy* being particularly vital in this context (Wang and Strong 1996). Incomplete data are imperfect data, especially when at least one attribute must be replaced by surrogate measures. Inaccurate data (if quantitative) or mislabeled data (if qualitative), next to any other data quality issue, also leads to imperfect data. Generally, imperfect input data cannot generate more

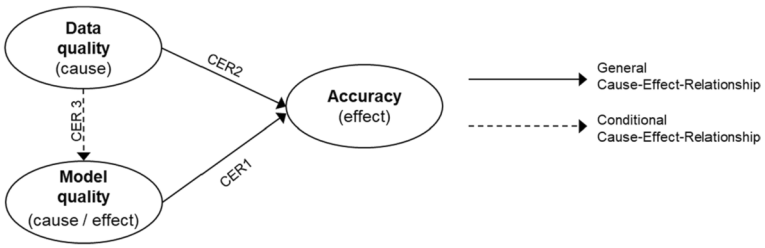

Fig. 2 Cause–effect relationships explaining the accuracy of EPCs

accurate results than perfect input data for any deterministic method without the model having a systematic error, i.e., bias (Schwarz and Köckler 2011; Bevington 1969).

CER3: Considering the influence of data quality on model quality, it is important to distinguish between engineering and data-driven EQMs. As the engineering EQMs are based on physical laws, human behavior and inaccuracies when collecting input data are exogenous to the models reviewed (Zhao and Magoulès 2012a; Foucquier et al. 2013; Mathew et al. 2015). Thus, they not only work best when input data is perfect but are also adversely affected by error propagation (Schwarz and Köckler 2011; Fornasini 2008). In contrast, data-driven EQMs require their underlying model to be trained on data. For data-driven EQMs, data quality can also affect the model quality and thus, in addition to CER2, output accuracy (Kaiser et al. 2022). The model training corresponds to optimizing an objective function over all observations. Thus, the more accurately the input data resembles reality (all possible observations), the more likely it is that the model is of high theoretical output accuracy, while a model that is trained on lower-quality data will have lower theoretical accuracy (Bi and Zhang 2005). However, training a model on inaccurate but unbiased data generates models more robust to lower data quality because parameters bearing greater uncertainty are weighted less (e.g., Pregenzer et al. 1994). In this context, adaptability corresponds to handling varying degrees of uncertainty. Data-driven models exhibit a certain degree of adaptability to data uncertainty. We sketch the considered CERs in Fig. 2.

We denote CER3 as a dashed line, as it has practically no influence in cases where the EQM's underlying model is purely based on physical laws. Thus, we consider this a conditional CER as the relationship is conditional to the model in place.

Design candidates based on data quality and model quality

Having established that data and model quality are central influencing factors on accuracy while simultaneously influencing one another, we derive the design candidates based on the possible combinations of both. To this end, we assume binary choices for both to keep the analyses tractable.

Data quality Several different parameters affect data quality in EPCs. First, the expertise of auditors is an important determinant of EPC's quality, as in most countries, qualifications and often an accreditation is required (Arcipowska et al. 2014). Second, there is the possibility of collecting input data through on-site visits by a qualified auditor. Third, next to expertise, previous research found that auditors' time constraints seem to limit output accuracy (Pasichnyi et al. 2019). We opt to use the design options "qualified

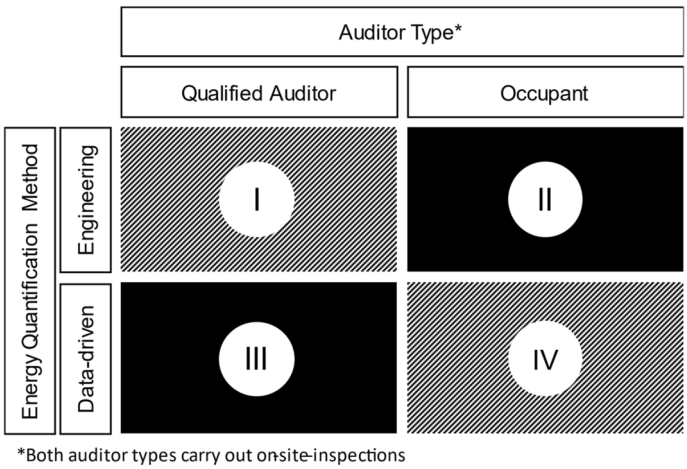

Fig. 3 Design candidates for EPAs

auditor" and "occupant" for data quality for our theoretical analysis. Without training, the average occupant might lack the expertise and tools to achieve the same data quality as a qualified auditor, which is consequently reflected in lower data quality (Claesson 2011). We assume both occupants and qualified auditors to carry out on-site visits.

Model quality There is the choice of the EQM. As outlined before, however, EQMs are not only distinguishable by theoretical accuracy but react differently to data uncertainty from imperfect data quality depending on the type, i.e., engineering or data-driven EQM. For our analyses, we choose engineering and data-driven EQMs as overarching design candidates for model quality without committing to any particular EQM.

This leaves the four conceptual abstract design candidates for our study representing the design model, namely: (I) a qualified auditor using an engineering EQM, (II) an occupant using an engineering EQM, (III) a qualified auditor using a data-driven EQM, and (IV) an occupant using a data-driven EQM. Figure 3 summarizes all four conceivable design candidates.

For the measurement model, we introduce (statistical) performance evaluation measures (PEM) required to meaningfully compare the BEP prediction performance of the different design candidates (Amasyali and El-Gohary 2018). PEMs quantify the goodness of fit for predictions against a ground truth, i.e., the predicted BEP using EQMs against the actual BEP. Some studies evaluate the impact of building parameters on BEP by using sensitivity or variable importance analyses (Ali et al. 2020; Yuan et al. 2019). In the following sections, we use the term "output accuracy" to describe BEP prediction performance, rather than restricting it to a specific PEM, such as the Coefficient of Variation (CV) most commonly used in BEP prediction studies (Amasyali and El-Gohary 2018). Higher values for output accuracy indicate better BEP prediction performance.

Characterizing output accuracy of the design candidates

As mentioned, we use output accuracy as the key measure to compare EQMs. Analogous to CER2 for model training, the output accuracy for model prediction also depends on data quality. When measuring data uncertainty using one single aggregated measure, we can then describe the output accuracy of one EQM as a function of uncertainty in

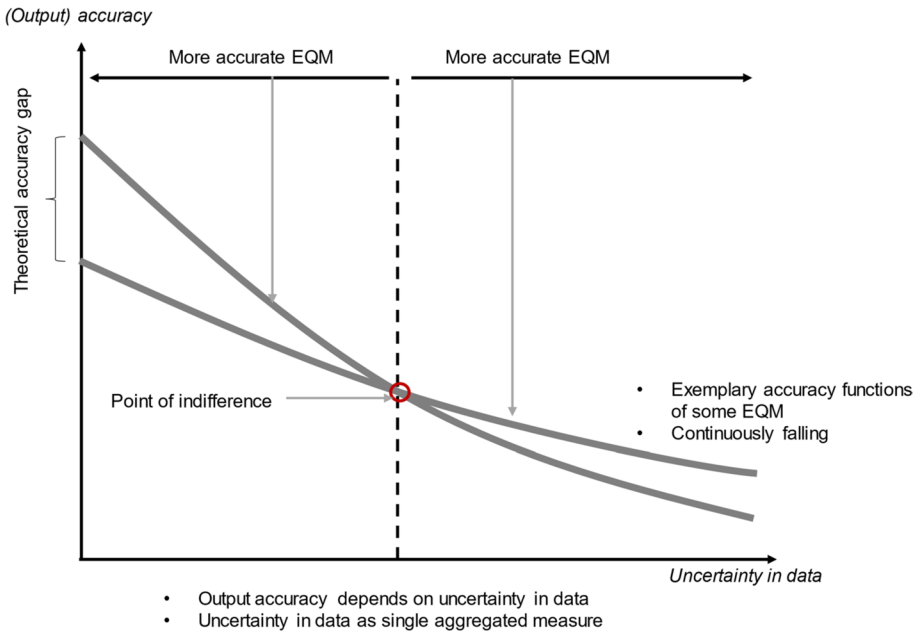

Fig. 4 Sketch of the conceptual analytic model

data. More precisely, let $\xi \in [0, 1]$ denote the underlying uncertainty in the data, whereby 1 corresponds to perfect data and 0 to random noise. We can then define a function $f^i(\xi) \rightarrow \mathbb{R}^+$ mapping the uncertainty in data ξ to the respective output accuracy in $\mathbb{R}^+$ for a specific EQM i and PEM. It holds that for $\xi = 1$ this function delivers the EQM's theoretical output accuracy, which, by CER1, can be made arbitrarily accurate. From CER2, however, it follows that with growing uncertainty, an EQM's output accuracy declines monotonously, i.e., for all EQM i. Finally, CER3 suggests that the slope of the function varies depending on the type of the EQM, i.e., for two different EQMs $i \neq j$, it may hold that $\frac{\partial f^i}{\partial \xi} \neq \frac{\partial f^j}{\partial \xi}$. We assume that engineering EQMs dispose of higher theoretical output accuracy, while data-driven EQMs dispose of higher output accuracy for high uncertainty in data as they are less susceptible to data quality problems. By further considering continuous functions, we can apply the mean value theorem and derive that there exists a fixed $\xi^* \in [0, 1]$ for which the corresponding functions of the two types of EQMs must intersect, i.e., $f^i(\xi^*) = f^j(\xi^*)$.[1] The resulting point of intersection indicates a point of indifference, i.e., where both EQMs deliver the same accuracy given a level of uncertainty in data.[2] Figure 4 sketches these conceptual analytical relationships visually for two EQMs.

We partition the continuous scale of uncertainty and then associate certain types of auditors (design option), or more generally, types of human agents, with the then divided segments of uncertainty. For our analyses, the differentiation between qualified auditors without time constraints, qualified auditors with time constraints, occupants

[1] Note, that continuity is a necessity for the mean value theorem yet must not necessarily hold. However, even for discontinuous function we can still determine an arbitrarily small ε-neighborhood around a fixed ξ^* where the comparative advantage in output accuracy switches.

[2] However, the point of intersection might also rest below the horizontal axis. Then, clearly, there is a dominant and an inferior EQM. Note that an EQM intersecting the horizontal axis is considered zero thereafter.

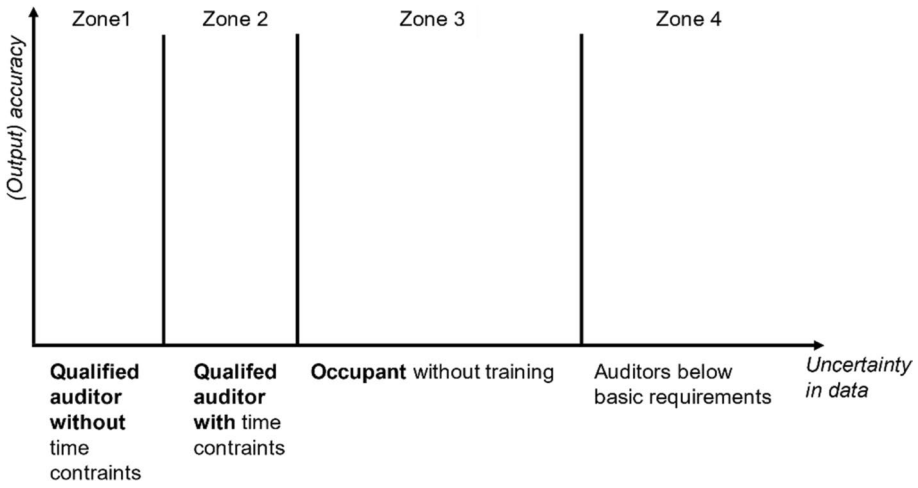

Fig. 5 Output accuracy-uncertainty plane partitioned in zones by the uncertainty levels of the four types of human agents

without training, and auditors below basic requirements is relevant. We assume these types of agents to produce results in decreasing quality, as depicted in Fig. 5. In that vein, auditors below basic requirements provide the lowest output accuracy. We describe the divided segments of uncertainty as zones listed from 1 to 4 as follows:

- Zone 1 corresponds to uncertainty levels in data that can only be reached if qualified auditors perform EPAs without time constraints.
- Zone 2 represents the continuum of what might be expected by qualified auditors with time constraints concerning uncertainty in data.
- Zone 3 represents a larger continuum of what might be expected by occupants concerning uncertainty in data. We allow an occupant without training but with basic general capabilities and knowledge to represent this type of human agent. This may be any person capable of keeping the house on one's own.
- Zone 4 corresponds to an uncertainty level below what can typically be expected by occupants.

This is relevant as the intersection between the (output) accuracy function of any two EQMs and its position, i.e., the zone in which the point of intersection falls, guides decision-making on what type of EQM serves a specific type of human agent best.

Conceptual analysis of design candidates

This subsection presents the different possible graphs of the output accuracy functions for the data-driven and engineering EQMs, building on the previous sections. As mentioned in "Literature" and "Cause–effect relationships between model quality, data quality, and output accuracy" sections, literature describes the theoretical output accuracy of engineering EQMs higher than for data-driven EQMs. At the same time, data-driven EQMs can weigh down imprecise input variables, making them less susceptible to low data quality. Therefore, we argue that an engineering EQM provides better results under perfect data quality, while a data-driven EQM provides better output accuracy

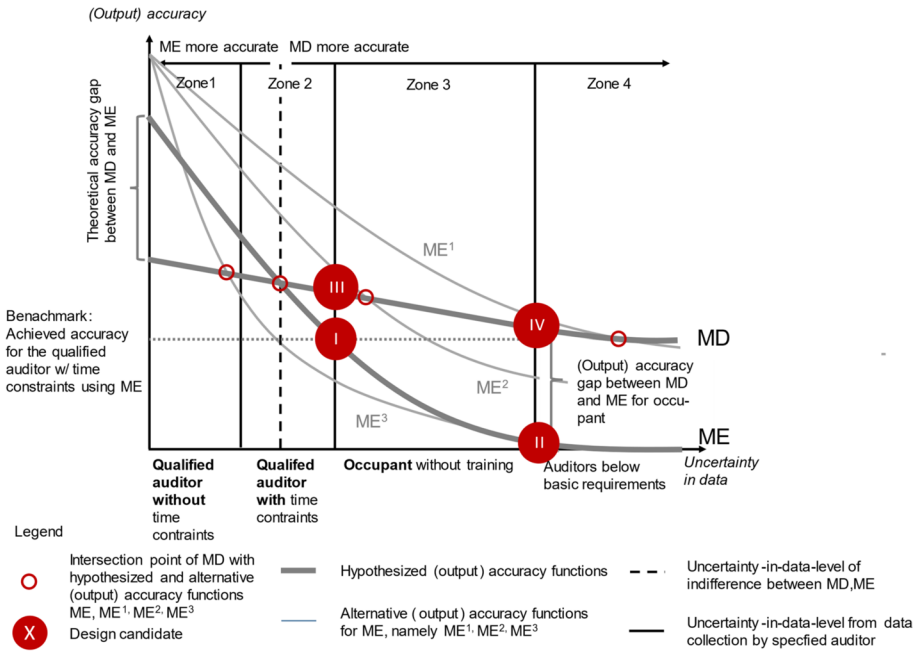

Fig. 6 Conceptual analysis of the abstract design candidates (cf. "Design candidates based on data quality and model quality" section)

under high data uncertainty. This leaves us with the question of where the two functions intersect.

First, we let MD denote the output accuracy function of a data-driven EQM. Then, we introduce four output accuracy functions each intersecting with MD in a different of the four zones. We let ME^4 denote the engineering EQM intersecting MD in zone 4, ME^3 the engineering EQM intersecting with MD in zone 3, ME^2 the engineering EQM intersecting with MD in zone 2, and ME^1 the engineering EQM intersecting in zone 1. In particular, if ME^4 was the valid hypothesis for the output accuracy function, then the engineering EQM would dominate the data-driven EQM for all considered types of auditors, i.e., all considered types of human agents should use the engineering EQM. Likewise, if ME^1 was the valid hypothesis for the output accuracy function, then all considered types of auditors should use the data-driven EQM. However, the interpretation is less clear when the point of intersection lies in zones 2 or 3. Assuming that it was in zone 3, qualified auditors with time constraints would be advised to use the engineering EQM as they currently do. However, there are reasons to be skeptical to that end: first, previous research has validated that some data-driven EQMs can infer parameters very well from non-physical building attributes (Berger et al. 2016). Second, while a qualified auditor might use (advanced) mechanical tools on-site to better assess parameters like heat transmittance of the building envelope, e.g., by thermography, there is some evidence that this is only done for a surcharge (Fox et al. 2016). This raises the question of why the auditor should do this if there is already sufficient certainty under limited time and effort. Both arguments suggest that such auditors might need more detailed information than standard on-site visit procedures would allow. This, in turn, indicates that even a qualified auditor with time constraints could prefer a data-driven EQM. Figure 6 depicts the different output functions with ME^2 as the hypothesized output function in contrast

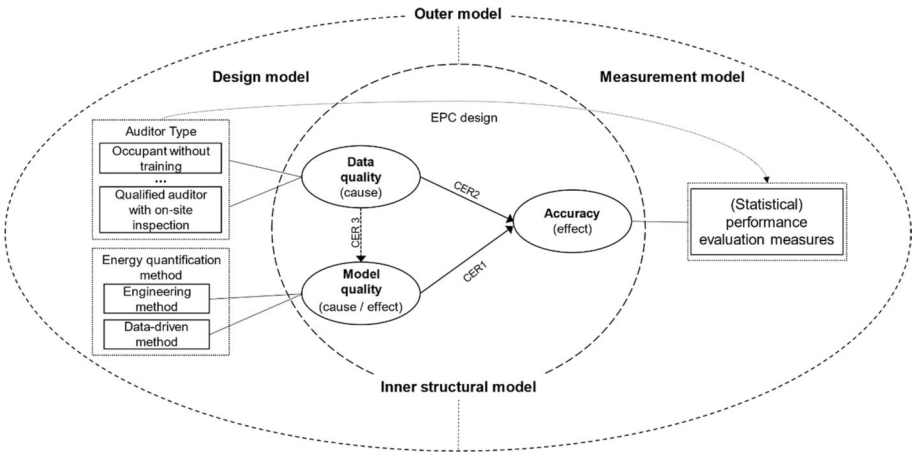

Fig. 7 Testable theory for EPCs based on Niehaves and Ortbach (2016)

to ME^4, ME^3, and ME^1 serving as alternative (output) accuracy functions. However, a rigorous validation of this hypothesis is not feasible analytically, but only empirically. As of that reason, we leave the zone in which there is the point of intersection as a hypothesis for now and refer to the empirical validation to "Evaluation" section. However, if this hypothesis holds, we can then infer the zonal positions of the design candidates (I), (II), (III), and (IV). We decide to place the design candidates at the conservative end of the zonal spectrum for reasons of consistency, while theoretically they might be placed anywhere along their output accuracy function within that zone. Nonetheless, design candidates referring to the same human agent must feature the same level of uncertainty, i.e., they must be placed on the same vertical line. We find that candidates (I) and (III) feature the same level of uncertainty in data, whereas candidate (III) will presumably have a higher output accuracy. Similarly, we find that candidates (II) and (IV) feature the same level of uncertainty in data, whereas candidate (IV) is considered to have significantly higher output accuracy. Also, we see that candidate (IV) is positioned above candidate (I) on the (output) accuracy axis, while yet being exposed to more uncertainty in data. We present the conceptual analysis as a testable theory in Fig. 6.

Summary of the analysis of the design candidates

Based on the conceptual analysis and previous sections, we summarize and illustrate our findings as a testable theory with the inner and outer model according to Niehaves and Ortbach (2016) illustrated in Fig. 7. The model consists of an outer and an inner structural model, whereby the outer model divides into the previously derived design and measurement model. The design model specifies the design candidates with the two design options of the auditor type and the EQM. The measurement model allows testing different combinations of design options, i.e., design candidates. For this purpose, PEMs allow evaluating the results with respect to the BEP prediction performance. The inner model contains the two options to manage accuracy in EPCs, data and model quality, as well as their relationships via CERs. The design option auditor type strongly influences the data quality, and the design option EQM the model quality.

The developed theory represents the central result of our work and provides insights into CERs of EQMs, aiming to explain status quo prediction performances of different EQMs used in practice and research. By allowing us to influence individual design options, which in turn affect data and model quality, our theory, in conjunction with the PEMs embodied in the measurement model, enables us to (empirically) test the accuracy functions and related hypotheses stated in "Conceptual analysis of design candidates" section. In "Evaluation" section, we validate and evaluate as far as possible the developed theory based on literature.

Evaluation

Evaluation design

Evaluation is a crucial step when developing artifacts to demonstrate the relevance of an artifact for practice and research (Sonnenberg and Brocke 2011). For our developed theory, evaluation is important to ensure that the theoretical insights regarding the CERs and the influence of design options on BEP prediction performance reflect the real world in the best possible way. Only then, we can derive robust and valid implications for practice and research. For this purpose, artifact evaluation usually involves the definition of evaluation criteria, which measure the degree to which a developed artifact achieves its goal or its quality. March and Smith (1995) propose the evaluation criteria completeness, fidelity with real-world phenomena, internal consistency, level of detail, and robustness for models and theories. We aim to draw on all five evaluation criteria. To this end, we apply logical reasoning supported by first evidence from existing research.

Evaluation results

Here, we present the evaluation results for both the components of our theory and our theory as a whole.

Completeness: In terms of the completeness of our theory, we consider the inner structural model and the design and measurement models that form the outer model. First, and regarding the inner structural model, we derived from literature that data quality and model quality are ultimately the only options to manage accuracy in issuing EPCs. We subsume the expertise and possible time constraints under which auditors issue EPCs. Therefore, we argue that for each EQM, there are only these two options to influence accuracy. To this end, the inner structural model is complete, even though there could be other preceding influencing factors such as available training programs and documentation for auditors, which are not captured in our theory. Second and regarding the design model, we restricted our study to four possible design candidates, even if there are multiple types of auditors and EQMs. Since our theory should be understandable and comprehensible, it is only possible to depict a well-defined range of real-world options. Thus, for both design options, we use terms established in literature. The distinction between data-driven and engineering EQMs, also often referred to as black-box and white-box EQMs, is the standard in literature (Amasyali and El-Gohary 2018). For the sake of completeness, it should be mentioned that so-called hybrid EQMs, which combine data-driven and engineering EQMs, are also available but are not widely used due to some drawbacks (Wei et al. 2018). Regarding the two auditor types, Wenninger and Wiethe (2021) already distinguished between occupants and qualified auditors in

an EQM benchmarking study. We further divided a theoretically continuous spectrum into four types of auditors for the evaluation of uncertainty in data. Even if there might be additional or differently grouped types of auditors, we argue that the split of auditor types is not crucial for the underlying principles of our theory. E.g., further levels in the quality of data recording could be achieved if, for example, house residents or auditors are further enabled to record the data in a quality-assured manner within the given time restrictions. We, therefore, consider our design model to be complete. Third and regarding the measurement model, we referred to common PEMs. Regarding completeness, we distinguish between an evaluation based on literature and an evaluation based on an EPC user perspective, e.g., occupants. Based on literature, our evaluation is complete since we use established PEMs in our measurement model. From an EPC user perspective, it might be relevant for individual EPCs to specify the risk of BEP misprediction for informed energy efficiency investments (Rockstuhl et al. 2021). In contrast to the PEMs we use, additional information such as the probability of deviation from the BEP prediction in addition to mean values would be useful. In the case of data-driven EQMs, quantile regression could solve this task, for example. An adaptation of the PEMs and the use of EQMs that can provide such information would make sense in the context of an EPC user perspective.

Fidelity with real-world phenomena: Regarding fidelity with real-world phenomena and our findings derived in "Conceptual analysis of design candidates" section, we rely on and refer to already existing empirical research instead of replicating their findings in this piece of re-search. Even if findings across studies trivially differ due to different EQMs applied and datasets used, we identify a pronounced trend: particularly helpful in this context are studies that create comparability between results by simultaneously examining both engineering EQMs with officially collected parameters and data-driven EQMs with parameters collected by building occupants. To evaluate our theory, we rely on a study by Wenninger and Wiethe (2021). They investigated how well the established data-driven EQMs ANN, D-vine copula quantile regression, Extreme Gradient Boosting (XGB), Random Forest (RF), and Support Vector Regression (SVR) perform in comparison to the legally required engineering EQM for issuing EPCs. Note that the engineering EQMs were performed by qualified energy auditors, whereas for the data-driven EQMs, homeowners, i.e., non-experts, collected fewer and simpler data. Figure 8 shows the study results for the engineering and data-driven EQMs in different building age classes and living spaces. The prediction accuracy on the y-axis is represented by the CV, with smaller values indicating higher prediction accuracy. We see that all data-driven EQMs exceed the engineering EQM regarding prediction accuracy and nearly halve the prediction error. For further details, we refer to Wenninger and Wiethe (2021).

Thus, from an evaluation perspective of our theory, first, we can confirm that the accuracy functions of data-driven and engineering EQMs differ. Second, the output accuracy functions displayed in Fig. 6 are likely to differ from the findings of Wenninger and Wiethe (2021). Since they tested design candidates I and IV with almost twice the accuracy for design candidate IV, MD's accuracy for uncertainty in data of design candidate IV should significantly exceed ME's accuracy for uncertainty in data of design candidate I. We conclude that, while our theory generally supports the findings of

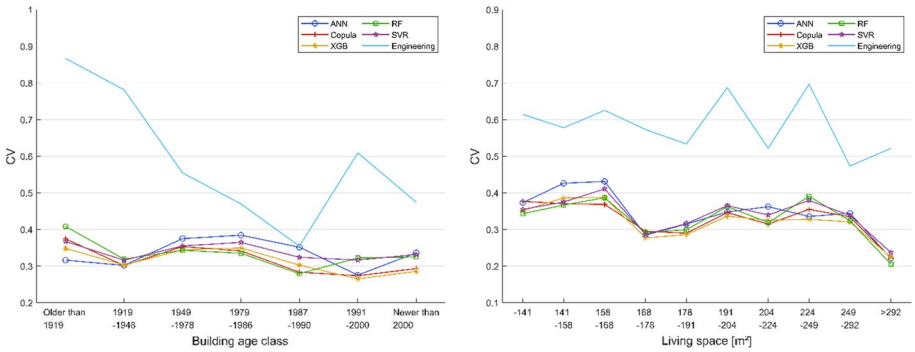

Fig. 8 Comparison of engineering EQMs with established data-driven EQMs regarding prediction performance measured with the coefficient of variation (CV) (Figure from Wenninger and Wiethe 2021)

existing empirical research, further empirical studies are needed for a complete evaluation, which we discuss in more detail in the next section.

Internal consistency: For internal consistency, we suppose that all sub-models—design model, measurement model, inner structural model—as well as the outer model must be consistent and coherent. To evaluate this criterion, we put ourselves in the role of a user of our theory (e.g., an empirical scientist). The research process will start with the selection of design candidates by selecting different design options, and the end is the measurement of the BEP prediction accuracy in the measurement model. By doing so, we could discover possible inconsistencies. After determining the design candidates, the inner structural model maps the interaction of data quality and model quality, which results in a BEP prediction. The measurement model then can be understood as a test mechanism for empirical validation of the influence of different design options on the BEP prediction accuracy. Based on the results in the measurement model, the process could be re-run with a further iterative loop, and design options could be selected in such a way that, for example, the BEP prediction accuracy is increased, or the influence of different options is investigated in the sense of a "feature importance analysis." For the inner structural model, we consider it to be consistent as long as our derived and defined assumptions hold. Using the mathematical derivation from "Characterizing output accuracy of the design candidates" section, we can describe the output accuracy of an EQM as a function of the uncertainty in data. This allows us to model the effect of design options on output accuracy and strengthens consistency. Regarding the design model, further empirical evidence is needed to assess consistency scientifically. This mainly concerns the theoretically optimal accuracy. In addition, there is little empirical evidence on how the auditor types cope with uncertainty in data. The same is likely true for the measurement model to be consistent with literature. Since the PEMs in the measurement model deterministically provide the same output for the same input, this sub model may also be considered consistent in itself. Nevertheless, from a process perspective, we consider the internal consistency as given to a large extent since all sub-models represent a continuous process.

Level of detail: The level of detail is relative to the current state of knowledge. Based on our literature analysis and best of our knowledge, there is no current attempt to synthesize the findings from empirical studies into a theory. For that reason, we argue that

even the simplest, lowest detailed theory can bring value to the scientific discourse and help guide EQM development. For the absence of the theory, evaluating completeness might be considered the more important evaluation criterion than the level of detail. Nonetheless, reading the detail, in this study, we describe the dimensions of the design candidates, the auditor types, and accuracy functions reflecting important constructs in the domain.

Robustness: Regarding robustness, our model is instance-agnostic and therefore generally applicable to any EQM. This, in turn, means that there are no specific instance problems, and we can first demonstrated robustness. In addition, our theory aims not to predict the location of the points of intersection. Instead, our theory demonstrates the existence of these as a function of data and model quality, influenced by the EQM and auditor type. This underlines that there is no dependence on an single instance so that high robustness can be assumed. Future research may empirically determine these points of intersection for different EQM and auditor types.

Summarizing, applying logical reasoning and evidence from existing research supports the validity of our proposed theory to a large extent. As pointed out, all sub-models and CERs can be empirically tested to further validate or falsify and refine the theory. Regardless of that, we discuss still vague or uncertain aspects found in the evaluation in the following section.

Discussion and implications

Our research attempts to explain why data-driven EQMs on simple data input can achieve higher levels of accuracy than engineering EQMs on data input from qualified auditors, as observed in empirical studies. This, as an artifact, is relevant because it provides a basis for conceiving and understanding a phenomenon potentially considered unreasonable before. In addition, this theory might inspire and guide the design of new and improved EQMs. In a similar vein, this theory intends to enable the informed use of data-driven EQMs. That might come in particularly handy given the ongoing policy-related discussion on EPCs, as well as the process and the tooling for issuing EPCs. Previously, there has been the conception that high-quality input data for EQMs can only exist when there is a considerable effort while collecting the data. However, data-driven EQMs exhibit some traits that help accommodate uncertainty in data on top of their ability to produce reliable outcomes with less specific/detailed information than needed for engineering EQMs.

In this regard, the theory suggests three implications that can guide the use of data-driven EQMs. First, data-driven EQMs can be a valid alternative to calculation-/simulation-based, i.e., engineering EQMs under some conditions. The theory in this research has underlined the reasons for this. Second, data-driven EQMs can differ largely in the amount of input they require, accuracy, and robustness against missing or anomalous data input. Therefore, choosing an appropriate machine learning algorithm is a relevant decision, similar to choosing an appropriate physical calculation/simulation model. Likewise, is the selection of variables to consider for the data-driven EQM an important issue. This is especially true when not only physical measures serve as variables. Third and inapplicable to engineering EQMs, it is important to train a data-driven EQM on data that will be most similar to the data applied to the data-driven EQM for inference.

For example, a data-driven EQM trained on data collected by a qualified auditor applying much tooling should not be considered for application by occupants, although on validation data, the data-driven EQM has performed very well. To that end, the theory suggests calibrating for uncertainty under realistic (work) environments. In this theory, we have considered archetypical auditor types. In reality, the capabilities of collecting data should be critically assessed and potentially revisited over time. In that sense, informed use of data-driven EQMs requires systematically identifying the building characteristics the type of auditor can reliably assess. As can be seen, the design and use of data-driven EQM are strongly coupled.

In this study, we especially stressed the role of CER1 to CER3 to link the design candidates to performance evaluation measures of prediction via an inner structural model. For CER1, understanding and the ability to measure variable importance help manage model quality for both engineering (via sensitivity analyses) and data-driven EQMs. Regarding CER2, data quality can be managed by expertise and physical presence, among others. CER3 is specific to data-driven EQMs and is key for a data-driven EQMs trait to somewhat accommodate for uncertainty in data.

Similar to the introduction of technology acceptance models (Davis 1989), which enjoy great popularity in the information technology domain, we derived the CERs from literature before empirically testing and validating these CERs jointly in a consecutive study. Discussing the general testability of the CERs is important already at this stage. For that reason, we present a testing approach for each CER:

- As outlined before, testing CER1 for engineering EQMs has already been carried out many times (Foucquier et al. 2013). A typical approach is to look at a series of sample buildings, where input data for the simplest but also the most advanced EQM are available. Then, we can perform the calculations of all engineering EQMs, which are to be compared. The accuracy rankings for each EQM should be stable over sample buildings to validate CER1, e.g., by rank correlation (Yilmaz et al. 2008). We can apply a similar test for the data-driven EQMs. However, for data-driven EQMs, an experiment needs to isolate effects related to CER3.
- Regarding CER2, for each engineering EQM, we can perform analytical and numerical analyses to study and evaluate how error terms in the inputs influence the outputs quite generally. For data-driven EQMs, we could use any trained model for inference on data describing the same buildings to test if it performs worse when (artificially added) distortion (noise) on the data increases. This test should consider various classes of machine learning algorithms, e.g., random forests, extreme gradient boost, Bayesian networks, and shallow and deep neural networks (Wenninger and Wiethe 2021).
- CER3 is relevant for data-driven EQMs, only. For testing purposes, we might provide a series of training datasets in decreasing levels of data quality on the same sufficient number of buildings. Any chosen machine learning algorithm should be applied consistently to train as many models as there are datasets. Testing the models on another sufficiently sized out-of-sample dataset should allow ordering the models according to the data quality of the data set they were trained. This might serve as a test of CER3.

From a further research point of view, there are open venues methodologically and regarding the adjacent (sub-) domains of the field. We subsequently suggest a three-pronged research agenda.

Testing, confirming, and falsifying the theory: As mentioned above, the scope of this research article allows for describing the explanatory theory and performing first validation based on reported findings from literature as well as logical reasoning, particularly with regard to the inner structural model. However, a theory needs to withstand the storm of testing. We have highlighted ways for further empirical validation of the proposed theory. Further research will be beneficial at testing all presented relationships. In particular, it will be interesting to conduct studies where empirical evidence is low at the moment. This particularly involves designing rigorous studies testing the slopes of the accuracy functions. Eventually, also if at some point, there is some evidence that this theory is flawed/falsified, from an epistemological point of view, the understanding of data-driven EQMs for socio-physical processes like issuing EPCs will benefit based on that scientific discourse.

Extending the scope and detail of the model: The presented explanatory model contributes particularly to giving a rationale for the reported surplus accuracy of data-driven EQMs. An apparent gap in these studies reporting surplus accuracy has made those EQMs more interpretable and eliminated potential pseudo-relationships and unfair biases. This is the purpose of explainable artificial intelligence (Wenninger et al. 2022a). It appears reasonable that CER3 can benefit from a thorough analysis in this regard. In addition, understanding these implications will give rise to incorporating the view on hybrid EQMs discussed in literature more recently.

Applying the concepts to adjacent domains: First, research in this field has mainly focused on one- and two-family homes. It will be of interest to research how well the presented conceptual analytic model generalizes over other building types: multi-family homes, condominiums, commercial buildings, and mixed-use buildings. Second, in this research, we have suggested that regional factors might play a role, such as in Ahlrichs et al. (2022). While this should not impact the general CERs, we may expect design candidates to dominate one type of auditor in some country/region while the same design candidate may be dominated by another elsewhere. This suggests conducting a cross-country study in the next step. Third, in this research, there is an emphasis on energy for heat loads. However, with the increasing intensity of sector-coupling, e.g., through heat pumps, alternative views on energy carriers and consumption may become prevalent. Fourth, with alternative building types or energy carriers, alternative time horizons become relevant. While in this study, we linked and compared studies typically researching annual consumption, this may be very different with alternative applications, e.g., on a daily or hourly basis. If this presented conceptual analytic model prevails over these time horizons will remain a question to be studied. Lastly, quantifying the savings from energetic retrofits continues to be a field of debate when it comes to identifying appropriate EQMs. This piece of research, though, might serve as a starting point for further considerations in that area.

Conclusion

In this study, we developed and discussed a theory on how data collection, the type of auditor, the method for building energy performance prediction, and its accuracy relate to one another. To overcome limitations of previous studies that only examined prediction accuracy for building energy performance, but where it was unclear which cause–effect relationships lead to the surplus accuracy for data-driven methods, we provide three important contributions. First, we presented a testable theory giving a well-grounded basis for why there is a good reason that data-driven methods on simple input data can outperform currently applied engineering methods with data gathered by qualified auditors. To that end we highlighted that data-driven methods compensate for difficulties in collecting data, e.g., due to lack of auditor expertise. Second, by our testable theory, we provide a design framework for future energy quantification methods. While considering different design options, such as different types of auditors and methods, we model cause–effect relationships, which can be tested in a measurement model using different performance evaluation measures to ensure validity and robustness. In that vein, we set up a research agenda and depict approaches that allow future research to test and validate/falsify our developed theory empirically. Third, as our findings come with practical caveats and limitations, we discuss and outline implications of designing data-driven energy quantification methods for practice and policymakers.

We conclude that analyzing the accuracy of building energy performance prediction methods as an outcome of a sociotechnical system is a complex task. However, data and model quality are the most relevant levers driving accuracy. We, therefore, argue that the development of energy quantification methods should be viewed as a holistic and interdisciplinary approach. It thus should not be limited to traditional engineering disciplines and enables the broad use of energy performance certificates for benchmarking building energy performance and providing retrofit recommendations. Our study is the first to provide theoretical insights on factors affecting building energy performance prediction accuracy and comes with its limitations. However, we are confident that it provides very relevant insights and implications for energy performance certificate design and thus for a successful heat transition in the important building sector.

Abbreviations

ANN	Artificial neural network
BEP	Building energy performance
CER	Cause–effect relationship
CV	Coefficient of variation
EPA	Energy performance assessment
EPC	Energy performance certificate
EQM	Energy quantification method
IS	Information system
MLA	Machine learning algorithm
PEM	Performance evaluation measure
RF	Random forest
SVM	Support vector machine
SVR	Support vector regression
XGB	Extreme gradient boosting

Acknowledgements

Supported by PayPal and the Luxembourg National Research Fund FNR (P17/IS/13342933/PayPal-FNR/Chair in DFS/Gilbert Fridgen).

Author contributions
LW, SW, CW: conceptualization, methodology, software, formal analysis, data curation, software, writing—original draft, writing—review and editing, visualization. GF: conceptualization, writing—review and editing, funding acquisition. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
The datasets generated and analyzed during the current study are not publicly available due contractual requirements.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 20 May 2022 Accepted: 8 June 2022
Published online: 17 June 2022

References

Ahlrichs J, Rockstuhl S, Tränkler T et al (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
Ahlrichs J, Wenninger S, Wiethe C et al (2022) Impact of socio-economic factors on local energetic retrofitting needs—a data analytics approach. Energy Policy 160:112646. https://doi.org/10.1016/j.enpol.2021.112646
Ali U, Shamsi MH, Bohacek M et al (2020) A data-driven approach for multi-scale GIS-based building energy modeling for analysis, planning and support decision making. Appl Energy 279:115834. https://doi.org/10.1016/j.apenergy.2020.115834
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
Andrade-Cabrera C, de Rosa M, Kathirgamanathan A et al (2018) A study on the trade-off between energy forecasting accuracy and computational complexity in lumped parameter building energy models. https://www.researchgate.net/publication/327562414_A_Study_on_the_Trade-off_between_Energy_Forecasting_Accuracy_and_Computational_Complexity_in_Lumped_Parameter_Building_Energy_Models. Accessed 04 Jan 2022
Arcipowska A, Anagnostopoulos F, Mariottini F et al. (2014) Energy performance certificates across the EU. https://bpie.eu/wp-content/uploads/2015/10/Energy-Performance-Certificates-EPC-across-the-EU.-A-mapping-of-national-approaches-2014.pdf. Accessed 04 Jan 2022
Berger J, Orlande HR, Mendes N et al (2016) Bayesian inference for estimating thermal properties of a historic building wall. Build Environ 106:327–339. https://doi.org/10.1016/j.buildenv.2016.06.037
Bevington PR (1969) Data reduction and error analysis for the physical sciences. McGraw Hill Book Co., New York
Bhattacherjee P (2004) Understanding changes in belief and attitude toward information technology usage: a theoretical model and longitudinal test. MIS Q 28:229. https://doi.org/10.2307/25148634
Bi J, Zhang T (2005) Support vector classification with input data uncertainty. In: Advances in neural information processing systems. pp 161–168
Bigalke U, Marcinek H (2016) Auswertung von Verbrauchskennwerten energieeffizienter Wohngebäude
Borgstein EH, Lamberts R, Hensen J (2016) Evaluating energy performance in non-domestic buildings: a review. Energy Build 128:734–755. https://doi.org/10.1016/j.enbuild.2016.07.018
Bourdeau M, Xq Z, Nefzaoui E et al (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
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
Calì D, Osterhage T, Streblow R et al (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
Claesson J (2011) CERBOF Projekt no. 72: Utfall och metodutvärdering av energideklaration av byggnader. https://www.researchgate.net/publication/237005861_CERBOF_Projekt_no_72_Utfall_och_metodutvardering_av_energideklaration_av_byggnader. Accessed 04 Jan 2022
Coakley D, Raftery P, Keane M (2014) A review of methods to match building energy simulation models to measured data. Renew Sustain Energy Rev 37:123–141. https://doi.org/10.1016/j.rser.2014.05.007
Davis FD (1985) A technology acceptance model for empirically testing new end-user information systems: theory and results. Ph.D. Thesis

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13:319. https://doi.org/10.2307/249008
de Wilde P (2014) The gap between predicted and measured energy performance of buildings: a framework for investigation. Autom Constr 41:40–49. https://doi.org/10.1016/j.autcon.2014.02.009
Deb C, Schlueter A (2021) Review of data-driven energy modelling techniques for building retrofit. Renew Sustain Energy Rev 144:110990. https://doi.org/10.1016/j.rser.2021.110990
Deutsche Energie-Agentur GmbH (2016) dena-Gebäudereport: Statistiken und Analysen zur Energieeffizienz im Gebäudebestand
Deutscher Bundestag (2013) Novelle der Energieeinsparverordnung und des Energieeinsparungsgesetzes
Doty DH, Glick WH (1994) Typologies as a unique form of theory building: toward improved understanding and modeling. Acad Manag Rev 19:230. https://doi.org/10.2307/258704
Droutsa KG, Kontoyiannidis S, Dascalaki EG et al (2016) Mapping the energy performance of hellenic residential buildings from EPC (energy performance certificate) data. Energy 98:284–295. https://doi.org/10.1016/j.energy.2015.12.137
Eicker U, Zirak M, Bartke N et al (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 Commission (2020) In focus: energy efficiency in buildings. https://ec.europa.eu/info/news/focus-energy-efficiency-buildings-2020-feb-17_en. Accessed 27 July 2021
European Commission (2021) Making our homes and buildings fit for a greener future. https://ec.europa.eu/commission/presscorner/api/files/attachment/869476/Buildings_Factsheet_EN_final.pdf.pdf. Accessed 27 July 2021
Fernandez I, Borges CE, Penya YK (2011) Efficient building load forecasting. In: ETFA2011. IEEE, pp 1–8
Fornasini P (2008) The uncertainty in physical measurements: an introduction to data analysis in the physics laboratory. Springer, New York
Foucquier A, Robert S, Suard F et al (2013) State of the art in building modelling and energy performances prediction: a review. Renew Sustain Energy Rev 23:272–288
Fox M, Goodhew S, de Wilde P (2016) Building defect detection: external versus internal thermography. Build Environ 105:317–331. https://doi.org/10.1016/j.buildenv.2016.06.011
German Energy Agency (2018) dena Concise building report: energy efficiency in the building stock—statistics and analyses
German Federal Ministry for Economic Affairs and Energy (2018) Energieeffizienz in Zahlen: Entwicklungen und Trends in Deutschland 2018
Gram-Hanssen K (2013) Efficient technologies or user behaviour, which is the more important when reducing households' energy consumption? Energ Effic 6:447–457. https://doi.org/10.1007/s12053-012-9184-4
Gregor S (2006) The nature of theory in information systems. MIS Q 30:611. https://doi.org/10.2307/25148742
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
Heo Y, Choudhary R, Augenbroe GA (2012) Calibration of building energy models for retrofit analysis under uncertainty. Energy Build 47:550–560
Herrando M, Cambra D, Navarro M et al (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
Hertle H, Duscha M, Eisenmann L et al (2005) Verbrauchs- oder Bedarfspass? Anforderungen an den Energiepass für Wohngebäude aus Sicht privater Käufer und Mieter
Kaiser M, Stirnweiß D, Wederhake L (2022) Hierarchische Eignungsprüfung von externen (open) data sets für unternehmensinterne analytics- und machine-learning-Projekte. HMD. https://doi.org/10.1365/s40702-022-00842-3
Kaymakci C, Wenninger S, Sauer A (2021) A holistic framework for AI systems in industrial applications. 16. Internationale Tagung Wirtschaftsinformatik 2021
Klobas JE (1995) Beyond information quality: fitness for purpose and electronic information resource use. J Inf Sci 21:95–114. https://doi.org/10.1177/016555159502100204
Lee AS (2001) Editor's comments. MIS Q
Li Y, Kubicki S, Guerriero A et al (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
Li Q, Ren P, Meng Q (2010) Prediction model of annual energy consumption of residential buildings. In: International conference on advances in energy engineering, pp 223–226
March ST, Smith GF (1995) Design and natural science research on information technology. Decis Support Syst 15:251–266. https://doi.org/10.1016/0167-9236(94)00041-2
Mathew PA, Dunn LN, Sohn MD et al (2015) Big-data for building energy performance: lessons from assembling a very large national database of building energy use. Appl Energy 140:85–93. https://doi.org/10.1016/j.apenergy.2014.11.042
Menezes AC, Cripps A, Bouchlaghem D et al (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
Niehaves B, Ortbach K (2016) The inner and the outer model in explanatory design theory: the case of designing electronic feedback systems. Eur J Inf Syst 25:303–316. https://doi.org/10.1057/ejis.2016.3
Papadopoulos S, Kontokosta CE (2019) Grading buildings on energy performance using city benchmarking data. Appl Energy 233–234:244–253. https://doi.org/10.1016/j.apenergy.2018.10.053
Pasichnyi O, Wallin J, Levihn F et al (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
Pipino LL, Lee YW, Wang RY (2002) Data quality assessment. Commun ACM 45:211–219

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
Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6:21–45. https://doi.org/10.1109/MCAS.2006.1688199
Pregenzer M, Flotzinger D, Pfurtscheller G (1994) Distinction sensitive learning vector quantisation-a new noise-insensitive classification method. In: The 1994 IEEE international conference on neural networks: IEEE World Congress on Computational Intelligence, June 27–June 29, 1994, Walt Disney World Dolphin Hotel, Orlando Florida. IEEE Neural Networks Council, New York, Piscataway, NJ, pp 2890–2894
Qiao Q, Yunusa-Kaltungo A, Edwards RE (2021) Towards developing a systematic knowledge trend for building energy consumption prediction. J Build Eng 35:101967. https://doi.org/10.1016/j.jobe.2020.101967
Rockstuhl S, Wenninger S, Wiethe C et al (2021) Understanding the risk perception of energy efficiency investments: investment perspective vs. energy bill perspective. Energy Policy 159:112616. https://doi.org/10.1016/j.enpol.2021.112616
Schwarz HR, Köckler N (2011) Numerische Mathematik, 8., aktualisierte Auflage. Vieweg+Teubner Verlag/Springer Fachmedien, Wiesbaden
Semple S, Jenkins D (2020) Variation of energy performance certificate assessments in the European Union. Energy Policy 137:111127. https://doi.org/10.1016/j.enpol.2019.111127
Sonnenberg C, Vom Brocke J (2011) Evaluation patterns for design science research artefacts. In: European design science symposium. Springer, pp 71–83
Strong DM, Lee YW, Wang RY (1997) Data quality in context. Commun ACM 40:103–110
Sutherland BR (2020) Driving data into energy-efficient buildings. Joule 4:2256–2258. https://doi.org/10.1016/j.joule.2020.10.017
The European Parliament and the Council of the European Union (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
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
Walter T, Price PN, Sohn MD (2014) Uncertainty estimation improves energy measurement and verification procedures. Appl Energy 130:230–236. https://doi.org/10.1016/j.apenergy.2014.05.030
Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 12:5–33
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
Watson RT, Boudreau MC, Chen AJ (2010) Information systems and environmentally sustainable development: energy informatics and new directions for the IS community. MIS Q 34:23. https://doi.org/10.2307/20721413
Wei Y, Zhang X, Shi Y et al (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
Wenninger S, Wiethe C (2021) Benchmarking energy quantification methods to predict heating energy performance of residential buildings in Germany. Bus Inf Syst Eng. https://doi.org/10.1007/s12599-021-00691-2
Wenninger S, Kaymakci C, Wiethe C (2022a) Explainable long-term building energy consumption prediction using QLattice. Appl Energy 308:118300. https://doi.org/10.1016/j.apenergy.2021.118300
Wenninger S, Kaymakci C, Wiethe C et al. (2022b) How sustainable is machine learning in energy applications? The sustainable machine learning balance sheet. In: 17th international conference on Wirtschaftsinformatik, Nürnberg, Germany
Yilmaz E, Aslam JA, Robertson S (2008) A new rank correlation coefficient for information retrieval. In: Chua T-S, Leong M-K, Myaeng SH et al (eds) Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval—SIGIR '08. ACM Press, New York, p 587
Yuan P, Duanmu L, Wang Z (2019) Coal consumption prediction model of space heating with feature selection for rural residences in severe cold area in China. Sustain Cities Soc 50:101643. https://doi.org/10.1016/j.scs.2019.101643
Zhao H, Magoulès F (2012a) 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
Zhao H, Magoulès F (2012b) Feature selection for predicting building energy consumption based on statistical learning method. J Algorithms Comput Technol 6:59–77. https://doi.org/10.1260/1748-3018.6.1.59

Publisher's Note